Proceedings of the 2025 Winter Simulation Conference
E. Azar, A. Djanatliev, A. Harper, C. Kogler, V. Ramamohan, A. Anagnostou, and S. J. E. Taylor, eds.

DISTRIBUTIONALLY ROBUST LOGISTIC REGRESSION WITH MISSING DATA

Weicong Chen[1], and Hoda Bidkhori[1]

[1]Dept. of Computational and Data Sciences, George Mason University, Fairfax, VA, USA

ABSTRACT

Missing data presents a persistent challenge in machine learning. Conventional approaches often rely on data imputation followed by standard learning procedures, typically overlooking the uncertainty introduced by the imputation process. This paper introduces Imputation-based Distributionally Robust Logistic Regression (I-DRLR)—a novel framework that integrates data imputation with class-conditional Distributionally Robust Optimization under the Wasserstein distance. I-DRLR explicitly models distributional ambiguity in the imputed data and seeks to minimize the worst-case logistic loss over the resulting uncertainty set. We derive a convex reformulation to enable tractable optimization and evaluate the method on the Breast Cancer and Heart Disease datasets from the UCI Repository. Experimental results demonstrate consistent improvements for out-of-sample performance in both prediction accuracy and ROC-AUC, outperforming traditional methods that treat imputed data as fully reliable.

1 INTRODUCTION

This paper presents a distributionally robust approach to address the issue of missing data in logistic regression, a widely used statistical method for classification. The motivation stems from real-world applications—such as healthcare analytics, industrial monitoring, traffic management, and customer behavior modeling—where datasets often contain incomplete feature observations due to various factors (Ehrlinger et al. 2018; Noh et al. 2004; Preda et al. 2005; Smith et al. 2003). These factors include sensor errors, communication delays, or data privacy restrictions.

Traditional approaches to handling missing data in machine learning models, including logistic regression, aim to construct a complete dataset through imputation methods such as mean substitution, maximum likelihood estimation, K-Nearest Neighbors (KNN), or multiple imputation (Austin and van Buuren 2022; Dempster et al. 1977; Martins et al. 2024; Jerez et al. 2010; Verchand and Montanari 2024), followed by standard model training. However, these approaches ignore the uncertainty introduced during the imputation process, which can lead to unreliable predictions (Schafer and Graham 2002; Zhang 2016).

In this paper, we propose a distributionally robust method that accounts for the uncertainty in imputed values, resulting in more reliable forecasts. Our approach incorporates distributional ambiguity in the imputed data distribution and minimizes the worst-case logistic loss within this ambiguity set.

Robust methods have been extensively used to tackle uncertainty in machine learning models, particularly in scenarios involving distributional shift and noise (Bertsimas et al. 2019; Blanchet et al. 2020). Among these, Distributionally Robust Optimization (DRO) has gained attention for its ability to hedge against distributional uncertainty by optimizing for worst-case performance within an ambiguous set of possible data distributions and utilized for machine learning models including logistic regression (Chen and Paschalidis 2018; Faccini et al. 2022; Lee and Mehrotra 2015; Shafieezadeh-Abadeh et al. 2015). While DRO has proven effective in handling distributional uncertainty, it has not yet been applied to model the uncertainty arising from imputed datasets—an issue that frequently occurs in real-world applications.

We propose a novel integrated framework that considers distributional uncertainty around the imputed data and leverages DRO to optimize the worst-case logistic loss in the corresponding ambiguity set. Rather than completely relying on empirical imputed data distribution, our method constructs class-conditional

979-8-3315-8726-0/25/$31.00 ©2025 IEEE

ambiguity sets—defined using the Wasserstein distance—centered around empirical distributions generated via imputation. The DRO formulation then seeks to minimize the worst-case logistic loss over these sets of data distributions, allowing for controlled perturbations in the feature space while preserving class labels (Delage and Ye 2010; Gao and Kleywegt 2016; Kuhn et al. 2024). This enables the model to hedge against distributional uncertainty introduced by the imputation process.

Contributions. The main contributions of this paper are as follows:

1. We propose a novel *Imputation-based Distributionally Robust Logistic Regression (I-DRLR)* framework that hedges against uncertainty arising from the imputation process. We consider missing data in the feature space and construct separate Wasserstein ambiguity sets for each class. The model then solves the worst-case logistic loss for each corresponding set. The framework is flexible, allowing control over the level of uncertainty and choice of ambiguity set.
2. We derive a tractable reformulation of the I-DRLR problem and prove its equivalence to a convex optimization problem. This enables efficient solutions using standard convex optimization libraries such as *CVXPY* and commercial solvers like Gurobi and CPLEX.
3. We conduct extensive computational experiments on the *Breast Cancer Wisconsin (Diagnostic)* and *Heart Disease* datasets from the UCI Machine Learning Repository (Wolberg et al. 1993; Janosi et al. 1989). Across varying missing data probabilities and robustness radii, I-DRLR consistently provides better out-of-sample performance than standard imputation-based methods in terms of both *prediction accuracy* and *ROC-AUC*. We also analyze the effect of ambiguity set design on generalization performance.

The remainder of the paper is organized as follows. Section 2 reviews related work on missing data, robust statistics, and DRO in machine learning. Section 3 provides the necessary background. Section 4 introduces our I-DRLR model based on class-conditional Wasserstein ambiguity sets. Section 5 presents a tractable convex reformulation for scalable implementation. Section 6 reports experimental results on real-world datasets from the UCI Machine Learning Repository under varying missing data probabilities and robustness levels. The Appendix includes the proof of the reformulation.

2 RELATED WORK

Missing data is a critical issue in machine learning and has been extensively studied in the literature. The most common strategy for addressing missing data is imputation, which involves estimating and filling in the missing values based on observed information. Various imputation techniques have been proposed, including single imputation, multiple imputation, maximum likelihood estimation, and K-Nearest Neighbors, among others (Austin and van Buuren 2022; Emmanuel et al. 2021; Martins et al. 2024; Jerez et al. 2010; Verchand and Montanari 2024; Zhang 2016).

Data uncertainty is another major challenge in machine learning applications, and ensuring robustness is essential for obtaining reliable predictions. As a result, robust methods have been widely studied in the field. Much of the existing literature on data uncertainty focuses on variations in the underlying data distribution, such as distributional shifts, noise, or adversarial perturbations (Bertsimas et al. 2019; Blanchet et al. 2020; Caramanis et al. 2012). DRO is a powerful framework that optimizes for the worst-case expected loss over an ambiguity set of possible data distributions (Delage and Ye 2010; Gao and Kleywegt 2016; Kuhn et al. 2024). DRO has been successfully applied to various machine learning models, including linear regression, support vector machines (SVM), and logistic regression (Blanchet et al. 2020; Blanchet et al. 2024; Chen and Paschalidis 2018; Duchi and Namkoong 2018; Faccini et al. 2022; Kuhn et al. 2019; Shafieezadeh-Abadeh et al. 2015; Taşkesen et al. 2020). While DRO methods in machine learning have been applied to address uncertainties arising from distributional shifts, noise, and adversarial perturbations, they have, to the best of our knowledge, not been utilized to tackle the issue of missing data in machine learning models.

This paper contributes to the literature on missing data in logistic regression by explicitly incorporating uncertainty in the imputed data distribution (Austin and van Buuren 2022; Jiang et al. 2018; Verchand and Montanari 2024). We propose a novel approach that integrates imputation with DRO to handle datasets containing missing values. Specifically, we introduce the *Imputation-based Distributionally Robust Logistic Regression (I-DRLR)* method, which combines imputation with a class-conditional DRO framework based on the Wasserstein distance. We further derive a tractable reformulation of the I-DRLR problem as a convex optimization, enabling efficient implementation using standard optimization solvers. Finally, we conduct comprehensive experiments to evaluate the performance of our method relative to conventional logistic regression, with a focus on out-of-sample prediction accuracy and ROC-AUC.

3 PRELIMINARY AND BACKGROUND

Standard logistic regression estimates class label probabilities based on feature vectors. Let the binary label be denoted by y. The model uses a feature vector $\mathbf{x} \in \mathbb{R}^d$, where $\mathbb{R}^d$ represents a d-dimensional real vector space. The relationship between predictors and outcomes is described using a coefficient vector $\beta \in \mathbb{R}^d$, with class probabilities estimated via the logistic function:

$$\text{Prob}(y \mid \mathbf{x}) = \left[1 + \exp\left(-y\mathbf{x}^\top \beta\right)\right]^{-1}. \tag{1}$$

In (1), the optimal coefficient vector β is found by minimizing the total logistic loss over the entire dataset. The logistic loss function is defined as follows:

$$\ell_\beta(\mathbf{x}, y) = \log\left(1 + \exp\left(-y\mathbf{x}^\top \beta\right)\right). \tag{2}$$

This formulation arises from the principle of maximum likelihood estimation under the assumption of independently and identically distributed (i.i.d.) data and a logistic loss for binary outcomes (Hastie et al. 2009). Considering the empirical distribution $\hat{\mathbf{P}}$, standard logistic regression can then be written as the following optimization problem:

$$\min_\beta \mathbb{E}_{\hat{\mathbf{P}}}[\ell_\beta(\mathbf{x}, y)] = \min_\beta \frac{1}{n} \sum_{i=1}^{n} \log(1 + \exp(-y_i \cdot \beta^\top \mathbf{x}_i)). \tag{3}$$

Assuming that our dataset has $n = n_0 + n_1$ i.i.d. samples, where n_0 and n_1 represent the number of training samples with labels 0 and 1, respectively. Let $\hat{\mathbf{P}}_0$ denote the empirical distribution of samples with $y = 0$, and $\hat{\mathbf{P}}_1$ the empirical distribution for $y = 1$. Each of these contributes to the overall empirical distribution $\hat{\mathbf{P}}$. Let $\mathbf{x}_i^{(0)}$ and $\mathbf{x}_i^{(1)}$ denote the i^{th} samples from classes 0 and 1, respectively. Define $\hat{p}_0 = n_0/n$ and $\hat{p}_1 = n_1/n$ as the empirical class proportions. Then, (3) can be equivalently expressed as:

$$\hat{p}_0 \cdot \mathbb{E}_{\hat{\mathbf{P}}_0}[\ell_\beta(\mathbf{x}, 0)] + \hat{p}_1 \cdot \mathbb{E}_{\hat{\mathbf{P}}_1}[\ell_\beta(\mathbf{x}, 1)] = \hat{p}_0 \cdot \frac{1}{n_0} \sum_{i=1}^{n_0} \ell_\beta(\mathbf{x}_i^{(0)}, 0) + \hat{p}_1 \cdot \frac{1}{n_1} \sum_{i=1}^{n_1} \ell_\beta(\mathbf{x}_i^{(1)}, 1). \tag{4}$$

Missing Data Assumption. In this paper, we assume that only the feature variables contain missing values—regardless of the underlying missingness mechanism—while the labels are fully observed (Josse and Reiter 2018; Schafer and Graham 2002).

Distributionally Robust Optimization. DRO is a framework for decision-making under uncertainty that seeks solutions robust to distributional ambiguity. Rather than optimizing for a fixed probability distribution, DRO optimizes for the worst-case expected loss over a set of possible distributions (Delage and Ye 2010; Gao and Kleywegt 2016; Kuhn et al. 2024). There are several ways to define this ambiguity set; in our case, we use a commonly used metric, the Wasserstein distance, which quantifies the cost of transporting mass to transform one probability distribution into another. It effectively measures the optimal transport cost between distributions (Villani 2008). In the following section, we present our *Imputation-based Distributionally Robust Logistic Regression (I-DRLR)* method, which applies the DRO framework to the distribution of imputed data. This is achieved by constructing class-conditional Wasserstein ambiguity sets to capture uncertainty in the imputed feature values.

4 MAIN MODEL

In this section, we employ a DRO approach to minimize the worst-case expected logistic loss across all distributions within the ambiguity sets of possible shifts of the imputed data distribution. We consider the distributions that are close to the imputed empirical distribution for each class, allowing perturbations only in the feature space while keeping labels fixed.

Recall $\hat{\mathbf{P}}_y$ denotes the empirical distribution of features with label $y \in \{0,1\}$, and let $\hat{p}_y$ be the empirical class prior, which is the proportion of training samples that belong to class y in the empirical dataset. For each class label y, we define a class-conditional ambiguity set with a shared Wasserstein radius $\tau > 0$:

$$\mathscr{P}_y = \left\{ \mathbf{P}_y \in \mathscr{P}(\mathbb{X}) : W(\mathbf{P}_y, \hat{\mathbf{P}}_y) \leq \tau \right\}, \tag{5}$$

where $\mathscr{P}(\mathbb{X})$ denotes the set of all probability distributions supported on the feature space $\mathbb{X} \subseteq \mathbb{R}^d$, and $W(\cdot,\cdot)$ denotes the Wasserstein distance over $\mathbb{X}$. The radius $\tau > 0$ is tunable and may be adjusted based on the problem context—for example, higher missing rates may suggest a larger uncertainty radius.

While any valid distance can define the ambiguity set, we use the Wasserstein distance, a widely adopted and foundational metric in distributionally robust optimization. The Wasserstein distance between two distributions $\mathbf{P}_y$ and $\hat{\mathbf{P}}_y$ is defined as:

$$W(\mathbf{P}_y, \hat{\mathbf{P}}_y) = \inf_{\pi \in \Pi(\mathbf{P}_y, \hat{\mathbf{P}}_y)} \mathbb{E}_{(\mathbf{x},\mathbf{x}') \sim \pi} \left[\|\mathbf{x} - \mathbf{x}'\| \right], \tag{6}$$

where $\Pi(\mathbf{P}_y, \hat{\mathbf{P}}_y)$ denotes the set of all couplings with marginals $\mathbf{P}_y$ and $\hat{\mathbf{P}}_y$, and $\|\cdot\|$ is a norm (e.g., $\ell_1, \ell_2, \ell_\infty$ norm) on the feature space $\mathbb{R}^d$. Intuitively, this formulation seeks the minimum expected cost of transporting "mass" from one distribution to another (Villani 2008). Formulation (6) ensures that perturbations are allowed only in the feature space, and the label remains fixed.

The following is a distributionally robust formulation associated to problem (3). Since missing data and uncertainty occur only in the feature space, we adopt a class-conditional distributionally robust optimization approach, using the ambiguity sets defined in (5). The resulting *Imputation-based Distributionally Robust Logistic Regression (I-DRLR)* problem is formulated as follows:

$$\min_{\beta} \left\{ \hat{p}_0 \cdot \max_{\mathbf{P}_0 \in \mathscr{P}_0} \mathbb{E}_{\mathbf{x} \sim \mathbf{P}_0} \left[\ell_\beta(\mathbf{x}, 0) \right] + \hat{p}_1 \cdot \max_{\mathbf{P}_1 \in \mathscr{P}_1} \mathbb{E}_{\mathbf{x} \sim \mathbf{P}_1} \left[\ell_\beta(\mathbf{x}, 1) \right] \right\} \tag{7}$$

where $\ell_\beta(\mathbf{x}, y)$ is the logistic loss function defined in the previous section. Although the Wasserstein DRO formulation of logistic regression can be reduced to regularized logistic regression under specific conditions, this equivalence does not hold in general, and it does not apply in our setting where we have utilized class-conditional ambiguity sets (Wu et al. 2022). More importantly, regularized logistic regression does not explicitly account for the uncertainty introduced by imputation. In contrast, Wasserstein DRO directly models distributional ambiguity in the feature space, providing robustness against worst-case perturbations space (Shafieezadeh-Abadeh et al. 2015).

5 TRACTABLE REFORMULATION

In this section, we present a tractable reformulation of the I-DRLR problem (7), influenced by the ideas introduced in (Shafieezadeh-Abadeh et al. 2015). The reformulated optimization problem is convex for most commonly used norms, and can be efficiently solved using standard convex optimization solvers such as those in *CVXPY*, a standard convex optimization library. A complete proof of the reformulation is provided in the Appendix.

Theorem 1 (Tractable Reformulation of I-DRLR) Let $\hat{\mathbf{x}}_i \in \mathbb{R}^d$ and $\hat{y}_i \in \{0,1\}$ denote the imputed feature vector and the corresponding observed label. Then, the Imputation-based Distributionally Robust Logistic Regression (I-DRLR) problem (7) is equivalent to the following convex optimization problem:

$$\begin{aligned}
\min_{\beta,\lambda_0,\lambda_1,s_i^{(0)},s_i^{(1)}} \quad & \hat{p}_0 \left(\lambda_0 \tau + \frac{1}{n_0} \sum_{i=1}^{n_0} s_i^{(0)} \right) + \hat{p}_1 \left(\lambda_1 \tau + \frac{1}{n_1} \sum_{i=1}^{n_1} s_i^{(1)} \right) \\
\text{s.t.} \quad & \ell_\beta(\hat{\mathbf{x}}_i^{(0)}, 0) \leq s_i^{(0)}, \quad \forall i = 1, \ldots, n_0, \\
& \ell_\beta(\hat{\mathbf{x}}_i^{(1)}, 1) \leq s_i^{(1)}, \quad \forall i = 1, \ldots, n_1, \\
& \|\beta\|_* \leq \lambda_0, \\
& \|\beta\|_* \leq \lambda_1.
\end{aligned} \tag{8}$$

In the above formulation, n_0 and n_1 are the number of training samples with labels 0 and 1. The quantities $\hat{p}_0$ and $\hat{p}_1$ are the empirical class proportions defined earlier. The vector $\beta \in \mathbb{R}^d$ represents the parameters of the logistic regression model, while the scalar variables $s_i^{(0)}$ and $s_i^{(1)}$ serve as upper bounds on the logistic loss for each class-specific training example $(\hat{\mathbf{x}}_i^{(0)}, 0)$ and $(\hat{\mathbf{x}}_i^{(1)}, 1)$. The dual variables λ_0 and λ_1 represent the model's sensitivity to perturbations in the class-conditional distributions. Lastly, $\|\cdot\|_*$ denotes the dual norm associated with the norm used in defining the Wasserstein distance.

6 EXPERIMENTAL RESULTS

In this section, we evaluate the effectiveness of the proposed I-DRLR method by comparing its out-of-sample performance against Imputation-based Logistic Regression (I-LR). We conduct a series of experiments under varying missing rates and uncertainty levels. In addition, we evaluate the performance of I-DRLR under different ambiguity set configurations. Across all settings, the I-DRLR approach consistently demonstrates superior performance in the presence of incomplete data. Our experiments are conducted on two benchmark datasets from the UCI Machine Learning Repository: the *Breast Cancer Wisconsin (Diagnostic)* dataset (Wolberg et al. 1993) and the *Heart Disease* dataset (Janosi et al. 1989). We first describe the experimental setup in detail, followed by the presentation of results in Subsections 6.1 and 6.2 for the two datasets. The result for comparison under different uncertainty sets is presented in Subsection 6.3.

Experiment Setup. Since the *Breast Cancer Wisconsin (Diagnostic)* dataset and the *Heart Disease* dataset are fully observed, we introduce missingness by randomly removing feature values at predefined rates. Specifically, 10% and 20% of all feature entries are removed uniformly at random across all feature dimensions. The imputed datasets utilized in I-LR and I-DRLR are then obtained using a K-Nearest Neighbors (KNN) imputer with $k = 5$ neighbors. All numerical features are subsequently standardized to ensure consistent scaling. Finally, each dataset is split into an 80%-20% training-test partition. Both I-LR and the proposed I-DRLR model are trained on the imputed training data. I-DRLR is formulated as a convex optimization problem that minimizes worst-case logistic loss over Wasserstein balls of radius $\tau \in \{0.05, 0.10, 0.15\}$. We use the same τ value for both class-conditional ambiguity sets as defined in the previous section. The norm used for the Wasserstein distance is the ℓ_2 norm, and the optimization is solved using the *CVXPY*, a python-based standard convex optimization library. To ensure the reliability of the results, we repeated the entire experimental procedure 30 times. In each iteration, missing values are introduced randomly and independently. The models are then retrained, and the out-of-sample performance is recorded. The out-of-sample performance of the models is evaluated using two standard metrics: *Accuracy* and *ROC-AUC score*, which measures the model's ability to distinguish between classes by calculating the area under the ROC curve. The ROC curve plots the true positive rate against the false positive rate. In addition to evaluating the performance on different robustness radii τ, we also experiment with different norm choices for the Wasserstein distance, including ℓ_1 and ℓ_2 norms. This allows us to study the effect of different uncertainty sets on I-DRLR performance under incomplete data. The following sections discuss the results of our experiments.

6.1 Results for the Breast Cancer Dataset

In this section, we compare the performance of I-LR and I-DRLR on the *Breast Cancer Wisconsin (Diagnostic)*, which contains 569 samples with 30 numerical characteristics representing cell nuclei characteristics from digitized images. The binary target indicates whether a tumor is malignant (1) or benign (0). We compare the out-of-sample performance across different missing probabilities, 10% and 20%. For each missing probability and uncertainty level, we repeated the experimental procedure 30 times to ensure the reliability of the results. The out-of-sample accuracy and ROC-AUC results for 10% missing probability are summarized using boxplots in Figure 1. Mean and standard deviation statistics are summarized in Tables 1 and 2. The results show that I-DRLR consistently outperforms I-LR in both accuracy and ROC-AUC. With increasing values of τ, I-DRLR produces more stable and reliable predictions, as evidenced by higher medians and tighter interquartile ranges in the boxplots. Tables 3 and 4, along with Figure 2, reaffirm the above observation for the 20% missing probability. In addition, they reveal that the performance gap between I-DRLR and I-LR widens as the missing probability increases.

Table 1: Accuracy with 10% missing probability.

Method	Mean	Std Dev
I-LR: I-DRLR($\tau = 0$)	0.900292	0.022744
I-DRLR ($\tau = 0.05$)	0.924854	0.021694
I-DRLR ($\tau = 0.1$)	0.933041	0.017466
I-DRLR ($\tau = 0.15$)	0.933041	0.018641

Table 2: ROC-AUC with 10% missing probability.

Method	Mean	Std Dev
I-LR: I-DRLR($\tau = 0$)	0.971427	0.010656
I-DRLR ($\tau = 0.05$)	0.983535	0.008318
I-DRLR ($\tau = 0.1$)	0.985610	0.008251
I-DRLR ($\tau = 0.15$)	0.986341	0.008117

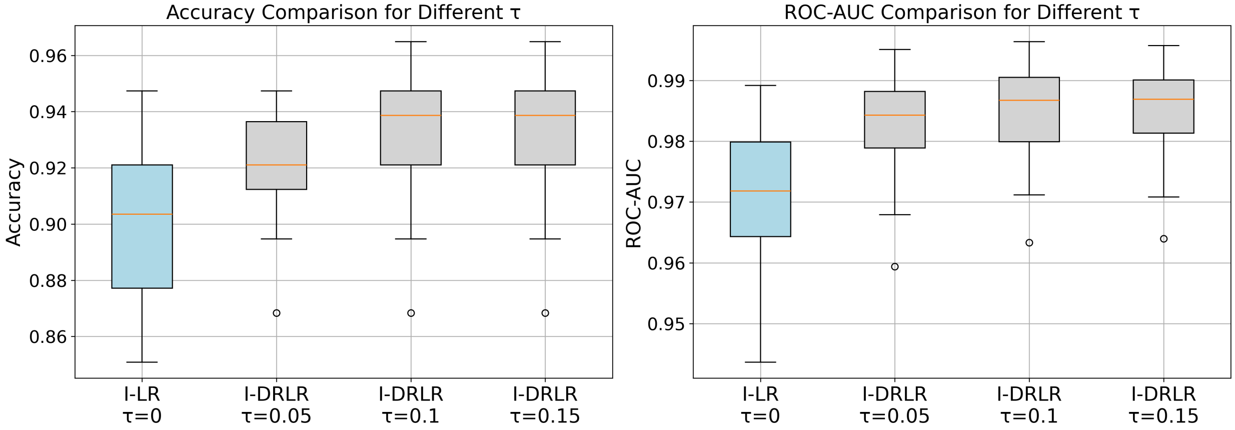

Figure 1: Accuracy & ROC-AUC with 10% missing probability.

Table 3: Accuracy with 20% missing probability.

Method	Mean	Std Dev
I-LR: I-DRLR($\tau = 0$)	0.896784	0.033278
I-DRLR ($\tau = 0.05$)	0.923684	0.024720
I-DRLR ($\tau = 0.1$)	0.929825	0.023265
I-DRLR ($\tau = 0.15$)	0.932749	0.020775

Table 4: ROC-AUC with 20% missing probability.

Method	Mean	Std Dev
I-LR: I-DRLR($\tau = 0$)	0.970816	0.012733
I-DRLR ($\tau = 0.05$)	0.985490	0.006793
I-DRLR ($\tau = 0.1$)	0.987258	0.006312
I-DRLR ($\tau = 0.15$)	0.987804	0.005887

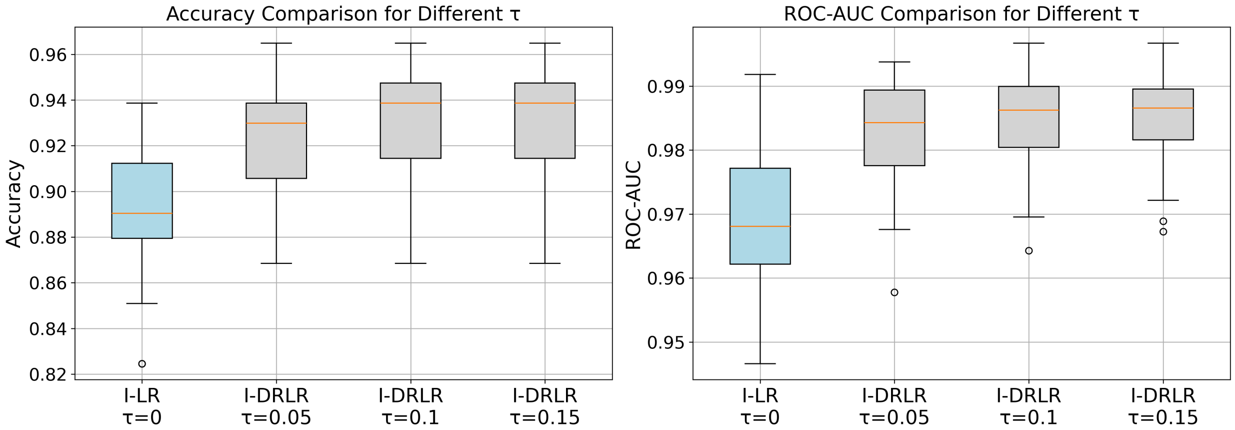

Figure 2: Accuracy & ROC-AUC with 20% missing probability.

6.2 Results for the Heart Disease Dataset

In this section, we apply the same experimental procedure to the *Heart Disease* dataset, which consists of 303 patient records. The binary target indicates the presence (1) or absence (0) of heart disease. Performance results for the 10% missing-data scenario are visualized as boxplots in Figure 3, while the corresponding mean and standard deviation statistics are summarized in Tables 5 and 6. As illustrated, I-DRLR consistently outperforms I-LR across all evaluation metrics. As τ increases, I-DRLR shows improved out-of-sample performance. We replicated the experiment with a 20% missing probability, and again, I-DRLR consistently outperforms I-LR across all metrics. Figure 4 further highlights that under 20% missingness, I-DRLR exhibits greater stability and superior performance.

Table 5: Accuracy with 10% missing probability.

Method	Mean	Std Dev
I-LR: I-DRLR($\tau = 0$)	0.761749	0.040324
I-DRLR ($\tau = 0.05$)	0.775410	0.036314
I-DRLR ($\tau = 0.1$)	0.784699	0.033826
I-DRLR ($\tau = 0.15$)	0.781967	0.036821

Table 6: ROC-AUC with 10% missing probability.

Method	Mean	Std Dev
I-LR: I-DRLR($\tau = 0$)	0.838578	0.032293
I-DRLR ($\tau = 0.05$)	0.860776	0.026883
I-DRLR ($\tau = 0.1$)	0.870833	0.025391
I-DRLR ($\tau = 0.15$)	0.872701	0.024471

Table 7: Accuracy with 20% missing probability.

Method	Mean	Std Dev
I-LR: I-DRLR($\tau = 0$)	0.777596	0.040313
I-DRLR ($\tau = 0.05$)	0.774317	0.034624
I-DRLR ($\tau = 0.1$)	0.780328	0.033236
I-DRLR ($\tau = 0.15$)	0.785246	0.035331

Table 8: ROC-AUC with 20% missing probability.

Method	Mean	Std Dev
I-LR: I-DRLR($\tau = 0$)	0.845546	0.028995
I-DRLR ($\tau = 0.05$)	0.857471	0.026922
I-DRLR ($\tau = 0.1$)	0.864583	0.028858
I-DRLR ($\tau = 0.15$)	0.865445	0.029422

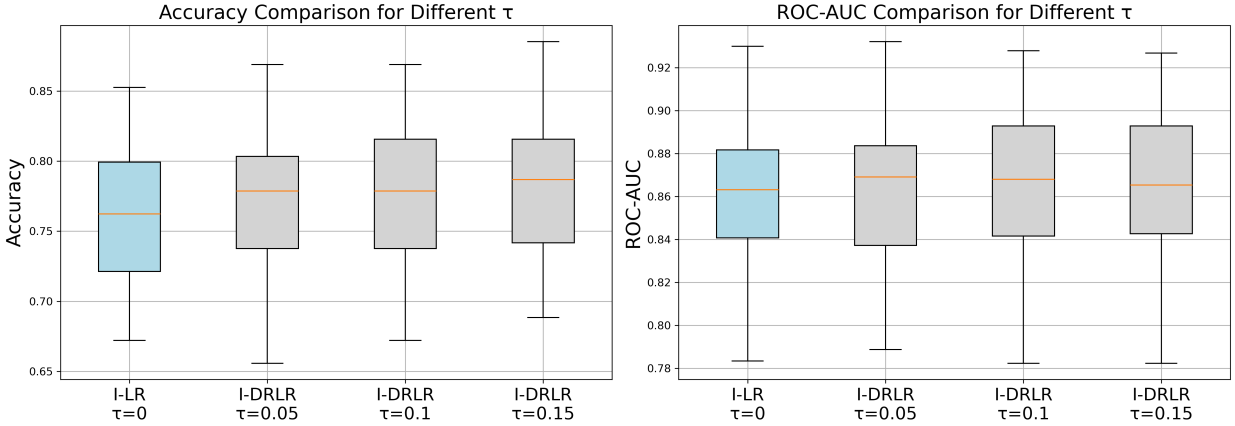

Figure 3: Accuracy & ROC-AUC with 10% missing probability.

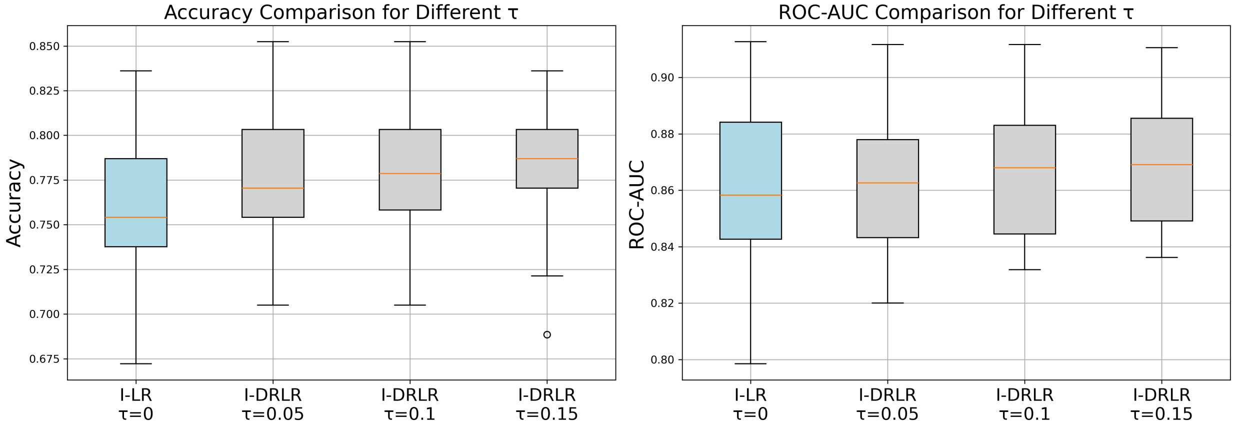

Figure 4: Accuracy & ROC-AUC with 20% missing probability.

6.3 Comparison under Different Ambiguity Sets

In this section, we compare the out-of-sample performance of I-LR and I-DRLR on the *Heart Disease* dataset for different norms that are incorporated in the Wasserstein ambiguity set (ℓ_1, ℓ_2). The performance comparisons under 10% and 20% missing probabilities and different robustness parameters ($\tau = 0.05, 0.1, 0.15$) are visualized in Figures 5 and 6. As illustrated in Figure 5, I-DRLR models outperform I-LR on the *Heart Disease* dataset with 10% missing probability for both ℓ_1 and ℓ_2 norms. These results consistently show improvements in both accuracy and ROC-AUC. I-DRLR with the ℓ_2 norm and a larger τ stands out with the most favorable results. We increase the missing probability to 20% and compare the results. Figure 6 shows that I-DRLR with ℓ_1 and ℓ_2 norms generally outperform I-LR, especially in ROC-AUC. The best performance is observed with the ℓ_2 norm at $\tau = 0.15$, which improves both accuracy and ROC-AUC.

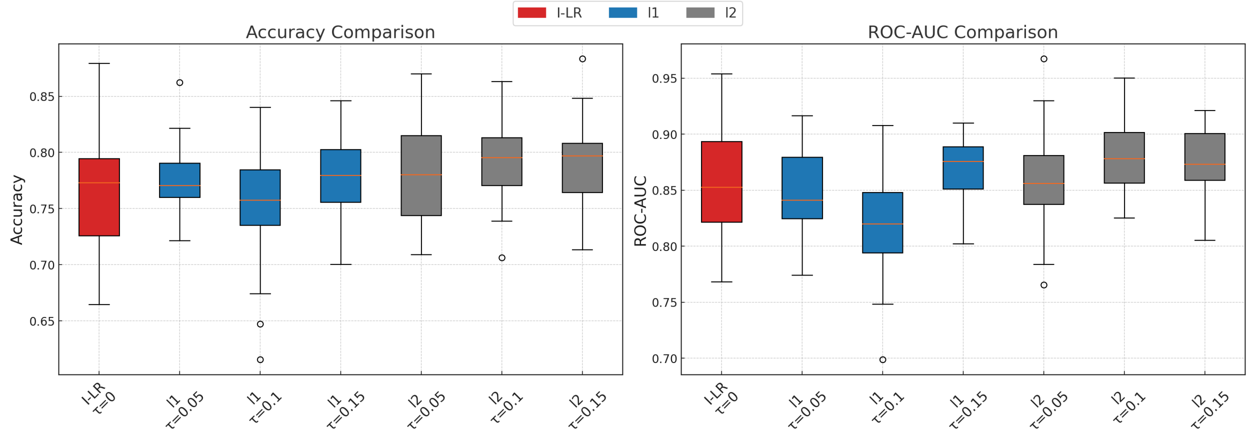

Figure 5: Accuracy & ROC-AUC on different norms with 10% missing probability.

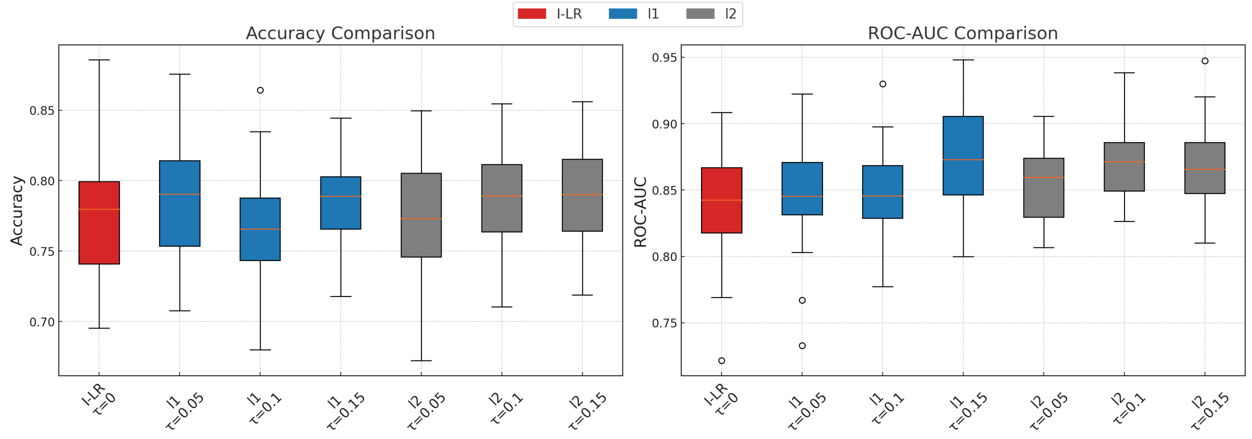

Figure 6: Accuracy & ROC-AUC on different norms with 20% missing probability.

7 CONCLUSION

This paper proposes Imputation-based Distributionally Robust Logistic Regression (I-DRLR), a principled framework that integrates data imputation with class-conditional Distributionally Robust Optimization (DRO) using the Wasserstein distance. I-DRLR explicitly models uncertainty in the imputed feature space by constructing class-wise ambiguity sets centered on the empirical imputed distributions. Empirical evaluations on benchmark datasets with induced missing data mechanism show that I-DRLR consistently outperforms Imputation-based Logistic Regression (I-LR) in both predictive accuracy and ROC-AUC, particularly under higher levels of missing probability and robustness radii. These results underscore the value of incorporating distributional robustness into learning procedures with missing data and motivate further exploration of robust methods that address imputation uncertainty. The proposed I-DRLR method is designed to operate under the assumption of a completely unknown missing data mechanism. However, more specialized models could be developed when partial knowledge of the missing mechanism is available. Exploring theoretical performance guarantees also remains a valuable direction for future research. While the current framework assumes fully observed labels, it can extend for settings involving partially observed labeled data. Moreover, future work could investigate alternative robust learning models (Blanchet et al. 2024)—beyond the Wasserstein distributionally robust optimization framework to effectively handle missing data.

8 APPENDIX

8.1 Proof for Theorem 1

Proof. We consider the Imputation-based Distributionally Robust Logistic Regression (I-DRLR) problem under the class-conditional Wasserstein ambiguity sets over the feature space $\mathbb{R}^d$, using a shared Wasserstein radius $\tau > 0$. For each class $y \in \{0,1\}$, the ambiguity set is defined as:

$$\mathscr{P}_y = \left\{ \mathbf{P}_y \in \mathscr{P}(\mathbb{R}^d) : W(\mathbf{P}_y, \hat{\mathbf{P}}_y) \le \tau \right\},$$

where $\hat{\mathbf{P}}_y$ is the empirical distribution of features for class y, and $W(\cdot,\cdot)$ is the Wasserstein distance defined as:

$$W(\mathbf{P}, \hat{\mathbf{P}}) = \inf_{\pi \in \Pi(\mathbf{P}, \hat{\mathbf{P}})} \int_{\mathbb{R}^d \times \mathbb{R}^d} \|\mathbf{x} - \mathbf{x}'\| \, \mathrm{d}\pi(\mathbf{x}, \mathbf{x}'),$$

with $\|\cdot\|$ a norm on $\mathbb{R}^d$, and $\Pi(\mathbf{P}, \hat{\mathbf{P}})$ the set of all couplings with marginals $\mathbf{P}$ and $\hat{\mathbf{P}}$.

Let $\ell_\beta(\mathbf{x}, y) = \log(1 + \exp(-y \cdot \beta^\top \mathbf{x}))$ denote the logistic loss. The I-DRLR objective is:

$$\min_{\beta \in \mathbb{R}^d} \sum_{y \in \{0,1\}} \hat{p}_y \sup_{\mathbf{P}_y \in \mathscr{P}_y} \mathbb{E}_{\mathbf{x} \sim \mathbf{P}_y} \left[\ell_\beta(\mathbf{x}, y) \right].$$

Fix a class $y \in \{0,1\}$. The robust expected loss over $\mathscr{P}_y$ can be expressed from strong duality for Wasserstein DRO (Zhang et al. 2024):

$$\sup_{\mathbf{P}_y \in \mathscr{P}_y} \mathbb{E}_{\mathbf{x} \sim \mathbf{P}_y} \left[\ell_\beta(\mathbf{x}, y) \right] = \inf_{\lambda_y \ge 0} \left\{ \lambda_y \tau + \mathbb{E}_{\hat{\mathbf{x}} \sim \hat{\mathbf{P}}_y} \left[\sup_{\mathbf{x} \in \mathbb{R}^d} \left(\ell_\beta(\mathbf{x}, y) - \lambda_y \|\mathbf{x} - \hat{\mathbf{x}}\| \right) \right] \right\}.$$

Since $\hat{\mathbf{P}}_y$ is empirical, the expectation becomes a finite sum:

$$\frac{1}{n_y} \sum_{i=1}^{n_y} \sup_{\mathbf{x} \in \mathbb{R}^d} \left(\ell_\beta(\mathbf{x}, y) - \lambda_y \|\mathbf{x} - \hat{\mathbf{x}}_i^{(y)}\| \right),$$

where $\hat{\mathbf{x}}_i^{(y)}$ denotes the i^{th} imputed sample from class y. Therefore, the DRO objective becomes:

$$\min_{\beta, \lambda_0, \lambda_1 \ge 0} \sum_{y \in \{0,1\}} \hat{p}_y \left(\lambda_y \tau + \frac{1}{n_y} \sum_{i=1}^{n_y} \sup_{\mathbf{x} \in \mathbb{R}^d} \left[\ell_\beta(\mathbf{x}, y) - \lambda_y \|\mathbf{x} - \hat{\mathbf{x}}_i^{(y)}\| \right] \right).$$

To evaluate the inner supremum, fix $\hat{\mathbf{x}}_i^{(y)}$. From convex analysis (Shafieezadeh-Abadeh et al. 2015), we have:

$$\sup_{\mathbf{x} \in \mathbb{R}^d} \left[\ell_\beta(\mathbf{x}, y) - \lambda_y \|\mathbf{x} - \hat{\mathbf{x}}_i^{(y)}\| \right] = \begin{cases} \ell_\beta(\hat{\mathbf{x}}_i^{(y)}, y), & \text{if } \left\| \nabla_{\mathbf{x}} \ell_\beta(\hat{\mathbf{x}}_i^{(y)}, y) \right\|_* \le \lambda_y, \\ +\infty, & \text{otherwise.} \end{cases}$$

We compute the gradient:

$$\nabla_{\mathbf{x}} \ell_\beta(\hat{\mathbf{x}}_i^{(y)}, y) = -y \cdot \sigma(-y \cdot \beta^\top \hat{\mathbf{x}}_i^{(y)}) \cdot \beta,$$

where $\sigma(t) = \frac{1}{1+e^{-t}}$ is the sigmoid function. Therefore, the dual norm becomes:

$$\left\| \nabla_{\mathbf{x}} \ell_\beta(\hat{\mathbf{x}}_i^{(y)}, y) \right\|_* = \sigma(-y \cdot \beta^\top \hat{\mathbf{x}}_i^{(y)}) \cdot \|\beta\|_*.$$

Since $\sigma(\cdot) < 1$, it suffices to enforce:

$$\|\beta\|_* \leq \lambda_y.$$

Under this constraint, the pointwise supremum becomes $\ell_\beta(\hat{\mathbf{x}}_i^{(y)}, y)$, and the DRO objective reduces to:

$$\min_{\beta,\lambda_0,\lambda_1 \geq 0} \left[\hat{p}_0 \left(\lambda_0 \tau + \frac{1}{n_0} \sum_{i=1}^{n_0} \ell_\beta(\hat{\mathbf{x}}_i^{(0)}, 0) \right) + \hat{p}_1 \left(\lambda_1 \tau + \frac{1}{n_1} \sum_{i=1}^{n_1} \ell_\beta(\hat{\mathbf{x}}_i^{(1)}, 1) \right) \right],$$

subject to:

$$\|\beta\|_* \leq \lambda_0, \quad \|\beta\|_* \leq \lambda_1.$$

Finally, introducing slack variables $s_i^{(0)} \geq \ell_\beta(\hat{\mathbf{x}}_i^{(0)}, 0)$ and $s_i^{(1)} \geq \ell_\beta(\hat{\mathbf{x}}_i^{(1)}, 1)$, we obtain the equivalent convex program:

$$\begin{aligned}
\min_{\beta,\lambda_0,\lambda_1,s_i^{(0)},s_i^{(1)}} \quad & \hat{p}_0 \left(\lambda_0 \tau + \frac{1}{n_0} \sum_{i=1}^{n_0} s_i^{(0)} \right) + \hat{p}_1 \left(\lambda_1 \tau + \frac{1}{n_1} \sum_{i=1}^{n_1} s_i^{(1)} \right) \\
\text{s.t.} \quad & \ell_\beta(\hat{\mathbf{x}}_i^{(0)}, 0) \leq s_i^{(0)}, \quad \forall i = 1, \ldots, n_0, \\
& \ell_\beta(\hat{\mathbf{x}}_i^{(1)}, 1) \leq s_i^{(1)}, \quad \forall i = 1, \ldots, n_1, \\
& \|\beta\|_* \leq \lambda_0, \\
& \|\beta\|_* \leq \lambda_1.
\end{aligned}$$

□

This is a convex optimization problem over $\beta, \lambda_0, \lambda_1, s_i^{(0)}, s_i^{(1)}$, and can be solved efficiently using standard convex solvers such as CVXPY. The dual norm $\|\cdot\|_*$ corresponds to the dual of the norm used to define the Wasserstein distance in the feature space.

REFERENCES

Austin, P. C., and S. van Buuren. 2022. "The Effect of High Prevalence of Missing Data on Estimation of the Coefficients of a Logistic Regression Model When Using Multiple Imputation". *BMC Medical Research Methodology* 22(196).

Bertsimas, D., J. Dunn, C. Pawlowski, and Y. D. Zhuo. 2019. "Robust Classification". *INFORMS Journal on Optimization* 1(1):2–34.

Blanchet, J., Y. Kang, and K. Murthy. 2020. "Robust Wasserstein Profile Inference and Applications to Machine Learning". *arXiv preprint arXiv:1610.05627*.

Blanchet, J., J. Li, S. Lin, and X. Zhang. 2024. "Distributionally Robust Optimization and Robust Statistics". *arXiv preprint arXiv:2401.14655*.

Caramanis, C., S. Mannor, and H. Xu. 2012. *Robust Optimization in Machine Learning*. Cambridge, Massachusetts: MIT Press.

Chen, R., and I. C. Paschalidis. 2018. "A Robust Learning Approach for Regression Models Based on Distributionally Robust Optimization". *Journal of Machine Learning Research* 19:1–48.

Delage, E., and Y. Ye. 2010. "Distributionally Robust Optimization Under Moment Uncertainty with Application to Data-Driven Problems". *Operations Research* 58(3):595–612.

Dempster, A. P., N. M. Laird, and D. B. Rubin. 1977. "Maximum Likelihood from Incomplete Data via the EM Algorithm". *Journal of the Royal Statistical Society: Series B (Methodological)* 39(1):1–38.

Duchi, J. C., and H. Namkoong. 2018. "Learning Models with Uniform Performance via Distributionally Robust Optimization". *arXiv preprint arXiv:1810.08750*.

Ehrlinger, L., T. Grubinger, B. Varga, M. Pichler, T. Natschläger, and J. Zeindl. 2018. "Treating Missing Data in Industrial Data Analytics". In *Proceedings of the Thirteenth International Conference on Digital Information Management (ICDIM)*, 148–155. Berlin, Germany, September 24–26.

Emmanuel, T., T. Maupong, D. Mpoeleng, T. Semong, B. Mphago, and O. Tabona. 2021. "A Survey on Missing Data in Machine Learning". *Journal of Big Data* 8(140).

Faccini, D., F. Maggioni, and F. A. Potra. 2022. "Robust and Distributionally Robust Optimization Models for Linear Support Vector Machine". *Computers & Operations Research* 147:105930.

Gao, R., and A. J. Kleywegt. 2016. "Distributionally Robust Stochastic Optimization with Wasserstein Distance". *arXiv preprint arXiv:1604.02199*.

Hastie, T., R. Tibshirani, and J. Friedman. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd ed. New York: Springer.

Janosi, A. and Steinbrunn, W. and Pfisterer, M. and Detrano, R. 1989. "Heart Disease [Dataset]". UCI Machine Learning Repository. Accessed April 11, 2025.

Jerez, J. M., I. Molina, P. J. García-Laencina, E. Alba, N. Ribelles, M. Martín *et al.* 2010. "Missing Data Imputation Using Statistical and Machine Learning Methods in a Real Breast Cancer Problem". *Artificial Intelligence in Medicine* 50(2):105–115.

Jiang, W., J. Josse, and M. Lavielle. 2018. "Logistic Regression with Missing Covariates—Parameter Estimation, Model Selection and Prediction Within a Joint-Modeling Framework". *arXiv preprint arXiv:1805.04602*.

Josse, J., and J. P. Reiter. 2018. "Consistency of Supervised Learning with Missing Values". *arXiv preprint arXiv:1806.01974*.

Kuhn, D., P. Mohajerin Esfahani, V. A. Nguyen, and S. Shafieezadeh-Abadeh. 2019. "Wasserstein Distributionally Robust Optimization: Theory and Applications in Machine Learning". *arXiv preprint arXiv:1908.08729*.

Kuhn, D., S. Shafiee, and W. Wiesemann. 2024. "Distributionally Robust Optimization". *arXiv preprint arXiv:2411.02549*.

Lee, C., and S. Mehrotra. 2015. "A Distributionally-Robust Approach for Finding Support Vector Machines". Unpublished manuscript, Optimization Online, accessed April 11, 2025.

Martins, S. R., J. de Uña-Álvarez, and M. d. C. Iglesias-Pérez. 2024. "Logistic Regression with Missing Responses and Predictors: A Review of Existing Approaches and a Case Study". *arXiv preprint arXiv:2302.03435*.

Noh, H.-J., M. Kwak, and I. Han. 2004. "Improving the Prediction Performance of Customer Behavior Through Multiple Imputation". *Intelligent Data Analysis* 8(6):563–577.

Preda, C., A. Duhamel, M. Picavet, and T. Kechadi. 2005. "Tools for Statistical Analysis with Missing Data: Application to a Large Medical Database". In *Connecting Medical Informatics and Bio-Informatics*, edited by R. e. a. Engelbrecht, 181–186. Amsterdam, The Netherlands: IOS Press.

Schafer, J. L., and J. W. Graham. 2002. "Missing Data: Our View of the State of the Art". *Psychological Methods* 7(2):147–177.

Shafieezadeh-Abadeh, S., P. Mohajerin Esfahani, and D. Kuhn. 2015. "Distributionally Robust Logistic Regression". In *Advances in Neural Information Processing Systems 28*, edited by C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, 1576–1584. Montreal, Canada: Curran Associates, Inc.

Smith, B. L., W. T. Scherer, and J. H. Conklin. 2003. "Exploring Imputation Techniques for Missing Data in Transportation Management Systems". *Transportation Research Record* 1836:132–142.

Taşkesen, B., V. A. Nguyen, D. Kuhn, and J. Blanchet. 2020. "A Distributionally Robust Approach to Fair Classification". *arXiv preprint arXiv:2007.09530*.

Verchand, K. A., and A. Montanari. 2024. "High-Dimensional Logistic Regression with Missing Data: Imputation, Regularization, and Universality". *arXiv preprint arXiv:2410.01093*.

Villani, C. 2008. *Optimal Transport: Old and New*, Volume 338 of *Grundlehren der mathematischen Wissenschaften*. Springer.

Wolberg, W. and Mangasarian, O. and Street, N. and Street, W. 1993. "Breast Cancer Wisconsin (Diagnostic) [Dataset]". UCI Machine Learning Repository. Accessed April 11, 2025.

Wu, Q., J. Y.-M. Li, and T. Mao. 2022. "On Generalization and Regularization via Wasserstein Distributionally Robust Optimization". Accessed April 11, 2025.

Zhang, L., J. Yang, and R. Gao. 2024. "A Short and General Duality Proof for Wasserstein Distributionally Robust Optimization". *arXiv preprint arXiv:2205.00362*.

Zhang, Z. 2016. "Missing Data Imputation: Focusing on Single Imputation". *Annals of Translational Medicine* 4(1):9.

AUTHOR BIOGRAPHIES

WEICONG CHEN is a Ph.D. student in the Department of Computational and Data Sciences at George Mason University. He earned his Master's degree in Business Analytics from the Johns Hopkins University. His research interests include distributionally robust optimization, robustness, and fairness in machine learning models and operations research applications. His email address is wchen27@gmu.edu.

HODA BIDKHORI is an assistant professor at the Department of Computational and Data Sciences at George Mason University. She earned her Ph.D. in Applied Mathematics from the Massachusetts Institute of Technology (MIT), where she subsequently spent several years as a postdoctoral researcher and lecturer in Operations Research and Statistics. Her current research focuses on the theory and applications of data-driven optimization and data analytics. Her e-mail address is hbidkhor@gmu.edu. Her website is https://sites.google.com/view/hoda-bidkhori/home.