Journal of Machine Learning Research 22 (2021) 1-46 Submitted 12/19; Revised 10/20; Published 02/21

Incorporating Unlabeled Data into Distributionally-Robust Learning

Charlie Frogner FROGNER@MIT.EDU

Sebastian Claici SCLAICI@MIT.EDU

Edward Chien EDCHIEN@MIT.EDU

Justin Solomon JSOLOMON@MIT.EDU

Computer Science & Artificial Intelligence Laboratory (CSAIL)
Massachusetts Institute of Technology
Cambridge, MA 02139, USA

Editor: Arnak Dalalyan

Abstract

We study a robust alternative to empirical risk minimization called distributionally robust learning (DRL), in which one learns to perform against an adversary who can choose the data distribution from a specified set of distributions. We illustrate a problem with current DRL formulations, which rely on an overly broad definition of allowed distributions for the adversary, leading to learned classifiers that are unable to predict with any confidence. We propose a solution that incorporates unlabeled data into the DRL problem to further constrain the adversary. We show that this new formulation is tractable for stochastic gradient-based optimization and yields a computable guarantee on the future performance of the learned classifier, analogous to—but tighter than—guarantees from conventional DRL. We examine the performance of this new formulation on 14 real data sets and find that it often yields effective classifiers with nontrivial performance guarantees in situations where conventional DRL produces neither. Inspired by these results, we extend our DRL formulation to active learning with a novel, distributionally-robust version of the standard model-change heuristic. Our active learning algorithm often achieves superior learning performance to the original heuristic on real data sets.

Keywords: Distributionally robust optimization, Wasserstein distance, optimal transport, supervised learning, active learning

1. Introduction

Human learning is robust in ways that statistical learning struggles to replicate. Small changes to image pixel values and audio waveforms, for example, can dramatically alter the outputs of classifiers trained by conventional empirical risk minimization, while remaining imperceptible to human observers (Szegedy et al., 2013; Carlini and Wagner, 2018). Robustness to artificial and natural variations, however, is critical when learning systems are deployed "in the wild," such as in self-driving vehicles (Huval et al., 2015; Bojarski et al., 2016) and speech recognition systems (Junqua and Haton, 2012; Hannun et al., 2014). Hence, the design of robust learning techniques is a key focus of recent machine learning research (Eykholt et al., 2017; Madry et al., 2017; Raghunathan et al., 2018; Singh et al., 2018; Sinha et al., 2018; Cohen et al., 2019; Yuan et al., 2019).

Distributionally robust learning (DRL) (Delage and Ye, 2010; Abadeh et al., 2015; Chen and Paschalidis, 2018) offers an alternative to empirical risk minimization in which one learns to perform against an adversary who chooses the data distribution from a specified set of distributions. This approach offers several benefits, including robust performance with respect to perturbations of the data distribution and computable guarantees on the generalization of the learned model—provided the adversary's decision set includes the true data distribution.

The robustness guarantees offered by DRL rely on selection of the adversary's decision set; if the set does not include the true data distribution, the guarantees do not necessarily hold.[1] Most previous work has chosen the decision set to be a norm ball around the empirical distribution of the training data (Abadeh et al., 2015; Chen and Paschalidis, 2018; Esfahani and Kuhn, 2018; Sinha et al., 2018). As we show in Section 5.2, however, in many cases this ball must be extremely large to contain the true data distribution. As a result, the distributionally-robust learner attempts to be robust to an overly broad set of data distributions, preventing it from making a prediction with any confidence. As a result, it can do no better than assigning equal probability to all of the classes.

In this paper, we address the problem of overly-large decision sets by using unlabeled data to further constrain the adversary. In essence, we can remove from the decision set distributions that are unrealistic in the sense that their marginals in feature space do not resemble the unlabeled data. With a smaller decision set, the distributionally-robust learner can provide a tighter bound on the generalization performance, yielding nontrivial predictors with non-vacuous performance guarantees in situations where conventional DRL offers neither.

Our mechanism for optimizing against an adversary constrained by unlabeled data is general-purpose and applicable beyond supervised learning. We use this same mechanism to formulate a novel distributionally-robust method for active learning; this method frequently outperforms both uniform random sampling and standard methods for active learning.

2. Background

2.1 Notation

For any Polish space $\mathcal{S}$, we use $\mathcal{B}(\mathcal{S})$ to denote the associated Borel σ-algebra and $\mathcal{M}(\mathcal{S})$ to denote set of Radon measures on $(\mathcal{S}, \mathcal{B}(\mathcal{S}))$. $\mathcal{M}_+(\mathcal{S})$ is the set of nonnegative Radon measures on $(\mathcal{S}, \mathcal{B}(\mathcal{S}))$, and $\mathcal{Q}(\mathcal{S})$ is the set of probability measures: $\mathcal{Q}(\mathcal{S}) = \{\pi \in \mathcal{M}_+(\mathcal{S}) : \pi(\mathcal{S}) = 1\}$. $C_b(\mathcal{S})$ is the set of continuous, bounded functions from $\mathcal{S}$ into $\mathbb{R}$.

2.2 Statistical Learning

Let $\mathcal{X}$ be an input space and $\mathcal{Y}$ a label space, and let $\mathbb{P}$ be the true data distribution, a probability measure over $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$. We focus on a classification setting, in which $\mathcal{Y} = \{\mathbf{y}^k\}_{k=1}^{N_\mathcal{Y}}$ is a finite collection of discrete labels, while $\mathcal{X}$ can be any compact Polish space.

[1] Here, the guarantee we refer to is the bound on expected loss in the objective of the DRL problem. If the true data distribution is light-tailed, then generalization bounds can hold even for decision sets not containing the true distribution, due to a regularization effect of DRL (Shafieezadeh-Abadeh et al., 2019).

The learning problem chooses a hypothesis $h_\theta : \mathcal{X} \to \mathcal{Q}(\mathcal{Y})$, parameterized by $\theta \in \Theta \subseteq \mathbb{R}^q$, that minimizes the expected risk, $\mathbf{E}^{\mathbb{P}}\, \ell(h_\theta(X), Y)$, where $\ell : \mathcal{Q}(\mathcal{Y}) \times \mathcal{Y} \to \mathbb{R}$ is a loss function measuring deviation of the prediction $h_\theta(X)$ from the true label Y.[2]

We cannot directly evaluate the expected risk, however, since $\mathbb{P}$ is unknown. We instead have a labeled sample $\hat{\mathcal{Z}}_l = \{(\mathbf{x}_l^i, \mathbf{y}_l^i)\}_{i=1}^{N_l} \subset \mathcal{X} \times \mathcal{Y}$ consisting of N_l i.i.d. samples from $\mathbb{P}$. If $\hat{\mathbb{P}}_l = \frac{1}{N_l} \sum_{i=1}^{N_l} \delta_{(\mathbf{x}_l^i, \mathbf{y}_l^i)}$ is the empirical distribution of the labeled data, traditional empirical risk minimization substitutes $\hat{\mathbb{P}}_l$ for $\mathbb{P}$ in the statistical learning problem, solving

$$\underset{\theta \in \Theta}{\text{minimize}} \;\; \mathbf{E}^{\hat{\mathbb{P}}_l}\, \ell(h_\theta(X), Y) = \frac{1}{N_l} \sum_{i=1}^{N_l} \ell(h_\theta(\mathbf{x}_l^i), \mathbf{y}_l^i). \tag{1}$$

To reduce variance of this approximation and promote generalization, often a regularization term (e.g., penalizing model complexity) is added to the loss.

2.3 Distributional Robustness

Distributionally-robust learning (DRL) (Delage and Ye, 2010; Abadeh et al., 2015; Chen and Paschalidis, 2018) is an alternative to empirical risk minimization that attempts to learn a predictor with minimal worst-case expected risk, against an adversary who chooses the distribution of the data from a specified decision set $\mathcal{P}$:

$$\underset{\theta \in \Theta}{\text{minimize}} \sup_{\mu \in \mathcal{P}} \mathbf{E}^{\mu}\, \ell(h_\theta(X), Y). \tag{2}$$

$\mathcal{P}$ is typically a norm ball centered at the empirical distribution of the labeled data $\hat{\mathbb{P}}_l$. If $\mathcal{P}$ is chosen such that it contains the true data distribution $\mathbb{P}$, the objective in (2) upper-bounds the expected risk of the hypothesis.

In this paper, we focus on Wasserstein distributional robustness (Abadeh et al., 2015; Chen and Paschalidis, 2018), in which the adversary's decision set $\mathcal{P}$ is a norm ball with respect to the Wasserstein distance:

Definition 1 (Wasserstein distance) *Let $c : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}_+$ be a lower-semicontinuous cost function. For any $\mu, \nu \in \mathcal{Q}(\mathcal{Z})$, the Wasserstein distance between μ and ν is*

$$\mathcal{W}_c(\mu, \nu) = \inf_{\pi \in \Pi(\mu,\nu)} \int_{\mathcal{Z} \times \mathcal{Z}} c(\mathbf{z}, \mathbf{z}')\, \mathrm{d}\pi(\mathbf{z}, \mathbf{z}'), \tag{3}$$

with $\Pi(\mu, \nu) = \{\pi \in \mathcal{M}_+(\mathcal{Z} \times \mathcal{Z}) : \pi(A \times \mathcal{Z}) = \mu(A), \pi(\mathcal{Z} \times B) = \nu(B), \forall A, B \in \mathcal{B}(\mathcal{Z})\}$, i.e., the set of all joint distributions on $\mathcal{Z} \times \mathcal{Z}$ having marginals μ and ν. π is sometimes also called a "transportation plan" for moving the mass in μ to match ν.

The Wasserstein distance differs from other common divergences on probability measures, such as the KL divergence, in that it takes into account the geometry of the domain $\mathcal{Z}$, via the transport cost c. For this reason, it can compare measures with disjoint support, for example. We derive our theoretical results for a general cost c, but in our experiment, we choose c identically to the previous work on Wasserstein distributionally robust logistic regression (Abadeh et al., 2015), setting $c((\mathbf{x}, \mathbf{y}), (\mathbf{x}', \mathbf{y}')) = \|\mathbf{x} - \mathbf{x}'\|_p + \kappa|\mathbf{y} - \mathbf{y}'|$.

[2] We will always write (X, Y) for the pair of random variables $X : \Omega \to \mathcal{X}$ and $Y : \Omega \to \mathcal{Y}$ over which we are taking the expectation. Their distributions are to be understood from the context.

2.4 Related Work

Distributionally robust optimization (Calafiore and El Ghaoui, 2006) has been explored extensively beyond the learning setting, for a broad variety of objective functions and decision sets. Often decision sets are defined by moment or support conditions (Delage and Ye, 2010; Goh and Sim, 2010; Wiesemann et al., 2014) or divergences on probability measures such as the Prokhorov metric (Erdoğan and Iyengar, 2006) or f-divergences (Ben-Tal et al., 2013; Duchi et al., 2016; Namkoong and Duchi, 2016; Bertsimas et al., 2018; Miyato et al., 2015). Directional deviation conditions have also been explored (Chen et al., 2007). Kuhn et al. (2019) give a recent review of applications in machine learning.

Distributionally robust learning over a Wasserstein ball was proposed for logistic regression (Abadeh et al., 2015), regularized linear regression (Chen and Paschalidis, 2018), and more general losses (Gao and Kleywegt, 2016; Sinha et al., 2018; Dziugaite and Roy, 2017; Esfahani and Kuhn, 2018; Blanchet et al., 2019). An equivalence to regularization, under various assumptions on the loss, was shown by Gao et al. (2017); Shafieezadeh-Abadeh et al. (2019). Blanchet et al. (2019) additionally discuss the limitation of Wasserstein DRL that we describe in Sections 3 and 5, arising from the size of the Wasserstein ball required to encompass the true data distribution, suggesting alternative robustness criteria that are sufficient for generalization.

Several recent works have investigated the use of unlabeled data in training distributionally robust models (Chen et al., 2019; Najafi et al., 2019; Blanchet and Kang, 2020). Carmon et al. (2019) demonstrate that unlabeled data can bolster l_∞-robustness greatly. None has used unlabeled data to constrained the ambiguity set in the way that we propose. Blanchet and Kang (2016) propose to use labeled data to constrain the ambiguity set, in a similar spirit as the current work.

In Section 6, we discuss an application of the proposed method to active learning, which is a well-studied topic that has inspired a wide variety of algorithms (Yang and Loog, 2018). We focus on a class of heuristics that seek to maximize the change in the learned model resulting from obtaining a labeled example (Settles et al., 2008; Freytag et al., 2014; Cai et al., 2017).

3. Distributionally-Robust Learning with Unlabeled Data

3.1 A Problem with the Existing Approach

In the "medium-data" regime, where the labeled sample may be far from the true data distribution $\mathbb{P}$ with respect to Wasserstein distance, Wasserstein distributionally-robust learning suffers from imprecision of the decision set $\mathcal{P}$, which is a Wasserstein ball centered at the empirical distribution of the labeled sample. The volume of this ball grows rapidly in its radius, requiring the learner to be robust to an enormous variety of data distributions. This problem manifests as low confidence of the distributionally robust learner, even when the radius ε is chosen to be much smaller than the true distance to the data distribution—thereby foregoing the performance guarantee implied by (2).

Figure 1 shows an example; additional illustrations and model details are in Section 5.2. We train a Wasserstein distributionally robust logistic regression model using 20 labeled samples from the Wisconsin breast cancer data set (Dua and Graff, 2019).

We plot both the test set likelihood and the maximum confidence of the learner over input samples as the radius ε of the decision set is varied.[3] We see that the confidence goes to 0.5—i.e., the classes are assigned equal probability for all test set samples—at a radius much smaller than the distance to the empirical distribution of the test set. Notably, the radius that maximizes the likelihood is approximately 1% of the distance to the test distribution. This maximum is often suggested as an appropriate target when choosing the radius ε in practice, as we will discuss in Section 5.1.

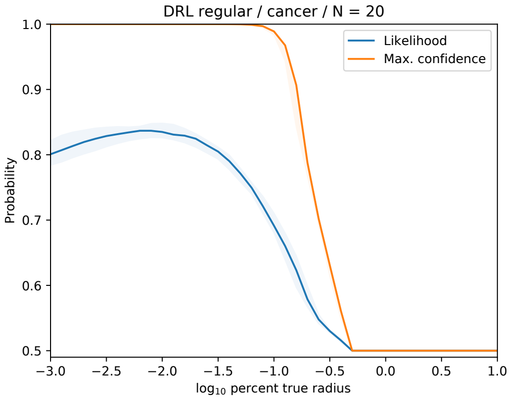


Figure 1: Wasserstein distributionally robust learning yields a no-confidence predictor at radius ε much smaller than the distance to the true data distribution.

3.2 Constraining the Adversary Using Unlabeled Data

We propose to deal with the overwhelming size of the decision set by constraining it further, pruning unrealistic potential data distributions while still allowing the set to contain the true data distribution. Specifically, we intersect two additional constraints with the Wasserstein ball.

The first constraint uses unlabeled data to constrain the marginal in $\mathcal{X}$ of the data distribution. As is common in many learning settings, we assume that unlabeled data are acquired much more readily than labeled data, giving the learner access to large set of unlabeled examples. Let $\mathbb{P}_{\mathcal{X}}$ be the **$\mathcal{X}$-marginal**, defined by $\mathbb{P}_{\mathcal{X}}(A) = \mathbb{P}(A \times \mathcal{Y})$ for all Borel subsets $A \in \mathcal{B}(\mathcal{X})$. Then our **unlabeled data** is a set $\hat{\mathcal{X}}_u \subset \mathcal{X}$ drawn i.i.d. from $\mathbb{P}_{\mathcal{X}}$.

The second constraint restricts the **$\mathcal{Y}$-marginal** of the data distribution, by defining intervals on the individual label probabilities. Let $\mathbb{P}_{\mathcal{Y}}$ be the $\mathcal{Y}$-marginal, which in a classification setting is discrete $\mathbb{P}_{\mathcal{Y}} = \sum_{k=1}^{N_{\mathcal{Y}}} \mathbf{p}_{\mathcal{Y}}^k \delta_{\mathbf{y}^k}$ for $\mathcal{Y} = \{\mathbf{y}^k\}_{k=1}^{N_{\mathcal{Y}}}$ the set of labels and $\mathbf{p}_{\mathcal{Y}}^k$ the corresponding label probabilities. The **interval** for each label is $[\underline{\mathbf{p}}_{\mathcal{Y}}^k, \overline{\mathbf{p}}_{\mathcal{Y}}^k]$. These interval constraints might come from prior knowledge, another data set as in the ecological inference setting (King, 2013; Frogner and Poggio, 2019), or directly from the training data, as described in Section 5.2.

3.3 Problem Formulation and Duality

If we restrict the decision set as described in Section 3.2, we need to establish that the distributionally robust learning problem is still tractable, particularly since one of the constraints we have added is infinite-dimensional. Recall that $\mathbb{P}_{\mathcal{X}}$ is the marginal of the unlabeled data on the feature space, and that $\underline{p}_{\mathcal{Y}}$ and $\overline{p}_{\mathcal{Y}}$ are the lower and upper bounds on the marginal on the label.

[3] The confidence of a hypothesis h_θ at a point $\mathbf{x} \in \mathcal{X}$ we define by $\max\{h_\theta(\mathbf{x}), 1 - h_\theta(\mathbf{x})\}$.

$\mathcal{X}$	Feature space
$\mathcal{Y} = \{\mathbf{y}^k\}_{k=1}^{N_\mathcal{Y}}$	Label space
$\mathcal{Z}$	$\mathcal{X} \times \mathcal{Y}$
h_θ	Hypothesis function, parameterized by θ
ℓ	Loss function
$\mathbb{P}$	True data distribution (over $\mathcal{Z}$)
$\mathbb{P}_\mathcal{X}$	$\mathcal{X}$-marginal of $\mathbb{P}$
$\mathbb{P}_\mathcal{Y}$	$\mathcal{Y}$-marginal of $\mathbb{P}$
$\hat{\mathbb{P}}_l$	Labeled data distribution (over $\mathcal{Z}$)
$\mathcal{B}_\varepsilon(\hat{\mathbb{P}}_l)$	Wasserstein ball of radius ε about $\hat{\mathbb{P}}_l$
$\overline{\mathbf{p}}_\mathcal{Y}^k$ ($\underline{\mathbf{p}}_\mathcal{Y}^k$)	Upper (lower) bound on marginal probability of $\mathbf{y}^k$
$\mathcal{U}(\mathbb{P}_\mathcal{X}, \underline{\mathbf{p}}_\mathcal{Y}, \overline{\mathbf{p}}_\mathcal{Y})$	Set of probability measures $\pi \in \mathcal{Q}(\mathcal{Z} \times \mathcal{Z})$ whose first $\mathcal{X}$-marginal is $\mathbb{P}_\mathcal{X}$ and first $\mathcal{Y}$-marginal satisfies $\pi((\mathcal{X} \times \{\mathbf{y}^k\}) \times \mathcal{Z}) \in [\underline{\mathbf{p}}_\mathcal{Y}^k, \overline{\mathbf{p}}_\mathcal{Y}^k]$, $\forall k$

Table 1: Notation.

We can define a set of possible joint distributions on $\mathcal{X} \times \mathcal{Y}$ that are consistent with this data,

$$\begin{aligned}\mathcal{U}(\mathbb{P}_\mathcal{X}, \underline{\mathbf{p}}_\mathcal{Y}, \overline{\mathbf{p}}_\mathcal{Y}) = \Big\{\mathbb{P} \in \mathcal{M}_+(\mathcal{X} \times \mathcal{Y}) : \;& \mathbb{P}(A \times \mathcal{Y}) = \mathbb{P}_\mathcal{X}(A), \\ & \mathbb{P}(\mathcal{X} \times B) \in [\underline{\mathbb{P}}_\mathcal{Y}(B), \overline{\mathbb{P}}_\mathcal{Y}(B)], \\ & \forall A \in \mathcal{B}(\mathcal{X}), B \subseteq \mathcal{Y}\Big\},\end{aligned} \tag{4}$$

with $\underline{\mathbb{P}}_\mathcal{Y} = \sum_{k=1}^{N_\mathcal{Y}} \underline{\mathbf{p}}_\mathcal{Y}^k \delta_{\mathbf{y}^k}$ and $\overline{\mathbb{P}}_\mathcal{Y} = \sum_{k=1}^{N_\mathcal{Y}} \overline{\mathbf{p}}_\mathcal{Y}^k \delta_{\mathbf{y}^k}$.

Suppose in addition we observe labeled data $\hat{\mathcal{Z}}_l = \{\mathbf{z}_\ell^i\}_{i=1}^{N_l} \subset \mathcal{X} \times \mathcal{Y}$, with $\mathbf{z}_\ell^i = (\mathbf{x}_l^i, \mathbf{y}_\ell^i)$, that define the empirical distribution $\hat{\mathbb{P}}_l = \frac{1}{N_l} \sum_{i=1}^{N_l} \delta_{\mathbf{z}_\ell^i}$. We define the adversary's decision set to be the intersection of the set of distributions $\mathcal{U}(\mathbb{P}_\mathcal{X}, \underline{\mathbf{p}}_\mathcal{Y}, \overline{\mathbf{p}}_\mathcal{Y})$ with a Wasserstein ball of radius ε around the empirical distribution $\hat{\mathbb{P}}_l$:

$$\mathcal{P} = \mathcal{U}(\mathbb{P}_\mathcal{X}, \underline{\mathbf{p}}_\mathcal{Y}, \overline{\mathbf{p}}_\mathcal{Y}) \cap \mathcal{B}_\varepsilon(\hat{\mathbb{P}}_l), \tag{5}$$

where $\mathcal{B}_\varepsilon(\hat{\mathbb{P}}_l) = \{\mu : \mathcal{W}_c(\mu, \hat{\mathbb{P}}_l) \leq \varepsilon\}$. Thus, our feasible set contains all distributions that have the correct data and label marginals (and are thus in $\mathcal{U}(\mathbb{P}_\mathcal{X}, \underline{\mathbf{p}}_\mathcal{Y}, \overline{\mathbf{p}}_\mathcal{Y})$), but which are also close to the known labeled distribution (and thus contained in $\mathcal{B}_\varepsilon(\hat{\mathbb{P}}_l)$).

The resulting distributionally-robust problem is defined identically to (2), using this decision set $\mathcal{P}$. The inner problem with fixed θ is that of evaluating a **worst-case expected loss**

$$f(\theta) = \sup_{\mu \in \mathcal{P}} \mathbf{E}^\mu\, \ell(h_\theta(X), Y). \tag{6}$$

For marginals $\mathbb{P}_\mathcal{X}$ with infinite support this an infinite-dimensional linear program as the marginal on $\mathcal{X}$ of the solution μ must contain the support of $\mathbb{P}_\mathcal{X}$.

We can rewrite (6) by casting it as an optimal transportation problem over the space $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ between our unknown distribution μ and the given data distribution such that the transport plan π satisfies the marginal constraints on μ:

$$f(\theta) = \begin{cases} \begin{array}{rll} \sup_{\pi \in \mathcal{M}(\mathcal{Z}\times\mathcal{Z})} & \int_{(\mathcal{X}\times\mathcal{Y})\times\mathcal{Z}} \ell(h_\theta(\mathbf{x}), \mathbf{y}) \, \mathrm{d}\pi((\mathbf{x},\mathbf{y}),\mathbf{z}') & \\ \text{s.t.} & \int_{\mathcal{Z}\times\mathcal{Z}} c(\mathbf{z},\mathbf{z}') \, \mathrm{d}\pi(\mathbf{z},\mathbf{z}') \leq \varepsilon & \\ & \pi(\mathcal{Z}, \mathbf{z}_\ell^i) = \frac{1}{N_l} & \forall i \in \{1,\ldots,N_l\}, \\ & \pi((A \times \mathcal{Y}) \times \mathcal{Z}) = \mathbb{P}_\mathcal{X}(A) & \forall A \in \mathcal{B}(\mathcal{X}), \\ & \pi((\mathcal{X},\mathbf{y}^k),\mathcal{Z}) \leq \overline{\mathbf{p}}_\mathcal{Y}^k & \forall k \in \{1,\ldots,N_\mathcal{Y}\}, \\ & \pi((\mathcal{X},\mathbf{y}^k),\mathcal{Z}) \geq \underline{\mathbf{p}}_\mathcal{Y}^k & \forall k \in \{1,\ldots,N_\mathcal{Y}\}, \\ & \pi(A) \geq 0 & \forall A \in \mathcal{B}(\mathcal{Z}\times\mathcal{Z}). \end{array} \end{cases} \tag{7}$$

Here the variable $\mathbf{z} = (\mathbf{x}, \mathbf{y})$ indexes the support of the worst-case measure while $\mathbf{z}'$ indexes the support of $\hat{\mathbb{P}}_l$. π is a transport plan that joins these two measures. Observe that only the constraint on the $\mathcal{X}$ marginal is infinite dimensional. We will show that this constraint corresponds in the dual problem to an expectation under $\mathbb{P}_\mathcal{X}$ of a finite dimensional cost.

While the program (6) is infinite dimensional, its dual is a problem in finite dimensions:

$$g(\theta) = \begin{cases} \inf_{\alpha,\beta,\underline{\lambda},\overline{\lambda}} \alpha\varepsilon + \frac{1}{N_l}\sum_{i=1}^{N_l} \beta^i + \sum_{k=1}^{N_\mathcal{Y}} \left(\overline{\lambda}^k \overline{\mathbf{p}}_\mathcal{Y}^k - \underline{\lambda}^k \underline{\mathbf{p}}_\mathcal{Y}^k\right) \\ \qquad + \mathbf{E}^{\mathbb{P}_\mathcal{X}} \left[\max_{\substack{k \in \{1,\ldots,N_\mathcal{Y}\} \\ i \in \{1,\ldots,N_l\}}} \ell(h_\theta, (X, \mathbf{y}^k)) - \left(\alpha c\left((X,\mathbf{y}^k), \mathbf{z}_\ell^i\right) + \beta^i\right) - (\overline{\lambda}^k - \underline{\lambda}^k) \right] \\ \text{s.t. } \alpha, \underline{\lambda}^k, \overline{\lambda}^k \geq 0 \quad \forall k \in \{1,\ldots,N_\mathcal{Y}\}. \end{cases} \tag{8}$$

Here $\alpha \in \mathbb{R}$, $\beta \in \mathbb{R}^{N_\ell}$, and $\underline{\lambda}, \overline{\lambda} \in \mathbb{R}^{N_\mathcal{Y}}$. Each of these dual variables corresponds to a primal constraint: α corresponds to the constraint on the transport cost, β the constraint that the second marginal be $\hat{\mathbb{P}}_l$, $\underline{\lambda}$ the lower bound on the worst-case label probabilities, and $\overline{\lambda}$ the upper bound. The infinite-dimensional constraint that the first marginal of the primal transport plan have $\mathcal{X}$-marginal $\mathbb{P}_\mathcal{X}$ corresponds here to the expectation in the objective. This correspondence is established in much more detail in the proof of Theorem 2 (Appendix A), which shows that the two problems $g(\theta)$ and $f(\theta)$ are in fact equivalent.

We state our main theoretical result, whose proof is deferred to Appendix A:

Theorem 2 (Strong duality) *Let $\mathcal{X}$ be a compact Polish space and $\mathcal{Y} = \{\mathbf{y}^k\}_{k=1}^{N_\mathcal{Y}}$ any finite set. Let $\mathbb{P}_\mathcal{X}$ be a probability measure over $\mathcal{X}$ and $\hat{\mathbb{P}}_l = \frac{1}{N_l}\sum_{i=1}^{N_l} \delta_{\mathbf{z}_\ell^i}$ an empirical probability measure over $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, and define intervals $[\underline{\mathbf{p}}_\mathcal{Y}^k, \overline{\mathbf{p}}_\mathcal{Y}^k] \subseteq [0,1]$, $k \in \{1,\ldots,N_\mathcal{Y}\}$. Let the transportation cost $c : \mathcal{Z} \times \mathcal{Z} \to [0,+\infty)$ be nonnegative and upper semicontinuous with $c(\mathbf{z},\mathbf{z}') = 0 \Leftrightarrow \mathbf{z} = \mathbf{z}'$. Assume $\ell(h_\theta(\cdot),\cdot) : \mathcal{Z} \to \mathbb{R}$ is upper semicontinuous. Define f as in* (6) *and g as in* (8). *If $\mathcal{U}(\mathbb{P}_\mathcal{X}, \underline{\mathbf{p}}_\mathcal{Y}, \overline{\mathbf{p}}_\mathcal{Y}) \cap \mathcal{B}_\varepsilon(\hat{\mathbb{P}}_l) \neq \emptyset$, then*

$$f(\theta) = g(\theta), \quad \forall \theta \in \Theta. \tag{9}$$

Furthermore, if $\operatorname{relint}(\mathcal{U}(\mathbb{P}_\mathcal{X}, \underline{\mathbf{p}}_\mathcal{Y}, \overline{\mathbf{p}}_\mathcal{Y}) \cap \mathcal{B}_\varepsilon(\hat{\mathbb{P}}_l)) \neq \emptyset$, then there exists a minimizer $(\alpha_*, \beta_*, \overline{\lambda}_*, \underline{\lambda}_*) \in \mathbb{R}_+ \times \mathbb{R}^{N_l} \times \mathbb{R}_+^{N_\mathcal{Y}} \times \mathbb{R}_+^{N_\mathcal{Y}}$ attaining the infimum in (8).

The take-away message of Theorem 2 is that distributionally-robust learning under the model proposed here amounts to minimizing g with respect to θ, a finite dimensional problem that can be tackled with stochastic gradient approaches.

We make three comments regarding Theorem 2. First, the Theorem assumes compact $\mathcal{X}$, convenient in several steps of the proof. The theorem remains practically relevant in the sense that truly unbounded data distributions are rare, due to physical constraints. We nevertheless expect the theorem continues to hold even with a noncompact $\mathcal{X}$; this is an area for future work. Second, existence of the minimizer in Theorem 2 requires the relative interior of the ambiguity set to be nonempty. This condition is well-defined with respect to the weak topology on the space of measures.

Third, it is common in DRL literature to use a cost $c((\mathbf{x},\mathbf{y}),(\mathbf{x}',\mathbf{y}')) = \|\mathbf{x}-\mathbf{x}'\|_p+\kappa|\mathbf{y}-\mathbf{y}'|$, setting $\kappa = +\infty$ to restrict the underlying optimal transport to move mass in $\mathcal{X}$ while leaving the corresponding labels fixed. Our strong duality result, however, in proving existence of a dual minimizer relies on boundedness of c on a compact domain $\mathcal{X} \times \mathcal{Y}$, which precludes the $\kappa = +\infty$ setting. This is as far as we know an artifact of the proof and not a fundamental limitation. We note also that κ can be arbitrarily large and Theorem 2 will still hold.

3.3.1 Relationship to Conventional DRL

The problem (6) is distinct from Wasserstein DRL in that we assume access to additional information: the marginal $\mathbb{P}_\mathcal{X}$ and bounds on $\mathbb{P}_\mathcal{Y}$. We note two properties of our formulation:

1. We will show in Section 4.1 that it suffices to access only samples from the marginal $\mathbb{P}_\mathcal{X}$. This is unlabeled data, meaning that our formulation in fact applies in a semi-supervised setting, rather than fully-supervised like conventional DRL.

2. This additional information constrains the ambiguity set further than in regular DRL, entirely explaining any performance difference. We will find in Section 5 that the performance guarantees obtained with the additional information are stronger than those obtained without it.

4. Algorithm and Analysis

4.1 Optimization by SGD

Problem (8) is a convex, finite dimensional optimization problem in $\alpha, \beta, \underline{\lambda}, \overline{\lambda}$ that is the sum of a linear term and an expectation under $\mathbb{P}_\mathcal{X}$. To apply stochastic gradient descent, we first need to compute derivatives under the variables $\alpha, \beta, \underline{\lambda}, \overline{\lambda}$.

We first compute derivatives of the term under the expectation. Define $\Phi^{i,k}(\cdot;\theta,\alpha,\beta,\underline{\lambda},\overline{\lambda})$ as the function

$$\Phi^{i,k}(\mathbf{x};\theta,\alpha,\beta,\underline{\lambda},\overline{\lambda}) = \ell(h_\theta,(X,\mathbf{y}^k)) - \left(\alpha c\left((X,\mathbf{y}^k),\mathbf{z}^i_\ell\right) + \beta^i\right) - (\overline{\lambda}^k - \underline{\lambda}^k).$$

The dual objective can be expressed as a function of $\theta, \alpha, \beta, \underline{\lambda}, \overline{\lambda}$ as

$$\alpha\varepsilon + \frac{1}{N_l}\sum_{i=1}^{N_l}\beta^i + \sum_{k=1}^{N_\mathcal{Y}}\left(\overline{\lambda}^k\overline{\mathbf{p}}^k_\mathcal{Y} - \underline{\lambda}^k\underline{\mathbf{p}}^k_\mathcal{Y}\right) + \mathbf{E}^{\mathbb{P}_\mathcal{X}}\left[\mathrm{max}_{\substack{k\in\{1,\ldots,N_\mathcal{Y}\}\\ i\in\{1,\ldots,N_l\}}}\Phi^{i,k}(\mathbf{x};\theta,\alpha,\beta,\underline{\lambda},\overline{\lambda})\right].$$

For a given choice of $i = i_0$ and $k = k_0$, there is a set of points $\mathbf{x}$ for which $\Phi^{i,k}(\mathbf{x}; \theta, \alpha, \beta, \underline{\lambda}, \overline{\lambda})$ is maximized at $(i, k) = (i_0, k_0)$. These points define a subset $V^{ik} \subseteq \mathcal{X}$,

$$V^{ik} = \left\{ \mathbf{x} \in \mathcal{X} : \Phi^{i,k}(\mathbf{x}; \theta, \alpha, \beta, \underline{\lambda}, \overline{\lambda}) \geq \Phi^{i',k'}(\mathbf{x}; \theta, \alpha, \beta, \underline{\lambda}, \overline{\lambda}) \; \forall \; i', k' \right\}. \tag{10}$$

The sets V^{ik} partition $\mathcal{X}$, up to boundary points where the sets meet one another. We can decompose the expectation above as a finite sum of integrals over domains V^{ik}, i.e.

$$\mathbf{E}^{\mathbb{P}_{\mathcal{X}}} \left[\max_{\substack{k \in \{1,\ldots,N_{\mathcal{Y}}\} \\ i \in \{1,\ldots,N_l\}}} \Phi^{i,k}(\mathbf{x}; \theta, \alpha, \beta, \underline{\lambda}, \overline{\lambda}) \right] = \sum_{i=1}^{N_l} \sum_{k=1}^{N_{\mathcal{Y}}} \int_{V^{ik}} \Phi^{i,k}(\mathbf{x}; \theta, \alpha, \beta, \underline{\lambda}, \overline{\lambda}) \, \mathrm{d}\mathbb{P}_{\mathcal{X}}(\mathbf{x}). \tag{11}$$

Note that V^{ik} changes depending on the parameters $(\theta, \alpha, \beta, \overline{\lambda}, \underline{\lambda})$. To evaluate a derivative with respect to one of these parameters, then, we need to differentiate under the integral sign. Applying Reynolds' Transport Theorem, we obtain that

$$\frac{\partial}{\partial \alpha} \sum_{i=1}^{N_l} \sum_{k=1}^{N_{\mathcal{Y}}} \int_{V^{ik}} \Phi^{i,k}(\mathbf{x}; \theta, \alpha, \beta, \underline{\lambda}, \overline{\lambda}) \, \mathrm{d}\mathbb{P}_{\mathcal{X}}(\mathbf{x}) = \sum_{i=1}^{N_l} \sum_{k=1}^{N_{\mathcal{Y}}} \int_{V^{ik}} \frac{\partial}{\partial \alpha} \Phi^{i,k}(\mathbf{x}; \theta, \alpha, \beta, \underline{\lambda}, \overline{\lambda}) \, \mathrm{d}\mathbb{P}_{\mathcal{X}}(\mathbf{x}), \tag{12}$$

and the same holds for the other parameters $\theta, \beta, \overline{\lambda}, \underline{\lambda}$. Reynolds' Theorem also specifies terms that are boundary integrals for the boundaries of the sets V^{ik}. In our case, these terms sum to zero, as almost every boundary point $\mathbf{x} \in \operatorname{int} \mathcal{X}$ is shared between exactly two sets V^{ik}, $V^{i'k'}$ and the integrands at $\mathbf{x}$ for the corresponding boundary integrals exactly cancel.

The exact forms for the derivatives of the dual objective are given in Appendix B. Although existence of subgradients is not required for Theorem 2, subgradients of $\Phi^{i,k}$ exist so long as $\ell \circ h_\theta$ is subdifferentiable in θ. To simplify notation further, we define

$$\Phi(\mathbf{x}; \theta, \alpha, \beta, \underline{\lambda}, \overline{\lambda}) = \max_{\substack{k \in \{1,\ldots,N_{\mathcal{Y}}\} \\ i \in \{1,\ldots,N_l\}}} \Phi^{i,k}(\mathbf{x}; \theta, \alpha, \beta, \underline{\lambda}, \overline{\lambda}). \tag{13}$$

We can optimize for the optimal dual parameters $\theta, \alpha, \beta, \underline{\lambda}, \overline{\lambda}$ by sampling $\mathbf{x}^1, \ldots, \mathbf{x}^{N_b}$ from $\mathbb{P}_{\mathcal{X}}$ and computing gradients of $\Phi(\mathbf{x}^j; \theta, \alpha, \beta, \underline{\lambda}, \overline{\lambda})$ with respect to the dual variables. These gradients define a stochastic gradient descent, subject to nonnegativity constraints on $\alpha, \underline{\lambda}$, and $\overline{\lambda}$. This approach is summarized in Algorithm 1.

4.2 Approximating the Performance Bound

An attractive feature of traditional Wasserstein DRL is that the optimal value of the objective upper-bounds the true expected risk $\mathbf{E}^{\mathbb{P}} \, \ell(h_\theta(X), Y)$, provided that the adversary's decision set contains the true data distribution $\mathbb{P}$.

The proposed formulation using unlabeled data provides a similar guarantee. By weak duality (Theorem 2), so long as $\mathbb{P} \in \mathcal{P}$, we have the following bound for all $\theta \in \Theta$, $(\alpha, \beta, \overline{\lambda}, \underline{\lambda}) \in \mathbb{R}_+ \times \mathbb{R}^{N_\ell} \times \mathbb{R}_+^{N_{\mathcal{Y}}} \times \mathbb{R}_+^{N_{\mathcal{Y}}}$.

$$\mathbf{E}^{\mathbb{P}} \, \ell(h_\theta(X), Y) \leq \alpha \varepsilon + \frac{1}{N_l} \sum_{i=1}^{N_l} \beta^i + \sum_{k=1}^{N_{\mathcal{Y}}} \left(\overline{\lambda}^k \overline{\mathbf{p}}_{\mathcal{Y}}^k - \underline{\lambda}^k \underline{\mathbf{p}}_{\mathcal{Y}}^k \right) + \mathbf{E}^{\mathbb{P}_{\mathcal{X}}} \left[\Phi(X) \right], \tag{14}$$

Algorithm 1 SGD for distributionally robust learning with unlabeled data

```
**Given**: $\varepsilon \geq 0$, $\overline{\mathbf{p}}_{\mathcal{Y}}, \underline{\mathbf{p}}_{\mathcal{Y}} \in [0,1]^{N_{\mathcal{Y}}}$, $\theta_0 \in \Theta$, step size $\gamma > 0$, batch size $N_b$.
$\theta \leftarrow \theta_0$, $\alpha, \beta, \overline{\lambda}, \underline{\lambda} \leftarrow \mathbf{0}$.
**while** not converged **do**
    # {Computation of subgradients of $\Phi$ is described in Appendix B.}
    Sample $\mathbf{x}^1, \ldots, \mathbf{x}^{N_b} \sim \mathbb{P}_{\mathcal{X}}$.
    $\theta \leftarrow \mathrm{Proj}_{\Theta}\left[\theta - \frac{\gamma}{N_b}\sum_{j=1}^{N_b} \nabla_\theta \Phi(\mathbf{x}^j; \theta, \alpha, \beta, \underline{\lambda}, \overline{\lambda})\right]$.
    $\alpha \leftarrow \max\left(0, \alpha - \gamma\left[\varepsilon + \frac{1}{N_b}\sum_{j=1}^{N_b} \nabla_\alpha \Phi(\mathbf{x}^j; \theta, \alpha, \beta, \underline{\lambda}, \overline{\lambda})\right]\right)$.
    $\beta \leftarrow \beta - \gamma\left[\frac{1}{N_l} + \frac{1}{N_b}\sum_{j=1}^{N_b} \nabla_\beta \Phi(\mathbf{x}^j; \theta, \alpha, \beta, \underline{\lambda}, \overline{\lambda})\right]$.
    $\overline{\lambda} \leftarrow \max\left(\mathbf{0}, \overline{\lambda} - \gamma\left[\overline{\mathbf{p}}_{\mathcal{Y}} + \frac{1}{N_b}\sum_{j=1}^{N_b} \nabla_{\overline{\lambda}} \Phi(\mathbf{x}^j; \theta, \alpha, \beta, \underline{\lambda}, \overline{\lambda})\right]\right)$.
    $\underline{\lambda} \leftarrow \max\left(\mathbf{0}, \underline{\lambda} - \gamma\left[-\underline{\mathbf{p}}_{\mathcal{Y}} + \frac{1}{N_b}\sum_{j=1}^{N_b} \nabla_{\underline{\lambda}} \Phi(\mathbf{x}^j; \theta, \alpha, \beta, \underline{\lambda}, \overline{\lambda})\right]\right)$.
**end while**
```

with $\Phi(\mathbf{x}) \triangleq \Phi(\mathbf{x}; \theta, \alpha, \beta, \underline{\lambda}, \overline{\lambda})$ from Section 4.1. For any fixed θ, α, β, $\overline{\lambda}$, $\underline{\lambda}$ we can approximate this bound by sampling from $\mathbb{P}_{\mathcal{X}}$ and replacing the expectation $\mathbf{E}^{\mathbb{P}_{\mathcal{X}}}\,\Phi(X)$ with a sample mean. When $(\alpha, \beta, \overline{\lambda}, \underline{\lambda})$ are optimal, this gives a concrete numerical estimate of the worst-case performance of h_θ. This error of this approximation can be bounded with high probability under a variety of assumptions. Assuming the random variable $\Phi(X)$ has finite variance (for instance in linear logistic regression with c the Euclidean distance in $\mathcal{X}$ and $\mathbb{P}_{\mathcal{X}}$ sub-Gaussian), for example, Chebyshev's inequality gives a tail bound that decreases with the number of unlabeled samples,

$$\Pr\left(\mathbf{E}^{\mathbb{P}_{\mathcal{X}}}\,\Phi(X) \geq \mathbf{E}^{\hat{\mathbb{P}}_{\mathcal{X}}}\,\Phi(X) + \epsilon\right) \leq \frac{\mathrm{Var}^{\mathbb{P}_{\mathcal{X}}}\,\Phi(X)}{N_u \epsilon^2}, \tag{15}$$

with $\mathrm{Var}^{\mathbb{P}_{\mathcal{X}}}\,\Phi(X)$ the variance, $\hat{\mathbb{P}}_u$ the empirical distribution of the N_u unlabeled samples, and $\epsilon > 0$ the error. Equivalently, with probability $1 - \delta$ we have $\mathbf{E}^{\mathbb{P}_{\mathcal{X}}}\,\Phi(X) \leq \mathbf{E}^{\hat{\mathbb{P}}_{\mathcal{X}}}\,\Phi(X) + \sqrt{\frac{\mathrm{Var}^{\mathbb{P}_{\mathcal{X}}}\,\Phi(X)}{N_u \delta}}$.

In some settings we can obtain tighter bounds on the error of the computed guarantee. Specifically, in many modelling scenarios, we have that $\ell(h_\theta(\cdot), \cdot)$ and $c(\cdot, \cdot)$ are bounded, implying that $\Phi(\cdot)$ is bounded (for fixed values of $\theta, \alpha, \beta, \overline{\lambda}, \underline{\lambda}$). If $\Phi(X) \in [a, b]$, Hoeffding's inequality (Hoeffding, 1963) states that

$$P\left(\mathbf{E}^{\mathbb{P}_{\mathcal{X}}}\,\Phi(X) \geq \mathbf{E}^{\hat{\mathbb{P}}_{\mathcal{X}}}\,\Phi(X) + \epsilon\right) \leq \exp\left(-\frac{2N_u \epsilon^2}{(b-a)^2}\right),$$

Equivalently, with probability $1 - \delta$ we have $\mathbf{E}^{\mathbb{P}_{\mathcal{X}}}\,\Phi(X) \leq \mathbf{E}^{\hat{\mathbb{P}}_{\mathcal{X}}}\,\Phi(X) + (b - a)\sqrt{\frac{\log(1/\delta)}{2N_u}}$.

Substituting $\mathbf{E}^{\hat{\mathbb{P}}_{\mathcal{X}}}$ for $\mathbf{E}^{\mathbb{P}_{\mathcal{X}}}$ in (14), then, we can approximate the guarantee on the expected loss provided by weak duality, with the error of the approximation improving with the square root of the number N_u of unlabeled samples. This differs from standard generalization bounds in that here the approximation error only appears when we attempt to numerically compute the bound, and does not characterize the robustness guarantee itself, stated in (14). It differs also in that it relies only on the number of unlabeled (rather than labeled) data, which we assume to be abundant.

In Section 5.2, we compute this bound for a number of data sets and compare it to the equivalent bound from traditional DRL (Figure 5). Since the bound relies on the true distribution $\mathbb{P}$ being included in the adversary's decision set, the choice of the radius ε of the Wasserstein ball around the labeled data distribution $\hat{\mathbb{P}}_l$ is very important. We comment on the impact of ε in Section 5.1.

5. Empirical Results

In this section, we investigate the empirical performance of our proposed formulation of distributionally robust learning in the particular case of logistic regression. First, we demonstrate an important limitation of the previously-proposed distributionally robust logistic regression (Abadeh et al., 2015): Namely, there is often no choice of the radius ε of the adversary's decision set that yields a classifier that is both robust and non-trivial, in the sense that it makes predictions with nonzero confidence. We then demonstrate that the formulation proposed here, which uses unlabeled data to restrict the adversary, can yield non-trivial classifiers with non-vacuous bounds on the generalization error.

5.1 How Important is the Choice of ε?

In practice, we do not know the radius necessary to include the true data distribution $\mathbb{P}$ in the Wasserstein ball $\mathcal{B}_\varepsilon(\hat{\mathbb{P}}_l)$. Standard practice for DRL is to choose ε by cross-validation, attempting to maximize a proxy for the out-of-sample performance. Implicitly, however, doing so relies on a *regularization* effect of traditional DRL, documented by Gao et al. (2017) and Shafieezadeh-Abadeh et al. (2019), which generates an inverted U-shaped out-of-sample performance curve with respect to ε. Maximizing cross-validation performance does not necessarily yield a robust classifier in the DRL sense: As we demonstrate in Section 5.2, for some data sets there is *no choice of* ε that both includes $\mathbb{P}$ in the adversary's decision set and yields a non-trivial classifier using traditional DRL. As a consequence, ε that maximizes generalization performance is much smaller than the distance between the labeled data $\hat{\mathbb{P}}_l$ and the true data distribution $\mathbb{P}$.

Here, we verify that the choice of ε matters critically for robustness in the sense of traditional DRL, meaning that a learned classifier that is robust to distributions within an ε-ball $\mathcal{B}_\varepsilon(\hat{\mathbb{P}}_l)$ is not robust to distributions even slightly outside the ε-ball. In this sense, choosing ε by cross-validation in traditional DRL can yield a classifier that is not robust to perturbations on the order of the distance between the labeled data $\hat{\mathbb{P}}_l$ and the true data distribution $\mathbb{P}$.

Figure 2 shows in blue the generalization performance (the average probability assigned by the learned model to the correct class, on an out-of-sample test set) as a function of the radius of robustness ε used during training the model, using the traditional distributionally robust logistic regression model of Abadeh et al. (2015). The figure also shows in orange the median confidence (the maximum of the predicted probabilities $h_\theta(\mathbf{x})$ and $1 - h_\theta(\mathbf{x})$) of the learned model evaluated on the test set. Further details are in Appendix D.2.

Figure 3 shows the performance of the learned model evaluated on the worst data distribution taken from the Wasserstein ball of radius $\varepsilon + \Delta$ around the distribution of training data. This ball is exactly the ambiguity set of traditional DRL for radius $\varepsilon + \Delta$, meaning that the performance shown in the figure measures robustness in the sense of traditional DRL—it is the worst we might perform if the data distribution is allowed to lie at a distance at most $\varepsilon + \Delta$ from the training data. Further details can be found in Appendix D.3.

We make three empirical observations:

1. The generalization performance curve for traditional DRL has an inverted U shape, with a maximum at a much smaller radius ε than is required to include the true data distribution $\mathbb{P}$ in the adversary's decision set. This is shown for several data sets in Figure 2 and further in Appendix D.2.

2. The confidence of the model in every case drops quickly as the radius ε increased, for radii much smaller than the distance required for the robustness guarantee to hold (which has log-value 0.0 in the figure), in most cases reaching zero confidence (i.e., assigning equal probability to the classes) for ε smaller than this distance.

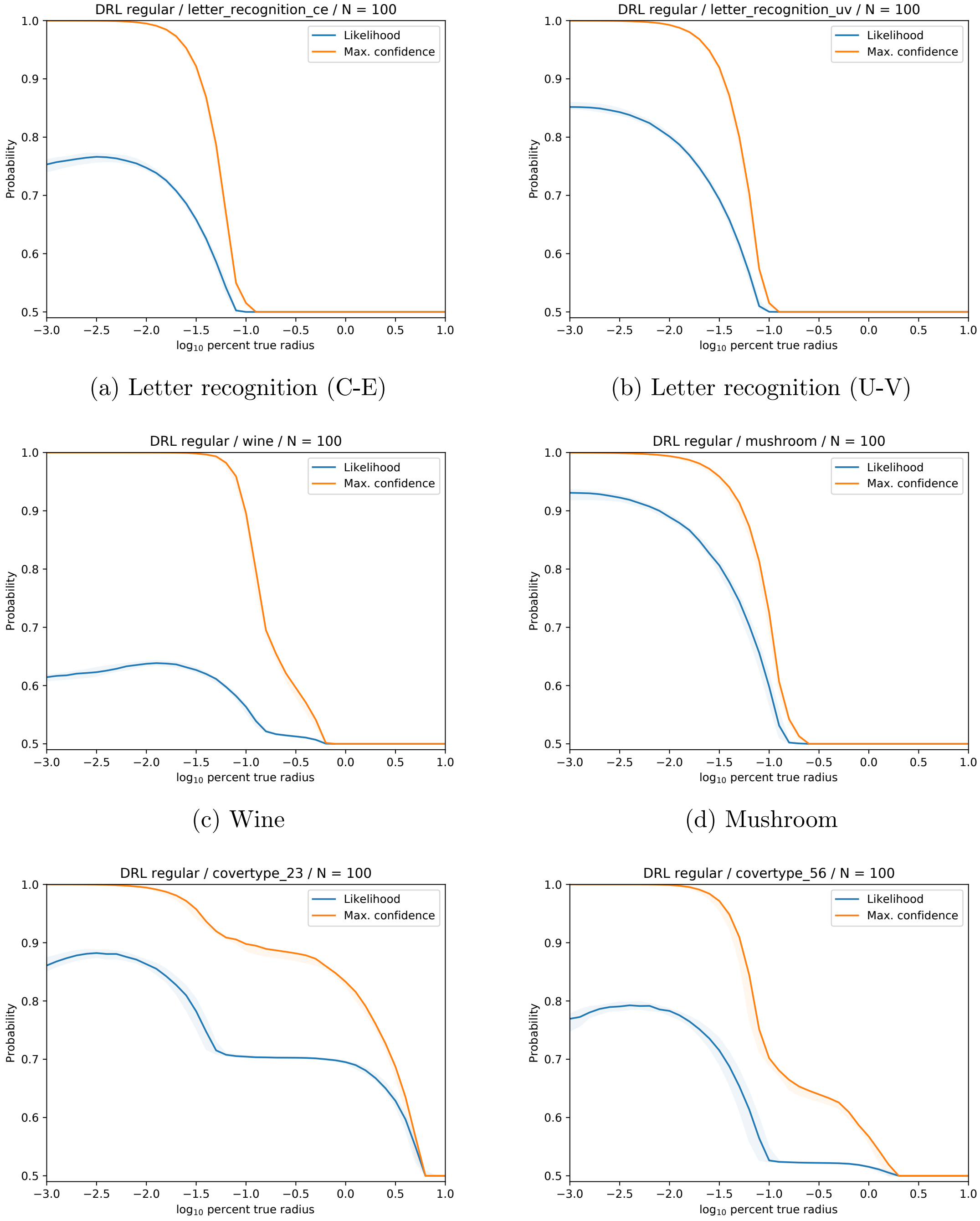
3. Wasserstein distributional robustness up to radius ε does not confer robustness to distributions even slightly outside $\mathcal{B}_\varepsilon(\hat{\mathbb{P}}_l)$, at distance $\varepsilon + \Delta$, in the sense that there exists a data distribution in the ball $\mathcal{B}_{\varepsilon+\Delta}(\hat{\mathbb{P}}_l)$ that yields poor performance for the traditional Wasserstein DRL predictor trained with radius of robustness ε. This is shown for several data sets in Figure 3 and further in Appendix D.3.

5.1.1 Choosing ε in the Proposed Method

The choice of ε for the proposed method is constrained by the fact that **the feasible set is empty** for ε below a threshold, as there might be no distribution in the ball $\mathcal{B}_\varepsilon(\hat{\mathbb{P}}_l)$ having the desired marginals $\mathbb{P}_\mathcal{X}$ and $\mathbb{P}_\mathcal{Y}$. This situation is easily detected in practice, as the value of the dual $g(\theta)$ becomes unbounded below.

Empirically, with the proposed method, we find no evidence of a bias-variance tradeoff as the radius ε is varied, unlike traditional Wasserstein DRL. Figure 4 shows out-of-sample performance as we vary the difference between the radius ε and the minimal such radius for which the feasible set is nonempty. The performance is flat out to a radius beyond which the confidence of the learner decreases quickly. Appendix D.4 contains further examples.

This last observation suggests a criterion for choosing ε under the proposed DRL model: One chooses the maximum ε such that the confidence of the learned classifier is above a threshold. This is the as-robust-as-possible selection, as opposed to the maximum-cross-validation-performance selection often used in traditional DRL. So long as the learned hypothesis has high confidence, the proposed DRL sees no tradeoff between out-of-sample performance and robustness so there's no cost to choosing ε solely according to the confidence. There are multiple possible implementations of this criterion. In our experiments, for example, thresholding the median confidence on the unlabeled set at 0.7 often suffices to ensure that ε is large enough to ensure $\mathbb{P} \in \mathcal{B}_\varepsilon(\hat{\mathbb{P}}_l)$, for reasonable values of N_l (Figure 6).



(a) Letter recognition (C-E)

(b) Letter recognition (U-V)

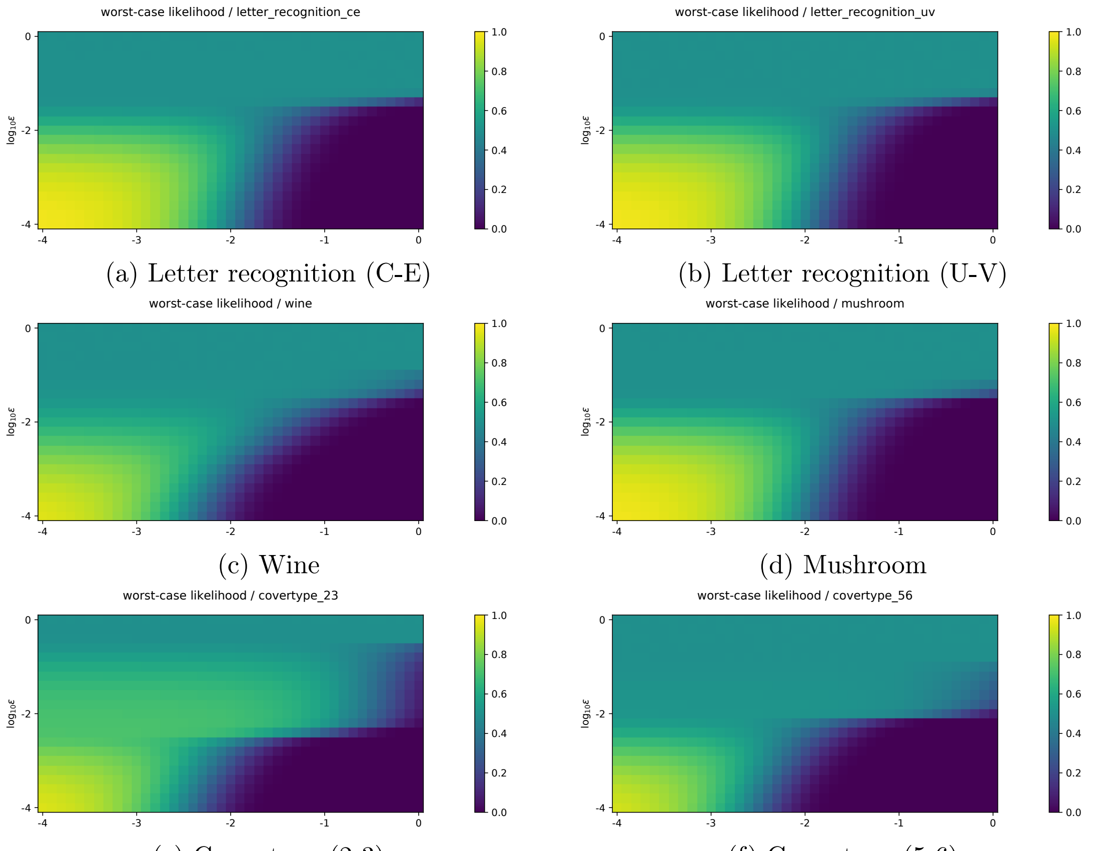
(c) Wine

(d) Mushroom

(e) Cover type (2-3)

(f) Cover type (5-6)

Figure 2: Traditional Wasserstein DRL. Out-of-sample performance (likelihood) and maximum confidence vs. radius of robustness ε as a percentage of the distance to the true data distribution $\mathbb{P}$. Performance shows a bias-variance tradeoff with peak at ε much smaller than the distance to $\mathbb{P}$. Confidence often drops sharply at a radius much smaller than the distance to $\mathbb{P}$.



(a) Letter recognition (C-E)

(b) Letter recognition (U-V)

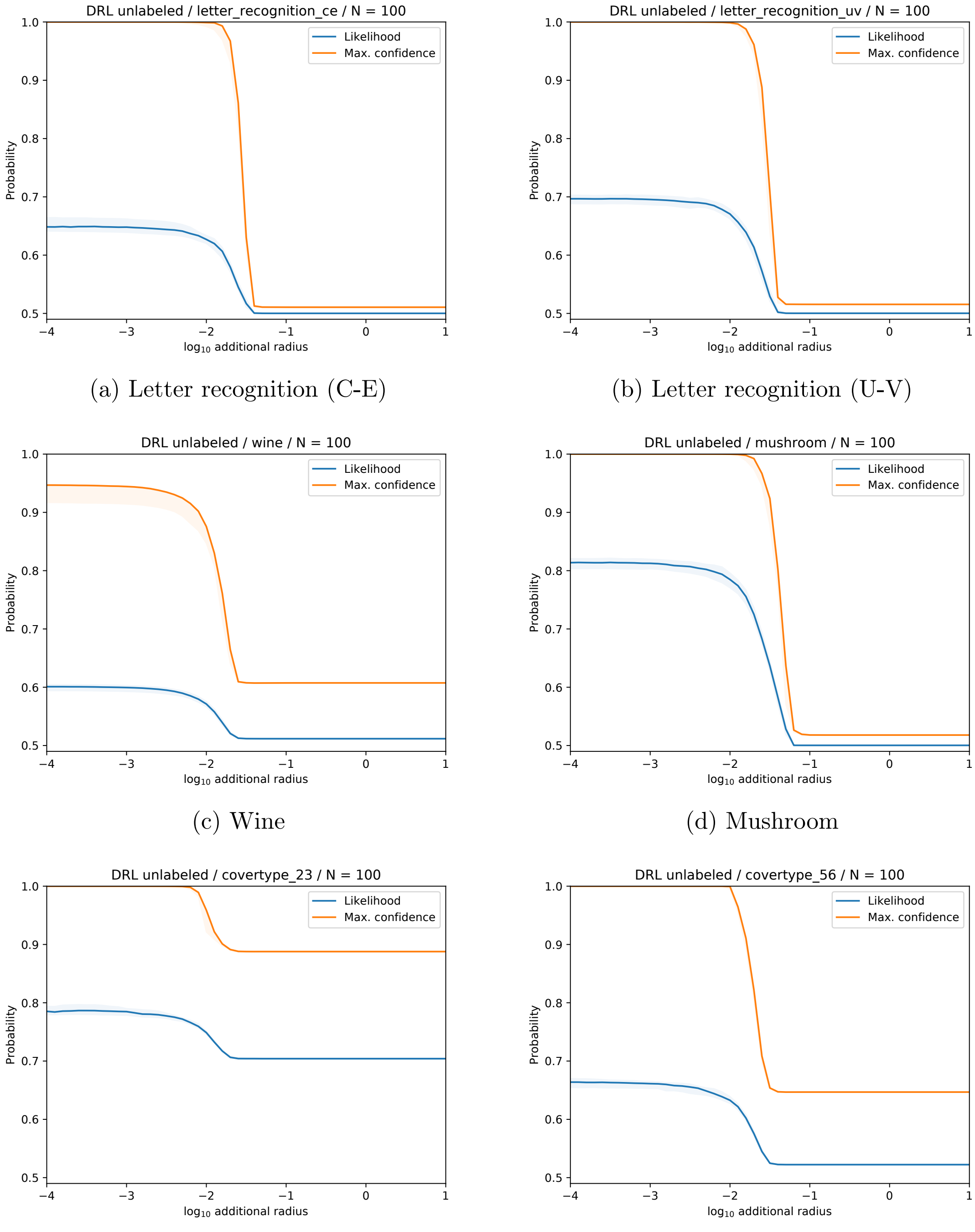
(c) Wine

(d) Mushroom

(e) Cover type (2-3)

(f) Cover type (5-6)

Figure 3: Traditional Wasserstein DRL. Worst-case performance (likelihood) vs. radius of robustness ε and test-time data radius $\varepsilon + \Delta$. Yellow indicates perfectly correct prediction (likelihood 1), blue perfectly incorrect (likelihood 0), and green perfectly indecisive prediction (likelihood 0.5). Training with radius ε confers little robustness beyond ε.



(a) Letter recognition (C-E)

(b) Letter recognition (U-V)

(c) Wine

(d) Mushroom

(e) Cover type (2-3)

(f) Cover type (5-6)

Figure 4: DRL with unlabeled data. Out-of-sample performance (likelihood) and maximum confidence vs. difference between the radius of robustness ε and the minimal radius necessary for the decision set to be nonempty. Unlike with traditional Wasserstein DRL, here there is no apparent bias-variance tradeoff. Performance is flat out to a radius at which the confidence drops sharply.

5.2 Empirical Performance of Learning with Unlabeled Data

In this section, we demonstrate the impact of the proposed method for constraining the adversary's decision set using unlabeled data. We evaluate the performance guarantee offered by the previously-proposed distributionally robust logistic regression model (Abadeh et al., 2015) on several binary classification data sets,[4] and we compare it to the guarantee offered by our model under the assumption that the radius ε of the Wasserstein ball $\mathcal{B}_\varepsilon(\hat{\mathbb{P}}_l)$ defining the adversary's decision set is chosen to include the true data distribution $\mathbb{P}$.

For each data set, we sample a small number N_l of labeled examples and compute the radius ε that is required to include the true (empirical) data distribution $\mathbb{P}$ in the Wasserstein ball $\mathcal{B}_\varepsilon(\hat{\mathbb{P}}_l)$. This is the smallest ε for which the performance guarantee from DRL holds. We use the labeled examples to compute the distributionally robust logistic regression under the traditional model (Abadeh et al., 2015) and additionally use the set of unlabeled examples to compute the same regression under the proposed model. We compare the performance guarantee (i.e. the dual objective value) computed under each DRL model. Identical values of ε are used for both methods, but a different value of ε is computed for each sampled set of labeled examples $\hat{\mathcal{Z}}_l$.

We examine two settings for the proposed method. The first assumes a **strong prior** that specifies the exact (true) label probabilities, such that $\underline{\mathbf{p}}_\mathcal{Y} = \overline{\mathbf{p}}_\mathcal{Y}$. In practice such a strong prior might come from auxiliary data, such as in ecological inference or with domain knowledge. The second setting assumes a **weak prior** that specifies only 95% confidence intervals for the label probabilities, estimated directly from the from labeled data $\hat{\mathcal{Z}}_l$ (Clopper and Pearson, 1934). Unlike the strong prior, the weak prior requires no information about labels outside the training set.

We vary the number of labeled examples and examine the computed performance guarantee, shown in Figure 5, as well as the median confidence of the learned predictor, shown in Figure 6. The former is the worst-case guarantee (6) and not the actual generalization performance of the learned classifier. We make three observations:

1. For all but one of the data sets, the performance bound computed by traditional DRL is vacuous (guaranteeing only likelihood greater than or equal to 0.5), while the learned classifier is trivial (assigning equal probability to both classes), for all tested numbers of labeled examples (maximum $N_l = 1000$).

2. For all data sets, the proposed DRL using unlabeled data and either a strong prior and a weak prior on the label probabilities yields a non-vacuous performance bound and a non-trivial classifier, for N_l at which traditional DRL is vacuous.

3. The strong prior on label probabilities can yield highly non-trivial performance bounds, for smaller numbers of labeled examples N_l than the weak prior. The weak prior yields a looser performance guarantee as it specifies only the 95% intervals computed from the training data, allowing a broader range of data distributions in the ambiguity set.

4. For 3 of the 14 data sets (see Figures 8 and 9 in the appendix) both regular DRL and the weak prior yield vacuous performance guarantees for all training set sizes. This occurs for the "difficult" data sets, in which even an non-robust linear logistic regression

[4] Data sets are from the UCI machine learning repository (Dua and Graff, 2019).

model struggles to achieve high out-of-sample performance. This can be seen in the out-of-sample performance curves in Figure 10 for very small radius of robustness ε. The strong prior nevertheless recovers a non-vacuous guarantee in all cases.

We have chosen ε as small as possible while ensuring the computed performance guarantee holds, and the performance bound computed by either algorithm gets monotonically worse as ε increases.

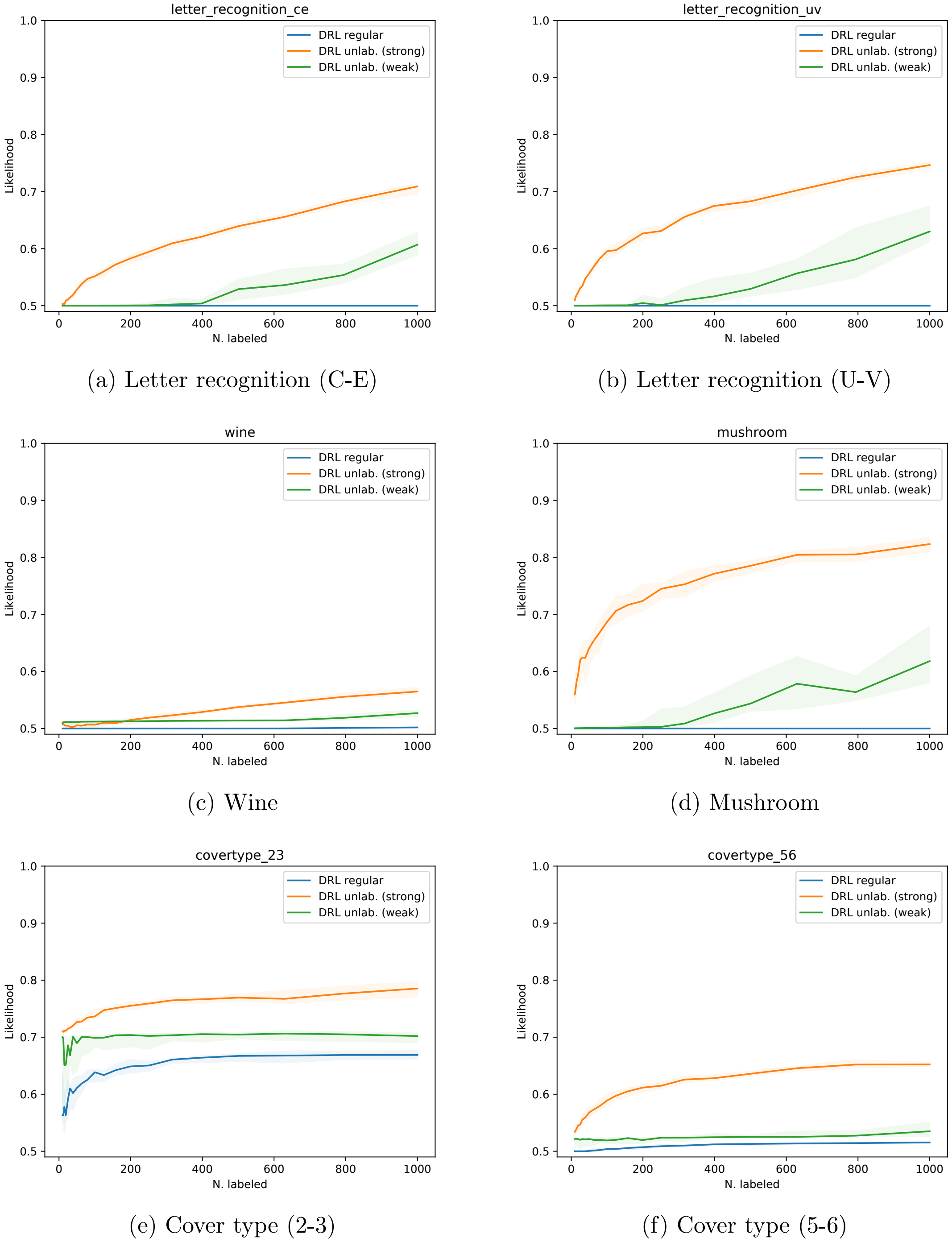
We emphasize that the results in Figures 5 and 6 do not address the generalization error for either algorithm but rather involve the guaranteed performance over a specific set of possible data distributions; this is the "robustness" in distributional robustness. Regular DRL attempts to be robust over a very large set of possible data distributions; this set can be made smaller using unlabeled data and prior bounds on label proportions. Our empirical results show that often this smaller set is small enough that the resulting performance guarantee is non-vacuous. The smaller set is obtained using additional information—from unlabeled data and prior bounds on label proportions—that is unavailable to regular DRL, placing it in a semi-supervised setting while the previous DRL is fully-supervised.

5.3 Discussion

The overwhelming size of an adversary's decision set is a weakness of Wasserstein DRL that prevents a reasonable tradeoff between robustness and confidence of the learned predictor. To circumvent this problem, we use unlabeled data to further constrain the decision set. Empirically, the proposed DRL problem produces non-trivial predictors having non-vacuous performance guarantees in cases where traditional Wasserstein DRL fails to do so.

One topic we have not addressed is computational complexity. The proposed DRL is computed via stochastic gradient descent. Each gradient computation scales linearly in the number of labeled examples N_l and this scaling might prohibit application to large labeled data sets. The key bottleneck is computing membership in the sets V^{ik} in (10), which relies on a maximization over labeled examples. This computation might be a fruitful target for performance improvement, possibly via parallelization or by leveraging the fact that the cost function c is a power of a metric. To give a sense of the overall complexity of the stochastic gradient descent procedure, in Figure 7 we show the progress of the dual objective value and ℓ^2 norm of the parameter gradient during the optimization, along with the average per-iteration wallclock time, for a single data set. The setting is the same as in Figure 5 and implementation details are given in Appendix D.1. Periodic rescaling of the step size plays an important role in the optimization, suggesting that the number of iterations required before convergence might significantly be reduced by using a smarter adaptation rule for the step size.

In addition, our model implicitly assumes access to unlabeled data that are uncorrupted by noise, as it constrains the data distribution to have marginal exactly equal to $\mathbb{P}_{\mathcal{X}}$. One can imagine a noise-tolerant version of the current model that replaces the exact marginal constraint defined by $\mathbb{P}_{\mathcal{X}}$ with another Wasserstein- (or other norm)-ball constraint. In the case of the additional Wasserstein ball constraint, the resulting problem is still a linear program, but it is no longer apparent that there is a dual problem having finitely many parameters. This problem will be interesting to study in future work.



(a) Letter recognition (C-E)

(b) Letter recognition (U-V)

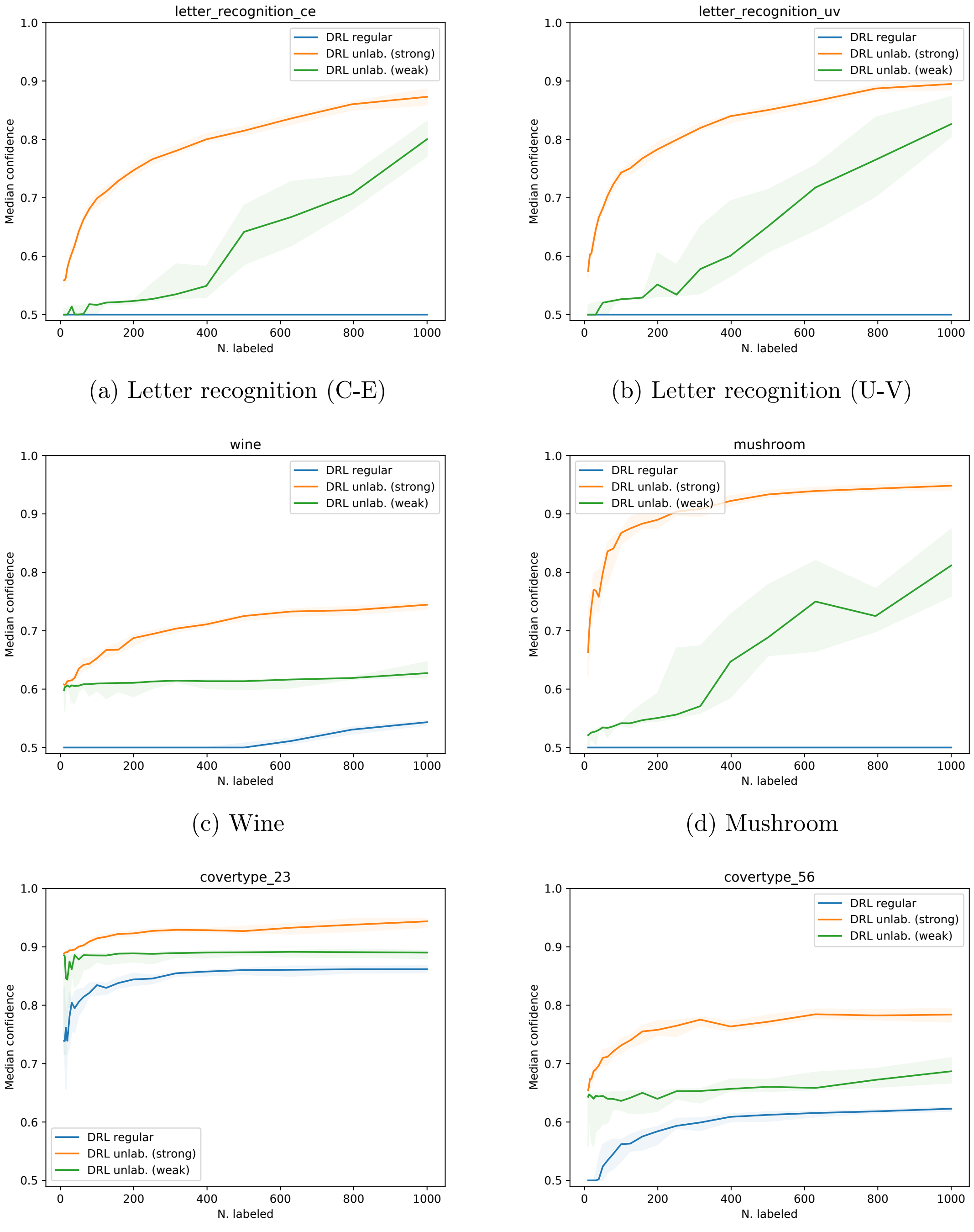
(c) Wine

(d) Mushroom

(e) Cover type (2-3)

(f) Cover type (5-6)

Figure 5: Worst-case performance bound (likelihood) vs. number of labeled data, setting ε to include the true test distribution. The regular DRL bound is often vacuous through $N_l = 1000$ while both DRL methods with unlabeled data yield non-vacuous bounds.



(a) Letter recognition (C-E)

(b) Letter recognition (U-V)

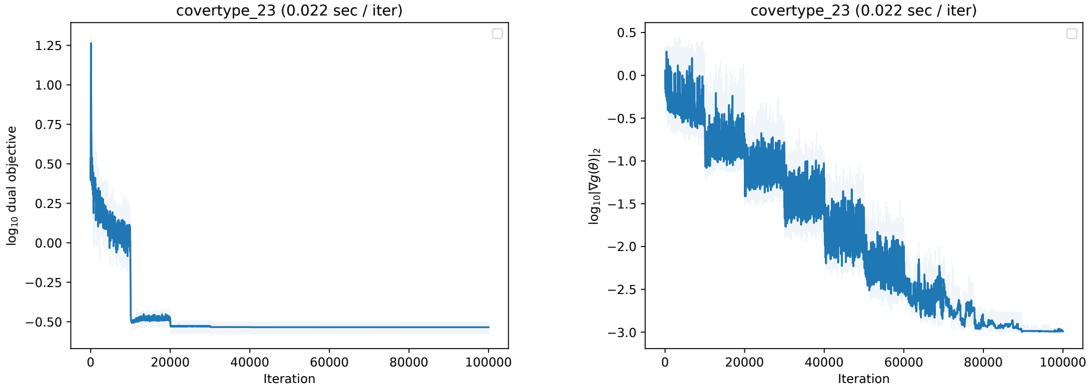
(c) Wine

(d) Mushroom

(e) Cover type (2-3)

(f) Cover type (5-6)

Figure 6: Median confidence vs. number of labeled data, setting ε to include the true test distribution. The regular DRL predictor often has confidence close to 0.5 in settings where both DRL methods using unlabeled data yield non-trivial predictors.



(a) Dual objective value

(b) ℓ^2 norm of the parameter gradient

Figure 7: Progress versus number of iterations for stochastic gradient descent optimization of the dual objective. Average wall clock time per iteration is 0.022 seconds on a Xeon E5-2690.

The dimensionality of the feature space $\mathcal{X}$ influences the performance guarantees obtained by both traditional DRL and the current method. In particular, empirical distributions of i.i.d. samples are known to converge to their continuous counterparts with their Wasserstein distance going as $\mathcal{O}(N^{-1/q})$, q being the dimension of the domain (Kloeckner, 2012; Claici et al., 2018). This scaling implies that the Wasserstein ball in the definition of both traditional DRL and in the currently proposed model gets "larger" as the dimension increases, likely diminishing the performance guarantee. This effect, of course, balances with the other factors that typically determine learning performance, such as the degree to which the chosen hypothesis class is well-tailored to the data set. In our experiments, we have explored data sets up to 617 dimensions (documented in Table 3).

6. Application: Distributionally-Robust Active Learning

Key to the learning algorithm of Section 4.1 is a mechanism for optimizing an objective over the intersection of a Wasserstein ball $\mathcal{B}_\varepsilon(\hat{\mathbb{P}}_l)$ with the set of distributions $\mathcal{U}(\mathbb{P}_\mathcal{X}, \overline{\mathbf{p}}_\mathcal{Y}, \underline{\mathbf{p}}_\mathcal{Y})$ that have prescribed marginals in $\mathcal{X}$ and $\mathcal{Y}$. Learning a classifier is just one possible application of this mechanism, however. In this section, we demonstrate another application, to active learning.

6.1 Model Change Heuristics

Given a set $\hat{\mathcal{Z}}_l$ of labeled data and a set $\hat{\mathcal{X}}_u$ of unlabeled data, an active learner attempts to choose the most beneficial example from $\hat{\mathcal{X}}_u$ for which to acquire a label. The goal of the active learner is to reduce the out-of-sample error of the predictor trained on $\hat{\mathcal{Z}}_l$ as rapidly as possible. Many active learning methods assign a score to each unlabeled example, indicating its predicted impact on the learned classifier if we choose to acquire its label. This score might represent various properties, such as model uncertainty, expected error reduction, or expected model change. In the current work we focus on **model change** criteria (Settles et al., 2008; Freytag et al., 2014; Cai et al., 2017), which are popular and often effective in practice (Yang and Loog, 2018).

In model change criteria, we define an **impact function** $f : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, which is large when acquiring the label $\mathbf{y}$ for point $\mathbf{x} \in \hat{\mathcal{X}}_u$ leads to a large change in the model parameters. Most often this is a norm of the parameter gradient (Yang and Loog, 2018),

$$f(\mathbf{x}, \mathbf{y}) = \|\nabla_\theta \ell(h_\theta(\mathbf{x}), \mathbf{y})\|,$$

for $\|\cdot\|$ a norm and h_θ the hypothesis trained on $\hat{\mathcal{Z}}_l$. The active learning heuristic selects $\mathbf{x}_* \in \hat{\mathcal{X}}_u$ that maximizes an estimate of the anticipated impact across possible labels at point $\mathbf{x}$. This might be the conditional expectation according to the model distribution at $\mathbf{x}$, the minimum over labels, or the maximum over labels:

- **Expected impact**: $\mathbf{x}_* = \operatorname{argmax}_{\mathbf{x} \in \hat{\mathcal{X}}_u} \sum_{k=1}^{N_\mathcal{Y}} h_\theta(\mathbf{x})_k f(\mathbf{x}, \mathbf{y}^k)$, with h_θ trained on $\hat{\mathcal{Z}}_l$.
- **Optimistic**: $\mathbf{x}_* = \operatorname{argmax}_{\mathbf{x} \in \hat{\mathcal{X}}_u} \max_{\mathbf{y} \in \mathcal{Y}} f(\mathbf{x}, \mathbf{y})$.
- **Conservative**: $\mathbf{x}_* = \operatorname{argmax}_{\mathbf{x} \in \hat{\mathcal{X}}_u} \min_{\mathbf{y} \in \mathcal{Y}} f(\mathbf{x}, \mathbf{y})$.

A potential problem with the expected model change criterion, which it shares with a number of other standard heuristics (Yang and Loog, 2018), is that it relies on the current hypothesis h_θ when predicting the impact of choosing a new point $\mathbf{x} \in \hat{\mathcal{X}}_u$ to label. Specifically, $h_\theta(\mathbf{x})$ is used in place of the conditional distribution over labels at the point $\mathbf{x}$. This is prone to error when the hypothesis is far from the true conditional distribution, incorrectly weighting the impact of obtaining labels at the points where the hypothesis is in error.

The "optimistic" and "conservative" estimates above are simple attempts to eliminate the hypothesis h_θ from the estimated impact. Notably, these ignore the labeled data $\mathcal{Z}_\ell$ entirely.

6.2 A Distributionally-Robust Approach

The machinery presented in Section 3 provides an alternative way to eliminate the hypothesis h_θ from our estimate of the impact of labeling point $\mathbf{x} \in \mathcal{X}_u$. We can formulate a distributionally-robust estimate of the impact, which computes a lower bound on the expected impact with respect to an entire set of plausible data distributions, rather than just the model distribution. This lower bound can be closer to the true expected impact (under $\mathbb{P}$) than the naïve conservative estimate, as our set of plausible distributions need not include those that are unreasonably far from the training set.

More precisely, we can formulate the problem of choosing the next sample to label as

$$\underset{\mathbf{x}_* \in \hat{\mathcal{X}}_u}{\text{maximize}} \inf_{\mu \in \mathcal{P}} \mathbf{E}^{\mu}_{Y|X=\mathbf{x}_*} f(X, Y), \tag{16}$$

with $\mathcal{P} = \mathcal{B}_\varepsilon(\hat{\mathbb{P}}_l) \cap \mathcal{U}(\mathbb{P}_\mathcal{X}, \overline{\mathbf{p}}_\mathcal{Y}, \underline{\mathbf{p}}_\mathcal{Y})$ as in Section 3, the intersection of a Wasserstein ball centered at the labeled data and the set of distributions having the prescribed marginals. The conditional expectation in (16) is linear in μ, since the $\mathcal{X}$-marginal of μ is exactly $\mathbb{P}_\mathcal{X}$, for all $\mu \in \mathcal{P}$. In practice, given the unlabeled data $\hat{\mathcal{X}}_u$, we can approximate this marginal by density estimation. We will use the notation $\hat{\varphi}(\mathbf{x}_*)$ for the approximate marginal density.

The inner problem in (16) estimates the impact of labeling the point $\mathbf{x}_* \in \mathcal{X}_u$. Just as in the DRL problem formulated in Section 3, this is the optimization of a linear objective—here $\mathbf{E}^\mu \, \mathbf{1}_{\mathbf{x}_*}(X) f(\mathbf{x}_*, Y)$—with respect to a probability measure constrained to the feasible set $\mathcal{P} = \mathcal{B}_\varepsilon(\hat{\mathbb{P}}_l) \cap \mathcal{U}(\mathbb{P}_\mathcal{X}, \overline{\mathbf{p}}_\mathcal{Y}, \underline{\mathbf{p}}_\mathcal{Y})$.[5] Just as in Section 3, we can solve this via its dual,

$$g(\mathbf{x}_*) = \begin{cases} -\inf_{\alpha,\beta,\underline{\lambda},\overline{\lambda}} & \alpha\varepsilon + \frac{1}{N_l}\sum_{i=1}^{N_l}\beta^i + \sum_{k=1}^{N_\mathcal{Y}}\left(\overline{\lambda}^k\overline{\mathbf{p}}_\mathcal{Y}^k - \underline{\lambda}^k\underline{\mathbf{p}}_\mathcal{Y}^k\right) + \mathbf{E}^{\mathbb{P}_\mathcal{X}}\,\Psi(X; \mathbf{x}_*, \alpha, \beta, \overline{\lambda}, \underline{\lambda}) \\ \text{s.t.} & \alpha, \underline{\lambda}^k, \overline{\lambda}^k \geq 0, \quad \forall k \in \{1, \ldots, N_\mathcal{Y}\} \end{cases} \tag{17}$$

with

$$\Psi(\mathbf{x}; \mathbf{x}_*, \alpha, \beta, \overline{\lambda}, \underline{\lambda}) = \max_{\substack{k\in\{1,\ldots,N_\mathcal{Y}\},\\ i\in\{1,\ldots,N_l\}}} -\frac{\mathbf{1}_{\mathbf{x}_*}(\mathbf{x}) f(\mathbf{x}, \mathbf{y}^k)}{\hat{\varphi}(\mathbf{x}_*)} - \alpha c((\mathbf{x}, \mathbf{y}^k), \mathbf{z}_\ell^i) - \beta^i - (\overline{\lambda}^k - \underline{\lambda}^k) \tag{18}$$

and f the impact function from Section 6.1.

The infimum in (17) corresponds exactly to (8) in Section 3.3 and we can likewise solve (17) via SGD, as shown in Algorithm 2, with $\Psi(\cdot; \mathbf{x}_*, \alpha, \beta, \overline{\lambda}, \underline{\lambda})$ here corresponding to $\Phi(\cdot; \theta, \alpha, \beta, \overline{\lambda}, \underline{\lambda})$ from Section 4.1, replacing $\ell(h_\theta, (\mathbf{x}, \mathbf{y}))$ with $\frac{-\mathbf{1}_{\mathbf{x}_*}(\mathbf{x}) f(\mathbf{x},\mathbf{y})}{\hat{\varphi}(\mathbf{x})}$. The relevant gradient computations for $\Psi(\mathbf{x}; \mathbf{x}_*, \alpha, \beta, \overline{\lambda}, \underline{\lambda})$ are identical to those for $\Phi(\mathbf{x}; \theta, \alpha, \beta, \overline{\lambda}, \underline{\lambda})$ and are included in Appendix B.

Algorithm 2 SGD for distributionally robust active learning

```
Given: $\theta \in \Theta$, $\varepsilon \geq 0$, $\overline{\mathbf{p}}_\mathcal{Y}, \underline{\mathbf{p}}_\mathcal{Y} \in [0,1]^{N_\mathcal{Y}}$, $\theta_0 \in \Theta$, step size $\gamma > 0$, batch size $N_b$.
for $\mathbf{x}_* \in \hat{\mathcal{X}}_u$ do
    $\alpha, \beta, \overline{\lambda}, \underline{\lambda} \leftarrow \mathbf{0}$.
    while not converged do
        $\mathbf{x}^1, \ldots, \mathbf{x}^{N_b} \sim \mathbb{P}_\mathcal{X}$.
        $\alpha \leftarrow \max\left(0, \alpha - \gamma\left[\varepsilon + \frac{1}{N_b}\sum_{j=1}^{N_b}\nabla_\alpha\Psi(\mathbf{x}^j; \mathbf{x}_*, \alpha, \beta, \overline{\lambda}, \underline{\lambda})\right]\right)$.
        $\beta \leftarrow \beta - \gamma\left[\frac{1}{N_l} + \frac{1}{N_b}\sum_{j=1}^{N_b}\nabla_\beta\psi(\mathbf{x}^j; \mathbf{x}_*, \alpha, \beta, \overline{\lambda}, \underline{\lambda})\right]$.
        $\overline{\lambda} \leftarrow \max\left(\mathbf{0}, \overline{\lambda} - \gamma\left[\overline{\mathbf{p}}_\mathcal{Y} + \frac{1}{N_b}\sum_{j=1}^{N_b}\nabla_{\overline{\lambda}}\psi(\mathbf{x}^j; \mathbf{x}_*, \alpha, \beta, \overline{\lambda}, \underline{\lambda})\right]\right)$.
        $\underline{\lambda} \leftarrow \max\left(\mathbf{0}, \underline{\lambda} - \gamma\left[-\underline{\mathbf{p}}_\mathcal{Y} + \frac{1}{N_b}\sum_{j=1}^{N_b}\nabla_{\underline{\lambda}}\psi(\mathbf{x}^j; \mathbf{x}_*, \alpha, \beta, \overline{\lambda}, \underline{\lambda})\right]\right)$.
    end while
    $\hat{g}[\mathbf{x}_*] \leftarrow -\left(\alpha\varepsilon + \frac{1}{N_l}\sum_{i=1}^{N_l}\beta^i + \sum_{k=1}^{N_\mathcal{Y}}(\overline{\lambda}^k\overline{\mathbf{p}}_\mathcal{Y}^k - \underline{\lambda}^k\underline{\mathbf{p}}_\mathcal{Y}^k) + \frac{1}{N_u}\sum_{j=1}^{N_u}\Psi(\mathbf{x}_u^j; \mathbf{x}_*, \alpha, \beta, \overline{\lambda}, \underline{\lambda})\right)$.
end for
Choose $\mathrm{argmax}_{\mathbf{x}_* \in \hat{\mathcal{X}}_u}\, \hat{g}[\mathbf{x}_*]$.
```

[5] Note that the objective here, $-\mathbf{1}_{\mathbf{x}_*}(\mathbf{x}) f(\mathbf{x}, \mathbf{y})$, is lower semicontinuous in $\mathbf{x}$, whereas Theorem 2 requires upper semicontinuity of $\ell(h_\theta(\mathbf{x}), \mathbf{y})$ in $\mathbf{x}$. We nevertheless use the duality proved in Theorem 2, as one can approximate our lower semicontinuous objective here arbitrarily well by a combination of continuous bump functions centered at $(\mathbf{x}_*, \mathbf{y}^k)$, for all k, and the strong duality holds for any such approximation.

6.3 Empirical Results

We evaluate active learning performance on the set of 14 binary classification data sets used in Section 3. Given a set of labeled examples, a linear classifier is trained by ℓ^2-regularized logistic regression, with the weight on the regularizer fixed a priori. Given this classifier and a set of unlabeled examples, the active learning algorithm selects the next example for which to acquire a label. The process is iterated, beginning with 20 examples chosen uniformly at random (the same initial examples for all active learning methods, but different initial examples between trials), and terminating after 100 labeled examples have been acquired.

After training the classifier at each step, we evaluate the error (the likelihood) on the combination of labeled and unlabeled data, to provide a score that is comparable between steps. This score has previously been proposed as a proxy for out-of-sample error in both the semi-supervised (Grandvalet and Bengio, 2005) and active (Guo and Schuurmans, 2008) learning settings. We use this score for consistency with (Yang and Loog, 2018), which surveys and benchmarks a number of standard algorithms.

We compare the proposed distributionally-robust active learning method to both random sampling and the existing model-change heuristics (described in Section 6.1) Specifically, we test five methods:

1. **Random**: We choose the next example uniformly at random.

2. **EMC**: We choose the example that maximizes the expected norm of the parameter gradient, under the hypothesis distribution (Settles et al., 2008).

3. **Min. MC**: We choose the example that maximizes the minimum (over possible labels) norm of the parameter gradient ("conservative" in Section 6.1).

4. **Max. MC**: We choose the example that maximizes the maximum (over possible labels) norm of the parameter gradient ("optimistic" in Section 6.1).

5. **DR (strong)**: We choose the example that maximizes the proposed distributionally-robust lower bound on the expected norm of the parameter gradient (Section 6.2). We use a strong prior on the label probabilities, being the true label probabilities.

6. **DR (weak)**: We choose the example that maximizes the proposed distributionally-robust lower bound on the expected norm of the parameter gradient (Section 6.2). We use a weak prior on the label probabilities, being 95% confidence bounds estimated from the labeled data.

Table 2 shows the area under the likelihood curve (from samples 20 through 100) for each method and data set, using the median likelihood over trials (i.e., initial training samples). We make several observations:

1. The existing model change heuristics yield inconsistent performance, with EMC, Min. MC, and Max. MC underperforming random sampling on 10 of 14 data sets.

2. The proposed distributionally-robust heuristics only underperformed random sampling on 5 of 14 data sets.

Table 2: Active learning: 100× area under the likelihood curve (median over trials).

Data set	Random	EMC	Min MC	Max MC	DR (strong)	DR (weak)
Abalone	68.7	66.6	66.0	64.2	**69.8**	69.8
Bank	92.6	94.5	94.5	86.0	95.3	**95.4**
Cover (2/3)	**79.9**	78.0	79.4	74.0	78.4	78.4
Cover (5/6)	72.7	72.3	71.4	66.3	**73.0**	72.9
Isolet	60.8	58.2	60.0	54.8	63.3	**63.8**
Letter (C/E)	71.6	67.6	67.6	64.4	73.2	**73.6**
Letter (U/V)	77.8	73.2	73.1	67.1	**82.3**	82.1
Magic	**52.2**	46.0	46.3	46.0	43.1	43.1
Mushroom	86.3	91.0	90.6	77.9	**92.2**	**92.2**
Pulsar	76.9	**79.7**	79.4	75.0	79.6	79.6
Spam	68.4	68.0	67.3	63.1	**69.9**	**69.9**
Thyroid	79.1	**80.7**	80.6	78.8	78.4	78.6
Wine	**56.7**	53.0	51.9	51.3	50.4	50.5
Yeast	**54.7**	47.8	47.3	53.2	46.6	46.8

3. The proposed distributionally-robust heuristics outperform the other model change-based heuristics (EMC, Min. MC, and Max. MC) on 8 of 14 data sets.

4. The distributionally-robust heuristics performed similarly to one another, using a strong prior and a weak prior.

6.4 Discussion

The mechanism used to solve the Wasserstein DRL problem with unlabeled data (Section 3), which relies on the duality result in 3.3, can have broader applications. We have demonstrated an application to the problem of active learning that yields a distributionally-robust model change heuristic that empirically often outperforms the existing model change heuristics.

Computational complexity is a potential impediment to deploying the proposed active learning method. For each unlabeled example that might be selected for labeling, the method requires solving a distributionally robust problem (Equation 17) by an iterative method (such as in Algorithm 2). This in our experiments required on the order of 50000 iterations per example considered, with the complexity of each iteration scaling linearly with the number of training example N_l (identically to the DRL method in Section 3). As in Section 3, this complexity might be a productive target for future work.

7. Conclusion

We have explored an alternative to Wasserstein distributionally robust learning that incorporates unlabeled data to restrict the adversary's decision set. In particular, we proposed to intersect the standard Wasserstein ball constraint with the set of probability measures having specified marginals in both feature space and label space. This latter constraint adds some complexity to the derivation of a tractable algorithm (Section 3.3), which follows the standard DRL framework of dualizing the problem, but requires some care as the dual is now infinite-dimensional due to specifying the feature-space marginal in the primal. We prove a strong duality theorem (Theorem 2) that guarantees we can solve our proposed Wasserstein DRL problem via a dual formulation. This dual problem we can treat as the minimization of an expectation with respect to the pre-specified feature-space marginal, which is amenable to stochastic gradient methods (Algorithm 1). Critically, such methods rely only on sampling unlabeled data from the feature-space marginal of the data distribution. Therefore, the resulting SGD algorithm is tractable whenever we have access to plentiful unlabeled data, which is frequently the case in machine learning settings.

The motivation for exploring this alternative approach is an empirical observation that in standard Wasserstein DRL the adversary's decision set grows overly large very quickly as the radius of robustness ε is increased. As a result, choosing any radius sufficiently large for the adversary's decision set to contain the true data distribution will yield a trivial classifier (predicting equal probability for every class). Moreover, the generalization performance guarantee implied by distributionally robust methods only holds when the decision set contains the true data distribution. This performance guarantee is one major motivation for using DRL methods and here we have shown that there is a gap between theory and practice.

Restricting the adversary's decision set even more than we have done here might be a profitable avenue for further research. There are likely other ways to incorporate side information into the problem that can be applicable in a variety of practical settings. Moreover, the practical application of the method proposed here is currently somewhat constrained by the computational complexity of the SGD iterations, which scale linearly in the number of labeled examples used. It is possible there are significant speedups to be obtained via parallelization and further assumptions about the structure of the transport cost c.

Finally, we have proved the strong duality theorem (Theorem 2) only for feature spaces $\mathcal{X}$ that are compact. This is a reasonable assumption from a pragmatic standpoint, as no data distribution in practice will have unbounded support, but we also expect an equivalent theorem to hold for non-compact $\mathcal{X}$. Proof techniques from duality theorems for optimal transport, as in (Villani, 2003, Theorem 1.3) and (Rachev and Rüschendorf, 1998, Theorem 4.6.14), and distributional model risk assessment (Blanchet and Murthy, 2019, Theorem 1) might be applicable here.

Acknowledgments

The authors acknowledge the generous support of Army Research Office grants W911NF1710068 and W911NF2010168, of Air Force Office of Scientific Research award FA9550-19-1-031, of National Science Foundation grant IIS-1838071, from the MIT-IBM Watson AI Laboratory, from the Toyota-CSAIL Joint Research Center, from the QCRI–CSAIL Computer Science Research Program, from the MIT CSAIL Systems that Learn initiative, from the Skoltech–MIT Next Generation Program, and from a gift from Adobe Systems. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of these organizations. The authors also thank Nestor Guillen for helping them to understand duality over spaces of measures.

Appendix A. Proof of strong duality

To show strong duality, we will make use of a fundamental convex analysis result: the Fenchel duality theorem (Borwein and Zhu, 2005, Theorem 4.4.3).

Theorem 3 (Fenchel duality) *Let Ξ, Γ be Banach spaces, with convex functions $\gamma : \Xi \to \mathbb{R} \cup \{+\infty\}$ and $\chi : \Gamma \to \mathbb{R} \cup \{+\infty\}$, and a continuous linear map $A : \Xi \to \Gamma$. Then*

$$\inf_{\xi \in \Xi} \gamma(\xi) + \chi(A\xi) \geq \sup_{u \in \Gamma^*} -\gamma^*(A^* u) - \chi^*(-u).$$

If γ and χ are lower semicontinuous and A dom $\gamma \cap$ cont $\chi \neq \emptyset$, then equality holds above and the supremum on the right-hand side is attained.

Theorem 2 (Strong duality) *Let $\mathcal{X}$ be a compact Polish space and $\mathcal{Y} = \{\mathbf{y}^k\}_{k=1}^{N_\mathcal{Y}}$ any finite set. Let $\mathbb{P}_\mathcal{X}$ be a probability measure over $\mathcal{X}$ and $\hat{\mathbb{P}}_l = \frac{1}{N_l} \sum_{i=1}^{N_l} \delta_{\mathbf{z}_\ell^i}$ an empirical probability measure over $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, and define intervals $[\underline{\mathbf{p}}_\mathcal{Y}^k, \overline{\mathbf{p}}_\mathcal{Y}^k] \subseteq [0,1]$, $k \in \{1, \ldots, N_\mathcal{Y}\}$. Let the transportation cost $c : \mathcal{Z} \times \mathcal{Z} \to [0, +\infty)$ be nonnegative and upper semicontinuous with $c(\mathbf{z}, \mathbf{z}') = 0 \Leftrightarrow \mathbf{z} = \mathbf{z}'$. Assume $\ell(h_\theta(\cdot), \cdot) : \mathcal{Z} \to \mathbb{R}$ is upper semicontinuous. Define f as in (6) and g as in (8). If $\mathcal{U}(\mathbb{P}_\mathcal{X}, \underline{\mathbf{p}}_\mathcal{Y}, \overline{\mathbf{p}}_\mathcal{Y}) \cap \mathcal{B}_\varepsilon(\hat{\mathbb{P}}_l) \neq \emptyset$, then*

$$f(\theta) = g(\theta), \quad \forall \theta \in \Theta. \tag{19}$$

If $\operatorname{relint}(\mathcal{U}(\mathbb{P}_\mathcal{X}, \underline{\mathbf{p}}_\mathcal{Y}, \overline{\mathbf{p}}_\mathcal{Y}) \cap \mathcal{B}_\varepsilon(\hat{\mathbb{P}}_l)) \neq \emptyset$*, then there exists a minimizer $(\alpha_*, \beta_*, \overline{\lambda}_*, \underline{\lambda}_*) \in \mathbb{R}_+ \times \mathbb{R}^{N_l} \times \mathbb{R}_+^{N_\mathcal{Y}} \times \mathbb{R}_+^{N_\mathcal{Y}}$ attaining the infimum in (8).*

Proof For convenience, in what follows we will write $\ell(h_\theta, \mathbf{z}) \triangleq \ell(h_\theta(\mathbf{x}), \mathbf{y})$ for $\mathbf{z} = (\mathbf{x}, \mathbf{y}) \in \mathcal{Z}$. We first recall the primal problem (7) and the dual problem (8) from the main text:

$$f(\theta) = \begin{cases} \sup_\pi & \int_{\mathcal{Z} \times \mathcal{Z}} \ell(h_\theta, \mathbf{z}) \, \mathrm{d}\pi(\mathbf{z}, \mathbf{z}') & \\ \text{s.t.} & \int_{\mathcal{Z} \times \mathcal{Z}} c(\mathbf{z}, \mathbf{z}') \, \mathrm{d}\pi(\mathbf{z}, \mathbf{z}') \leq \varepsilon & \\ & \int_{\mathcal{Z} \times \mathcal{Z}} \delta_{\mathbf{z}_\ell^i}(\mathbf{z}') d\pi(\mathbf{z}, \mathbf{z}') = \frac{1}{N_l} & \forall i \in \{1, \ldots, N_l\}, \\ & \pi((A \times \mathcal{Y}) \times \mathcal{Z}) = \mathbb{P}_\mathcal{X}(A) & \forall A \in \mathcal{B}(\mathcal{X}), \\ & \int_{(\mathcal{X} \times \mathcal{Y}) \times \mathcal{Z}} \delta_{\mathbf{y}^k}(\mathbf{y}) \, \mathrm{d}\pi((\mathbf{x}, \mathbf{y}), \mathbf{z}') \leq \overline{\mathbf{p}}_\mathcal{Y}^k & \forall k \in \{1, \ldots, N_\mathcal{Y}\}, \\ & \int_{(\mathcal{X} \times \mathcal{Y}) \times \mathcal{Z}} \delta_{\mathbf{y}^k}(\mathbf{y}) \, \mathrm{d}\pi((\mathbf{x}, \mathbf{y}), \mathbf{z}') \geq \underline{\mathbf{p}}_\mathcal{Y}^k & \forall k \in \{1, \ldots, N_\mathcal{Y}\}, \\ & \pi(A) \geq 0 & \forall A \in \mathcal{B}(\mathcal{Z} \times \mathcal{Z}). \end{cases} \tag{20}$$

More succinctly, we may write the primal problem as:

$$f(\theta) = \sup_{\pi \in \Phi} F_\theta(\pi). \tag{21}$$

with F_θ the objective above and Φ the feasible set.

The Lagrangian dual to (21) is

$$\inf_{(\alpha, \beta, \phi, \overline{\lambda}, \underline{\lambda}) \in \Lambda_\theta} G(\alpha, \beta, \phi, \overline{\lambda}, \underline{\lambda}), \tag{22}$$

with

$$\begin{aligned} G(\alpha, \beta, \phi, \overline{\lambda}, \underline{\lambda}) &= \alpha\varepsilon + \frac{1}{N_l}\sum_{i=1}^{N_l} \beta^i + \int_{\mathcal{X}} \phi(\mathbf{x})\, \mathrm{d}\mathbb{P}_{\mathcal{X}}(\mathbf{x}) + \sum_{k=1}^{N_{\mathcal{Y}}} \left(\overline{\lambda}^k \overline{\mathbf{p}}^k_{\mathcal{Y}} - \underline{\lambda}^k \underline{\mathbf{p}}^k_{\mathcal{Y}} \right), \\ \Lambda_\theta = \Big\{ & (\alpha, \beta, \phi, \overline{\lambda}, \underline{\lambda}) \in \mathbb{R} \times \mathbb{R}^{N_l} \times C_b(\mathcal{X}) \times \mathbb{R}^{N_{\mathcal{Y}}} \times \mathbb{R}^{N_{\mathcal{Y}}} : \\ & \phi(\mathbf{x}) \geq \ell(h_\theta, (\mathbf{x}, \mathbf{y}^k)) - \alpha c((\mathbf{x}, \mathbf{y}^k), \mathbf{z}^i_\ell) - \beta^i - (\overline{\lambda}^k - \underline{\lambda}^k), \\ & \alpha, \overline{\lambda}^k, \underline{\lambda}^k \geq 0, \\ & \forall \mathbf{x} \in \mathcal{X}, k \in \{1, \ldots, N_{\mathcal{Y}}\}, i \in \{1, \ldots, N_l\} \Big\}. \end{aligned} \tag{23}$$

We make two claims:

1. $\sup_{\pi \in \Phi} F_\theta(\pi) = \inf_{(\alpha,\beta,\phi,\overline{\lambda},\underline{\lambda}) \in \Lambda_\theta} G(\alpha, \beta, \phi, \overline{\lambda}, \underline{\lambda})$.
2. There exists a dual optimizer $(\alpha, \beta, \phi_{\theta,\alpha,\beta,\overline{\lambda},\underline{\lambda}}, \overline{\lambda}, \underline{\lambda}) \in \mathbb{R} \times \mathbb{R}^{N_l} \times L^0(\mathcal{X}, \mathbb{P}_{\mathcal{X}}) \times \mathbb{R}^{N_{\mathcal{Y}}} \times \mathbb{R}^{N_{\mathcal{Y}}}$ with
$$\phi_{\theta,\alpha,\beta,\overline{\lambda},\underline{\lambda}}(\mathbf{x}) = \max_{k \in \{1,\ldots,N_{\mathcal{Y}}\}} \ell(h_\theta, (\mathbf{x}, \mathbf{y}^k)) - \min_{i \in \{1,\ldots,N_l\}} \left(\alpha c((\mathbf{x}, \mathbf{y}^k), \mathbf{z}^i_\ell) + \beta^i \right) - (\overline{\lambda}^k - \underline{\lambda}^k). \tag{24}$$
If $\ell(h_\theta, \cdot)$ and $c(\cdot, \mathbf{z}^i_\ell)$ are continuous, for all i, then $(\alpha, \beta, \phi_{\theta,\alpha,\beta,\overline{\lambda},\underline{\lambda}}, \overline{\lambda}, \underline{\lambda}) \in \Lambda_\theta$.

We start with the first claim, and apply Fenchel duality to the proposed dual problem (23) to show the desired equality. This strategy of dualizing the proposed dual mirrors the arguments of Villani (2003) in proving strong duality for regular optimal transport. On a compact Polish space, the dual of the space of finite, signed measures is larger than the space of continuous, bounded functions, necessitating this approach. On the other hand, the Riesz representation theorem allows us to move from the proposed dual to the primal in a rigorous manner. It tells us that the space of finite, signed measures is isomorphic to the dual of the space of continuous, bounded functions (on a compact Polish space).

We will actually show the first claim holds for a slight generalization of the dual problem (22). Consider the spaces $\Xi = \mathbb{R} \times C_b(\mathcal{Z}) \times C_b(\mathcal{X}) \times C_b(\mathcal{Y}) \times C_b(\mathcal{Y})$ and $\Gamma = C_b(\mathcal{Z} \times \mathcal{Z})$, which are Banach spaces. The norm on Ξ is given by $\|(\alpha, \beta, \phi, \overline{\lambda}, \underline{\lambda})\| \triangleq |\alpha| + \|\beta\|_\infty + \|\phi\|_\infty + \|\overline{\lambda}\|_\infty + \|\underline{\lambda}\|_\infty$, while the norm on Γ is $\|\cdot\|_\infty$.

Let $\nu_{\mathcal{Z}} \in \mathcal{M}(\mathcal{Z}), \nu_{\mathcal{X}} \in \mathcal{M}(\mathcal{X}), \overline{\nu}_{\mathcal{Y}} \in \mathcal{M}(\mathcal{Y}), \underline{\nu}_{\mathcal{Y}} \in \mathcal{M}(\mathcal{Y})$. The dual problem we will rewrite as

$$\inf_{\xi \in \Xi} \tilde{G}(\xi) + \chi(A\xi), \tag{25}$$

with

$$\begin{aligned} &\tilde{G} : \xi = (\alpha, \beta, \phi, \overline{\lambda}, \underline{\lambda}) \in \Xi \mapsto \begin{cases} \alpha\varepsilon + \langle \nu_{\mathcal{Z}}, \beta \rangle + \langle \nu_{\mathcal{X}}, \phi \rangle + \langle \overline{\nu}_{\mathcal{Y}}, \overline{\lambda} \rangle - \langle \underline{\nu}_{\mathcal{Y}}, \underline{\lambda} \rangle & \alpha, \overline{\lambda}, \underline{\lambda} \geq 0 \\ +\infty & \text{otherwise} \end{cases}, \\ &\chi : u \in \Gamma \mapsto \begin{cases} 0 & u(\mathbf{z}, \mathbf{z}') \geq \ell(h_\theta, \mathbf{z}) \ \forall \mathbf{z}, \mathbf{z}' \in \mathcal{Z} \\ +\infty & \text{otherwise} \end{cases}, \end{aligned} \tag{26}$$

and $A : \Xi \to \Gamma$ a linear operator defined by

$$(A\xi)(\mathbf{z}, \mathbf{z}') = \alpha c(\mathbf{z}, \mathbf{z}') + \beta(\mathbf{z}') + \phi(\mathbf{x}) + \overline{\lambda}(\mathbf{y}) - \underline{\lambda}(\mathbf{y}), \tag{27}$$

where $\mathbf{z} = (\mathbf{x}, \mathbf{y})$. Optimization problem (25) is identical to the dual problem (22) when $\nu_{\mathcal{Z}} = \hat{\mathbb{P}}_l$, $\nu_{\mathcal{X}} = \mathbb{P}_{\mathcal{X}}$, $\overline{\nu}_{\mathcal{Y}} = \overline{\mathbb{P}}_{\mathcal{Y}}$, and $\underline{\nu}_{\mathcal{Y}} = \underline{\mathbb{P}}_{\mathcal{Y}}$.

$\tilde{G}$ is convex and lower semi-continuous as a function of $(\alpha, \beta, \phi, \overline{\lambda}, \underline{\lambda})$, because it is linear on a closed, convex domain. A is clearly continuous and χ is convex and lower semi-continuous as the indicator of a closed, convex domain. Note also that $A\,\mathrm{dom}\,\tilde{G} \cap \mathrm{cont}\chi$ is nonempty as $\ell(h_\theta, \cdot)$ is an upper semi-continuous function on a compact domain and is bounded. In particular, $\ell(h_\theta, \cdot) < M$ for some $M \in \mathbb{R}$, so by choosing $\beta = M$ with $\alpha = \phi = \overline{\lambda} = \underline{\lambda} = 0$ we have that $A(\alpha, \beta, \phi, \overline{\lambda}, \underline{\lambda}) \in \mathrm{cont}\chi$.

As the underlying domain is compact, the topological duals of Ξ and Γ are $\Xi' = \mathbb{R} \times \mathcal{M}(\mathcal{Z}) \times \mathcal{M}(\mathcal{X}) \times \mathcal{M}(\mathcal{Y}) \times \mathcal{M}(\mathcal{Y})$ and $\Gamma' = \mathcal{M}(\mathcal{Z} \times \mathcal{Z})$. This duality allows us to define an adjoint for the operator A, given by $A^* : \Gamma' \to \Xi'$, with

$$A^*(\pi) = \left(\int c\,\mathrm{d}\pi, \pi_{Z'}, \pi_X, \pi_Y, -\pi_Y \right),$$

such that

$$\begin{aligned} A^*(\pi)(\alpha, \beta, \phi, \overline{\lambda}, \underline{\lambda}) &= \int_{\mathcal{Z} \times \mathcal{Z}} (\alpha c((\mathbf{x}, \mathbf{y}), \mathbf{z}') + \beta(\mathbf{z}') + \phi(\mathbf{x}) + \overline{\lambda}(\mathbf{y}) - \underline{\lambda}(\mathbf{y})) d\pi((\mathbf{x}, \mathbf{y}), \mathbf{z}') \\ &= \int_{\mathcal{Z} \times \mathcal{Z}} A(\alpha, \beta, \phi, \overline{\lambda}, \underline{\lambda})\,\mathrm{d}\pi. \end{aligned} \tag{28}$$

Here, π_X, π_Y denote $\mathcal{X}$- and $\mathcal{Y}$-marginals of the first marginal of π, while $\pi_{Z'}$ is the second marginal of π.

We can compute the convex conjugates of $\tilde{G}$ and χ.

$$\begin{aligned} \tilde{G}^*(a, \sigma, \tau, \zeta, \omega) &= \sup_{(\alpha, \beta, \phi, \overline{\lambda}, \underline{\lambda}) \in \Xi} \langle (a, \sigma, \tau, \zeta, \omega), (\alpha, \beta, \phi, \overline{\lambda}, \underline{\lambda}) \rangle - \tilde{G}(\alpha, \beta, \phi, \overline{\lambda}, \underline{\lambda}) \\ &= \sup_{(\alpha, \beta, \phi, \overline{\lambda}, \underline{\lambda}) \in \Xi} \begin{cases} (a - \varepsilon)\alpha + \int_{\mathcal{Z}} \beta\,(\mathrm{d}\sigma - \mathrm{d}\nu_{\mathcal{Z}}) + \int_{\mathcal{X}} \phi\,(\mathrm{d}\tau - \mathrm{d}\nu_{\mathcal{X}}) \\ \qquad + \int_{\mathcal{Y}} \overline{\lambda}\,(\mathrm{d}\zeta - \mathrm{d}\overline{\nu}_{\mathcal{Y}}) + \int_{\mathcal{Y}} \underline{\lambda}\,(\mathrm{d}\omega + \mathrm{d}\underline{\nu}_{\mathcal{Y}}) \\ \qquad \text{if } \alpha, \overline{\lambda}, \underline{\lambda} \geq 0 \\ -\infty \quad \text{otherwise} \end{cases} \\ &= \begin{cases} 0 & \text{if } a \leq \varepsilon, \mathrm{d}\sigma = \mathrm{d}\nu_{\mathcal{Z}}, \mathrm{d}\tau = \mathrm{d}\nu_{\mathcal{X}}, \mathrm{d}\zeta \leq \mathrm{d}\overline{\nu}_{\mathcal{Y}}, \mathrm{d}\omega \leq -\mathrm{d}\underline{\nu}_{\mathcal{Y}} \\ +\infty & \text{otherwise.} \end{cases} \\ \chi^*(\pi) &= \sup_{u \in \Gamma} \langle \pi, u \rangle - \chi(u) \\ &= \sup_{u \geq l(h_\theta(\cdot), \cdot)} \int_{\mathcal{Z} \times \mathcal{Z}} u\,\mathrm{d}\pi \\ &= \begin{cases} \int_{\mathcal{Z}} l(h_\theta, \mathbf{z})\,d\pi(\mathbf{z}) & \text{if } \mathrm{d}\pi \leq 0 \\ +\infty & \text{otherwise.} \end{cases} \end{aligned}$$

Above we have again used the fact that $l(h_\theta, \cdot)$ is bounded as an upper-semicontinuous function on a compact set. The resulting optimization problem, given by $\sup_{\pi \in \Gamma^*} -\tilde{G}^*(A^*\pi) -$

$\xi^*(-\pi)$, is the primal problem (21) with $-\tilde{G}^*(A^*\pi)$ expressing the Wasserstein and marginal constraints, and $-\chi^*(-\pi)$ expressing the objective and the positivity constraints on π. Strong duality follows from direct application of Theorem 3, which also gives us existence of a primal maximizer π^* to (21).

The second claim states that there exists a dual minimizer $(\alpha, \beta, \phi_{\theta,\alpha,\beta,\overline{\lambda},\underline{\lambda}}, \overline{\lambda}, \underline{\lambda}) \in \mathbb{R} \times \mathbb{R}^{N_l} \times L^0(\mathcal{X}, \mathbb{P}_{\mathcal{X}}) \times \mathbb{R}^{N_{\mathcal{Y}}} \times \mathbb{R}^{N_{\mathcal{Y}}}$ to (23) with

$$\phi_{\theta,\alpha,\beta,\overline{\lambda},\underline{\lambda}}(\mathbf{x}) = \max_{k\in\{1,\ldots,N_{\mathcal{Y}}\}, i\in\{1,\ldots,N_l\}} \ell(h_\theta, (\mathbf{x}, \mathbf{y}^k)) - \alpha c((\mathbf{x}, \mathbf{y}^k), \mathbf{z}_\ell^i) - \beta^i - (\overline{\lambda}^k - \underline{\lambda}^k).$$

We start by noting that the function $\phi_{\theta,\alpha,\beta,\overline{\lambda},\underline{\lambda}}$ is a pointwise maximum over a finite collection of functions $\ell(h_\theta, \cdot) - \alpha c(\cdot, \mathbf{z}_\ell^i)$, plus a constant term. If both $\ell(h_\theta, (\cdot, \mathbf{y}^k))$ and $c((\cdot, \mathbf{y}^k), \mathbf{z}_\ell^i)$, for all k, i, are measurable with respect to $\mathbb{P}_{\mathcal{X}}$, then $\phi_{\theta,\alpha,\beta,\overline{\lambda},\underline{\lambda}} \in L^0(\mathcal{X}, \mathbb{P}_{\mathcal{X}})$ as well. Moreover, if $\ell(h_\theta, \cdot)$ and $c(\cdot, \mathbf{z}_\ell^i)$ are continuous, then they are bounded (under the assumption that $\mathcal{X}$ is compact) and so $\phi_{\theta,\alpha,\beta,\overline{\lambda},\underline{\lambda}} \in C_b(\mathcal{X})$. In that case, for any finite $(\alpha, \beta, \overline{\lambda}, \underline{\lambda})$ satisfying $\alpha, \overline{\lambda}, \underline{\lambda} \geq 0$, the element $(\alpha, \beta, \phi_{\theta,\alpha,\beta,\overline{\lambda},\underline{\lambda}}, \overline{\lambda}, \underline{\lambda})$ is in Λ_θ.

For any probability measure $\pi \in \mathcal{Q}(\mathcal{Z} \times \mathcal{Z})$ whose first marginal satisfies $\pi((A \times \mathcal{Y}) \times \mathcal{Z}) = \mathbb{P}_{\mathcal{X}}(A)$, $\forall A \in \mathcal{B}(\mathcal{X})$, and whose second marginal satisfies $\operatorname{supp} \pi \subseteq \mathcal{Z} \times \hat{\mathcal{Z}}_l$, the following holds:

$$\mathbf{E}^{\mathbb{P}_{\mathcal{X}}} \phi_{\theta,\alpha,\beta,\overline{\lambda},\underline{\lambda}}(X) \geq \mathbf{E}^{\pi}[\ell(h_\theta, Z) - \alpha c(Z, Z') - \beta(Z') - (\overline{\lambda}(Y) - \underline{\lambda}(Y))], \tag{29}$$

with $(Z, Z') \sim \pi$ and $Z = (X, Y)$. Here we abuse notation slightly and write β as a function on $\mathcal{Z}$, with $\beta(\mathbf{z}_\ell^i) \triangleq \beta^i$, and $\overline{\lambda}, \underline{\lambda}$ as functions on $\mathcal{Y}$, with $\overline{\lambda}(\mathbf{y}^k) \triangleq \overline{\lambda}^k$ and $\underline{\lambda}(\mathbf{y}^k) \triangleq \underline{\lambda}^k$. The inequality holds necessarily because $\mathbf{E}^{\mathbb{P}_{\mathcal{X}}} \phi_{\theta,\alpha,\beta,\overline{\lambda},\underline{\lambda}}(\mathbf{x})$ is exactly the maximal value of the righthand side of the inequality, over π satisfying the above constraints.

Define $\Lambda_{\theta,*} = \{(\alpha, \beta, \overline{\lambda}, \underline{\lambda}) \in \mathbb{R} \times \mathbb{R}^{N_l} \times \mathbb{R}^{N_{\mathcal{Y}}} \times \mathbb{R}^{N_{\mathcal{Y}}} : \alpha, \overline{\lambda}, \underline{\lambda} \geq 0\}$ and

$$G_\theta : (\alpha, \beta, \overline{\lambda}, \underline{\lambda}) \in \Lambda_{\theta,*} \mapsto G(\alpha, \beta, \phi_{\theta,\alpha,\beta,\overline{\lambda},\underline{\lambda}}, \overline{\lambda}, \underline{\lambda}). \tag{30}$$

It is clear from statement of the dual problem in (23) that for any $(\alpha, \beta, \phi, \overline{\lambda}, \underline{\lambda}) \in \Lambda_\theta$, $G_\theta(\alpha, \beta, \overline{\lambda}, \underline{\lambda}) \leq G(\alpha, \beta, \phi, \overline{\lambda}, \underline{\lambda})$. This is because $\phi_{\theta,\alpha,\beta,\overline{\lambda},\underline{\lambda}}$ is the smallest function that satisfies the constraints in (23). Moreover, $G_\theta(\alpha, \beta, \overline{\lambda}, \underline{\lambda}) \geq \sup_{\pi\in\Phi} F_\theta(\pi)$ (i.e. the optimal value of the primal), as Fenchel duality guarantees existence of a primal optimizer π_* and the above fact bounding $\mathbf{E}^{\mathbb{P}_{\mathcal{X}}} \phi_{\theta,\alpha,\beta,\overline{\lambda},\underline{\lambda}}$ guarantees

$$\begin{aligned}
&G_\theta(\alpha, \beta, \overline{\lambda}, \underline{\lambda}) \\
&\qquad \geq \mathbf{E}^{\pi_*}\Bigg[\alpha(\varepsilon - c(Z, Z')) + \sum_{i=1}^{N_l} \left(\frac{1}{N_l} - \delta_{\mathbf{z}_\ell^i}(Z')\right) \beta(\mathbf{z}_\ell^i) \\
&\qquad\qquad + \ell(h_\theta, Z) + \sum_{k=1}^{N_{\mathcal{Y}}} (\overline{\mathbf{p}}_{\mathcal{Y}}^k - \delta_{\mathbf{y}^k}(Y))\overline{\lambda}(\mathbf{y}^k) + (\delta_{\mathbf{y}^k}(Y) - \underline{\mathbf{p}}_{\mathcal{Y}}^k)\underline{\lambda}(\mathbf{y}^k)\Bigg] \\
&\qquad \geq \mathbf{E}^{\pi_*} \ell(h_\theta, Z) \\
&\qquad = F_\theta(\pi_*).
\end{aligned} \tag{31}$$

The second inequality follows from feasibility of π_*, meaning that it satisfies the constraints (20). As we have shown, though, $\sup_{\pi \in \Phi} F_\theta(\pi) = \inf_{(\alpha,\beta,\phi,\overline{\lambda},\underline{\lambda}) \in \Lambda_\theta} G(\alpha, \beta, \phi, \overline{\lambda}, \underline{\lambda})$, so it must be that $\sup_{\pi \in \Phi} F_\theta(\pi) = \inf_{(\alpha,\beta,\overline{\lambda},\underline{\lambda}) \in \Lambda_{\theta,*}} G_\theta(\alpha, \beta, \overline{\lambda}, \underline{\lambda})$.

The claim therefore reduces to existence of a finite minimizer of G_θ. Suppose $\overline{\mathbf{p}}^k_{\mathcal{Y}} > \underline{\mathbf{p}}^k_{\mathcal{Y}}$ for all k and choose any $\pi \in \operatorname{relint} \Phi$, meaning that

$$\begin{aligned} &\mathbf{E}^\pi\, c(Z, Z') < \varepsilon, \\ &\mathbf{E}^\pi\, \delta_{\mathbf{z}^i_\ell}(Z') = \frac{1}{N_l}, \quad \forall i \in \{1, \dots, N_l\}, \\ &\mathbf{E}^\pi\, \delta_{\mathbf{y}^k}(Y) \in (\underline{\mathbf{p}}^k_{\mathcal{Y}}, \overline{\mathbf{p}}^k_{\mathcal{Y}}), \quad \forall k \in \{1, \dots, N_{\mathcal{Y}}. \end{aligned} \tag{32}$$

Substituting π for π_*, the bound in (31) still holds. Strict feasibility of π therefore implies that G_θ is lower bounded by a function linear in $(\alpha, \beta, \overline{\lambda}, \underline{\lambda})$ that is increasing in $\alpha, \overline{\lambda}, \underline{\lambda}$. Since these variables are constrained to be nonnegative, G_θ is therefore coercive in α, $\overline{\lambda}$, and $\underline{\lambda}$.

Unfortunately, any feasible π yields a lower bound that is independent of β. We can remedy this, however, as follows. Let $\{\alpha, \beta, \overline{\lambda}, \underline{\lambda}\} \in \Lambda_{\theta,*}$. Note that G_θ is invariant under shifts of β by a constant, so we can assume $\sum_{i=1}^{N_l} \beta^i = 0$. Therefore for any $\beta \neq 0$ there exists at least one pair of indices i, i' such that $\operatorname{sgn} \beta^i \neq \operatorname{sgn} \beta^{i'}$.

We will define a set of probability measures π_+ that have first marginal identical to that of π but that have a very small amount of mass shifted so as to alter the second marginal. Call this set $\Psi[\pi]$,

$$\Psi[\pi] = \mathcal{B}_a(\pi) \cap \{\pi_+ \in \mathcal{Q}(\mathcal{Z} \times \mathcal{Z}) : \pi_+(A \times \mathcal{Z}) = \pi(A \times \mathcal{Z}),\ \forall A \in \mathcal{B}(\mathcal{X})\}, \tag{33}$$

with $\mathcal{B}_a(\pi)$ the total variation norm ball of radius $a < \frac{\varepsilon - \mathbf{E}^\pi c(Z,Z')}{C}$, where $C = \max_{\mathbf{z},\mathbf{z}' \in \mathcal{Z}} c(\mathbf{z}, \mathbf{z}')$. Note that C is finite due to compactness of $\mathcal{X}$, as we assumed c is semicontinuous. And a is a radius sufficiently small to guarantee $\mathbf{E}^{\pi_+} c(Z, Z') < \varepsilon$, for all $\pi_+ \in \mathcal{B}_a(\pi)$, despite the shifted mass. a is positive due to strict feasibility of π.

For any $\pi_+ \in \Psi[\pi]$, the lower bound stated in (31) still holds for π_+ in place of π^*. This means that

$$\begin{aligned} &G_\theta(\alpha, \beta, \overline{\lambda}, \underline{\lambda}) \\ &\quad \geq \sup_{\pi_+ \in \Psi[\pi]} \mathbf{E}^{\pi_+} \Bigg[\alpha(\varepsilon - c(Z, Z')) + \sum_{i=1}^{N_l} \left(\frac{1}{N_l} - \delta_{\mathbf{z}^i_\ell}(Z') \right) \beta(\mathbf{z}^i_\ell) \\ &\qquad\qquad + \ell(h_\theta, Z) + \sum_{k=1}^{N_{\mathcal{Y}}} (\overline{\mathbf{p}}^k_{\mathcal{Y}} - \delta_{\mathbf{y}^k}(Y)) \overline{\lambda}(\mathbf{y}^k) + (\delta_{\mathbf{y}^k}(Y) - \underline{\mathbf{p}}^k_{\mathcal{Y}}) \underline{\lambda}(\mathbf{y}^k) \Bigg], \end{aligned} \tag{34}$$

for all $(\alpha, \beta, \overline{\lambda}, \underline{\lambda}) \in \Lambda_{\theta,*}$. This lower bound is a pointwise supremum over linear functions in $(\alpha, \beta, \overline{\lambda}, \underline{\lambda})$, induced by measures $\pi_+ \in \Psi[\pi]$. Importantly, the bound is still increasing in $\alpha, \overline{\lambda}, \underline{\lambda}$, by definition of $\Psi[\pi]$. And for any β, there exists $\pi_+ \in \Psi[\pi]$ such that $\operatorname{sgn}(\frac{1}{N_l} - \mathbf{E}^{\pi_+} \delta_{\mathbf{z}^i_\ell}(Z')) = \operatorname{sgn} \beta^i$, for all i. This pointwise supremum is increasing in $\alpha, \overline{\lambda}, \underline{\lambda}$, and increasing in $|\beta^i|$ for all i. So the lower bound is coercive and therefore G_θ is coercive. This suffices for existence of a finite $(\alpha, \beta, \overline{\lambda}, \underline{\lambda}) \in \Lambda_{\theta,*}$ that optimizes G_θ.

The above holds when $\overline{\mathbf{p}}^k_{\mathcal{Y}} > \underline{\mathbf{p}}^k_{\mathcal{Y}}$ for all k. Suppose now that there exists k such that $\overline{\mathbf{p}}^k_{\mathcal{Y}} = \underline{\mathbf{p}}^k_{\mathcal{Y}}$. Then we face the same problem as we did with β, with the linear lower bound defined by substituting strictly feasible π for π_* in (31) now independent of the $\overline{\lambda}^k$ and $\underline{\lambda}^k$ terms. We will deal with this problem analogously to the approach above.

We again start with $\pi \in \operatorname{relint} \Phi$. Now, however, we relax the constraints defining $\Psi[\pi]$, to allow small shifts of mass that preserve only the first $\mathcal{X}$-marginal. We define

$$\Psi_{\mathcal{X}}[\pi] = \mathcal{B}_a(\pi) \cap \{\pi_+ \in \mathcal{Q}(\mathcal{Z} \times \mathcal{Z}) : \pi_+((A \times \mathcal{Y}) \times \mathcal{Z}) = \pi(A)\ \forall A \in \mathcal{B}(\mathcal{X})\}, \tag{35}$$

with ball $\mathcal{B}_a(\pi)$ having radius $a < \min\left\{\frac{\varepsilon - \mathbf{E}^\pi c(Z,Z')}{C}, (\overline{\mathbf{p}}_{\mathcal{Y}}^k - \mathbf{E}^\pi \delta_{\mathbf{y}^k}(Y)), (\mathbf{E}^\pi \delta_{\mathbf{y}^k}(Y) - \underline{\mathbf{p}}_{\mathcal{Y}}^k)\right\}$ with k ranging over the indices such that $\overline{\mathbf{p}}_{\mathcal{Y}}^k \neq \underline{\mathbf{p}}_{\mathcal{Y}}^k$ and C defined as above.

For any $\pi_+ \in \Psi_{\mathcal{X}}[\pi]$, the lower bound in (31) still holds, when we substitute π_+ for π_*. So G_θ is lower bounded by a pointwise supremum over linear functions in $(\alpha, \beta, \overline{\lambda}, \underline{\lambda})$, identically to (34), substituting $\Psi_{\mathcal{X}}[\pi]$ for $\Psi[\pi]$. For each k with $\overline{\mathbf{p}}_{\mathcal{Y}}^k = \underline{\mathbf{p}}_{\mathcal{Y}}^k \triangleq \mathbf{p}_{\mathcal{Y}}^k$, however, the linear terms depending on $\overline{\lambda}^k, \underline{\lambda}^k$ simplify slightly and each one becomes $(\mathbf{p}_{\mathcal{Y}}^k - \mathbf{E}^{\pi_+} \delta_{\mathbf{y}^k}(Y))(\overline{\lambda}^k - \underline{\lambda}^k)$. Although $\overline{\lambda}^k$ and $\underline{\lambda}^k$ are nonnegative, their difference is unconstrained. G_θ is independent of shifts of $\overline{\lambda}^k, \underline{\lambda}^k$ by a constant, so we will assume $\min\{\overline{\lambda}^k, \underline{\lambda}^k\} = 0$ and discuss the single unconstrained variable $\lambda^k = \overline{\lambda}^k - \underline{\lambda}^k$.

If there exists at least one k' such that $\overline{\mathbf{p}}_{\mathcal{Y}}^{k'} \neq \underline{\mathbf{p}}_{\mathcal{Y}}^{k'}$ then for any $(\alpha, \beta, \overline{\lambda}, \underline{\lambda}) \in \Lambda_{\theta,*}$ it is possible to choose $\pi_+ \in \Psi_{\mathcal{X}}[\pi]$ such that $\operatorname{sgn}(\mathbf{p}^k - \mathbf{E}^{\pi_+} \delta_{\mathbf{y}^k}(Y)) = \operatorname{sgn} \lambda^k$, for all k such that $\overline{\mathbf{p}}_{\mathcal{Y}}^k = \underline{\mathbf{p}}_{\mathcal{Y}}^k$, by shifting mass between the set $\mathcal{X} \times \{\mathbf{y}^{k'} : \overline{\mathbf{p}}_{\mathcal{Y}}^{k'} \neq \underline{\mathbf{p}}_{\mathcal{Y}}^{k'}\} \times \mathcal{Z}$ and the set $\mathcal{X} \times \{\mathbf{y}^k : \overline{\mathbf{p}}_{\mathcal{Y}}^k = \underline{\mathbf{p}}_{\mathcal{Y}}^k\} \times \mathcal{Z}$, such that the corresponding terms have the correct sign. The radius a for $\Psi_{\mathcal{X}}[\pi]$ was chosen explicitly so that this movement of mass (combined with that described above for β) will yield $\pi_+ \in \Psi_{\mathcal{X}}[\pi]$ that is increasing in all of α, $\overline{\lambda}^{k'}$, and $\underline{\lambda}^{k'}$, $\overline{\mathbf{p}}_{\mathcal{Y}}^{k'} \neq \underline{\mathbf{p}}_{\mathcal{Y}}^{k'}$, while possessing the correct sign for β^i and λ^k, $\overline{\mathbf{p}}_{\mathcal{Y}}^k = \underline{\mathbf{p}}_{\mathcal{Y}}^k$. Therefore the supremum over $\Psi_{\mathcal{X}}[\pi]$ lower bounds G_θ and is coercive in $\alpha, \beta, \overline{\lambda}, \underline{\lambda}$.

If $\overline{\mathbf{p}}_{\mathcal{Y}}^k = \underline{\mathbf{p}}_{\mathcal{Y}}^k$ for all k, then there is an additional symmetry: We can shift λ^k by a constant without impacting G_θ. Therefore we can assume $\sum_{k=1}^{N_{\mathcal{Y}}} \lambda^k = 0$. The rest of the proof proceeds identically to that for β.

Note that this proof of the second claim relies on the existence of $\pi \in \operatorname{relint} \Phi$. Suppose no such π exists, meaning that for all $\pi \in \Phi$ either $\mathbf{E}^\pi c(Z, Z') = \varepsilon$ or $\mathbf{E}^\pi \delta_{\mathbf{y}^k}(Y) \in \{\overline{\mathbf{p}}_{\mathcal{Y}}^k, \underline{\mathbf{p}}_{\mathcal{Y}}^k\}$, for some k (or both). Suppose the latter is the case. Assuming the pairs $\{(\overline{\mathbf{p}}_{\mathcal{Y}}^k, \underline{\mathbf{p}}_{\mathcal{Y}}^k)\}_{k=1}^{N_{\mathcal{Y}}}$ are not degenerate in the sense that there exists only one probability vector $\mathbf{p}_{\mathcal{Y}} \in \Delta^{N_{\mathcal{Y}}}$ that satisfies the constraints $\mathbf{p}_{\mathcal{Y}}^k \in [\underline{\mathbf{p}}_{\mathcal{Y}}^k, \overline{\mathbf{p}}_{\mathcal{Y}}^k]$, for all k, we can shift any small amount of mass δ in π between the set $\mathcal{X} \times \{\mathbf{y}^k : \mathbf{E}^\pi \delta_{\mathbf{y}^k}(Y) = \overline{\mathbf{p}}_{\mathcal{Y}}^k \neq \underline{\mathbf{p}}_{\mathcal{Y}}^k\} \times \mathcal{Z}$ and the set $\mathcal{X} \times \{\mathbf{y}^k : \mathbf{E}^\pi \delta_{\mathbf{y}^k}(Y) < \overline{\mathbf{p}}_{\mathcal{Y}}^k \neq \underline{\mathbf{p}}_{\mathcal{Y}}^k\} \times \mathcal{Z}$ and likewise between the set $\mathcal{X} \times \{\mathbf{y}^k : \mathbf{E}^\pi \delta_{\mathbf{y}^k}(Y) = \underline{\mathbf{p}}_{\mathcal{Y}}^k \neq \overline{\mathbf{p}}_{\mathcal{Y}}^k\} \times \mathcal{Z}$ and the set $\mathcal{X} \times \{\mathbf{y}^k : \mathbf{E}^\pi \delta_{\mathbf{y}^k}(Y) > \underline{\mathbf{p}}_{\mathcal{Y}}^k \neq \underline{\mathbf{p}}_{\mathcal{Y}}^k\} \times \mathcal{Z}$, with the resulting altered probability measure now strictly feasible for radius ε increased by $C\delta$ (C defined as above, the maximal value of c). Moreover, if there is no such k with $\mathbf{E}^\pi \delta_{\mathbf{y}^k}(Y) \in \{\overline{\mathbf{p}}_{\mathcal{Y}}^k, \underline{\mathbf{p}}_{\mathcal{Y}}^k\}$, then $\mathbf{E}^\pi c(Z, Z') = \varepsilon$ and π will be strictly feasible for any radius ε that is increased by $\delta > 0$. So, assuming nondegenerate $\{(\underline{\mathbf{p}}_{\mathcal{Y}}^k, \overline{\mathbf{p}}_{\mathcal{Y}}^k)\}_{k=1}^{N_{\mathcal{Y}}}$, there exists at most one value of ε for which $\Phi \neq \emptyset$ but there is no $\pi \in \operatorname{relint} \Phi$. ■

Appendix B. Subgradients of dual program

Recall that $\Phi(\mathbf{x};\theta,\alpha,\beta,\overline{\lambda},\underline{\lambda})$ is defined in (13) as

$$\Phi(\mathbf{x};\theta,\alpha,\beta,\overline{\lambda},\underline{\lambda}) = \max_{\substack{k\in\{1,\ldots,N_{\mathcal{Y}}\}\\ i\in\{1,\ldots,N_l\}}} \ell(h_\theta,(X,\mathbf{y}^k)) - \left(\alpha c\left((X,\mathbf{y}^k),\mathbf{z}_\ell^i\right)+\beta^i\right) - (\overline{\lambda}^k-\underline{\lambda}^k).$$

For fixed $\mathbf{x}\in\mathcal{X}$, we can view $\Phi(\mathbf{x};\theta,\alpha,\beta,\overline{\lambda},\underline{\lambda})$ as a function of the dual variables. Algorithm 1 relies on computing a subderivative of $\Phi(\mathbf{x};\theta,\alpha,\beta,\overline{\lambda},\underline{\lambda})$ with respect to the variables $\theta,\alpha,\beta,\underline{\lambda},\overline{\lambda}$. Assuming that $\ell(h_\theta,\cdot)$ admits a subderivative for any θ, we can use the following expressions:

$$\begin{aligned}
\frac{\partial}{\partial\theta^j}\Phi(\mathbf{x};\theta,\alpha,\beta,\overline{\lambda},\underline{\lambda}) &\in \sum_{i=1}^{N_l}\sum_{k=1}^{N_{\mathcal{Y}}}\mathbb{1}_{V^{ik}}(\mathbf{x})\frac{\partial}{\partial\theta^j}\ell(h_\theta,(\mathbf{x},\mathbf{y}^k)),\\
\frac{\partial}{\partial\alpha}\Phi(\mathbf{x};\theta,\alpha,\beta,\overline{\lambda},\underline{\lambda}) &= -\sum_{i=1}^{N_l}\sum_{k=1}^{N_{\mathcal{Y}}}\mathbb{1}_{V^{ik}}(\mathbf{x})c((\mathbf{x},\mathbf{y}^k),\mathbf{z}_\ell^i),\\
\frac{\partial}{\partial\beta^i}\Phi(\mathbf{x};\theta,\alpha,\beta,\overline{\lambda},\underline{\lambda}) &= -\sum_{k=1}^{N_{\mathcal{Y}}}\mathbb{1}_{V^{ik}}(\mathbf{x}),\\
\frac{\partial}{\partial\overline{\lambda}^k}\Phi(\mathbf{x};\theta,\alpha,\beta,\overline{\lambda},\underline{\lambda}) &= -\sum_{i=1}^{N_l}\mathbb{1}_{V^{ik}}(\mathbf{x}),\\
\frac{\partial}{\partial\underline{\lambda}^k}\Phi(\mathbf{x};\theta,\alpha,\beta,\overline{\lambda},\underline{\lambda}) &= \sum_{i=1}^{N_l}\mathbb{1}_{V^{ik}}(\mathbf{x}).
\end{aligned}\tag{36}$$

For $\mathbf{x}$ lying on the boundary between two of the sets V^{ik}, we can obtain a subgradient by arbitrarily selecting only one of these V^{ik} to contain $\mathbf{x}$ when evaluating $\mathbb{1}_{V^{ik}}(\mathbf{x})$. In most practical settings, the boundaries should have lower dimension than $\mathcal{X}$ and therefore $\mathbb{P}_{\mathcal{X}}$-measure zero.

Appendix C. Data sets

In the experiments described in Sections 5.2, 5.1, and 6.3, we use 14 real data sets, taken from the UCI repository (Dua and Graff, 2019). Data sets were chosen from amongst those of "multivariate classification" type having more than 1000 examples. We attempted to select frequently-downloaded data sets from a variety of domains. We excluded data sets on which a ℓ^2-regularized linear logistic regression using 100 randomly selected training samples could not achieve likelihood greater than 0.55. Table 3 shows the data sets along with the number of examples and class balance in each data set.

In every experiment, the full data set was first standardized by subtracting out the mean and dividing by the standard deviation, per-feature, then scaling the resulting data matrix by dividing out the maximum absolute value over all entries.

Some caveats apply:

- **Abalone**: We excluded all examples from classes 9 and 10, to ensure non-trivial classification performance was possible with a small number of labeled examples.

Table 3: Data sets used in this paper. All are from the UCI repository (Dua and Graff, 2019).

Data set	Full name	N. features	N. examples	% positive
Abalone	Abalone	10	2854	50.7
Bank	Bank Marketing	53	10000	50.0
Cover (2/3)	Cover Type	54	10000	11.2
Cover (5/6)	Cover Type	54	10000	64.7
Isolet	Isolet	617	7797	19.2
Letter (C/E)	Letter Recognition	256	1504	48.9
Letter (U/V)	Letter Recognition	256	1577	51.6
Magic	MAGIC Gamma Telescope	10	13376	50.0
Mushroom	Mushroom	117	8124	48.2
Pulsar	HTRU2	8	3278	50.0
Spam	Spambase	57	4601	39.4
Thyroid	Thyroid Disease	21	7200	92.6
Wine	Wine	11	2700	39.3
Yeast	Yeast	8	1484	28.9

- **Bank** and **Cover**: We chose 10000 examples uniformly without replacement.
- **Cover**: We classified types 2 vs 3 and 5 vs. 6.
- **Isolet**: We classified vowels vs. the rest.
- **Letter recognition**: We classified “C” vs. “E” and “U” vs. “V.”
- **Bank**, **Magic**, and **Pulsar**: We chose examples uniformly without replacement from the larger class to achieve exact balance.
- **Thyroid**: We classified “3” vs. the rest.
- **Wine**: We excluded all examples with rating 6, to ensure non-trivial classification performance was possible with a small number of labeled examples.
- **Yeast**: We classified “nuclear” vs. the rest.

Appendix D. Description of experiments and additional empirical results

In all experiments we use the transport cost $c((\mathbf{x}, \mathbf{y}), (\mathbf{x}', \mathbf{y}')) = \|\mathbf{x} - \mathbf{x}'\|_2 + \frac{\kappa}{2}|\mathbf{y} - \mathbf{y}'|$ with $\kappa = 1$ and $\mathbf{y} \in \{+1, -1\} \subset \mathbb{R}$.

D.1 Performance When the Decision Set Contains the True Data Distribution

The following describes Figures 5 and 6 in Section 5.2 as well as the supplementary Figures 8 and 9. Figure 5 shows the worst-case likelihood bound output by each algorithm, varying the number of training examples. Specifically, we choose ε to be the smallest radius such that the Wasserstein ball $\mathcal{B}_\varepsilon(\hat{\mathbb{P}}_l)$ contains the true data distribution $\mathbb{P}$ (i.e. the empirical distribution defined by the full data set), and compute a linear logistic regression under the traditional Wasserstein DRL model (Equation (2), setting $\mathcal{P} = \mathcal{B}_\varepsilon(\hat{\mathbb{P}}_l)$) and under the proposed model (with $\mathcal{P} = \mathcal{B}_\varepsilon(\hat{\mathbb{P}}_l) \cap \mathcal{U}(\mathbb{P}_\mathcal{X}, \overline{\mathbf{p}}_\mathcal{Y}, \underline{\mathbf{p}}_\mathcal{Y})$). In both cases the problem can be written

$$\operatorname*{minimize}_{\theta\in\mathbb{R}^{q+1}} \sup_{\mu\in\mathcal{P}} \mathbf{E}^\mu\, Y \log(1+\exp\{-\langle X,\theta\rangle\}) + (1-Y)\log(1+\exp\{\langle X,\theta\rangle\}), \tag{37}$$

with $\mathcal{X} = \mathbb{R}^q \times \{1\}$ the feature space and $\mathcal{Y} = \{0,1\}$ the label space. There are two settings for the intervals $[\overline{\mathbf{p}}^k_\mathcal{Y}, \underline{\mathbf{p}}^k_\mathcal{Y}]$ constraining the label probabilities. The first is the "strong" prior, in which we know the exact label marginal $\mathbb{P}_\mathcal{Y} = \frac{1}{N_\mathcal{Y}} \sum_{k=1}^{N_\mathcal{Y}} \mathbf{p}^k_\mathcal{Y} \delta_{\mathbf{y}^k}$ and we set $\overline{\mathbf{p}}^k_\mathcal{Y} = \underline{\mathbf{p}}^k_\mathcal{Y} = \mathbf{p}^k_\mathcal{Y}$ for all k. The second setting is the "weak" prior, in which we estimate a 95% confidence interval for the label probability, directly from the labeled sample $\hat{\mathcal{Z}}_l$, using the method of Clopper and Pearson (1934).

We solve (37) via its dual (Section 3.3), using the Adam optimizer (Kingma and Ba, 2014) with $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-8}$, and a batch size of 100 and decreasing the learning rate by a factor of 8 every 10000 steps. The resulting dual objective value is used as the negative log of the worst-case likelihood bound shown in Figure 5.

Figure 6 shows the median over unlabeled input examples of the confidence of the learned predictor evaluated on those examples. For example $\mathbf{x} \in \mathcal{X}$ the confidence is $\max\{h_\theta(\mathbf{x}), 1 - h_\theta(\mathbf{x})\}$.

In both figures, the solid line is the median over 40 independent trials and the shaded region is the 95% interval of the median. Each trial represents a single independent sample of N_l labeled examples from the given data set, taken uniformly without replacement.

Figure 8 shows the worst-case likelihood bound for additional data sets and Figure 9 the median confidence.

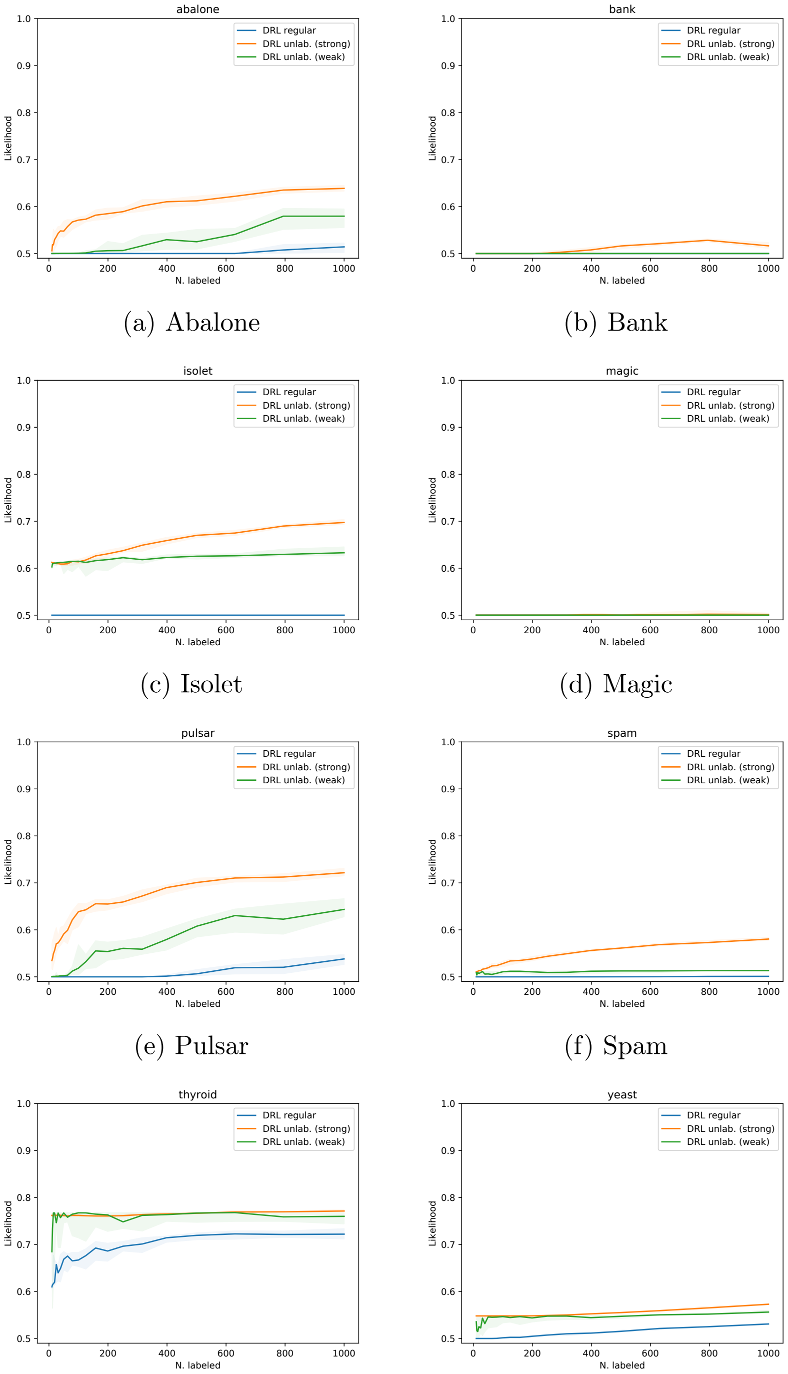
D.2 Traditional Wasserstein DRL Performance as We Vary ε

The following describes Figure 2 in Section 5.1 and Figure 10 in the appendix. The figures show out-of-sample generalization performance as well as confidence of predictors trained using traditional Wasserstein DRL, as we vary the radius of robustness ε. In particular, for each trial, we sample a data set of $N_l = 100$ labeled examples, which we use to compute the Wasserstein distributionally robust logistic regression (Abadeh et al., 2015), setting the radius ε to be a fixed percentage of the distance to the data distribution $\mathbb{P}$, i.e., the empirical distribution of the full data set. The log of this percentage is shown on the horizontal axis of the figure.

The figures show both the test set likelihood and the maximum over input examples of the confidence of the learned predictor, defined by $\max\{h_\theta(\mathbf{x}), 1 - h_\theta(\mathbf{x})\}$ for each $\mathbf{x} \in \mathcal{X}$. The solid line is the median over 100 trials while the shaded region shows the 95% interval of the median.

D.3 Traditional Wasserstein DRL Robustness Beyond the Decision Set

The following describes Figure 3 in Section 5.1 and 11 in the appendix. The figures show the worst-case performance of a predictor trained by traditional Wasserstein DRL with



(a) Abalone

(b) Bank

(c) Isolet

(d) Magic

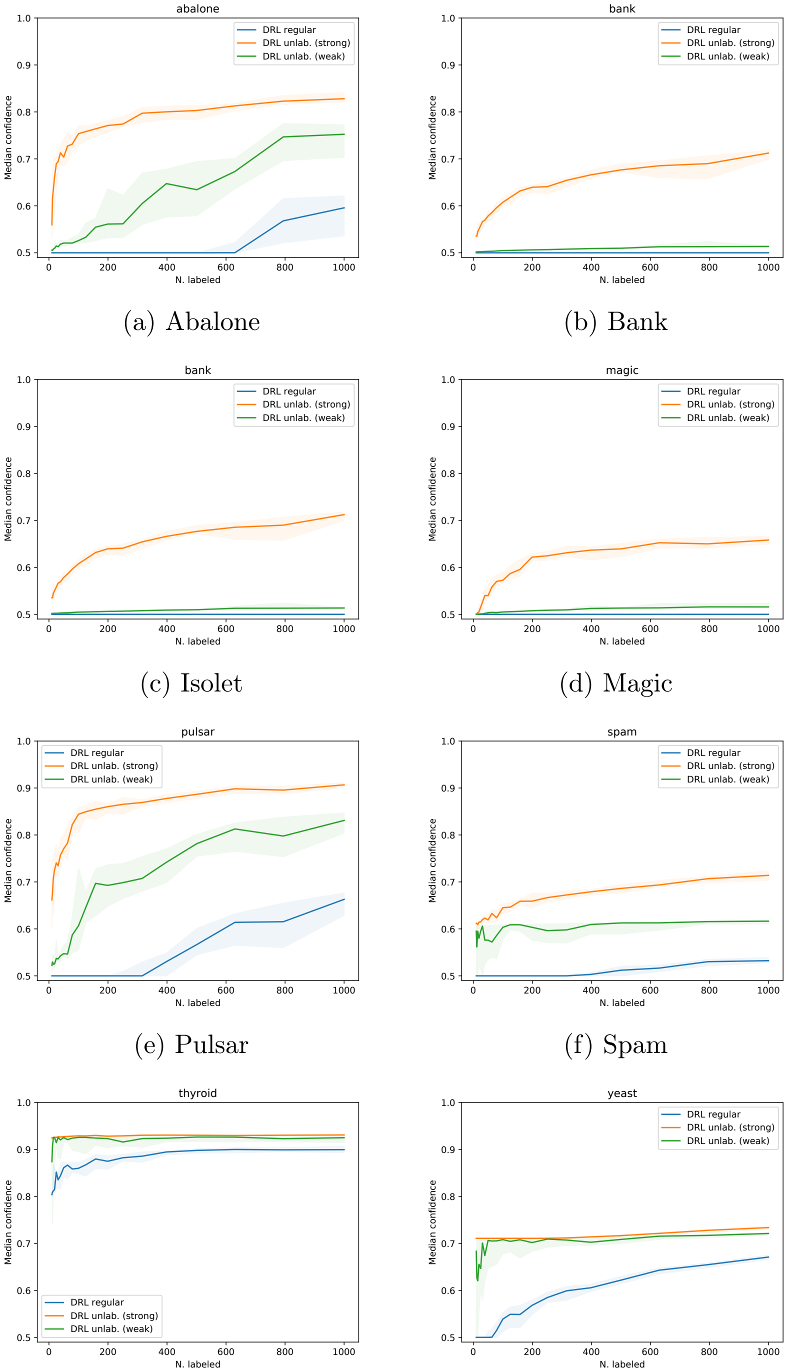
(e) Pulsar

(f) Spam

(g) Thyroid

(h) Yeast

Figure 8: Worst-case performance bound (likelihood) vs. number of labeled data, setting ε to include the true test distribution. The regular DRL bound is often vacuous through $N_l = 1000$ while both DRL methods with unlabeled data often yield non-vacuous bounds.



(a) Abalone

(b) Bank

(c) Isolet

(d) Magic

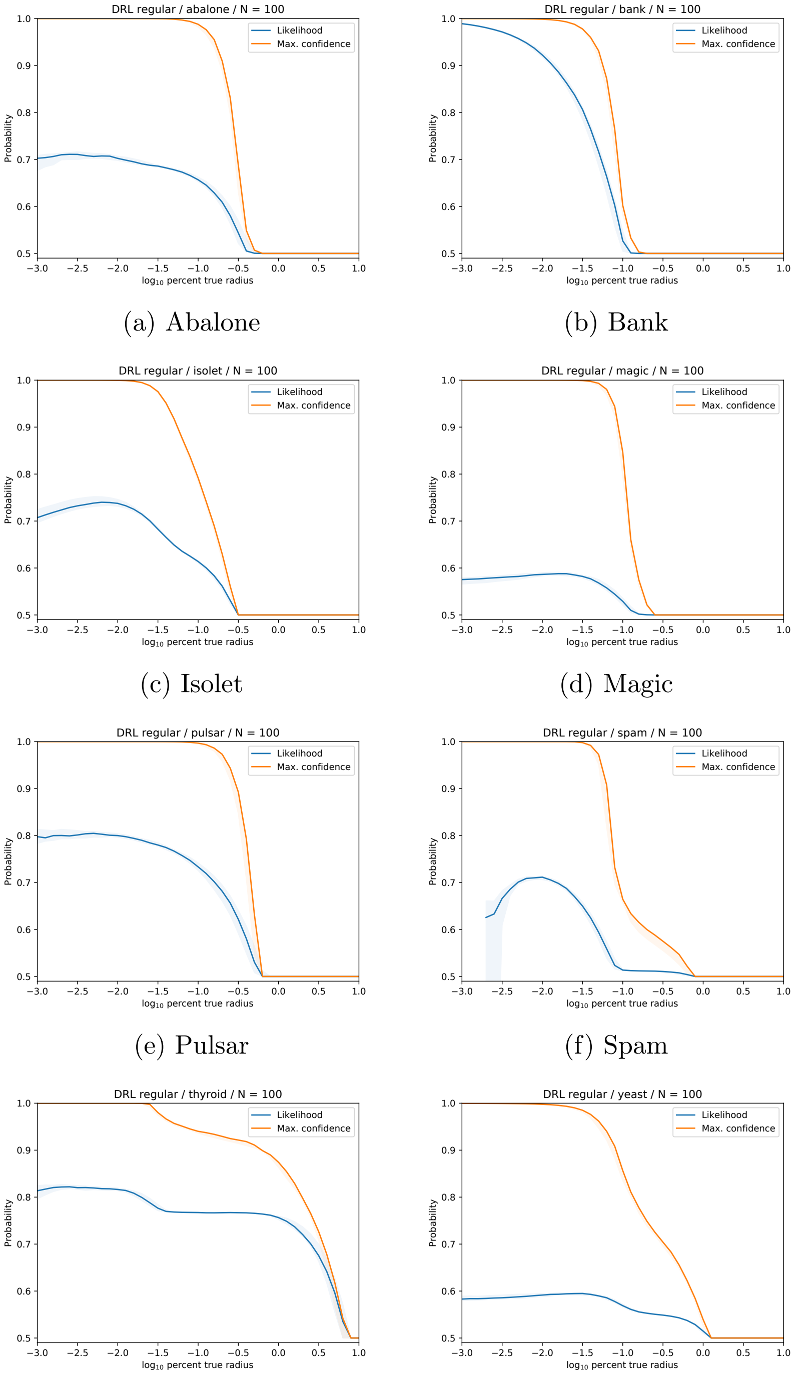
(e) Pulsar

(f) Spam

(g) Thyroid

(h) Yeast

Figure 9: Median confidence vs. number of labeled data, setting ε to include the true test distribution. The regular DRL predictor often has confidence close to 0.5 in settings where both DRL methods using unlabeled data yield non-trivial predictors.



(a) Abalone

(b) Bank

(c) Isolet

(d) Magic

(e) Pulsar

(f) Spam

(g) Thyroid

(h) Yeast

Figure 10: Traditional Wasserstein DRL. Out-of-sample performance (likelihood) and maximum confidence vs. radius of robustness ε as a percentage of the distance to the true data distribution $\mathbb{P}$. Performance shows a bias-variance tradeoff with peak at ε much smaller than the distance to $\mathbb{P}$. Confidence often drops sharply at a radius much smaller than the distance to $\mathbb{P}$.

a radius of robustness ε, when the test-time data distribution is allowed to come from a Wasserstein ball having the same center and slightly larger radius $\varepsilon + \Delta$. Specifically, for each trial, we sample a training set of $N_l = 100$ labeled examples, which we use to compute the Wasserstein distributionally robust logistic regression (Abadeh et al., 2015), fixing the radius ε of the underlying Wasserstein ball to a value between 10^{-3} and 10^0. For each radius ε, we obtain a set of learned parameters $\hat{\theta}_\varepsilon \in \Theta$. We then evaluate the worst-case value of the negative log-likelihood, fixing these parameters $\hat{\theta}_\varepsilon$, but increasing the radius of the underlying Wasserstein ball to $\varepsilon + \Delta$, for $\Delta \in [10^{-3}, 10^0]$. This worst-case value can be written

$$f(\hat{\theta}_\varepsilon, \Delta) = \sup_{\mu \in \mathcal{B}_{\varepsilon+\Delta}(\hat{\mathbb{P}}_l)} \mathbf{E}^\mu Y \log(1 + \exp\{-\langle X, \hat{\theta}_\varepsilon \rangle\}) + (1 - Y) \log(1 + \exp\{\langle X, \hat{\theta}_\varepsilon \rangle\}), \quad (38)$$

with $\mathcal{B}_{\varepsilon+\Delta}(\hat{\mathbb{P}}_l)$ the Wasserstein ball of radius $\varepsilon + \Delta$ centered at the empirical distribution of the labeled data $\hat{\mathbb{P}}_l$. This is exactly the inner problem of traditional Wasserstein DRL and is solved by the same mechanism, fixing the parameters $\hat{\theta}_\varepsilon$.

Each figure shows the median over trials of the resulting worst-case likelihood value, $\exp(-f(\hat{\theta}_\varepsilon, \Delta))$, as we vary ε (vertical axis) and Δ (horizontal axis). Each axis shows the base-10 log of the respective value. The color encodes the likelihood value, with blue indicating value 0, green value 0.5, and yellow value 1.

D.4 DRL with Unlabeled Data, Lack of Bias-Variance Tradeoff

The following describes Figure 4 in Section 5.1 as well as Figure 12 in the appendix. The figures show the out-of-sample performance of linear logistic regression models learned by the proposed DRL method (Section 3). Specifically, for each trial we sample 100 examples uniformly from the full data set, to form the training set $\hat{\mathcal{Z}}_l$. We then find a minimal radius ε_0 such that the feasible set $\mathcal{P} = \mathcal{B}_{\varepsilon_0}(\hat{\mathbb{P}}_l) \cap \mathcal{U}(\mathbb{P}_\mathcal{X}, \overline{\mathbf{p}}_\mathcal{Y}, \underline{\mathbf{p}}_\mathcal{Y})$ is non-empty, by doing a binary search for the minimal radius for which the value of the objective $g(\theta)$ (Section 3.3) is nonnegative. Here, $\mathbb{P}_\mathcal{X}$ is the $\mathcal{X}$-marginal of the true data distribution $\mathbb{P}$, which is taken to be the empirical distribution of the full data set. Given this radius ε_0, then we select radius $\varepsilon = \varepsilon_0 + \Delta$, for $\Delta \in [10^{-4}, 10^1]$, and solve the DRL problem

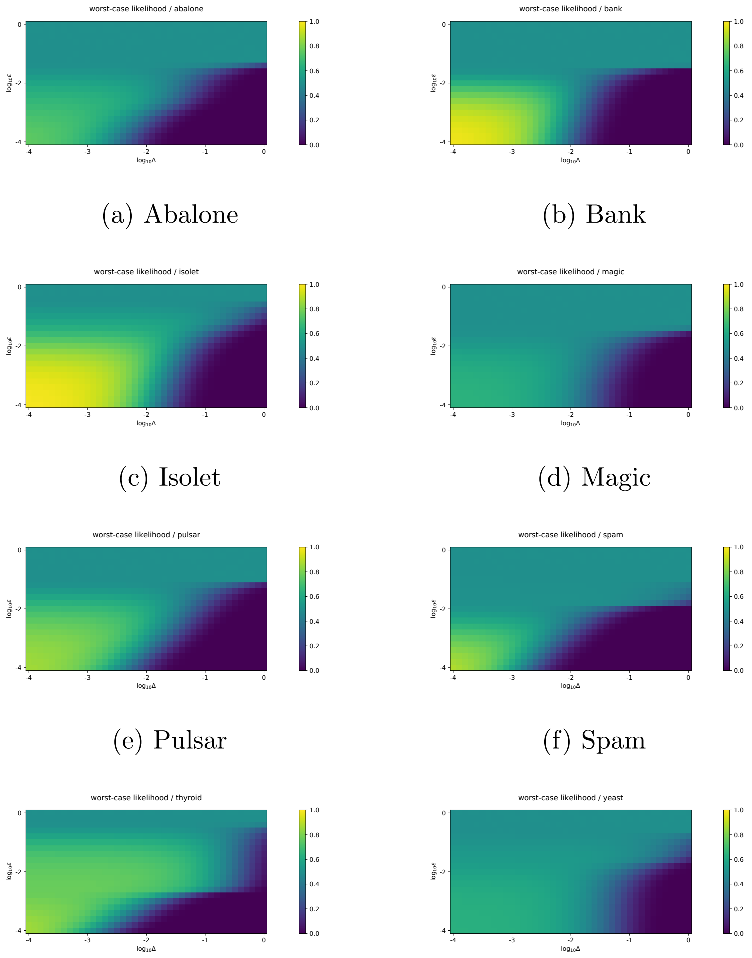
$$\underset{\theta \in \Theta}{\text{minimize}} \sup_{\mu \in \mathcal{B}_\varepsilon(\hat{\mathbb{P}}_l)} \mathbf{E}^\mu Y \log(1 + \exp\{-\langle X, \theta \rangle\}) + (1 - Y) \log(1 + \exp\{\langle X, \theta \rangle\}), \quad (39)$$

with $\mathcal{X} = \mathbb{R}^q \times \{1\}$ the feature space and $\mathcal{Y} = \{0, 1\}$ the label space. Here, and when finding ε_0, we choose $\overline{\mathbf{p}}_\mathcal{Y} = \underline{\mathbf{p}}_\mathcal{Y} = \mathbf{p}_\mathcal{Y}$, the true label probabilities from $\mathbb{P}$. This DRL problem is solved as described in Appendix D.1.

The solid lines in Figure 4 in Section 5.1 and Figure 12 in the appendix show the median over 100 trials of likelihood on the test set for the learned model as well as the median confidence, defined as in Appendix D.1. The shaded regions are 95% confidence intervals for the median. The horizontal axis shows the base-10 log of the excess radius Δ beyond ε_0.

D.5 Active Learning

The following describes Table 2 in Section 6.3. The table shows the area under the likelihood curve for model-change active learning heuristics and a random baseline applied to 14 binary



(a) Abalone

(b) Bank

(c) Isolet

(d) Magic

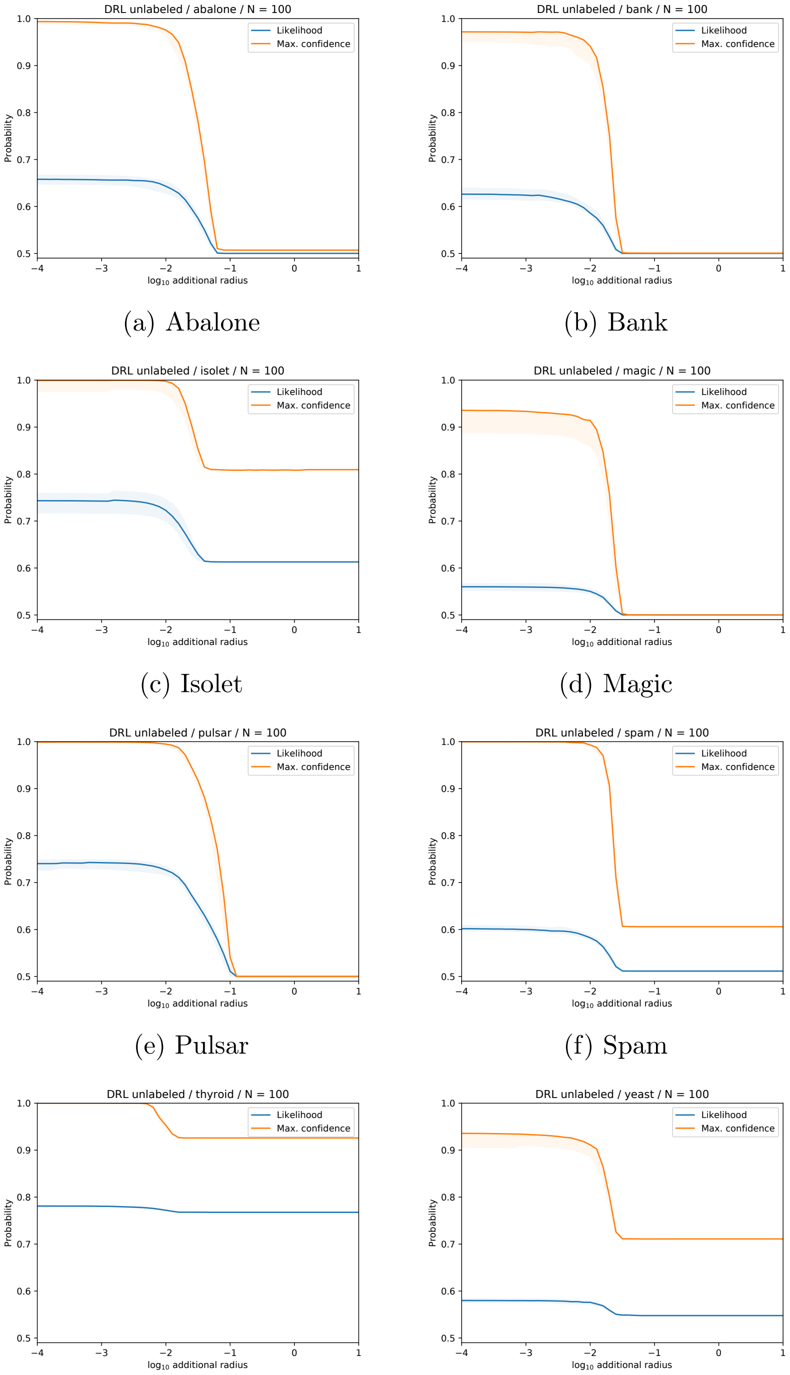
(e) Pulsar

(f) Spam

(g) Thyroid

(h) Yeast

Figure 11: Traditional Wasserstein DRL. Worst-case performance (likelihood) vs. radius of robustness ε and test-time data radius $\varepsilon + \Delta$. Yellow indicates perfectly correct prediction (likelihood 1), blue perfectly incorrect (likelihood 0), and green perfectly indecisive prediction (likelihood 0.5). Training with radius ε confers little robustness beyond ε.



(a) Abalone

(b) Bank

(c) Isolet

(d) Magic

(e) Pulsar

(f) Spam

(g) Thyroid

(h) Yeast

Figure 12: DRL with unlabeled data. Out-of-sample performance (likelihood) and maximum confidence vs. difference between the radius of robustness ε and the minimal radius necessary for the decision set to be nonempty. Unlike with traditional Wasserstein DRL, here there is no apparent bias-variance tradeoff. Performance is flat out to a radius at which the confidence drops sharply.

classification data sets. For each data set and each trial, we sample an initial set $\hat{\mathcal{Z}}_l$ of 20 labeled examples, with the remaining samples forming the unlabeled set $\hat{\mathcal{X}}_u$. We then use this labeled set to learn a linear logistic regression model, solving

$$\hat{\theta} = \operatorname*{argmin}_{\theta \in \Theta} \frac{1}{N_l} \sum_{i=1}^{N_l} \mathbf{y}_\ell^i \log(1 + \exp(-\langle \mathbf{x}_l^i, \theta \rangle)) + (1 - \mathbf{y}_\ell^i) \log(1 + \exp(\langle \mathbf{x}_l^i, \theta \rangle)) + \gamma \|\theta\|_2^2, \quad (40)$$

where $\hat{\mathcal{Z}}_l = \{(\mathbf{x}_l^i, \mathbf{y}_\ell^i)\}_{i=1}^{N_l}$ and $\gamma = 0.001$. Given $\hat{\theta}$, then, we apply each of the given active learning methods to select a new sample $\mathbf{x}_* \in \hat{\mathcal{X}}_u$ to label. Specifically,

1. **Random**. We sample $\mathbf{x}_*$ uniformly from $\hat{\mathcal{X}}_u$.

2. **EMC**. We compute for each $\mathbf{x}_u \in \hat{\mathcal{X}}_u$

$$\hat{g}(\mathbf{x}_u) = \frac{2\|\mathbf{x}_u\|_2}{(1 + \exp(-\langle \hat{\theta}, \mathbf{x}_u \rangle))(1 + \exp(\langle \hat{\theta}, \mathbf{x}_u \rangle))}, \quad (41)$$

   and select $\mathbf{x}_* = \operatorname{argmax}_{\mathbf{x}_u \in \hat{\mathcal{X}}_u} \hat{g}(\mathbf{x}_u)$.

3. **Min. MC**. We compute for each $\mathbf{x}_u \in \hat{\mathcal{X}}_u$

$$\hat{g}(\mathbf{x}_u) = \min \left\{ (1 + \exp(-\langle \mathbf{x}_u, \hat{\theta} \rangle))^{-1}, (1 + \exp(\langle \mathbf{x}_u, \hat{\theta} \rangle))^{-1} \right\}, \quad (42)$$

   and select $\mathbf{x}_* = \operatorname{argmax}_{\mathbf{x}_u \in \hat{\mathcal{X}}_u} \hat{g}(\mathbf{x}_u)$.

4. **Max. MC**. We compute for each $\mathbf{x}_u \in \hat{\mathcal{X}}_u$

$$\hat{g}(\mathbf{x}_u) = \max \left\{ (1 + \exp(-\langle \mathbf{x}_u, \hat{\theta} \rangle))^{-1}, (1 + \exp(\langle \mathbf{x}_u, \hat{\theta} \rangle))^{-1} \right\} \quad (43)$$

   and select $\mathbf{x}_* = \operatorname{argmax}_{\mathbf{x}_u \in \hat{\mathcal{X}}_u} \hat{g}(\mathbf{x}_u)$.

5. **DR (strong)**. We take 100 examples $\hat{\mathcal{X}}_u'$ uniformly at random from $\hat{\mathcal{X}}_u$ and compute for each $\mathbf{x}_u \in \hat{\mathcal{X}}_u'$

$$\hat{g}(\mathbf{x}_u) = \inf_{\mu \in \mathcal{P}} \mathbf{E}^\mu \frac{\delta_{\mathbf{x}_u}(X) \|X\|_2}{\hat{\varphi}(X)(1 + \exp(-Y \langle X, \hat{\theta} \rangle))}, \quad (44)$$

   with $\mathcal{P} = \mathcal{B}_\varepsilon(\hat{\mathbb{P}}_l) \cap \mathcal{U}(\mathbb{P}_\mathcal{X}, \overline{\mathbf{p}}_\mathcal{Y}, \underline{\mathbf{p}}_\mathcal{Y})$, $\hat{\varphi}(\mathbf{x}_u) = \frac{1}{N_u}$ for all $\mathbf{x}_u \in \hat{\mathcal{X}}_u$, and $\overline{\mathbf{p}}_\mathcal{Y} = \underline{\mathbf{p}}_\mathcal{Y} = \mathbf{p}_\mathcal{Y}$ the true label probabilities from $\mathbb{P}$. We select ε via a binary search for the minimal radius at which the feasible set $\mathcal{P}$ is non-empty, setting ε to be greater than this radius by a fixed margin $\Delta = 10^{-3}$. We select $\mathbf{x}_* = \operatorname{argmax}_{\mathbf{x}_u \in \hat{\mathcal{X}}_u} \hat{g}(\mathbf{x}_u)$.

6. **DR (weak)**. We take 100 examples $\hat{\mathcal{X}}_u'$ uniformly at random from $\hat{\mathcal{X}}_u$ and compute for each $\mathbf{x}_u \in \hat{\mathcal{X}}_u'$

$$\hat{g}(\mathbf{x}_u) = \inf_{\mu \in \mathcal{P}} \mathbf{E}^\mu \frac{\delta_{\mathbf{x}_u}(X) \|X\|_2}{\hat{\varphi}(X)(1 + \exp(-Y \langle X, \hat{\theta} \rangle))}, \quad (45)$$

with $\mathcal{P} = \mathcal{B}_\varepsilon(\hat{\mathbb{P}}_l) \cap \mathcal{U}(\mathbb{P}_\mathcal{X}, \overline{\mathbf{p}}_\mathcal{Y}, \underline{\mathbf{p}}_\mathcal{Y})$, $\hat{\varphi}(\mathbf{x}_u) = \frac{1}{N_u}$ for all $\mathbf{x}_u \in \hat{\mathcal{X}}_u$, and $[\underline{\mathbf{p}}^k_\mathcal{Y}, \overline{\mathbf{p}}^k_\mathcal{Y}]$ the 95% confidence intervals estimated from the original set of 20 labeled examples by the method of Clopper and Pearson (1934). We select ε as follows. We estimate ε_ℓ and ε_h via a binary search for the minimal radius at which the feasible set $\mathcal{P}$ is non-empty, setting $\overline{\mathbf{p}}^k_\mathcal{Y} = \underline{\mathbf{p}}^k_\mathcal{Y}$, i.e. the lower bound of the interval, for ε_ℓ, and vice versa, for ε_h. ε is then $\max\{\varepsilon_\ell, \varepsilon_h\} + \Delta$ with Δ a fixed margin $\Delta = 10^{-3}$. We select $\mathbf{x}_* = \operatorname{argmax}_{\mathbf{x}_u \in \hat{\mathcal{X}}_u} \hat{g}(\mathbf{x}_u)$.

We then acquire the true label $\mathbf{y}_*$ corresponding to $\mathbf{x}_*$, remove $\mathbf{x}_*$ from $\hat{\mathcal{X}}_u$, and add $(\mathbf{x}_*, \mathbf{y}_*)$ to the labeled set $\hat{\mathcal{Z}}_l$. We repeat the process above (beginning by estimating $\hat{\theta}$) until $\mathcal{Z}_\ell$ contains 100 samples.

Each time we estimate $\hat{\theta}$, we evaluate the performance of the learned classifier via the likelihood on the full data set (including $\hat{\mathcal{Z}}_l$). Taking the median over 50 trials, we obtain a likelihood curve (likelihood vs. number of labeled samples) for each active learning method, with the number of samples ranging from 20 to 100. The area under this curve is computed by the trapezoidal rule. The value shown in Table 2 is 100 times this area.

For the proposed distributionally robust heuristics, we solve the dual problem (Section 6.2) using the Adam optimizer (Kingma and Ba, 2014), setting $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-8}$, and a batch size of 100 and decreasing the learning rate by a factor of 10 every 5000 steps.

References

Soroosh Shafieezadeh Abadeh, Peyman Mohajerin Mohajerin Esfahani, and Daniel Kuhn. Distributionally robust logistic regression. In *Advances in Neural Information Processing Systems*, pages 1576–1584, 2015.

Aharon Ben-Tal, Dick Den Hertog, Anja De Waegenaere, Bertrand Melenberg, and Gijs Rennen. Robust solutions of optimization problems affected by uncertain probabilities. *Management Science*, 59(2):341–357, 2013.

Dimitris Bertsimas, Vishal Gupta, and Nathan Kallus. Data-driven robust optimization. *Mathematical Programming*, 167(2):235–292, 2018.

Jose Blanchet and Yang Kang. Sample out-of-sample inference based on wasserstein distance. *arXiv preprint arXiv:1605.01340*, 2016.

Jose Blanchet and Yang Kang. Semi-supervised learning based on distributionally robust optimization. *Data Analysis and Applications 3: Computational, Classification, Financial, Statistical and Stochastic Methods*, 5:1–33, 2020.

Jose Blanchet and Karthyek Murthy. Quantifying distributional model risk via optimal transport. *Mathematics of Operations Research*, 44(2):565–600, 2019.

Jose Blanchet, Yang Kang, and Karthyek Murthy. Robust wasserstein profile inference and applications to machine learning. *Journal of Applied Probability*, 56(3):830–857, 2019.

Mariusz Bojarski, Davide Del Testa, Daniel Dworakowski, Bernhard Firner, Beat Flepp, Prasoon Goyal, Lawrence D Jackel, Mathew Monfort, Urs Muller, Jiakai Zhang, et al. End to end learning for self-driving cars. *arXiv:1604.07316*, 2016.

Jonathan Borwein and Qiji Zhu. *Techniques of Variational Analysis.* Springer-Verlag New York, 2005.

Wenbin Cai, Yexun Zhang, Ya Zhang, Siyuan Zhou, Wenquan Wang, Zhuoxiang Chen, and Chris Ding. Active learning for classification with maximum model change. *ACM Transactions on Information Systems (TOIS)*, 36(2):15, 2017.

Giuseppe Carlo Calafiore and Laurent El Ghaoui. On distributionally robust chance-constrained linear programs. *Journal of Optimization Theory and Applications*, 130 (1):1–22, 2006.

Nicholas Carlini and David Wagner. Audio adversarial examples: Targeted attacks on speech-to-text. In *2018 IEEE Security and Privacy Workshops (SPW)*, pages 1–7. IEEE, 2018.

Yair Carmon, Aditi Raghunathan, Ludwig Schmidt, John C Duchi, and Percy S Liang. Unlabeled data improves adversarial robustness. In *Advances in Neural Information Processing Systems*, pages 11192–11203, 2019.

Kaixuan Chen, Lina Yao, Dalin Zhang, Xiaojun Chang, Guodong Long, and Sen Wang. Distributionally robust semi-supervised learning for people-centric sensing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3321–3328, 2019.

Ruidi Chen and Ioannis Ch Paschalidis. A robust learning approach for regression models based on distributionally robust optimization. *The Journal of Machine Learning Research*, 19(1):517–564, 2018.

Xin Chen, Melvyn Sim, and Peng Sun. A robust optimization perspective on stochastic programming. *Operations Research*, 55(6):1058–1071, 2007.

Sebastian Claici, Edward Chien, and Justin Solomon. Stochastic wasserstein barycenters. In *International Conference on Machine Learning*, pages 999–1008, 2018.

Charles J Clopper and Egon S Pearson. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, 26(4):404–413, 1934.

Jeremy M Cohen, Elan Rosenfeld, and J Zico Kolter. Certified adversarial robustness via randomized smoothing. *arXiv:1902.02918*, 2019.

Erick Delage and Yinyu Ye. Distributionally robust optimization under moment uncertainty with application to data-driven problems. *Operations Research*, 58(3):595–612, 2010.

Dheeru Dua and Casey Graff. UCI machine learning repository, 2019. URL `http://archive.ics.uci.edu/ml`.

John Duchi, Peter Glynn, and Hongseok Namkoong. Statistics of robust optimization: A generalized empirical likelihood approach. *arXiv:1610.03425*, 2016.

Gintare Karolina Dziugaite and Daniel M Roy. Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. *arXiv:1703.11008*, 2017.

Emre Erdoğan and Garud Iyengar. Ambiguous chance constrained problems and robust optimization. *Mathematical Programming*, 107(1-2):37–61, 2006.

Peyman Mohajerin Esfahani and Daniel Kuhn. Data-driven distributionally robust optimization using the Wasserstein metric: Performance guarantees and tractable reformulations. *Mathematical Programming*, 171(1-2):115–166, 2018.

Kevin Eykholt, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Chaowei Xiao, Atul Prakash, Tadayoshi Kohno, and Dawn Song. Robust physical-world attacks on deep learning models. *arXiv:1707.08945*, 2017.

Alexander Freytag, Erik Rodner, and Joachim Denzler. Selecting influential examples: Active learning with expected model output changes. In *European Conference on Computer Vision*, pages 562–577. Springer, 2014.

Charlie Frogner and Tomaso Poggio. Fast and flexible inference of joint distributions from their marginals. In *International Conference on Machine Learning*, pages 2002–2011, 2019.

Rui Gao and Anton J Kleywegt. Distributionally robust stochastic optimization with Wasserstein distance. *arXiv:1604.02199*, 2016.

Rui Gao, Xi Chen, and Anton J Kleywegt. Wasserstein distributional robustness and regularization in statistical learning. *arXiv:1712.06050*, 2017.

Joel Goh and Melvyn Sim. Distributionally robust optimization and its tractable approximations. *Operations Research*, 58(4-part-1):902–917, 2010.

Yves Grandvalet and Yoshua Bengio. Semi-supervised learning by entropy minimization. In *Advances in Neural Information Processing Systems*, pages 529–536, 2005.

Yuhong Guo and Dale Schuurmans. Discriminative batch mode active learning. In *Advances in Neural Information Processing Systems*, pages 593–600, 2008.

Awni Hannun, Carl Case, Jared Casper, Bryan Catanzaro, Greg Diamos, Erich Elsen, Ryan Prenger, Sanjeev Satheesh, Shubho Sengupta, Adam Coates, et al. Deep speech: Scaling up end-to-end speech recognition. *arXiv:1412.5567*, 2014.

Wassily Hoeffding. Probability Inequalities for Sums of Bounded Random Variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963. ISSN 0162-1459. doi: 10.2307/2282952. URL `https://www.jstor.org/stable/2282952`. Publisher: [American Statistical Association, Taylor & Francis, Ltd.].

Brody Huval, Tao Wang, Sameep Tandon, Jeff Kiske, Will Song, Joel Pazhayampallil, Mykhaylo Andriluka, Pranav Rajpurkar, Toki Migimatsu, Royce Cheng-Yue, et al. An empirical evaluation of deep learning on highway driving. *arXiv:1504.01716*, 2015.

Jean-Claude Junqua and Jean-Paul Haton. *Robustness in automatic speech recognition: fundamentals and applications*, volume 341. Springer Science & Business Media, 2012.

Gary King. *A solution to the ecological inference problem: Reconstructing individual behavior from aggregate data.* Princeton University Press, 2013.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv:1412.6980*, 2014.

Benoit Kloeckner. Approximation by finitely supported measures. *ESAIM: Control, Optimisation and Calculus of Variations*, 18(2):343–359, 2012.

Daniel Kuhn, Peyman Mohajerin Esfahani, Viet Anh Nguyen, and Soroosh Shafieezadeh-Abadeh. Wasserstein distributionally robust optimization: Theory and applications in machine learning. In *Operations Research & Management Science in the Age of Analytics*, pages 130–166. INFORMS, 2019.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv:1706.06083*, 2017.

Takeru Miyato, Shin-ichi Maeda, Masanori Koyama, Ken Nakae, and Shin Ishii. Distributional smoothing with virtual adversarial training. *arXiv:1507.00677*, 2015.

Amir Najafi, Shin-ichi Maeda, Masanori Koyama, and Takeru Miyato. Robustness to adversarial perturbations in learning from incomplete data. In *Advances in Neural Information Processing Systems*, pages 5541–5551, 2019.

Hongseok Namkoong and John C Duchi. Stochastic gradient methods for distributionally robust optimization with f-divergences. In *Advances in Neural Information Processing Systems*, pages 2208–2216, 2016.

Svetlozar T Rachev and Ludger Rüschendorf. *Mass Transportation Problems: Volume I: Theory*, volume 1. Springer Science & Business Media, 1998.

Aditi Raghunathan, Jacob Steinhardt, and Percy S Liang. Semidefinite relaxations for certifying robustness to adversarial examples. In *Advances in Neural Information Processing Systems*, pages 10877–10887, 2018.

Burr Settles, Mark Craven, and Soumya Ray. Multiple-instance active learning. In *Advances in Neural Information Processing Systems*, pages 1289–1296, 2008.

Soroosh Shafieezadeh-Abadeh, Daniel Kuhn, and Peyman Mohajerin Esfahani. Regularization via mass transportation. *Journal of Machine Learning Research*, 20(103):1–68, 2019.

Gagandeep Singh, Timon Gehr, Matthew Mirman, Markus Püschel, and Martin Vechev. Fast and effective robustness certification. In *Advances in Neural Information Processing Systems*, pages 10802–10813, 2018.

Aman Sinha, Hongseok Namkoong, and John Duchi. Certifying some distributional robustness with principled adversarial training. In *International Conference on Learning Representations*, 2018.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv:1312.6199*, 2013.

C. Villani. *Topics in Optimal Transportation Theory.* American Mathematical Society, 2003.

Wolfram Wiesemann, Daniel Kuhn, and Melvyn Sim. Distributionally robust convex optimization. *Operations Research*, 62(6):1358–1376, 2014.

Yazhou Yang and Marco Loog. A benchmark and comparison of active learning for logistic regression. *Pattern Recognition*, 83:401–415, 2018.

Xiaoyong Yuan, Pan He, Qile Zhu, and Xiaolin Li. Adversarial examples: Attacks and defenses for deep learning. *IEEE Transactions on Neural Networks and Learning Systems*, 2019.