Distributionally Robust Data Join

Pranjal Awasthi
Google Research, NY, USA

Christopher Jung
Stanford University, CA, USA

Jamie Morgenstern
University of Washington, Seattle, WA, USA

Abstract

Suppose we are given two datasets: a labeled dataset and unlabeled dataset which also has additional auxiliary features not present in the first dataset. What is the most principled way to use these datasets together to construct a predictor?

The answer should depend upon whether these datasets are generated by the same or different distributions over their mutual feature sets, and how similar the test distribution will be to either of those distributions. In many applications, the two datasets will likely follow different distributions, but both may be close to the test distribution. We introduce the problem of building a predictor which minimizes the maximum loss over all probability distributions over the original features, auxiliary features, and binary labels, whose Wasserstein distance is r_1 away from the empirical distribution over the labeled dataset and r_2 away from that of the unlabeled dataset. This can be thought of as a generalization of distributionally robust optimization (DRO), which allows for two data sources, one of which is unlabeled and may contain auxiliary features.

2012 ACM Subject Classification Theory of computation → Machine learning theory

Keywords and phrases Distributionally Robust Optimization, Semi-Supervised Learning, Learning Theory

Digital Object Identifier 10.4230/LIPIcs.FORC.2023.10

Related Version *Full Version*: `https://arxiv.org/abs/2202.05797`

1 Introduction

For a variety of prediction tasks, a number of sources of data may be available on which to train, each possibly following a distinct distribution. For example, health records might be available from at a number of geographically and demographically distinct hospitals. How should one combine these data sources to build the best possible predictor?

If the datasets S_1, S_2 follow different distributions D_1, D_2, the test distribution D will necessarily differ from at least one. A refinement of our prior question is to ask for which test distributions, then, can training with S_1, S_2 give a good predictor?

More generally, very common issues of mismatch between training and test distributions (and uncertainty over which test distribution one might face) has led to a great deal of interest in applying tools from distributionally robust optimization (DRO) to machine learning [12, 28, 24, 26]. In contrast to classical statistical learning theory, DRO picks a function f whose maximum loss (over a set of distributions near S) is minimized. This set of potential test distributions, often referred to as the ambiguity or uncertainty set, captures the uncertainty over the test distribution, along with knowledge that the test distribution will be close to the training distribution.

The ambiguity set is usually defined as a set of distributions with distance at most r from the empirical distribution over the training data: $B(\tilde{\mathcal{P}}_S, r) = \left\{\mathcal{Q} : D(\tilde{\mathcal{P}}_S, \mathcal{Q}) \leq r\right\}$ where $\tilde{\mathcal{P}}_S$ is the empirical distribution over training dataset S and D is some distance

measure between two probability distributions. Then, DRO aims to find a model θ such that for some loss ℓ, $\theta = \arg\min_\theta \sup_{\mathcal{Q} \in B(\tilde{\mathcal{P}}_S, r)} \mathbb{E}_{(x,y)\sim\mathcal{Q}}[\ell(\theta, (x, y))]$ – that is, minimize the loss over the worst case distribution in the ball of distributions $B(\tilde{\mathcal{P}}_S, r)$. The larger r, the more distributions over which DRO hedges its performance, leading to a tension between performance (minimizing worst-case error) and robustness (over the set of distributions on which performance is measured).

In this work, we introduce a natural extension of distributionally robust learning, *two anchor* distributionally robust learning, which we also refer to as the distributionally robust data join problem. Two anchor distributionally robust learning has access to two sources of training data, the first source containing labels, and the second source without labels but with auxiliary features not present in the first source. The optimization is then over the set of distributions close to *both* the labeled and auxiliary data distributions.

Formally, suppose one has two training datasets. The first dataset S_1 consists of feature vectors $\mathcal{X} = \mathbb{R}^{m_1}$ and binary prediction labels for some task $\mathcal{Y} = \{\pm 1\}$. The other dataset S_2 contains feature vectors $\mathcal{X}$ and auxiliary features $\mathcal{A} = \mathbb{R}^{m_2}$ but *not* the labels. The goal is to find a model θ that hedges its performance against any distribution $\mathcal{Q}$ over $(\mathcal{X}, \mathcal{A}, \mathcal{Y})$ whose Wasserstein distance is r_1 away from the empirical distribution over S_1 and r_2 away from that of S_2. Note that our setting is a strict generalization of semi-supervised setting: for $m_2 = 0$, there are no additional features in the second dataset, and S_2 is simply some additional unlabeled dataset. In contrast to pure semi-supervised settings, our method and setting both allow the learner to take advantage of the additional auxiliary features and to learn a model robust to additional distribution shift. We also emphasize that having the common features x between S_1 and S_2 help learn about the relationship between the auxiliary features a and the label y indirectly. Consider the following example where we actually have one dataset that contains the feature vector, auxiliary features, and the label altogether $S^{\text{combined}} = \{(x_i, a_i, y_i)\}_{i=1}^n$. From this dataset, we may form $S_1 = \{(x_i, a_i)\}_{i=1}^n$ and $S_2 = \{(x_i, y_i)\}_{i=1}^n$ where for every point (x_i, a_i) in S_1 and there's a corresponding (x_i, y_i) such that they share the same feature. In fact, instantiating our framework with $r_1 = 0$ and $r_2 = 0$ corresponds exactly to performing empirical risk minimization over S^{combined}. In other words, the quality of how well feature vectors x's match between S_1 and S_2 determine how well we may be able to learn the relationship between the auxiliary features a and the label y.

In practice, it is quite common to have the datasets fragmented as our setting captures. For instance, suppose some dataset has been collected at a hospital in order to build a predictive model that is to be used at a nearby hospital. After collecting this data, some other research may find other features that could have been useful for the prediction task but unfortunately were not collected during the contruction of this dataset. Fortunately, another nearby hospital may have data that contains both the original features and the useful auxiliary features but does not have labels for this prediction task. Our data join approach allows to find a model that utilizes such auxiliary features and explicitly considers the distribution mismatch between the hospital where the model is deployed and the hospitals from which these two datasets have been collected.

Auxiliary features may be useful not only for improving accuracy of the model but for guaranteeing additional properties including notions of fairness. In the appendix of the full version of the paper, we show that our distributionally robust data join problem encompasses a two-anchor distributionally robust learning instance where one can try to minimize not just the model's overall loss but also penalize the model for its difference in performance across demographic groups, even in situations where demographic information is present only in

one dataset and the label is only present in the other dataset. This extension is motivated by designing equitable predictors (e.g., which equalize false positive rate over a collection of demographic groups) where one training set contains labels for the relevant task but no demographic information, and another training set contains demographic information but may not contain task labels. Such settings are quite common in practice, where demographic data is not collected for every dataset – indeed, collection of demographic data is difficult to do well or sometimes even illegal [1, 15, 32, 34].

The contribution of our work can be summarized as follows:

1. New Problem Formulation of Distributionally Robust Data Join: we introduce and precisely formulate the distributionally robust data join problem in Section 2 and exactly characterize its feasibility in Section 3.1.
2. Application to Fairness: we further show how our original problem can be slightly modified to capture the problem of enforcing fairness when demographic group information is not available in the original labeled dataset (In the appendix of the full version of the paper).
3. Tractable Reformulation with an Approximation Guarantee (Theorem 7 in Section 3): we show how to approximate the distributionally robust data join problem with two convex optimization problems with an approximation guarantee.
4. Experiments (Section 4): we design and perform a synthetic experiment that shows how our distributionally robust data join method performs much better than the baselines. Additionally, we show some preliminary results on the experiments on a few real world datasets.

1.1 Related Work

Distributionally Robust Optimization: Prior work has looked at many different ways to define the ambiguity set: characterizing the set with moment and support information [8, 16, 33], or using various distance measures on probability space and defined the ambiguity set to be all the probability measures that are within certain distance ϵ of the empirical distribution: [12] use f-divergence, [18] the Kullback-Leibler divergence, [13] the Prohorov metric, and [28, 3, 2, 14] the Wasserstein distance, [17] chi-square divergence, and so forth. Defining ambiguity sets with divergence measures suffers from the fact that they do not incorporate the underlying geometry between the points – i.e. almost all divergence measures require the distribution in the ambiguity set to be absolutely continuous with respect to the anchor distribution. Therefore, because the distributions in the ambiguity set are simple re-weighting of the anchor distribution, divergence based ambiguity sets don't include distributions where the empirical distributions are perturbed a little bit and hence aren't robust to "black swan" outliers [23]. By contrast, the Wasserstein distance allows one to take advantage of the natural geometry of the points (e.g. L_p space). Furthermore, when we consider ambiguity sets defined by *two* anchor distributions as we do in this work, the two empirical distributions that are the anchors of the ambiguity set are almost surely not continuous with respect to each other. For these reasons, we focus on the Wasserstein distance in this work.

Most relevant to our work from the distributionally robust optimization literature is [28]. They show that regularizing the model parameter of the logistic regression has the effect of robustly hedging the model's performance against distributions whose distribution over just the covariates is slightly different than that of the empirical distribution over the training data. Distributionally robust logistic regression is a generalization of p-norm regularized logistic regression because it allows for not only distribution shift in the convariates but also the distribution shift over the labels. In a couple of real world datasets, they show that distributionally robust logistic regression seems to outperform regularized logistic regression

by the same amount that regularized logistic regression outperforms vanilla logistic regression. Our work is a natural extension of this work in that we take additional unlabeled dataset with auxiliary features into account. However, we remark that our contributions go beyond the contributions of [28]. In particular, reasoning about couplings between 3 distributions (labeled dataset, unlabeled dataset, and unknown target dataset) as shown later in Section 2.2 is *a priori* not obvious and rather novel. Existing 2 distribution coupling approach used in [28] (e.g., creating one coupling between labeled and unlabeled, and another between one of these and the test distribution) will not give empirically or theoretically good matchings between all three distributions and will generally also not be computationally tractable in our case. We further discuss new technical difficulties that have to be overcome in order to solve our problem later in Section 3 and the appendix of the full version of the paper. [30] extend [28] by adding a fairness regularization term, but the demographic information is available in the original labeled dataset in their setting unlike our setting.

Semi-supervised Learning: There have been significant advances in semi-supervised learning where the learner has access not only labeled data but also unlabeled data [36, 35, 7]. While our setting is similar to semi-supervised settings, we capture a broader class of possible problems in two ways. First, our approach allows the unlabeled dataset to have additional auxiliary features, and second, we explicitly take distribution shift into account.

Imputation: Numerous imputation methods for missing values in data exist, many of which have few or no theoretical guarantees [11, 27]. Many of these methods work best (or only have guarantees) when data values are missing at random. Our work, on the other hand, assumes all prediction labels are missing from the second dataset and all auxiliary features are missing from the first dataset. Another related problem is the matrix factorization problem which is also referred to as matrix completion problem [25, 22, 4]: here the goal is to find a low rank matrix that can well approximate the given data matrix with missing values. Our problem is different in that we don't make such structural assumption about the data matrix effectively being of low rank, but instead we assume all the auxiliary features are only available from a separate unlabeled dataset.

Fairness: Many practical prediction tasks have disparate performance across demographic groups, and explicit demographic information may not be available in the original training data. Several lines of work aim to reduce the gap in performance of a predictor between groups even without group information for training.

[17] show that the chi-square divergence between the overall distribution and the distribution of any subgroup can be bounded by the size of the subgroup: e.g. for any sufficiently large subgroup, its divergence to the overall distribution cannot be too big. Therefore, by performing distributionally robust learning with ambiguity set defined by chi-square divergence, they are able to optimize for the worst-case risk over all possible sufficiently large subgroups even when the demographic information is not available. [9] provide provably convergence oracle-efficient learning algorithms with the same kind of minimax fairness guarantees when the demographic group information is available.

One may naively think that given auxiliary demographic group information data, the most accurate imputation for the demographic group may be enough to not only estimate the unfairness of given predictor but also build a predictor with fairness guarantees. However, [1] show that due to different underlying base rates across groups, the Bayes optimal predictor for the demographic group information can result in maximally biased estimate of unfairness.

[10] demonstrate that one can rely on a multi-accurate regressor, which was first introduced by [21], as opposed to a 0-1 classifier in order to estimate the unfairness without any bias and also build a fair classifier for downstream tasks. When only some data points are missing demographic information, [19] show how to bypass the need to explicitly impute the missing values and instead rely on some decision tree based approach in order to optimize a fairness-regularized objective function. [20], given two separate datasets like in our setting, show how to construct confidence intervals for unfairness that is consistent with the given datasets via Fréchet and Hoeffding inequalities; our work is different in that we allow a little bit of slack by forming a Wasserstein ball around both datasets and can actually construct a fair model as opposed to only measuring unfairness.

[5] and [6] have shown when the demographic group information is available but possibly noisy, stochastically and adversarially respectively, how to build a fair classifier.

2 Preliminaries

2.1 Notations

We have two kinds of datasets, the auxiliary feature dataset and the prediction label dataset denoted in the following way: $S_A = \{(x_i^A, a_i^A)\}_{i=1}^{n_A}, \quad S_P = \{(x_i^P, y_i^P)\}_{i=1}^{n_P}$ where the domain for feature vector x is $\mathcal{X} = \mathbb{R}^{m_1}$, domain for auxiliary features a is $\mathcal{A} = \mathbb{R}^{m_2}$, and the label space is $y \in \mathcal{Y} = \{\pm 1\}$. For any vector $v \in \mathbb{R}^m$ and $d_1, d_2 \in [m]$, we write $v[d_1 : d_2]$ to denote the coordinates from d_1 to d_2 of vector v and $v[d]$ to denote the dth coordinate. We assume both $\mathcal{X}$ and $\mathcal{A}$ are compact and convex. For convenience, we write $S_A^{\mathcal{X}} = \{x : (x, a) \in S_A\}, \quad S_P^{\mathcal{X}} = \{x : (x, y) \in S_P\}$ to denote just the feature vectors of the dataset.

Given any dataset $S = \{z_i\}_{i=1}^n$, we will write $\tilde{\mathcal{P}}_S = \frac{1}{n}\sum_{i=1}^n \delta(z_i)$ to denote the empirical distribution over the dataset S where δ is the Dirac delta funcion. We'll write $\mathbb{P}_Z$ to denote the set of all probability distributions over Z. Similarly, we write $\mathbb{P}_{(Z,Z')}$ to denote a set of all possible joint distributions over Z and Z'. Also, given a joint distribution $\mathcal{P} \in \mathbb{P}_{(Z,Z')}$, we write $\mathcal{P}_Z$ and $\mathcal{P}_{Z'}$ to denote the marginal distribution over Z and Z' respectfully, meaning $\mathcal{P}_Z(z) = \int \mathcal{P}(z, dz')$ and $\mathcal{P}_{Z'}(z') = \int \mathcal{P}(dz, z')$. We extend the notation when the joint distribution is over more than two sets: e.g. $\mathcal{P}_{z,z'}((z, z')) = \int \mathcal{P}(z, z', dz'')$ where we have marginalized over Z'' for $\mathcal{P}$ which is a joint distribution over Z, Z', Z''.

We write the set of all possibly couplings between two distributions $\mathcal{P} \in \mathbb{P}_Z$ and $\mathcal{P}' \in \mathbb{P}_{Z'}$ as $\Pi(\mathcal{P}, \mathcal{P}') = \left\{\pi \in \mathbb{P}_{(Z,Z')} : \pi_Z = \mathcal{P}, \pi_{Z'} = \mathcal{P}'\right\}$. For a coupling between more than two distributions, we use the same convention and write $\Pi(\mathcal{P}, \mathcal{P}', \mathcal{P}'')$ for instance.

Given any metric $d : Z \times Z \to \mathbb{R}$ and two probability distributions $\mathcal{P}, \mathcal{P}' \in \mathbb{P}_Z$, we write the Wasserstein distance between them as $D_d(\mathcal{P}, \mathcal{P}') = \inf_{\pi \in \Pi(\mathcal{P}, \mathcal{P}')} \mathbb{E}_{(z,z') \sim \pi}[d(z, z')]$.

Given some distribution $\mathcal{P} \in \mathbb{P}$ over some set Z, metric $d : Z \times Z \to \mathbb{R}$, a radius $r > 0$, we will write $B_d(\mathcal{P}, r) = \{\mathcal{Q} \in \mathbb{P} : D_d(\mathcal{P}, \mathcal{Q}) \le r\}$ to denote the Wasserstein ball of radius r around the given distribution $\mathcal{P}$. When the metric is obvious from the context, we may simply write $B(\mathcal{P}, r)$.

In our case, the relevant metrics that are used to measure distance between points are

$$d_{\mathcal{X}}(x, x') = ||x - x'||_p, \quad d_A((x, a), (x', a')) = ||x - x'||_p + \kappa_A ||a - a'||_{p'}$$
$$d_P((x, y), (x', y')) = ||x - x'||_p + \kappa_P |y - y'|$$

where $||v||_p = (\sum_d |v[d]|^p)^{\frac{1}{p}}$ is some p-norm and $\kappa_A, \kappa_P \ge 0$ are the coefficients that control how much we care about the $||a - a'||_{p'}$ and $|y - y'|$. We'll write $||v||_{p,*} = \sup_{||v'||_p \le 1} \langle v, v' \rangle$ to denote dual norm for p-norm. Also, for convenience, given any vector v, we'll write

$\overline{v}_p = \frac{v}{||v||_p}$ and $\overline{v}_{p,*} = \frac{v}{||v||_{p,*}}$ to denote the normalized vectors. When it's clear from the context which norm is being used, we write $||\cdot||$, $||\cdot||_*$, $\overline{v}$, and $\overline{v}_*$. Now, we are ready to describe distributionally robust data join problem.

2.2 Distributionally Robust Data Join

We are given an auxiliary dataset S_A and a prediction label dataset S_P. We are interested in a joint distribution $\mathcal{Q}$ over (x, a, y) such that

1. its marginal distribution over (x, a) is at most r_A away from $\tilde{\mathcal{P}}_{S_A}$ in Wasserstein distance: $\mathcal{D}_{d_A}(\mathcal{Q}_{\mathcal{X},\mathcal{A}}, \tilde{\mathcal{P}}_{S_A}) \leq r_A$
2. its marginal distribution over (x, y) is at most r_P away from $\tilde{\mathcal{P}}_{S_P}$ in Wasserstein distance: $D_{d_P}(\mathcal{Q}_{\mathcal{X},\mathcal{Y}}, \tilde{\mathcal{P}}_{S_p}) \leq r_P$

Combining them together, the set of distributions we are interested in is

$$\begin{aligned} W(S_A, S_P, r_A, r_P) &= \{\mathcal{Q} \in \mathbb{P}_{(\mathcal{X},\mathcal{A},\mathcal{Y})} : \mathcal{D}_{d_A}(\mathcal{Q}_{\mathcal{X},\mathcal{A}}, \tilde{\mathcal{P}}_{S_A}) \leq r_A, D_{d_P}(\mathcal{Q}_{\mathcal{X},\mathcal{Y}}, \tilde{\mathcal{P}}_{S_p}) \leq r_P\} \\ &= \{\mathcal{Q} \in \mathbb{P}_{(\mathcal{X},\mathcal{A},\mathcal{Y})} : Q_{\mathcal{X},\mathcal{A}} \in B_{d_A}(\tilde{\mathcal{P}}_{S_A}, r_A), Q_{\mathcal{X},\mathcal{Y}} \in B_{d_P}(\tilde{\mathcal{P}}_{S_P}, r_P)\}. \end{aligned}$$

Now, we consider some learning task where the performance is measured according to the worst case distribution in the above set of distributions. We want to find some model parameter θ such that its loss against the worst-case distribution among $W(S_A, S_P, r_A, r_P)$ is minimized:

$$\min_{\theta \in \Theta} \sup_{\mathcal{Q} \in W(S_A, S_P, r_A, r_P)} \mathop{\mathbb{E}}_{(x,a,y)\sim\mathcal{Q}} [\ell(\theta, (x, a, y))]. \tag{1}$$

where $\ell : \Theta \times (\mathcal{X} \times \mathcal{A} \times \mathcal{Y}) \to \mathbb{R}$ is a convex loss function evaluated at θ. For the sake of concreteness, we focus on logistic loss[1] $\ell(\theta, (x, a, y)) = \log(1 + \exp(-y\langle\theta, (x, a)\rangle))$.

Also, we sometimes make use of the following functions $f(t) = \log(1 + \exp(t))$ and $h(\theta, (x, a)) = f(-\langle\theta, (x, a)\rangle)$ instead of ℓ, as it is more convenient due to not having to worry about y in certain cases: $\ell(\theta, (x, a, +1)) = h(\theta, (x, a))$ and $\ell(\theta, (x, a, -1)) = h(-\theta, (x, a))$. We write the convex conjugate of f as $f^*(b) = \sup_x \langle x^*, x\rangle - f(x)$, which in our case evaluates to $b \log b + (1 - b) \log(1 - b)$ when $b \in (0, 1)$, 0 if $b = 0$ or 1, and ∞ otherwise.

3 Tractable Reformulation

Let us give an overview of this section. Note that the optimization problem in (1) is a saddle point problem. In Section 3.1, we first make the coupling in the optimal transport more explicit in the inner sup term. Then, by leveraging Kantorovich duality, we replace the sup term with its dual problem which is a minimization problem, thereby making the original saddle problem into minimization problem. However, the resulting dual problem has constraints that involve some supremum term, meaning it's an semi-infinite program (i.e. $\sup_{z \in Z}$ constraint(z) ≤ 0 is equivalent to constraint(z) $\leq 0, \forall z \in Z$). Finally, in Section 3.3, we show how each supremum term can be approximated by some other closed-form constraint. And we finally show that the resulting problem can be decomposed into two convex optimization problems and its optimal solution has additional approximation guarantee to the original optimal solution (Theorem 7).

[1] All our results still hold for any other convex loss with minimal modifications

3.1 Formulation through Coupling

We show how to rewrite the problem (1) using the underlying coupling between the "anchor" distributions (S_A, S_P) and $\mathcal{Q} \in W(S_A, S_P, r_A, r_P)$. For simplicity, instead of $\pi\left((x_i^A, a_i^A), (x_j^P, y_j^P), (x, a, y)\right)$ which is a coupling between $\tilde{\mathcal{P}}_{S_A}$, $\tilde{\mathcal{P}}_{S_P}$, and some joint distribution $\mathcal{Q} \in \mathbb{P}_{\mathcal{X},\mathcal{A},\mathcal{Y}}$, we write $\pi_{i,j}^y(x, a) = \pi\left((x_i^A, a_i^A), (x_i^P, y_i^P), (x, a, y)\right)$. Then, since the "anchor" distributions $\tilde{\mathcal{P}}_{S_A}$ and $\tilde{\mathcal{P}}_{S_P}$ are discrete distributions, we can rewrite the problem (1) as choosing $\theta \in \Theta$ that minimizes the following value:

$$\sup_{\pi_{i,j}^{a,y}} \sum_{i=1}^{n_A}\sum_{j=1}^{n_P}\sum_{y\in\mathcal{Y}}\int_{\mathcal{X},\mathcal{A}} \ell(\theta,(x,a,y))\pi_{i,j}^y(dx,da) \tag{2}$$

$$\text{s.t.} \quad \sum_{i=1}^{n_A}\sum_{j=1}^{n_P}\sum_{y\in\mathcal{Y}}\int_{\mathcal{X},\mathcal{A}} d_A^i(x,a)\pi_{i,j}^y(dx,da) \le r_A, \quad \sum_{i=1}^{n_A}\sum_{j=1}^{n_P}\sum_{y\in\mathcal{Y}}\int_{\mathcal{X},\mathcal{A}} d_P^j(x,y)\pi_{i,j}^y(dx,da) \le r_P$$

$$\sum_{j=1}^{n_P}\sum_{y\in\mathcal{Y}}\int_{\mathcal{X},\mathcal{A}} \pi_{i,j}^y(dx,da) = \frac{1}{n_A} \quad \forall i \in [n_A], \quad \sum_{i=1}^{n_A}\sum_{y\in\mathcal{Y}}\int_{\mathcal{X},\mathcal{A}} \pi_{i,j}^y(dx,da) = \frac{1}{n_P} \quad \forall j \in [n_P]$$

where $d_A^i(x, a) = d_A((x_i^A, a_i^A), (x, a))$ and $d_P^j(x, y) = d_P((x_j^P, y_j^P), (x, y))$. We defer intuitive explanations and derivation of this problem to the appendix of the full version of the paper. For any fixed parameter θ, we'll denote the optimal value of the above problem (2) as $p^*(\theta, r_A, r_P)$ and $p^*(r_A, r_P) = \inf_\theta p^*(\theta, r_A, r_P)$.

It can be shown that minimizing over the above supremum value in (1) and the optimization problem (2) are equivalent as shown in the following theorem. We also provide a tight characterization of the feasibility of (2). The proof of Theorem 1 and 2 can be found in the appendix of the full version of the paper.

▶ **Theorem 1.** *For any fixed $\theta \in \Theta$,*

$$p^*(\theta, r_A, r_P) = \sup_{\mathcal{Q}\in W(S_A,S_P,r_A,r_P)} \mathbb{E}_{(x,a,y)\sim Q}[\ell(\theta,(x,a,y))].$$

▶ **Theorem 2.** *$D_{d_{\mathcal{X}}}(\tilde{\mathcal{P}}_{S_A^{\mathcal{X}}}, \tilde{\mathcal{P}}_{S_P^{\mathcal{X}}}) \le r_A + r_P$, if and only if there exists a feasible solution for (2).*

3.2 Strong Duality

We claim that the following problem is the dual to problem (2) and show that strong duality holds between them:

$$\inf_{\substack{\alpha_A,\alpha_P,\\ \{\beta_i\},\{\beta'_j\}}} \alpha_A r_A + \alpha_P r_P + \frac{1}{n_A}\sum_{i\in[n_a]}\beta_i + \frac{1}{n_P}\sum_{j\in[n_P]}\beta' \tag{3}$$

$$\text{s.t.} \sup_{(x,a)} \left(\ell(\theta,(x,a,y)) - \alpha_A d_A^i(x,a) - \alpha_P d_P^j(x,y)\right) \le \beta_i + \beta'_j \quad \forall i \in [n_A], j \in [n_P], y \in \mathcal{Y}$$

For fixed θ, we'll write $d^*(\theta, r_A, r_P)$ to denote the optimal value for the above dual problem (3). As in [28], strong duality directly follows from [29], but to be self-contained, we include the proof in the appendix of the full version of the paper, which follows the same proof structure presented in [31].

▶ **Theorem 3.** *If there exists a feasible solution for the primal problem (2), then we have that strong duality holds between the primal problem (2) and its dual problem (3): $p^*(\theta, r_A, r_P) = d^*(\theta, r_A, r_P)$ for fixed θ.*

In other words, we have successfully transformed the saddle point problem (1) into a minimization problem over θ and the dual variables $\alpha_A, \alpha_P, \{\beta_i\}$ and $\{\beta'_j\}_j$:

$$\min_{\substack{\theta\in\Theta,\alpha_A,\alpha_P,\\ \{\beta_i\},\{\beta'_j\}}} \alpha_A r_A + \alpha_P r_P + \frac{1}{n_A}\sum_{i\in[n_a]}\beta_i + \frac{1}{n_P}\sum_{j\in[n_P]}\beta'_j \tag{4}$$

$$\text{s.t.} \max_{y\in\{\pm 1\}} \sup_{(x,a)} \Big(\ell(\theta,(x,a,y)) - \alpha_A d_A^i(x,a) - \alpha_P d_P^j(x,y)\Big) \le \beta_i + \beta'_j \quad \forall i\in[n_A], j\in[n_P]$$

3.3 Replacing the sup Term

Note that $\sup_{(x,a)}$ in the constraint makes it hard to actually compute the expression: it's neither concave or convex in terms of (x,a) as it's the difference between convex functions $\ell(\theta,(x,a,y))$ and $\alpha_A d_A^i(x,a) + \alpha_P d_P^j(x,y)$. In that regard, we show how to approximate the sup term in the constraint of dual problem (3) with some closed form expression by extending the techniques used in [28] who study when there's only one "anchor" point – i.e. $\sup_x \ell(\theta,x) - \alpha d_{\mathcal{X}}(x_i,x)$ as opposed to in our case with two anchor points.

First, let's focus only on the terms that actually depend on (x,a) and ignore our dependence on y briefly:

$$\begin{aligned}&\sup_{(x,a)} \ell(\theta,(x,a,y)) - \alpha_A d_A^i(x,a) - \alpha_P d_P^j(x,y)\\ &= \kappa_P\alpha_P|y_j^P - y| + \left(\sup_{(x,a)} h(y\theta,(x,a)) - \alpha_A||x_i^A - x||_p - \alpha_P||x_j^P - x||_p + \alpha_A\kappa_A||a_i^A - a||_{p'}\right).\end{aligned}$$

We obtain an upper bound for the supremum term in the lemma below whose full proof can be found in the appendix of the full version of the paper.

▶ **Theorem 4.** *Fix any $y\in\mathcal{Y}$ and θ. Write $\theta_1 = \theta[1:m_1]$ and $\theta_2 = [m_1+1:m_1+m_2]$. Suppose $p\neq 1$ and $p\neq\infty$. If $||\theta_1||_{p,*} \le \alpha_A + \alpha_P$ and $||\theta_2||_{p',*} \le \kappa_A\alpha_A$, then*

$$\begin{aligned}&\sup_{(x,a)} h(y\theta,(x,a)) - \alpha_A||x_i^A - x||_p - \alpha_P||x_j^P - x||_p - \alpha_A\kappa_A||a_i^A - a||_{p'}\\ &\le f\left(\left(\frac{\min(\alpha_A,\alpha_P)||\theta_1||_*||x_i^A - x_j^P||}{\alpha_A+\alpha_P} + \frac{\langle y\theta_1, \alpha_A x_i^A + \alpha_P x_j^P\rangle}{\alpha_A+\alpha_P}\right) + \langle y\theta_2, a_i^A\rangle\right)\\ &\quad - \min(\alpha_A,\alpha_P)||x_i^A - x_j^P||_p.\end{aligned}$$

Otherwise, $\sup_{(x,a)} h(y\theta,(x,a)) - \alpha_A||x - x_i^A||_p - \alpha_P||x - x_j^P||_p - \alpha_A\kappa_A||a_i^A - a||_{p'}$ *evaluates to* ∞.

Proof Sketch. Similar to [28], we leverage convex conjugacy in order to re-express the sup term. However, because we have multiple anchor points, the re-expression results in an infimal convolution of *two* linear functions with bounded norm constraints as opposed to the case of [28] where they only have to handle a convex conjugate of a *single* linear function with bounded norm constraint and hence find an exact closed form expression. Therefore, in the appendix of the full version of the paper, we develop new techniques where we show (1) infimal convolution of linear functions with norm constraints is convex, (2) obtain a closed form solution of the infimal convolution at two extreme points, and (3) use linear interpolation of these extreme points to obtain an upper-bound, as a line segment of the two extreme points sits above the graph for convex functions. ◀

Equipped with the above upper bound on the supremum term, we can imagine trying to replace the supremum term with the above upper bound in order to get a feasible dual solution to the dual problem (4). However, one may worry that there is a big gap between the original supremum term and our upperbound in Theorem 4.

To this end, we further show that we can in fact approximate the supremum term with one more trick and hence obtain an approximate dual solution. Suppose we write $\hat{x}_{i,j} = \begin{cases} x_j^P & \text{if } \alpha_A < \alpha_P \\ x_i^A \end{cases}$ and $\hat{\alpha} = \min(\alpha_A, \alpha_P)$. Note that by definition, the value measured at $(\hat{x}_{i,j}, a_i^A)$ is a lower bound on the supremum. In other words, we have

$$\begin{aligned} &h(y\theta, (\hat{x}_{i,j}, a_i^A)) - \alpha_A ||x_i^A - \hat{x}_{i,j}||_p - \alpha_P ||x_j^P - \hat{x}_{i,j}||_p = f(\langle y\theta, (\hat{x}_{i,j}, a_i^A)\rangle) - \hat{\alpha}||x_i^A - x_j^P||_p \\ &\leq \sup_{(x,a)} h(y\theta, (x,a)) - \alpha_A ||x_i^A - x||_p - \alpha_P ||x_j^P - x||_p - \alpha_A \kappa_A ||a_i^A - a||_{p'} \\ &\leq f\left(\left(\frac{\min(\alpha_A, \alpha_P)||\theta_1||_* ||x_i^A - x_j^P||}{\alpha_A + \alpha_P} + \frac{\langle y\theta_1, \alpha_A x_i^A + \alpha_P x_j^P\rangle}{\alpha_A + \alpha_P}\right) + \langle y\theta_2, a_i^A\rangle\right) - \hat{\alpha}||x_i^A - x_j^P||_p. \end{aligned}$$

Now, via Hölder's inequality, we can show the lower bound and the upper bound above on the supremum term are in fact very close, meaning by using either the upper bound or the lower bound, we can approximate the supremum very well. Here's a lemma that shows that the value evaluated at $(\hat{x}_{i,j}, a_i^A)$ is pretty close to the upper bound in Theorem 4:

▶ **Lemma 5.**

$$\begin{aligned} &f\left(\left(\frac{\min(\alpha_A, \alpha_P)||\theta_1||_* ||x_i^A - x_j^P||}{\alpha_A + \alpha_P} + \frac{\langle y\theta_1, \alpha_A x_i^A + \alpha_P x_j^P\rangle}{\alpha_A + \alpha_P}\right) + \langle y\theta_2, a_i^A\rangle\right) - f(\langle y\theta, (\hat{x}_{i,j}, a_i^A)\rangle) \\ &\leq 2\hat{\alpha}||x_i^A - x_j^P||. \end{aligned}$$

In other words, replacing the original supremum constraint with a constraint evaluated at $(\hat{x}_{i,j}, a_i^A)$ will not incur too much additional error. Finally, using the fact that $f(-t) = f(t) + t$ for logistic function f, we can bring back the terms that depend on y and approximate the original supremum constraint in the following manner:

▶ **Corollary 6.**

$$\begin{aligned} &\left(\max_{y\in\{\pm 1\}} \sup_{(x,a)} \left(\ell(\theta, (x,a,y)) - \alpha_A d_A^i(x,a) - \alpha_P d_P^j(x,y)\right)\right) \\ &\quad - \left(f(\langle y_j^P \theta, (\hat{x}_{i,j}, a_i^A)\rangle) + \max(y_j^P \langle \theta, (\hat{x}_{i,j}, a_i^A)\rangle - \alpha_P \kappa_P, 0) - \hat{\alpha}||x_i^A - x_j^P||\right) \\ &\leq 2\hat{\alpha}||x_i^A - x_j^P|| \end{aligned}$$

In other words, replacing the supremum constraint with the constraint evaluated at $(\hat{x}_{i,j}, a_i^A)$ and using the above trick to remove the max over y will arrive at the following problem, for which we provide an approximation guarantee in Theorem 7.

$$\begin{aligned} &\min_{\alpha_A, \alpha_P, \theta_1, \theta_2, \{\beta_i\}, \{\beta'_j\}} (\alpha_A r_A + \alpha_P r_P) + \frac{1}{n_A} \sum_{i\in[n_a]} \beta_i + \frac{1}{n_P} \sum_{j\in[n_P]} \beta'_j \qquad (5) \\ &\text{s.t.} \quad f(y_j^P \langle \theta, (\hat{x}_{i,j}, a_i^A)\rangle) + \max(y_j^P \langle \theta, (\hat{x}_{i,j}, a_i^A)\rangle - \alpha_P \kappa_P, 0) - \hat{\alpha}||x_i^A - x_j^P|| \\ &\qquad \leq \beta_i + \beta'_j \quad \forall i \in [n_A], j \in [n_P] \\ &\qquad ||\theta_1||_* \leq \alpha_A + \alpha_P, ||\theta_2||_* \leq \kappa_A \alpha_A. \end{aligned}$$

▶ **Theorem 7.** *We can solve problem* (5) *by solving two convex optimization problems. And the optimal* θ^* *for the above problem* (5) *is such that its objective value for the original problem* (1) *is at most* $2\hat{\alpha} \max_{i\in[n_A], j\in[n_P]} ||x_i^A - x_j^P||$ *greater than the optimal solution:*

$$\begin{aligned} &\sup_{\mathcal{Q}\in W(S_A, S_P, r_A, r_P)} \mathbb{E}_{(x,a,y)\sim\mathcal{Q}} [\ell(\theta^*, (x,a,y))] - 2\hat{\alpha} \max_{i\in[n_A], j\in[n_P]} ||x_i^A - x_j^P|| \\ &\leq \min_{\theta\in\Theta} \sup_{\mathcal{Q}\in W(S_A, S_P, r_A, r_P)} \mathbb{E}_{(x,a,y)\sim\mathcal{Q}} [\ell(\theta, (x,a,y))]. \end{aligned}$$

Table 1 Average accuracy of each method over 10 experiment runs and standard deviations for synthetic dataset with a distribution shift.

	LR	RLR	DRLR	DJ
Accuracy	0.4126 ± 0.1049	0.5786 ± 0.3992	0.9068 ± 0.0076	$\mathbf{0.9923} \pm 0.0057$

Just as in [28], two convex optimization problems that problem (5) decomposes into can be solved by IOPT and YALMIP. In addition, we remark that $2\hat{\alpha} \max_{i \in [n_A], j \in [n_P]} ||x_i^A - x_j^P||$ is a reasonable approximation guarantee because this value should be in the same order as $\alpha_A r_A + \alpha_P r_P$: recall that we have argued in Theorem 2, a feasible solution exists if and only if $D_{d_{\mathcal{X}}}(\tilde{\mathcal{P}}_{S_A^{\mathcal{X}}}, \tilde{\mathcal{P}}_{S_P^{\mathcal{X}}}) \leq r_A + r_P$. Additionally, the worst case pairwise distance can actually be improved with an additional assumption: since any underlying coupling for the Wasserstein distance most likely transports non-zero probability mass between only close points, we can imagine considering only the k-nearest-neighbors of each point as opposed to all possible pairs between two datasets, hence decreasing the approximation error to the maximal pairwise distance between some point and its k-nearest-neighbor. We make this point more formal in the appendix of the full verison of the paper.

4 Experiments

We now describe an experimental evaluation of our method on a synethetic dataset and real world datasets. In all our epxeriments, we use the approach discussed in the appendix of the full version in which we make practically simplifying assumptions in order to solve the problem (5) via projected gradient descent. We use 2-norm throughout the experiments: i.e. $p, p' = 2$.

4.1 Synthetic Data

We briefly discuss how we create the snythetic dataset. We want our synthetic data generation process to encompass the components that are unique to our robust data join setting – namely, distribution shift and auxiliary unlabeled dataset that contains additional features that should help with the prediction task.

To that end, we discuss the data generation process at a high level here and more fully in Appendix B. We have two groups such that the ideal hyperplane that distinguishes the positive and negative points is different for each group. We introduce distribution shift into the setting by having the original labeled training dataset consist mostly of points from the first group and the test dataset consist mostly from the second group. As for specific details of the data generation process that are important for our setting, we have one of the features to carry information regarding which group the point belongs to.

As for the unlabeled dataset with auxiliary features, the points will mostly come from the second group, hence being closer to the test distribution. Furthermore, we include additional features that are present in the unlabeled dataset to be highly correlated with the true label, although this unlabeled dataset doesn't contain the true label of each point.

Because we want our baselines that compare our distributionally robust data join approach (DJ) against to be in the same model class (i.e. logistic regression) as our method for fair comparison, we consider the following baselines:

1. LR: Vanilla logistic regression trained on labeled dataset S_P
2. RLR: Regularized logistic regression trainined on labeled dataset S_P
3. DRLR: Distributionally robust logistic regression trainied on S_P

Table 2 Average accuracy of each method over 10 experiment runs and standard deviations for three UCI datasets.

	BC ($m_1 = 5$)	BC ($m_1 = 25$)	IO ($m_1 = 4$)	IO ($m_1 = 25$)	HD	1vs8
DJ	**0.9140** $\pm$ 0.0368	0.9281 $\pm$ 0.0155	**0.8208** $\pm$ 0.0816	**0.7896** $\pm$ 0.04885	**0.7495** $\pm$ 0.0374	**0.90841** $\pm$ 0.0270
LR	0.9012 $\pm$ 0.0294	0.9140 $\pm$ 0.0393	0.7764 $\pm$ 0.1560	0.7868 $\pm$ 0.0653	0.7286 $\pm$ 0.0504	0.8729 $\pm$ 0.0337
RLR	0.9053 $\pm$ 0.0228	**0.9287** $\pm$ 0.0199	0.7915 $\pm$ 0.1417	0.7868 $\pm$ 0.0690	0.7363 $\pm$ 0.0565	0.8953 $\pm$ 0.0250
LRO	0.8789 $\pm$ 0.0318	0.8789 $\pm$ 0.0318	0.7330 $\pm$ 0.0788	0.7330 $\pm$ 0.0788	0.6626 $\pm$ 0.0569	0.7766 $\pm$ 0.0599
RLRO	0.8953 $\pm$ 0.0212	0.8953 $\pm$ 0.0212	0.7377 $\pm$ 0.0800	0.7377 $\pm$ 0.0800	0.6714 $\pm$ 0.0568	0.8710 $\pm$ 0.0450
FULL	0.9684 $\pm$ 0.0143	0.9684 $\pm$ 0.0143	0.8754 $\pm$ 0.0764	0.8754 $\pm$ 0.0764	0.8319 $\pm$ 0.0311	0.9495 $\pm$ 0.0222

The result of this experiment can be found in Table 1. There are few plausible reasons as to why our approach (DJ) does extremely well in this synthetic experiment. Our distributionally robust data join is definitely taking advantage of the proximity of unlabeled dataset to the test distribution in that the majority of points are both from the second group. Although regularized and distributionally robust logistic regression is trying to be robust against some form of distribution shift, the set of distributions they are hedging against may be too big as they are hedging against all distributions that are close to the empirical distribution over the labeled dataset. By contrast, the set of distributions that distributionally robust data join may be smaller because it's hedging against the set of distributions that are close to the labeled dataset *and* the unlabeled dataset. Finally, auxiliary features in the unlabeled dataset are providing information very relevant for the prediction task.

4.2 UCI Datasets

Here we discuss some experiments we have run and show that as a proof of concept, our distributionally robust data join framework has the potential to be practical empirically. However, we remark unlike in the synthetic data experiment, we do not introduce any distribution shift (i.e. training and test are iid samples from the same distribution) and also choose the additional features for the unlabeled dataset in an arbitrary way because of our lack of contextual expertise of the features in each dataset. Therefore, the gaps between our method and the baselines we consider are not as impressive as the performance gap we see in the synthetic experiments.

We use four UCI datasets for our real world dataset experiment: Breast Cancer dataset (BC), Ionosphere dataset (IO), Heart disease dataset (HD), and Handwritten Digits dataset with 1's and 8's (1vs8). We provide more details about these datasets in Appendix B. For all these datasets, each experiment run consists of the following: (1) randomly divide the dataset into $S_{\text{train}} = \{(x_i, a_i, y_i)\}_{i=1}^{n_{\text{train}}}$ and S_{test}, (2) create the prediction label dataset and auxiliary dataset where v data points belong to both datasets: $S_P = \{(x_i, y_i)\}_{i=1}^{n_P+v}$ and $S_A = \{(x_i, a_i)\}_{n_P+1}^{n_{\text{train}}}$.

We compare our method of joining S_A and S_P, which we denote as DJ, to the following baselines:

1. LR: Logistic regression trained on S_P
2. RLR: Regularized logistic regression on S_P
3. LRO: Logistic regression on overlapped data $\{(x_i, a_i, y_i)\}_{i=n_P+1}^{n_P+v}$
4. RLRO: Regularized logistic regression on overlapped data $\{(x_i, a_i, y_i)\}_{i=n_P+1}^{n_P+v}$.
5. FULL: full training on $\{(x_i, a_i, y_i)\}_{i=1}^{n_{\text{train}}}$

where FULL is simply to show the highest accuracy we could have achieved if the labeled dataset actually had the auxiliary features and the unlabeled dataset had the labels. The results of the experiment can be found in Table 2, and we include further details of the

experiment in Appendix B. Without any distribution shift, the distributionally robust data join method is solving a somewhat harder problem than the other baselines because of its hedging against other nearby distributions. Yet it can be seen that the use of the additional auxiliary features through our data join method helps achieve better accuracy than the baselines.

References

1 Pranjal Awasthi, Alex Beutel, Matthäus Kleindessner, Jamie Morgenstern, and Xuezhi Wang. Evaluating fairness of machine learning models under uncertain and incomplete information. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 206–214, 2021.

2 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.

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

4 Emmanuel J Candès and Benjamin Recht. Exact matrix completion via convex optimization. *Foundations of Computational mathematics*, 9(6):717–772, 2009.

5 L Elisa Celis, Lingxiao Huang, Vijay Keswani, and Nisheeth K Vishnoi. Fair classification with noisy protected attributes: A framework with provable guarantees. In *International Conference on Machine Learning*, pages 1349–1361. PMLR, 2021.

6 L Elisa Celis, Anay Mehrotra, and Nisheeth K Vishnoi. Fair classification with adversarial perturbations. *arXiv preprint*, 2021. `arXiv:2106.05964`.

7 Olivier Chapelle, Bernhard Scholkopf, and Alexander Zien. Semi-supervised learning (chapelle, o. et al., eds.; 2006)[book reviews]. *IEEE Transactions on Neural Networks*, 20(3):542–542, 2009.

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

9 Emily Diana, Wesley Gill, Michael Kearns, Krishnaram Kenthapadi, and Aaron Roth. Minimax group fairness: Algorithms and experiments. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 66–76, 2021.

10 Emily Diana, Wesley Gill, Michael Kearns, Krishnaram Kenthapadi, Aaron Roth, and Saeed Sharifi-Malvajerdi. Multiaccurate proxies for downstream fairness. *arXiv preprint*, 2021. `arXiv:2107.04423`.

11 A Rogier T Donders, Geert JMG Van Der Heijden, Theo Stijnen, and Karel GM Moons. A gentle introduction to imputation of missing values. *Journal of clinical epidemiology*, 59(10):1087–1091, 2006.

12 John C Duchi and Hongseok Namkoong. Learning models with uniform performance via distributionally robust optimization. *The Annals of Statistics*, 49(3):1378–1406, 2021.

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

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

15 Allen Fremont, Joel S Weissman, Emily Hoch, and Marc N Elliott. When race/ethnicity data are lacking. *RAND Health Q*, 6:1–6, 2016.

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

17 Tatsunori Hashimoto, Megha Srivastava, Hongseok Namkoong, and Percy Liang. Fairness without demographics in repeated loss minimization. In *International Conference on Machine Learning*, pages 1929–1938. PMLR, 2018.

18 Zhaolin Hu and L Jeff Hong. Kullback-leibler divergence constrained distributionally robust optimization. *Available at Optimization Online*, 2013.

19 Haewon Jeong, Hao Wang, and Flavio P Calmon. Fairness without imputation: A decision tree approach for fair prediction with missing values. *arXiv preprint*, 2021. `arXiv:2109.10431`.

20 Nathan Kallus, Xiaojie Mao, and Angela Zhou. Assessing algorithmic fairness with unobserved protected class using data combination. *Management Science*, 2021.

21 Michael P Kim, Amirata Ghorbani, and James Zou. Multiaccuracy: Black-box post-processing for fairness in classification. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 247–254, 2019.

22 Yehuda Koren, Robert Bell, and Chris Volinsky. Matrix factorization techniques for recommender systems. *Computer*, 42(8):30–37, 2009.

23 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.

24 Jaeho Lee and Maxim Raginsky. Minimax statistical learning with wasserstein distances. In Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 2692–2701, 2018. URL: `https://proceedings.neurips.cc/paper/2018/hash/ea8fcd92d59581717e06eb187f10666d-Abstract.html`.

25 Andriy Mnih and Russ R Salakhutdinov. Probabilistic matrix factorization. In *Advances in neural information processing systems*, pages 1257–1264, 2008.

26 Hamed Rahimian and Sanjay Mehrotra. Distributionally robust optimization: A review. *arXiv preprint*, 2019. `arXiv:1908.05659`.

27 Patrick Royston. Multiple imputation of missing values. *The Stata Journal*, 4(3):227–241, 2004.

28 Soroosh Shafieezadeh-Abadeh, Peyman Mohajerin Esfahani, and Daniel Kuhn. Distributionally robust logistic regression. In Corinna Cortes, Neil D. Lawrence, Daniel D. Lee, Masashi Sugiyama, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 1576–1584, 2015. URL: `https://proceedings.neurips.cc/paper/2015/hash/cc1aa436277138f61cda703991069eaf-Abstract.html`.

29 Alexander Shapiro. On duality theory of conic linear problems. In *Semi-infinite programming*, pages 135–165. Springer, 2001.

30 Bahar Taskesen, Viet Anh Nguyen, Daniel Kuhn, and Jose Blanchet. A distributionally robust approach to fair classification. *arXiv preprint*, 2020. `arXiv:2007.09530`.

31 Cédric Villani. *Topics in optimal transportation*. American Mathematical Soc., 2003.

32 Joel S Weissman and Romana Hasnain-Wynia. Advancing health care equity through improved data collection. *The New England journal of medicine*, 364(24):2276–2277, 2011.

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

34 Yan Zhang. Assessing fair lending risks using race/ethnicity proxies. *Management Science*, 64(1):178–197, 2018.

35 Xiaojin Zhu and Andrew B Goldberg. Introduction to semi-supervised learning. *Synthesis lectures on artificial intelligence and machine learning*, 3(1):1–130, 2009.

36 Xiaojin Jerry Zhu. Semi-supervised learning literature survey. Technical report, University of Wisconsin-Madison Department of Computer Sciences, 2005.

A Possible Negative Societal Impact and Limitations

We do not foresee any direct negative societal impact of our work. However, just as other distributionally robust optimization methods, our robust guarantees may come at the price of achieving slightly worse accuracy. However, we note that this trade-off between more robustness and higher utility can be controlled by setting r_A and r_P appropriately. On a related note, another limitation of our approach is that it requires specifying r_A and r_P; one needs to have some knowledge about how "far" the distributions (i.e. labeled dataset, unlabeled dataset with auxiliary features, and test distribution) may be, which is a limitation as in other methods that require setting some hyperparameters appropriately.

B Missing Details from Section 4

All the experiments were performed on one of the authors' personal computer, MacBook Pro 2017, and every experiment took less than an hour.

We note that as it's standard in practice to output the last iterate instead of the averaged iterate, we use the last iterate of the projected gradient descent instead of the averaged one for all our experiments. Now, the total number of points and the features for each dataset is here along with where the dataset can be found:

1. BC (`https://archive.ics.uci.edu/ml/datasets/breast+cancer`): 569 points with 30 features
2. IO (`https://archive.ics.uci.edu/ml/datasets/ionosphere`): 351 points with 34 features
3. HD (`https://archive.ics.uci.edu/ml/datasets/Heart+Disease`): 300 points with 13 features
4. 1vs8 (`https://scikit-learn.org/stable/modules/generated/sklearn.datasets.load_digits.html#sklearn.datasets.load_digits`): This is a copy of the test dataset from `https://archive.ics.uci.edu/ml/datasets/Optical+Recognition+of+Handwritten+Digits`). It originally contains 1797 points with 64 points. But after filtering out all the digits except for 1's and 8's, there are 356 points.

For every dataset, we preprocess the data by standardizing each feature – that is, removing the mean and scaling to unit variance.

We take the common feature to be the first 5 features for (BC, HD) and 4 for IO – i.e. $m_1 = 5$ and 4 respectively. For 1vs8, we have $m_1 = 32$, the first half bits of the 8x8 image. And the remaining features are the auxiliary features $\mathcal{A}$: $m_2 = 25, 30, 8$, and 32 for BC, IO, HD, and 1vs8 respectively. For all datasets, we set the test size to be 30% of the entire dataset. Then, we set $(n_P, v) = (20, 5)$, $(20, 10)$, $(30, 5)$, $(30, 10)$ for BC, IO, HD, 1vs8 respectively. In other words, we imagine the total number of points in our labeled sets S_P and the number of features to be very small. For BC and IO, we also try a case when the number of common features is a lot more (i.e. $m_1 = 25$).

Now we report the best regularization penalties that maximize the accuracy of RLR and RLRO respectively over all experiment runs at the granularity level of 10^{-2}. The best regularization penalty for RLR and RLRO were $\lambda = (0.07, 0.04)$ for BC ($m_1 = 5$), $(0.04, 0.04)$ for BC ($m_1 = 25$), $(0.02, 0.02)$ for IO ($m_1 = 4$), $(0.01, 0.02)$ for IO ($m_1 = 25$), $(0.08, 0.03)$ for HD, and $(0.08, 0.08)$ for 1vs8. The parameters for data join used for each of the datasets can be found in the table below:

For all of the methods (logistic regression, regularized logistic regression, distributionally robust logistic regression, and our distributionally robust data join), the learning rate used was $7 * 10^{-2}$ and the total number of iterations was 1500.

Table 3 Parameters used for distributionally data join (DJ) for UCI datasets.

	BC ($m_1 = 5$)	BC ($m_1 = 25$)	IO ($m_1 = 4$)	IO ($m_1 = 25$)	HD	1vs8
r_A	0.65	1.65	0.3	1.5	0.65	1.85
r_P	0.65	1.65	0.3	1.5	0.65	1.85
κ_A	5	5	10	5	10	5
κ_P	2.5	2.5	5	2.5	5	15
k	1	1	1	1	1	1

Finally, we describe how we generated the data that was used to test how well DJ handles distribution shift. First, define

$$\beta_1 = [1, 0, 0, 0, 0, 0, 0, 0, 0, 0] \quad \text{and} \quad \beta_2 = [1, 1, 1, 1, 1, 1, 1, 1, 1, 1].$$

For the first group $g = 1$, the positive points and negative points were drawn from a multivariate normal distribution with mean β_1 and $-\beta_1$ respectively both with the standard deviation of 0.2:

$$x|y = +1, g = 1 \sim N(\beta_1, 0.2) \quad \text{and} \quad x|y = -1, g = 1 \sim N(-\beta_1, 0.2).$$

For the second group $g = 2$, the positive points and negative points were drawn from a multivariate normal distribution with mean β_2 and $-\beta_2$ respectively both with the standard deviation of 0.3:

$$x|y = +1, g = 2 \sim N(\beta_2, 0.2) \quad \text{and} \quad x|y = -1, g = 2 \sim N(-\beta_2, 0.2).$$

Now, for the first dataset $S_1 = \{(x_j^1, y_j^1)\}_{j=1}^{n_1}$, we had the number of points from group 1 and from group 2 was 400 and 20 respectively. And we had it so that the number of positive and negative points in each group was exactly the same: i.e. 200 positive and negative points for group 1, and 10 positive and 10 negative points for group 2.

For the second dataset, $S_2 = \{(x_i^2, y_i^2)\}_{i=1}^{n_2}$, the number of points from group 1 and from group 2 was 200 and 2000 respectively. The number of positive and negative points in each group was exactly the same once again here.

Our labeled dataset will be the first two coordinates of the fist dataset, meaning $m_1 = 2$:

$$S_P = \{(x_j^1[0:2], y_j^1)\}_{j=1}^{n_1}.$$

Then, we will randomly divide the second dataset so that the 70% of it will be used as unlabeled dataset S_A and the other 30% is to be used as the test dataset S_{test}.

$$S_A = \{x_i^2\}_{i=1}^{0.7n_2} \quad \text{and} \quad S_{\text{test}} = \{(x_i^2, y_i^2)\}_{i=0.7n_2+1}^{n_2}.$$

Note that $m_2 = 10$.

The baselines that we consider for this synthetic data experiment are

1. Logistic regression trained (LR) on S_P
2. Regularized regression trained (RLR) on S_P with $\lambda = 10$
3. Distributionally logistic regression (DLR) trained on S_P with $r = 100, \kappa = 10$.