What is the Right Notion of Distance between Predict-then-Optimize Tasks?

Paula Rodriguez-Diaz[1] **Lingkai Kong**[1] **Kai Wang**[2] **David Alvarez-Melis**[1,3,4] **Milind Tambe**[1]

[1]Harvard University
[2]Georgia Institute of Technology
[3]Kempner Institute
[4]Microsoft Research

Abstract

Comparing datasets is a fundamental task in machine learning, essential for various learning paradigms—from evaluating train and test datasets for model generalization to using dataset similarity for detecting data drift. While traditional notions of dataset distances offer principled measures of similarity, their utility has largely been assessed through prediction error minimization. However, in Predict-then-Optimize (PtO) frameworks, where predictions serve as inputs for downstream optimization tasks, model performance is measured through decision regret rather than prediction error. In this work, we propose OTD^3 (Optimal Transport Decision-aware Dataset Distance), a novel dataset distance that incorporates downstream decisions in addition to features and labels. We show that traditional feature-label distances lack informativeness in PtO settings, while OTD^3 more effectively captures adaptation success. We also derive a PtO-specific adaptation bound based on this distance. Empirically, we show that our proposed distance accurately predicts model transferability across three different PtO tasks from the literature. Code is available at https://github.com/paularodr/OTD3

1 INTRODUCTION

Comparing datasets is a fundamental task in machine learning and a crucial component of various downstream tasks. Understanding the *similarity* (or *dissimilarity*) of datasets can inform decisions in transfer learning [Tran et al., 2019, Ben-David et al., 2010], multitask learning [Janati et al., 2019, Shui et al., 2019], and data valuation [Just et al., 2023, Jiang et al., 2023], among other applications. For example, selecting a pre-training dataset that is similar to a data-poor target domain can lead to better fine-tuning performance. Notions of *dataset distance* have emerged as a principled way of quantifying these similarities and differences [Mercioni and Holban, 2019, Janati et al., 2019, Alvarez-Melis and Fusi, 2020]. Such distances provide insights into the relation and correspondence between data distributions, help in evaluating model performance, and guide the selection of appropriate learning algorithms.

The concept of *dataset* can vary based on context and objectives. In classical statistics, it generally refers to feature vectors, focusing on the distribution and relationships within a feature space $\mathcal{X}$. Classic distributional distances offer formal measures of dataset similarity: the Total Variation distance [Verdú, 2014] quantifies the maximum discrepancy between distributions; Wasserstein distance, or Earth Mover's Distance, measures the cost of transforming one distribution into another [Villani, 2008]; and Integral Probability Metrics (IPM) measure how well a class of classifiers can distinguish samples from the two distributions [Müller, 1997].

In supervised learning, datasets include both features from space $\mathcal{X}$ and labels from space $\mathcal{Y}$. The distance between two such datasets involves measuring both the feature and label differences. This can be challenging when the label space $\mathcal{Y}$ is not a metric space. Approaches such as those proposed by Courty et al. [2014], Alvarez-Melis et al. [2018], and Alvarez-Melis and Fusi [2020] offer a principled method for computing dataset distances considering the joint feature-label distribution $\mathcal{P}(\mathcal{X} \times \mathcal{Y})$. These methods ensure that both the features and labels are adequately accounted for in the distance measure, offering a more holistic comparison between datasets.

However, the Predict-then-Optimize (PtO) framework introduces a unique challenge by using machine learning predictions as inputs for a downstream optimization problem, shifting the focus from minimizing prediction error to minimizing decision regret [Donti et al., 2017, Elmachtoub and Grigas, 2022, Wilder et al., 2019, Mandi et al., 2023]. This results in PtO tasks involving not just a feature-label dataset,

but also a decision space Ω of optimization solutions, creating a feature-label-decision dataset with samples in $\mathcal{X} \times \mathcal{Y} \times \Omega$. The decision space Ω may not be a metric space; for example, decisions related to the solution of a top-k problem do not necessarily form a metric space. Moreover, decisions might need to be evaluated under various criteria, such as minimizing travel distance or maximizing safety. Even if Ω were a metric space, it is uncertain whether its associated distance would be meaningful for assessing the adaptability of a PtO task across different domains. This complexity underscores the need for new distance measures that incorporate decisions to accurately capture the nature of PtO tasks.

In this work we introduce OTD^3, a decision-aware dataset distance based on Optimal Transport (OT) techniques [Villani, 2008] that incorporates features, labels, and decisions. OTD^3 is the first distance metric designed to account for downstream decisions, directly addressing the unique challenges posed by PtO tasks. We evaluate its utility as a learning-free criterion for assessing the transferability of models trained within a *decision-focused learning* (DFL) framework under distribution shift. Within this framework, models are commonly trained using surrogate loss functions that aim to minimize decision regret [Wilder et al., 2019, Mandi et al., 2023]. In the context of domain adaptation in PtO, where we measure performance through decision regret, we derive a generalization bound that highlights the importance of considering features, labels, and decisions jointly. Our empirical analysis spans three PtO tasks from the literature—Linear Model Top-K, Warcraft Shortest Path, and Inventory Stock Problem—demonstrating that our decision-aware distance better predicts transfer performance compared to feature-label distances alone.

In summary, we make the following contributions:

- We introduce a decision-aware dataset distance
- We derive an adaptation bound for PtO tasks in terms of this distance
- We empirically validate our approach on three PtO tasks

2 RELATED WORK

Dataset Distances via Optimal Transport Optimal Transport (OT)-based distances have gained traction as an effective method for comparing datasets. These methods characterize datasets as empirical probability distributions supported in finite samples, and require a cost function between pairs of samples to be provided as an input. Most OT-based dataset distance approaches define this cost function solely in terms of the features of the data, either directly or in a latent embedding space. For example, Muzellec and Cuturi [2018] proposed representing objects as elliptical distributions and scaling these computations, while Frogner et al. [2019] extended this to discrete measures. Delon and Desolneux [2020] introduced a Wasserstein-type distance for Gaussian mixture models. These approaches are useful mostly in unsupervised learning settings since they do not take into account labels or classes associated with data points. To address this limitation, a different line of work has proposed extensions of OT amenable to supervised or semi-supervised learning settings that explicitly incorporate label information in the cost function. Courty et al. [2014] used group-norm penalties to guide OT towards class-coherent matches while Alvarez-Melis et al. [2018] employed submodular cost functions to integrate label information into the OT objective. For discrete labels, Alvarez-Melis and Fusi [2020] proposed using a hierarchical OT approach to compute label-to-label distances as distances between the conditional distributions of features defined by the labels.

Predict-then-Optimize (PtO) The PtO framework has seen significant advancements in integrating machine learning with downstream optimization. The frameworks proposed by Amos and Kolter [2017], Donti et al. [2017], Wilder et al. [2019] and Elmachtoub and Grigas [2022] have been instrumental in this integration. Subsequent work has focused on differentiating through the parameters of optimization problems with various structures, including learning appropriate loss functions [Wang et al., 2020, Shah et al., 2022, 2023, Bansal et al., 2023] and handling nonlinear objectives [Qi et al., 2023, Elmachtoub et al., 2025]. Recent efforts have addressed data-centric challenges within PtO, including including worst-case distribution shifts [Ren et al., 2024], robustness to adversarial label drift [Johnson-Yu et al., 2023] and active learning for data acquisition [Liu et al., 2025]. While these works propose task-specific learning algorithms, they all share a common underlying principle: dataset similarity. In distribution shifts and label drift, the key challenge lies in the (dis)similarity between training and test datasets, whereas in data acquisition, it concerns the (dis)similarity between the training dataset and the acquisition source.

3 BACKGROUND

3.1 OPTIMAL TRANSPORT

OT theory provides an elegant and powerful mathematical framework for measuring the distance between probability distributions by characterizing similarity in terms of correspondence and transfer [Villani, 2008, Kantorovitch, 1942]. In a nutshell, OT addresses the problem of transferring probability mass from one distribution to another while minimizing a cost function associated with the transportation.

Formally, given two probability distributions α and β defined on measurable spaces $\mathcal{X}$ and $\mathcal{Y}$, respectively, the OT problem seeks a transport plan π (defined as a coupling between α and β) that minimizes the total transportation cost. According to the Kantorovich formulation [Kantorovitch, 1942], for any coupling π, the transport cost between α and

β with respect to π is defined as:

$$d_T(\alpha, \beta; \pi) := \int_{\mathcal{X}\times\mathcal{Y}} c(x, y)\, \mathrm{d}\pi(x, y), \tag{1}$$

where $c(x, y)$ is the cost function representing the cost of transporting mass from point $x \in \mathcal{X}$ to point $y \in \mathcal{Y}$. The transport cost $d_T(\alpha, \beta; \pi)$ defines a distance, known as the *transport distance* with respect to π, between α and β. The OT problem then minimizes the transport cost over all possible couplings between α and β, defining the *optimal transport distance* as:

$$d_{OT}(\alpha, \beta; c) := \min_{\pi\in\Pi(\alpha,\beta)} \int_{\mathcal{X}\times\mathcal{Y}} c(x, y)\, \mathrm{d}\pi(x, y), \tag{2}$$

where $\Pi(\alpha, \beta)$ denotes the set of all possible couplings (transport plans) that have α and β as their marginals. This formulation finds the optimal way to transform one distribution into another by minimizing the total transportation cost.

3.2 OPTIMAL TRANSPORT DATASET DISTANCE

In supervised machine learning, datasets can be represented as empirical joint distributions over a feature-label space $\mathcal{X} \times \mathcal{Y}$. OT distances can be used to measure the similarity between these empirical distributions, thus providing a principled way to compare datasets. Given two datasets $\mathcal{D}$ and $\mathcal{D}'$ consisting of feature-label tuples (x, y) and (x', y'), respectively, the challenge of defining a transport distance between $\mathcal{D}$ and $\mathcal{D}'$ lies in the challenge of defining an appropriate cost function between (x, y) and (x', y') pairs. A straightforward way to define the feature-label pairwise cost is via the individual metrics in $\mathcal{X}$ and $\mathcal{Y}$ if available. If $d_\mathcal{X}$ and $d_\mathcal{Y}$ are metrics on $\mathcal{X}$ and $\mathcal{Y}$, respectively, the cost function can be defined as:

$$c_{\mathcal{X}\mathcal{Y}}((x, y), (x', y')) = \left(d_\mathcal{X}(x, x')^p + d_\mathcal{Y}(y, y')^p\right)^{1/p} \tag{3}$$

for $p \geq 1$. This point-wise cost function defines a valid metric on $\mathcal{X} \times \mathcal{Y}$. However, it is uncommon for $d_\mathcal{Y}$ to be readily available. To address this, Courty et al. [2017] propose replacing $d_\mathcal{Y}(y, y')$ with a loss function $\mathcal{L}(y, y')$ that measures the discrepancy between y and y' while Alvarez-Melis and Fusi [2020] suggest using a p-Wasserstein distance between the conditional distributions of features defined by y and y' as an alternative to $d_\mathcal{Y}(y, y')$. The latter is known as the *Optimal Transport Dataset Distance* (OTDD). We also use this term when referring to the dataset distance $d_{OT}(\mathcal{D}, \mathcal{D}'; c_{\mathcal{X}\mathcal{Y}})$.

3.3 PREDICT-THEN-OPTIMIZE

The Predict-then-Optimize (PtO) framework involves two sequential steps: prediction and optimization. First, a predictive model f is used to predict costs based on some features $x_1, \ldots, x_N \in \mathcal{X}$, represented as $\hat{\boldsymbol{y}} = [\hat{y}_1, \ldots, \hat{y}_N] = [f(x_1), \ldots, f(x_N)]$. Second, an optimization model uses these predicted costs $\hat{\boldsymbol{y}}$ as the objective function costs:

$$M(\hat{\boldsymbol{y}}) := \text{argmax}_w\, g(w; \hat{\boldsymbol{y}}), \quad s.t.\ \ w \in \Omega, \tag{4}$$

where Ω is the space of feasible solutions. We assume that $w^*_M : \mathbb{R}^d \to \Omega$ acts as an oracle for solving this optimization problem, such that $w^*_M(\hat{\boldsymbol{y}})$ represents the optimal solution for $M(\hat{\boldsymbol{y}})$. However, the solution $w^*_M(\hat{\boldsymbol{y}})$ is optimal for $M(\hat{\boldsymbol{y}})$ but might not be optimal for $M(\boldsymbol{y})$, where $\boldsymbol{y}$ represents the true costs.

Given a hypothesis function $f : \mathcal{X} \to \mathcal{Y}$, we measure its performance on the optimization problem $M(\boldsymbol{y})$ using the predicted cost vector $\hat{\boldsymbol{y}} = [f(x_1), \ldots, f(x_N)]$ and the true cost vector $\boldsymbol{y} = [y_1, \ldots, y_N]$. This is quantified as the *decision quality* $q(\hat{\mathbf{y}}, \boldsymbol{y}) := g(w^*_M(\hat{\mathbf{y}}); \boldsymbol{y})$, reflecting the quality of decisions made using $w^*_M(\hat{\mathbf{y}})$ as a solution to $M(\boldsymbol{y})$. The *decision quality regret*, which evaluates the performance of f, is defined as:

$$q_{\text{reg}}(\hat{\boldsymbol{y}}, \boldsymbol{y}) = |q(\boldsymbol{y}, \boldsymbol{y}) - q(\hat{\boldsymbol{y}}, \boldsymbol{y})|. \tag{5}$$

The goal of decision-focused learningWilder et al. [2019] in a PtO task is to learn a predictive model f_θ that minimizes the decision quality regret, ensuring that the decisions derived from the predictions are as close to optimal as possible.

4 MOTIVATING EXAMPLE

To illustrate the role of decisions in PtO task comparisons, we look at correspondence between task similarity and zero-shot transfer performance in a simple PtO task: the Linear Model Top-K setting from Shah et al. [2022]. This task consists of two stages:

Predict: Given a resource's feature $x_n \sim \mathcal{P}_\mathcal{X}$, where $\mathcal{P}_\mathcal{X} = \text{Unif}[-1, 1]$, a linear model predicts its utility $\hat{y}_n$, where the true utility follows $y_n = p(x_n)$, a cubic polynomial. Predictions for N resources form $\hat{\boldsymbol{y}} = [\hat{y}_1, \ldots, \hat{y}_N]$.

Optimize: Select the top $K = 1$ resource by solving $M(\hat{\boldsymbol{y}}) = \max_{\boldsymbol{z}\in[0,1]^N}\{\boldsymbol{z} \cdot \sigma_x(\hat{\boldsymbol{y}})\}$ such that $||\boldsymbol{z}||_0 = K$, where σ_x orders $\hat{\boldsymbol{y}}$ in ascending order of $\boldsymbol{x} = [x_1, \ldots, x_N]$.

We analyze this task under target shifts—where label distributions change while feature distributions remain constant—parametrized by γ, where the shifted utility function is given by $p_\gamma(x) = 10(x^3 - \gamma x)$. We define two source domains, A and B, with shifts $\gamma = 0$ and $\gamma = 1.2$, respectively. The target domain C is characterized by $\gamma = 0.65$.[1]

Assume we have only a few instance $\mathcal{D}_C = \{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^n \sim C$, but want to learn how to perform Top-K selection in domain C when the true costs $\boldsymbol{y}$

[1]We choose $\gamma = 0.65$ for consistency with Shah et al. [2022].

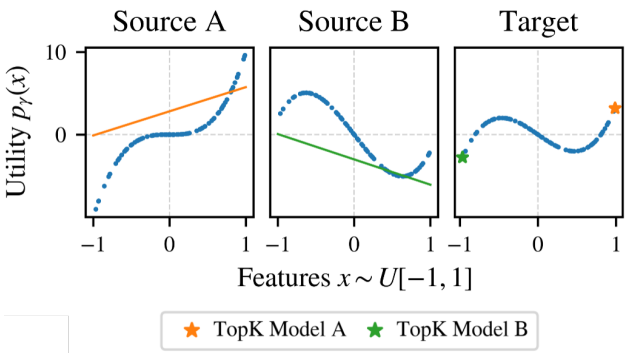


Figure 1: Linear Model Top-K instances under target shift

are unknown. This limited data is insufficient for directly learning weights in this domain. However, we have access to datasets $\mathcal{D}_A$ and $\mathcal{D}_B$ drawn form source domains A and B, respectively, allowing us to learn decision-focused weights θ_A and θ_B.[2] The key question is: which weights, θ_A or θ_B, should be used for the PtO task in $\mathcal{D}_C$?

A natural approach is to use the OTDD to identify the source dataset closest to the target. However, in this case, the computed distances $d_{OT}(\mathcal{D}_A, \mathcal{D}_C; c_{\mathcal{XY}})$ and $d_{OT}(\mathcal{D}_B, \mathcal{D}_C; c_{\mathcal{XY}})$ are equal, suggesting no preference between θ_A and θ_B. Yet, in practice, θ_A yields zero regret on $\mathcal{D}_C$, while θ_B results in a regret close to 4, making θ_A the clear choice for the PtO task on $\mathcal{D}_C$ (see Appendix Fig. 8 for details). Figure 1 illustrates this discrepancy: the model with θ_A (orange) successfully selects the correct Top-K resource in $\mathcal{D}_C$, while the model with θ_B (green) fails to do so. Since regret differs significantly, dataset distances should reflect that $\mathcal{D}_C$ is closer to $\mathcal{D}_A$ than $\mathcal{D}_B$. We argue that feature-label distances alone are insufficient, and incorporating decision components is necessary for distances to accurately reflect similarities, and hence adaptability, in PtO tasks.

5 DECISION-AWARE DATASET DISTANCE

A dataset for a PtO task with downstream optimization $M(\cdot)$ and oracle w^*_M consists of feature-label-decision triplets $(x, y, z) \in \mathcal{X} \times \mathcal{Y} \times \Omega$, where the decision z is the precomputed optimal solution to the optimization task parametrized by the true label y, i.e., $z = M(y)$. Our objective is to formalize a notion of similarity between PtO tasks by defining a distance $d(\mathcal{D}, \mathcal{D}')$ between datasets $\mathcal{D} = \{(x, y, M(y))\}_{(x,y)\sim\mathcal{P}}$ and $\mathcal{D}' = \{(x, y, M'(y))\}_{(x,y)\sim\mathcal{P}'}$ for any distributions $\mathcal{P}$ and $\mathcal{P}'$ over the joint feature-label space $\mathcal{X} \times \mathcal{Y}$ and optimization problems M and M' parametrized by $\mathcal{Y}$.

OT provides a natural framework for comparing datasets by leveraging the geometry of the underlying space and establishing correspondences between distributions. It has been used as the foundation for OTDD, a dataset distance defined over features and labels [Alvarez-Melis and Fusi, 2020]. We extend this idea to PtO tasks, where dataset distances must also account for differences in decisions arising from the downstream optimization process. Unlike standard settings where similarity is assessed based only on feature-label distributions, PtO tasks introduce an additional layer of complexity: decisions z are solutions to an optimization problem dependent on y, and their quality directly impacts task performance.

In the following sections, we formalize our proposed decision-aware dataset distance by extending OTDD to incorporate decision quality. This formulation provides a principled way to compare PtO datasets, ensuring that the resulting distance reflects meaningful differences in feature-label-decision distributions while remaining sensitive to the structure of the underlying optimization problem.

5.1 DATASET DISTANCE FORMULATION

To apply OT to datasets in PtO settings, we need a well-defined metric for the joint space of features, labels and decisions, $\mathcal{W} := \mathcal{X} \times \mathcal{Y} \times \Omega$, to serve as the ground cost function in the OT problem. A natural approach is to construct distances in $\mathcal{W}$ by combining metrics from the feature space $\mathcal{X}$, the label space $\mathcal{Y}$, and the decision space Ω. In most well-studied PtO settings, $\mathcal{X}$ and $\mathcal{Y}$ are equipped with metrics $d_{\mathcal{X}}$ and $d_{\mathcal{Y}}$, respectively, providing a natural foundation for measuring feature-label distances. However, defining an appropriate metric for the decision space Ω requires special consideration.

While some decision spaces naturally admit standard metrics, others—such as those arising in resource allocation or scheduling—do not align with conventional distance measures. Even when a metric exists for Ω, it may fail to capture decision quality regret, the ultimate objective in PtO tasks. For example, in a $p \times q$ grid, Euclidean or Manhattan distances can measure geometric differences between paths but fail to account for task-specific objectives, such as minimizing costs or maximizing safety.

To ensure the ground cost function for $\mathcal{W}$ properly reflects both decision quality and feature-label relationships, we introduce the concept of *decision quality disparity*. This extends traditional metrics by comparing decisions not just in terms of their spatial or structural differences but also in terms of their effectiveness under different labels. Specifically, decision quality disparity measures the extent to which two decisions $z, z' \in \Omega$ differ in performance when evaluated under labels y and y' respectively.

Definition 5.1. For an optimization problem $M(\cdot)$ with objective function g, the *decision quality disparity* function $l_g(\,\cdot\,; y, y') : \Omega^2 \to \mathbb{R}$ measures the difference in decision

[2] We use github.com/sanketkshah/LODLs

quality between two decisions $z, z' \in \Omega$ given the labels $y, y' \in \mathcal{Y}$. It is defined as:

$$l_g(z, z'; y, y') := |g(z; y) - g(z'; y')|. \tag{6}$$

Note that decision quality regret (Section 3.3) is a special case of decision quality disparity, where $q_{\text{reg}}(\hat{y}, y) = l_g(w^*(\hat{y}), w^*(y); y, y)$ for an optimization oracle w^*. We use decision quality disparity to define a point-wise distance in the joint feature-label-decision space $\mathcal{W}$. The resulting ground cost function $c^{\boldsymbol{\alpha}}_{PtO}$ for the OT problem is given by:

$$\begin{aligned} c^{\boldsymbol{\alpha}}_{PtO}((x, y, z), (x', y', z')) := \; & \alpha_X \cdot d_{\mathcal{X}}(x, x') \\ & + \alpha_Y \cdot d_{\mathcal{Y}}(y, y') \\ & + \alpha_W \cdot l_g(z, z'; y', y'), \end{aligned} \tag{7}$$

for $\boldsymbol{\alpha} = [\alpha_X, \alpha_Y, \alpha_W] \in \mathbb{R}^3_{\geq 0}$ such that $||\boldsymbol{\alpha}|| = 1$. Here, $d_{\mathcal{X}}$ and $d_{\mathcal{Y}}$ represent metrics for the feature space $\mathcal{X}$ and label space $\mathcal{Y}$, while l_g captures differences in decisions. The additive combination in c_{PtO} ensures simplicity and validity as a metric, with each component independently reflecting a distinct aspect of similarity. This design avoids complex, application-specific interactions and prioritizes interpretability. In the appendix, we show that c_{PtO} is a proper distance in $\mathcal{W}$. Notably, we set $\alpha_Y > 0$ to analyze scenarios where both labels and decisions contribute to PtO similarity versus cases where labels may be redundant.

We extend this point-wise distance to a distance between datasets $\mathcal{D}$ and $\mathcal{D}'$ by solving the OT cost with ground cost $c^{\boldsymbol{\alpha}}_{PtO}$, denoted as $d_{OT}(\mathcal{D}, \mathcal{D}'; c^{\boldsymbol{\alpha}}_{PtO})$. We refer to this distance as the *Optimal Transport Decision-Aware Dataset Distance (OTD^3).*

Proposition 5.2. *For any $\boldsymbol{\alpha} = (\alpha_X, \alpha_Y, \alpha_W)$ with $\alpha_X, \alpha_Y, \alpha_W > 0$, $d_{OT}(\mathcal{D}, \mathcal{D}'; c^{\boldsymbol{\alpha}}_{PtO})$ is a valid metric on $\mathcal{P}(\mathcal{X} \times \mathcal{Y} \times \Omega)$, the space of measures over joint distributions of features $\mathcal{X}$, labels $\mathcal{Y}$, and decisions Ω. If $\alpha_Y = 0$, $d_{OT}(\mathcal{D}, \mathcal{D}'; c^{\boldsymbol{\alpha}}_{PtO})$ is at least a pseudometric.*

This decision-aware dataset distance compares decisions z and z' by evaluating their decision quality disparity in $\mathbb{R}$ relative to a pair of fixed labels, rather than directly comparing them in the decision space Ω. Intuitively, comparing decisions based on their quality, i.e., comparing $g(z; y)$ with $g(z'; y)$, rather than comparing z and z' directly using some metric in Ω, if available, is reasonable because similar decisions might yield significantly different outcomes in the objective function. In Section 5.2 we show that comparing decision in this way offers a principled means of assessing adaptation success of PtO tasks across distributions in the feature-label-decision space.

Component Weights are Task-Specific Hyperparameters. The weights $\boldsymbol{\alpha}$ on the ground cost component (Eq. 7), are pivotal in defining the OTD^3, offering a flexible framework to account for the varying importance of features, labels, and decisions in PtO tasks. Unlike previous OT-based dataset distances that did not differentiate between the weights of feature and label components in the ground cost function—often because both were measured in the same space Alvarez-Melis and Fusi [2020] or were weighted equally Courty et al. [2017]— our method allows for distinct weights, enabling a more nuanced evaluation of dataset similarity tailored to each specific task. This flexibility ensures that the distance metric reflects the relative significance of each dataset component according to its impact on the PtO task, which can vary widely in practice depending on the application.

The impact of each component—features, labels, and decisions—on the overall distance can vary across PtO tasks. In particular, the decision and label components may sometimes capture overlapping structure. When such alignment occurs, the added value of decision information may be diminished, while in other cases, decisions encode complementary information. This mirrors the intuition from multivariate modeling where high correlation between variables can reduce the sensitivity to their individual weights. While our current formulation provides flexibility via weighting, understanding when and how much each component contributes remains an open and important question. We return to this empirically in Section 7.1.

5.2 DECISION REGRET ADAPTATION BOUND

Given source and target distributions $\mathcal{P}_S$ and $\mathcal{P}_T$ over $\mathcal{X} \times \mathcal{Y}$, we study domain adaptation from $\mathcal{P}_S$ to $\mathcal{P}_T$ in a PtO framework where decisions are generated by a downstream optimization problem $M(\cdot)$ parametrized in $\mathcal{Y}$. Let $f : \mathcal{X} \to \mathcal{Y}$ be a labeling function. We define the expected cost of f under a distribution $\mathcal{P}$ over $\mathcal{X} \times \mathcal{Y}$ with respect to any cost function $l : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ as

$$err(f; l, \mathcal{P}) := \mathbb{E}_{(x,y)\sim\mathcal{P}}\, l(f(x), y). \tag{8}$$

In PtO tasks, the performance of f over a distribution $\mathcal{P}$ is quantified as the *expected decision quality regret*, given by $err(f; q_{\text{reg}}, \mathcal{P})$. Our goal is to bound this error on the target distribution, $\text{err}(f; q_{\text{reg}}, \mathcal{P}_T)$, in terms of the distance between $\mathcal{P}_T$ and the source distribution $\mathcal{P}_S$. We use OTD^3 to achieve this.

Prior work by Courty et al. [2017] provided adaptation bounds for an expected target error $\text{err}(f; l, \mathcal{P}_T)$ with a cost function l that is bounded, symmetric, k-Lipschitz, and satisfies the triangle inequality. However, decision quality regret q_{reg}, the key cost function in PtO tasks, is inherently non-symmetric, making these bounds inapplicable to $\text{err}(f; q_{\text{reg}}, \mathcal{P})$. To address this, we introduce the notion of *decision quality disparity* l_q (Definition 5.1) to bound decision quality regret q_{reg}. Additionally, we assume that the de-

cision quality function q has a bounded rate of change with respect to both the predicted and true cost vectors (Assumption 5.3). Under these conditions, we derive an adaptation bound for $err(f; q_{\text{reg}}, \mathcal{P}_T)$ using the OTD^3 (Theorem 5.5). As demonstrated in lemmas B.1 and B.2 in the Appendix, Assumption 5.3 holds for common PtO task structures.

Assumption 5.3. The decision quality function q is k_1, k_2-Lipschitz. This means that for any $y, y^*, z, z^* \in \mathcal{Y}$ the following inequality holds:

$$|q(y, y^*) - q(z, z^*)| \leq k_1\|y - z\| + k_2\|y^* - z^*\|$$

.

Definition 5.4 (Courty et al. [2017])**.** Let μ_1 and μ_2 be distributions over some metric space $\mathcal{X}$ with metric $d_{\mathcal{X}}$. Let $\Pi(\mu_1, \mu_2)$ be a joint distribution over $\mu_1 \times \mu_2$. Let $\phi : \mathbb{R} \to [0, 1]$. A labeling function $f : \mathcal{X} \to \mathbb{R}$ is ϕ-*Lipschitz transferable with respect to* Π if for all $\lambda > 0$:

$$\Pr_{(x_1,x_2)\sim\Pi(\mu_1,\mu_2)}[|f(x_1) - f(x_2)| > \lambda d_{\mathcal{X}}(x_1, x_2)] \leq \phi(\lambda).$$

Theorem 5.5. *Suppose Assumption 5.3 holds for an optimization problem $M(\cdot)$ with optimization oracle w^*. Let $f : \mathcal{X} \to \mathcal{Y}$ be a labeling function, and define the distributions $\mathcal{P}_T^f := (x, y, w^*(f(x)))_{(x,y)\sim\mathcal{P}_T}$ and $\mathcal{P}_S^* := (x, y, w^*(y))_{(x,y)\sim\mathcal{P}_S}$ over the joint feature-label-decision space $\mathcal{W}$. Let Π^* denote the optimal coupling for the OT problem with ground cost $c_{PtO}^{\boldsymbol{\alpha}}$ between $\mathcal{P}_T^f$ and $\mathcal{P}_S^*$. Let $\tilde{f}$ be a labeling function that is ϕ-Lipschitz transferable with respect to Π^*. Assume that the feature space $\mathcal{X}$ is bounded by K and that $\tilde{f}$ is l-Lipschitz, satisfying $|\tilde{f}(x_1) - \tilde{f}(x_2)| \leq 2lK = L$.*

For any $\lambda > 0$ and $\alpha_W \in (0, 1)$ such that $(\lambda k_1 + k_2 + 1)\alpha_W = 1$, with $\alpha_X = \lambda k_1 \alpha_W$ and $\alpha_Y = k_2 \alpha_W$, the following bound holds with probability at least $1 - \delta$:

$$err(f; q_{reg}, \mathcal{P}_T) \leq err(\tilde{f}; q_{reg}, \mathcal{P}_S) + err(\tilde{f}; q_{reg}, \mathcal{P}_T) + k_1 L\phi(\lambda) + \frac{1}{\alpha_W} d_{OT}(\mathcal{P}_T^f, \mathcal{P}_S^*; c_{PtO}^{\boldsymbol{\alpha}}).$$

The proof of Theorem 5.5 is provided in the supplementary material. The first two terms in the bound represent the joint decision quality regret minimizer across the source and target distributions. This indicates that successful domain adaptation in the PtO framework requires predictions that achieve low regret in both domains simultaneously. This result aligns with the findings of Courty et al. [2017], Mansour et al. [2009] and Ben-David et al. [2010] in the context of domain adaptation for supervised learning. The third term $k_1 L\phi(\lambda)$ captures the extend to which Lipschitz continuity between the source and target distributions may fail.

The final term measures the discrepancy between the source domain $\mathcal{P}_S^*$ and the predicted target domain $\mathcal{P}_T^f$ using the optimal transport distance between their joint distributions of features, labels, and decisions. The bound relies on two key parameters: λ and α_W. λ controls the weight of the Lipschitz term and is valid for any $\lambda > 0$, while α_W determines the weight assigned to decisions in the convex combination $c_{PtO}^{\boldsymbol{\alpha}}$. Note that the bound holds for any combination of weights $\alpha_X, \alpha_Y, \alpha_W$, as λ can always be adjusted to ensure a valid convex combination.

Our approach recognizes the necessity of incorporating decisions into dataset distances when used for domain adaptation purposes for PtO tasks. Our OT-based dataset distance, defined by the ground cost function $c_{PtO}^{\boldsymbol{\alpha}}$, jointly accounts for differences in all key components—features, labels, and decisions—providing a comprehensive measure that is meaningful for adaptability of PtO tasks.

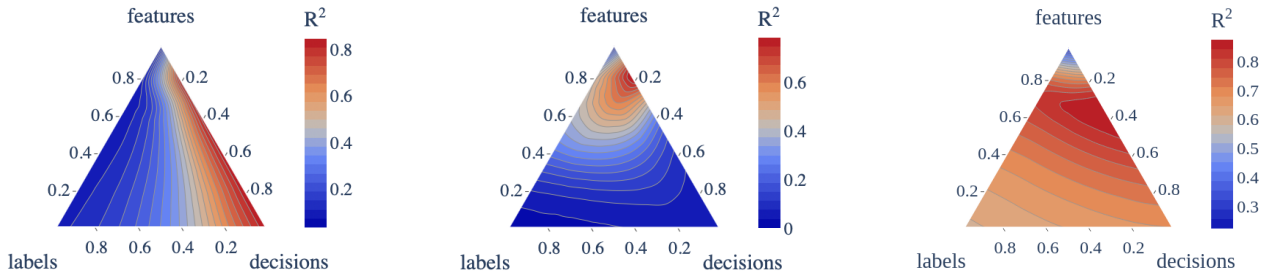
6 EXPERIMENTAL SETTINGS

We conduct experiments on three PtO settings with diverse structures and sensitivity to distribution shifts, making them well-suited for analyzing dataset distances in domain adaptation. Additional details are provided in the Appendix.

Linear Model Top-K [Shah et al., 2022]. This setting involves training a linear model to map features $x_n \sim U[-1, 1]$ to true utilities based on a cubic polynomial $p(x_n) = 10(x_n^3 - 0.65x_n)$. The downstream task is selecting the K elements with highest utility. We introduce synthetic distribution shifts by modifying the original feature-label distribution $\mathcal{P} = (\text{Id}, p)_* U[-1, 1]$. Specifically, for various values of $\gamma \in [0, 1.3]$, we define the feature-label distributions $\mathcal{P}_\gamma = (\text{Id}, p_\gamma)_* U[-1, 1]$ where $p_\gamma(x_n) = 10(x_n^3 - \gamma x_n)$, using $\mathcal{P}_{0.65}$ as the target distribution.

Warcraft Shortest Path [Vlastelica et al., 2020]. This task involves finding the minimum-cost path on RGB grid maps from the Warcraft II tileset dataset, where each pixel has an unknown travel cost. The goal is to predict these costs and then determine the optimal path from the top-left to the bottom-right pixel. The target distribution $\mathcal{P}$ is defined over $\mathbb{R}^{d\times d\times 3} \times \mathbb{R}^{p\times p}$, with $d = 96$ and $p = 12$. To simulate distribution shifts, we generate synthetic distributions $\mathcal{P}_\gamma$ by uniformly sampling pixel class costs from the same range as $\mathcal{P}$.

Inventory Stock Problem [Donti et al., 2017]. This task involves determining the order quantity z to minimize costs given a stochastic demand y, influenced by features x. The cost function f_{stock} includes linear and quadratic costs for both ordering and deviations (over-orders and under-orders) from demand. We generate problem instances by randomly sampling $x \in \mathbb{R}^n$ and then generating $p(y|x; \theta)$ according to $p(y|x; \theta) \propto \exp\big((\theta^T x)^2\big)$. Distribution shifts are introduced in features x and labels y: x is sampled from a Gaussian distribution with a mean sampled from $U[-0.5, 0.5]$, and θ is also sampled from a Gaussian distribution.



(a) Linear Model Top-K (b) Warcraft Shortest Path (c) Inventory Stock

Figure 2: **Component weighting and transferability prediction**. The color scale represents the R-square value from a linear regression of OTD^3—with all possible weight combinations for features, labels, and decisions—against regret transferability. The left border of the triplot shows R-square values when using OTD^3 with $\alpha_W = 0$, equivalent to OTDD.

7 EXPERIMENTS

7.1 SELECTING SOURCE DATASETS FOR TRANSFER LEARNING

Dataset distances for feature-label datasets, such as OTDD, have shown to be predictive of classification error/accuracy transfer—i.e., the error/accuracy on a target dataset $\mathcal{D}_T^{test}$ for a model adapted from a source $\mathcal{D}_S$ to a target $\mathcal{D}_T$. Alvarez-Melis and Fusi [2020] demonstrated that OTDD effectively predicts transferability and used this measure to select the best source dataset in a transfer learning task. We extend this source dataset selection experiment to the PtO setting, evaluating how well the OTD^3 predicts *regret transferability* between PtO tasks.

We analyze the correlation between the distance from a source dataset $\mathcal{D}_S$ to a target dataset $\mathcal{D}_T$ and the regret incurred on unseen target data $\mathcal{D}_T^{test}$ when adapting a model from $\mathcal{D}_S$ to $\mathcal{D}_T$—*i.e.* pretraining on $\mathcal{D}_S$ and fine-tuning on $\mathcal{D}_T$. We compare OTD$^3(\mathcal{D}_S, \mathcal{D}_T)$ against regret transferability $\mathcal{T}$, which quantifies the relative reduction in regret when transferring from $\mathcal{D}_S$ to $\mathcal{D}_T$:

$$\mathcal{T}(S \to T) = 100 \times \frac{\text{reg}(\mathcal{D}_T) - \text{reg}(\mathcal{D}_S \to \mathcal{D}_T)}{\text{reg}(\mathcal{D}_T)},$$

where $\text{reg}(\mathcal{D}_T)$ represents the mean regret when training directly on $\mathcal{D}_T$, and $\text{reg}(\mathcal{D}_S \to \mathcal{D}_T)$ represents the mean regret when adapting from $\mathcal{D}_S$ to $\mathcal{D}_T$. Each regret term is computed on $\mathcal{D}_T^{test}$, ensuring that transferability is evaluated based on the model's performance on unseen target data.

For every experimental setting we generate K source datasets $\mathcal{D}_{S_1}, \ldots, \mathcal{D}_{S_K}$, each sampled from a different distribution $\mathcal{P}_{S_i}$, along with training and test datasets $\mathcal{D}_T$ and $\mathcal{D}_T^{test}$ drawn from a target distribution $\mathcal{P}_T$. For each source-target pair $(\mathcal{D}_{S_i}, \mathcal{D}_T)$, we compute the regret transferability $\mathcal{T}(S_i \to T)$ by training models using standard DFL approaches (Appendix D) and analyze its relationship with the OTD^3.

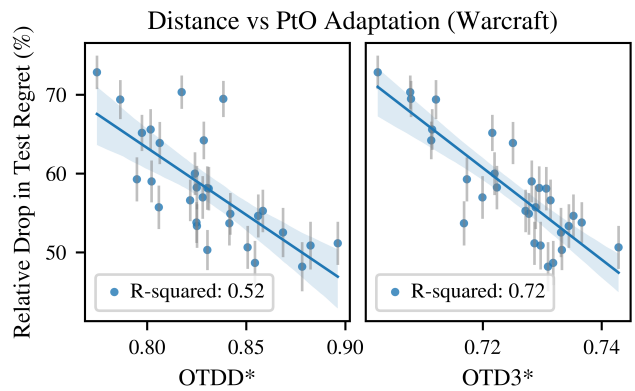


Figure 3: Dataset distance vs PtO adaptation in Warcraft. Results for a target dataset with 100 samples against 30 source datasets with 1,000 samples. Dataset distances OTDD and OTD^3 are computed with weights that maximize the correlation between distance and regret transferability.

Predicting transferability. Figure 2 shows the correlation strength (R^2 from linear regression) between regret transfer and dataset distance for different weighting combinations $\boldsymbol{\alpha}$. In the Linear Model Top-K and Warcraft settings, incorporating the decision component ($\alpha_W > 0$) significantly enhances the correlation between dataset distance and regret transfer, even when the label component is excluded ($\alpha_Y = 0$). Conversely, omitting the decision component ($\alpha_W = 0$, left side of the triplot) weakens this correlation. This trend is further emphasized when comparing the highest achievable correlation. In Warcraft, the OTD^3 with maximizing weights is far more predictive of regret transfer than the OTDD (or the OTD^3 with $\alpha_W = 0$) under its best weighting (Figure 3). The best-performing weights in this case were $\alpha_X = 0.8$ and $\alpha_Y = 0.2$ for the OTDD, and $\alpha_X = 0.75$, $\alpha_Y = 0$, and $\alpha_W = 0.25$. We denote these optimized versions as the OTDD$*$ and the OTD3$*$.

In Figure 4 we extend our analysis to varying sizes of the target dataset, ranging from 10 to 100 samples, which are used for dataset distance computation and fine-tuning,

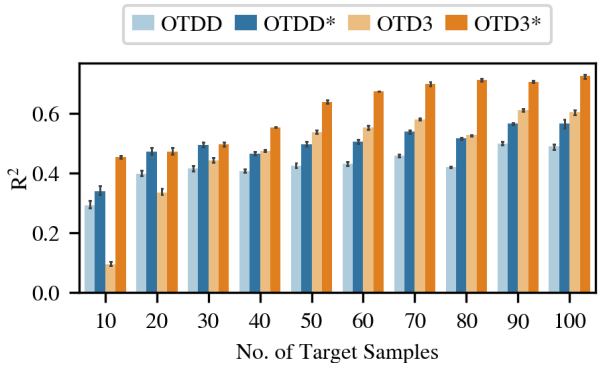
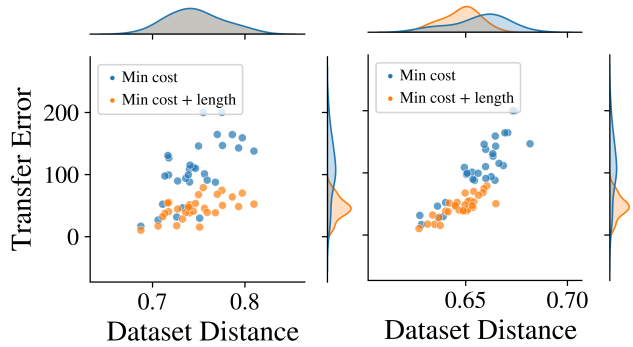


Figure 4: Correlation between dataset distance and regret transferability (R^2) vs. target sample size, for four distance variants: OTDD with equal feature-label weights and optimized weights (OTDD*), and OTD^3 with equal output weights (0.5, 0.25, 0.25) and optimized weights (OTD^3*).



(a) Not including decisions (b) Including decisions

Figure 5: Distance vs. Adaptation for two tasks in the Warcraft setting. Dataset distance is computed (a) without incorporating decisions, and (b) with decision incorporation.

while keeping the source datasets and target test set fixed at 1,000 samples each. We compute dataset distances using a three-dimensional weight grid and compare the correlation achieved with equal input-output weighting ($\alpha_X = 0.5, \alpha_Y = 0.5$ for OTDD, $\alpha_X = 0.5, \alpha_Y = 0.25, \alpha_W = 0.25$ for OTD^3) against OTDD* and OTD^3*.

Incorporating decisions into the dataset distance, with appropriate weighting, consistently improves the predictability of PtO transferability across all target sample sizes. While tuning the decision component weight significantly boosts correlation, rapidly reaching $R^2 > 0.6$, OTD^3 outperforms OTDD from as few as 30 target samples onward, demonstrating its effectiveness even without extensive data for weight optimization.

Decision vs label component. The advantage of including the decision component over the label component is less pronounced in the Inventory Stock problem (Fig. 2c). Here, either the label or decision component with features still maintains a strong correlation between regret transfer and dataset distance. To explore this further, we examine how differences in the label space $d_y(y, y')$ correlate with differences in the decision space $l_q(y, y', z, z')$. In the Inventory Stock problem, there is a strong correlation between these differences (Appendix Fig. F.1), suggesting that decisions are closely tied to the labels. In contrast, the Warcraft domain lacks this strong correlation (Appendix Fig. F.1), making the decision component more critical for accurately predicting transferability.

7.2 CHARACTERIZING TARGET SHIFT IMPACT

Target shift—where label distributions change while feature distributions remain constant—creates mismatches between training and test datasets, often degrading performance in supervised learning. However, our experimental results (Fig. 3[left]) show that some source datasets with significant target shift—characterized by high feature-label distance—can still achieve low regret in the PtO task. This suggests that target shift may not impact PtO performance in the same way it affects purely predictive tasks.

To further explore the impact of target shift in PtO tasks, we analyze the Warcraft setting under two downstream optimization tasks: minimizing path cost alone and minimizing both path cost and length. We apply the same transfer learning experiment from Section 7.1 it to these two tasks. Although the same target shifts are applied on both tasks, their effect on PtO transferability is less severe for minimizing path cost and length compared to minimizing cost alone (Fig. 5). Our decision-aware dataset distance, using weights from Section 7.1, effectively captures this behavior. The distance distribution for the task less impacted by the target shift is more left-skewed (Fig. 5b). In contrast, the dataset distance that only accounts for features and labels, is unable to differentiate between these two tasks (Fig. 5a).

7.3 ROBUSTNESS TO MODEL COMPLEXITY

We assess the robustness of OTD^3 by evaluating its performance across five model architectures of increasing complexity in the Warcraft setting (Figure 6). Although OTD^3 is model-agnostic by design, we measure its effectiveness through its ability to predict model transferability, specifically via R^2 values across weight configurations. In this setting, we find that intermediate-complexity models (Mobilenet, Partial ResNet18, Partial ResNet34) exhibit both high R^2 and broad regions of strong performance. This suggests that OTD^3 can reliably identify informative weightings when the underlying regret landscape is structured yet stable. The smoothness of these regions also indicates robustness to variations in component weights.

At the lower end of the complexity spectrum, the Small

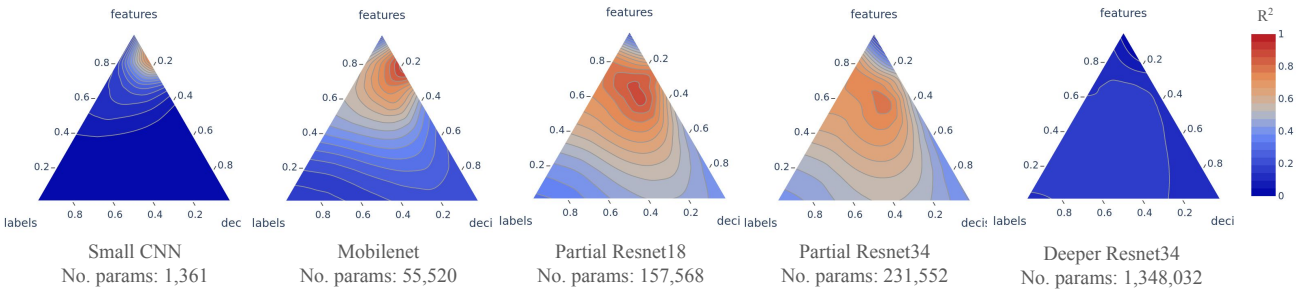


Figure 6: **OTD^3 Performance as Model Complexity Increases**. Predictability of OTD^3 (measured via R^2) across feature/label/decision weightings, shown for increasingly expressive model architectures on the Warcraft Shortest Path setting.

CNN displays low regret variance overall. Still, OTD^3 is able to highlight a narrow region of relative predictive strength—helping distinguish among otherwise uniformly weak configurations. In contrast, the most complex model (Deeper ResNet34) presents high regret variance, likely due to over-parameterization relative to the limited data. In this case, OTD^3 struggles to recover consistent patterns, reflecting the difficulty of transferability prediction in noisy, unstable regret landscapes.

These results suggest OTD^3 is most effective when regret variation is meaningful but not overly erratic, performing optimally when relationships between dataset distance and transferability are discernible and not obscured by noise from inappropriate model complexity. The robust performance across reasonably complex architectures highlights OTD^3's practical utility in PtO scenarios.

8 DISCUSSION

We introduce the first dataset distance tailored to PtO tasks, integrating features, labels, and decisions to better assess prediction-to-decision similarities. Our experiments show that incorporating decisions significantly improves transfer predictability, particularly in complex settings where label shifts do not directly correlate with decisions. This approach effectively captures task dynamics dictated by downstream optimization structures without requiring explicit analysis. Moreover, our framework is adaptable, allowing flexible weighting of components to provide meaningful comparisons across diverse PtO tasks—an essential feature for real-world applications where datasets vary not only in features and labels but also in decision complexity.

Several promising directions can extend our framework. Handling decision components of varying structures and dimensions using techniques like the Gromov-Wasserstein [Mémoli, 2011] distance could bridge gaps between non-comparable decision spaces. Further refining the weighting of features, labels, and decisions—particularly through tuning methods independent of transferability measures—could enhance its utility. Additionally, adapting our approach to more intricate PtO structures, such as those where multiple feature-label pairs define a single decision, through a hierarchical OT framework [Yurochkin et al., 2019], could further improve its applicability.

By establishing this first notion of dataset distance designed for PtO tasks, our work lays a foundation for future research, opening avenues for more robust and versatile transferability metrics in decision-aware learning.

ACKNOWLEDGMENTS

We thank Sanket Shah for insightful discussions throughout the development of this work. We also thank the reviewers and participants at the DMLR Workshop and the Humans, Algorithmic Decision-Making, and Society Workshop at ICML 2024 for their feedback on earlier versions, and the UAI reviewers for their useful feedback that significantly improved this work.

PRD acknowledges support from the NSF under the AI Institute for Societal Decision Making (AI-SDM), Award No. 2229881. KW acknowledges support from NSF IIS-2403240 and the Schmidt Sciences AI2050 Fellowship. DAM acknowledges support from the Kempner Institute, the Aramont Fellowship Fund, and the FAS Dean's Competitive Fund for Promising Scholarship.

References

David Alvarez-Melis and Nicolo Fusi. Geometric Dataset Distances via Optimal Transport. In *Proceedings of the 34th Conference on Neural Information Processing Systems*, 2020.

David Alvarez-Melis, Tommi Jaakkola, and Stefanie Jegelka. Structured Optimal Transport. In *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*. PMLR, 2018.

Brandon Amos and J. Zico Kolter. OptNet: Differentiable Optimization as a Layer in Neural Networks. In *Proceed-*

ings of the 34th International Conference on Machine Learning, 2017.

Dishank Bansal, Ricky T. Q. Chen, Mustafa Mukadam, and Brandon Amos. TaskMet: Task-Driven Metric Learning for Model Learning. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, 2023.

Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A Theory of Learning from Different Domains. *Machine Learning*, 79:151–175, 2010.

Nicolas Courty, Rémi Flamary, and Devis Tuia. Domain Adaptation with Regularized Optimal Transport. In *Machine Learning and Knowledge Discovery in Databases*, pages 274–289, 2014.

Nicolas Courty, Rémi Flamary, Amaury Habrard, and Alain Rakotomamonjy. Joint Distribution Optimal Transportation for Domain Adaptation. In *Proceedings of the 31st Conference on Neural Information Processing Systems*, 2017.

Julie Delon and Agnès Desolneux. A Wasserstein-Type Distance in the Space of Gaussian Mixture Models. *SIAM Journal on Imaging Sciences*, 13(2):936–970, 2020.

Priya Donti, Brandon Amos, and J. Zico Kolter. Task-based End-to-end Model Learning in Stochastic Optimization. In *Proceedings of the 31st Conference on Neural Information Processing Systems*, 2017.

Adam N. Elmachtoub and Paul Grigas. Smart "Predict, then Optimize". *Management Science*, 68(1):9–26, 2022.

Adam N. Elmachtoub, Henry Lam, Haofeng Zhang, and Yunfan Zhao. Estimate-Then-Optimize versus Integrated-Estimation-Optimization versus Sample Average Approximation: A Stochastic Dominance Perspective, 2025.

Charlie Frogner, Farzaneh Mirzazadeh, and Justin Solomon. Learning Embeddings into Entropic Wasserstein Spaces. In *International Conference on Learning Representations*, 2019.

Hicham Janati, Marco Cuturi, and Alexandre Gramfort. Wasserstein Regularization for Sparse Multi-Task Regression. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, 2019.

Kevin Fu Jiang, James Zou, Weixin Liang, and Yongchan Kwon. OpenDataVal: a Unified Benchmark for Data Valuation. In *Proceedings of the 37th Conference on Neural Information Processing Systems*, 2023.

Sonja Johnson-Yu, Jessie Finocchiaro, Kai Wang, Yevgeniy Vorobeychik, Arunesh Sinha, Aparna Taneja, and Milind Tambe. Characterizing and Improving the Robustness of Predict-Then-Optimize Frameworks. In *Decision and Game Theory for Security*, 2023.

Hoang Anh Just, Feiyang Kang, Jiachen T. Wang, Yi Zeng, Myeongseob Ko, Ming Jin, and Ruoxi Jia. LAVA: Data Valuation without Pre-Specified Learning Algorithms. In *International Conference on Learning Representations*, 2023.

L. Kantorovitch. On the Translocation of Masses. *Dokl. Akad. Nauk SSSR*, 37(1):227–229, 1942. ISSN 0002-3264.

Mo Liu, Paul Grigas, Heyuan Liu, and Zuo-Jun Max Shen. Active Learning in the Predict-then-Optimize Framework: A Margin-Based Approach, 2025. arXiv:2305.06584.

Jayanta Mandi, James Kotary, Senne Berden, Maxime Mulamba, Victor Bucarey, Tias Guns, and Ferdinando Fioretto. Decision-Focused Learning: Foundations, State of the Art, Benchmark and Future Opportunities, 2023.

Yishay Mansour, Mehryar Mohri, and Afshin Rostamizadeh. Domain Adaptation: Learning Bounds and Algorithms. In *Proceedings of The 22nd Annual Conference on Learning Theory*, 2009.

Marina Adriana Mercioni and Stefan Holban. A Survey of Distance Metrics in Clustering Data Mining Techniques. In *Proceedings of the 3rd International Conference on Graphics and Signal Processing*, 2019.

Boris Muzellec and Marco Cuturi. Generalizing Point Embeddings using the Wasserstein Space of Elliptical Distributions. In *Proceedings of the 32nd Conference on Neural Information Processing Systems*, 2018.

Facundo Mémoli. Gromov–Wasserstein Distances and the Metric Approach to Object Matching. *Foundations of Computational Mathematics*, 11(4):417–487, 2011. ISSN 1615-3383.

Alfred Müller. Integral Probability Metrics and Their Generating Classes of Functions. *Advances in Applied Probability*, 29(2):429–443, 1997. ISSN 0001-8678.

Meng Qi, Paul Grigas, and Zuo-Jun Max Shen. Integrated Conditional Estimation-Optimization, 2023. arXiv:2110.12351.

Kevin Ren, Yewon Byun, and Bryan Wilder. Decision-Focused Evaluation of Worst-Case Distribution Shift. In *Proceedings of the Fortieth Conference on Uncertainty in Artificial Intelligence*, pages 3076–3093. PMLR, 2024.

Sanket Shah, Kai Wang, Bryan Wilder, Andrew Perrault, and Milind Tambe. Decision-Focused Learning without Differentiable Optimization: Learning Locally Optimized Decision Losses. In *Proceedings of the 36th Conference on Neural Information Processing Systems*, 2022.

Sanket Shah, Andrew Perrault, Bryan Wilder, and Milind Tambe. Leaving the Nest: Going Beyond Local Loss Functions for Predict-Then-Optimize. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2023.

Changjian Shui, Mahdieh Abbasi, Louis-Émile Robitaille, Boyu Wang, and Christian Gagné. A Principled Approach for Learning Task Similarity in Multitask Learning. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, 2019.

Anh Tran, Cuong Nguyen, and Tal Hassner. Transferability and Hardness of Supervised Classification Tasks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

Sergio Verdú. Total Variation Distance and the Distribution of Relative Information. In *Information Theory and Applications Workshop (ITA)*, pages 1–3, 2014.

Cédric Villani. *Optimal Transport, Old and New*, volume 338. Springer Science & Business Media, Berlin, Heidelberg, 2008. ISBN 978-3-540-71049-3.

Marin Vlastelica, Anselm Paulus, Vit Musil, Georg Martius, and Michal Rolinek. Differentiation of Blackbox Combinatorial Solvers. In *Eighth International Conference on Learning Representations*, 2020.

Kai Wang, Bryan Wilder, Andrew Perrault, and Milind Tambe. Automatically Learning Compact Quality-aware Surrogates for Optimization Problems. In *Proceedings of the 34th Conference on Neural Information Processing Systems*, 2020.

Bryan Wilder, Bistra Dilkina, and Milind Tambe. Melding the Data-Decisions Pipeline: Decision-Focused Learning for Combinatorial Optimization. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence*, 2019.

Yujia Xie, Hanjun Dai, Minshuo Chen, Bo Dai, Tuo Zhao, Hongyuan Zha, Wei Wei, and Tomas Pfister. Differentiable Top-k Operator with Optimal Transport. In *Proceedings of the 34th Conference on Neural Information Processing Systems*, 2020.

Mikhail Yurochkin, Sebastian Claici, Edward Chien, Farzaneh Mirzazadeh, and Justin M Solomon. Hierarchical Optimal Transport for Document Representation. In *Proceedings of the 33rd Conference on Neural Information Processing Systems*, 2019.

What is the Right Notion of Distance between Predict-then-Optimize Tasks? (Supplementary Material)

Paula Rodriguez-Diaz[1] **Lingkai Kong**[1] **Kai Wang**[2] **David Alvarez-Melis**[1,3,4] **Milind Tambe**[1]

[1]Harvard University
[2]Georgia Institute of Technology
[3]Kempner Institute
[4]Microsoft Research

A PROOF OF PROPOSITION 5.2

To demonstrate that the OTD[3], $d_{OT}(\cdot,\cdot;c_{PtO})$ is a valid metric, it is sufficient to verify that the ground cost function c_{PtO} used in the optimal transport problem is a metric on $\mathcal{X}\times\mathcal{Y}\times\Omega$. If c_{PtO} is indeed a metric, then $d_{OT}(\cdot,\cdot;c_{PtO})$ corresponds to the Wasserstein distance Villani [2008]. In Equation 7, $d_{OT}(\cdot,\cdot;c_{PtO})$ is defined as a convex combination of $d_{\mathcal{X}}$ and $d_{\mathcal{Y}}$, which are metrics on $\mathcal{X}$ and $\mathcal{Y}$ respectively, and the decision quality disparity l_q. To show that c_{PtO} is a metric, it suffices to show that l_q satisfies the four metric properties: non-negativity, identity of indiscernibles, symmetry, and the triangle inequality. If l_q does not individually satisfy these properties, we must demonstrate that the convex combination of $d_{\mathcal{X}}$, $d_{\mathcal{Y}}$, and l_q satisfies these properties collectively under the assumption that $\alpha_X,\alpha_Y,\alpha_W>0$.

First, l_q is clearly non-negative because it is defined as an absolute value. It is symmetric in the convex combination of c_{PtO} because it is taken as the absolute difference between two decision qualities with fixed true costs.

$$\begin{aligned} l_g(z,z';y',y') &= \left|q(z;y')-q(z';z')\right| \\ &= \left|q(z';y')-q(z;z')\right| \\ &= l_g(z',z;y',y') \end{aligned}$$

Moreover, l_q satisfies triangle inequality due to the triangle inequality property of the absolute value.

$$\begin{aligned} & l_g(z_1,z_2;y_1,y_2)+l_g(z_2,z_3;y_2,y_3) \\ &= \left|g(z_1;y_1)-g(z_2;y_2)\right|+\left|g(z_2;y_2)-g(z_3;y_3)\right| \\ &\le \left|g(z_1;y_1)-g(z_2;y_2)+g(z_2;y_2)-g(z_3;y_3)\right| \\ &= \left|g(z_1;y_1)-g(z_3;y_3)\right| \\ &= l_g(z_1,z_3;y_1,y_3) \end{aligned}$$

Lastly, while l_q might not satisfy the identity of indiscernibles in isolation (specifically, $l_q(y,y';z,z)=0$ does not necessarily imply $y=y'$; meaning two different decisions can lead to the same objective value), c_{PtO} does satisfy this property for $\alpha_X,\alpha_Y,\alpha_W>0$. If $(x,y,z)=(x',y',z')$, then $l_g(z,z';y',y')=\left|g(z;y)-g(z';y)\right|=0$ because $z=z'$ implies $g(z;y)=g(z';y)$ and hence $c_{PtO}((x,y,z),(x',y',z'))=0$. Conversely, if $c_{PtO}((x,y,z),(x',y',z'))=0$, then $d_{\mathcal{X}}(x,x')=0$, $d_{\mathcal{Y}}(y,y')=0$, and $l_q(y,y';z,z)=0$ because $\alpha_X,\alpha_Y,\alpha_W>0$. Since $d_{\mathcal{Y}}(y,y')=0$ implies $y=y'$ (because $d_{\mathcal{Y}}$ is a metric), it follows that $w^*(y)=w^*(y')$ and hence $z=z'$.

Therefore, c_{PtO} satisfies the identity of indiscernibles. Consequently, since l_q satisfies non-negativity, symmetry, and the triangle inequality, and since c_{PtO} satisfies the identity of indiscernibles, $d_{OT}(\cdot,\cdot;c_{PtO})$ is indeed a valid metric with c_{PtO} a valid metric on $\mathcal{X}\times\mathcal{Y}\times\Omega$.

B PREAMBLE FOR THEOREM 5.5

B.1 VALIDITY ASSUMPTION 5.3

Assumption 5.3 imposes a specific structure on the downstream optimization problem by assuming that the decision quality function has a bounded rate of change with respect to both the predicted and true cost vectors. This is a reasonable assumption for certain downstream optimization tasks, as highlighted in the following lemmas.

Lemma B.1. *If $M(\cdot)$ is a convex program with a strongly convex objective and constraints with independent derivatives (Linear Independence Constraint Qualification (LICQ)), Assumption 5.3 holds.*

The strong convexity of the objective ensures that the gradient is Lipschitz continuous, while the LICQ guarantees that the optimal solutions depend continuously on the parameters. By the smoothness of the objective and the continuity of the optimal solutions, the difference in the decision quality function q between two sets of parameters and their corresponding optimal solutions can be bounded by a linear combination of the distances between the parameters and the distances between the optimal solutions.

Lemma B.2. *If $M(\cdot)$ has a linear optimization objective with a strongly convex feasible region, Assumption 5.3 holds.*

When $M(\cdot)$ has a linear optimization objective and a strongly convex feasible region, the decision quality function q satisfies the k_1, k_2-Lipschitz property. The linearity of the objective ensures that changes in the parameters lead to proportional changes in the objective value, while the strong convexity of the feasible region guarantees that the optimal solutions are unique and vary smoothly with respect to the parameters. This smooth dependence, combined with the linear structure of the objective, implies that the difference in q between two sets of parameters and their corresponding optimal solutions can be bounded by a linear combination of the distances between the parameters and the distances between the optimal solutions.

B.2 LIPSCHITZNESS OF THE DECISION QUALITY DISPARITY FUNCTION

To establish the bound presented in Theorem 5.5, we rely on the fact that l_g is k_1, k_2-Lipschitz under Assumption 5.3. The following proposition demonstrates that l_g indeed satisfies the Lipschitz condition given this assumption.

Proposition B.3. *If g, the objective function of the downstream optimization problem, is k_1, k_2-Lipschitz (Assumption 5.3), then l_g is also k_1, k_2-Lipschitz.*

Proof.

$$
\begin{aligned}
&\left|l_g(z, z_1; y, y_1) - l_g(z, z_2; y, y_2)\right| \\
&= \left||g(z; y) - g(z_1; y_1)| - |g(z; y) - g(z_2; y_2)|\right| \\
&\leq \left|g(z; y) - g(z_1; y_1) - g(z; y) + g(z_2; y_2)\right| && (9) \\
&= \left|g(z_2; y_2) - g(z_1; y_1)\right| \\
&= \left|g(z_2; y_2) - g(z_1; y_2) + g(z_1; y_2) - g(z_1; y_1)\right| \\
&\leq \left|g(z_2; y_2) - g(z_1; y_2)\right| + \left|g(z_1; y_2) - g(z_1; y_1)\right| && (10) \\
&\leq k_1\|z_1 - z_2\| + k_2\|y_1 - y_2\| && (11)
\end{aligned}
$$

Inequalities (9) and (10) are a result of the triangle inequality of the absolute value. Inequality (11) is due to the $k_1 - k_2$-lipschitzness of g. □

C PROOF OF THEOREM 5.5

Theorem C.1. *Suppose Assumption 5.3 holds. For a feature space $\mathcal{X}$, a label space $\mathcal{Y}$, and a decision set Ω, let $\mathcal{W} := \mathcal{X} \times \mathcal{Y} \times \Omega$. Let $\mathcal{P}_S$ and $\mathcal{P}_T$ be the source and target distributions over $\mathcal{X} \times \mathcal{Y}$ respectively. For any labeling function $f : \mathcal{X} \to \mathcal{Y}$, let $\mathcal{P}_T^f$ and $\mathcal{P}_S^*$ be distributions over $\mathcal{W}$ given by $\mathcal{P}_T^f := (x, y, w^*(f(x)))_{(x,y)\sim\mathcal{P}_T}$ and $\mathcal{P}_S^* := (x, y, w^*(y))_{(x,y)\sim\mathcal{P}_S}$. For a ground cost function of the form*

$$c^{\boldsymbol{\alpha}}_{PtO}((x, y, z), (x', y', z')) = \alpha_X d_{\mathcal{X}}(x, x') + \alpha_Y d_{\mathcal{Y}}(y, y') + \alpha_W l_g(z, z'; y', y'),$$

let Π^* be the coupling that minimizes the OT problem with ground cost $c^{\boldsymbol{\alpha}}_{PtO}$ between $\mathcal{P}^f_T$ and $\mathcal{P}^*_S$. Let $\tilde{f}$ be a labeling function that is ϕ-Lipschitz transferable w.r.t. Π^*. We assume $\mathcal{X}$ is bounded by K and $\tilde{f}$ is l-Lipschitz, such that $|\tilde{f}(x_1) - \tilde{f}(x_2)| \leq 2lK = L$. Then, for all $\lambda > 0$ and $\alpha_W \in (0,1)$ such that $(\lambda k_1 + k_2 + 1)\alpha_W = 1$, and $\alpha_X = \lambda k_1 \alpha_W$ and $\alpha_Y = k_2 \alpha_W$, we have with probability at least $1-\delta$ that:

$$err(f; q_{reg}, \mathcal{P}_T) \leq \; err(\tilde{f}; q_{reg}, \mathcal{P}_S) + err(\tilde{f}; q_{reg}, \mathcal{P}_T) + k_1 L\phi(\lambda) + (1/\alpha_W) d_{OT}(\mathcal{P}^f_T, \mathcal{P}^*_S\,; c^{\boldsymbol{\alpha}}_{PtO})$$

Proof.

$$\begin{aligned}
&err(f; q_{\text{reg}}, \mathcal{P}_T) \\
&\quad= \mathbb{E}_{(x,y)\sim\mathcal{P}_T}\; l_g(w^*(f(x)), w^*(y); y, y) \\
&\quad\leq \mathbb{E}_{(x,y)\sim\mathcal{P}_T}\; l_g(w^*(f(x)), w^*(\tilde{f}(x)); y, y) + \mathbb{E}_{(x,y)\sim\mathcal{P}_T}\; l_g(w^*(\tilde{f}(x)), w^*(y); y, y) && (12)\\
&\quad= \mathbb{E}_{(x,y)\sim\mathcal{P}_T}\; l_g(w^*(\tilde{f}(x)), w^*(f(x)); y, y) + \mathbb{E}_{(x,y)\sim\mathcal{P}_T}\; l_g(w^*(\tilde{f}(x)), w^*(y); y, y) && (13)\\
&\quad= \mathbb{E}_{(x,y,z)\sim\mathcal{P}^f_T}\; l_g(w^*(\tilde{f}(x)), z; y, y) + \mathbb{E}_{(x,y)\sim\mathcal{P}_T}\; l_g(w^*(\tilde{f}(x)), w^*(y); y, y) && (14)\\
&\quad= \mathbb{E}_{(x,y,z)\sim\mathcal{P}^f_T}\; l_g(w^*(\tilde{f}(x)), z; y, y) - err(\tilde{f}; q_{\text{reg}}, \mathcal{P}_S) + err(\tilde{f}; q_{\text{reg}}, \mathcal{P}_S) + err(\tilde{f}; q_{\text{reg}}, \mathcal{P}_T) \\
&\quad= \mathbb{E}_{(x,y,z)\sim\mathcal{P}^f_T}\; l_g(w^*(\tilde{f}(x)), z; y, y) - \mathbb{E}_{(x,y,z)\sim\mathcal{P}^*_S}\; l_g(w^*(\tilde{f}(x)), z; y, y) + err(\tilde{f}; q_{\text{reg}}, \mathcal{P}_S) + err(\tilde{f}; q_{\text{reg}}, \mathcal{P}_T) \\
&\quad\leq \left|\mathbb{E}_{(x,y,z)\sim\mathcal{P}^f_T}\; l_g(w^*(\tilde{f}(x)), z; y, y) - \mathbb{E}_{(x,y,z)\sim\mathcal{P}^*_S}\; l_g(w^*(\tilde{f}(x)), z; y, y)\right| + err(\tilde{f}; q_{\text{reg}}, \mathcal{P}_S) + err(\tilde{f}; q_{\text{reg}}, \mathcal{P}_T)
\end{aligned}$$

Inequality (12) uses the fact that $l_g(\,\cdot\,; y, y)$ satisfies the triangle inequality and line (13) is due to the symmetry of $l_g(\,\cdot\,; y, y)$ for any $y \in \mathcal{Y}$. Line (14) comes from the fact that $\mathcal{P}^f_T := (x, f(x), y)_{(x,y)\sim\mathcal{P}_T}$. We continue by bounding the first term.

$$\begin{aligned}
&\left|\mathbb{E}_{(x,y,z)\sim\mathcal{P}^f_T}\; l_g(w^*(\tilde{f}(x)), z; y, y) - \mathbb{E}_{(x,y,z)\sim\mathcal{P}^*_S}\; l_g(w^*(\tilde{f}(x)), z; y, y)\right| \\
&= \left|\int_{\mathcal{W}} l_g(w^*(\tilde{f}(x)), z; y, y)(\mathcal{P}^f_T(X = x, Y = y, Z = z) - \mathcal{P}^*_S(X = x, Y = y, Z = z))\,\mathrm{d}x\,\mathrm{d}y\,\mathrm{d}z\right| \\
&= \left|\int_{\mathcal{W}} l_g(w^*(\tilde{f}(x)), z; y, y)\,\mathrm{d}\Pi^*((x_s, y_s, z_s), (x_t, y_t, z^f_t))\right| \\
&\leq \int_{\mathcal{W}^2} \left|l_g(\tilde{z}_t, z^f_t; y_t, y_t) - l_g(\tilde{z}_s, z_s; y_s, y_s)\right| \mathrm{d}\Pi^*(\mathbf{w}_s, \mathbf{w}^f_t) && (15)\\
&\leq \int_{\mathcal{W}^2} \left|l_g(\tilde{z}_t, z^f_t; y_t, y_t) - l_g(\tilde{z}_s, z^f_t; y_s, y_t)\right| + \left|l_g(\tilde{z}_s, z^f_t; y_s, y_t) - l_g(\tilde{z}_s, z_s, ; y_s, y_s)\right| \mathrm{d}\Pi^*(\mathbf{w}_s, \mathbf{w}^f_t) && (16)\\
&\leq \int_{\mathcal{W}^2} k_1 d_{\mathcal{Y}}(\tilde{f}(x_t), \tilde{f}(x_s)) + k_2 d_{\mathcal{Y}}(y_t, y_s) + \left|l_g(\tilde{z}_s, z^f_t; y_s, y_t) - l_g(\tilde{z}_s, y_s; y_s, y_s)\right| \mathrm{d}\Pi^*(\mathbf{w}_s, \mathbf{w}^f_t) && (17)\\
&\leq\; k_1 L\phi(\lambda) + \int_{\mathcal{W}^2} \lambda k_1 d_{\mathcal{X}}(x_t, x_s) + k_2 d_{\mathcal{Y}}(y_t, y_s) + \left|l_g(\tilde{z}_s, z^f_t; y_s, y_t) - l_g(\tilde{z}_s, y_s; y_s, y_s)\right| \mathrm{d}\Pi^*(\mathbf{w}_s, \mathbf{w}^f_t) && (18)\\
&\leq\; k_1 L\phi(\lambda) + \int_{\mathcal{W}^2} \lambda k_1 d_{\mathcal{X}}(x_t, x_s) + k_2 d_{\mathcal{Y}}(y_t, y_s) + l_g(z^f_t, z_s; y_s, y_s)\,\mathrm{d}\Pi^*(\mathbf{w}_s, \mathbf{w}^f_t)
\end{aligned}$$

From line (15) onwards we take $\mathbf{w}_s := (x_s, y_s, y_s), \mathbf{w}^f_t := (x_t, y^f_t, y_t)$ and $\tilde{z}_s = w^*(\tilde{f}(x_s)), \tilde{z}_t = w^*(\tilde{f}(x_t))$ for ease of notation. Given a weight α_W, we now normalize the last term such that the ground cost function is a convex combination of $d_{\mathcal{X}}$, $d_{\mathcal{Y}}$m and l_g.

$$\begin{aligned}
&\int_{\mathcal{W}^2} \lambda k_1 d_{\mathcal{X}}(x_t, x_s) + k_2 d_{\mathcal{Y}}(y_t, y_s) + l_g(z^f_t, z_s; y_s, y_s)\,\mathrm{d}\Pi^*(\mathbf{w}_s, \mathbf{w}^f_t) \\
&= \frac{1}{\alpha_W} \int_{\mathcal{W}^2} \lambda k_1 \alpha_W d_{\mathcal{X}}(x_t, x_s) + k_2 \alpha_W d_{\mathcal{Y}}(x_t, x_s) + \alpha_W l_g(z^f_t, z_s; y_s, y_s)\,\mathrm{d}\Pi^*(\mathbf{w}_s, \mathbf{w}^f_t) \\
&= \frac{1}{\alpha_W}\; d_{OT}(\mathcal{P}^f_T, \mathcal{P}^*_S; c^{\boldsymbol{\alpha}}_{PtO})
\end{aligned}$$

□

D EXPERIMENTAL SETTINGS DETAILS

D.1 LINEAR MODEL TOP-K Shah et al. [2022]

PtO task description. The Linear Model Top-K setting is a learning task designed to evaluate decision-focused learning approaches in scenarios where the true relationship between features and outcomes is nonlinear, yet the model used for prediction is constrained to be linear. Specifically, the objective is to train a linear model to perform top–K selection when the underlying data is generated by a cubic polynomial function. This controlled setup enables an assessment of how well decision-focused methods handle model misspecification. The predict-then-optimize (PtO) task in this setting is defined as follows:

Predict: Given the feature $x_n \sim \mathcal{P}_{\mathcal{X}}$, where $\mathcal{P}_{\mathcal{X}} = \text{Unif}[-1, 1]$, of a resource n, the prediction tasks consists of using a linear model to predict the corresponding utility $\hat{y}_n$, where the true utility $y_n = p(x_n)$ is a cubic polynomial in x_n. The predictions for N resources are aggregated into a vector $\hat{\boldsymbol{y}} = [\hat{y}_1, \ldots, \hat{y}_N]$, where each element corresponds to the predicted utility of a resource.

Optimize: The optimization task involves selecting the K out of N resources with the highest utility. This corresponds to solving the optimization problem $M(\hat{\boldsymbol{y}}) = \max_{\boldsymbol{z} \in [0,1]^N} \{\boldsymbol{z} \cdot \sigma_x(\hat{\boldsymbol{y}})\}$ such that $||\boldsymbol{z}||_0 = K$, where σ_x is the permutation that orders $\hat{\boldsymbol{y}}$ in ascending order of $\boldsymbol{x} = [x_1, \ldots, x_N]$.

Synthetic distribution shift We introduce synthetic distribution shifts to create a scenario for transfer learning. We modify the original feature-label distribution $\mathcal{P} = (\text{Id}, p)_* U[-1, 1]$. Specifically, for various values of $\gamma \in [0, 1.3]$, we define the feature-label distributions $\mathcal{P}_\gamma = (\text{Id}, p_\gamma)_* U[-1, 1]$ where $p_\gamma(x_n) = 10(x_n^3 - \gamma x_n)$, using $\mathcal{P}_{0.65}$ as the target distribution.

Training details We use the implementation from Shah et al. [2022][1] to train models by setting `loss="DFL"`. This implementation uses an entropy regularized Top-K loss function proposed by Xie et al. [2020] that reframes the Top-K problem with entropy regularization as an optimal transport problem, enabling end-to-end learning.

D.2 WARCRAFT SHORTEST PATH Vlastelica et al. [2020]

PtO task description. This setting involves finding the minimum-cost path on $d \times d$ RGB grid maps from the Warcraft II tileset dataset, where each pixel represents terrain with an unknown traversal cost. The task is to first predict these costs from an input image and then determine the shortest path from the top-left to the bottom-right corner based on the predicted cost map. This benchmark is particularly notable because it involves image inputs, a modality not widely explored in other shortest-path learning tasks. Following **?**, we use 96×96 RGB images as input, with the shortest path being computed on a coarser 12×12 grid representation of the predicted costs.

Predict: Given the feature $x_n \in \mathbb{R}^{d \times d \times 3}$, predict the travel cost grid $\hat{y}_n \in \mathbb{R}^{p \times p}$.

Optimize: Solve a shortest-path problem over the predicted cost grid. Specifically, find the path $\boldsymbol{z}$ that minimizes the total traversal cost: $M(\hat{\boldsymbol{y}}) = \min_{\boldsymbol{z} \in [0,1]^p} \{\boldsymbol{z} \cdot \hat{y}\}$ subject to boundary conditions $\boldsymbol{z}_{0,0} = \boldsymbol{z}_{p,p} = 1$ and connectivity constraints ensuring that $\boldsymbol{z}$ represents a valid path from the top-left to the bottom-right corner.

Synthetic distribution shift. The original distribution $\mathcal{P}$, which we treat as the target distribution, is defined over $\mathbb{R}^{d \times d} \times \mathbb{R}^{p \times p}$, where $d = 96$ and $p = 12$. Here, $\mathbb{R}^{d \times d}$ represents the feature space depicting maps, while $\mathbb{R}^{p \times p}$ represents the traveling costs on these maps. We induce a target shift for $\mathcal{P}_\gamma$ by uniformly sampling the costs for different pixel classes from the same range as $\mathcal{P}$ ([0.8, 9.2] for the Warcraft II tileset dataset). Figure 7 illustrates the costs coming from two different distributions over one same feature while highlighting the different decisions (shortest path) that these costs yield.

Training details. We use `pyepo`[2] implementation with SPO+ loss function on a truncated ResNet-18 consisting of the first five layers, followed by a final convolutional layer that reduces the number of output channels to one. Finally, we use an adaptive max-pooling layer to obtain a fixed $p \times p$ spatial resolution, allowing for a structured representation of the extracted features.

[1] github.com/sanketkshah/LODLs
[2] github.com/khalil-research/PyEPO

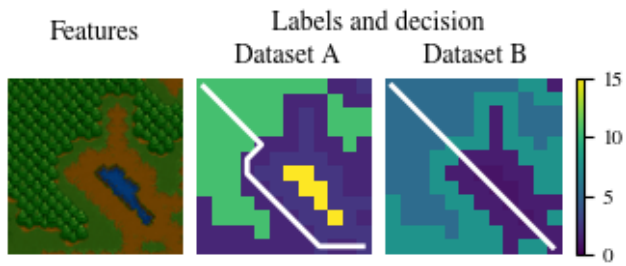


Figure 7: *Synthetic distribution shift in Warcraft Shortest Path*. The white line illustrates the decision, corresponding to the shortest path, on dataset A (center) and dataset B (right) for a sample with the same features (left map).

D.3 INVENTORY STOCK PROBLEM Donti et al. [2017]

PtO task description. In this problem a company must determine the optimal order quantity z of a product to minimize costs given a stochastic demand y, which is influenced by observed features x. The cost structure includes both linear and quadratic costs for the amount of product ordered, as well as different linear and quadratic costs for over-orders $[z-y]^+$ and under-orders $[y-z]^+$. The objective function is:

$$\begin{aligned} f_{\text{stock}}(y,z) =& c_0 z + \frac{1}{2} q_0 z^2 + c_b [y-z]_+ + \frac{1}{2} q_b ([y-z]_+)^2 \\ &+ c_h [z-y]_+ + \frac{1}{2} q_h ([z-y]_+)^2 \end{aligned} \tag{19}$$

where $[v]_+ \equiv \max\{v, 0\}$. In our paper, we use $c_0 = 30, q_0 = 10, c_b = 10, q_b = 2, c_h = 30, q_h = 25$. For a given probability model $p(y|x;\theta)$, the proxy stochastic programming problem can be formulated as: $\underset{z}{\text{minimize}} \quad \mathbf{E}_{y \sim p(y|x;\theta)} \left[f_{\text{stock}}(y,z) \right]$.

To simplify the setting, we assume that the demands are discrete, taking on values $d_1, \ldots, d_k$ with probabilities (conditional on x) $(p_\theta)_i \equiv p(y = d_i | x; \theta)$. Thus, our stochastic programming problem can be succinctly expressed as a joint quadratic program:

$$\begin{aligned} \underset{z \in \mathbb{R}, z_b, z_h \in \mathbb{R}^k}{\text{minimize}} \Big\{ & c_0 z + \frac{1}{2} q_0 z^2 + \sum_{i=1}^{k} (p_\theta)_i \big(c_b (z_b)_i \\ & + \frac{1}{2} q_b (z_b)_i^2 + c_h (z_h)_i + \frac{1}{2} q_h (z_h)_i^2 \big) \Big\} \\ \text{subject to} \quad & d - z\mathbf{1} \le z_b, \quad z\mathbf{1} - d \le z_h, \quad z, z_h, z_b \ge 0 \end{aligned} \tag{10}$$

Synthetic distribution shift We generate problem instances by randomly sampling $x \in \mathbb{R}^n$ and then generating $p(y|x;\theta)$ according to $p(y|x;\theta) \propto \exp\big((\theta^T x)^2\big)$. We introduce distribution shifts for both x and y. Specifically, x is sampled from a Gaussian distribution where the mean is sampled from $U[-0.5, 0.5]$, and θ is also sampled from a Gaussian distribution.

Training details We use the implementation from Donti et al. [2017][3] following their Inventory Stock Problem experiments.

E OTD^3 IMPLEMENTATION DETAILS

Our implementation of the OTD^3 relies on the `POT`[4] package. The computation of dataset distance involves two main steps:

1. **Computing Pairwise Pointwise Distances:** We first compute the pairwise distances between samples in the source and target datasets. This involves calculating distances separately for features, labels, and decisions, weighted according to the selected component weights $(\alpha_X, \alpha_Y, \alpha_W)$. Feature and label distances are computed using standard metric spaces (e.g., Euclidean or cosine distance), while decision distances are computed using decision quality disparity.

[3] github.com/locuslab/e2e-model-learning
[4] pythonot.github.io/

2. **Solving the Optimal Transport Problem:** Given the computed pairwise distances, we compute the dataset distance using Earth Mover's Distance (EMD) via `POT`'s `emd` solver. EMD finds the exact optimal transport plan, making it well-suited for capturing true correspondences between source and target datasets without introducing regularization bias. This approach was computationally feasible in our experiments due to the relatively small dataset sizes.

Additionally, for experiments involving hyperparameter tuning, we evaluate multiple weight combinations on a predefined grid and select the setting that maximizes correlation with regret transferability.

F ADDITIONAL RESULTS

F.1 SELECTING SOURCE DATASETS FOR TRANSFER LEARNING

In Section 7.1 we analyzed the correlation between dataset distance and transferability in PtO. The plots presented in Figure 8 show this correlation for the Linear Model TopK setting and the Inventory Stock problem under two weighting profiles: one where decision-related features are excluded (left) and one where they are included (right). In both settings, incorporating decisions into the distance metric leads to improved predictability of transfer performance. This effect is more pronounced in the Linear Model TopK task than in the Inventory Stock problem.

For these settings, we do not perform fine-tuning on the target dataset. Instead, we assess transferability in a zero-shot setting, where a model trained on the source dataset is directly applied to the target domain without further adaptation. This choice is motivated by the relative simplicity of the feature spaces involved, which enables a meaningful evaluation of dataset distances without introducing potential confounding effects from additional training steps. Accordingly, rather than plotting dataset distance against the relative drop in regret after fine-tuning, we plot it against $\mathcal{T}(S \to T) = (\text{reg}(\mathcal{D}_S) - \text{reg}(\mathcal{D}_T))/\text{reg}(\mathcal{D}_T)$, where $\text{reg}(\mathcal{D}_S)$ denotes the decision regret when applying the source-trained model to the target dataset, and $\text{reg}(\mathcal{D}_S)$ is the regret of a model trained directly on the target.

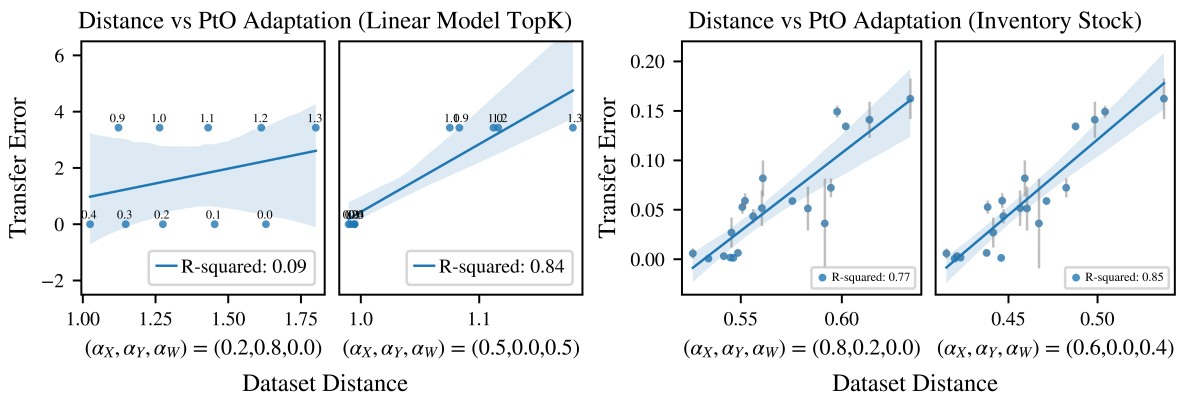
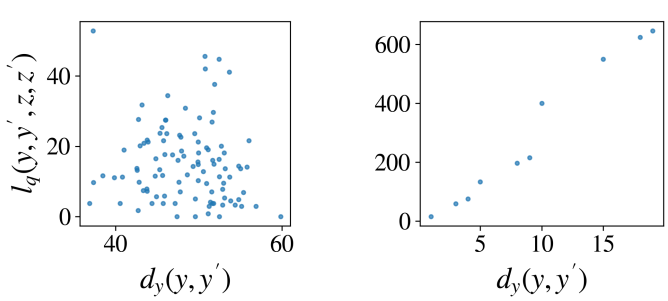


Figure 8: *Distance vs Adaptation*. OT distance for the best feature-label and feature-label-decision weighting against regret transferability.



(a) Warcraft Shortest Path

(b) Inventory Stock

Figure 9: Difference in labels against difference in decisions.

To illustrate the relationship between label space differences $d_y(y, y')$ and decision space differences $l_q(y, y', z, z')$ in different PtO tasks, we provide the following visualizations. Figure F.1 shows this correlation for the Inventory Stock problem, while Figure F.1 presents the same analysis for the Warcraft domain.