Distributionally Robust Data Join

Pranjal Awasthi* Christopher Jung† Jamie Morgenstern‡

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.

*Google

†University of Pennsylvania. Part of this work done while an intern at Google Research New York.

‡University of Washington

Contents

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 $\mathcal{P}_1, \mathcal{P}_2$, the test distribution $\mathcal{P}$ 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) have led to a great deal of interest in applying tools from distributionally robust optimization (DRO) to machine learning [Duchi and Namkoong, 2021, Shafieezadeh-Abadeh et al., 2015, Lee and Raginsky, 2018, Rahimian and Mehrotra, 2019]. 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,

$$\theta = \arg\min_{\theta} \sup_{\mathcal{Q} \in B(\tilde{\mathcal{P}}_S, r)} \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{Q}} [\ell(\theta, (x, y))].$$

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 auxilliary features not present in the first source. The optimization is then over the set of distributions close to *both* the labeled and auxilliary data distributions.

Formally, suppose one has two training datasets. The first dataset S_1 consists of feature vectors $\mathcal{X} \subseteq \mathbb{R}^{m_1}$ and binary prediction labels for some task $\mathcal{Y} = \{\pm 1\}$. The other dataset S_2 contains feature vectors $\mathcal{X}$ and auxiliary features $\mathcal{A} \subseteq \mathbb{R}^{m_2}$ 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 allow the learner to not only take advantage of the additional auxiliary features but also learn a model robust to additional distribution shift.

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 have found other useful features that could have been collected for the prediction task. 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.

Auxilliary features may be useful not only for improving accuracy of the model but for guaranteeing additional properties including notions of fairness. We show that one can solve a two-anchor distributionally

robust learning instance penalizing models for their difference in performance across demographic groups, where demographic information is present only in one 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 [Awasthi et al., 2021, Fremont et al., 2016, Weissman and Hasnain-Wynia, 2011, Zhang, 2018].

The contribution of our work can be summarized as follows:

1. New problem formulation of distributionally robust data join (Section 2.2),
2. Tractable reformulation with an approximation guarantee (Section 3 and Section 4): we show how to approximate the distributionally robust data join problem with a tractable convex optimization problem with an approximation guarantee,
3. Applications to fairness with missing demographic group information (Section 5): with slight modifications, we show how to penalize the model for its unfairness even when the labeled dataset lacks the demographic group information,
4. Experiments (Section 6): we perform some experiments to demonstrate the usefulness of our distributionally robust data join method.

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 [Delage and Ye, 2010, Goh and Sim, 2010, Wiesemann et al., 2014], 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: Duchi and Namkoong [2021] use f-divergence, Hu and Hong [2013] the Kullback-Leibler divergence, Erdoğan and Iyengar [2006] the Prohorov metric, and Shafieezadeh-Abadeh et al. [2015], Blanchet and Murthy [2019], Blanchet et al. [2019], Esfahani and Kuhn [2018] the Wasserstein distance, Hashimoto et al. [2018] chi-square divergence, and so forth. In this work, we focus on the Wasserstein distance.

Most relevant to our work within literature on distributionally robust optimization literature is that of Shafieezadeh-Abadeh et al. [2015]. 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 a distribution shift not only in the covariates but also 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. Taskesen et al. [2020] extend Shafieezadeh-Abadeh et al. [2015] by adding a fairness regularization term as we also do, but the demographic information is not available in the original training data in their setting.

Semi-supervised Learning: There have been significant advances in semi-supervised learning where the learner has access to not only labeled data but also unlabeled data [Zhu, 2005, Zhu and Goldberg, 2009, Chapelle et al., 2009]. While our model subsumes 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 [Donders et al., 2006, Royston, 2004]. 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 [Mnih and Salakhutdinov, 2008, Koren et al., 2009, Candès and Recht, 2009]: 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 when the group information may not be directly available during training.

Hashimoto et al. [2018] 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. Diana et al. [2020] 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, Awasthi et al. [2021] 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. Diana et al. [2021] demonstrate that one can rely on a multi-accurate regressor, which was first introduced by Kim et al. [2019], 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, Jeong et al. [2021] 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. Kallus et al. [2021], 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. Celis et al. [2021a] and Celis et al. [2021b] show 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 $\mathcal{X} \subseteq \mathbb{R}^{m_1}$, the domain for auxiliary features a is $\mathcal{A} \subseteq \mathbb{R}^{m_2}$, 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. For convenience, we write $S_A^{\mathcal{X}} = \{x : (x, a) \in S_A\}$, $S_P^{\mathcal{X}} = \{x : (x, y) \in S_P\}$ 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

$$D_d(\mathcal{P}, \mathcal{P}') = \inf_{\pi \in \Pi(\mathcal{P}, \mathcal{P}')} \mathop{\mathbb{E}}_{(z,z') \sim \pi} [d(z, z')].$$

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 $B_d(\mathcal{P}, r) = \{\mathcal{Q} \in \mathbb{P} : D_d(\mathcal{P}, \mathcal{Q}) \leq r\}$ 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 are

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

where $||v||_p = (\sum_d |v[d]|^p)^{\frac{1}{p}}$ is some p-norm. We'll write $||v||_{p,*} = \sup_{||v'||_p \leq 1} \langle v, v' \rangle$ to denote its dual 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 over (x, a, y) whose marginal distribution over (x, a) is at most r_A away from $\tilde{\mathcal{P}}_{S_A}$ in Wasserstein distance and similarly whose marginal distribution over (x, y) is at most r_P away from $\tilde{\mathcal{P}}_{S_P}$ in Wasserstein distance.

More formally, the set of distributions we are interested in is

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

Now, we consider some learning task where the performance is measured according to the worst case distribution in the above set of distributions:

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

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

$$\begin{aligned} f^*(b) &= \sup_x \langle x^*, x \rangle - f(x) \\ &= \begin{cases} b \log b + (1 - b) \log(1 - b) & \text{if } b \in (0, 1) \\ 0 & \text{if } b = 0 \text{ or } 1 \\ \infty & \text{otherwise} \end{cases} \end{aligned}$$

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

3 Tractable Optimization

Note that the optimization problem in (1) is a saddle point problem. In Section 3.1, we make the coupling in the optimal transport more explicit in the inner sup term. Then, as in Shafieezadeh-Abadeh et al. [2015], 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 each involve some sup term, meaning it's an semi-infinite program (i.e. $\sup_{z\in Z}$ constraint(z) ≤ 0 is equivalent to constraint(z) $\leq 0, \forall z \in Z$). However, in Section 3.3, we show how each sup term can be approximated and be replaced by a single constraint.

3.1 Formulation through Coupling

We show how to rewrite the problem (1) by surfacing the underlying coupling π between the "anchor" distributions (S_A, S_P) and our target distribution $\mathcal{Q} \in W(S_A, S_P, r_A, r_P)$. Because $\pi \in \Pi(\tilde{\mathcal{P}}_{S_A}, \tilde{\mathcal{P}}_{S_P}, \mathcal{Q})$ is a coupling between $\tilde{\mathcal{P}}_{S_A}$, $\tilde{\mathcal{P}}_{S_P}$, and some distribution $\mathcal{Q}$, we must have the following for π:

1. Marginalizing π over $i \in [n_A]$ must yield a coupling $\pi_{S_P,(\mathcal{X},\mathcal{A},\mathcal{Y})}$ between $\tilde{\mathcal{P}}_{S_P}$ and Q:

$$\pi_{S_P,(\mathcal{X},\mathcal{A},\mathcal{Y})}((x_j^P, y_j^P), (x, a, y)) = \sum_{i=1}^{n_A} \pi\left((x_i^A, a_i^A), (x_j^P, y_j^P), (x, a, y)\right)$$

2. Marginalizing over $j \in [n_P]$ must yield a coupling $\pi_{S_A,(\mathcal{X},\mathcal{A},\mathcal{Y})}$ between $\tilde{\mathcal{P}}_{S_A}$ and $\mathcal{Q}$:

$$\pi_{S_A,(\mathcal{X},\mathcal{A},\mathcal{Y})}((x_i^A, a_i^A), (x, a, y)) = \sum_{j=1}^{n_P} \pi\left((x_i^A, a_i^A), (x_j^P, y_j^P), (x, a, y)\right)$$

3. π's marginal distribution over $(\mathcal{X}, \mathcal{A}, \mathcal{Y})$, S_A and S_P is exactly $\mathcal{Q}$, $\tilde{\mathcal{P}}_{S_A}$, $\tilde{\mathcal{P}}_{S_P}$ respectively:

$$\pi_{S_A}(x_i^A, a_i^A) = \sum_{j=1}^{n_P} \int \pi\left((x_i^A, a_i^A), (x_j^P, y_j^P), (dx, da, dy)\right) = \frac{1}{n_A}$$

$$\pi_{S_P}(x_j^P, y_j^P) = \sum_{i=1}^{n_A} \int \pi\left((x_i^A, a_i^A), (x_j^P, y_j^P), (dx, da, dy)\right) = \frac{1}{n_P}$$

$$\mathcal{Q} = \pi_{(\mathcal{X},\mathcal{A},\mathcal{Y})}(x, a, y) = \sum_{i=1}^{n_A}\sum_{j=1}^{n_P} \pi\left((x_i^A, a_i^A), (x_j^P, y_j^P), (x, a, y)\right)$$

Using the above notations, we can re-write the constraint in $W(S_A, S_P, r_A, r_P)$ where π's marginal distribution over $(\mathcal{X}, \mathcal{A})$ must be at most r_A away from $\tilde{\mathcal{P}}_{S_A}$ in Wasserstein distance as follows:

$$\begin{aligned}
&\mathop{\mathbb{E}}_{(x_i^A,a_i^A),(x,a,y))\sim\pi_{S_A,(\mathcal{X},\mathcal{A},\mathcal{Y})}} \left[d_A((x_i^A, a_i^A), (x, a))\right] \\
&= \sum_{i=1}^{n_A} \int d_A((x_i^A, a_i^A), (x, a))\pi_{S_A,(\mathcal{X},\mathcal{A},\mathcal{Y})}(x_i^A, a_i^A, (dx, da, dy)) \\
&= \sum_{i=1}^{n_A}\sum_{j=1}^{n_P} \int d_A((x_i^A, a_i^A), (x, a))\pi\left((x_i^A, a_i^A), (x_j^P, y_j^P), (dx, da, dy)\right) \leq r_A.
\end{aligned}$$

Similarly, we can write the other constraint that π's marginal distribution over $(\mathcal{X}, \mathcal{Y})$ must be at most r_P away from $\tilde{\mathcal{P}}_{S_P}$ as

$$\begin{aligned}&\mathop{\mathbb{E}}_{(x_j^P, y_j^P), (x,a,y)) \sim \pi_{S_P, (\mathcal{X}, \mathcal{A}, \mathcal{Y})}} \left[d_P((x_j^P, y_j^P), (x, y)) \right] \\ &= \sum_{i=1}^{n_A} \sum_{j=1}^{n_P} \int d_P((x_j^P, a_j^P), (x, a)) \pi \left((x_i^A, a_i^A), (x_j^P, y_j^P), (dx, da, dy) \right) \leq r_P.\end{aligned}$$

Lastly, the constraint that in order π to be a valid coupling, its marginal distribution over S_A and S_P should be exactly $\frac{1}{n_A}$ and $\frac{1}{n_P}$ over its support is equivalent to

$$\begin{aligned}\sum_{j=1}^{n_P} \sum_{a \in A} \sum_{y \in \mathcal{Y}} \int \pi \left((x_i^A, a_i^A), (x_j^P, y_j^P), (dx, da, dy) \right) &= \frac{1}{n_A} \quad \forall i \in [n_A] \\ \sum_{i=1}^{n_A} \sum_{a \in A} \sum_{y \in \mathcal{Y}} \int \pi \left((x_i^A, a_i^A), (x_j^P, y_j^P), (dx, da, dy) \right) &= \frac{1}{n_P} \quad \forall j \in [n_P].\end{aligned}$$

For simplicity, instead of $\pi \left((x_i^A, a_i^A), (x_j^P, y_j^P), (x, a, y) \right)$, 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, combining all these together, we can rewrite the problem (1) as choosing $\theta \in \Theta$ that minimizes the following value:

$$\begin{aligned}\sup_{\pi_{i,j}^{a,y}} \quad & \sum_{i=1}^{n_A} \sum_{j=1}^{n_P} \sum_{y \in \mathcal{Y}} \int_{\mathcal{X}, \mathcal{A}} \ell(\theta, (x, a, y)) \pi_{i,j}^y(dx, da) \\ \text{s.t.} \quad & \sum_{i=1}^{n_A} \sum_{j=1}^{n_P} \sum_{y \in \mathcal{Y}} \int_{\mathcal{X}, \mathcal{A}} d_A^i(x, a) \pi_{i,j}^y(dx, da) \leq r_A \\ & \sum_{i=1}^{n_A} \sum_{j=1}^{n_P} \sum_{y \in \mathcal{Y}} \int_{\mathcal{X}, \mathcal{A}} d_P^j(x, y) \pi_{i,j}^y(dx, da) \leq r_P \\ & \sum_{j=1}^{n_P} \sum_{y \in \mathcal{Y}} \int_{\mathcal{X}, \mathcal{A}} \pi_{i,j}^y(dx, da) = \frac{1}{n_A} \quad \forall i \in [n_A] \\ & \sum_{i=1}^{n_A} \sum_{y \in \mathcal{Y}} \int_{\mathcal{X}, \mathcal{A}} \pi_{i,j}^y(dx, da) = \frac{1}{n_P} \quad \forall j \in [n_P]\end{aligned} \tag{2}$$

where $d_A^i(x, a) = d_A((x_i^A, a_i^A), (x, a))$ and $d_P^j(x, y) = d_P((x_j^P, y_j^P), (x, y))$. 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 3.1 and 3.2 can be found in Appendix A.1.

Theorem 3.1. *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 3.2. *$D_{d_\mathcal{X}}(\tilde{\mathcal{P}}_{S_A^\mathcal{X}}, \tilde{\mathcal{P}}_{S_P^\mathcal{X}}) \leq r_A + r_P$, 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:

$$\begin{aligned} \inf_{\substack{\alpha_A,\alpha_P,\\ \{\beta_i\}_{i\in[n_A]},\\ \{\beta'_j\}_{j\in[n_P]}}} \quad & \alpha_A r_A + \alpha_P r_P + \frac{1}{n_A}\sum_{i\in[n_a]}\beta_i + \frac{1}{n_P}\sum_{j\in[n_P]}\beta' \\ \text{s.t.} \quad & \sup_{(x,a)}\Big(\ell(\theta,(x,a,y)) - \alpha_A d_A^i(x,a) - \alpha_P d_P^j(x,y)\Big) \le \beta_i + \beta'_j \forall i\in[n_A], j\in[n_P], y\in\mathcal{Y} \end{aligned} \tag{3}$$

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 Shafieezadeh-Abadeh et al. [2015] and Esfahani and Kuhn [2018], strong duality directly follows from proposition 3.4 of Shapiro [2001], but to be self-contained, we include the proof in Appendix A.3 which follows the same proof structure presented in Villani [2003]. For clarity, we assume in the proof that $\mathcal{X}$ and $\mathcal{A}$ is compact, but for more interested readers, we refer to the strong duality proof in Theorem 1.3 of Villani [2003] to see how to remove the compactness assumption on $\mathcal{X}$ and $\mathcal{A}$.

Theorem 3.3. *Assume $\mathcal{X}$ and $\mathcal{A}$ are compact spaces. 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 into a minimization problem:

$$\begin{aligned} \min_{\substack{\theta\in\Theta,\\ \alpha_A,\alpha_P,\\ \{\beta_i\}_{i\in[n_A]},\\ \{\beta'_j\}_{j\in[n_P]}}} \quad & \alpha_A r_A + \alpha_P r_P + \frac{1}{n_A}\sum_{i\in[n_a]}\beta_i + \frac{1}{n_P}\sum_{j\in[n_P]}\beta' \\ \text{s.t.} \quad & \sup_{(x,a)}\Big(\ell(\theta,(x,a,y)) - \alpha_A d_A^i(x,a) - \alpha_P d_P^j(x,y)\Big) \le \beta_i + \beta'_j \forall i\in[n_A], j\in[n_P], y\in\mathcal{Y} \end{aligned}$$

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 Shafieezadeh-Abadeh et al. [2015] who study when there's only one "anchor" point — i.e. $\sup_x \ell(\theta,x) - \alpha d_{\mathcal{X}}(x_i,x)$.

With some rearranging, let's focus only on the terms that actually depend on (x,a).

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

During this discussion, we drop y by using h[2] and also for simplicity, we write R to denote

$$R = \sup_{(x,a)} h(\theta,(x,a)) - \alpha_A||x_i^A - x||_p - \alpha_P||x_j^P - x||_p - \alpha_A\kappa_A||a_i^A - a||_{p'}.$$

We now rearrange some terms of R and use convex conjugate of h to represent the supremum term with what is known as an infimal convolution:

[2]Note that all our arguments are based on some fixed θ, so if $y = +1$, proceed with the original θ, and for $y = -1$, proceed with a new fixed $\theta' = -\theta$.

Lemma 3.1. *Fix any* θ, (x_i^A, a_i^A, x_j^P), *and* $(\alpha_A, \alpha_P, \kappa_A)$. *If* $||\theta[1:m_1]||_{p,*} > \alpha_A + \alpha_P$ *or* $||\theta[m_1+1 : m_1+m_2]||_{p',*} > \kappa_A\alpha_A$, *then* $\sup_{(x,a)} h(\theta,(x,a)) - \alpha_A||x_i^A - x||_p - \alpha_P||x_j^P - x||_p - \alpha_A\kappa_A||a_i^A - a||_{p'} = \infty$. *Otherwise, we have*

$$\begin{aligned}&\sup_{(x,a)} h(\theta,(x,a)) - \alpha_A||x_i^A - x||_p - \alpha_P||x_j^P - x||_p - \alpha_A\kappa_A||a_i^A - a||_{p'} \\ &= \sup_{b\in[0,1]} -f^*(b) + (g_1^i \square g_2^j)(b\theta[1:m_1]) + \langle b\theta[m_1+1:m_1+m_2], a_i^A\rangle\end{aligned}$$

where

$$g_1^i(\theta) = \begin{cases}\langle\theta, x_i^A\rangle & \text{if } ||\theta||_{p,*} \le \alpha_A \\ \infty & \text{otherwise}\end{cases} \qquad g_2^j(\theta) = \begin{cases}\langle\theta, x_j^P\rangle & \text{if } ||\theta||_{p,*} \le \alpha_P \\ \infty & \text{otherwise}\end{cases}$$

and $(g_1^i \square g_2^j)(\theta) = \inf_{\theta_1+\theta_2=\theta} g_1^i(\theta_1) + g_2^j(\theta_2)$ *is the infimal convolution of* g_1^i *and* g_2^j.

Then, by noting that an infimal convolution of two linear functions over bounded norm domain is convex, we show how to upper bound the infimal convolution with a linear term:

Theorem 3.4. *Suppose the norm* $||\cdot||$ *is some p-norm where* $p \neq 1$ *and* $p \neq \infty$. *Fix* θ *where* $||\theta||_* \le \alpha_A + \alpha_P$. *Then, for any* $b \in [0,1]$,

$$(g_1^i \square g_2^j)(b\theta) \le \left(\frac{b}{\alpha_A+\alpha_P}\right)\left(||\theta||_* \min(\alpha_A,\alpha_P)||x_i^A - x_j^P|| + \langle\theta, \alpha_A x_i^A + \alpha_P x_j^P\rangle\right) - \min(\alpha_A,\alpha_P)||x_i^A - x_j^P||$$

Combining Lemma 3.1 and Theorem 3.4, we can show the following upper bound on R:

Theorem 3.5. *We 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*

$$R \le f\left(\left(\frac{\min(\alpha_A,\alpha_P)||\theta_1||_*||x_i^A - x_j^P||}{\alpha_A+\alpha_P} + \frac{\langle\theta_1, \alpha_A x_i^A + \alpha_P x_j^P\rangle}{\alpha_A+\alpha_P}\right) + \langle\theta_2, a_i^A\rangle\right) - \min(\alpha_A,\alpha_P)||x_i^A - x_j^P||_p.$$

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

Suppose we write $\hat{x}_{i,j} = \begin{cases} x_j^P & \text{if } \alpha_A < \alpha_P \\ x_i^A & \end{cases}$ and $\hat{\alpha} = \min(\alpha_A, \alpha_P)$.

Then, via Hölder's inequality, we can show that evaluating the constraint at $(\hat{x}_{i,j}, a_i^A)$ is pretty close to to the upperbound of R in Theorem 3.5 and hence, it is also close to R because the constraint evaluated at $(\hat{x}_i^A, a_i^A)$ is a lower bound for R.

Theorem 3.6. *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*

$$\left(h(\theta,(\hat{x}_{i,j}, a_i^A)) - \hat{\alpha}||x_i^A - x_j^P||_p\right) - R \le 2\hat{\alpha}||x_i^A - x_j^P||.$$

Therefore, we can approximate R with $\left(h(\theta,\hat{x}_{i,j},a_i^A)) - \hat{\alpha}||x_i^A - x_j^P||_p\right)$. In the next section, we try to justify why the approximation error $2\hat{\alpha}||x_i^A - x_j^P||$ is reasonable.

4 Optimization

4.1 Approximation

We will first try to reformulate the original problem by making some structural assumption about the optimal transport $\pi_{i,j}^y(x,a)$. Because it is an optimal transport, we most likely have that for every (x,a,y) whose

measure is non-zero (i.e. $\pi^y_{i,j}(x,a) > 0$), its distance to (x^A_i, a^A_i) and (x^P_j, y^P_j) should be small. In other words, we most likely have that for any (i,j) where $||x^A_i - x^P_j||$ is big, $\pi^y_{i,j}(x,a)$ will be zero. Therefore, we assume that for every $i \in [n_A]$, we only consider its k-closest neighbors out of $\{x^P_j\}_{j\in[n_P]}$ and do the same for $j \in [n_P]$. We will denote this set of pairs by M

$$M = \left\{(i,j) : \begin{array}{l} x^A_i \text{ is one of } x^P_j\text{'s } k\text{-nearest neighbors among } \{x^A_{i'}\}_{i'} \\ \text{or } x^P_j \text{ is one of } x^A_i\text{'s k-nearest neighbors among } \{x^P_{j'}\}_{j'}. \end{array}\right\}$$

Noting that the dual constraint for each $i \in [n_A], j \in [n_P]$, and $y \in \mathcal{Y}$ corresponds to the primal variable $\pi^y_{i,j}$, this assumption allows us to only consider constraints $(i,j) \in M$. Then, after multiplying the objective by $n_A n_P$ with some rearranging, the dual problem becomes

$$\begin{aligned} \min_{\substack{\theta,\alpha_A,\alpha_P,\\ \{\beta_i,\beta'_j\}_{(i,j)\in M}}} \quad & n_A n_P(\alpha_A r_A + \alpha_P r_P) + \sum_{(i,j)\in M}(\beta_i + \beta'_j) \\ \text{s.t.} \quad & \max_{y\in\mathcal{Y}} \sup_{(x,a)} \Big(\ell(\theta,(x,a,y)) - \alpha_A d^i_A(x,a) - \alpha_P d^j_P(x,y)\Big) \le \beta_i + \beta'_j \quad \forall (i,j) \in M. \end{aligned} \tag{4}$$

In the case where the k-nearest-neighbor graph M between S_A and S_P is nicely structured[3], we should be always able to find $\{\beta_i, \beta_j\}$ such that for each $(i,j) \in M$

$$\max_{y\in\mathcal{Y}} \sup_{(x,a)} \Big(\ell(\theta,(x,a,y)) - \alpha_A d^i_A(x,a) - \alpha_P d^j_P(x,y)\Big) = \beta_i + \beta'_j. \tag{5}$$

Note that if there exists $\{\beta_i, \beta_j\}$ that satisfy (5), the optimal solution to (4) must satisfy (5). Therefore, assuming such $\{\beta_i, \beta_j\}$ exists, we get to re-write the optimization problem as

$$\min_{\alpha_A,\alpha_P,\theta} n_A n_P(\alpha_A r_A + \alpha_P r_P) + \sum_{(i,j)\in M} \max_{y\in\mathcal{Y}} \sup_{(x,a)} \big(\ell(\theta,(x,a,y)) - \alpha_A d^i_A(x,a) - \alpha_P d^j_P(x,y)\big).$$

Using the following fact about logistic function $f(-t) = f(t) + t$, we know that

$$\max(f(t), f(-t)) = f(t) + \max(t, 0).$$

In other words, $\max\big(f(y^P_j\langle\theta,(\hat{x}_{i,j},a^A_i)\rangle), f(-y^P_j\langle\theta,(\hat{x}_{i,j},a^A_i)\rangle) - \alpha_P\kappa_P\big) = f(y^P_j\langle\theta,(\hat{x}_{i,j},a^A_i)\rangle) + \max(y^P_j\langle\theta,(\hat{x}_{i,j},a^A_i)\rangle - \alpha_P\kappa_P, 0)$. Using our approximation of the supremum term as in Theorem 3.6 and the above fact, the problem then becomes

$$\begin{aligned} \min_{\alpha_A,\alpha_P,\theta_1,\theta_2} \quad & (\alpha_A r_A + \alpha_P r_P) + \frac{1}{n_A n_P}\sum_{i=1}^{n_A}\sum_{j=1}^{n_P}(f(y^P_j\langle\theta,(\hat{x}_{i,j},a^A_i)\rangle) \\ & + \max(y^P_j\langle\theta,(\hat{x}_{i,j},a^A_i)\rangle - \alpha_P\kappa_P, 0) - \hat{\alpha}||x^A_i - x^P_j||) \\ \text{s.t.} \quad & ||\theta_1||_* \le \alpha_A + \alpha_P, ||\theta_2||_* \le \kappa_A\alpha_A. \end{aligned} \tag{6}$$

Note that because we have restricted our attention only to pairs who are close to one another, the additive approximation error due to using evaluating the constraint only at $(\hat{x}_{i,j}, a^A_i)$ which amounts to $\frac{2\hat{\alpha}}{n_A n_P}\sum_{(i,j)\in M}||x^A_i - x^P_j||$ in the objective must be small.

[3]More formally, this is equivalent to assuming that there exists a feasible solution to the following system following linear equations. Suppose A is a $|M| \times (n_A + n_P)$ matrix where for every lth pair (i,j) in M, $M[l,i] = 1$ and $M[l, n_A + j] = 1$. And b is a vector of length $|M|$ where for every lth pair $(i,j) \in M$,

$$b[l] = \max_{y\in\mathcal{Y}} \sup_{(x,a)} \Big(\ell(\theta,(x,a,y)) - \alpha_A d^i_A(x,a) - \alpha_P d^j_P(x,y)\Big).$$

Our assumption is equivalent to assuming that there exists a vector x of length $n_A + n_P$ such that $Ax = b$ or equivalently, A is left-invertible. Note that the very first n_A coordinates correspond to $\{\beta_i\}$ and the last n_P coordinates correspond to $\{\beta'_j\}$.

4.2 Projected Gradient Descent

To solve the optimization problem (6), we employ first-order projected gradient descent. In order to handle $\hat{\alpha} = \min(\alpha_A, \alpha_P)$, we can just solve the optimization twice: once with $\alpha_A < \alpha_P$ as one of the constraints and the other time with $\alpha_A \geq \alpha_P$. Suppose $\hat{\alpha} = \alpha_A$, meaning $\hat{x}_{i,j} = x_j^P$. Then we write the objective function as

$$\Omega^A(\alpha_A, \alpha_P, \theta) = (\alpha_A r_A + \alpha_P r_P) + \frac{1}{n_A n_P} \sum_{(i,j)\in M} (f(y_j^P \langle \theta, (x_j^P, a_i^A)\rangle) + \max(y_j^P \langle \theta, (x_j^P, a_i^A)\rangle - \alpha_P \kappa_P, 0) - \alpha_A ||x_i^A - x_j^P||)$$

and the constraint set is

$$C^A = \{(\alpha_A, \alpha_P, \theta) : ||\theta_1||_* \leq \alpha_A + \alpha_P, ||\theta_2|| \leq \kappa_A \alpha_A, \alpha_A < \alpha_P\}.$$

Similarly, when $\hat{\alpha} = \alpha_P$, we write $\Omega^P(\alpha_A, \alpha_P, \theta)$ and C^P where the α constraint is replaced by $\alpha_A \geq \alpha_P$. Note that in both cases, we have a convex optimization problem.

Claim 4.1. *The objective functions $\Omega^A(\alpha_A, \alpha_P, \theta)$ and $\Omega^P(\alpha_A, \alpha_P, \theta)$ are convex in $(\alpha_A, \alpha_P, \theta)$. The constraint sets C^A and C^P are also convex in $(\alpha_A, \alpha_P, \theta)$.*

Suppose we write

$$(\alpha'_A, \alpha'_P, \theta') = \arg \min_{(\alpha_A, \alpha_P, \theta)\in C^A} \Omega^A(\alpha_A, \alpha_P, \theta)$$
$$(\alpha''_A, \alpha''_P, \theta'') = \arg \min_{(\alpha_A, \alpha_P, \theta)\in C^P} \Omega^P(\alpha_A, \alpha_P, \theta).$$

Claim 4.2. *The optimal solution to problem (6) is $(\alpha'_A, \alpha'_P, \theta')$ if $\Omega^A(\alpha'_A, \alpha'_P, \theta') \leq \Omega^P(\alpha''_A, \alpha''_P, \theta'')$ and $(\alpha''_A, \alpha''_P, \theta'')$ otherwise.*

Typical regularized models either constrain the norm of the parameter θ to be directly bounded by some constants specified initially or include the norm as part of the objective multiplied by some multiplicative penalty constant. However, our optimization problem is a hybrid of both as (1) the norms of the parameter θ are to be bounded by α_A and α_P but (2) (α_A, α_P) are part of the optimization variables that are multiplied by some penalty constants r_A and r_P in the objective function.

Nevertheless, the constraint set is convex so Euclidean projection can be solved via any convex solver, and in the case of $p = 2$, we have exactly characterized a closed form solution of the output of the projection in Appendix B.1. Therefore, in order to solve (6), we can use projected gradient descent (PGD)

$$(\alpha_A^{t+1}, \alpha_P^{t+1}, \theta^{t+1}) = \text{Project}_C\left((\alpha_A^t, \alpha_P^t, \theta^t) - \eta \nabla \Omega(\alpha_A^t, \alpha_P^t, \theta^t)\right).$$

It is well known that the rate of convergence for PGD is $O(\frac{1}{\sqrt{T}})$ with appropriately chosen step size η. We

present the overall algorithm to solve problem (6) in Algorithm 1.

Algorithm 1: Distributionally Robust Data Join

Input: S_A, S_P, r_A, r_P, κ_A, κ_P, k, T
Run k-nearest neighbors on S_A and S_P to calculate the matching pairs M
choose arbitrary $\theta, \alpha_A, \alpha_P$
Set $\theta_A^1 = \theta, \alpha_A^1 = \alpha_A, \alpha_P^1 = \alpha_P$
Set ${\theta'}_A^1 = \theta, {\alpha'}_A^1 = \alpha_A, \alpha_P^1$
for $i = 1$ **to** $T - 1$ **do**
  $(\alpha_A^{t+1}, \alpha_P^{t+1}, \theta^{t+1}) = \text{Project}_{C^A}\left((\alpha_A^t, \alpha_P^t, \theta^t) - \eta \nabla \Omega^A(\alpha_A^t, \alpha_P^t, \theta^t)\right)$
  $({\alpha'}_A^{t+1}, {\alpha'}_P^{t+1}, {\theta'}^{t+1}) = \text{Project}_{C^P}\left(({\alpha'}_A^t, {\alpha'}_P^t, {\theta'}^t) - \eta \nabla \Omega^P({\alpha'}_A^t, {\alpha'}_P^t, {\theta'}^t)\right)$
$\overline{\alpha_A} = \frac{1}{T}\sum_{t=1}^T \alpha_A^t, \overline{\alpha_P} = \frac{1}{T}\sum_{t=1}^T \alpha_P^t, \overline{\theta} = \frac{1}{T}\sum_{t=1}^T \theta^t$
$\overline{\alpha_A}' = \frac{1}{T}\sum_{t=1}^T {\alpha'}_A^t, \overline{\alpha_P}' = \frac{1}{T}\sum_{t=1}^T {\alpha'}_P^t, \overline{\theta}' = \frac{1}{T}\sum_{t=1}^T {\theta'}^t$
if $\Omega^A(\overline{\alpha_A}, \overline{\alpha_P}, \overline{\theta}) < \Omega^P(\overline{\alpha_A}', \overline{\alpha_P}', \overline{\theta}')$ **then**
  Return $(\overline{\alpha_A}, \overline{\alpha_P}, \overline{\theta})$
else
  Return $(\overline{\alpha_A}', \overline{\alpha_P}', \overline{\theta}')$

Write $\Omega(\alpha_A, \alpha_P, \theta) = \Omega^A(\alpha_A, \alpha_P, \theta)$ if $\alpha_A < \alpha_P$ and $\Omega^P(\alpha_A, \alpha_P, \theta)$ otherwise to denote the objective solution to problem (6). Then, the optimal value $(\alpha_A^*, \alpha_P^*, \theta^*)$ of problem (6) is $(\alpha_A^*, \alpha_P^*, \theta^*) = \arg\min_{(\alpha_A, \alpha_P, \theta) \in C^A \cup C^P} \Omega(\alpha_A, \alpha_P, \theta$

Theorem 4.1. *With appropriately chosen step size η, Algorithm 1 returns $(\alpha_A, \alpha_P, \theta)$ such that $\Omega(\alpha_A, \alpha_P, \theta) \leq \Omega(\alpha_A^*, \alpha_P^*, \theta^*) + O\left(\frac{1}{\sqrt{T}}\right)$.*

5 Application: Fairness

In many situations, the actual demographic group information may not be available in the original labeled dataset, but another auxiliary unlabeled dataset may contain the needed demographic group information. We can leverage our data join method in order to incorporate this auxiliary dataset to penalize the model for model's unfairness. Suppose $\mathcal{A}$ represents two different groups that an individual can belong to — $\mathcal{A} = \{0, 1\}$.

Given θ, we define its unfairness with respect to distribution $\mathcal{P}$ over $\mathcal{X}, \mathcal{A}, \mathcal{Y}$ as

$$\mathcal{U}(\theta, \mathcal{P}) = \left| \Pr_{(x,a,y)\sim\mathcal{P}}[u(h_\theta(x)) | a = 1, y = 1] - \Pr_{(x,a,y)\sim\mathcal{P}}[u(h_\theta(x)) | a = 0, y = 1] \right|$$

where $u(t) = \log(t)$ and $h_\theta(x) = \frac{1}{1+\exp(-\langle\theta, x\rangle)}$ as in Taskesen et al. [2020]. This term is similar to the difference in true positive rates as in the case of equal opportunity, but it differs in that it looks at the log-probability — this fairness criterion is referred to as log-probabilistic equalized opportunity in Taskesen et al. [2020].

Also, as in Taskesen et al. [2020], we suppose that we know the underlying positive rates for each group and constrain the joint distribution's marginal distribution over $\mathcal{A}$ and $\mathcal{Y}$ in the following manner: given some $p_0, p_1 \in (0, 1)$, we define

$$W_{(p_0,p_1)}(S_A, S_P, r_A, r_P) = \left\{ \mathcal{Q} \in W(S_A, S_P, r_A, r_P) : \Pr_{(x,a,y)\sim\mathcal{Q}}[a = 0, y = 1] = p_0, \Pr_{(x,a,y)\sim\mathcal{Q}}[a = 0, y = 1] = p_1 \right\}.$$

Then, the problem we are interested in is

$$\min_{\theta\in\Theta} \sup_{\mathcal{Q}\in W_{(p_0,p_1)}(S_A,S_P,r_A,r_P)} \mathbb{E}_{(x,a,y)\sim Q}[\ell(\theta, (x, a, y))] + \eta\mathcal{U}(\theta, \mathcal{Q}) \tag{7}$$

where we are adding a fairness regularization term multiplied by some constant η where $|\eta| < \min(p_0, p_1)$.

Following the same argument as in Section 3.1, we can re-write the problem as

$$\begin{aligned}
&\sup_{\pi_{i,j}^{a,y}} \sum_{i=1}^{n_A}\sum_{j=1}^{n_P}\sum_{a\in A}\sum_{y\in\mathcal{Y}}\int \left(\ell(\theta,(x,a,y)) + \eta\left(u(h_\theta((x,a)))\frac{\mathbb{1}[a=1,y=1]}{p_1} - u(h_\theta((x,a)))\frac{\mathbb{1}[a=0,y=1]}{p_0}\right)\right)\pi_{i,j}^{a,y}(dx)\\
\text{s.t.}\quad &\sum_{i=1}^{n_A}\sum_{j=1}^{n_P}\sum_{a\in A}\sum_{y\in\mathcal{Y}}\int d_A^i(x,a)\pi_{i,j}^{a,y}(dx) \le r_a\\
&\sum_{i=1}^{n_A}\sum_{j=1}^{n_P}\sum_{a\in A}\sum_{y\in\mathcal{Y}}\int d_P^j(x,y)\pi_{i,j}^{a,y}(dx) \le r_P\\
&\sum_{j=1}^{n_P}\sum_{a\in A}\sum_{y\in\mathcal{Y}}\int \pi_{i,j}^{a,y}(dx) = \frac{1}{n_A} \quad \forall i\in[n_A]\\
&\sum_{i=1}^{n_A}\sum_{a\in A}\sum_{y\in\mathcal{Y}}\int \pi_{i,j}^{a,y}(dx) = \frac{1}{n_P} \quad \forall j\in[n_P]\\
&\sum_{i=1}^{n_A}\sum_{a\in A}\sum_{y\in\mathcal{Y}}\int \mathbb{1}[a=0,y=1]\pi_{i,j}^{a,y}(dx) = p_0\\
&\sum_{i=1}^{n_A}\sum_{a\in A}\sum_{y\in\mathcal{Y}}\int \mathbb{1}[a=1,y=1]\pi_{i,j}^{a,y}(dx) = p_1
\end{aligned} \tag{8}$$

Denoting the value of the above optimization as $p^{\mathsf{fair}}(p_0,p_1,\eta)$, the same argument as in Theorem 3.1 can be used to see that the value of (7) is exactly $\max(p^{\mathsf{fair}}(p_0,p_1,\eta), p^{\mathsf{fair}}(p_0,p_1,-\eta))$ where we need to try out η and $-\eta$ in order to handle the absolute value in $\mathcal{U}$.

As in Section 3.2, the dual problem of (8) can be derived by looking at the Lagrangian, which after rearranging the terms a little bit is as follows:

$$\begin{aligned}
&\mathcal{L}(\pi,\alpha_A,\alpha_P,\{\beta_i\},\{\beta'_j\}\},\gamma_0,\gamma_1)\\
&= \sum_{i=1}^{n_A}\sum_{j=1}^{n_P}\sum_{a\in\mathcal{A}}\sum_{y\in\mathcal{Y}}\int \left(\ell(\theta,(x,a,y)) + \eta\left(u(h_\theta((x,a)))\frac{\mathbb{1}[a=1,y=1]}{p_1} - u(h_\theta((x,a)))\frac{\mathbb{1}[a=0,y=1]}{p_0}\right)\right.\\
&- \alpha_A d_A^i(x,a) - \alpha_P d_P^j(x,y) - \beta_i - \beta'_j\\
&\left. - \gamma_0\mathbb{1}[a=0,y=1] - \gamma_1\mathbb{1}[a=1,y=1]\right)\pi_{i,j}^{y,a}(dx)\\
&+ \alpha_A r_A + \alpha_P r_P + \frac{1}{n_A}\sum_{i=1}^{n_A}\beta_i + \frac{1}{n_P}\sum_{j=1}^{n_P}\beta'_j + p_0\gamma_0 + p_1\gamma_1.
\end{aligned}$$

The dual problem is then

$$\begin{aligned}\inf_{\substack{\alpha_A,\alpha_P,\\ \{\beta_i\}_{i\in[n_A]},\{\beta'_j\}_{j\in[n_P]}}} \quad & \alpha_A r_A + \alpha_P r_P + \frac{1}{n_A}\sum_{i\in[n_A]}\beta_i + \frac{1}{n_P}\sum_{j\in[n_P]}\beta'_j + p_0\gamma_0 + p_1\gamma_1 \\ \text{s.t.} \quad & \sup_x \left(\ell(\theta,(x,a,y)) + \eta u(h_\theta((x,a))) \left(\frac{\mathbb{1}[a=1,y=1]}{p_1} - \frac{\mathbb{1}[a=0,y=1]}{p_0}\right) \right. \\ & \left. - \alpha_A d_A^i(x,a) - \alpha_P d_P^j(x,y) \right) - \beta_i - \beta'_j \\ & - \gamma_0 \mathbb{1}[a=0,y=1] - \gamma_1 \mathbb{1}[a=1,y=1] \le 0 \quad i\in[n_A], j\in[n_P], a\in\mathcal{A}, y\in\mathcal{Y}\end{aligned} \tag{9}$$

Note that when $y=-1$, then the term in the supremum simply becomes

$$\ell(\theta,(x,a,y)) - \alpha_A d_A^i(x,a) - \alpha_P d_P^j(x,y).$$

When $y=1$, then we get

$$\begin{aligned}&\log(1+\exp(-\langle\theta,(x,a)\rangle)) - \eta\log(1+\exp(-\langle\theta,(x,a)\rangle)\left(\frac{\mathbb{1}[a=1]}{p_1} - \frac{\mathbb{1}[a=0]}{p_0}\right) - \alpha_A d_A^i(x,a) - \alpha_P d_P^j(x,y) \\ &= \log(1+\exp(-\langle\theta,(x,a)\rangle))\left(1-\eta\left(\frac{\mathbb{1}[a=1]}{p_1} - \frac{\mathbb{1}[a=0]}{p_0}\right)\right) - \alpha_A d_A^i(x,a) - \alpha_P d_P^j(x,y) \\ &= \left(1-\eta\left(\frac{\mathbb{1}[a=1]}{p_1} - \frac{\mathbb{1}[a=0]}{p_0}\right)\right)\ell(\theta,(x,a,y)) - \alpha_A d_A^i(x,a) - \alpha_P d_P^j(x,y)\end{aligned}$$

For simplicity, we write

$$c(a,y) = 1 - \mathbb{1}[y=1]\eta\left(\frac{\mathbb{1}[a=1]}{p_1} - \frac{\mathbb{1}[a=0]}{p_0}\right).$$

Note that $c(a,y) > 0$ because $\eta > \min(p_0,p_1)$. Write $\overline{c} = \max_{a,y} c(a,y)$.

Then, the above dual problem can be re-written as

$$\begin{aligned}\inf_{\substack{\alpha_A,\alpha_P,\\ \{\beta_i\}_{i\in[n_A]},\{\beta'_j\}_{j\in[n_P]}}} \quad & \alpha_A r_A + \alpha_P r_P + \frac{1}{n_A}\sum_{i\in[n_A]}\beta_i + \frac{1}{n_P}\sum_{j\in[n_P]}\beta' + p_0\gamma_0 + p_1\gamma_1 \\ \text{s.t.} \quad & \sup_x \left(c(a,y)\cdot\ell(\theta,(x,a,y)) - \alpha_A d_A^i(x,a) - \alpha_P d_P^j(x,y) \right) - \beta_i - \beta'_j \\ & - \gamma_0\mathbb{1}[a=0,y=1] - \gamma_1\mathbb{1}[a=1,y=1] \le 0 \quad i\in[n_A], j\in[n_P], a\in\mathcal{A}, y\in\mathcal{Y}.\end{aligned} \tag{10}$$

We remark that the $c(a,y)$ that folds the fairness constraint into the original loss is essentially equivalent to the cost plugged into the cost-sensitive oracle in Agarwal et al. [2018] and Kearns et al. [2018].

Note that the constant can be taken out of the sup as $c(a,y)$ is always positive and the same proof as in Lemma 3.1 can be used:

$$\sup_x \left(c(a,y)\cdot\ell(\theta,(x,a,y)) - \alpha_A d_A^i(x,a) - \alpha_P d_P^j(x,y)\right) = c(a,y)\cdot\sup_x\left(\ell(\theta,(x,a,y)) - \frac{\alpha_A}{c(a,y)}d_A^i(x,a) - \frac{\alpha_P}{c(a,y)}d_P^j(x,y)\right)$$

We have intentionally taken a outside the sup to not worry about $c(a,y)$ in the supremum. Just as Lemma 3.1, we write down the supremum using the convex conjugate f^*.

Lemma 5.1. *Fix any θ, (x_i^A, a, x_j^P), and $(\alpha_A, \alpha_P, \kappa_A)$. If $||\theta[1:m_1]||_{p,*} > \alpha_A + \alpha_P$, then $\sup_x h(\theta,(x,a)) - \alpha_A||x_i^A - x||_p - \alpha_P||x_j^P - x||_p = \infty$. Otherwise, we have*

$$\begin{aligned}&\sup_x h(\theta,(x,a)) - \alpha_A||x_i^A - x||_p - \alpha_P||x_j^P - x||_p \\ &= \sup_{b\in[0,1]} -f^*(b) + (g_1^i \square g_2^j)(-b\theta[1:m_1]) + \langle b\theta[m_1+1:m_1+m_2], a\rangle\end{aligned}$$

where g_1^i and g_2^j is the same as defined in Lemma 3.1.

As before, via the convexity of infimal convolution of two linear functions (Lemma A.2), we can upper bound the supremum. The only difference is that a_i^A has been replaced with a. For simplicity, in the following lemma and theorem, we use $\alpha_A := \frac{\alpha_A}{c(a,y)}$ and $\alpha_P := \frac{\alpha_P}{c(a,y)}$.

Theorem 5.1. *Fix any θ, (x_i^A, a, x_j^P), and $(\alpha_A, \alpha_P, \kappa_A)$. If $||\theta_1||_{p,*} > \alpha_A + \alpha_P$, then $\sup_x h(\theta,(x,a)) - \alpha_A||x_i^A - x||_p - \alpha_P||x_j^P - x||_p = \infty$ Otherwise, we have*

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

Now, note that depending on (a,y), $\alpha_A := \frac{\alpha_A}{c(a,y)}$ and $\alpha_P := \frac{\alpha_P}{c(a,y)}$ in the above lemma and theorem changes. Therefore, unless $||\theta||_{p,*} \le \min_{(a,y)} \frac{\alpha_A}{c(a,y)} + \frac{\alpha_P}{c(a,y)}$,

$$\max_{a,y} \sup_x \Big(c(a,y)\cdot\ell(\theta,(x,a,y)) - \alpha_A d_A^i(x,a) - \alpha_P d_P^j(x,y)\Big) = \infty.$$

In other words, we need

$$||\theta_1||_{p,*} \le \frac{\alpha_A+\alpha_P}{\overline{c}}$$

or the term evaluates to ∞ otherwise. Then, via our approximation with $\hat{x}_{i,j}$ as in Section 3, the optimization problem is

$$\begin{aligned}\min_{\substack{\alpha_A,\alpha_P,\\ \{\beta_i\}_{i\in[n_A]},\{\beta'_j\}_{j\in[n_P]}}} \quad & n_A n_P(\alpha_A r_A + \alpha_P r_P + p_0\gamma_0 + p_1\gamma_1) + n_P\sum_{i\in[n_A]}\beta_i + n_A\sum_{j\in[n_P]}\beta' \\ \text{s.t.} \quad & c(a,y)\cdot\ell(\theta,(\hat{x}_{i,j},a,y)) + \alpha_A\kappa_A|a_i^A - a| + \alpha_P\kappa_P|y_j^P - y| - \hat{\alpha}||x_i^A - x_j^P|| \\ & + \gamma_0\mathbb{1}[a=0,y=1] + \gamma_1\mathbb{1}[a=1,y=1] \le \beta_i + \beta'_j \quad i\in[n_A], j\in[n_P], a\in\mathcal{A}, y\in\{\pm1\}, \\ & ||\theta_1||_* \le \frac{\alpha_A+\alpha_P}{\overline{c}}\end{aligned}$$

Under the same assumption as in Section 4[4], the optimization problem can be re-written as

$$\begin{aligned}\min_{\substack{\alpha_A,\alpha_P,\\ \{\beta_i\}_{i\in[n_A]},\{\beta'_j\}_{j\in[n_P]}}} \quad & (\alpha_A r_A + \alpha_P r_P + p_0\gamma_0 + p_1\gamma_1) + \frac{1}{n_A n_P}\sum_{(i,j)\in M}\max_{a\in\mathcal{A},y\in\{\pm1\}} c(a,y)\cdot\ell(\theta,(\hat{x}_{i,j},a,y)) + \alpha_A\kappa_A|a_i^A - a| \\ & + \alpha_P\kappa_P|y_j^P - y| - \hat{\alpha}||x_i^A - x_j^P|| \\ \text{s.t.} \quad & ||\theta_1||_* \le \frac{\alpha_A+\alpha_P}{\overline{c}}.\end{aligned}$$

Note that this is still a convex optimization problem as taking max still preserves convexity of the functions inside. As before, we can get the same kind of approximation error. For each fixed (a,y), approximating the $\sup_x$ term with $\hat{x}_{i,j}$ will result in approximation error of $2c(a,y)||x_i^A - x_j^P||$ as in Theorem 3.6. Therefore, even when we take the max over all (a,y), the overall gap must be bounded by $2c(a,y)||x_i^A - x_j^P|| \le 4||x_i^A - x_j^P||$.

[4]That is, the k-nearest-neighbor matching matrix A as formed as in Section 4 is left-invertible

Theorem 5.2. *Suppose $p \neq 1$ and $p \neq \infty$. If $||\theta_1||_{p,*} \leq \alpha_A + \alpha_P$, then*

$$\max_{a,y} \sup_{x \in \mathcal{X}} \Big(c(a,y) \cdot \ell(\theta, (x,a,y)) - \alpha_A d_A^i(x,a) - \alpha_P d_P^j(x,y) \Big)$$
$$- \max_{a,y} \big(c(a,y) \cdot \ell(\theta, (\hat{x}_{i,j}, a, y)) + \alpha_A \kappa_A |a_i^A - a| + \alpha_P \kappa_P |y_j^P - y| - \min(\alpha_A, \alpha_P) ||x_i^A - x_j^P|| \big)$$
$$\leq 4\hat{\alpha} ||x_i^A - x_j^P||.$$

We remark that solving for $\sup_{\mathcal{Q} \in W_{(p_0,p_1)}(S_A, S_P, r_A, r_P)} \mathcal{U}(\theta, \mathcal{Q})$ for some fixed θ, which can be indeed solved with minimal modifications, corresponds to estimating the worst case unfairness of θ over all distributions $\mathcal{Q} \in W_{(p_0,p_1)}(S_A, S_P, r_A, r_P)$. Kallus et al. [2021] consider a special case where $r_A, r_P = 0, 0$, but they can handle various fairness measures.

6 Experiments

In all our experiments, we use 2-norm for our data join method: i.e. $p, p' = 2$. We note that as it's standard in practice to use the last iterate instead of the averaged iterate, we use the last iterate of the projected gradient descent steps instead of the averaged one for all our experiments — we use $(\alpha^T, \alpha^T, \theta^T)$ if $\Omega^A(\alpha^T, \alpha^T, \theta^T) < \Omega^P(\alpha'^T, \alpha'^T, \theta'^T)$ and $(\alpha'^T, \alpha'^T, \theta'^T)$ otherwise. The code used for the experiments can be found at `https://github.com/chrisjung/Distributionally-Robust-Data-Join`.

6.1 UCI datasets

Here we discuss some experiments we have run on UCI datasets. The UCI datasets that we used are the following:

1. Breast Cancer dataset (BC)[5]: 569 points with 30 features,
2. Ionosphere dataset (IO)[6]: 351 points with 34 features,
3. Heart Disease dataset (HD)[7]: 300 points with 13 features,
4. Handwritten Digits dataset (1vs8)[8]: 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. The task we considered was distinguishing between 1's and 8's.

For every dataset, we preprocess the data by standardizing each feature — that is, removing the mean and scaling to unit variance. After standardizing our dataset, 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 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 used $m_1 = 32$, 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 — $m_1 = 25$.

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

[5] `https://archive.ics.uci.edu/ml/datasets/breast+cancer`
[6] `https://archive.ics.uci.edu/ml/datasets/ionosphere`
[7] `https://archive.ics.uci.edu/ml/datasets/Heart+Disease`
[8] `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`

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.

	BC ($m_1 = 5$)	BC ($m_1 = 25$)	IO ($m_1 = 4$)	IO ($m_1 = 25$)	HD	1vs8
DJ	**0.9199** $\pm$ 0.0283	**0.9415** $\pm$ 0.0165	**0.8226** $\pm$ 0.0764	**0.7906** $\pm$ 0.0484	**0.7495** $\pm$ 0.0374	**0.9206** $\pm$ 0.0322
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

Table 1: Average accuracy of each method over 10 experiment runs and standard deviations for three UCI datasets

We include the parameters used for each of these baselines and our method (DJ) and how they were chosen in Appendix D. It can be seen that the use of the additional auxiliary features through our data join method seems to help achieve better accuracy than the baselines that we considered.

6.2 Synthetic Dataset

Through a simple experiment on synthetically generated data, we demonstrate how our approach (DJ) can handle distribution shifts well. Note that in the previous experiment with the UCI datasets, each points have been all drawn iid, so how well our method can handle distribution shift wasn't really tested in those experiments.

	LR	RLR	DRLR	DJ
Accuracy	0.4126 $\pm$ 0.1049	0.5786 $\pm$ 0.3992	0.9068 $\pm$ 0.0076	**0.9923** $\pm$ 0.0057

Table 2: Average accuracy of each method over 10 experiment runs and standard deviations for synthetic dataset with a distribution shift

We first describe the data generation process. At a high level, there are two groups whose covariate and label distributions are different. The majority of the points in the labeled dataset S_P is the first group, but in the unlabeled and test dataset (S_A, S_{test}), the majority is the second group. More specifically, 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, 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, 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 0.2:

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

Now, for the first dataset $S_1 = \{(x_j^1, y_j^1)\}_{j=1}^{n_1}$, we set the number of points from group 1 and from group 2 to be 400 and 20 respectively. And we had the number of positive and negative points in each group to be 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 $m_2 = 10 - m_1 = 8$.

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 $r = 100, \kappa = 10$.

Depending on which group is the majority in the dataset, the ideal hyperplane is different. If the majority is the first group, the ideal hyperplane is such that it ignores all the features except for the fist one and returns 1 if the first feature is positive and -1 otherwise. However, if the majority is the second group, the ideal hyperplane is such that it does the same process for all the features: predict +1 if all the features all mostly positive -1 otherwise. This is the reason why vanilla logistic regression puts most of its weights only on the first feature and not the second, but because the majority has been flipped in the test distribution, it performs very poorly. Regularized and distributionally logistic regression seems to mitigate against this effect. However, they still need to hedge against all the distributions that is nearby the empirical distribution over S_P, so the accuracy isn't as high.

By contrast, the reason why our data join approach method does well as compared to other methods is mainly due to its k-nearest-neighbor matching. The group identity is actually encoded in the second feature: the second feature of a point is 0 if it's from the first group, and if it's from the second group, it is -1 or 1 depending on the label. Therefore, k-nearest-neighbor should be able to match each of the points to another point that belongs to the same group and the correct label. And as a result of joining the second unlabeled dataset via this knn matching, which group is the majority included in the dataset must have been flipped. Furthermore, with the availability of the auxiliary features, namely feature 3 to 10, the data join can nearly predict the label of each point perfectly (i.e. 99.45% as shown in Table 2).

In other words, one can expect our distributionally robust data join method to perform well, when the information embedded in the common features $\mathcal{X}$ allows the k-nearest-neighbors to match the points very well. Nevertheless, we remark that the k-nearest-neighbor's matching doesn't have to be perfect as the regularization of the model parameters (i.e. $||\theta_1|| \leq \alpha_A + \alpha_P$ and $||\theta_2|| \leq \kappa_A \alpha_A$) and the label uncertainty (i.e. $\max(f(y\langle\theta, \hat{x}_{i,j}, a)\rangle), f(-y\langle\theta, \hat{x}_{i,j}, a)\rangle)) + \alpha_P \kappa_P)$) should be able to tolerate some amount of the mismatches that happen from k-nearest-neighbor.

References

Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna Wallach. A reductions approach to fair classification. In *International Conference on Machine Learning*, pages 60–69. PMLR, 2018.

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.

Dimitri P Bertsekas. *Convex optimization theory*.

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.

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

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, 2021a.

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

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.

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

Emily Diana, Wesley Gill, Michael Kearns, Krishnaram Kenthapadi, and Aaron Roth. Convergent algorithms for (relaxed) minimax fairness. *arXiv e-prints*, pages arXiv–2011, 2020.

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

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.

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

Emre Erdoğan and Garud Iyengar. Ambiguous chance constrained problems and robust optimization. *Mathematical Programming*, 107(1):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):115–166, 2018.

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

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

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.

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

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

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

Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In *International Conference on Machine Learning*, pages 2564–2572. PMLR, 2018.

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.

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

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.

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

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

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

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.

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

Thomas Strömberg. *A study of the operation of infimal convolution*. PhD thesis, Luleå tekniska universitet, 1994.

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

Cédric Villani. *Topics in optimal transportation*. Number 58. American Mathematical Soc., 2003.

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.

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

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

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

Xiaojin Jerry Zhu. Semi-supervised learning literature survey. 2005.

Appendices

A Missing Details from Section 3

A.1 Missing Details from Section 3.1

Theorem 3.1. *For any fixed $\theta \in \Theta$, $p^*(\theta, r_A, r_P) = \sup_{\mathcal{Q} \in W(S_A, S_P, r_A, r_P)} \mathbb{E}_{(x,a,y)\sim Q}[\ell(\theta, (x, a, y))]$*

Proof. It's clear that for any feasible solution π for (2), we must have that

$$\pi_{(\mathcal{X},\mathcal{A},\mathcal{Y})} \in W(S_A, S_P, r_A, r_P)$$

as we have a coupling $\pi_{S_A,(\mathcal{X},\mathcal{A},\mathcal{Y})}$ between $\tilde{\mathcal{P}}_{S_A}$ and $\pi_{(\mathcal{X},\mathcal{A},\mathcal{Y})}$ such that

$$\mathop{\mathbb{E}}_{(x_i^A, a_i^A),(x,a,y))\sim\pi_{S_A,(\mathcal{X},\mathcal{A},\mathcal{Y})}} \left[d_A((x_i^A, a_i^A), (x, a))\right] \leq r_A$$

and

$$\mathop{\mathbb{E}}_{(x_j^P, y_j^P),(x,a,y))\sim\pi_{S_P,(\mathcal{X},\mathcal{A},\mathcal{Y})}} \left[d_P((x_j^P, y_j^P), (x, y))\right] \leq r_P.$$

Also, for any $\mathcal{Q} \in W(S_A, S_P, r_A, r_P)$, let's write the optimal transport between $\tilde{\mathcal{P}}_{S_A}$ and $\mathcal{Q}$ as $\pi^*_{S_A,(\mathcal{X},\mathcal{A},\mathcal{Y})}$ and the optimal transport between $\tilde{\mathcal{P}}_{S_P}$ and $\mathcal{Q}$ as $\pi^*_{S_P,(\mathcal{X},\mathcal{A},\mathcal{Y})}$. Then consider the following coupling between $\tilde{\mathcal{P}}_{S_A}, \tilde{\mathcal{P}}_{S_P}$, and $\mathcal{Q}$:

$$\pi((x_i^A a_i^A), (x_j^P, y_j^P), (x, a, y)) = \pi^*_{S_A,(\mathcal{X},\mathcal{A},\mathcal{Y})}((x_i^A a_i^A), (x, a, y)) \cdot \pi^*_{S_P,(\mathcal{X},\mathcal{A},\mathcal{Y})}((x_j^P, y_j^P), (x, a, y)).$$

which is a product of $\pi^*_{S_A,(\mathcal{X},\mathcal{A},\mathcal{Y})}$ and $\pi^*_{S_P,(\mathcal{X},\mathcal{A},\mathcal{Y})}$. This π is clearly a feasible solution for (2). $\pi_{S_A,(\mathcal{X},\mathcal{A},\mathcal{Y})} = \pi^*_{S_A,(\mathcal{X},\mathcal{A},\mathcal{Y})}$ which witnesses that its Wasserstein distance to $\tilde{\mathcal{P}}_{S_A}$ is at most r_A, and the same argument applies for $\tilde{\mathcal{P}}_{S_P}$. Also, its marginal distribution over S_A and S_P will be exactly $\tilde{\mathcal{P}}_{S_A}$ and $\tilde{\mathcal{P}}_{S_P}$ respectively because both $\pi^*_{S_A,(\mathcal{X},\mathcal{A},\mathcal{Y})}$ and $\pi^*_{S_P,(\mathcal{X},\mathcal{A},\mathcal{Y})}$ is a valid coupling for $\tilde{\mathcal{P}}_{S_A}$ and $\tilde{\mathcal{P}}_{S_P}$ respectively.

Therefore, as their feasible solution spaces are equivalent and the objective functions are the same, we must have

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

□

A.2 Feasibility of Problem (2)

Here we focus on the feasibility of problem (2): more specifically, how big r_A and r_P needs to be in order for $W(S_A, S_P, r_A, r_P)$ to be a non-empty set.

Theorem 3.2. *$D_{d_{\mathcal{X}}}(\tilde{\mathcal{P}}_{S_A^{\mathcal{X}}}, \tilde{\mathcal{P}}_{S_P^{\mathcal{X}}}) \leq r_A + r_P$, if and only if there exists a feasible solution for (2).*

Proof. ($\Rightarrow$) **direction:** Suppose $\pi^* = \arg\min_{\pi \in \Pi(\tilde{\mathcal{P}}_a^X, \tilde{\mathcal{P}}_a^X)} \mathbb{E}_\pi[d(x, x')]$ is the coupling between $\tilde{\mathcal{P}}_{S_a^X}$ and $\tilde{\mathcal{P}}_{S_p^X}$ from that results in the Wasserstein distance $D_{d_{\mathcal{X}}}(\tilde{\mathcal{P}}_{S_A^{\mathcal{X}}}, \tilde{\mathcal{P}}_{S_P^{\mathcal{X}}}) = \mathbb{E}_{(x_i^A, x_j^P)\sim\pi^*}[d_X(x_i^A, x_j^P)]$.

For every $i \in [n_A]$ and $j \in [n_P]$, define

$$\begin{aligned} x^*_{i,j} &= x_i^A - \frac{r_A}{r_A + r_P}(x_i^A - x_j^P) \\ &= x_j^P + \frac{r_p}{r_A + r_P}(x_i^A - x_j^P) \end{aligned}$$

which is essentially a weighted average of x_i^A and x_j^P.

Note that we have

$$\begin{aligned}||x_{i,j}^* - x_i^A|| &= ||x_i^A - \frac{r_A}{r_A + r_P}(x_i^A - x_j^P) - x_i^A|| \\ &= \frac{r_A}{r_A + r_P}||x_i^A - x_j^P||\end{aligned}$$

$$\begin{aligned}||x_{i,j}^* - x_j^P|| &= ||x_j^P + \frac{r_P}{r_A + r_P}(x_i^A - x_j^P) - x_i^P|| \\ &= \frac{r_P}{r_A + r_P}||x_i^A - x_j^P||\end{aligned}$$

Then, construct $\pi_{i,j}^{a,y}$ as follows:

$$\pi_{i,j}^{y_j^P}(x_{i,j}^*, a_i^A) = \pi^*(x_i^A, x_j^P)$$

and 0 otherwise: in other words, for each (i,j), there's a point mass of $\pi^*(x_i^A, x_j^P)$ at $x_{i,j}^*, a_{i,j}^A$ with $y = y_j^P$. We now show that the constructed coupling $\pi_{i,j}^y$ is a feasible solution for (12).

First, note that we can prove that its marginal distribution transport cost is bounded by r_A and r_P. In the case of S_A, we have

$$\begin{aligned}\sum_{i=1}^{n_A}\sum_{j=1}^{n_P}\sum_{y\in\mathcal{Y}}\int_{\mathcal{X},\mathcal{A}} d_A^i(x,a)\pi_{i,j}^y(dx,da) &= \sum_{i=1}^{n_A}\sum_{j=1}^{n_p} d_A^i(x_{i,j}^*)\pi_{i,j}^{y_j^p}(x_{i,j}^*, a_i^A) \\ &= \sum_{i=1}^{n_A}\sum_{j=1}^{n_p}\left(||x_{i,j}^* - x_i^A|| + \kappa_A||a_i^A - a_i^A||\right)\pi^*(x_i^A, x_j^P) \\ &= \frac{r_A}{r_A + r_P}\sum_{i=1}^{n_A}\sum_{j=1}^{n_p}||x_j^P - x_i^A||\pi^*(x_i^A, x_j^P) \\ &\leq r_A.\end{aligned}$$

For S_P, we can similarly show

$$\begin{aligned}\sum_{i=1}^{n_A}\sum_{j=1}^{n_P}\sum_{y\in\mathcal{Y}}\int_{\mathcal{X},\mathcal{A}} d_P^j(x,a)\pi_{i,j}^y(dx,da) &= \sum_{i=1}^{n_A}\sum_{j=1}^{n_p} d_P^j(x_{i,j}^*)\pi_{i,j}^{y_j^p}(x_{i,j}^*, a_i^A) \\ &= \sum_{i=1}^{n_A}\sum_{j=1}^{n_p}\left(||x_{i,j}^* - x_j^P|| + \kappa_P|y_j^P - y_j^P|\right)\pi^*(x_i^A, x_j^P) \\ &= \frac{r_P}{r_A + r_P}\sum_{i=1}^{n_A}\sum_{j=1}^{n_p}||x_j^P - x_i^A||\pi^*(x_i^A, x_j^P) \\ &\leq r_P.\end{aligned}$$

Finally, the constructed $\pi_{i,j}^{a,y}$ is a valid coupling:

$$\begin{aligned}\sum_{j=1}^{n_P}\sum_{y\in\mathcal{Y}}\int_{\mathcal{X},\mathcal{A}}\pi_{i,j}^y(dx,da) &= \sum_{j=1}^{n_P}\pi^*(x_i^A, x_j^P) = \frac{1}{n_A} \quad \forall i \in [n_A] \\ \sum_{i=1}^{n_A}\sum_{y\in\mathcal{Y}}\int_{\mathcal{X},\mathcal{A}}\pi_{i,j}^y(dx,da) &= \sum_{i=1}^{n_A}\pi^*(x_i^A, x_j^P) = \frac{1}{n_P} \quad \forall j \in [n_P],\end{aligned}$$

as π^* was a valid coupling between $\tilde{\mathcal{P}}_{S_A}$ and $\tilde{\mathcal{P}}_{S_P}$.

($\Leftarrow$) **direction:** We'll use $\pi^a_{i,j}$ to denote the feasible solution to (2). Now, construct a coupling π such that the expected transport cost between $\tilde{\mathcal{P}}_{S^{\mathcal{X}}_A}$ and $\tilde{\mathcal{P}}_{S^{\mathcal{X}}_P}$ under π is at most $r_A + r_P$, meaning the Wasserstein distance is at most $\min(r_a, r_p)$.

Construct the coupling π between $\tilde{\mathcal{P}}_{S^{\mathcal{X}}_A}$ and $\tilde{\mathcal{P}}_{S^{\mathcal{X}}_P}$ as

$$\pi(x^A_i, x^P_j) = \sum_{y\in\mathcal{Y}} \int_{\mathcal{X},\mathcal{A}} \pi^y_{i,j}(dx, da).$$

It's easy to see that π is a valid coupling as

$$\sum_{i=1}^{n_A} \pi(x^A_i, x^P_j) = \sum_{i=1}^{n_A}\sum_{y\in\mathcal{Y}} \int_{\mathcal{X},\mathcal{A}} \pi^y_{i,j}(dx, da) = \frac{1}{n_P} \quad \text{and} \quad \sum_{j=1}^{n_P} \pi(x^A_i, x^P_j) = \sum_{j=1}^{n_P}\sum_{y\in\mathcal{Y}} \int_{\mathcal{X},\mathcal{A}} \pi^y_{i,j}(dx, da) = \frac{1}{n_A}$$

for each $i \in [n_A]$ and $j \in [n_P]$.

Finally, due to its feasibility, we get

$$\begin{aligned}
\sum_{i=1}^{n_A}\sum_{j=1}^{n_P}\sum_{y\in\mathcal{Y}} \int_{\mathcal{X},\mathcal{A}} d^i_A(x,y)\pi^y_{i,j}(dx, da) &\le r_A \\
\sum_{i=1}^{n_A}\sum_{j=1}^{n_P}\sum_{y\in\mathcal{Y}} \int \left(||x^A_i - x|| + \kappa_a||a^A_i - a||\right)\pi^y_{i,j}(dx, da) &\le r_A \\
\sum_{i=1}^{n_A}\sum_{j=1}^{n_P}\sum_{y\in\mathcal{Y}} \int_{\mathcal{X},\mathcal{A}} ||x^A_i - x||\pi^y_{i,j}(dx, da) &\le r_A \qquad (11)
\end{aligned}$$

Similarly, we get

$$\sum_{i=1}^{n_A}\sum_{j=1}^{n_P}\sum_{y\in\mathcal{Y}} \int_{\mathcal{X},\mathcal{A}} ||x^P_j - x||\pi^y_{i,j}(dx, da) \le r_P \qquad (12)$$

By adding (11) and (12), we get

$$\begin{aligned}
\sum_{i=1}^{n_A}\sum_{j=1}^{n_P}\sum_{y\in\mathcal{Y}} \int_{\mathcal{X},\mathcal{A}} \left(||x^A_i - x|| + ||x^P_j - x||\right)\pi^y_{i,j}(dx, da) &\le r_A + r_P \\
\sum_{i=1}^{n_A}\sum_{j=1}^{n_P}\sum_{y\in\mathcal{Y}} \int_{\mathcal{X},\mathcal{A}} ||x^A_i - x^P_j||\pi^y_{i,j}(dx, da) &\le r_A + r_P \\
\sum_{i=1}^{n_A}\sum_{j=1}^{n_P} ||x^A_i - x^P_j||\pi(x^A_i, x^P_j) &\le r_A + r_P.
\end{aligned}$$

The second line follows from the triangle inequality $||x^A_i - x^P_j|| \le ||x^A_i - x|| + ||x^P_j - x||$. Therefore, we have $D_{d_{\mathcal{X}}}(\tilde{\mathcal{P}}_{S^{\mathcal{X}}_A}, \tilde{\mathcal{P}}_{S^{\mathcal{X}}_P}) \le r_A + r_P$. □

A.3 Missing Details from Section 3.2

Theorem 3.3. *Assume $\mathcal{X}$ and $\mathcal{A}$ are compact spaces. 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.*

Proof. This theorem essentially follows from Fenchel-Rokafellar Duality which is formally stated later in the proof. Before applying the duality theorem, it is instructive to take a look at the corresponding Lagrangian for (2):

$$\begin{aligned}\mathcal{L}(\pi, \alpha_A, \alpha_P, \{\beta_{i,j}\}_{i,j}) = &\sum_{i=1}^{n_A}\sum_{j=1}^{n_P}\sum_{y\in\mathcal{Y}}\int_{\mathcal{X},\mathcal{A}} \ell(\theta,(x,a,y))\pi_{i,j}^y(dx,da)) \\ &+ \alpha_A\left(r_A - \sum_{i=1}^{n_A}\sum_{j=1}^{n_P}\sum_{y\in\mathcal{Y}}\int_{\mathcal{X},\mathcal{A}} d_A^i(x,a)\pi_{i,j}^y(dx,da)\right) \\ &+ \alpha_P\left(r_P - \sum_{i=1}^{n_A}\sum_{j=1}^{n_P}\sum_{y\in\mathcal{Y}}\int d_P^j(x,y)\pi_{i,j}^y(dx,da)\right) \\ &+ \sum_{i=1}^{n_A}\beta_i\left(\frac{1}{n_A} - \sum_{j=1}^{n_P}\sum_{y\in Y}\int_{\mathcal{X},\mathcal{A}}\pi_{i,j}^y(dx,da)\right) \\ &+ \sum_{j=1}^{n_P}\beta_j'\left(\frac{1}{n_P} - \sum_{i=1}^{n_A}\sum_{y\in Y}\int_{\mathcal{X},\mathcal{A}}\pi_{i,j}^y(dx,da)\right).\end{aligned}$$

Rearranging the terms yields

$$\begin{aligned}&\mathcal{L}(\pi, \alpha_A, \alpha_P, \{\beta_i\}, \{\beta_j'\}\}) \\ &= \sum_{i=1}^{n_A}\sum_{j=1}^{n_P}\sum_{y\in\mathcal{Y}}\int_{\mathcal{X},\mathcal{A}}\left(\ell(\theta,(x,a,y)) - \alpha_A d_A^i(x,a) - \alpha_P d_P^j(x,y) - \beta_i - \beta_j'\right)\pi_{i,j}^y(dx,da) \\ &+ \alpha_A r_A + \alpha_P r_P + \frac{1}{n_A}\sum_{i=1}^{n_A}\beta_i + \frac{1}{n_P}\sum_{j=1}^{n_P}\beta_j'.\end{aligned}$$

Note that the optimal primal value can be written in terms of its Lagrangian:

$$p^*(\theta, r_A, r_P) = \sup_{\pi}\inf_{\alpha_A,\alpha_P,\{\beta_i\},\{\beta_j'\}}\mathcal{L}(\pi,\alpha_A,\alpha_P,\{\beta_i\},\{\beta_j'\}).$$

For notational economy, we'll write

$$\psi(\alpha_A,\alpha_P,\{\beta_i\},\{\beta_j'\},x,a,y) = -\alpha_A d_A^i(x,a) - \alpha_P d_P^j(x,y) - \beta_i - \beta_j'$$

Now, we state Fenchel's duality theorem:

Theorem A.1 (Fenchel-Rokafellar Duality)**.** *Let E be a normed vector space, and let $f, g : E \to \mathbb{R}\cup\{+\infty\}$ be two convex functions. Assume there exists $z_0 \in E$ such that $f(z_0) < \infty$ and $g(z_0) < \infty$, and f and g are continuous at z_0. Then,*

$$\inf_E(g+f) = \sup_{z^*\in E^*}(-g^*(-z^*) - f^*(z^*))$$

By Riesz's theorem, we have that the dual space of the Radon measure $\pi_{i,j}^y$ is the continuous bounded functions which we denote as $u(i,j,x,a,y)$. In our case, define

$$f(u) = \begin{cases}0 & \text{if } u(i,j,x,a,y) + \ell(\theta,(x,a,y)) \le 0 \text{ for all } i\in[n_A] \text{ and } j\in[n_P] \\ \infty & \text{otherwise}\end{cases}$$

$$
\begin{aligned}
&g(u) \\
&= \begin{cases} \left(\alpha_A r_A + \alpha_P r_P + \frac{1}{n_A}\sum_{i=1}^{n_A}\beta_i + \frac{1}{n_P}\sum_{j=1}^{n_P}\beta'_j\right) & \text{if } u(i,j,x,a,y) = \psi(\alpha_A,\alpha_P,\{\beta_i\},\{\beta'_j\},x,a,y) \\ & \text{for some } \alpha_A,\alpha_P,\{\beta_i\},\{\beta'_j\} \\ \infty & \text{otherwise} \end{cases}
\end{aligned}
$$

Note that both f and g are convex:

1. f is convex
   Consider any u, v such that $f(u) < \infty$ and $f(v) < \infty$, then $u(i,j,x,a,y) \leq -\ell(\theta,x,a,y)$ and $v(i,j,x,a,y) \leq -\ell(\theta,x,a,y)$. Then, because $tu(i,j,x,a,y) + (1-t)v(i,j,x,a,y) \leq -\ell(\theta,x,a,y)$, we have
   $$tf(u) + (1-t)f(v) = 0 = f(t(u) + (1-t)v).$$
   If either $f(u) = \infty$ or $f(v) = \infty$, then
   $$f(t(u) + (1-t)v) \leq tf(u) + (1-t)f(v).$$

2. g is convex
   Suppose u, v is such that $g(u) < \infty$ and $g(v) < \infty$ and $g(u) = \alpha_A^u r_A + \alpha_P^u r_P + \frac{1}{n_A}\sum_{i\in[n_A]}\beta_i^u + \frac{1}{n_P}\sum_{j\in[n_P]}\beta'^u_j$ and $g(v) = \alpha_A^v r_A + \alpha_P^v r_P + \frac{1}{n_A}\sum_{i\in[n_A]}\beta_i^v + \frac{1}{n_P}\sum_{j\in[n_P]}\beta'^v_j$. Then, we have $tg(u) + (1-t)g(v) = g(tu + (1-t)v)$. If $g(u) = \infty$ or $g(v) = \infty$, it's easy to see that $g(tu + (1-t)v) \leq \infty$ as well.

Note that

$$
\begin{aligned}
&\inf_u (f(u) + g(u)) \\
&= \inf_{\ell(\theta,(x,a,y)) - \alpha_A d_A^i(x,a) - \alpha_P d_P^j(x,y) - \beta_i - \beta'_j \leq 0} \left(\alpha_A r_A + \alpha_P r_P + \frac{1}{n_A}\sum_{i\in[n_A]}\beta_i + \frac{1}{n_P}\sum_{j\in[n_P]}\beta'_j\right) \\
&= d^*(r_A, r_P)
\end{aligned}
$$

We derive their convex conjugates:

$$
\begin{aligned}
f^*(\{\pi_{i,j}^{a,y}\}) &= \sup_{u+\ell\leq 0} \sum_{i=1}^{n_A}\sum_{j=1}^{n_P}\sum_{y\in\mathcal{Y}}\int_{\mathcal{X},\mathcal{A}} u(i,j,x,a,y)\pi_{i,j}^y(dx,da) \\
&= -\sum_{i=1}^{n_A}\sum_{j=1}^{n_P}\sum_{y\in\mathcal{Y}}\int_{\mathcal{X},\mathcal{A}} \ell(\theta,(x,a,y))\pi_{i,j}^y(dx,da)
\end{aligned}
$$

$$
\begin{aligned}
&g^*(\{\pi_{i,j}^{a,y}\}) \\
&= \sup_{u(i,j,x,a,y)=\psi(\alpha_A,\alpha_P,\{\beta_i\},\{\beta'_j\},x,a,y)} \sum_{i=1}^{n_A}\sum_{j=1}^{n_P}\sum_{y\in\mathcal{Y}}\int_{\mathcal{X},\mathcal{A}} u(i,j,x,a,y)\pi_{i,j}^y(dx,da) \\
&- \left(\alpha_A r_A + \alpha_P r_P + \frac{1}{n_A}\sum_{i\in[n_A]}\beta_i + \frac{1}{n_P}\sum_{j\in[n_P]}\beta'_j\right) \\
&= \sup_{\alpha_A,\alpha_P,\{\beta_i\},\{\beta'_j\}_j} \sum_{i=1}^{n_A}\sum_{j=1}^{n_P}\sum_{y\in\mathcal{Y}}\left(\int_{\mathcal{X},\mathcal{A}} -\alpha_A d_A^i(x,a) - \alpha_P d_P^j(x,y) - \beta_i - \beta'_j\right)\pi_{i,j}^y(dx,da) \\
&- \left(\alpha_A r_A + \alpha_P r_P + \frac{1}{n_A}\sum_{i\in[n_A]}\beta_i + \frac{1}{n_P}\sum_{j\in[n_P]}\beta'_j\right).
\end{aligned}
$$

Also, note that

$$\begin{aligned}
&\sup_{\pi}(-g^*(-\pi) - f^*(\pi)) \\
&= \sup_{\pi} \inf_{\alpha_A,\alpha_P,\{\beta_{i,j}\}_{i,j}} \sum_{i=1}^{n_A}\sum_{j=1}^{n_P}\sum_{y\in\mathcal{Y}} \left(\int_{\mathcal{X},\mathcal{A}} -\alpha_A d_A^i(x,a) - \alpha_P d_P^j(x,y) - \beta_i - \beta'_j\right) \pi_{i,j}^y(dx,da) \\
&+ \left(\alpha_A r_A + \alpha_P r_P + \frac{1}{n_A}\sum_{i\in[n_A]}\beta_i + \frac{1}{n_P}\sum_{j\in[n_P]}\beta'_j\right) \\
&+ \sum_{i=1}^{n_A}\sum_{j=1}^{n_P}\sum_{y\in\mathcal{Y}}\int_{\mathcal{X},\mathcal{A}} \ell(\theta,(x,a,y))\pi_{i,j}^y(dx,da) \\
&= \sup_{\pi} \inf_{\alpha_A,\alpha_P,\{\beta_i\},\{\beta'_j\}} \mathcal{L}(\pi,\alpha_A,\alpha_P,\{\beta_i\},\{\beta'_j\}) \\
&= p^*(\theta, r_A, r_P).
\end{aligned}$$

Therefore, by Theorem A.1, we see that $p^*(\theta, r_A, r_P) = d^*(\theta, r_A, r_P)$. □

A.4 Missing Details from Section 3.3

Lemma 3.1. *Fix any θ, (x_i^A, a_i^A, x_j^P), and $(\alpha_A, \alpha_P, \kappa_A)$. If $||\theta[1:m_1]||_{p,*} > \alpha_A + \alpha_P$ or $||\theta[m_1+1 : m_1+m_2]||_{p',*} > \kappa_A\alpha_A$, then $\sup_{(x,a)} h(\theta,(x,a)) - \alpha_A||x_i^A - x||_p - \alpha_P||x_j^P - x||_p - \alpha_A\kappa_A||a_i^A - a||_{p'} = \infty$. Otherwise, we have*

$$\begin{aligned}
&\sup_{(x,a)} h(\theta,(x,a)) - \alpha_A||x_i^A - x||_p - \alpha_P||x_j^P - x||_p - \alpha_A\kappa_A||a_i^A - a||_{p'} \\
&= \sup_{b\in[0,1]} -f^*(b) + (g_1^i \square g_2^j)(b\theta[1:m_1]) + \langle b\theta[m_1+1:m_1+m_2], a_i^A\rangle
\end{aligned}$$

where

$$g_1^i(\theta) = \begin{cases} \langle\theta, x_i^A\rangle & \text{if } ||\theta||_{p,*} \le \alpha_A \\ \infty & \text{otherwise} \end{cases} \qquad g_2^j(\theta) = \begin{cases} \langle\theta, x_j^P\rangle & \text{if } ||\theta||_{p,*} \le \alpha_P \\ \infty & \text{otherwise} \end{cases}$$

and $(g_1^i \square g_2^j)(\theta) = \inf_{\theta_1+\theta_2=\theta} g_1^i(\theta_1) + g_2^j(\theta_2)$ is the infimal convolution of g_1^i and g_2^j.

Proof. Noting that h is convex and thus h is equal to its biconjugate h^{**}, we have

$$\sup_{(x,a)} h(\theta,(x,a)) - \alpha_A||x_i^A - x||_p - \alpha_P||x_j^P - x||_p - \alpha_A\kappa_A||a_i^A - a||_{p'}$$

$$= \sup_{(x,a)} \sup_{b\in[0,1]} \langle b\theta,(x,a)\rangle - f^*(b) - \alpha_A||x_i^A - x||_p - \alpha_P||x_j^P - x||_p - \alpha_A\kappa_A||a_i^A - a||_{p'}$$

$$= \sup_{b\in[0,1]} \sup_{(x,a)} \langle b\theta,(x,a)\rangle - f^*(b) - \sup_{||q_1||_{p,*}\le\alpha_A} \langle q_1, x_i^A - x\rangle - \sup_{||q_2||_{p,*}\le\alpha_P} \langle q_2, x_j^P - x\rangle - \sup_{||q_3||_{p',*}\le\alpha_A\kappa_A} \langle q_3, a_i^A - a\rangle$$

$$= \sup_{b\in[0,1]} \sup_{(x,a)} \langle b\theta,(x,a)\rangle - f^*(b)$$

$$- \sup_{||q_1||_{p,*}\le\alpha_A} \langle (q_1,0),(x_i^A,a)-(x,a)\rangle - \sup_{||q_2||_{p,*}\le\alpha_P} \langle (q_2,0),(x_j^P,a)-(x,a)\rangle - \sup_{||q_3||_{p',*}\le\alpha_A\kappa_A} \langle (0,q_3),(x,a_i^A)-(x,a)\rangle$$

$$= \sup_{b\in[0,1]} \sup_{(x,a)} \inf_{\substack{||q_1||_{p,*}\le\alpha_A,\\ ||q_2||_{p,*}\le\alpha_P,\\ ||q_3||_{p',*}\le\alpha_A\kappa_A}} \langle b\theta,(x,a)\rangle - f^*(b)$$

$$- \langle (q_1,0),(x_i^A,a)-(x,a)\rangle - \langle (q_2,0),(x_j^P,a)-(x,a)\rangle - \langle (0,q_3),(x,a_i^A)-(x,a)\rangle$$

$$= \sup_{b\in[0,1]} \sup_{(x,a)} \inf_{\substack{||q_1||_{p,*}\le\alpha_A,\\ ||q_2||_{p,*}\le\alpha_P,\\ ||q_3||_{p',*}\le\alpha_A\kappa_A}} \langle b\theta + (q_1,0) + (q_2,0) + (0,q_3),(x,a)\rangle - f^*(b) - \langle q_1, x_i^A\rangle - \langle q_2, x_j^P\rangle - \langle q_3, a_i^A\rangle.$$

We can swap the order of inf and sup due to proposition 5.5.4 of Bertsekas.

$$= \sup_{b\in[0,1]} \inf_{\substack{||q_1||_{p,*}\le\alpha_A,\\ ||q_2||_{p,*}\le\alpha_P,\\ ||q_3||_{p',*}\le\alpha_A\kappa_A}} \sup_{(x,a)} \langle b\theta + (q_1,0) + (q_2,0) + (0,q_3),(x,a)\rangle - f^*(b) - \langle q_1, x_i^A\rangle - \langle q_2, x_j^P\rangle - \langle q_3, a_i^A\rangle.$$

Note that unless $b\theta + (q_1,0) + (q_2,0) + (0,q_3) = 0$, (x,a) can be chosen arbitrarily big. Also, if $\theta + (q_1,0) + (q_2,0) + (0,q_3) \neq 0$, then b can be chosen to be 1. Therefore, if there doesn't exist (q_1,q_2,q_3) such that $\theta + (q_1,0) + (q_2,0) + (0,q_3) = 0$, everything evaluates to ∞. In other words, the expression evaluates to ∞ unless both of the following conditions are true:

1. $||\theta[1:m_1]||_{p,*} \le \alpha_A + \alpha_P$
2. $||\theta[m_1+1:m_1+m_2]||_{p',*} \le \kappa_A\alpha_A$

as $q_1 = \frac{-\alpha_A}{||\theta[1:m_1]||}\theta[1:m_1]$, $q_2 = \frac{-\alpha_P}{||\theta[1:m_1]||}\theta[1:m_1]$, and $q_3 = \frac{\kappa_A\alpha_A}{||\theta[m_1+1:m_1+m_2]||_*}\theta[m_1+1:m_1+m_2]$ is one such triplet that satisfy $\theta + (q_1,0) + (q_2,0) + (0,q_3) = 0$.

Now, suppose θ satisfies the above conditions as we know it evaluates to ∞ otherwise. Then, we get

$$= \sup_{b\in[0,1]} -f^*(b)$$

$$+ \inf_{\substack{||q_1||_{p,*}\le\alpha_A,\\ ||q_2||_{p,*}\le\alpha_P,\\ ||-b\theta[m_1+1:m_1+m_2]||_{p',*}\le\alpha_A\kappa_A}} \begin{cases} -\langle q_1, x_i^A\rangle - \langle q_2, x_j^P\rangle + \langle b\theta[m_1+1:m_1+m_2], a_i^A\rangle & \text{if } b\theta[1:m_1] + q_1 + q_2 = 0 \\ \infty & \text{otherwise} \end{cases}$$

$$= \sup_{b\in[0,1]} -f^*(b) + \inf_{\substack{||q_1||_{p,*}\le\alpha_A,\\ ||q_2||_{p,*}\le\alpha_P}} \begin{cases} \langle q_1, x_i^A\rangle + \langle q_2, x_j^P\rangle & \text{if } q_1 + q_2 = b\theta[1:m_1] \\ \infty & \text{otherwise} \end{cases} + \langle b\theta[m_1+1:m_1+m_2], a_i^A\rangle$$

$$= \sup_{b\in[0,1]} -f^*(b) + (g_1^i \square g_2^j)(b\theta[1:m_1]) + \langle b\theta[m_1+1:m_1+m_2], a_i^A\rangle$$

□

Now, using the fact that the infimal convolution of linear functions is convex which we prove in Appendix A.5, we show how to upperbound the supremum term.

Theorem 3.5. *We 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 $||\theta_1||_{p,*} \leq \alpha_A + \alpha_P$ and $||\theta_2||_{p',*} \leq \kappa_A \alpha_A$, then*

$$R \leq f\left(\left(\frac{\min(\alpha_A, \alpha_P)||\theta_1||_*||x_i^A - x_j^P||}{\alpha_A + \alpha_P} + \frac{\langle \theta_1, \alpha_A x_i^A + \alpha_P x_j^P\rangle}{\alpha_A + \alpha_P}\right) + \langle \theta_2, a_i^A\rangle\right) - \min(\alpha_A, \alpha_P)||x_i^A - x_j^P||_p.$$

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

Proof. Because f is a convex function, its biconjugate is itself, so

$$\sup_{b\in[0,1]} -f^*(b) + b \cdot X = f(X).$$

Therefore, we have

$$\begin{aligned}
&\sup_{(x,a)} h(\theta, (x, a)) - \alpha_A||x - x_i^A||_p - \alpha_P||x - x_j^P||_p - \alpha_A\kappa_A||a_i^A - a||_{p'} \\
&= \sup_{b\in[0,1]} -f^*(b) + (g_1^i \square g_2^j)(b\theta_1) + \langle b\theta_2, a_i^A\rangle \\
&\leq \sup_{b\in[0,1]} -f^*(b) + \left(\frac{b}{\alpha_A + \alpha_P}\right)\left(\min(\alpha_A, \alpha_P)||\theta_1||_*||x_i^A - x_j^P|| + \langle \theta_1, \alpha_A x_i^A + \alpha_P x_j^P\rangle\right) - \min(\alpha_A, \alpha_P)||x_i^A - x_j^P||_p + b\langle \theta_2, a_i^A\rangle \\
&= \sup_{b\in[0,1]} -f^*(b) \\
&+ b\left(\left(\frac{1}{\alpha_A + \alpha_P}\right)\left(\min(\alpha_A, \alpha_P)||\theta_1||_*||x_i^A - x_j^P|| + \langle \theta_1, \alpha_A x_i^A + \alpha_P x_j^P\rangle\right) + \langle \theta_2, a_i^A\rangle\right) - \min(\alpha_A, \alpha_P)||x_i^A - x_j^P||_p \\
&= f\left(\left(\frac{\min(\alpha_A, \alpha_P)||\theta_1||_*||x_i^A - x_j^P|| + \langle \theta, \alpha_A x_i^A + \alpha_P x_j^P\rangle}{\alpha_A + \alpha_P}\right) + \langle \theta_2, a_i^A\rangle\right) - \min(\alpha_A, \alpha_P)||x_i^A - x_j^P||_p.
\end{aligned}$$

The first inequality follows from Theorem 3.4. ☐

Lemma A.1.

$$\inf_x(\alpha_A||x - x_i^A|| + \alpha_P||x - x_j^P||) = \min(\alpha_A, \alpha_P)||x_i^A - x_j^P||.$$

and when $\alpha_A < \alpha_P$, the infimum is achieved at $x = x_j^P$ and otherwise at x_i^A.

Proof.

$$\begin{aligned}
&\inf_x \sup_{||q_1||_*\leq\alpha_A} \langle q_1, x - x_i^A\rangle + \sup_{||q_2||_*\leq\alpha_P} \langle q_2, x - x_j^P\rangle \\
&= \inf_x \sup_{\substack{||q_1||_*\leq\alpha_A, \\ ||q_2||_*\leq\alpha_P}} \langle q_1 + q_2, x\rangle + \langle q_1, -x_i^A\rangle + \langle q_2, -x_j^P\rangle
\end{aligned}$$

We are able to swap inf and sup due to proposition 5.5.4 of Bertsekas.

$$\begin{aligned}
&= \sup_{\substack{||q_1||_*\leq\alpha_A, \\ ||q_2||_*\leq\alpha_P}} \inf_x \langle q_1 + q_2, x\rangle + \langle q_1, -x_i^A\rangle + \langle q_2, -x_j^P\rangle \\
&= \sup_{||q||_*\leq\min(\alpha_A,\alpha_P)} \langle q, -x_i^A + x_j^P\rangle \\
&= \min(\alpha_A, \alpha_P)||x_i^A - x_j^P||.
\end{aligned}$$

The second inequality holds true because The sum of two norms has to be non-negative, and unless $q_1 = q_2$, the inf term can be made arbitrarily small, meaning we need to set $q_1 = -q_2$. ☐

Theorem 3.6. *Suppose $p \neq 1$ and $p \neq \infty$. If $||\theta_1||_{p,*} \leq \alpha_A + \alpha_P$ and $||\theta_2||_{p',*} \leq \kappa_A \alpha_A$, then*

$$\big(h(\theta, (\hat{x}_{i,j}, a_i^A)) - \hat{\alpha}||x_i^A - x_j^P||_p\big) - R \leq 2\hat{\alpha}||x_i^A - x_j^P||.$$

Proof. Fix $i, j, \theta, \alpha_A, \alpha_P$. For convenience, we write

$$\hat{x} = \begin{cases} x_j^P & \text{if } \alpha_A < \alpha_P \\ x_i^A & \end{cases} \qquad \text{and} \quad \hat{\alpha} = \min(\alpha_A, \alpha_P).$$

Also, we write the supremum $\sup_{(x,a)} f(\langle\theta, (x,a)\rangle) - \alpha_A||x - x_i^A|| - \alpha_P||x - x_j^P|| - \kappa_A\alpha_A|a_i^A - a|$ is achieved at (x^*, a^*). We write $U(x,a) = -\alpha_A||x - x_i^A|| - \alpha_P||x - x_j^P|| - \kappa_A\alpha_A|a_i^A - a|$, meaning $(x^*, a^*) = \arg\max f(\langle\theta, (x,a)\rangle) + U(x,a)$.

From Theorem 3.5, we have

$$f(\langle\theta, (x^*, a^*)\rangle) + U(x^*, a^*) \leq f(\langle\theta, (\check{x}, a_i^A)\rangle) + U(\hat{x}, a_i^A)$$

where

$$\check{x} = \frac{\alpha_A x_i^A + \alpha_P x_j^P + \hat{\alpha} v(\theta_1)||x_i^A - x_j^P||}{\alpha_A + \alpha_P}.$$

Therefore, we have

$$f(\langle\theta, (\hat{x}, a_i^A)\rangle) + U(\hat{x}, a_i^A) \leq f(\langle\theta, (x^*, a^*)\rangle) + U(x^*, a^*) \leq f(\langle\theta, (\check{x}, a_i^A)\rangle) + U(\hat{x}, a_i^A).$$

In other words,

$$(f(\langle\theta, (x^*, a^*)\rangle) + U(x^*, a^*)) - (f(\langle\theta, (\hat{x}, a_i^A)\rangle) + U(\hat{x}, a_i^A)) \leq f(\langle\theta, (\check{x}, a_i^A)\rangle) - f(\langle\theta, (\hat{x}, a_i^A)\rangle).$$

Hölder's inequality gives us

$$\left|\sum_{c\in[m_1]} \theta[c](x_i^A - x_j^P)[c]\right| \leq \sum_{c\in[m_1]} \big|\theta[c](x_i^A - x_j^P)[c]\big| \leq ||\theta_1||_*||x_i^A - x_j^P||$$

Suppose $\alpha_A < \alpha_P$, meaning $\hat{x} = x_j^P$

$$\begin{aligned}
&f(\langle\theta, (\check{x}, a_i^A)\rangle) - f(\langle\theta, (\hat{x}, a_i^A)\rangle) \\
&\leq \frac{\alpha_A||\theta_1||_*||x_i^A - x_j^P|| + \langle\theta_1, \alpha_A x_i^A + \alpha_P x_j^P\rangle}{\alpha_A + \alpha_P} + \langle\theta_2, a_i^A\rangle - \langle\theta, (\hat{x}, a_i^A)\rangle \\
&= \frac{\alpha_A||\theta_1||_*||x_i^A - x_j^P|| + \langle\theta_1, \alpha_A x_i^A + \alpha_P x_j^P\rangle}{\alpha_A + \alpha_P} - \frac{\alpha_A \sum_{c\in[m_1]} \theta[c](x_j^P - x_i^A)[c] + \langle\theta, \alpha_A x_i^A + \alpha_P x_j^P\rangle}{\alpha_A + \alpha_P} \\
&= \frac{\alpha_A||\theta_1||_*||x_i^A - x_j^P|| - \alpha_A(\sum_{c\in[m_1]} \theta[c](x_j^P - x_i^A)[c])}{\alpha_A + \alpha_P} \\
&\leq \frac{2\alpha_A||\theta_1||_*||x_i^A - x_j^P||}{\alpha_A + \alpha_P} \\
&\leq 2\alpha_A||x_i^A - x_j^P||
\end{aligned}$$

where the first inequality follows from f's 1-Lipschitzness — i.e. $|f(x) - f(x')| \leq |x - x'|$. The same argument works when $\alpha_A \geq \alpha_P$. ∎

A.5 Infimal Convolution

Now, we prove a few lemmas regarding the infimal convolution of two linear functions[9]. Since the domain of g_1^i and g_2^j invovles the same p-norm, we elide p in the following lemmas.

Lemma A.2. *$(g_1^i \square g_2^j)(\theta)$ is convex in θ.*

Proof. In order to show a function is convex, it suffices to show that its epigraph is convex. Note that epigraphs of g_1^i and g_2^j are both convex:

$$S_1 = \text{epi } g_1^i = \{(q, r) : ||q||_* \leq \alpha_A, r \geq g_1^i(q)\}$$

$$S_2 = \text{epi } g_2^j = \{(q, r) : ||q||_* \leq \alpha_P, r \geq g_2^j(q)\}.$$

Note that the epigraph of $(g_1^i \square g_2^j)(\theta)$ is the Minkowski sum of S_1 and S_2 [Strömberg, 1994]:

$$S_3 = \text{epi } (g_1^i \square g_2^j) = \{(x_1 + x_2, r_1 + r_2) : (x_1, r_1) \in S_1, (x_2, r_2) \in S_2\}.$$

For any $(x_1+x_2, r_1+r_2) \in S_3$ and $(x_1'+x_2', r_1'+r_2') \in S_3$ where $(x_1, r_1), (x_1', r_1') \in S_1$ and $(x_2, r_2), (x_2', r_2') \in S_2$, the convex combination with $t \in [0, 1]$

$$(t(x_1 + x_2) + (1 - t)(x_1' + x_2'), t(r_1 + r_2) + (1 - t)(r_1' + r_2')$$

must belong in S_3 because $(tx_1 + (1 - t)x_1', tr_1 + (1 - t)r_1') \in S_1$ and $(tx_2 + (1 - t)x_2', tr_2 + (1 - t)r_2') \in S_2$ due to the convexity of S_1 and S_2.

□

Lemma A.3.

$$(g_1^i \square g_2^j)(0) = -\min(\alpha_A, \alpha_P)||x_j^P - x_i^A||.$$

Proof.

$$\begin{aligned}
(g_1^i \square g_2^j)(0) &= \inf_{q:||q||_* \leq \min(\alpha_A, \alpha_P)} g_1^i(q) + g_2^j(-q) \\
&= \inf_{q:||q||_* \leq \min(\alpha_A, \alpha_P)} \langle q, x_i^A \rangle - \langle q, x_j^P \rangle \\
&= - \sup_{q:||q||_* \leq \min(\alpha_A, \alpha_P)} \langle q, -x_i^A + x_j^P \rangle \\
&= -\min(\alpha_A, \alpha_P)||x_i^A - x_j^P||.
\end{aligned}$$

□

Now, for any q, we write

$$v(q) = \arg \max_{v:||v|| \leq 1} \langle v, q \rangle.$$

Note that $\langle v(q), q \rangle = ||q||_*$. In words, $v(q)$ is the vector whose inner product with q evaluates to the dual norm of q. Note that for any scalar $c > 0$, $v(q) = v(cq)$, meaning only the direction matters.

In the lemma below, we show that given two different directions (q, q'), we must have $v(q) \neq v(q')$.

Lemma A.4. *Suppose the norm $||\cdot||$ is some p-norm where $p \neq 1$ and $p \neq \infty$, meaning corresponding dual norm $||\cdot||_{p,*}$ is r-norm where $r \neq 1$ and $r \neq \infty$. Given q and q' where $||q||_* = ||q'||_* = 1$ and $q \neq q'$, we must have $v(q) \neq v(q')$.*

[9] Readers more interested in the properties of infimal convolution may refer to Strömberg [1994].

Proof. For any q where $||q||_* = 1$, let's consider $v(q) = \max_{v:||v||\leq 1}\langle v, q\rangle$. Because the linear objective forces the optimal solution be at the boundary of the feasible convex set, it is equivalent to solving $\max_{v:||v||=1}\langle v, q\rangle$. Lagrange multiplier approach yields the following conditions for the optimal solution:

$$\begin{aligned} q[i] + \lambda \cdot \text{sign}(v(q)[i]) \cdot \left(\frac{|v(q)[i]|}{||v(q)||}\right)^{p-1} &= 0 \quad \forall i \in [n] \\ ||v(q)|| &= 1 \end{aligned} \tag{13}$$

where λ corresponds to the Lagrange multiplier.

Consider the following two unnormalized vectors v^{+1} and v^{-1}:

$$\begin{aligned} v^{+1}(q) &= \left(\text{sign}(q[i]) \cdot \left||q[1]|^{\frac{1}{p-1}}\right|, \ldots, \text{sign}(q[n]) \cdot \left||q[n]|^{\frac{1}{p-1}}\right|\right) \\ v^{-1}(q) &= \left(\text{sign}(-q[i]) \cdot \left||q[1]|^{\frac{1}{p-1}}\right|, \ldots, \text{sign}(-q[n]) \cdot \left||q[n]|^{\frac{1}{p-1}}\right|\right). \end{aligned}$$

The solutions to the equations in (13) are the normalized $\frac{v^{+1}(q)}{||v^{+1}(q)||}$ and $\frac{v^{-1}(q)}{||v^{-1}(q)||}$, meaning they are the local optima.

Because $\langle \frac{v^{+1}(q)}{||v^{+1}(q)||}, q\rangle = -\langle \frac{v^{-1}(q)}{||v^{-1}(q)||}, q\rangle$ and $\langle \frac{v^{+1}(q)}{||v^{+1}(q)||}, q\rangle > 0$, we must have that

$$v(q) = \arg\max_{v:||v||\leq 1} \langle v, q\rangle = \frac{v^{+1}(q)}{||v^{+1}(q)||}.$$

Hence, for any two different directions q and q', we must have that $\frac{v^{+1}(q)}{||v^{+1}(q)||}$ will be different by construction, as long as $p \neq 1$ or $p \neq \infty$. Hence, $v(q) \neq v(q')$. □

Corollary A.1. *Suppose the norm $||\cdot||$ is some p-norm where $p \neq 1$ and $p \neq \infty$. For any q where $||q||_* = \alpha$, we have that for any other q' where $q' \neq q$ and $||q'||_* \leq \alpha$,*

$$\langle v(q), q\rangle > \langle v(q), q'\rangle.$$

Proof. As said in Section 2, given any vector q, we'll write $\overline{q}_* = \frac{q}{||q||_*}$. If $\overline{q}_* = \overline{q'}_*$, then there exists some scalar $c > 1$ such that $q = cq'$ since $||q||_* > ||q'||_*$. Then, we must have

$$\langle v(q), q\rangle = c\langle v(q), q'\rangle > \langle v(q), q'\rangle$$

as $v(q) = v(q')$ in this case.

Now, in the case where $\overline{q}_* \neq \overline{q'}_*$, we see that

$$\langle v(q), q\rangle = \alpha \geq ||q'||_* = \langle v(q'), q'\rangle > \langle v(q), q'\rangle.$$

□

Lemma A.5. *Suppose the norm $||\cdot||$ is some p-norm where $p \neq 1$ and $p \neq \infty$. Fix some direction $\overline{\theta}_*$ where $||\overline{\theta}_*||_* = 1$. Then,*

$$(g_1^i \square g_2^j)((\alpha_A + \alpha_P)\overline{\theta}_*) = \langle \overline{\theta}_*, \alpha_A x_i^A + \alpha_P x_j^P\rangle.$$

Proof. We first claim that when given $(\alpha_A + \alpha_P)\overline{\theta}_*$, there exists only one pair (q_1, q_2) such that $||q_1||_* \leq \alpha_A$, $||q_2||_* \leq \alpha_P$, and $q_1 + q_2 = (\alpha_a + \alpha_P)\overline{\theta}_*$: namely,

$$q_1 = \alpha_A\overline{\theta}_* \quad \text{and} \quad q_2 = \alpha_P\overline{\theta}_*.$$

By construction, $||q_1||_* = \alpha_A$, $||q_2||_* = \alpha_P$, and $q_1 + q_2 = (\alpha_A + \alpha_P)\overline{\theta}_*$.

Now, for the sake of contradiction, suppose there exists another (q_1', q_2') such that the above condition holds true. Because $q_1' + q_2' = (\alpha_A + \alpha_P)\overline{\theta}$, let's say that $q_1' = q_1 + u$ and $q_2' = q_2 - u$ for some $u \neq 0$. However, we argue that it must be the case that either $||q_1'||_* > \alpha_A$ or $||q_2'||_* > \alpha_P$. Without loss of generality, suppose $||q_1 + u||_* = \alpha_A - \epsilon$ for some $\epsilon \geq 0$.

Now, consider $v(\overline{\theta}) = v(q_1) = v(q_2)$. Corollary A.1 tells us that for any other q' where $||q'||_* \leq \alpha_A$,

$$\langle v(\overline{\theta}_*), q' \rangle < \langle v(\overline{\theta}_*), q_1 \rangle = \alpha_A.$$

Because the dual norm of $q_1 + u$ is still bounded by α_A, we have

$$\begin{aligned} \langle v(\overline{\theta}_*), q_1 + u \rangle &< \langle v(\overline{\theta}_*), q_1 \rangle \\ \langle v(\overline{\theta}_*), u \rangle &< 0. \end{aligned}$$

Then, we must have

$$||q_2'||_* = \langle v(q_2'), q_2 - u \rangle > \langle v(\overline{\theta}_*), q_2 - u \rangle = \alpha_P - \langle v(\overline{\theta}_*), u \rangle > \alpha_P,$$

giving us the needed contradiction.

Therefore, because there's only pair $(q_1, q_2) = (\alpha_A \overline{\theta}_*, \alpha_P \overline{\theta}_*)$ where $||q_1||_* \leq \alpha_A$, $||q_2||_* \leq \alpha_P$, and $q_1 + q_2 = (\alpha_A + \alpha_P)\overline{\theta}_*$, we must have

$$\begin{aligned} (g_1^i \square g_2^j)((\alpha_A + \alpha_P)\overline{\theta}_*) &= g_1^i(\alpha_A \overline{\theta}_*) + g_2^j(\alpha_P \overline{\theta}_*) \\ &= \langle \overline{\theta}_*, \alpha_A x_i^A + \alpha_P x_j^P \rangle \end{aligned}$$

□

Theorem 3.4. *Suppose the norm $||\cdot||$ is some p-norm where $p \neq 1$ and $p \neq \infty$. Fix θ where $||\theta||_* \leq \alpha_A + \alpha_P$. Then, for any $b \in [0, 1]$,*

$$(g_1^i \square g_2^j)(b\theta) \leq \left(\frac{b}{\alpha_A + \alpha_P}\right)\left(||\theta||_* \min(\alpha_A, \alpha_P)||x_i^A - x_j^P|| + \langle \theta, \alpha_A x_i^A + \alpha_P x_j^P \rangle\right) - \min(\alpha_A, \alpha_P)||x_i^A - x_j^P||$$

Proof. Because Lemma A.2 tells us that the infimal convolution of g_1^i and g_2^j is convex, we know that $(g_1^i \square g_2^j)(b\overline{\theta}_*)$ must be convex in b. By convexity, we have that for any $b, b' \in [0, \alpha_A + \alpha_P]$ and $t \in [0, 1]$

$$(g_1^i \square g_2^j)(((1-t)b + tb')\overline{\theta}_*) \leq (1-t)(g_1^i \square g_2^j)(b\overline{\theta}_*) + t(g_1^i \square g_2^j)(b'\overline{\theta}_*).$$

When we set $(b, b') = (0, \alpha_A + \alpha_P)$ and use the above upper bound, we get for any $t \in [0, 1]$

$$\begin{aligned} (g_1^i \square g_2^j)(t(\alpha_A + \alpha_P)\overline{\theta}_*) &\leq -(1-t)\min(\alpha_A, \alpha_P)||x_i^A - x_j^P|| + t\langle \overline{\theta}_*, \alpha_A x_i^A + \alpha_P x_j^P \rangle \\ &= t(\min(\alpha_A, \alpha_P)||x_i^A - x_j^P|| + \langle \overline{\theta}_*, \alpha_A x_i^A + \alpha_P x_j^P \rangle) - \min(\alpha_A, \alpha_P)||x_i^A - x_j^P|| \end{aligned}$$

due to Lemma A.3 and A.5.

In other words, given any θ where $||\theta||_* \leq \alpha_A + \alpha_P$, we can upper bound the infimal convolution as

$$\begin{aligned} (g_1^i \square g_2^j)(b\theta) &= (g_1^i \square g_2^j)(b||\theta||_* \overline{\theta}_*) \\ &= (g_1^i \square g_2^j)\left(\frac{b||\theta||_*}{\alpha_A + \alpha_P}(\alpha_A + \alpha_P)\overline{\theta}_*\right) \\ &\leq b\left(\frac{||\theta||_*}{\alpha_A + \alpha_P}\right)(\min(\alpha_A, \alpha_P)||x_i^A - x_j^P|| + \langle \overline{\theta}, \alpha_A x_i^A + \alpha_P x_j^P \rangle) - \min(\alpha_A, \alpha_P)||x_i^A - x_j^P|| \\ &= \left(\frac{b}{\alpha_A + \alpha_P}\right)\left(\min(\alpha_A, \alpha_P)||\theta||_*||x_i^A - x_j^P|| + \langle \theta, \alpha_A x_i^A + \alpha_P x_j^P \rangle\right) - \min(\alpha_A, \alpha_P)||x_i^A - x_j^P|| \end{aligned}$$

□

B Missing Details from Section 4

B.1 Missing Details from Section 4.2

Theorem 4.1. *With appropriately chosen step size η, Algorithm 1 returns $(\alpha_A, \alpha_P, \theta)$ such that $\Omega(\alpha_A, \alpha_P, \theta) \leq \Omega(\alpha_A^*, \alpha_P^*, \theta^*) + O\left(\frac{1}{\sqrt{T}}\right)$.*

Proof. Note that due to convergence rate of projected gradient descent, we have

$$\Omega^A(\overline{\alpha_A}, \overline{\alpha_P}, \overline{\theta}) \leq \Omega^A(\alpha'_A, \alpha'_P, \theta') + O\left(\frac{1}{\sqrt{T}}\right)$$
$$\Omega^P(\overline{\alpha_A}', \overline{\alpha_P}', \overline{\theta}') \leq \Omega^P(\alpha''_A, \alpha''_P, \theta'') + O\left(\frac{1}{\sqrt{T}}\right)$$

Also, we have

$$\Omega^A(\overline{\alpha_A}, \overline{\alpha_P}, \overline{\theta}) = \Omega(\overline{\alpha_A}, \overline{\alpha_P}, \overline{\theta})$$
$$\Omega^A(\alpha'_A, \alpha'_P, \theta') = \Omega(\alpha'_A, \alpha'_P, \theta')$$
$$\Omega^P(\overline{\alpha_A}', \overline{\alpha_P}', \overline{\theta}') = \Omega(\overline{\alpha_A}', \overline{\alpha_P}', \overline{\theta}')$$
$$\Omega^P(\alpha''_A, \alpha''_P, \theta'') = \Omega(\alpha''_A, \alpha''_P, \theta'')$$

Therefore, we must have

$$\Omega(\alpha_A, \alpha_P, \theta) \leq \Omega(\alpha_A^*, \alpha_P^*, \theta^*) + O\left(\frac{1}{\sqrt{T}}\right)$$

□

Here we try to give a characterization of the projection when $p = 2$. It is not immediate clear how to perform a projection onto C: given $\theta, \alpha_A, \alpha_P$, we need to find

$$\arg\min_{\theta', \alpha'_A, \alpha'_P \in C_1} ||(\theta, \alpha_A, \alpha_P) - (\theta', \alpha'_A, \alpha'_P)||_2^2 = \arg\min_{\theta', \alpha'_A, \alpha'_P \in C_1} ||\theta - \theta'||_2^2 + |\alpha_A - \alpha'_A|^2 + |\alpha_P - \alpha'_P|^2.$$

Suppose we are given $(\theta, \alpha_A, \alpha_P)$ such that $||\theta_1||_2 > \alpha_A + \alpha_P$ and/or $||\theta_2||_2 > \kappa_A \alpha_A$. The Lagrangian for the above optimization problem we are interested in is the following:

$$\begin{aligned}
&\mathcal{L}(\theta'_1, \theta'_2, \alpha'_A, \alpha'_P) \\
&= \frac{1}{2}\sum_i (\theta'_1[i] - \theta_1[i])^2 + \frac{1}{2}\sum_i (\theta'_2[i] - \theta_2[i])^2 + \frac{1}{2}(\alpha'_A - \alpha_A)^2 + \frac{1}{2}(\alpha'_P - \alpha_P)^2 \\
&+ \lambda_1((\sum_i (\theta'_1[i])^2)^{1/2} - \alpha'_A - \alpha'_P) + \lambda_2((\sum_i (\theta'_2[i])^2)^{1/2} - \kappa_A \alpha'_A) + \lambda_3(\alpha'_A - \alpha'_P).
\end{aligned}$$

The stationary part of the KKT condition requires that the gradient with respect to $\theta'_1, \theta'_2, \alpha'_A$ and α'_P is 0. In other words, we have

$$\begin{aligned}
\nabla_{\theta'_1[i]}\mathcal{L} &= (\theta'_1[i] - \theta_1[i]) + \frac{\lambda_1}{2}\frac{2\theta'_1[i]}{(\sum_i (\theta'_1[i])^2)^{1/2}} = (\theta'_1[i] - \theta_1[i]) + \frac{\lambda_1 \theta'_1[i]}{||\theta'_1||_2} = 0 \\
\nabla_{\theta'_2[i]}\mathcal{L} &= (\theta'_2[i] - \theta_2[i]) + \frac{\lambda_1}{2}\frac{2\theta'_2[i]}{(\sum_i (\theta'_2[i])^2)^{1/2}} = (\theta'_2[i] - \theta_2[i]) + \frac{\lambda_2 \theta'_1[i]}{||\theta'_2||_2} = 0 \\
\nabla_{\alpha'_A}\mathcal{L} &= \alpha'_A - \alpha_A - \lambda_1 - \lambda_2 \kappa_A + \lambda_3 = 0 \\
\nabla_{\alpha'_P}\mathcal{L} &= \alpha'_P - \alpha_P - \lambda_1 - \lambda_3 = 0
\end{aligned}$$

With some arranging, we get

$$\begin{aligned}
&\theta_1' + \frac{\lambda_1 \theta_1'}{||\theta_1'||} = \theta_1 \\
\implies &||\theta_1'||\bar{\theta_1'} + \lambda_1 \bar{\theta_1'} = \theta_1 \\
\implies &\bar{\theta_1'} = \frac{\theta_1}{||\theta_1'|| + \lambda_1} \\
\implies &\theta_1' = \frac{||\theta_1'||\theta_1}{||\theta_1'|| + \lambda_1} \\
\implies &||\theta_1'|| = \left|\left| \frac{||\theta_1'||\theta_1}{||\theta_1'|| + \lambda_1} \right|\right| \\
\implies &||\theta_1'|| = \frac{||\theta_1'||}{||\theta_1'|| + \lambda_1} ||\theta_1|| \\
\implies &||\theta_1'|| + \lambda_1 = ||\theta_1||
\end{aligned}$$

Similarly, we have

$$\begin{aligned}
&\theta_2' + \frac{\lambda_2 \theta_2'}{||\theta_2'||} = \theta_2 \\
\implies &||\theta_2'||\bar{\theta_2'} + \lambda_2 \bar{\theta_2'} = \theta_1 \\
\implies &\bar{\theta_2'} = \frac{\theta_2}{||\theta_2'|| + \lambda_2} \\
\implies &\theta_2' = \frac{||\theta_2'||\theta_2}{||\theta_2'|| + \lambda_2} \\
\implies &||\theta_2'|| = ||\frac{||\theta_2'||\theta_2}{||\theta_2'|| + \lambda_2}|| \\
\implies &||\theta_2'|| = \frac{||\theta_2'||}{||\theta_2'|| + \lambda_2} ||\theta_2|| \\
\implies &||\theta_2'|| + \lambda_2 = ||\theta_2||.
\end{aligned}$$

Note that θ_1' is simply a rescaling of θ_1:

$$\theta_1' = \frac{||\theta_1|| - \lambda_1}{||\theta_1||} \theta_1.$$

The complementary slack conditions require that

$$\lambda_1(||\theta_1'|| - \alpha_A' - \alpha_P') = 0.$$

In other words, either $\theta_1' = \theta_1$ or $||\theta_1'|| = \alpha_A' + \alpha_P'$. The same argument applies for θ_2': either $\theta_2' = \theta_2$ or $||\theta_2'|| = \kappa_A \alpha_A$. Now, we consider all four cases, and for each of those cases, we repeatedly consider the case where $\lambda_3 = 0$ and $\lambda_3 > 0$ (i.e. $\alpha_A' - \alpha_P' = 0$ from the complementary slack condition).

Case $\theta_1' = \theta_1$ and $\theta_2' = \theta_2$: In this case, we need only concern ourselves with how to set α_A' and α_P'. Because we have $\lambda_1, \lambda_2 = 0$,

$$\begin{aligned}
\alpha_A' - \alpha_A + \lambda_3 &= 0 \\
\lambda_3 &= \alpha_P' - \alpha_P.
\end{aligned}$$

The complementary slackness condition requires $\lambda_3(\alpha_A' - \alpha_P') = 0$. In other words, when $\lambda_3 = 0$, we have $(\alpha_A', \alpha_P') = (\alpha_A, \alpha_P)$. In other case where $\alpha_A' = \alpha_P'$, we have $(\alpha_A', \alpha_P') = (\frac{\alpha_A + \alpha_P}{2}, \frac{\alpha_A + \alpha_P}{2})$.

Case $\theta_1' = \theta_1$ and $||\theta_2'|| = \kappa_A \alpha_A'$: We have $\lambda_1 = 0$ and

$$\lambda_2 = ||\theta_2|| - ||\theta_2'|| = ||\theta_2|| - \kappa_A \alpha_A'$$

Plugging in $\lambda_1 = 0$, we have

$$\begin{aligned} \alpha_A' - \alpha_A - \lambda_2 \kappa_A + \lambda_3 &= 0 \\ \alpha_P' - \alpha_P - \lambda_3 &= 0. \end{aligned}$$

Substituting in λ_2 value, we get

$$\begin{aligned} & \alpha_A' - \alpha_A - \kappa_A(||\theta_2|| - \kappa_A \alpha_A') + \lambda_3 = 0 \\ \implies & \alpha_A'(1 + \kappa_A^2) = \alpha_A + \kappa_A ||\theta_2|| - \lambda_3 \\ \implies & \alpha_A' = \frac{\alpha_A + \kappa_A ||\theta_2|| - \lambda_3}{1 + \kappa_A^2} \end{aligned}$$

If $\lambda_3 = 0$, we have

$$\begin{aligned} \alpha_A' &= \frac{\alpha_A + \kappa_A ||\theta_2||}{1 + \kappa_A^2} \\ \alpha_P' &= \alpha_P. \end{aligned}$$

If $\lambda_3 \neq 0$ and hence $\alpha_A' = \alpha_P'$, then we have

$$\begin{aligned} & \alpha_A'(1 + \kappa_A^2) = \alpha_A + \kappa_A ||\theta_2|| - (\alpha_A' - \alpha_P) \\ \implies & \alpha_P' = \alpha_A' = \frac{\alpha_A + \alpha_P + \kappa_A ||\theta_2||}{2 + \kappa_A^2} \end{aligned}$$

Case $||\theta_1'|| = \alpha_A' + \alpha_P'$ and $\theta_2' = \theta_2$: We have that $\lambda_2 = 0$ and $\lambda_1 > 0$ and also

$$\lambda_1 = ||\theta_1|| - ||\theta_1'|| = ||\theta_1|| - (\alpha_A' + \alpha_P').$$

Plugging in $\lambda_2 = 0$, we have

$$\begin{aligned} \alpha_A' - \alpha_A - \lambda_1 + \lambda_3 &= 0 \\ \alpha_P' - \alpha_P - \lambda_1 - \lambda_3 &= 0 \end{aligned}$$

If $\lambda_3 = 0$, then

$$\begin{aligned} & \alpha_A' - \alpha_A - \lambda_1 = 0 \quad \text{and} \quad \alpha_P' - \alpha_P - \lambda_1 = 0 \\ \implies & \lambda_1 = \alpha_A' - \alpha_A = \alpha_P' - \alpha_P \end{aligned}$$

Substituting $\alpha_A' = \alpha_A + \alpha_P' - \alpha_P$ into $\alpha_P' - \alpha_P = \lambda_1 = ||\theta_1|| - (\alpha_A' + \alpha_P')$, we get

$$\begin{aligned} & \alpha_P' - \alpha_P = ||\theta_1|| - (\alpha_A + 2\alpha_P' - \alpha_P) \\ \implies & \alpha_P' - \alpha_P = ||\theta_1|| - \alpha_A - 2\alpha_P' + \alpha_P \\ \implies & 3\alpha_P' = ||\theta_1|| - \alpha_A + 2\alpha_P \\ \implies & \alpha_P' = \frac{||\theta_1|| - \alpha_A + 2\alpha_P}{3}. \end{aligned}$$

α'_A is then calculated as

$$\alpha'_A = \alpha_A + \frac{||\theta_1|| - \alpha_A + 2\alpha_P}{3} - \alpha_P.$$

If $\lambda_3 \neq 0$ and hence $\alpha'_A = \alpha'_P$, then

$$||\theta_1|| - 2\alpha'_A = \alpha'_A - \alpha_A + \lambda_3 = \alpha'_A - \alpha_P - \lambda_3 = \lambda_1$$

From the first equation, we get

$$\lambda_3 = ||\theta_1|| - 2\alpha'_A - (\alpha'_A - \alpha_A) = ||\theta_1|| - 3\alpha'_A + \alpha_A.$$

Plugging in this value for λ_3 into the second equation, we get

$$\begin{aligned} ||\theta_1|| - 2\alpha'_A &= \alpha'_A - \alpha_P - (||\theta_1|| - 3\alpha'_A + \alpha_A) \\ \implies -2\alpha'_A &= 4\alpha'_A - \alpha_P - \alpha_A - 2||\theta_1|| \\ \implies \frac{\alpha_A + \alpha_P + 2||\theta_1||}{6} &= \alpha'_A. \end{aligned}$$

Case $||\theta'_1|| = \alpha'_A + \alpha'_P$ and $||\theta'_2|| = \kappa_A \alpha'_A$:

$$\begin{aligned} \lambda_1 &= ||\theta_1|| - ||\theta'_1|| = ||\theta_1|| - (\alpha'_A + \alpha'_P) \\ \lambda_2 &= ||\theta_2|| - ||\theta'_2|| = ||\theta_2|| - \kappa_A \alpha'_A \end{aligned}$$

Putting these equations altogether with variables $\alpha'_A, \alpha'_P, \lambda_1, \lambda_2, \lambda_3$, we have

$$\begin{aligned} ||\theta_1|| &= \alpha'_A + \alpha'_P + \lambda_1 \\ ||\theta_2|| &= \kappa_A \alpha'_A + \lambda_2 \\ \alpha'_A - \alpha_A - \lambda_1 - \lambda_2 \kappa_A + \lambda_3 &= 0 \\ \alpha'_P - \alpha_P - \lambda_1 - \lambda_3 &= 0 \end{aligned}$$

We'll use the first equation to substitute in $\lambda_1 = ||\theta_1|| - \alpha'_A - \alpha'_P$ to get

$$\begin{aligned} ||\theta_2|| &= \kappa_A \alpha'_A + \lambda_2 \\ 2\alpha'_A - \alpha_A - ||\theta_1|| + \alpha'_P - \lambda_2 \kappa_A + \lambda_3 &= 0 \\ 2\alpha'_P - \alpha_P - ||\theta_1|| + \alpha'_A - \lambda_3 &= 0 \end{aligned}$$

Similarly, use the last equation to substitute in $\lambda_3 = 2\alpha'_P - \alpha_P - ||\theta_1|| + \alpha'_A$.

$$\begin{aligned} ||\theta_2|| &= \kappa_A \alpha'_A + \lambda_2 \\ 3\alpha'_A + 3\alpha'_P - \alpha_A - \alpha_P - 2||\theta_1|| - \lambda_2 \kappa_A &= 0 \end{aligned}$$

Finally, plug in $\lambda_2 = ||\theta_2|| - \kappa_A \alpha'_A$.

$$\begin{aligned} 3\alpha'_A + 3\alpha'_P - \alpha_A - \alpha_P - 2||\theta_1|| - \kappa_A(||\theta_2|| - \kappa_A \alpha'_A) &= 0 \\ \implies (3 + \kappa_A^2)\alpha'_A + 3\alpha'_P - \alpha_A - \alpha_P - 2||\theta_1|| - \kappa_A||\theta_2|| &= 0 \end{aligned}$$

As before, when $\lambda_3 = 0$, we get

$$\begin{aligned} 2\alpha'_P - \alpha_P - ||\theta_1|| + \alpha'_A &= 0 \\ \implies \alpha'_P &= \frac{||\theta_1|| + \alpha_P - \alpha'_A}{2}. \end{aligned}$$

Then, we get

$$
\begin{aligned}
&(3+\kappa_A^2)\alpha'_A + 3\left(\frac{||\theta_1|| + \alpha_P - \alpha'_A}{2}\right) - \alpha_A - \alpha_P - 2||\theta_1|| - \kappa_A||\theta_2|| = 0 \\
\implies &\left(\frac{3}{2}+\kappa_A^2\right)\alpha'_A = -3\left(\frac{||\theta_1|| + \alpha_P}{2}\right) + \alpha_A + \alpha_P + 2||\theta_1|| + \kappa_A||\theta_2|| \\
\implies &\alpha'_A = \frac{-3\left(\frac{||\theta_1||+\alpha_P}{2}\right) + \alpha_A + \alpha_P + 2||\theta_1|| + \kappa_A||\theta_2||}{\frac{3}{2}+\kappa_A^2} \\
\implies &\alpha'_A = \frac{-3\left(\frac{||\theta_1||+\alpha_P}{2}\right) + \alpha_A + \alpha_P + 2||\theta_1|| + \kappa_A||\theta_2||}{\frac{3}{2}+\kappa_A^2} \\
\implies &\alpha'_A = \frac{2\alpha_A + \alpha_P + ||\theta_1|| + 2\kappa_A||\theta_2||}{3+2\kappa_A^2}.
\end{aligned}
$$

Consequently, we have

$$
\alpha'_P = \frac{||\theta_1|| + \alpha_P}{2} - \left(\frac{2\alpha_A + \alpha_P + ||\theta_1|| + 2\kappa_A||\theta_2||}{6+4\kappa_A^2}\right).
$$

Otherwise, when $\lambda_3 > 0$, we have $\alpha'_A = \alpha'_P$. In this case, we get

$$
\begin{aligned}
(6+\kappa_A^2)\alpha'_A &= \alpha_A + \alpha_P + 2||\theta_1|| + \kappa_A||\theta_2|| \\
\implies \alpha'_A = \alpha'_P &= \frac{\alpha_A + \alpha_P + 2||\theta_1|| + \kappa_A||\theta_2||}{6+\kappa_A^2}.
\end{aligned}
$$

We summarize the results in the following tables:

Cases	$\lambda_3 = 0$
$(\theta'_1, \theta'_2) = (\theta_1, \theta_2)$	$(\alpha'_A, \alpha'_P) = (\alpha_A, \alpha_P)$
$(\theta'_1, \theta'_2) = (\theta_1, \kappa_A\alpha'_A\overline{\theta}_2)$	$(\alpha'_A, \alpha'_P) = (\frac{\alpha_A+\kappa_A\|\|\theta_2\|\|}{1+\kappa_A^2}, \alpha_P)$
$(\theta'_1, \theta'_2) = ((\alpha'_A+\alpha'_P)\overline{\theta}_1, \theta_2)$	$(\alpha'_A, \alpha'_P) = (\alpha_A + \alpha'_P - \alpha_P, \frac{\|\|\theta_1\|\|-\alpha_A+2\alpha_P}{3})$
$(\theta'_1, \theta'_2) = ((\alpha'_A+\alpha'_P)\overline{\theta}_1, \kappa_A\alpha'_A\overline{\theta}_2)$	$(\alpha'_A, \alpha'_P) = \left(\frac{2\alpha_A+\alpha_P+\|\|\theta_1\|\|+2\kappa_A\|\|\theta_2\|\|}{3+2\kappa_A^2}, \frac{\|\|\theta_1\|\|+\alpha_P}{2} - \left(\frac{2\alpha_A+\alpha_P+\|\|\theta_1\|\|+2\kappa_A\|\|\theta_2\|\|}{6+4\kappa_A^2}\right)\right)$

Cases	$\lambda_3 > 0$
$(\theta'_1, \theta'_2) = (\theta_1, \theta_2)$	$\alpha'_A = \alpha'_P = \frac{\alpha_A+\alpha_P}{2}$
$(\theta'_1, \theta'_2) = (\theta_1, \kappa_A\alpha'_A\overline{\theta}_2)$	$\alpha'_A = \alpha'_P = \frac{\alpha_A+\alpha_P+\kappa_A\|\|\theta_2\|\|}{2+\kappa_A^2}$
$(\theta'_1, \theta'_2) = ((\alpha'_A+\alpha'_P)\overline{\theta}_1, \theta_2)$	$\alpha'_A = \alpha'_P = \frac{\alpha_A+\alpha_P+2\|\|\theta_1\|\|}{6}$
$(\theta'_1, \theta'_2) = ((\alpha'_A+\alpha'_P)\overline{\theta}_1, \kappa_A\alpha'_A\overline{\theta}_2)$	$\alpha'_A = \alpha'_P = \frac{\alpha_A+\alpha_P+2\|\|\theta_1\|\|+\kappa_A\|\|\theta_2\|\|}{6+\kappa_A^2}$

C Missing Details from Section 5

Lemma 5.1. *Fix any θ, (x_i^A, a, x_j^P), and $(\alpha_A, \alpha_P, \kappa_A)$. If $||\theta[1:m_1]||_{p,*} > \alpha_A + \alpha_P$, then $\sup_x h(\theta,(x,a)) - \alpha_A||x_i^A - x||_p - \alpha_P||x_j^P - x||_p = \infty$. Otherwise, we have*

$$
\begin{aligned}
&\sup_x h(\theta,(x,a)) - \alpha_A||x_i^A - x||_p - \alpha_P||x_j^P - x||_p \\
&= \sup_{b\in[0,1]} -f^*(b) + (g_1^i \square g_2^j)(-b\theta[1:m_1]) + \langle b\theta[m_1+1:m_1+m_2], a\rangle
\end{aligned}
$$

where g_1^i and g_2^j is the same as defined in Lemma 3.1.

Proof. Noting that h is convex and thus h is equal to its biconjugate h^{**}, we have

$$\begin{aligned}
&\sup_x h(\theta,(x,a)) - \alpha_A||x_i^A - x||_p - \alpha_P||x_j^P - x||_p \\
&= \sup_x \sup_{b\in[0,1]} \langle b\theta,(x,a)\rangle - f^*(b) - \alpha_A||x_i^A - x||_p - \alpha_P||x_j^P - x||_p \\
&= \sup_{b\in[0,1]} \sup_x \langle b\theta,(x,a)\rangle - f^*(b) - \sup_{||q_1||_{p,*}\le\alpha_A} \langle q_1, x_i^A - x\rangle - \sup_{||q_2||_{p,*}\le\alpha_P} \langle q_2, x_j^P - x\rangle \\
&= \sup_{b\in[0,1]} \sup_x \langle b\theta,(x,a)\rangle - f^*(b) - \sup_{||q_1||_{p,*}\le\alpha_A} \langle (q_1,\xi),(x_i^A,a)-(x,a)\rangle - \sup_{||q_2||_{p,*}\le\alpha_P} \langle (q_2,0),(x_j^P,a)-(x,a)\rangle \\
&= \sup_{b\in[0,1]} \sup_x \inf_{\substack{||q_1||_{p,*}\le\alpha_A,\\ ||q_2||_{p,*}\le\alpha_P}} \langle b\theta,(x,a)\rangle - f^*(b) - \langle (q_1,\xi),(x_i^A,a)-(x,a)\rangle - \langle (q_2,0),(x_j^P,a)-(x,a)\rangle \\
&= \sup_{b\in[0,1]} \sup_x \inf_{\substack{||q_1||_{p,*}\le\alpha_A,\\ ||q_2||_{p,*}\le\alpha_P}} \langle b\theta + (q_1,\xi) + (q_2,0),(x,a)\rangle - f^*(b) - \langle q_1, x_i^A\rangle - \langle q_2, x_j^P\rangle
\end{aligned}$$

where ξ can be chosen arbitrarily.

We appeal to proposition 5.5.4 of Bertsekas to swap inf and sup:

$$= \sup_{b\in[0,1]} \inf_{\substack{||q_1||_{p,*}\le\alpha_A,\\ ||q_2||_{p,*}\le\alpha_P}} \sup_x \langle b\theta + (q_1,\xi) + (q_2,0),(x,a)\rangle - f^*(b) - \langle q_1, x_i^A\rangle - \langle q_2, x_j^P\rangle.$$

Note that unless $b\theta + (q_1,\xi) + (q_2,0) = 0$, x can be chosen arbitrarily big. Also, if $\theta + (q_1,\xi) + (q_2,0) \neq 0$, then b can be chosen to be 1. Therefore, if there doesn't exist (q_1,q_2) such that $\theta + (q_1,\xi) + (q_2,0) = 0$, everything evaluates to ∞. In other words, the expression evaluates to ∞ unless $||\theta[1:m_1]||_{p,*} \le \alpha_A + \alpha_P$ and $\xi = \theta[m_1+1 : m_1+m_2]$.

Now, suppose θ satisfies the above condition as we know it evaluates to ∞ otherwise. Then, we get

$$\begin{aligned}
&= \sup_{b\in[0,1]} -f^*(b) + \langle b\theta[m_1+1:m_1+m_2], a\rangle + \inf_{\substack{||q_1||_{p,*}\le\alpha_A,\\ ||q_2||_{p,*}\le\alpha_P}} \begin{cases} -\langle q_1, x_i^A\rangle - \langle q_2, x_j^P\rangle & \text{if } b\theta[1:m_1] + q_1 + q_2 = 0 \\ \infty & \text{otherwise} \end{cases} \\
&= \sup_{b\in[0,1]} -f^*(b) + (g_1^i \square g_2^j)(-b\theta[1:m_1]) + \langle b\theta[m_1+1:m_1+m_2], a\rangle.
\end{aligned}$$

□

Theorem 5.1. *Fix any θ, (x_i^A, a, x_j^P), and $(\alpha_A, \alpha_P, \kappa_A)$. If $||\theta_1||_{p,*} > \alpha_A + \alpha_P$, then $\sup_x h(\theta,(x,a)) - \alpha_A||x_i^A - x||_p - \alpha_P||x_j^P - x||_p = \infty$ Otherwise, we have*

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

Proof. Because f is a convex function, its biconjugate is itself, so

$$\sup_{b\in[0,1]} -f^*(b) + b\cdot X = f(X).$$

Therefore, we have

$$
\begin{aligned}
&\sup_x h(\theta,(x,a)) - \alpha_A||x - x_i^A||_p - \alpha_P||x - x_j^P||_p \\
&= \sup_{b\in[0,1]} -f^*(b) + (g_1^i \square g_2^j)(b\theta_1) + \langle b\theta_2, a\rangle \\
&\leq \sup_{b\in[0,1]} -f^*(b) + \left(\frac{b}{\alpha_A+\alpha_P}\right)\left(\min(\alpha_A,\alpha_P)||\theta_1||_*||x_i^A - x_j^P|| + \langle\theta_1, \alpha_A x_i^A + \alpha_P x_j^P\rangle\right) - \min(\alpha_A,\alpha_P)||x_i^A - x_j^P||_p + b\langle\theta_2, a\rangle \\
&= \sup_{b\in[0,1]} -f^*(b) \\
&+ b\left(\left(\frac{1}{\alpha_A+\alpha_P}\right)\left(\min(\alpha_A,\alpha_P)||\theta_1||_*||x_i^A - x_j^P|| + \langle\theta_1, \alpha_A x_i^A + \alpha_P x_j^P\rangle\right) + \langle\theta_2, a\rangle\right) - \min(\alpha_A,\alpha_P)||x_i^A - x_j^P||_p \\
&= f\left(\left(\frac{\min(\alpha_A,\alpha_P)||\theta_1||_*||x_i^A - x_j^P|| + \langle\theta, \alpha_A x_i^A + \alpha_P x_j^P\rangle}{\alpha_A+\alpha_P}\right) + \langle\theta_2, a\rangle\right) - \min(\alpha_A,\alpha_P)||x_i^A - x_j^P||_p.
\end{aligned}
$$

The first inequality follows from Theorem 3.4. □

Theorem 5.2. *Suppose $p \neq 1$ and $p \neq \infty$. If $||\theta_1||_{p,*} \leq \alpha_A + \alpha_P$, then*

$$
\begin{aligned}
&\max_{a,y}\sup_{x\in\mathcal{X}}\left(c(a,y)\cdot\ell(\theta,(x,a,y)) - \alpha_A d_A^i(x,a) - \alpha_P d_P^j(x,y)\right) \\
&- \max_{a,y}\left(c(a,y)\cdot\ell(\theta,(\hat{x}_{i,j},a,y)) + \alpha_A\kappa_A|a_i^A - a| + \alpha_P\kappa_P|y_j^P - y| - \min(\alpha_A,\alpha_P)||x_i^A - x_j^P||\right) \\
&\leq 4\hat{\alpha}||x_i^A - x_j^P||.
\end{aligned}
$$

Proof. First, fix any (a,y). Using the same argument as in Theorem 3.6,

$$
\begin{aligned}
&\left(\sup_x \ell(\theta,(x,a,y)) - \frac{\alpha_A}{c(a,y)}||x_i^A - x||_p - \frac{\alpha_P}{c(a,y)}||x_j^P - x||_p\right) - \left(\ell(\theta,(\hat{x}_{i,j},a,y)) - \frac{\hat{\alpha}}{c(a,y)}||x_i^A - x_j^P||\right) \\
&\leq \left(\ell(\theta,(\check{x},a,y)) - \frac{\hat{\alpha}}{c(a,y)}||x_i^A - x_j^P||\right) - \left(\ell(\theta,(\hat{x}_{i,j},a,y)) - \frac{\hat{\alpha}}{c(a,y)}||x_i^A - x_j^P||\right) \\
&\leq 2||x_i^A - x_j^P||
\end{aligned}
$$

where $\check{x}$ is the same as in the proof of Theorem 3.6. Multiplying by $c(a,y)$, we have

$$
\begin{aligned}
&\sup_x\left(c(a,y)\cdot\ell(\theta,(x,a,y)) - \alpha_A||x_i^A - x||_p - \alpha_P||x_j^P - x||_p\right) - \left(c(a,y)\cdot\ell(\theta,(\hat{x}_{i,j},a,y)) - \alpha_A||x_i^A - \hat{x}_{i,j}||_p - \alpha_P||x_j^P - \hat{x}_{i,j}||_p\right) \\
&\leq 2c(a,y)||x_i^A - x_j^P|| \leq 4||x_i^A - x_j^P||.
\end{aligned}
$$

because $c(a,y) \leq 2$ for any (a,y).

Finally, write

$$
(x^*,a^*,y^*) = \arg\max_{x,a,y}\left(c(a,y)\cdot\ell(\theta,(x,a,y)) - \alpha_A d_A^i(x,a) - \alpha_P d_P^j(x,y)\right).
$$

Then, we have

$$
\begin{aligned}
&\Big(c(a^*,y^*)\cdot\ell(\theta,(x^*,a^*,y^*)) - \alpha_A d_A^i(x^*,a^*) - \alpha_P d_P^j(x^*,y^*)\Big)\\
&-\max_{(a,y)}\big(c(a,y)\cdot\ell(\theta,(\hat{x}_{i,j},a,y)) + \alpha_A\kappa_A|a_i^A - a| + \alpha_P\kappa_P|y_j^P - y| - \min(\alpha_A,\alpha_P)||x_i^A - x_j^P||\big)\\
&\le \Big(c(a^*,y^*)\cdot\ell(\theta,(x^*,a^*,y^*)) - \alpha_A d_A^i(x^*,a^*) - \alpha_P d_P^j(x^*,y^*)\Big)\\
&-\big(c(a^*,y^*)\cdot\ell(\theta,(\hat{x}_{i,j},a^*,y^*)) + \alpha_A\kappa_A|a_i^A - a^*| + \alpha_P\kappa_P|y_j^P - y^*| - \min(\alpha_A,\alpha_P)||x_i^A - x_j^P||\big)\\
&= \big(c(a^*,y^*)\cdot\ell(\theta,(x^*,a^*,y^*)) - \alpha_A||x_i^A - x^*||_p - \alpha_P||x_j^P - x^*||_p\big)\\
&-\big(c(a^*,y^*)\cdot\ell(\theta,(\hat{x}_{i,j},a^*,y^*)) - \alpha_A||x_i^A - \hat{x}_{i,j}||_p - \alpha_P||x_j^P - \hat{x}_{i,j}||_p\big)\\
&\le 4||x_i^A - x_j^P||_p
\end{aligned}
$$

where the first inequality follows because $-\max_{a,y}$ term cannot be greater than when the inner term is evaluated at (a^*,y^*), and the last inequality follows because for (a^*,y^*), $\max_x$ is achieved at x^*. □

D Missing Details from Section 6

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:

	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	5	5	10	5	10	5
k	1	1	1	1	1	1

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

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.