Distributionally Robust Universal Classification: Bypassing the Curse of Dimensionality

Siyuan Chen
H. Milton Stewart School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA, USA, schen904@gatech.edu.

Weijun Xie
H. Milton Stewart School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA, USA, wxie@gatech.edu.

The Universal Classification (UC) problem aims to find an optimal classifier from a universal policy space that minimizes the expected 0-1 loss, also known as the misclassification risk. However, conventional empirical risk minimization often results in overfitting and poor out-of-sample performance. To address this limitation, we propose the Distributionally Robust Universal Classification (DRUC) formulation, which enhances generalization by incorporating distributional robustness through a Wasserstein distance-based ambiguity set centered at the empirical distribution. To manage the infinite-dimensional nature of the DRUC policy space, we develop its in-sample DRUC counterpart, which allows for a more tractable reformulation while preserving robustness properties. We prove that, asymptotically, the in-sample DRUC formulation converges to the original UC formulation and is equivalent to the DRUC formulation. Under mild conditions, we provide non-asymptotic finite-sample performance guarantees. Furthermore, we derive a mixed-integer linear programming (MILP) reformulation to obtain the optimal in-sample DRUC policy and propose an efficient 2-approximation algorithm. Our numerical experiments show the efficiency of the proposed approximation algorithm and demonstrate the superior out-of-sample performance of the in-sample DRUC formulation.

Key words: Universal classification; asymptotic convergence; sample size; distributionally robust optimization; approximation algorithm.

History:

1. Introduction

Let us consider the following multi-class Universal Classification (UC) problem:

$$v_{UC} = \inf_{f \in \mathcal{F}} \left\{ \mathbb{E}_{(\boldsymbol{X},Y)\sim\mathbb{P}_0} \left[\mathbb{1}_{\{f(\boldsymbol{X})\neq Y\}} \right] \right\}, \tag{1}$$

where the random variable Y denotes the class label, taking values in a finite set $\mathcal{Y} = [m] := \{1, 2, \ldots, m\}$, and the random vector $\boldsymbol{X}$ represents the feature vector, supported on a closed subset $\mathcal{X} \subseteq \mathbb{R}^d$ (we assume closedness in the sense of measure theory; see Teschl 2014). The function class $\mathcal{F}$ consists of all Lebesgue measurable functions mapping from $\mathcal{X}$ to $\mathcal{Y}$, and $\mathbb{P}_0$ denotes the true joint distribution of the feature-label pair $(\boldsymbol{X}, Y)$. The objective of the UC problem (1) is to identify a measurable function that minimizes the expected misclassification risk.

If the true distribution $\mathbb{P}_0$ were known, an optimal classification policy for problem (1) would assign each feature vector $\boldsymbol{x}$ to the class label that maximizes its conditional true probability (Lin 2002), as stated in our Lemma 3. However, in real-world applications, the true underlying distribution is typically unknown. Instead, we usually have access only to a finite empirical dataset $\{(\widehat{\boldsymbol{x}}^i, \widehat{y}^i)\}_{i\in[n]}$, sampled from $(\boldsymbol{X}, Y)$. One practical approach, therefore, is to approximate the unknown distribution $\mathbb{P}_0$ with the empirical distribution $\widehat{\mathbb{P}}_n$, leading to the following Empirical Universal Classification (EUC) problem:

$$v_{EUC} = \inf_{f\in\mathcal{F}} \left\{ \mathbb{E}_{(\boldsymbol{X},Y)\sim\widehat{\mathbb{P}}_n} \left[\mathbb{1}_{\{f(\boldsymbol{X})\neq Y\}} \right] \right\}, \tag{2}$$

where the empirical distribution is $\widehat{\mathbb{P}}_n = \frac{1}{n}\sum_{i\in[n]} \delta_{(\widehat{\boldsymbol{x}}^i, \widehat{y}^i)}$, where $\{(\widehat{\boldsymbol{x}}^i, \widehat{y}^i)\}_{i\in[n]}$ are empirical data and $\delta_{(\cdot)}$ denotes the Dirac measure. The EUC problem (2) corresponds to the sampling average approximation (SAA) of the UC problem (1) (Kleywegt et al. 2002). However, when the feature space $\mathcal{X}$ is continuous and the random feature vector $\boldsymbol{X}$ follows a continuous distribution, the optimal policy for this EUC problem (2) degenerates to simply assigning their observed labels to in-sample data points, while arbitrarily labeling all out-of-sample feature points. Unlike the SAA method for traditional stochastic programs, the solution to the EUC problem (2) severely overfits the empirical dataset. Thus, the EUC problem (2) may not be appropriate for approximating the original UC problem (1).

To address this overfitting issue, we propose a Distributionally Robust Universal Classification (DRUC) formulation as follows:

$$v_n^{\mathcal{X}} = \inf_{f\in\mathcal{F}} \sup_{\mathbb{P}\in\mathcal{P}_W(\mathcal{X}\times\mathcal{Y})} \left\{ \mathbb{E}_{(\boldsymbol{X},Y)\sim\mathbb{P}} \left[\mathbb{1}_{\{f(\boldsymbol{X})\neq Y\}} \right] \right\}. \tag{3}$$

In this formulation, rather than only committing to the empirical distribution, we introduce an ambiguity set $\mathcal{P}_W(\mathcal{X}\times\mathcal{Y})$ and seek to minimize the worst-case expected classification loss among all distributions within this set. Specifically, we use the Wasserstein-based ambiguity set, defined as the set of all probability distributions whose type-1 Wasserstein distance from the empirical distribution $\widehat{\mathbb{P}}_n$ is no greater than a predetermined positive radius $\theta > 0$; i.e., $\mathcal{P}_W(\mathcal{X}\times\mathcal{Y}) = \{\mathbb{P} : \mathbb{P}(\mathcal{X}\times\mathcal{Y}) = 1, \mathcal{W}_1(\mathbb{P}, \widehat{\mathbb{P}}_n) \leq \theta\}$. The type-1 Wasserstein distance between two probability distributions μ and μ' supported on $\mathcal{X}\times\mathcal{Y}$ is defined as follows:

$$\mathcal{W}_1(\mu,\mu') = \inf\left\{ \mathbb{E}[\|\boldsymbol{X}-\boldsymbol{X}'\| + \kappa \mathbb{1}_{\{Y\neq Y'\}}] : (\boldsymbol{X},Y)\sim\mu, (\boldsymbol{X}',Y')\sim\mu' \right\}.$$

Following the literature such as Shafieezadeh Abadeh et al. (2015), we adopt the feature-label distance metric between two data points $(\boldsymbol{x}, y)$ and $(\boldsymbol{x}', y')$ to be

$$\|\boldsymbol{x}-\boldsymbol{x}'\| + \kappa \mathbb{1}_{\{y\neq y'\}},$$

where κ is a parameter that quantifies the relative importance of label differences regarding covariate differences. Unfortunately, it is difficult to obtain a finite-dimensional reformulation of DRUC (3) since the policy space is infinite-dimensional.

To bypass this curse of dimensionality, we introduce a more tractable variant called the *in-sample DRUC*. Specifically, we define the in-sample feature space as $\widehat{\mathcal{X}}_n = \{\widehat{\boldsymbol{x}}^i\}_{i\in[n]}$, and let set $\widehat{\mathcal{F}}_n$ denote the class of all Lebesgue measurable functions mapping from the finite set $\widehat{\mathcal{X}}_n$ to the label set $\mathcal{Y}$. The resulting in-sample DRUC formulation admits the following form:

$$v_n^{\widehat{\mathcal{X}}_n} = \inf_{f\in\mathcal{F}} \sup_{\mathbb{P}\in\mathcal{P}_W(\widehat{\mathcal{X}}_n\times\mathcal{Y})} \left\{\mathbb{E}_{(\boldsymbol{X},Y)\sim\mathbb{P}}\left[\mathbb{1}_{\{f(\boldsymbol{X})\neq Y\}}\right]\right\}. \tag{4}$$

In the in-sample DRUC formulation, we restrict the ambiguity set to distributions that are supported exclusively on the in-sample feature space, rather than the entire feature space. Thus, we have $v_n^{\widehat{\mathcal{X}}_n} \leq v_n^{\mathcal{X}}$, since the in-sample DRUC formulation considers the supremum over a smaller ambiguity set. Without loss of generality, we can restrict any policy $f \in \mathcal{F}$ to the in-sample policy space $\widehat{\mathcal{F}}_n$ in the in-sample DRUC formulation (4) since the formulation does not depend on the policy assignment outside the support $\widehat{\mathcal{X}}_n$. For any in-sample policy $\widehat{f} \in \widehat{\mathcal{F}}_n$, we can extend it to a general policy $\widetilde{f} \in \mathcal{F}$ using a "closest-label" approach. That is, for each feature point $\boldsymbol{x} \in \mathcal{X}$, we define $\widetilde{f}(\boldsymbol{x})$ to assign the same label as that of the nearest in-sample feature point under the in-sample policy $\widehat{f}$. We will show that this in-sample DRUC formulation is computationally more tractable than the original DRUC, as the number of decision variables scales linearly with the sample size and remains independent of feature dimensions. Based on the closest-label extension, we can show that the optimal in-sample DRUC value converges asymptotically to the optimal DRUC value as the sample size approaches infinity. We will also demonstrate that the optimal in-sample DRUC value converges to the optimal original UC value as the sample size approaches infinity and the Wasserstein radius approaches zero.

1.1. Literature Review

Supervised learning (James et al. 2013) is a fundamental field in machine learning, where classification involves assigning each observation to one label. Many classification approaches have been developed, each demonstrating advantages tailored to different data patterns. Logistic regression (Hosmer Jr et al. 2013) estimates the conditional probability of each label as a logistic function of a linear combination of features. Support Vector Machines (SVMs) (Wang 2005, Xu et al. 2009) attempt to determine optimal separating hyperplanes by solving an optimization problem that minimizes the misclassified hinge loss. Both standard logistic regression and SVM assume linear separability among labeled feature points and thus have limitations, particularly in complex or

multi-class classification scenarios. Classification and Regression Trees (CART) (Breiman 2017) recursively partition data according to feature thresholds, producing interpretable decision trees. More recently, deep learning methods (LeCun et al. 2015), which leverage highly adaptable neural networks and flexible architectures, have found extensive applications in classification. Other nonlinear approaches, including CART-based ensemble models (e.g., Random Forests, Boosting methods), provide substantial flexibility and typically exhibit superior empirical performance. Although empirically performing well, these methods often lack interpretability and rigorous mathematical guarantees regarding performance.

Distributionally Robust Optimization (DRO) (Rahimian and Mehrotra 2022, Kuhn et al. 2024) is a framework extensively used for decision-making under uncertainty (Zhang et al. 2024, Perakis et al. 2023, Carlsson et al. 2018, Ghosal and Wiesemann 2020, Blanchet et al. 2022, Xin and Goldberg 2022, Yin and Zhao 2021, Du et al. 2020, Wang et al. 2019, Chang et al. 2019, Feng et al. 2022, Sun et al. 2023). Its application spans various areas, including vehicle routing (Ghosal and Wiesemann 2020, Yin and Zhao 2021, Carlsson et al. 2018), portfolio optimization (Blanchet et al. 2022, Du et al. 2020), pricing (Zhang et al. 2024, Perakis et al. 2023), scheduling (Wang et al. 2019, Chang et al. 2019), and resource allocation (Feng et al. 2022, Sun et al. 2023). Particularly, the work of Zhang et al. (2024) proposed a Wasserstein-based DRO framework for the Newsvendor problem with covariate information under a universal policy space. They considered an in-sample formulation, proved its equivalence to the original problem, and introduced a slope-minimization method to obtain the optimal out-of-sample policy from the in-sample policy. Motivated by their approach, we leverage the in-sample formulation strategy to circumvent the infinite-dimensional complexity associated with universal policy spaces. Nevertheless, due to the discrete nature of the output (label) space and the non-convexity of the 0-1 loss in our problem, the method proposed by Zhang et al. cannot be directly applied to our DRUC problem.

DRO frameworks have also been widely used in supervised learning literature (see the comprehensive review by Kuhn et al. 2019). Existing research has integrated DRO into classical parametric models. For instance, Shafieezadeh Abadeh et al. (2015) studied a Distributionally Robust Logistic Regression with a Wasserstein ambiguity set. They reformulated the minimax optimization problem as a tractable convex program and demonstrated its improved performance compared to the conventional logistic regression method. Recently, Belbasi et al. (2023) studied distributionally robust linear classification and regression problems that involve both continuous and discrete features, proving polynomial-time solvability in multiple cases and demonstrating superior performance over traditional methods on benchmark datasets. Additionally, using DRO to integrate decision-making problems with machine learning methodologies has also attracted much attention. Kannan et al. (2024) proposed a residual-based DRO framework that extends their earlier SAA

formulation (Kannan et al. 2022). Their framework aims to minimize the worst-case post-prediction cost, which can accommodate various regression models. Zhu et al. (2022) studied a weighted k-nearest neighbors (k-NN) classifier within a minimax DRO framework. By incorporating metric learning techniques, their method exhibited robust empirical performance, especially under limited training data. Quite differently, their approach did not use 0-1 loss and did not consider potential shifts in labels, whereas our formulation focuses on 0-1 loss and explicitly incorporates label alterations into the transportation metric. Moreover, Zhu et al. (2022) showed the equivalence between their proposed k-NN classifier and the in-sample DRO model. Still, they did not show the connection between in-sample DRO and the original DRO models. Our formulation differs from theirs, as we consider the deterministic policy space. More importantly, we rigorously prove that the optimal in-sample DRUC value converges to the optimal DRUC value when sample sizes tend to infinity and obtain finite-sample guarantees. Our proposed approximation algorithm is also more scalable, based on a linear programming relaxation whose total number of variables and constraints scales linearly with both the sample size and the number of labels.

1.2. Summary of Contributions

In this paper, we study a Distributionally Robust Universal Classification (DRUC) problem with the type-1 Wasserstein ambiguity set. Our main contributions are summarized as follows.

(i) Our DRUC formulation does not impose any restrictive assumptions (e.g., linearity, Lipschitz continuity) on the optimal policy of the relationship between features and labels.

(ii) We formulate the in-sample DRUC problem, proving that asymptotically, the optimal in-sample DRUC value, the optimal DRUC value, and the optimal UC value coincide. This resolves the overfitting issue of the SAA method. We also provide finite-sample guarantees under mild conditions.

(iii) We bypass the curse of dimensionality of the original DRUC by presenting a big-M free reformulation of the in-sample DRUC problem as a Mixed-Integer Linear Program (MILP), whose number of variables and constraints grows linearly with the sample size and the number of labels. We also propose a 2-approximation algorithm based on the linear programming (LP) relaxation solution.

(iv) Through numerical study, we validate the effectiveness and convergence properties of our approximation algorithm and numerically demonstrate the superior performance and robustness of our proposed approach compared to the closest-point method, logistic regression model, and neural networks.

Organization. The rest of the paper is organized as follows. Section 2 shows both original DRUC and in-sample DRUC reformulations. Sections 3 and 4 present the main convergence results of the

optimal UC value, the optimal DRUC value, and the optimal in-sample DRUC values, and also show the finite sample guarantees under mild conditions. Section 5 develops a MILP reformulation of the in-sample DRUC problem and proposes an approximation algorithm based on its continuous relaxation. Finally, Section 6 presents numerical experiments, and Section 7 concludes the paper.

Notation. Let $[n] := \{1, \ldots, n\}$, $\|\cdot\|_p$ denote the ℓ_p-norm for $p \in [1, +\infty]$, δ denote the Dirac measure. We take $\|\cdot\|$ as the shorthand for $\|\cdot\|_2$ in this paper. We set $\operatorname{dist}(v, \widehat{B}) := \inf_{w \in \widehat{B}} \|v - w\|$ for a point $v \in \mathbb{R}^d$ and a set $\widehat{B} \subseteq \mathbb{R}^d$. We let $B_{\boldsymbol{x}}(R)$ denote the open ball with origin $\boldsymbol{x}$ and radius R. We let $\operatorname{int}(\widehat{B})$, $\operatorname{bd}(\widehat{B})$, and $\operatorname{cl}(\widehat{B})$ denote the interior, boundary, and closure of a set $\widehat{B} \subseteq \mathbb{R}^d$, respectively. We let $\mathbb{R}_{++}$ denote the set of strictly positive real numbers and $\mathbb{Z}_+$ denote the set of nonnegative integers. We let $\mathcal{B}(\mathcal{X})$ be the Borel σ-algebra of the feature space $\mathcal{X}$. We let the indicator function $\mathbb{1}_A$ be equal to one if an event A occurs, 0, otherwise. The abbreviations "i.e.," and "i.i.d." are shorthand for "that is" and "independent and identically distributed."

2. Equivalent Representations of the DRUC and In-sample DRUC Problems

In this section, we present equivalent representations of the DRUC problem (3) and the in-sample DRUC problem (4), and derive their properties, which will be instrumental for proving the convergence results in the next section.

For notational convenience, we denote the feature space as set $\mathscr{X}$, which corresponds to $\mathcal{X}$ in the DRUC problem and to $\widehat{\mathcal{X}}_n$ in the in-sample DRUC problem. Let $\mathscr{F}$ denote the set of all measurable functions mapping the set $\mathscr{X}$ to $\mathcal{Y}$. Given a policy $f \in \mathscr{F}$, we let $v_n^{\mathscr{X}}(f)$ denote the worst-case objective value of the corresponding problem. Specifically,

$$v_n^{\mathscr{X}}(f) = \sup_{\mathbb{P} \in \mathcal{P}_W(\mathscr{X} \times \mathcal{Y})} \left\{ \mathbb{E}_{(\boldsymbol{X}, Y) \sim \mathbb{P}} \left[\mathbb{1}_{\{f(\boldsymbol{X}) \neq Y\}} \right] \right\}. \tag{5a}$$

With these definitions, the optimal DRUC and in-sample DRUC objective values can be expressed as $v_n^{\mathscr{X}} = \inf_{f \in \mathscr{F}} v_n^{\mathscr{X}}(f)$.

2.1. An Equivalent Representation of the Worst-case Objective

According to the strong duality result in Wasserstein DRO literature (Gao and Kleywegt 2023a, Mohajerin Esfahani and Kuhn 2018, Blanchet and Murthy 2019), we can represent the worst-case objectives as

$$v_n^{\mathscr{X}}(f) = \inf_{\lambda \geq 0} \left\{ \lambda\theta + \frac{1}{n} \sum_{i \in [n]} \left[\sup_{\boldsymbol{x} \in \mathscr{X}} \sup_{y \in \mathcal{Y}} \left\{ \mathbb{1}_{\{f(\boldsymbol{x}) \neq y\}} - \lambda \|\boldsymbol{x} - \widehat{\boldsymbol{x}}^i\| - \lambda\kappa \mathbb{1}_{\{y \neq \widehat{y}^i\}} \right\} \right] \right\}. \tag{5b}$$

We further simplify the worst-case objective. Given a policy f with domain $\mathscr{X}$, let us define two index sets

$$I_n(f) = \left\{ i \in [n] : f(\widehat{\boldsymbol{x}}^i) = \widehat{y}^i \right\}, J_n(f) = \left\{ i \in [n] : f(\widehat{\boldsymbol{x}}^i) \neq \widehat{y}^i \right\}, \tag{6a}$$

and a quantity

$$s_i^n(f) = \min\left\{\kappa, \operatorname{dist}\left(\widehat{\boldsymbol{x}}^i, f^{-1}(\mathcal{Y}\setminus\{\widehat{y}^i\})\right)\right\}, \forall i \in [n], \tag{6b}$$

where $f^{-1}(\cdot)$ denotes the inverse image of the function $f(\cdot)$ and $\operatorname{dist}(\cdot,\cdot)$ denotes the minimum distance, under the norm $\|\cdot\|$, between a point and a set. By the definition, we know that $s_i^n(f) = 0$ when $i \in J_n(f)$. This is because when $f(\widehat{\boldsymbol{x}}^i) \neq \widehat{y}^i$, we have $\widehat{\boldsymbol{x}}^i \in f^{-1}(\mathcal{Y}\setminus\{\widehat{y}^i\})$ and hence $\operatorname{dist}\left(\widehat{\boldsymbol{x}}^i, f^{-1}(\mathcal{Y}\setminus\{\widehat{y}^i\})\right) = 0$. We show in the following results that the worst-case objective is closely connected with the values $\{s_i^n(f)\}_{i\in[n]}$.

Since the label space $\mathcal{Y}$ is finite, we can directly compute the suprema in (5b) by enumerating the possible label values and representing them using the $\{s_i^n(f)\}_{i\in[n]}$ defined in (6b), respectively.

Proposition 1 *Given a policy $f \in \mathscr{F}$, for any $i \in [n]$, we have*

$$\sup_{\boldsymbol{x}\in\mathscr{X}}\sup_{y\in\mathcal{Y}}\left\{\mathbb{1}_{\{f(\boldsymbol{x})\neq y\}} - \lambda\|\boldsymbol{x}-\widehat{\boldsymbol{x}}^i\| - \lambda\kappa\mathbb{1}_{\{y\neq\widehat{y}^i\}}\right\} = \left(1-\lambda s_i^n(f)\right)^+.$$

Specifically, for any $i \in J_n(f)$, we have

$$\sup_{\boldsymbol{x}\in\mathscr{X}}\sup_{y\in\mathcal{Y}}\left\{\mathbb{1}_{\{f(\boldsymbol{x})\neq y\}} - \lambda\|\boldsymbol{x}-\widehat{\boldsymbol{x}}^i\| - \lambda\kappa\mathbb{1}_{\{y\neq\widehat{y}^i\}}\right\} = 1.$$

Proof: See Appendix A.1. □

In Proposition 1, we simplify the double supremum in (5b) into an explicit expression involving parameters λ and $s_i^n(f)$. Using this simplified form, we characterize the outer infimum by finding the minimizer λ^* within the domain $[0,+\infty)$. Thus, we show that the worst-case objective equals the proportion of the ascendingly sorted values $\{s_i^n(f)\}_{i\in[n]}$ that accumulates a total sum of $n\theta$. This result is formally stated in the following theorem.

Theorem 1 *Given a policy $f \in \mathscr{F}$, suppose that the values $\{s_i^n(f)\}_{i\in[n]}$ are sorted in ascending order, that is, $0 \le s_{(1)}^n(f) \le s_{(2)}^n(f) \le \cdots \le s_{(n)}^n(f) \le \kappa$. Let us define $k_n(f,\theta)$ as*

(i) if $\sum_{i\in[n]} s_i^n(f) < n\theta$, then we let $k_n(f,\theta) = n$; or

(ii) if $\sum_{i\in[n]} s_i^n(f) \ge n\theta$, then $k_n(f,\theta)$ is the unique value in the range $(0,n]$ such that:

$$\sum_{i\in[\lceil k_n(f,\theta)\rceil-1]} s_{(i)}^n(f) + (k_n(f,\theta)+1-\lceil k_n(f,\theta)\rceil)s_{(\lceil k_n(f,\theta)\rceil)}^n(f) = n\theta. \tag{7}$$

Then the worst-case objective is equal to $v_n^{\mathscr{X}}(f) = \frac{1}{n}k_n(f,\theta)$.

Proof: See Appendix A.2. □

The theorem provides a simple approach for computing the worst-case objective value using the sorted values of $\{s_i^n(f)\}_{i\in[n]}$.

2.2. Important Corollaries

Building on the result in the previous subsection, we present two important properties of the quantity $k_n(f,\theta)$ in the following corollary. These properties will play a crucial role in proving the convergence results later in the paper. First, we introduce the monotonicity between the values $\{s_i^n(f)\}_{i\in[n]}$ and the worst-case objective $v_n^{\mathscr{X}}(f)$ according to Theorem 1.

Corollary 1 *Given two policies $f,g\in\mathscr{F}$, suppose that for all $i\in[n]$, $s_i^n(f)\le s_i^n(g)$. Then we have $v_n^{\mathscr{X}}(f)\ge v_n^{\mathscr{X}}(g)$.*

Proof: See Appendix A.3. □

Beyond the monotonicity, we derive two more corollaries about the worst-case objective based on Theorem 1.

Corollary 2 *Given a policy $f\in\mathscr{F}$, let $k_n(f,\theta)$ be the same quantity defined in Theorem 1. Then we have*

(i) $k_n(f,\theta)\ge|\{i\in[n]:s_i^n(f)\le\theta\}|$; and

(ii) when $\theta\ge\kappa$, the worst-case objective $v_n^{\mathscr{X}}(f)$ is equal to 1.

Proof: See Appendix A.4. □

The first part of Corollary 2 gives a lower bound for $k_n(f,\theta)$, using the number of $s_i^n(f)$ values that are smaller than θ. In the second part, we discuss the scenario when θ is larger than κ and prove that the worst-case objective value is always equal to 1 in this case. This result suggests that in practice, the Wasserstein radius θ should always be chosen no larger than κ to ensure the effectiveness of the optimal policy.

Besides the case where θ exceeds κ, we also study the case when θ approaches 0. Due to the discrete nature of the sample feature space $\widehat{\mathcal{X}}_n$, we propose the following corollary regarding the optimal in-sample DRUC policy for small values of the radius θ. Specifically, when θ is smaller than the minimum distance between any two in-sample feature points divided by n, changing labels does not yield an improved in-sample DRUC objective value. Thus, in this case, the optimal in-sample DRUC policy must preserve the original labels.

Corollary 3 *Suppose that $\{\widehat{\boldsymbol{x}}^i\}_{i=1}^n$ are distinct and $\theta\le\frac{1}{n}\min_{i,j\in[n]:i\neq j}\|\widehat{\boldsymbol{x}}^i-\widehat{\boldsymbol{x}}^j\|$. Then the unique optimal solution to the dual of the in-sample DRUC problem*

$$v_n^{\widehat{\mathcal{X}}_n}=\inf_{\widehat{f}\in\widehat{\mathcal{F}}_n}\inf_{\lambda\ge0}\left\{\lambda\theta+\frac{1}{n}\sum_{i\in[n]}\left[\sup_{\boldsymbol{x}\in\widehat{\mathcal{X}}_n}\sup_{y\in\mathcal{Y}}\left\{\mathbb{1}_{\{\widehat{f}(\boldsymbol{x})\neq y\}}-\lambda\|\boldsymbol{x}-\widehat{\boldsymbol{x}}^i\|-\lambda\kappa\mathbb{1}_{\{y\neq\widehat{y}^i\}}\right\}\right]\right\}$$

is $\widehat{f}^*\in\widehat{\mathcal{F}}_n$ such that $\widehat{f}^*(\widehat{\boldsymbol{x}}^i)=\widehat{y}^i$ for all $i\in[n]$.

Proof: See Appendix A.5. □

This result implies that when the Wasserstein radius θ approaches 0, the optimal in-sample DRUC policy assigns each feature point by its label, which is precisely the SAA solution. This observation motivates our choice of the SAA solution as a natural baseline in the numerical study, as it represents an intuitive method for the problem and aligns precisely with the optimal in-sample DRUC policy when the Wasserstein radius θ is close to 0.

3. Asymptotic Properties of the Optimal In-sample DRUC Value

In this section, we study the equivalence between the optimal DRUC value $v_n^{\mathcal{X}}$ and the optimal in-sample DRUC value $v_n^{\widehat{\mathcal{X}}_n}$ as well as the optimal UC value v_{UC} as the sample size goes to infinity.

3.1. Asymptotic Equivalence between the Optimal In-sample DRUC Value and the Optimal DRUC Value

Recall that for a given in-sample policy $\widehat{f}$, its "closest-label" extension $\widetilde{f}$ is defined as follows: For any $\boldsymbol{x} \in \mathcal{X}$, $\widetilde{f}(\boldsymbol{x})$ takes the same value as $\widehat{f}$ at the nearest in-sample feature point to $\boldsymbol{x}$; i.e.,

$$\widetilde{f}(\boldsymbol{x}) = \widehat{f}(\widehat{\boldsymbol{x}}^i),\ \widehat{\boldsymbol{x}}^i \in \arg\min_{\boldsymbol{\chi} \in \widehat{\mathcal{X}}_n} \|\boldsymbol{x} - \boldsymbol{\chi}\|. \tag{8}$$

The in-sample policy and its closest-label extension coincide on all in-sample points. Thus, applying definition (6b) to both the in-sample policy and its extension, we obtain the following values:

$$s_i^n(\widehat{f}) = \min\left\{\kappa, \mathrm{dist}\left(\widehat{\boldsymbol{x}}^i, \widehat{f}^{-1}(\mathcal{Y} \setminus \{\widehat{y}^i\})\right)\right\}, \quad s_i^n(\widetilde{f}) = \min\left\{\kappa, \mathrm{dist}\left(\widehat{\boldsymbol{x}}^i, \widetilde{f}^{-1}(\mathcal{Y} \setminus \{\widehat{y}^i\})\right)\right\}.$$

Since the in-sample policy $\widehat{f}$ and its extension $\widetilde{f}$ have identical values on all in-sample feature points, it follows that the inverse image of $\widehat{f}$ over the sample feature space is contained within that of its extension $\widetilde{f}$. Thus, we have $s_i^n(\widetilde{f}) \leq s_i^n(\widehat{f})$. These properties of the "closest-label" extension are summarized and rigorously proven in the following proposition.

Proposition 2 *Given an in-sample policy $\widehat{f}$ and its extension $\widetilde{f}$ as defined in* (8)*, we have*

$$I_n(\widetilde{f}) = I_n(\widehat{f}), J_n(\widetilde{f}) = J_n(\widehat{f}), s_i^n(\widetilde{f}) \leq s_i^n(\widehat{f}), \forall i \in [n], k_n(\widetilde{f}, \theta) \geq k_n(\widehat{f}, \theta).$$

Proof: See Appendix A.6 □

We leverage the closest-label extension to show that the optimal in-sample DRUC value $v_n^{\widehat{\mathcal{X}}_n}$ converges to the original optimal DRUC value $v_n^{\mathcal{X}}$ as the sample size n increases. Before presenting this result, we introduce the following assumption about the underlying true distribution.

Assumption 1 *The data points $\{(\widehat{\boldsymbol{x}}^i, \widehat{y}^i)\}_{i=1}^{\infty}$ are i.i.d. random vectors generated from the underlying true distribution $\mathbb{P}_0$. The true distribution $\mathbb{P}_0$ generates each label in $\mathcal{Y}$ with positive probability, i.e., $p_y := \mathbb{P}_{(\boldsymbol{X},Y)\sim\mathbb{P}_0}(Y = y) > 0$ for all $y \in \mathcal{Y}$.*

Assumption 1 is quite standard in the data-driven DRO literature. The requirement that data points are i.i.d. from a common underlying distribution, along with the condition that each label occurs with positive probability, is commonly adopted in works such as Mohajerin Esfahani and Kuhn (2018), Gao and Kleywegt (2023b). Similar assumptions appear in DRO formulations involving 0-1 loss, either through the assumption of a strictly positive density (Xie et al. 2021, Yang and Gao 2022) or by restricting attention to discrete feature spaces (Ho-Nguyen and Wright 2023). Our generalization accommodates these settings and supports more flexible structures, such as categorical variables and degenerate supports.

We study the asymptotic convergence of the optimal in-sample DRUC value $v_n^{\widehat{\mathcal{X}}_n}$ to the original optimal DRUC value $v_n^{\mathcal{X}}$. To prove this convergence, we begin with two technical lemmas. Under Assumption 1, we show that the in-sample feature points become nearly dense in the feature space $\mathcal{X}$ as the sample size increases. The idea is to divide the feature space $\mathcal{X}$ into a grid of closed hypercubes. Since the samples are generated i.i.d. from a distribution with support $\mathcal{X}$, each closed hypercube has a positive probability of containing a sample. By the strong law of large numbers (Kolmogorov 1933, Billingsley 1995), when the sample size grows, every hypercube within a bounded region eventually contains at least one in-sample feature point. This result is summarized in the following lemma.

Lemma 1 *Under Assumption 1, given any $\epsilon \in \mathbb{R}_{++}$ and $R \in \mathbb{R}_{++}$, consider the partition $\{A^\epsilon_{\boldsymbol{k}}\}_{\boldsymbol{k}\in\mathbb{Z}^d}$ of the space $\mathbb{R}^d$, where $A^\epsilon_{k_1,k_2,\cdots,k_d} = [k_1\epsilon,(k_1+1)\epsilon)\times[k_2\epsilon,(k_2+1)\epsilon)\times\cdots\times[k_d\epsilon,(k_d+1)\epsilon)$. Let us denote a set $\bar{\Lambda}_\epsilon(R) = \{\boldsymbol{k}\in\mathbb{Z}^d : \mathbb{P}_{(\boldsymbol{X},Y)\sim\mathbb{P}_0}(\boldsymbol{X}\in B_{\boldsymbol{0}}(R)\cap\mathcal{X}\cap\mathrm{cl}(A^\epsilon_{\boldsymbol{k}}))>0\}$. Then, almost surely, there exists an integer n_0 such that for all $n\geq n_0$, each set in the collection $\{\mathcal{X}\cap\mathrm{cl}(A^\epsilon_{\boldsymbol{k}})\}_{\boldsymbol{k}\in\bar{\Lambda}_\epsilon(R)}$ contains at least one in-sample feature point from $\{\widehat{\boldsymbol{x}}^i\}_{i\in[n]}$.*

Proof: See Appendix A.7. □

With each hypercube containing at least one in-sample feature point, we can then bound the difference between $s_i^n(\widehat{f})$ and $s_i^n(\widetilde{f})$ for any in-sample policy $\widehat{f}$ and its closest-label extension $\widetilde{f}$ by the diameter of the hypercubes. This result is shown in the following lemma. We illustrate the proof idea in Figure 1.

Lemma 2 *Under Assumption 1, given an in-sample policy $\widehat{f}$ and its closest-label extension $\widetilde{f}$ defined in* (8)*, given any $\epsilon>0$ and $R>0$, there almost surely exists an n_0 such that for all $n\geq n_0$ and for every $\widehat{\boldsymbol{x}}^i\in\widehat{\mathcal{X}}_n\cap B_0(R)$, the following inequalities hold: $s_i^n(\widehat{f})\geq s_i^n(\widetilde{f})\geq s_i^n(\widehat{f})-\epsilon$.*

Proof: See Appendix A.8. □

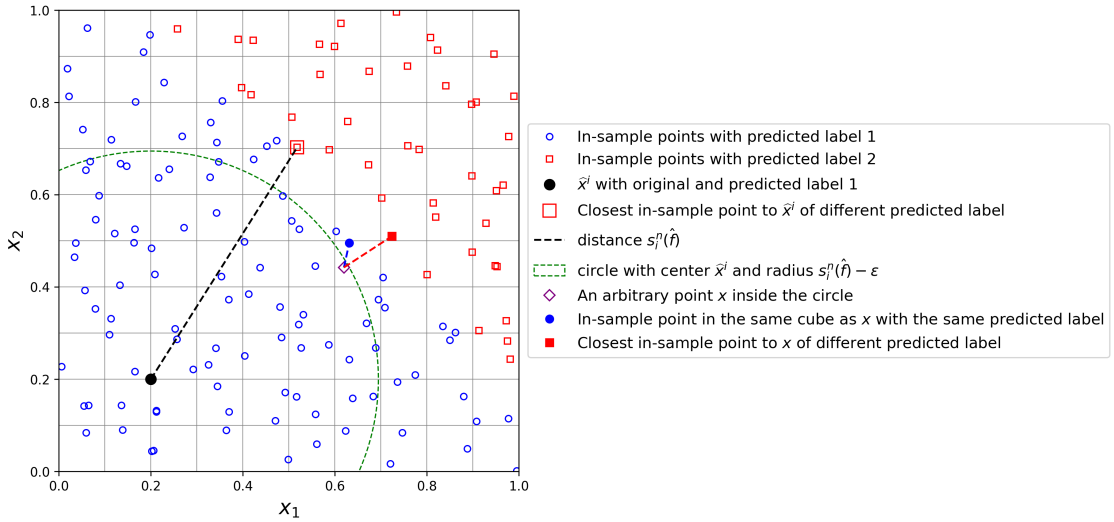


Figure 1 Illustration of the proof of Lemma 2. The feature space is partitioned into hypercubes of side length $\epsilon/\sqrt{2}$. Blue circles and red squares represent two different labels. Given an in-sample policy $\widehat{f}$, suppose that an in-sample point $\widehat{\boldsymbol{x}}^i$ (solid black circle) takes its original label $\widehat{y}^i$. Consider an arbitrary point $\boldsymbol{x}$ (purple diamond) within the green dashed circle centered at $\widehat{\boldsymbol{x}}^i$, with radius $s_i^n(\widehat{f})-\epsilon$. There must exist an in-sample point (solid blue circle) in the same hypercube, having the same predicted label as $\widehat{y}^i$, and located closer than any point with a different label (e.g., solid red square). Therefore, the extension $\widetilde{f}$ assigns the value $\widehat{y}^i$ throughout the green dashed region.

Lemma 2 bounds the gap between $s_i^n(\widehat{f})$ and $s_i^n(\widetilde{f})$. With that, we are ready to show the convergence of the optimal in-sample DRUC value $v_n^{\widehat{\mathcal{X}}_n}$ to the optimal DRUC value $v_n^{\mathcal{X}}$. The key idea is to refine the hypercube partition by letting the diameter of each hypercube in the grid become arbitrarily small. This allows us to uniformly bound the difference between $s_i^n(\widehat{f})$ and $s_i^n(\widetilde{f})$ for every possible in-sample policy $\widehat{f}$ and its closest-label extension $\widetilde{f}$. This uniform bound translates into a vanishing gap between the in-sample DRUC objective value $v_n^{\widehat{\mathcal{X}}_n}(\widehat{f})$ and the original DRUC objective value $v_n^{\mathcal{X}}(\widetilde{f})$ for any in-sample policy and its closest-label extension. That is, as the sample size increases, this gap converges to zero. By evaluating the original DRUC objective at the closest-label extension of the optimal in-sample policy, we obtain the convergence of $v_n^{\widehat{\mathcal{X}}_n}$ to $v_n^{\mathcal{X}}$. This result is formally proven in the following theorem.

Theorem 2 *Under Assumption 1, given any $\theta\in\mathbb{R}_{++},\kappa\in\mathbb{R}_{++}$, for $n\in\mathbb{Z}^+$, we have $0\le v_n^{\mathcal{X}}(\widetilde{f})-v_n^{\widehat{\mathcal{X}}_n}(\widehat{f})<\eta_n$ for any $\widehat{f}\in\widehat{\mathcal{F}}_n$ and its extension $\widetilde{f}$, where η_n depends only on $\widehat{\mathcal{X}}_n$ and not on $\widehat{f}$, and $\lim_{n\to\infty}\eta_n=0$ almost surely.*

Proof: See Appendix A.9. □

Remark 1 Theorem 2 shows a uniform upper bound on the gap between the in-sample DRUC objective value $v_n^{\widehat{\mathcal{X}}_n}(\widehat{f})$ and the corresponding DRUC value of its extension $v_n^{\mathcal{X}}(\widetilde{f})$, and this gap

converges to zero almost surely as the sample size grows. Based on this result, we now consider an optimal policy $\widehat{f}_n^{\min}$ for the in-sample DRUC problem, along with its "closest-label" extension $\widetilde{f}_n^{\min}$. Note that

$$v_n^{\widehat{\mathcal{X}}_n}(\widehat{f}_n^{\min}) = v_n^{\widehat{\mathcal{X}}_n} \le v_n^{\mathcal{X}} \le v_n^{\mathcal{X}}(\widetilde{f}_n^{\min}) \le v_n^{\widehat{\mathcal{X}}_n}(\widehat{f}_n^{\min}) + \eta_n.$$

Therefore, we see that almost surely

$$\lim_{n\to\infty}(v_n^{\mathcal{X}} - v_n^{\widehat{\mathcal{X}}_n}) = 0.$$

We also observe that the extension $\widetilde{f}_n^{\min}$ of the optimal in-sample policy $\widehat{f}_n^{\min}$ attains a DRUC objective value that converges to the optimal DRUC value as the sample size increases. Therefore, $\widetilde{f}_n^{\min}$ serves as an asymptotically optimal policy for the DRUC problem. In the numerical study, we adopt $\widetilde{f}_n^{\min}$ as an approximate optimal policy for evaluating the performance of the proposed formulation.

3.2. Convergence of the Optimal DRUC Value to the Optimal UC Value

In this subsection, we analyze the gap between the optimal DRUC value, whether from the original or in-sample formulation, and the optimal UC value, denoted by v_{UC}. Recall that we let p_y be the true probability of label y, i.e., $p_y = \mathbb{P}_{(\boldsymbol{X},Y)\sim\mathbb{P}_0}(Y = y)$. For notational convenience, for each $y \in \mathcal{Y}$, we let $\mathbb{P}_y$ denote the true probability distribution conditioning on the label y. That is,

$$\mathbb{P}_y(A) = \mathbb{P}_{(\boldsymbol{X},Y)\sim\mathbb{P}_0}(\boldsymbol{X} \in A | Y = y), \forall y \in \mathcal{Y}, A \in \mathcal{B}(\mathcal{X}). \tag{9a}$$

By total probability, the $\boldsymbol{x}$-coordinate $\mathbb{P}_0^{\mathcal{X}}$ of the true distribution $\mathbb{P}_0$ can be expressed as

$$\mathbb{P}_0^{\mathcal{X}}(A) = \mathbb{P}_{(\boldsymbol{X},Y)\sim\mathbb{P}_0}(\boldsymbol{X} \in A) = \sum_{y\in\mathcal{Y}} p_y \mathbb{P}_y(A), \forall A \in \mathcal{B}(\mathcal{X}). \tag{9b}$$

Assumption 1 ensures that $p_y > 0$ for all $y \in \mathcal{Y}$. Therefore, for any measurable set $A \subseteq \mathcal{X}$, the condition $\mathbb{P}_0^{\mathcal{X}}(A) = 0$ implies $\mathbb{P}_y(A) = 0$ for each $y \in \mathcal{Y}$. This shows that $\mathbb{P}_y$ is absolutely continuous with respect to $\mathbb{P}_0^{\mathcal{X}}$. According to the Radon-Nikodym theorem (see, e.g., Folland 1999, chapter 3.8), this implies the existence of a Borel measurable Radon-Nikodym derivative $\frac{\mathrm{d}\mathbb{P}_y}{\mathrm{d}\mathbb{P}_0^{\mathcal{X}}}$ on $\mathcal{X}$ for each $y \in \mathcal{Y}$.

Recall that, under the true distribution $\mathbb{P}_0$, the optima UC value v_{UC} is defined as

$$v_{UC} = \inf_{f\in\mathcal{F}} \mathbb{E}_{(\boldsymbol{X},Y)\sim\mathbb{P}_0}[\mathbb{1}_{\{f(\boldsymbol{X})\neq Y\}}] = \inf_{f\in\mathcal{F}} \mathbb{P}_{(\boldsymbol{X},Y)\sim\mathbb{P}_0}(f(\boldsymbol{X}) \neq Y).$$

To attain this optimal value, the optimal classification policy assigns each feature point to the label that maximizes the conditional probability of correctness. Specifically, we define the region

$$U_y = \left\{ \boldsymbol{x} \in \mathcal{X} : \begin{array}{ll} p_y \dfrac{\mathrm{d}\mathbb{P}_y}{\mathrm{d}\mathbb{P}_0^{\mathcal{X}}}(\boldsymbol{x}) > p_{y'} \dfrac{\mathrm{d}\mathbb{P}_{y'}}{\mathrm{d}\mathbb{P}_0^{\mathcal{X}}}(\boldsymbol{x}) & \text{for } y' = 1, \ldots, y-1; \\ p_y \dfrac{\mathrm{d}\mathbb{P}_y}{\mathrm{d}\mathbb{P}_0^{\mathcal{X}}}(\boldsymbol{x}) \ge p_{y'} \dfrac{\mathrm{d}\mathbb{P}_{y'}}{\mathrm{d}\mathbb{P}_0^{\mathcal{X}}}(\boldsymbol{x}) & \text{for } y' = y+1, \ldots, m. \end{array} \right\} \tag{9c}$$

By construction, the sets $U_{y_{y\in\mathcal{Y}}}$ are disjoint and form a partition of the entire feature space $\mathcal{X}$. We consider the classification policy that assigns all feature vectors in U_y to label y for each $y\in\mathcal{Y}$. The following lemma formally demonstrates that this policy is measurable, minimizes the misclassification probability, and is therefore optimal for the UC problem.

Lemma 3 *Under Assumption 1, let* $\{p_y\}_{y\in\mathcal{Y}}$, $\{\mathbb{P}_y\}_{y\in\mathcal{Y}}$ *be as defined in* (9a). *Then, the optimal UC value is given by*

$$v_{UC}=1-\int_{\mathcal{X}}\max_{y\in\mathcal{Y}}\left\{p_y\frac{\mathrm{d}\mathbb{P}_y}{\mathrm{d}\mathbb{P}_0^{\mathcal{X}}}\right\}\mathrm{d}\mathbb{P}_0^{\mathcal{X}},$$

with an optimal UC policy $f_0(\boldsymbol{x})=\min\{\arg\max_{y\in\mathcal{Y}}(p_y\frac{\mathrm{d}\mathbb{P}_y}{\mathrm{d}\mathbb{P}_0^{\mathcal{X}}}(\boldsymbol{x}))\}$.

Proof: See Appendix A.10. □

We present the following theorems to show that both the original and in-sample optimal DRUC values, $v_n^{\mathcal{X}}$ and $v_n^{\widehat{\mathcal{X}}_n}$, converge to the optimal UC value v_{UC} under Assumption 1. We first prove that the optimal in-sample DRUC value is asymptotically lower bounded by v_{UC} using a grid-based argument similar to the proof of Theorem 2. In the proof, we analyze how many data points within each hypercube contribute to the DRUC objective– whether through the quantity $k_n(\widehat{f},\theta)$ or the set $J_n(\widehat{f})$ (see Theorem 1). In particular, within each hypercube, all in-sample feature points must either be assigned the same label or be included in the quantity $k_n(\widehat{f},\theta)$. This characterization yields a valid lower bound on the in-sample DRUC value, which converges to v_{UC} as the sample size grows. We illustrate the proof idea in Figure 2. The result is formally stated below.

Theorem 3 *Under Assumption 1, given a* $\kappa>0$ *and* $\theta>0$ *such that* $0<\theta<\frac{\kappa}{m}$, *we have*

$$\liminf_{n\to\infty}v_n^{\mathcal{X}}\geq\liminf_{n\to\infty}v_n^{\widehat{\mathcal{X}}_n}\geq v_{UC}$$

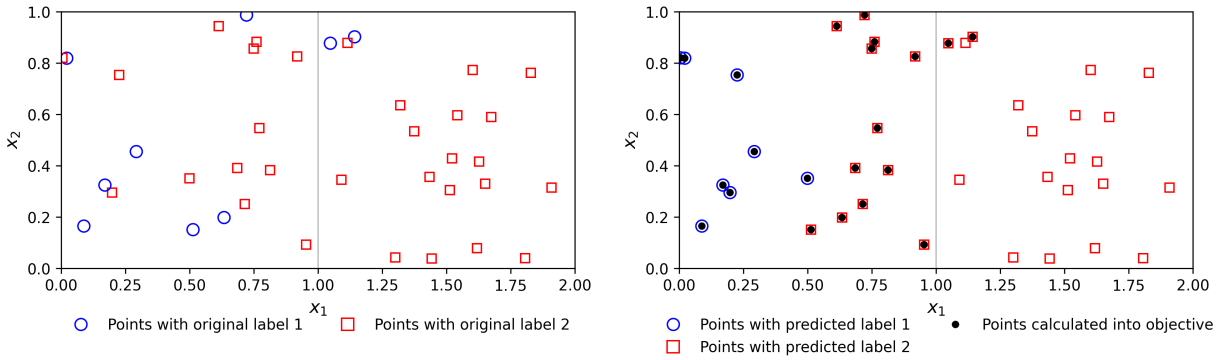
almost surely.

Proof: See Appendix A.11. □

To derive an upper bound on the optimal (original or in-sample) DRUC value, we analyze the DRUC value under a particular policy. Specifically, we consider the optimal UC policy $f_0(\boldsymbol{x})$ defined in Lemma 3. Before proceeding, we introduce a mild technical assumption: the decision boundaries of the optimal UC policy have zero probability.

Assumption 2 *For each* $y\in\mathcal{Y}$, *the relative boundary of* U_y *with respect to* $\mathcal{X}$ *is a zero-measure set regarding the measure* $\mathbb{P}_0^{\mathcal{X}}$, *which is*

$$\mathbb{P}_0^{\mathcal{X}}(\mathrm{bd}^{\mathcal{X}}(U_y))=\mathbb{P}_0^{\mathcal{X}}(\{\boldsymbol{x}\in\mathcal{X}:\forall r>0,\,B_{\boldsymbol{x}}(r)\cap U_y\neq\varnothing,B_{\boldsymbol{x}}(r)\cap(\mathcal{X}\setminus U_y)\neq\varnothing\})=0.$$



(a) Original labels

(b) A lower bound of the worst-case objective of a given in-sample policy $\widehat{f}$

Figure 2 Illustration of the proof of Theorem 3. Each subfigure depicts two hypercubes. The first subfigure shows the original labels of the in-sample feature points, where blue circles represent label 1 and red squares represent label 2. The second subfigure illustrates a specific in-sample policy: a blue circle indicates a predicted label of 1, while a red circle indicates a predicted label of 2. In the left hypercube of the second subfigure, both predicted labels are present. Thus, all points are included in the objective according to Corollary 2, since the Wasserstein radius satisfies $\theta > \epsilon$. In contrast, the right hypercube contains only a single predicted label. Therefore, points with predicted labels differing from their original labels are counted in the objective, as their $s_i^n(\widehat{f})$ values are zero. The solid-filled points mark those that must be included in the objective. The proportion of such points among all n in-sample points provides a lower bound on the worst-case objective value of this in-sample policy.

This assumption restricts the decision boundary between different sets $\{U_y\}_{y\in\mathcal{Y}}$ to have zero measure. It is a very mild condition, which holds in various settings, including cases where, for each label, the conditional distribution of the feature vector admits an analytic density function.

According to the definition (9c) of U_y, we further define the sets $V_y(r)$ for each $y \in \mathcal{Y}$ and $r > 0$ as follows:

$$V_y(r) = \{\boldsymbol{x} \in U_y : B_{\boldsymbol{x}}(r) \cap \mathcal{X} \subseteq U_y\}. \tag{10}$$

For each $y \in \mathcal{Y}$, we have $V_y(r) \subseteq U_y$. Let us define

$$\gamma(r) = \mathbb{P}_{(\boldsymbol{X},Y)\sim\mathbb{P}_0}\left(\boldsymbol{X} \in \bigcup_{y\in\mathcal{Y}}(U_y \setminus V_y(r))\right) = \sum_{y\in\mathcal{Y}} \mathbb{P}_0^{\mathcal{X}}(U_y \setminus V_y(r)). \tag{11}$$

The following lemma shows that under Assumptions 1 and 2, the probability $\gamma(r)$ vanishes as the radius r tends to zero. This confirms that the boundary regions become negligible in the limit.

Lemma 4 *Under Assumptions 1 and 2, we have* $\lim_{r\to 0_+}\gamma(r) = 0$.

Proof: See Appendix A.12 □

With the boundary probability $\gamma(r)$ shrinking to zero as $r \to 0$, we can provide an upper bound on the optimal DRUC value in the following theorem. As discussed earlier, this bound is obtained by evaluating the DRUC value of the optimal UC policy f_0. Under this policy, the proportion of misclassified samples $\frac{1}{n}|J_n(f_0)|$ converges to v_{UC}. The remaining component of the DRUC value, namely $\frac{1}{n}(k_n(f_0,\theta) - |J_n(f_0)|)$ in Theorem 1, can be shown to converge to zero as $\theta \to 0_+$. We illustrate the proof in Figure 3.

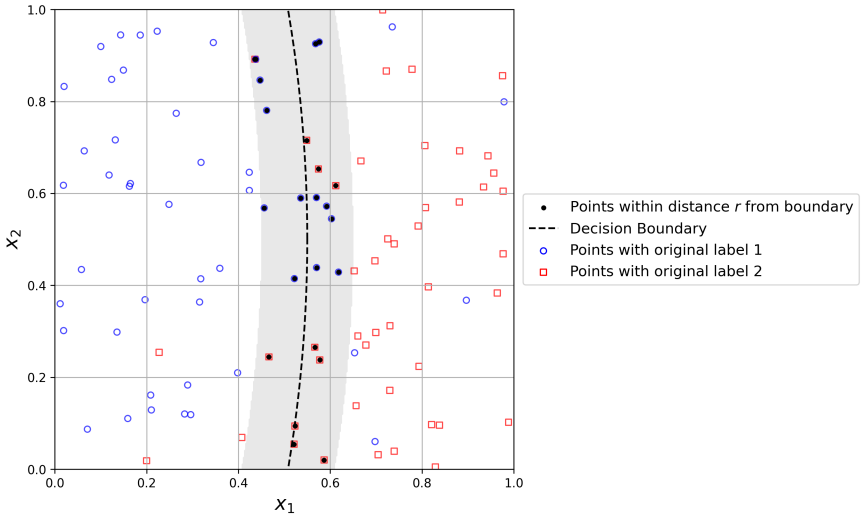


Figure 3 Illustration of the proof of Theorem 4. The black dashed curve represents the theoretical decision boundary of $\{U_y\}_{y\in\mathcal{Y}}$, and the shaded region consists of points within distance $r > 0$ of this boundary. The radius r is chosen so that $\gamma(r) \le \epsilon/2$, as guaranteed by Lemma 4. Thus, as $\epsilon \to 0_+$, the contribution to the objective from in-sample points within the shaded region converges to 0. For in-sample points outside the shaded region, the proportion of misclassified points $\frac{1}{n}|J_n(f_0)|$ converges to v_{UC} as the sample size tends to infinity.

Theorem 4 *Under Assumptions 1 and 2, given any $\kappa \in \mathbb{R}_{++}$ and $\epsilon \in \mathbb{R}_{++}$, there exists a $\bar{\theta}$ such that for any $\theta \in (0, \bar{\theta}]$, we have*

$$\limsup_{n\to\infty} v_n^{\widehat{\mathcal{X}}_n} \le \limsup_{n\to\infty} v_n^{\mathcal{X}} \le v_{UC} + \epsilon.$$

almost surely.

Proof: See Appendix A.13. □

3.3. Summary of Results and Discussions

We summarize the asymptotic convergence results in Figure 4.

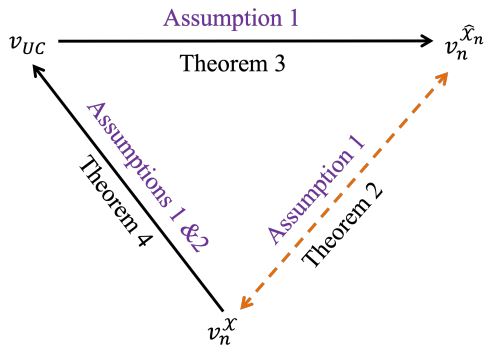


Figure 4 Summary of asymptotic convergence results. A solid arrow denotes an asymptotic inequality, while a bidirectional dashed arrow denotes an asymptotic equality.

According to Theorems 2 through 4, we establish a tight asymptotic relationship among the optimal DRUC value $v_n^{\mathcal{X}}$, the optimal in-sample DRUC value $v_n^{\widehat{\mathcal{X}}_n}$, and the optimal UC value v_{UC}. In particular, there exists a positive function $\epsilon(\theta)$, depending only on the Wasserstein radius θ, such that

$$v_{UC} \le \lim_{n\to\infty} v_n^{\widehat{\mathcal{X}}_n} = \lim_{n\to\infty} v_n^{\mathcal{X}} \le v_{UC} + \epsilon(\theta), \quad \text{with} \quad \lim_{\theta\to 0_+} \epsilon(\theta) = 0,$$

where the first equality follows from Theorem 2, the first inequality is due to Theorem 3, and the second inequality is implied by Theorem 4.

These convergence results provide theoretical justification for the in-sample DRUC formulation. It is worth noting that they rely only on Assumptions 1 and 2, which are mild and hold for a broad class of true distributions. In the subsequent sections, we further analyze the in-sample DRUC problem by establishing non-asymptotic performance guarantees and developing tractable reformulations.

4. Finite Sample Guarantees

In this section, we study finite-sample guarantees for the convergence theorems presented in Section 3. Specifically, we analyze the convergence rates between the optimal UC value, the optimal in-sample DRUC value, and the optimal DRUC value, and provide the corresponding finite-sample bounds. These results theoretically quantify the convergence behavior among these values.

Throughout this section, we assume that the feature distribution is light-tailed sub-Gaussian.

Assumption 3 *The marginal distribution of the feature vector $\boldsymbol{X}$ under the true distribution $\mathbb{P}_0$ is sub-Gaussian; that is, there exists a constant $c>0$ such that for any $t>0$, we have $\mathbb{P}_{(\boldsymbol{X},Y)\sim\mathbb{P}_0}(|\boldsymbol{X}| \ge t) \le 2\exp(-\frac{t^2}{c^2})$.*

Assumption 3 is standard in the statistical learning and DRO literature (e.g., Mohajerin Esfahani and Kuhn 2018), and it ensures sufficiently fast tail decay of the feature distribution.

4.1. Finite Sample Guarantees from the Optimal In-sample DRUC Value to the Optimal DRUC Value

We derive a high-probability finite-sample bound on the gap between the optimal in-sample DRUC value $v_n^{\widehat{\mathcal{X}}_n}$ and the optimal DRUC value $v_n^{\mathcal{X}}$. Unlike the asymptotic result in Theorem 2, which depends on a grid-based argument and thus the geometry of the feature space $\mathcal{X}$, our finite-sample analysis considers specific cases where $\mathcal{X}=\mathbb{R}^d$ or where the conditional distributions are supported on (possibly distinct) bounded hyperrectangles. These cases are detailed in the main results in this section. Note that the underlying proof techniques can be extended to other cases, such as when the feature space is polyhedral. However, these extensions require similar but more technical arguments, which we omit for brevity.

In the following proposition, we derive our finite-sample analysis using Lemma 2 and Theorem 2. For a gap tolerance η, we choose a radius R to control the tail probability $\beta(R)$ and a side length ϵ of the grid according to Theorem 2. Then by the contrapositive statement of Lemma 2, the probability that the gap is larger than η is bounded by the probability that one of the cubes does not contain any in-sample feature point. We can then obtain a finite sample guarantee on this probability.

Proposition 3 *Under Assumptions 1 and 3, suppose that the feature space is $\mathcal{X}=\mathbb{R}^d$, the marginal distribution $\mathbb{P}_0^{\mathcal{X}}$ of the true distribution $\mathbb{P}_0$ has a continuous density function g, and there exists a positive function $\phi:\mathbb{R}^+\to\mathbb{R}^+$ such that for any $R>0$ and any $\boldsymbol{x}\in B_{\mathbf{0}}(R)$, we have $g(\boldsymbol{x})\geq\phi(R)$. Then, under the conditions of Theorem 2, for any $\eta>0$, if*

$$n\geq\frac{-\kappa^2(3+\eta)^2\ln\frac{\alpha}{2}}{2\eta^2\theta^2}\vee\frac{\ln\frac{2(\pi d)^{d/2}(R+\kappa)^d(3+\eta)^d}{\Gamma\left(\frac{d}{2}+1\right)\alpha(\eta\theta)^d}}{\phi(R+\kappa)\left(\frac{\eta\theta}{(3+\eta)\sqrt{d}}\right)^d},\ \text{where } R=c\sqrt{-\ln\frac{\eta\theta}{2\kappa(3+\eta)}},$$

then we obtain the following finite-sample guarantee: $\mathbb{P}(v_n^{\mathcal{X}}-v_n^{\widehat{\mathcal{X}}_n}\geq\eta)\leq\alpha$.

Proof: See Appendix A.14. □

The above proposition provides a general statement for the case that the true distribution is supported on $\mathbb{R}^d$, including the Gaussian distribution. Besides this unbounded case, we also derive a finite-sample guarantee for a bounded case, where the support of the conditional true distribution of each label is a rectangle. In this case, we do not require the tail probability. The proof idea is similar to the proof of Proposition 3. But in this case, we adjust the side length of each small cube so that each cube evenly divides the corresponding side length of the support. Additionally, we create a grid division for each support rectangle and consider all cubes over the supports of all labels to ensure coverage of the entire feature space.

Proposition 4 *Under Assumption 1,suppose that for each $y \in \mathcal{Y}$, the conditional distribution of $(\boldsymbol{X}, Y) \sim \mathbb{P}_0$ given $Y = y$ is supported on a hyper-rectangle $H_y = \prod_{j=1}^{d}(a_{y,j}, b_{y,j})$. Let the marginal distribution $\mathbb{P}_0^{\mathcal{X}}$ have a continuous density function g, and suppose that the side lengths of all hyper-rectangles are uniformly bounded, i.e., $L \le b_{y,j} - a_{y,j} \le U$ for all $j \in [d]$ and $y \in \mathcal{Y}$. Furthermore, suppose there exists a constant $\zeta > 0$ such that $g(\boldsymbol{x}) \ge \zeta$ for all $\boldsymbol{x} \in \bigcup_{y\in\mathcal{Y}} H_y$. Then, under the conditions of Theorem 2, given any $\eta > 0$, if*

$$n \ge \frac{\ln \frac{m}{\alpha} + d \ln \frac{(1+\eta)U\sqrt{d}}{\eta\theta}}{\zeta \left(\frac{\eta\theta L}{L\sqrt{d}+\eta(L\sqrt{d}+\theta)} \right)^d},$$

then we have $\mathbb{P}(v_n^{\mathcal{X}} - v_n^{\widehat{\mathcal{X}}_n} \ge \eta) \le \alpha$.

Proof: See Appendix A.15. □

It is worth noting that the grid-based partitioning technique used to provide the finite-sample guarantee can be adapted to accommodate various shapes of the feature space. For instance, a polar coordinate-based partitioning scheme can be employed when the feature space is a Euclidean ball, while facet-aligned hypercube divisions are suitable for polyhedral feature spaces. Although these extensions follow similar proofs, we omit the derivations of their sample sizes for brevity.

4.2. Finite Sample Guarantees from the Optimal In-sample DRUC Value to the Optimal UC Value

We next strengthen Theorem 3 to obtain a finite-sample guarantee for the in-sample DRUC value $v_n^{\widehat{\mathcal{X}}_n}$ to be lower bounded by the optimal UC value v_{UC} within a tolerance level η, with high probability.

Proposition 5 *Under Assumptions 1, 3, and the conditions of Theorem 3, given any $\eta > 0$, when the sample size satisfies*

$$n \ge \frac{2}{\eta^2} \left(\frac{(\pi d)^{d/2} (c\sqrt{-\ln \frac{\eta}{4}} + \theta)^d}{\Gamma\left(\frac{d}{2}+1\right)\theta^d} \ln m - \ln \alpha \right),$$

then we have $\mathbb{P}(v_n^{\widehat{\mathcal{X}}_n} \le v_{UC} - \eta) \le \alpha$.

Proof: See Appendix A.16. □

4.3. Finite Sample Guarantees from the Optimal DRUC Value to the Optimal UC Value

To further refine the result in Theorem 4, we derive explicit conditions on the Wasserstein radius θ and the required sample size n such that the optimal DRUC value $v_n^{\mathcal{X}}$ is upper-bounded by v_{UC} within a tolerance η.

Proposition 6 *Under Assumptions 1, 2 and the conditions of Theorem 4, given any $\eta \in (0,1)$, if $\gamma(r) \leq \frac{\eta}{3}$, $\theta = \frac{\eta r}{3}$, and $n \geq \frac{-9\ln\alpha}{2\eta^2}$, then we have $\mathbb{P}(v_n^{\mathcal{X}} \geq v_{UC} + \eta) \leq \alpha$.*

Proof: See Appendix A.17. □

Because of the complexity of the function $\gamma(r)$, it is generally difficult to characterize its closed-form relationship with r. We provide insights into this relationship by studying two representative cases: the Gaussian distribution and the uniform distribution, which we analyze in the next two propositions.

Proposition 7 *Suppose that the feature space $\mathcal{X} = \mathbb{R}^d$, and the conditional random variable $(\boldsymbol{X}|Y=y)_{(\boldsymbol{X},Y)\sim\mathbb{P}_0}$ follows a multivariate normal distribution $\mathcal{N}(\boldsymbol{\mu}_y, \boldsymbol{\Sigma})$ for each $y \in \mathcal{Y}$, where the mean vectors $\{\boldsymbol{\mu}_y\}_{y\in\mathcal{Y}}$ are distinct. Denote the smallest eigenvalue of Σ by σ_d^2. Then we have*

(i) the quantity $\gamma(r) \leq \frac{m(m-1)r}{\sqrt{2\pi}\sigma_d}$;

(ii) the parameter θ in Proposition 6 can be chosen as $\theta = \frac{\eta^2\sqrt{2\pi}\sigma_d}{9m(m-1)}$; and

(iii) the least sample size n needed to guarantee that

$$\mathbb{P}(v_{UC} - \eta \leq v_n^{\widehat{\mathcal{X}}_n} \leq v_n^{\mathcal{X}} \leq v_{UC} + \eta) \geq 1 - 2\alpha,$$

is $n = \mathcal{O}(C^d d^{d/2}(d + \ln(\eta^{-1}))^{d/2}\eta^{-2d-2}m^{2d+1}\ln m + \eta^{-2}\ln(\alpha^{-1}))$, where C is a constant that does not depend on m, η, α and d.

Proof: See Appendix A.18. □

Proposition 8 *Suppose that for every $y \in \mathcal{Y}$, the underlying true distribution $(\boldsymbol{X}, Y) \sim \mathbb{P}_0$ conditioned on $Y = y$ is a uniform distribution supported on a hyper-rectangle $H_y = \prod_{j=1}^d [a_{y,j}, b_{y,j}]$. Suppose that there is a uniform bound $L \leq b_{y,j} - a_{y,j} \leq U$ on the side length for all $j \in [d]$ and $y \in \mathcal{Y}$. Then when $r \leq \frac{\ln 2}{2d}L$, we have*

(i) the quantity $\gamma(r) \leq \frac{8md}{L}r$;

(ii) the parameter θ in Proposition 6 can be chosen as

$$\theta = \frac{\eta^2 L}{72md};$$

(iii) the least sample size n needed to guarantee that

$$\mathbb{P}(v_{UC} - \eta \leq v_n^{\widehat{\mathcal{X}}_n} \leq v_n^{\mathcal{X}} \leq v_{UC} + \eta) \geq 1 - 2\alpha,$$

is $n = \mathcal{O}(C^d d^{3d/2}\eta^{-2d-2}m^{d+1}\ln m + \eta^{-2}\ln(\alpha^{-1}))$, where C is a constant that does not depend on m, η, α, and d.

Proof: See Appendix A.19. □

4.4. Summary of Results and Discussions

We summarize the finite-sample guarantee results in Figure 5. We remark the following:

(i) The required sample size scales proportionally to the d-exponential and d^d terms. The d-exponential term is due to the complexity of the universal policy space, and is also seen in the generalization bound in Zhang et al. (2024). The d^d term arises from the grid partition of the feature space, which contains $\mathcal{O}(d^d)$ grids. To derive a finite-sample guarantee, we bound the probability over each grid and apply a union bound, which leads to this sample complexity. Note that standard concentration inequalities are generally not applicable, as the classifier policy space is infinite-dimensional.

(ii) By definition, we always have $v_n^{\mathcal{X}} \geq v_n^{\widehat{\mathcal{X}}_n}$. To obtain the reverse inequality, Propositions 3 and 4 show that very mild conditions—namely, Assumptions 1 and 3, along with a lower bound on the probability density function—are sufficient to ensure the finite-sample guarantees.

(iii) Since incorporating distributional robustness leads to a larger objective value, we only require Assumptions 1 and 3 to establish the inequality $v_n^{\widehat{\mathcal{X}}_n} \gtrapprox v_{UC}$ in Proposition 5.

(iv) To establish the upper bound $v_n^{\mathcal{X}} \lessapprox v_{UC}$, if the function $\gamma(r)$—which characterizes the measure of the decision boundaries of the true classifier—is easy to bound, then only Assumptions 1 and 2 are needed. In this case, the finite-sample guarantee is relatively tight, since standard concentration inequalities are applicable.

(v) However, Assumptions 1, 2, and 3 together may still be insufficient to directly bound the function $\gamma(r)$. We believe that $\gamma(r)$ is, in general, independent of the feature dimension and instead depends on the true underlying distribution $\mathbb{P}_0$. To this end, Propositions 7 and 8 provide upper bounds on $\gamma(r)$ in the cases of Gaussian and uniform distributions, respectively.

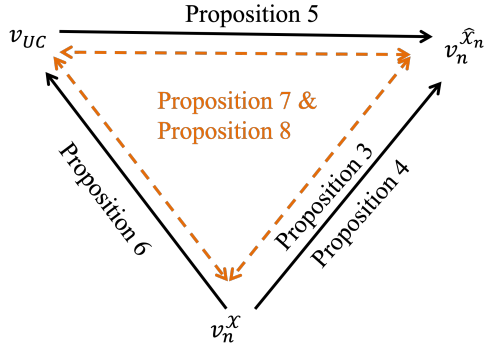


Figure 5 Summary of finite-sample guarantee results. A solid arrow denotes an approximate inequality ($\lessapprox$), while a bidirectional dashed arrow denotes an approximate equality ($\simeq$). Each proposition requires a specific assumption on the true distribution $\mathbb{P}_0$, as stated in the corresponding proposition.

5. Mixed-Integer Linear Programming (MILP) Formulation of the In-Sample DRUC Problem and Its Linear Programming Relaxation

To address the in-sample DRUC problem, we present a mixed-integer linear programming (MILP) reformulation. We show the equivalence between the MILP formulation and the in-sample DRUC problem, propose an approximation algorithm, and derive its approximation ratio.

5.1. MILP Formulation

Let us denote $d_{ij} = \|\widehat{\boldsymbol{x}}^i - \widehat{\boldsymbol{x}}^j\|$ for each $i, j \in [n]$. We first derive an LP reformulation for the in-sample DRUC value for a fixed in-sample policy $\widehat{f} \in \widehat{\mathcal{F}}_n$. Define w_{iy} as an indicator variable such that $w_{iy} = 1$ if $\widehat{f}(\widehat{\boldsymbol{x}}^i) = y$ and $w_{iy} = 0$ otherwise. Recall from Theorem 1 that the in-sample DRUC value $v_n^{\widehat{\mathcal{X}}_n}(\widehat{f})$ corresponds to the number of ascending transportation costs $s^n_{(i)}(\widehat{f})$ needed to reach a cumulative value $n\theta$, divided by n. This is equivalent to the maximum number of transportation costs that can be incorporated into the sum, divided by n.

By the definition (6b) of $s_i^n(\widehat{f})$, we have

$$s_i^n(\widehat{f}) = \min\{\kappa, \mathrm{dist}(\widehat{\boldsymbol{x}}^i, \widehat{f}^{-1}(\mathcal{Y} \setminus \{\widehat{y}^i\}))\}.$$

Thus, the transportation cost associated with $\widehat{\boldsymbol{x}}^i$ is either κ or d_{ij} for some j such that $\widehat{f}(\widehat{\boldsymbol{x}}^j) \neq \widehat{y}^i$ (equivalently, $w_{j\widehat{y}^i} = 0$).

We introduce decision variables r_{i0} and r_{ij} for each $i, j \in [n]$, where r_{i0} corresponds to selecting the cost κ for $\widehat{\boldsymbol{x}}^i$, and r_{ij} corresponds to selecting the cost d_{ij} if $\widehat{f}(\widehat{\boldsymbol{x}}^j) \neq \widehat{y}^i$. Under this setup, the cumulative transportation constraint becomes

$$\sum_{i\in[n]} \kappa r_{i0} + \sum_{i,j\in[n]} d_{ij} r_{ij} \leq \theta.$$

Moreover, to ensure that d_{ij} is selected only if $\widehat{f}(\widehat{\boldsymbol{x}}^j) \neq \widehat{y}^i$, we set the objective coefficient of r_{ij} to be $(1 - w_{j\widehat{y}^i})$. This leads to the following LP reformulation for computing $v_n^{\widehat{\mathcal{X}}_n}(\widehat{f})$.

Theorem 5 *Given any $\kappa \in \mathbb{R}_{++}, \theta \in \mathbb{R}_{++}$, $n \in \mathbb{Z}^+$, and an in-sample policy $\widehat{f} \in \widehat{\mathcal{F}}_n$, for each $i \in [n]$ and $y \in [m]$, define*

$$w_{iy} = \begin{cases} 1, & \text{if } \widehat{f}(\widehat{\boldsymbol{x}}^i) = y; \\ 0, & \text{if } \widehat{f}(\widehat{\boldsymbol{x}}^i) \neq y. \end{cases}$$

Then, the in-sample DRUC value $\widehat{v}_n(\widehat{f})$ can be computed by solving the following linear program:

$$\begin{aligned} \widehat{v}_n(\widehat{f}) = \max\ & \sum_{i,j\in[n]} (1 - w_{j\widehat{y}^i}) r_{ij} + \sum_{i\in[n]} r_{i0} \\ \text{s.t.}\ & \sum_{i\in[n]} \kappa r_{i0} + \sum_{i,j\in[n]} d_{ij} r_{ij} \leq \theta, \\ & r_{i0} + \sum_{j\in[n]} r_{ij} \leq \frac{1}{n}, \qquad \forall i \in [n], \\ & r_{i0} \geq 0, r_{ij} \geq 0, \qquad \forall i, j \in [n]. \end{aligned} \tag{12}$$

Given the maximization LP formulation (12), we then take the dual and minimize over all possible in-sample policies $\widehat{f} \in \widehat{\mathcal{F}}_n$ to obtain the optimal in-sample DRUC value $v_n^{\widehat{\mathcal{X}}_n}$.

Corollary 4 *The optimal in-sample DRUC value $v_n^{\widehat{\mathcal{X}}_n}$ is equal to the optimal value of the following MILP formulation:*

$$\begin{array}{ll} \min\ \theta\alpha + \sum\limits_{i\in[n]} \dfrac{1}{n}\lambda_i & \\ \text{s.t.} \ \ d_{ij}\alpha + \lambda_i + w_{j\widehat{y}^i} \geq 1, & \forall i,j \in [n], \\ \quad \kappa\alpha + \lambda_i \geq 1, & \forall i \in [n], \\ \quad \sum_{y\in[m]} w_{iy} = 1, & \forall i \in [n], \\ \quad \alpha \geq 0, \lambda_i \geq 0, w_{iy} \in \{0,1\}, & \forall i \in [n], y \in [m]. \end{array} \tag{13}$$

Proof: See Appendix A.21 □

In the MILP formulation (13), the number of binary decision variables is equal to the product of the number of in-sample data points n and the number of distinct labels m. This structure ensures that the formulation remains solvable and computationally scalable when n is on the order of hundreds and m is within a moderate range, such as $m \leq 10$.

5.2. Approximation Algorithm through Linear Programming Relaxation

For larger instances where n exceeds several thousand, solving the MILP formulation (13) to optimality becomes computationally challenging, particularly during parameter tuning processes. To address this challenge, we propose an alternative approximation algorithm, which we refer to as "MaxLin". The key idea is first to solve the LP relaxation of the MILP formulation (13), where the binary variables $\{w_{iy}\}_{i\in[n],y\in[m]}$ are relaxed to be continuous.

Let $\{w_{iy}^*\}_{i\in[n],y\in[m]}$ denote an optimal solution to the LP relaxation. We then construct an integer solution by assigning each in-sample feature point to the label corresponding to the largest relaxed value. Specifically, we define the integer solution $\{\widetilde{w}_{iy}\}_{i\in[n],y\in[m]}$ as

$$\widetilde{w}_{iy} = \begin{cases} 1, & \text{if } y = \min\left(\arg\max_{y'\in[m]} w_{iy'}^*\right); \\ 0, & \text{otherwise.} \end{cases} \tag{14}$$

We provide a theoretical approximation ratio of the objective value associated with the constructed integer solution $\{\widetilde{w}_{iy}\}_{i\in[n],y\in[m]}$ with respect to the optimal value of the MILP in the following theorem.

Theorem 6 *Let $\{\widetilde{w}_{iy}\}_{i\in[n],y\in[m]}$ be the solution obtained through MaxLin, and its corresponding objective value in* (13) *be $\widetilde{v}_n$. Then we have $v_n^{\widehat{\mathcal{X}}_n} \leq \widetilde{v}_n \leq 2v_n^{\widehat{\mathcal{X}}_n}$.*

Proof: See Appendix A.22. □

The approximation algorithm MaxLin is particularly effective for solving the large-scale in-sample DRUC problem. To assess its effectiveness, we conduct a thorough evaluation using both synthetic and real datasets in the next section.

6. Numerical Experiments

In this section, we construct synthetic datasets as follows for numerical experiments.

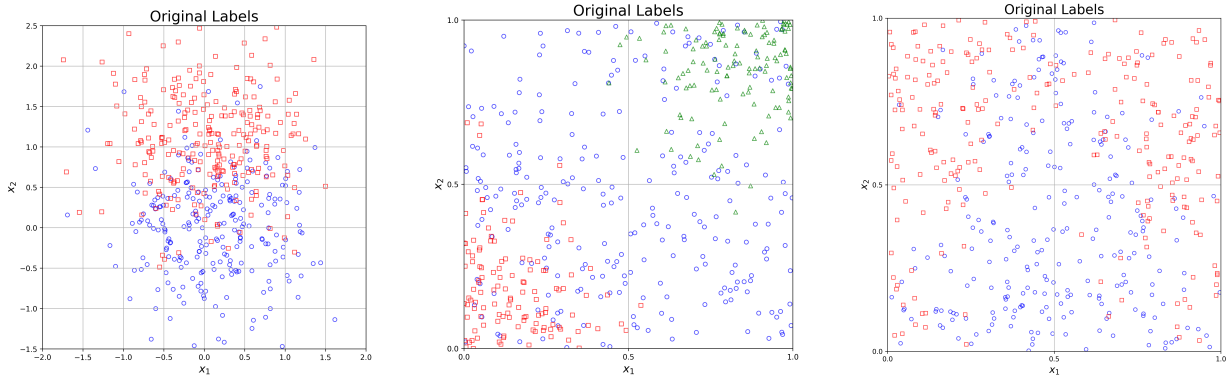
(i) For the Gaussian Dataset, we generate two labels: 0 and 1, with equal probability. Feature points with label 0 are normally distributed, centered at $(0,0)^\top$, with the covariance matrix $0.6I_2$; feature points with label 1 are also normally distributed with the same covariance matrix, but centered at $(0,1)^\top$.

(ii) For the Two-Corner Dataset, we generate three labels: 0, 1, and 2. Feature points with label 0 are generated with probability 0.5 and are uniformly distributed on $[0,1]^2$. Feature points with label 1 are generated with probability 0.25 and follow a folded normal distribution (folded into $[0,1]^2$) with center $(0,0)^\top$ and covariance $0.2\boldsymbol{I}_2$. Feature points with label 2 are similarly generated as label 1, except that they are centered at $(1,1)^\top$.

(iii) For Wave Dataset, we first generate feature points uniformly on $[0,1]^2$. Then we assign the label based on the following conditional probability:

$$\mathbb{P}_0(Y=1|X=(x_1,x_2))=\frac{1}{1+\exp(5(x_2-\sin^2(\pi x_1)))}.$$

This dataset has a soft boundary $x_2=\sin^2(\pi x_1)$ between two labels, and is therefore highly nonlinearly separable.



(a) Gaussian Dataset (b) Two-Corner Dataset (c) Wave Dataset

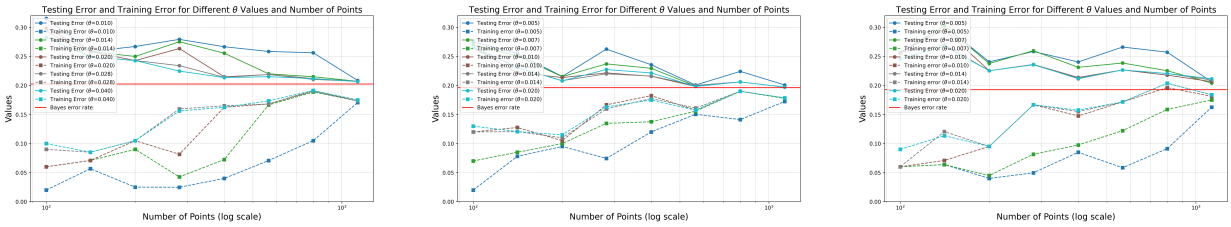
Figure 6 Illustration of synthetic datasets. There are two labels in the Gaussian Dataset and the Wave Dataset, and three labels in the Two-Corner Dataset.

Figure 6 illustrates the distribution of the feature points with different labels of the synthetic datasets. We validate the theoretical results and examine the performance of our model on the aforementioned synthetic datasets, as well as real-world UCI and image datasets.

6.1. Convergence to the Optimal UC Value

We validate the convergence theorem presented in Section 3.2 through numerical experiments. Specifically, we generate a set of in-sample data points (training data) of size $\lfloor 100 \times 2^{0.5\times\{0,1,\cdots,8\}} \rfloor$ and a fixed set of 10,000 out-of-sample data points (testing data) based on the synthetic distribution setting. We set the parameters $\kappa = 0.25$ and $\theta = 0.01 \times 2^{\frac{i}{2}}$ for $i \in [5]$, and solve the MILP (13) under various in-sample sizes and values of θ. To approximate the value of v_{UC}, we relabel the 10,000 out-of-sample data points using the label with the highest theoretical conditional probability density at each corresponding feature point. Additionally, we track both the in-sample misclassification risk (training error) and the out-of-sample misclassification risk (testing error) throughout the experiments.

Figure 7 displays the out-of-sample misclassification rate (testing error) and in-sample misclassification rate (training error) across different sample sizes and values of θ on synthetic datasets. The results demonstrate that as the number of data points increases, both the in-sample and out-of-sample misclassification rates converge to the optimal UC value v_{UC}. This observation confirms that the policies obtained through our in-sample DRUC formulation effectively mitigate overfitting and exhibit reliable generalization performance.



(a) Gaussian Dataset (b) Two-Corner Dataset (c) Wave Dataset

Figure 7 Convergence of the testing errors to the optimal UC value for different datasets.

6.2. Strength of the Approximation Algorithm MaxLin

To verify the effectiveness of the approximation algorithm MaxLin, we compare its performance with that of directly solving the MILP formulation on synthetic datasets under different parameter settings. The comparison results are reported in Table 1.

Particularly, for each case, we compute the testing accuracy by generating 1,000 out-of-sample data points from the same distribution as the training set and calculating the corresponding accuracy rate. The objective gap of the MaxLin solution is defined as the relative error between the MaxLin objective value and the best lower bound of the MILP formulation:

$$\text{ObjGap of MaxLin} = \frac{\text{Obj of MaxLin} - \text{Best Lower Bound}}{\text{Best Lower Bound}} \times 100\%.$$

Table 1 Performance comparison between MaxLin and MILP formulation.

Dataset	#Points	κ	ρ	Obj Gap of MaxLin	Testing Acc of MaxLin	Testing Acc of MILP	Time of MaxLin (s)	Time of MILP (s)
Gaussian	200	0.25	0.1	$7.6\% \pm 1.9\%$	$74.0\% \pm 1.8\%$	$\mathbf{76.1\% \pm 1.9\%}$	0.10 ± 0.01	0.45 ± 0.30
		0.25	0.05	$11.4\% \pm 3.4\%$	$73.8\% \pm 2.1\%$	$\mathbf{76.9\% \pm 1.9\%}$	0.11 ± 0.02	0.36 ± 0.18
		0.25	0.025	$6.1\% \pm 3.1\%$	$73.6\% \pm 2.3\%$	$\mathbf{74.6\% \pm 3.1\%}$	0.12 ± 0.06	1.16 ± 1.01
		0.5	0.1	$8.8\% \pm 4.5\%$	$77.8\% \pm 1.4\%$	$\mathbf{78.6\% \pm 1.2\%}$	0.23 ± 0.03	6.10 ± 4.30
		0.5	0.05	$7.1\% \pm 3.1\%$	$77.0\% \pm 2.2\%$	$\mathbf{77.5\% \pm 1.7\%}$	0.29 ± 0.06	19.9 ± 16.3
		0.5	0.025	$5.6\% \pm 2.3\%$	$76.0\% \pm 2.2\%$	$\mathbf{76.1\% \pm 2.4\%}$	0.28 ± 0.10	19.6 ± 18.4
Two-Corner	200	0.1	0.05	$5.2\% \pm 1.7\%$	$76.4\% \pm 1.4\%$	$\mathbf{77.8\% \pm 1.5\%}$	0.20 ± 0.11	0.45 ± 0.27
		0.1	0.025	$11.5\% \pm 3.6\%$	$76.0\% \pm 1.8\%$	$\mathbf{78.3\% \pm 1.6\%}$	0.12 ± 0.01	0.31 ± 0.08
		0.1	0.0125	$9.6\% \pm 3.6\%$	$76.3\% \pm 1.7\%$	$\mathbf{77.7\% \pm 1.9\%}$	0.12 ± 0.01	0.57 ± 0.37
		0.2	0.05	$8.7\% \pm 3.5\%$	$78.8\% \pm 1.3\%$	$\mathbf{78.9\% \pm 1.8\%}$	0.23 ± 0.02	6.60 ± 5.80
		0.2	0.025	$6.9\% \pm 2.5\%$	$78.4\% \pm 1.8\%$	$\mathbf{78.4\% \pm 1.5\%}$	0.28 ± 0.05	18.6 ± 24.7
		0.2	0.0125	$4.9\% \pm 2.4\%$	$\mathbf{78.3\% \pm 2.1\%}$	$78.0\% \pm 2.5\%$	0.34 ± 0.18	14.6 ± 14.4
Wave	200	0.1	0.05	$5.2\% \pm 1.5\%$	$74.0\% \pm 1.7\%$	$\mathbf{75.4\% \pm 1.5\%}$	0.11 ± 0.01	0.35 ± 0.18
		0.1	0.025	$8.7\% \pm 2.8\%$	$74.5\% \pm 2.3\%$	$\mathbf{76.4\% \pm 2.1\%}$	0.11 ± 0.01	0.44 ± 0.19
		0.1	0.0125	$6.6\% \pm 2.4\%$	$73.2\% \pm 1.6\%$	$\mathbf{74.7\% \pm 1.6\%}$	0.12 ± 0.01	0.82 ± 0.36
		0.2	0.05	$9.0\% \pm 2.8\%$	$\mathbf{76.4\% \pm 1.8\%}$	$75.2\% \pm 2.8\%$	0.31 ± 0.14	28.7 ± 19.2
		0.2	0.025	$5.4\% \pm 2.0\%$	$\mathbf{77.2\% \pm 1.6\%}$	$76.3\% \pm 3.1\%$	0.35 ± 0.14	59.3 ± 55.5
		0.2	0.0125	$5.1\% \pm 1.8\%$	$74.7\% \pm 1.8\%$	$\mathbf{74.8\% \pm 2.0\%}$	0.27 ± 0.02	16.6 ± 10.2

Table 1 summarizes the performance of MaxLin on the synthetic datasets across different parameter settings. The results indicate that the objective value of the approximate solution obtained by MaxLin is within 10% of the best MILP lower bound. This demonstrates the effectiveness of MaxLin as an approximation algorithm for the MILP formulation. Furthermore, by comparing the testing accuracies of MaxLin and MILP, we observe that the solutions from MaxLin achieve testing accuracies that are very close to, and in some cases even better than, those of the MILP solutions. These findings suggest that MaxLin provides approximate solutions with nearly equivalent out-of-sample performance. Given that solving the MILP formulation becomes computationally difficult for datasets with thousands of data points, we adopt MaxLin as the default approach for obtaining in-sample solutions in the subsequent numerical studies on both large-scale synthetic and real-world datasets.

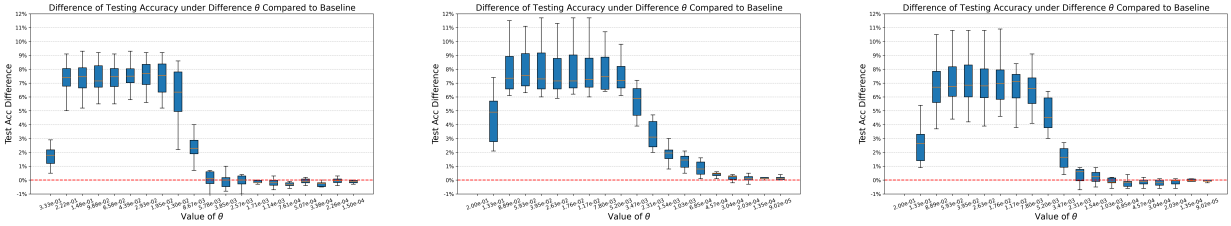
6.3. Model Performance on Synthetic Datasets

Similar to the previous two experiments, in the synthetic setting, we generate training and testing datasets with 1,000 instances each, sampled from the synthetic distribution, and compute the testing accuracies on the corresponding testing datasets.

Performance regarding the parameter θ: We first conduct a series of experiments to evaluate the performance of our proposed method regarding the parameter θ. It is worth noting that we use the "closest-point" method as our baseline approach, where the label of a testing feature point is predicted to be the same as that of its nearest training point. This method is theoretically equivalent to the case where θ is sufficiently small, as established in Corollary 3.

Figure 8 illustrates how the testing accuracy varies with the parameter θ across the synthetic data. We observe a consistent pattern across all plots: when θ is very small, the testing accuracy

closely matches that of the baseline, and label assignments deviate only slightly from their original values, which is consistent with Corollary 3. As θ increases, the testing accuracy initially increases, reaching a maximum near a certain threshold. However, when θ approaches κ, the testing accuracy decreases again. This U-shaped pattern is observed consistently across all synthetic settings. These results suggest that optimal performance is typically achieved when the parameter θ is chosen to be approximately one-tenth to one-third of the parameter κ.



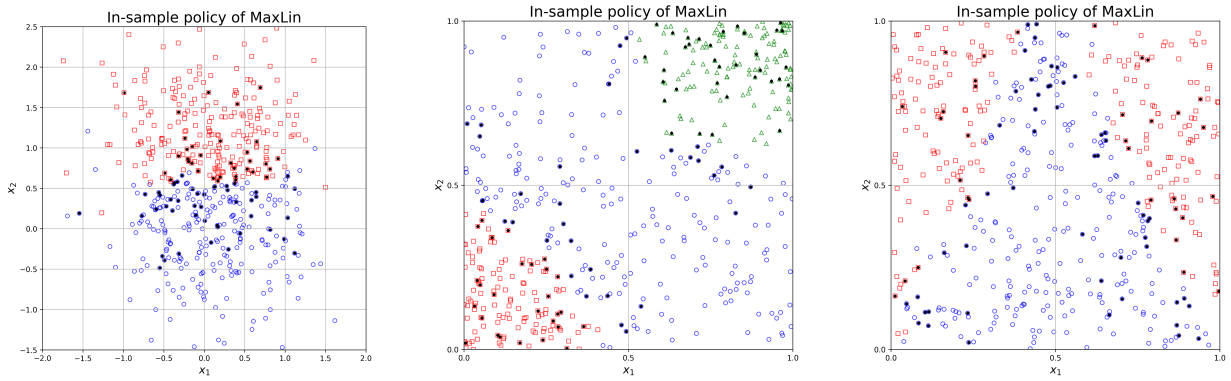
(a) Gaussian Dataset (b) Two-Corner Dataset (c) Wave Dataset

Figure 8 Boxplot of accuracy gaps of MaxLin under different θ compared with the baseline method. The red dashed line represents the performance of the closest-point (baseline) method. A positive accuracy gap means that MaxLin with the parameter has higher testing accuracy than the closest-point method, and vice versa.

Performance on the synthetic datasets: We then evaluate the performance of our method on synthetic datasets using the optimal parameter settings, and compare it with the baseline (closest-point) method and logistic regression. Table 2 presents the testing accuracies of each method. The results show that our method significantly outperforms the baseline method by effectively relabeling the in-sample feature points. For highly linearly separable data (e.g., Gaussian), our method performs slightly better than logistic regression. In other cases, particularly when the data exhibits highly nonlinear decision boundaries (e.g., Wave), our method outperforms logistic regression. These findings indicate that our method generalizes well and offers greater flexibility than linear models. We also visualize the approximate optimal in-sample policies obtained by MaxLin in Figure 9. The results show that the proposed MaxLin method tends to relabel feature points that are surrounded by points of a different label. This behavior allows the closest-label extension of the in-sample policy to achieve strong out-of-sample performance.

Table 2 The testing accuracies of different methods on synthetic data.

Dataset	#Instances	#Classes	MaxLin	Closest Point	LogReg
Gaussian	1000	2	$\mathbf{81.04\% \pm 0.38\%}$	$73.49\% \pm 1.23\%$	$80.71\% \pm 0.30\%$
Two Corner	1000	3	$\mathbf{80.51\% \pm 1.19\%}$	$72.48\% \pm 1.25\%$	$80.05\% \pm 0.98\%$
Wave	1000	2	$\mathbf{79.37\% \pm 1.35\%}$	$72.30\% \pm 1.76\%$	$67.32\% \pm 1.32\%$



(a) Gaussian Dataset (b) Two-Corner Dataset (c) Wave Dataset

Figure 9 Approximate optimal in-sample policies obtained by MaxLin. Different colors and shapes represent different policy assignments. Solid dots indicate feature points whose assigned labels differ from their original labels.

6.4. Model Performance on UCI and Image Datasets

We further evaluate our method on real-world UCI and image datasets using a neural network as the feature extractor. We employ the Feedforward Neural Network (FFNN) for UCI tabular datasets and the Convolutional Neural Network (CNN) for image datasets. We compare the following five methods based on neural networks on the real datasets: (i) A vanilla network trained with standard cross-entropy loss; (ii) A network with a low-dimensional feature embedding layer; (iii) A network with a low-dimensional feature embedding layer trained jointly with center loss (Wen et al. 2016); (iv) A method that uses network (iii) for feature extraction and applies the baseline (closest-point) classifier to the extracted features; and (v) Our proposed method, which uses network (iii) for feature extraction and applies MaxLin to the extracted features. Hyperparameters are selected based on validation accuracy. For methods with a feature embedding structure, we insert an additional linear layer that extracts two features, followed by a batch normalization layer before the final output layer. All models are trained for 200 epochs using the Adam optimizer with an initial learning rate of 0.01, which is halved every 50 epochs. When training the network with center loss, we set the center loss weight to be 1.5. The learning rate of the center parameters is chosen based on the number of classes: 0.2 for 2-class tasks, 0.3 for 3-class tasks, 0.4 for 4-class tasks, 0.5 for 5-class tasks. These settings are consistent with the recommendations in Wen et al. (2016).

Performance on the UCI datasets: We choose 3 UCI datasets of binary classification tasks, respectively credit card (Yeh 2009), rice (ric 2019), and spambase (Hopkins et al. 1999) for experiments. For consistency and comparability, we adopt a typical single-layer FFNN consisting of a fully connected layer, a ReLU activation function, a dropout layer, and an output layer. The fully

connected layer has an output dimension of 8, and dropout is applied with a rate of 0.4. For each dataset, we conduct two types of experiments. In the *balanced* setting, we randomly select 500 samples from each class for both the training and testing sets. In the *imbalanced* setting, we use 200 samples from one class and 800 from another (for the rice dataset, we select 300 and 700 samples due to data insufficiency).

The testing accuracies of different methods on the UCI datasets are reported in Table 3. Our proposed MaxLin method consistently outperforms the baseline approach (i.e., the closest-point method). Among all methods, MaxLin achieves the best performance in five out of six cases and is nearly tied with the best-performing method in the remaining one. These results indicate that our method is competitive with or superior to all other alternatives. They also highlight the effectiveness and practical value of MaxLin for real-world classification tasks.

Table 3 The testing accuracies of different methods on UCI datasets.

Dataset	Balanced	Vanilla FFNN	Low-dim FFNN	Centerloss FFNN	Closest Point	MaxLin
Credit Card	balanced	$68.48\% \pm 0.80\%$	$68.05\% \pm 1.44\%$	$68.95\% \pm 1.10\%$	$62.65\% \pm 1.22\%$	$\mathbf{69.15\% \pm 1.25\%}$
	unbalanced	$82.47\% \pm 0.91\%$	$82.42\% \pm 1.09\%$	$82.22\% \pm 0.89\%$	$75.33\% \pm 1.54\%$	$\mathbf{83.02\% \pm 1.56\%}$
Rice	balanced	$92.19\% \pm 0.80\%$	$92.18\% \pm 0.70\%$	$\mathbf{92.32\% \pm 0.60\%}$	$88.95\% \pm 1.08\%$	$92.25\% \pm 0.72\%$
	unbalanced	$93.25\% \pm 0.55\%$	$93.31\% \pm 0.54\%$	$\mathbf{93.32\% \pm 0.65\%}$	$90.40\% \pm 0.67\%$	$\mathbf{93.32\% \pm 0.64\%}$
Spam	balanced	$93.34\% \pm 0.70\%$	$\mathbf{93.80\% \pm 0.77\%}$	$93.79\% \pm 0.97\%$	$91.64\% \pm 1.04\%$	$\mathbf{93.80\% \pm 0.90\%}$
	unbalanced	$94.45\% \pm 0.68\%$	$94.81\% \pm 0.72\%$	$94.75\% \pm 0.76\%$	$93.86\% \pm 0.83\%$	$\mathbf{95.40\% \pm 0.47\%}$

Performance on the image datasets: We experiment on three image datasets: CIFAR-10 (Krizhevsky et al. 2009), Fashion-MNIST (Xiao et al. 2017), and SVHN (Netzer et al. 2011). To ensure comparability, we adopt a common CNN backbone consisting of two sequences of `Convolution`–`BatchNorm`–`ReLU`–`MaxPooling`–`Dropout`, followed by a `Linear`–`BatchNorm`–`ReLU`–`Dropout` sequence and a final output layer. The two convolutional layers have 32 and 64 output channels, respectively, with a kernel size of 3 and a stride of 1. The fully connected linear layer has an output dimension of 64, and the dropout rate is set to 0.25. We conduct experiments for 3-class, 4-class, and 5-class tasks. For each class, we randomly select 300 images for the training set and use the full test set for evaluation.

Table 4 presents the testing accuracies of the five methods. To better illustrate the instance-wise performance differences, we also report the accuracy gaps between our method and the baselines using boxplots in Figure 10. The results show that our proposed MaxLin method significantly outperforms the closest-point baseline in all cases. MaxLin achieves the highest accuracy in six out of nine experiments, and in the remaining three cases, its performance is comparable to that of the best-performing method. These findings indicate that our method achieves competitive, and often superior, performance compared to direct linear classifiers integrated within CNN architectures, such as those with feature embeddings or center loss. Overall, the results suggest that MaxLin is an effective classification strategy when used as the final-layer classifier in neural networks, demonstrating its ability to extract and exploit structure within the embedded feature space.

Table 4 The testing accuracies of different methods on image datasets.

Dataset	#Classes	Vanilla CNN	Low-dim CNN	Centerloss CNN	Closest Point	MaxLin
Cifar-10	3	$84.00\% \pm 1.26\%$	$82.54\% \pm 2.00\%$	$84.65\% \pm 1.47\%$	$81.54\% \pm 1.69\%$	$\mathbf{84.79\% \pm 1.05\%}$
	4	$75.31\% \pm 1.26\%$	$75.81\% \pm 0.87\%$	$76.82\% \pm 1.10\%$	$73.40\% \pm 1.18\%$	$\mathbf{77.08\% \pm 0.76\%}$
	5	$\mathbf{68.20\% \pm 2.14\%}$	$67.63\% \pm 3.12\%$	$67.15\% \pm 2.78\%$	$62.82\% \pm 2.52\%$	$68.09\% \pm 2.77\%$
SVHN	3	$92.72\% \pm 0.82\%$	$93.06\% \pm 0.97\%$	$93.78\% \pm 0.55\%$	$92.58\% \pm 0.70\%$	$\mathbf{93.80\% \pm 0.49\%}$
	4	$90.28\% \pm 0.98\%$	$90.49\% \pm 0.64\%$	$90.57\% \pm 0.97\%$	$89.34\% \pm 0.76\%$	$\mathbf{90.61\% \pm 0.93\%}$
	5	$\mathbf{89.05\% \pm 0.52\%}$	$88.35\% \pm 0.65\%$	$88.88\% \pm 0.91\%$	$86.73\% \pm 0.83\%$	$88.84\% \pm 1.05\%$
Fashion-MNIST	3	$96.44\% \pm 0.36\%$	$96.59\% \pm 0.33\%$	$\mathbf{96.82\% \pm 0.45\%}$	$96.40\% \pm 0.51\%$	$96.74\% \pm 0.43\%$
	4	$93.70\% \pm 0.55\%$	$94.00\% \pm 0.58\%$	$94.18\% \pm 0.48\%$	$93.28\% \pm 0.60\%$	$\mathbf{94.20\% \pm 0.38\%}$
	5	$88.95\% \pm 0.93\%$	$89.46\% \pm 1.09\%$	$89.39\% \pm 0.95\%$	$87.25\% \pm 1.33\%$	$\mathbf{89.66\% \pm 0.92\%}$

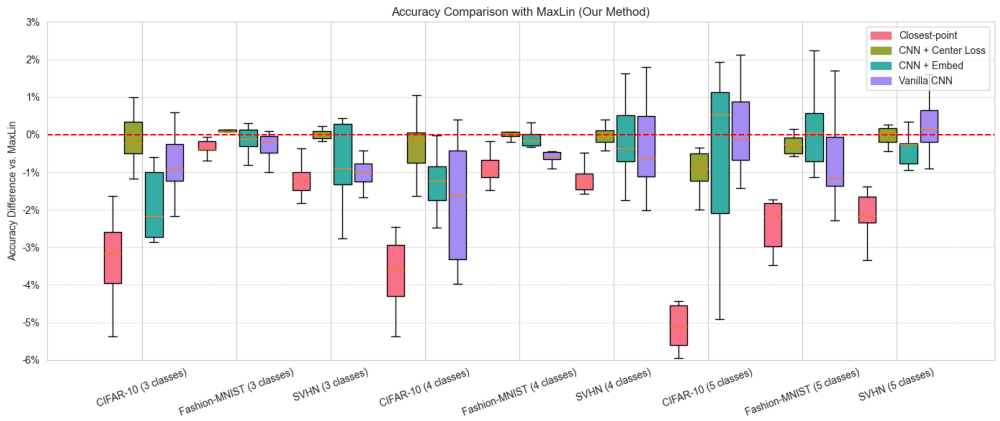


Figure 10 Boxplot of the accuracy gaps. The red dashed line represents the performance of MaxLin. A negative accuracy gap indicates that the alternative method performs worse than MaxLin, and vice versa.

7. Conclusion

This paper studied the multi-class Distributionally Robust Universal Classification (DRUC) problem, where a classifying policy can be any measurable function. To address the infinite-dimensional policy space of the DRUC formulation, we proposed an in-sample counterpart whose policy space is finite-dimensional. We proved asymptotic convergence relationships among the optimal Universal Classification (UC) value, the optimal DRUC value, and the optimal in-sample DRUC value, while also providing finite-sample guarantees under mild conditions. Additionally, we developed a mixed-integer linear programming formulation to efficiently solve the in-sample DRUC problem and proposed a 2-approximation algorithm. Through numerical experiments, we empirically validated the reliability of the proposed method, demonstrating its strong out-of-sample performance and ability to mitigate overfitting. Furthermore, we numerically showed that our method outperforms both the closest-point, logistic regression, and neural network methods on synthetic, UCI, and image datasets. A promising future direction is to relax the assumptions for proving the convergence.

References

(2019) Rice (Cammeo and Osmancik). UCI Machine Learning Repository, DOI: https://doi.org/10.24432/C5MW4Z.

Belbasi R, Selvi A, Wiesemann W (2023) It's all in the mix: Wasserstein machine learning with mixed features. *arXiv preprint arXiv:2312.12230* .

Billingsley P (1995) *Probability and Measure* (Wiley), 3rd edition.

Blanchet J, Chen L, Zhou XY (2022) Distributionally robust mean-variance portfolio selection with wasserstein distances. *Management Science* 68(9):6382–6410.

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

Breiman L (2017) *Classification and regression trees* (Routledge).

Carlsson JG, Behroozi M, Mihic K (2018) Wasserstein distance and the distributionally robust tsp. *Operations Research* 66(6):1603–1624.

Chang Z, Ding JY, Song S (2019) Distributionally robust scheduling on parallel machines under moment uncertainty. *European Journal of Operational Research* 272(3):832–846.

Du N, Liu Y, Liu Y (2020) A new data-driven distributionally robust portfolio optimization method based on wasserstein ambiguity set. *IEEE Access* 9:3174–3194.

Durrett R (2010) *Probability: Theory and Examples* (Cambridge University Press), 4th edition.

Feng Y, Liu YK, Chen Y (2022) Distributionally robust location–allocation models of distribution centers for fresh products with uncertain demands. *Expert Systems with Applications* 209:118180.

Folland GB (1999) *Real analysis: modern techniques and their applications* (John Wiley & Sons).

Gao R, Kleywegt A (2023a) Distributionally robust stochastic optimization with wasserstein distance. *Mathematics of Operations Research* 48(2):603–655, URL `http://dx.doi.org/10.1287/moor.2022.1275`.

Gao R, Kleywegt A (2023b) Distributionally robust stochastic optimization with wasserstein distance. *Mathematics of Operations Research* 48(2):603–655.

Ghosal S, Wiesemann W (2020) The distributionally robust chance-constrained vehicle routing problem. *Operations Research* 68(3):716–732.

Ho-Nguyen N, Wright SJ (2023) Adversarial classification via distributional robustness with wasserstein ambiguity. *Mathematical Programming* 198(2):1411–1447.

Hopkins M, Reeber E, Forman G, Suermondt J (1999) Spambase. UCI Machine Learning Repository, DOI: https://doi.org/10.24432/C53G6X.

Hosmer Jr DW, Lemeshow S, Sturdivant RX (2013) *Applied logistic regression* (John Wiley & Sons).

James G, Witten D, Hastie T, Tibshirani R, et al. (2013) *An introduction to statistical learning*, volume 112 (Springer).

Kannan R, Bayraksan G, Luedtke JR (2022) Data-driven sample average approximation with covariate information. *arXiv preprint arXiv:2207.13554* .

Kannan R, Bayraksan G, Luedtke JR (2024) Residuals-based distributionally robust optimization with covariate information. *Mathematical Programming* 207(1):369–425.

Kleywegt AJ, Shapiro A, Homem-de Mello T (2002) The sample average approximation method for stochastic discrete optimization. *SIAM Journal on optimization* 12(2):479–502.

Kolmogorov N A (1933) Sulla legge dei grandi numeri. *Giornale dell'Istituto Italiano degli Attuari* 4:83–91.

Krizhevsky A, Hinton G, et al. (2009) Learning multiple layers of features from tiny images .

Kuhn D, Esfahani PM, Nguyen VA, Shafieezadeh-Abadeh S (2019) Wasserstein distributionally robust optimization: Theory and applications in machine learning. *Operations research & management science in the age of analytics*, 130–166 (Informs).

Kuhn D, Shafiee S, Wiesemann W (2024) Distributionally robust optimization. URL `https://arxiv.org/abs/2411.02549`.

Laurent B, Massart P (2000) Adaptive estimation of a quadratic functional by model selection. *Annals of statistics* 1302–1338.

LeCun Y, Bengio Y, Hinton G (2015) Deep learning. *nature* 521(7553):436–444.

Lin Y (2002) Support vector machines and the bayes rule in classification. *Data Mining and Knowledge Discovery* 6:259–275.

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

Netzer Y, Wang T, Coates A, Bissacco A, Wu B, Ng AY, et al. (2011) Reading digits in natural images with unsupervised feature learning. *NIPS workshop on deep learning and unsupervised feature learning*, volume 2011, 4 (Granada).

Perakis G, Sim M, Tang Q, Xiong P (2023) Robust pricing and production with information partitioning and adaptation. *Management Science* 69(3):1398–1419.

Rahimian H, Mehrotra S (2022) Frameworks and results in distributionally robust optimization. *Open Journal of Mathematical Optimization* 3:1–85.

Shafieezadeh Abadeh S, Mohajerin Esfahani PM, Kuhn D (2015) Distributionally robust logistic regression. *Advances in neural information processing systems* 28.

Sun L, Xie W, Witten T (2023) Distributionally robust fair transit resource allocation during a pandemic. *Transportation science* 57(4):954–978.

Teschl G (2014) *Mathematical methods in quantum mechanics*, volume 157 (American Mathematical Soc.).

Wang L (2005) *Support vector machines: theory and applications*, volume 177 (Springer Science & Business Media).

Wang Y, Zhang Y, Tang J (2019) A distributionally robust optimization approach for surgery block allocation. *European Journal of Operational Research* 273(2):740–753.

Wen Y, Zhang K, Li Z, Qiao Y (2016) A discriminative feature learning approach for deep face recognition. *Computer vision–ECCV 2016: 14th European conference, amsterdam, the netherlands, October 11–14, 2016, proceedings, part VII 14*, 499–515 (Springer).

Xiao H, Rasul K, Vollgraf R (2017) Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747* .

Xie L, Gao R, Xie Y (2021) Robust hypothesis testing with wasserstein uncertainty sets. *arXiv preprint arXiv:2105.14348* .

Xin L, Goldberg DA (2022) Distributionally robust inventory control when demand is a martingale. *Mathematics of Operations Research* 47(3):2387–2414.

Xu H, Caramanis C, Mannor S (2009) Robustness and regularization of support vector machines. *Journal of machine learning research* 10(7).

Yang Z, Gao R (2022) Wasserstein regularization for 0–1 loss. *Optimization Online Preprint* .

Yeh IC (2009) Default of Credit Card Clients. UCI Machine Learning Repository, DOI: https://doi.org/10.24432/C55S3H.

Yin F, Zhao Y (2021) Optimizing vehicle routing via stackelberg game framework and distributionally robust equilibrium optimization method. *Information Sciences* 557:84–107.

Zhang L, Yang J, Gao R (2024) Optimal robust policy for feature-based newsvendor. *Management Science* 70(4):2315–2329.

Zhu S, Xie L, Zhang M, Gao R, Xie Y (2022) Distributionally robust weighted k-nearest neighbors. *Advances in Neural Information Processing Systems* 35:29088–29100.

Appendix A. Proofs

A.1 Proof of Proposition 1

Proposition 1 *Given a policy $f \in \mathscr{F}$, for any $i \in [n]$, we have*

$$\sup_{\boldsymbol{x}\in\mathscr{X}} \sup_{y\in\mathcal{Y}} \left\{ \mathbb{1}_{\{f(\boldsymbol{x})\neq y\}} - \lambda\|\boldsymbol{x}-\widehat{\boldsymbol{x}}^i\| - \lambda\kappa \mathbb{1}_{\{y\neq\widehat{y}^i\}} \right\} = \left(1-\lambda s_i^n(f)\right)^+.$$

Specifically, for any $i \in J_n(f)$, we have

$$\sup_{\boldsymbol{x}\in\mathscr{X}} \sup_{y\in\mathcal{Y}} \left\{ \mathbb{1}_{\{f(\boldsymbol{x})\neq y\}} - \lambda\|\boldsymbol{x}-\widehat{\boldsymbol{x}}^i\| - \lambda\kappa \mathbb{1}_{\{y\neq\widehat{y}^i\}} \right\} = 1.$$

Proof: For each $i \in [n]$, let us denote

$$\psi_i(\boldsymbol{x}) = \sup_{y\in\mathcal{Y}} \left\{ \mathbb{1}_{\{f(\boldsymbol{x})\neq y\}} - \lambda\|\boldsymbol{x}-\widehat{\boldsymbol{x}}^i\| - \lambda\kappa \mathbb{1}_{\{y\neq\widehat{y}^i\}} \right\}.$$

We first compute the inner supremum over y by discussing whether $f(\boldsymbol{x}) = \widehat{y}^i$ for any given $\boldsymbol{x} \in \boldsymbol{x} \in \mathscr{X}$ as follows:

(i) When $f(\boldsymbol{x}) = \widehat{y}^i$, by discussing whether $y = \widehat{y}^i$ or not, we have

$$\mathbb{1}_{\{f(\boldsymbol{x})\neq y\}} - \lambda\|\boldsymbol{x}-\widehat{\boldsymbol{x}}^i\| - \lambda\kappa \mathbb{1}_{\{y\neq\widehat{y}^i\}} = \begin{cases} -\lambda\|\boldsymbol{x}-\widehat{\boldsymbol{x}}^i\|, & \text{if } y = \widehat{y}^i; \\ 1-\lambda\kappa-\lambda\|\boldsymbol{x}-\widehat{\boldsymbol{x}}^i\|, & \text{if } y \neq \widehat{y}^i. \end{cases}$$

Therefore, in this case, we have

$$\psi_i(\boldsymbol{x}) = \max\{0, 1-\lambda\kappa\} - \lambda\|\boldsymbol{x}-\widehat{\boldsymbol{x}}^i\|.$$

(ii) When $f(\boldsymbol{x}) \neq \widehat{y}^i$, similarly, by discussing whether $y = \widehat{y}^i$ or not, we have

$$\mathbb{1}_{\{f(\boldsymbol{x})\neq y\}} - \lambda\|\boldsymbol{x}-\widehat{\boldsymbol{x}}^i\| - \lambda\kappa \mathbb{1}_{\{y\neq\widehat{y}^i\}} \begin{cases} = 1-\lambda\|\boldsymbol{x}-\widehat{\boldsymbol{x}}^i\|, & \text{if } y = \widehat{y}^i; \\ \leq 1-\lambda\kappa-\lambda\|\boldsymbol{x}-\widehat{\boldsymbol{x}}^i\|, & \text{if } y \neq \widehat{y}^i. \end{cases}$$

Since $1-\lambda\kappa-\lambda\|\boldsymbol{x}-\widehat{\boldsymbol{x}}^i\| \leq 1-\lambda\|\boldsymbol{x}-\widehat{\boldsymbol{x}}^i\|$, we have

$$\psi_i(\boldsymbol{x}) = 1-\lambda\|\boldsymbol{x}-\widehat{\boldsymbol{x}}^i\|.$$

Putting these two cases together, we have

$$\psi_i(\boldsymbol{x}) = \begin{cases} \max\{0, 1-\lambda\kappa\} - \lambda\|\boldsymbol{x}-\widehat{\boldsymbol{x}}^i\|, & \text{if } f(\boldsymbol{x}) = \widehat{y}^i; \\ 1-\lambda\|\boldsymbol{x}-\widehat{\boldsymbol{x}}^i\|, & \text{if } f(\boldsymbol{x}) \neq \widehat{y}^i. \end{cases}$$

Next, we observe that the outer supremum over $\boldsymbol{x}$ is equivalent to

$$\sup_{\boldsymbol{x}\in\mathscr{X}} \psi_i(\boldsymbol{x}) = \max\left\{ \sup_{\boldsymbol{x}\in f^{-1}(\widehat{y}^i)} \psi_i(\boldsymbol{x}), \sup_{\boldsymbol{x}\in f^{-1}(\mathcal{Y}\setminus\{\widehat{y}^i\})} \psi_i(\boldsymbol{x}) \right\}.$$

The two suprema can be simplified as

$$\begin{aligned}\sup_{\boldsymbol{x}\in f^{-1}(\widehat{y}^i)} \psi_i(\boldsymbol{x}) &= \sup_{\boldsymbol{x}\in f^{-1}(\widehat{y}^i)} \left\{\max\{0,1-\lambda\kappa\} - \lambda\|\boldsymbol{x}-\widehat{\boldsymbol{x}}^i\|\right\} \\ &= \max\{0,1-\lambda\kappa\} - \lambda \mathrm{dist}\left(\widehat{\boldsymbol{x}}^i, f^{-1}(\widehat{y}^i)\right) \\ &= \begin{cases} \max\{0,1-\lambda\kappa\}, & \text{if } f(\widehat{\boldsymbol{x}}^i)=\widehat{y}^i; \\ \max\{0,1-\lambda\kappa\} - \lambda \mathrm{dist}\left(\widehat{\boldsymbol{x}}^i, f^{-1}(\widehat{y}^i)\right), & \text{if } f(\widehat{\boldsymbol{x}}^i)\neq\widehat{y}^i; \end{cases}\end{aligned}$$

and

$$\begin{aligned}\sup_{\boldsymbol{x}\in f^{-1}(\mathcal{Y}\setminus\{\widehat{y}^i\})} \psi_i(\boldsymbol{x}) &= \sup_{\boldsymbol{x}\in f^{-1}(\mathcal{Y}\setminus\{\widehat{y}^i\})} \left\{\max\{0,1-\lambda\kappa\} - \lambda\|\boldsymbol{x}-\widehat{\boldsymbol{x}}^i\|\right\} \\ &= 1-\lambda \mathrm{dist}\left(\widehat{\boldsymbol{x}}^i, f^{-1}(\mathcal{Y}\setminus\{\widehat{y}^i\})\right) \\ &= \begin{cases} 1-\lambda \mathrm{dist}\left(\widehat{\boldsymbol{x}}^i, f^{-1}(\mathcal{Y}\setminus\{\widehat{y}^i\})\right), & \text{if } f(\widehat{\boldsymbol{x}}^i)=\widehat{y}^i; \\ 1, & \text{if } f(\widehat{\boldsymbol{x}}^i)\neq\widehat{y}^i. \end{cases}\end{aligned}$$

Since $\max\{0,1-\lambda\kappa\} - \lambda\mathrm{dist}(\widehat{\boldsymbol{x}}^i, f^{-1}(\widehat{y}^i)) \leq 1$, we can combine the two suprema under the cases $f(\widehat{\boldsymbol{x}}^i)=\widehat{y}^i$ and $f(\widehat{\boldsymbol{x}}^i)\neq\widehat{y}^i$ respectively and obtain

$$\begin{aligned}\sup_{\boldsymbol{x}\in\mathscr{X}} \psi_i(\boldsymbol{x}) &= \begin{cases} \max\left\{0,1-\lambda\kappa,1-\lambda \mathrm{dist}\left(\widehat{\boldsymbol{x}}^i, f^{-1}(\mathcal{Y}\setminus\{\widehat{y}^i\})\right)\right\}, & \text{if } f(\widehat{\boldsymbol{x}}^i)=\widehat{y}^i \\ 1, & \text{if } f(\widehat{\boldsymbol{x}}^i)\neq\widehat{y}^i \end{cases} \\ &= \begin{cases} \left(1-\lambda\min\left\{\kappa, \mathrm{dist}\left(\widehat{\boldsymbol{x}}^i, f^{-1}(\mathcal{Y}\setminus\{\widehat{y}^i\})\right)\right\}\right)^+, & \text{if } f(\widehat{\boldsymbol{x}}^i)=\widehat{y}^i \\ 1, & \text{if } f(\widehat{\boldsymbol{x}}^i)\neq\widehat{y}^i \end{cases} \\ &= \left(1-\lambda s_i^n(f)\right)^+,\end{aligned}$$

where the last equality follows from the fact that $s_i^n(f)=0$ when $f(\widehat{\boldsymbol{x}}^i)\neq\widehat{y}^i$. This completes the proof. □

A.2 Proof of Theorem 1

Theorem 1 *Given a policy $f\in\mathscr{F}$, suppose that the values $\{s_i^n(f)\}_{i\in[n]}$ are sorted in ascending order, that is, $0\leq s_{(1)}^n(f)\leq s_{(2)}^n(f)\leq\cdots\leq s_{(n)}^n(f)\leq\kappa$. Let us define $k_n(f,\theta)$ as*

(i) if $\sum_{i\in[n]} s_i^n(f) < n\theta$, then we let $k_n(f,\theta)=n$; or

(ii) if $\sum_{i\in[n]} s_i^n(f)\geq n\theta$, then $k_n(f,\theta)$ is the unique value in the range $(0,n]$ such that:

$$\sum_{i\in[\lceil k_n(f,\theta)\rceil-1]} s_{(i)}^n(f) + (k_n(f,\theta)+1-\lceil k_n(f,\theta)\rceil)s_{(\lceil k_n(f,\theta)\rceil)}^n(f) = n\theta. \tag{7}$$

Then the worst-case objective is equal to $v_n^{\mathscr{X}}(f)=\frac{1}{n}k_n(f,\theta)$.

Proof: By Proposition 1, we have

$$v_n^{\mathscr{X}}(f) = \frac{1}{n}\inf_{\lambda\geq 0}\left\{L(\lambda) := n\lambda\theta + \sum_{i\in[n]}\left(1-\lambda s_i^n(f)\right)^+\right\}.$$

It suffices to show that

$$\inf_{\lambda\geq 0} L(\lambda) = k_n(f,\theta).$$

Note that $L(\lambda)$ is a piecewise linear convex function of λ. Additionally, for $\ell = 0, 1 \cdots, n$, we have

$$\frac{\mathrm{d}L(\lambda)}{\mathrm{d}\lambda} = n\theta - \sum_{i\in[\ell]} s^n_{(i)}(f), \text{ when } \lambda \in \left(\frac{1}{s^n_{(\ell+1)}(f)}, \frac{1}{s^n_{(\ell)}(f)}\right),$$

where for the notational convenience, we let $s^n_{(0)}(f) = 0$ and $s^n_{(n+1)}(f) = +\infty$. Consider the unique $k^* \in \{1, 2, \cdots, n+1\}$ that satisfies

$$\sum_{i\in[k^*-1]} s^n_{(i)}(f) < n\theta \le \sum_{i\in[k^*]} s^n_{(i)}(f).$$

Then $\lambda^* = \frac{1}{s^n_{(k^*)}(f)}$ is a minimizer of $L(\lambda)$. Hence, we have

$$\inf_{\lambda\ge 0} L(\lambda) = L(\lambda^*) = \begin{cases} \dfrac{n\theta}{s^n_{(k^*)}(f)} + \displaystyle\sum_{i\in[k^*]} \left(1 - \frac{s^n_{(i)}(f)}{s^n_{(k^*)}(f)}\right), & \text{if } k^* = 1, 2, \cdots, n; \\ n, & \text{if } k^* = n+1. \end{cases}$$

It remains to discuss the value of k^*.

(i) If $k^* = n+1$, then $\sum_{i\in[n]} s^n_i(f) < n\theta$. In this case, we have $L(\lambda^*) = n = k_n(f,\theta)$.

(ii) If $k^* \le n$, then $\sum_{i\in[n]} s^n_i(f) \ge \sum_{i\in[k^*]} s^n_{(i)}(f) \ge n\theta$. Rearranging terms, we have

$$L(\lambda^*) = \frac{n\theta}{s^n_{(k^*)}(f)} + \sum_{i\in[k^*]} \left(1 - \frac{s^n_{(i)}(f)}{s^n_{(k^*)}(f)}\right) = k^* - \frac{\sum_{i\in[k^*]} s^n_{(i)}(f) - n\theta}{s^n_{(k^*)}(f)}.$$

By the definition of k^*, we know

$$0 \le \frac{\sum_{i\in[k^*]} s^n_{(i)}(f) - n\theta}{s^n_{(k^*)}(f)} < 1.$$

Therefore, we can get a unique $L(\lambda^*)$ by

$$\begin{cases} \lceil L(\lambda^*)\rceil = k^*, \\ \lceil L(\lambda^*)\rceil - L(\lambda^*) = \dfrac{\sum_{i\in[k^*]} s^n_{(i)}(f) - n\theta}{s^n_{(k^*)}(f)}. \end{cases} \tag{15a}$$

By (7), since $\theta > 0$, we must have $s^n_{(\lceil k_n(f,\theta)\rceil)}(f) > 0$. If not, all the $s^n_{(i)}(f)$ appearing on the left-hand side of (7) are equal to 0, leading to a contradiction. And we also have $0 < k_n(f,\theta) + 1 - \lceil k_n(f,\theta)\rceil \le 1$, so

$$\sum_{i\in[\lceil k_n(f,\theta)\rceil - 1]} s^n_{(i)}(f) < n\theta \le \sum_{i\in[\lceil k_n(f,\theta)\rceil]} s^n_{(i)}(f).$$

Therefore, we must have $\lceil k_n(f,\theta)\rceil = k^*$. Rearranging (7), we have

$$\lceil k_n(f,\theta)\rceil - k_n(f,\theta) = \frac{\sum_{i\in[\lceil k_n(f,\theta)\rceil]} s^n_{(i)}(f) - n\theta}{s^n_{(\lceil k_n(f,\theta)\rceil)}(f)} = \frac{\sum_{i\in[k^*]} s^n_{(i)}(f) - n\theta}{s^n_{(k^*)}(f)}.$$

Therefore, (7) defines a unique $k_n(f,\theta)$ that satisfies (15a) and$k_n(f,\theta) = L(\lambda^*)$.

Combining the two cases yields the result. □

A.3 Proof of Corollary 1

Corollary 1 *Given two policies $f, g \in \mathscr{F}$, suppose that for all $i \in [n]$, $s_i^n(f) \le s_i^n(g)$. Then we have $v_n^{\mathscr{X}}(f) \ge v_n^{\mathscr{X}}(g)$.*

Proof: We consider the two cases discussed in Theorem 1 for the policy f. When $\sum_{i\in[n]} s_i^n(f) < n\theta$, we have $v_n^{\mathscr{X}}(f) = 1$, so $v_n^{\mathscr{X}}(f) \ge v_n^{\mathscr{X}}(g)$.

When $\sum_{i\in[n]} s_i^n(f) \ge n\theta$, we also have $\sum_{i\in[n]} s_i^n(g) \ge \sum_{i\in[n]} s_i^n(f) \ge n\theta$. By (7), we have

$$\sum_{i\in[\lceil k_n(f,\theta)\rceil -1]} s_{(i)}^n(f) + (k_n(f,\theta) + 1 - \lceil k_n(f,\theta)\rceil) s_{(\lceil k_n(f,\theta)\rceil)}^n(f) = n\theta; \tag{16a}$$

$$\sum_{i\in[\lceil k_n(g,\theta)\rceil -1]} s_{(i)}^n(g) + (k_n(g,\theta) + 1 - \lceil k_n(g,\theta)\rceil) s_{(\lceil k_n(g,\theta)\rceil)}^n(g) = n\theta. \tag{16b}$$

Since $s_i^n(f) \le s_i^n(g)$ for all $i \in [n]$, the monotonicity holds also for the respective sorted values, i.e., $s_{(i)}^n(f) \le s_{(i)}^n(g)$, replacing $s_{(i)}^n(f)$'s by $s_{(i)}^n(g)$'s in (16a), we obtain

$$\sum_{i\in[\lceil k_n(f,\theta)\rceil -1]} s_{(i)}^n(g) + (k_n(f,\theta) + 1 - \lceil k_n(f,\theta)\rceil) s_{(\lceil k_n(f,\theta)\rceil)}^n(g) \ge n\theta. \tag{16c}$$

Comparing (16b) and (16c), we have $k_n(f,\theta) \ge k_n(g,\theta)$ and hence $v_n^{\mathscr{X}}(f) \ge v_n^{\mathscr{X}}(g)$. □

A.4 Proof of Corollary 2

Corollary 2 *Given a policy $f \in \mathscr{F}$, let $k_n(f,\theta)$ be the same quantity defined in Theorem 1. Then we have*

(i) $k_n(f,\theta) \ge |\{i \in [n] : s_i^n(f) \le \theta\}|$; and

(ii) when $\theta \ge \kappa$, the worst-case objective $v_n^{\mathscr{X}}(f)$ is equal to 1.

Proof: The two results come directly from the definition (7) of $k_n(f,\theta)$.

(i) To prove this property, let us denote

$$T_n(f,\theta) = \{i \in [n] : s_i^n(f) \le \theta\}. \tag{17}$$

When $\sum_{i\in[n]} s_i^n(f) < n\theta$, by Theorem 1, we have

$$k_n(f,\theta) = n \ge |T_n(f,\theta)|.$$

When $\sum_{i\in[n]} s_i^n(f) \ge n\theta$, by (17), we know that

$$s_{(1)}^n(f) \le s_{(2)}^n(f) \le \cdots \ s_{(|T_n(f,\theta)|)}^n(f) \le \theta.$$

Therefore,

$$\sum_{i\in[|T_n(f,\theta)|]} s_{(i)}^n(f) \le |T_n(f,\theta)|\theta \le n\theta.$$

By (7), we know that

$$\sum_{i\in[\lceil k_n(f,\theta)\rceil-1]} s^n_{(i)}(f) + (k_n(f,\theta)+1-\lceil k_n(f,\theta)\rceil)s^n_{(\lceil k_n(f,\theta)\rceil)}(f) = n\theta$$

Comparing the left-hand sides of the two equations, we have

$$k_n(f,\theta) \geq |T_n(f,\theta)|\,.$$

Combining the two cases yields the result.

(ii) When $\theta \geq \kappa$, for any $i \in [n]$, we have $s^n_i(f) \leq \kappa \leq \theta$. So $T_n(f,\theta) = [n]$. Therefore, $k_n(f,\theta) = n$ and hence $v^{\mathscr{X}}_n(f) = 1$.

This completes the proof. □

A.5 Proof of Corollary 3

Corollary 3 *Suppose that $\{\widehat{\boldsymbol{x}}^i\}_{i=1}^n$ are distinct and $\theta \leq \frac{1}{n}\min_{i,j\in[n]:\, i\neq j}\|\widehat{\boldsymbol{x}}^i - \widehat{\boldsymbol{x}}^j\|$. Then the unique optimal solution to the dual of the in-sample DRUC problem*

$$v^{\widehat{\mathcal{X}}_n}_n = \inf_{\widehat{f}\in\widehat{\mathcal{F}}_n}\inf_{\lambda\geq 0}\left\{\lambda\theta + \frac{1}{n}\sum_{i\in[n]}\left[\sup_{\boldsymbol{x}\in\widehat{\mathcal{X}}_n}\sup_{y\in\mathcal{Y}}\left\{\mathbb{1}_{\{\widehat{f}(\boldsymbol{x})\neq y\}} - \lambda\|\boldsymbol{x}-\widehat{\boldsymbol{x}}^i\| - \lambda\kappa\mathbb{1}_{\{y\neq\widehat{y}^i\}}\right\}\right]\right\}$$

is $\widehat{f}^*\in\widehat{\mathcal{F}}_n$ such that $\widehat{f}^*(\widehat{\boldsymbol{x}}^i) = \widehat{y}^i$ for all $i\in[n]$.

Proof: Given any in-sample policy $\widehat{f}\in\widehat{\mathcal{F}}_n$, according to (6a), for set $J_n(\widehat{f})$, we have

$$\begin{cases} |J_n(\widehat{f})| = 0, & \text{if } \widehat{f} = \widehat{f}^*; \\ |J_n(\widehat{f})| \geq 1, & \text{if } \widehat{f}\neq\widehat{f}^*. \end{cases}$$

For all $i\in I_n(\widehat{f})$, we have $\widehat{f}^{-1}(\mathcal{Y}\setminus\{\widehat{y}^i\}) \subseteq \{\widehat{\boldsymbol{x}}^j\}_{j\in[n],j\neq i}$, so

$$s^n_i(f) \geq \operatorname{dist}(\widehat{\boldsymbol{x}}^i, \widehat{f}^{-1}(\mathcal{Y}\setminus\{\widehat{y}^i\})) \geq \min_{i,j\in[n]:\, i\neq j}\|\widehat{\boldsymbol{x}}^i - \widehat{\boldsymbol{x}}^j\|.$$

Since $s^n_i(\widehat{f}) = 0$ for $i\in J_n(\widehat{f})$ and $\theta\leq\frac{1}{n}\min_{i,j\in[n]:i\neq j}\|\widehat{\boldsymbol{x}}^i-\widehat{\boldsymbol{x}}^j\|$, we have

$$0 = \sum_{i\in[|J_n(\widehat{f})|]} s^n_{(i)}(\widehat{f}) < n\theta \leq \min_{i,j\in[n]:i\neq j}\|\widehat{\boldsymbol{x}}^i-\widehat{\boldsymbol{x}}^j\| \leq \sum_{i\in[|J_n(\widehat{f})|+1]} s^n_{(i)}(\widehat{f}).$$

Letting $\mathscr{X} = \widehat{\mathcal{X}}_n$ in Theorem 1, we have $|J_n(\widehat{f})| < k_n(\widehat{f},\theta) \leq |J_n(\widehat{f})|+1$. Therefore, for any $\widehat{f}\neq\widehat{f}^*$, we have

$$v^{\widehat{\mathcal{X}}_n}_n(\widehat{f}^*) = \frac{k_n(\widehat{f}^*,\theta)}{n} \leq \frac{1}{n} < \frac{k_n(\widehat{f},\theta)}{n} = v^{\widehat{\mathcal{X}}_n}_n(\widehat{f})$$

Therefore, $\widehat{f}^*$ is the unique optimal solution. □

A.6 Proof of Proposition 2

Proposition 2 *Given an in-sample policy $\widehat{f}$ and its extension $\widetilde{f}$ as defined in* (8)*, we have*

$$I_n(\widetilde{f}) = I_n(\widehat{f}), J_n(\widetilde{f}) = J_n(\widehat{f}), s_i^n(\widetilde{f}) \leq s_i^n(\widehat{f}), \forall i \in [n], k_n(\widetilde{f}, \theta) \geq k_n(\widehat{f}, \theta).$$

Proof: The equivalence between the two pairs of sets follows directly from the definition (8) of $\widetilde{f}$, as it implies that $\widetilde{f}$ and $\widehat{f}$ coincide on the set $\widehat{\mathcal{X}}_n$.

For each $i \in I_n(\widehat{f})$, since the set $\widehat{\mathcal{X}}_n$ is finite, it follows from the definition (6b) of $s_i^n(\widehat{f})$ that there exists a point $\widehat{\boldsymbol{x}}^j \in \widehat{\mathcal{X}}_n$ satisfying

$$\min\{\|\widehat{\boldsymbol{x}}^j - \widehat{\boldsymbol{x}}^i\|, \kappa\} = s_i^n(\widehat{f}), \widehat{f}(\widehat{\boldsymbol{x}}^j) \neq \widehat{y}^i.$$

We know that $\widetilde{f}(\widehat{\boldsymbol{x}}^j) = \widehat{f}(\widehat{\boldsymbol{x}}^j) \neq \widehat{y}^i$, therefore, by the definition (6b) of $s_i^n(\widetilde{f})$, we have

$$s_i^n(\widetilde{f}) = \min\left\{\operatorname{dist}\left(\widehat{\boldsymbol{x}}^i, \widetilde{f}^{-1}(\mathcal{Y} \setminus \{\widehat{y}^i\})\right), \kappa\right\} \leq \min\{\|\widehat{\boldsymbol{x}}^j - \widehat{\boldsymbol{x}}^i\|, \kappa\} = s_i^n(\widehat{f}).$$

Hence we get $s_i^n(\widetilde{f}) \leq s_i^n(\widehat{f})$. Then, following Corollary 1, we have $k_n(\widetilde{f}, \theta) \geq k_n(\widehat{f}, \theta)$. □

A.7 Proof of Lemma 1

Lemma 1 *Under Assumption 1, given any $\epsilon \in \mathbb{R}_{++}$ and $R \in \mathbb{R}_{++}$, consider the partition $\{A_{\boldsymbol{k}}^\epsilon\}_{\boldsymbol{k} \in \mathbb{Z}^d}$ of the space $\mathbb{R}^d$, where $A_{k_1,k_2,\cdots,k_d}^\epsilon = [k_1\epsilon, (k_1+1)\epsilon) \times [k_2\epsilon, (k_2+1)\epsilon) \times \cdots \times [k_d\epsilon, (k_d+1)\epsilon)$. Let us denote a set $\bar{\Lambda}_\epsilon(R) = \{\boldsymbol{k} \in \mathbb{Z}^d : \mathbb{P}_{(\boldsymbol{X},Y)\sim\mathbb{P}_0}(\boldsymbol{X} \in B_{\boldsymbol{0}}(R) \cap \mathcal{X} \cap \operatorname{cl}(A_{\boldsymbol{k}}^\epsilon)) > 0\}$. Then, almost surely, there exists an integer n_0 such that for all $n \geq n_0$, each set in the collection $\{\mathcal{X} \cap \operatorname{cl}(A_{\boldsymbol{k}}^\epsilon)\}_{\boldsymbol{k} \in \bar{\Lambda}_\epsilon(R)}$ contains at least one in-sample feature point from $\{\widehat{\boldsymbol{x}}^i\}_{i\in[n]}$.*

Proof: We know that the set $\bar{\Lambda}_\epsilon(R)$ is finite since $B_{\boldsymbol{0}}(R)$ is bounded. We have $\mathbb{P}(\widehat{\boldsymbol{x}}^1 \in \operatorname{cl}(A_{\boldsymbol{k}}^\epsilon)) > 0$ for any $\boldsymbol{k} \in \bar{\Lambda}_\epsilon(R)$. According to the strong law of large numbers, it follows that

$$\frac{\sum_{i\in[n]} \mathbb{1}_{\{\widehat{\boldsymbol{x}}^i \in \operatorname{cl}(A_{\boldsymbol{k}}^\epsilon)\}}}{n} \to \mathbb{P}(\widehat{\boldsymbol{x}}^1 \in \operatorname{cl}(A_{\boldsymbol{k}}^\epsilon)) > 0$$

as $n \to \infty$ almost surely. Therefore, $\sum_{i\in[n]} \mathbb{1}_{\{\widehat{\boldsymbol{x}}^i \in \operatorname{cl}(A_{\boldsymbol{k}}^\epsilon)\}} \to \infty$ almost surely as $n \to \infty$. Since $\bar{\Lambda}_\epsilon(R)$ is a finite set and each sequence $\sum_{i\in[n]} \mathbb{1}_{\{\widehat{\boldsymbol{x}}^i \in \operatorname{cl}(A_{\boldsymbol{k}}^\epsilon)\}}$ is non-decreasing in n for any $\boldsymbol{k} \in \bar{\Lambda}_\epsilon(R)$, there exists an $n_0 \in \mathbb{N}$ such that, almost surely, for all $n \geq n_0$, there is at least one in-sample feature point in each set $\{\mathcal{X} \cap \operatorname{cl}(A_{\boldsymbol{k}}^\epsilon)\}_{\boldsymbol{k} \in \bar{\Lambda}_\epsilon(R)}$. □

A.8 Proof of Lemma 2

Lemma 2 *Under Assumption 1, given an in-sample policy $\widehat{f}$ and its closest-label extension $\widetilde{f}$ defined in* (8)*, given any $\epsilon > 0$ and $R > 0$, there almost surely exists an n_0 such that for all $n \geq n_0$ and for every $\widehat{\boldsymbol{x}}^i \in \widehat{\mathcal{X}}_n \cap B_0(R)$, the following inequalities hold: $s_i^n(\widehat{f}) \geq s_i^n(\widetilde{f}) \geq s_i^n(\widehat{f}) - \epsilon$.*

Proof: The first inequality holds due to Proposition 2. The remainder of the proof is to prove the second inequality. Consider the same partition as in Lemma 1, using hypercubes with side length $\frac{\epsilon}{\sqrt{d}}$. According to Lemma 1, let n_0 be sufficiently large such that, for all $n \geq n_0$, every hypercube indexed by $\bar{\Lambda}_{\epsilon/\sqrt{d}}(R+\kappa-\epsilon)$ contains at least one in-sample feature point.

For any $\widehat{\boldsymbol{x}}^i \in B_0(R) \cap \widehat{\mathcal{X}}_n$, let $s = \operatorname{dist}(\widehat{\boldsymbol{x}}^i, \widehat{f}^{-1}(\mathcal{Y} \setminus \widehat{y}^i))$. If $\min\{s,\kappa\} \leq \epsilon$, then the result is trivial. Suppose that $\min\{s,\kappa\} > \epsilon$. In this case, we know that $s > 0$, so $\widehat{f}(\widehat{\boldsymbol{x}}^i) = \widehat{y}^i$. For any $\boldsymbol{x} \in \mathcal{X}$ such that $\|\boldsymbol{x} - \widehat{\boldsymbol{x}}^i\| < \min\{s,\kappa\} - \epsilon$, we must have $\boldsymbol{x} \in B_{\boldsymbol{0}}(R+\kappa-\epsilon)$.

Next, we show that

Claim 1 *The feature point $\boldsymbol{x} \in \operatorname{cl}(A_{\boldsymbol{k}}^{\epsilon/\sqrt{d}})$ for some $\boldsymbol{k} \in \bar{\Lambda}_{\epsilon/\sqrt{d}}(R+\kappa-\epsilon)$.*

Proof: We prove by contradiction and assume that the feature point $\boldsymbol{x} \notin \operatorname{cl}(A_{\boldsymbol{k}}^{\epsilon/\sqrt{d}})$ for all $\boldsymbol{k} \in \bar{\Lambda}_{\epsilon/\sqrt{d}}(R+\kappa-\epsilon)$. Then since the set $\bar{\Lambda}_{\epsilon/\sqrt{d}}(R+\kappa-\epsilon)$ is finite, the closure $\operatorname{cl}(A_{\boldsymbol{k}}^{\epsilon/\sqrt{d}})$ is closed for any $\boldsymbol{k} \in \bar{\Lambda}_{\epsilon/\sqrt{d}}(R+\kappa-\epsilon)$, and $\boldsymbol{x} \in B_{\boldsymbol{0}}(R+\kappa-\epsilon)$, there exists $r > 0$ such that $B_{\boldsymbol{x}}(r) \subseteq B_{\boldsymbol{0}}(R+\kappa-\epsilon)$ and $B_{\boldsymbol{x}}(r) \bigcap \operatorname{cl}(A_{\boldsymbol{k}}^{\epsilon/\sqrt{d}}) = \varnothing$ for all $\boldsymbol{k} \in \bar{\Lambda}_{\epsilon/\sqrt{d}}(R+\kappa-\epsilon)$. Note that we can decompose $B_{\boldsymbol{x}}(r) \cap \mathcal{X}$ to

$$B_{\boldsymbol{x}}(r) \cap \mathcal{X} = \bigcup_{\boldsymbol{k} \in \mathbb{Z}^d} \left(B_{\boldsymbol{x}}(r) \cap \mathcal{X} \cap \operatorname{cl}\left(A_{\boldsymbol{k}}^{\epsilon/\sqrt{d}} \right) \right).$$

For any $\boldsymbol{k} \in \bar{\Lambda}_{\epsilon/\sqrt{d}}(R+\kappa-\epsilon)$, since $B_{\boldsymbol{x}}(r) \cap \operatorname{cl}(A_{\boldsymbol{k}}^{\epsilon/\sqrt{d}}) = \varnothing$, we must have $\mathbb{P}_{(\boldsymbol{X},Y)\sim\mathbb{P}_0}(\boldsymbol{X} \in B_{\boldsymbol{x}}(r) \cap \mathcal{X} \cap \operatorname{cl}(A_{\boldsymbol{k}}^{\epsilon/\sqrt{d}})) = 0$. For any $\boldsymbol{k} \notin \bar{\Lambda}_{\epsilon/\sqrt{d}}(R+\kappa-\epsilon)$, by the definition of $\bar{\Lambda}_{\epsilon/\sqrt{d}}(R+\kappa-\epsilon)$, we have $\mathbb{P}_{(\boldsymbol{X},Y)\sim\mathbb{P}_0}(\boldsymbol{X} \in B_{\boldsymbol{0}}(R+\kappa-\epsilon) \cap \mathcal{X} \cap \operatorname{cl}(A_{\boldsymbol{k}}^{\epsilon/\sqrt{d}})) = 0$. Since $B_{\boldsymbol{x}}(r) \subseteq B_{\boldsymbol{0}}(R+\kappa-\epsilon)$, we have $\mathbb{P}_{(\boldsymbol{X},Y)\sim\mathbb{P}_0}(\boldsymbol{X} \in B_{\boldsymbol{x}}(r) \cap \mathcal{X} \cap \operatorname{cl}(A_{\boldsymbol{k}}^{\epsilon/\sqrt{d}})) = 0$. Therefore, we have $\mathbb{P}_{(\boldsymbol{X},Y)\sim\mathbb{P}_0}(\boldsymbol{X} \in B_{\boldsymbol{x}}(r) \cap \mathcal{X}) = 0$, which contradicts that $\boldsymbol{x} \in \mathcal{X}$ and the feature space $\mathcal{X}$ is closed. Therefore, the feature point $\boldsymbol{x}$ must be included in $\operatorname{cl}(A_{\boldsymbol{k}}^{\epsilon/\sqrt{d}})$ for some $\boldsymbol{k} \in \bar{\Lambda}_{\epsilon/\sqrt{d}}(R+\kappa-\epsilon)$. ◇

Note that every feature point $\boldsymbol{x}' \in \operatorname{cl}(A_{\boldsymbol{k}}^{\epsilon/\sqrt{d}})$ satisfies $\|\boldsymbol{x} - \boldsymbol{x}'\| \leq \epsilon$. According to the triangle inequality, we then have $\operatorname{cl}(A_{\boldsymbol{k}}^{\epsilon/\sqrt{d}}) \subseteq B_{\widehat{\boldsymbol{x}}^i}(\min\{s,\kappa\}) \subseteq B_{\widehat{\boldsymbol{x}}^i}(s)$. Since $\operatorname{dist}(\widehat{\boldsymbol{x}}^i, \widehat{f}^{-1}(\mathcal{Y} \setminus \{\widehat{y}^i\})) = s$, any in-sample feature point $\widehat{\boldsymbol{x}}^j$ in the hypercube $\operatorname{cl}(A_{\boldsymbol{k}}^{\epsilon/\sqrt{d}})$ must satisfy that $\widehat{f}(\widehat{\boldsymbol{x}}^j) = \widehat{y}^i = \widehat{f}(\widehat{\boldsymbol{x}}^i)$.

For any in-sample feature point $\widehat{\boldsymbol{x}}^k$ such that $\widehat{f}(\widehat{\boldsymbol{x}}^k) \neq \widehat{y}^i$, we have $\|\widehat{\boldsymbol{x}}^k - \widehat{\boldsymbol{x}}^i\| \geq s$, hence

$$\|\widehat{\boldsymbol{x}}^k - \boldsymbol{x}\| \geq \|\widehat{\boldsymbol{x}}^k - \widehat{\boldsymbol{x}}^i\| - \|\boldsymbol{x} - \widehat{\boldsymbol{x}}^i\| \geq s - (\min\{s,\kappa\} - \epsilon) \geq \epsilon.$$

Therefore, by the definition of the extension $\widetilde{f}$, we have $\widetilde{f}(\boldsymbol{x}) = \widehat{f}(\widehat{\boldsymbol{x}}^j) = \widehat{f}(\widehat{\boldsymbol{x}}^i)$. This proof is illustrated in Figure 1. This identity holds for any $\boldsymbol{x} \in \mathcal{X}$ such that $\|\boldsymbol{x} - \widehat{\boldsymbol{x}}^i\| \leq \min\{s,\kappa\} - \epsilon$. Thus, we must have

$$\operatorname{dist}\left(\widehat{\boldsymbol{x}}^i, \widetilde{f}^{-1}(\mathcal{Y} \setminus \{\widehat{y}^i\})\right) \geq \min\left\{\operatorname{dist}\left(\widehat{\boldsymbol{x}}^i, \widehat{f}^{-1}(\mathcal{Y} \setminus \{\widehat{y}^i\})\right), \kappa\right\} - \epsilon = \min\{s,\kappa\} - \epsilon.$$

Therefore, we have

$$s_i^n(\widetilde{f}) = \min\left\{\kappa, \operatorname{dist}\left(\widehat{\boldsymbol{x}}^i, \widetilde{f}^{-1}(\mathcal{Y} \setminus \{\widehat{y}^i\})\right)\right\} \geq \min\{\kappa, \min\{s,\kappa\} - \epsilon\} = \min\{s,\kappa\} - \epsilon = s_i^n(\widehat{f}) - \epsilon.$$

The result also holds for any sample size $n \geq n_0$ due to monotonicity. This completes the proof. □

A.9 Proof of Theorem 2

Theorem 2 *Under Assumption 1, given any $\theta \in \mathbb{R}_{++}, \kappa \in \mathbb{R}_{++}$, for $n \in \mathbb{Z}^+$, we have $0 \le v_n^{\mathcal{X}}(\widetilde{f}) - v_n^{\widehat{\mathcal{X}}_n}(\widehat{f}) < \eta_n$ for any $\widehat{f} \in \widehat{\mathcal{F}}_n$ and its extension $\widetilde{f}$, where η_n depends only on $\widehat{\mathcal{X}}_n$ and not on $\widehat{f}$, and $\lim_{n\to\infty} \eta_n = 0$ almost surely.*

Proof: According to Proposition 2, we have $k_n(\widetilde{f},\theta) \ge k_n(\widehat{f},\theta)$ and hence $v_n^{\mathcal{X}}(\widetilde{f}) - v_n^{\widehat{\mathcal{X}}_n}(\widehat{f}) \ge 0$.

We split the remainder of the proof into three steps.

Step I. The second inequality holds trivially when $v_n^{\widehat{\mathcal{X}}_n}(\widehat{f}) = 1$, since in this case $v_n^{\mathcal{X}}(\widetilde{f}) = v_n^{\widehat{\mathcal{X}}_n}(\widehat{f}) = 1$. Let us consider any $\widehat{f}$ such that $v_n^{\widehat{\mathcal{X}}_n}(\widehat{f}) < 1$ and the corresponding extension $\widetilde{f}$. Let us define the values $\{s_i^n(\widetilde{f})\}_{i\in[n]}$ and $\{s_i^n(\widehat{f})\}_{i\in[n]}$ as in (6b), and suppose that $\{s_i^n(\widehat{f})\}_{i\in[n]}$ are sorted in ascending order as $s_{(1)}^n(\widehat{f}) \le s_{(2)}^n(\widehat{f}) \le \cdots \le s_{(n)}^n(\widehat{f})$.

Denote $\beta(R) = \mathbb{P}(\widehat{\boldsymbol{x}}^1 \notin B_{\mathbf{0}}(R))$, $\mathcal{X}_R = \mathcal{X} \cap B_{\mathbf{0}}(R)$ and $r_n(R) = \sum_{i\in[n]} \mathbb{1}_{\{\widehat{\boldsymbol{x}}^i \notin B_{\mathbf{0}}(R)\}}$. According to Lemma 2, without loss of generality, we can assume that

$$s_i^n(\widehat{f}) - \epsilon \le s_i^n(\widetilde{f}) \le s_i^n(\widehat{f}),\ \forall i \in [n] : \widehat{\boldsymbol{x}}^i \in B_{\mathbf{0}}(R). \tag{18a}$$

We consider the adjusted value $\{\widetilde{s}_i^n(\widehat{f},R)\}_{i\in[n]}$ defined as

$$\widetilde{s}_i^n(\widehat{f},R) = \begin{cases} s_i^n(\widetilde{f}), & \text{if } \widehat{\boldsymbol{x}}^i \in B_{\mathbf{0}}(R); \\ s_i^n(\widehat{f}), & \text{if } \widehat{\boldsymbol{x}}^i \notin B_{\mathbf{0}}(R). \end{cases} \tag{18b}$$

Then we know that $s_i^n(\widehat{f}) - \epsilon \le \widetilde{s}_i^n(\widehat{f},R) \le s_i^n(\widehat{f})$ for all $i \in [n]$. Let the sequence $\{\widetilde{s}_i^n(\widehat{f},R)\}_{i\in[n]}$ be sorted as $\widetilde{s}_{(1)}^n(\widehat{f},R) \le \widetilde{s}_{(2)}^n(\widehat{f},R) \le \cdots \le \widetilde{s}_{(n)}^n(\widehat{f},R)$. Then we have $s_{(i)}^n(\widehat{f}) - \epsilon \le \widetilde{s}_{(i)}^n(\widehat{f},R) \le s_{(i)}^n(\widehat{f})$.

Let $k_n(\widetilde{f},\theta)$ and $k_n(\widehat{f},\theta)$ be as defined in Theorem 1 and its in-sample counterpart, respectively. Then, by the in-sample counterpart of Theorem 1, we know that

$$k_n(\widehat{f},\theta) = n v_n^{\widehat{\mathcal{X}}_n}(\widehat{f}) < n, \tag{18c}$$

the strict inequality is due to $v_n^{\widehat{\mathcal{X}}_n}(\widehat{f}) < 1$. Thus, this case satisfies Part (ii) of Theorem 1, we have $\sum_{i\in[n]} s_i^n(\widehat{f}) \ge n\theta$. According to Part (ii) of Theorem 1, it follows that

$$\sum_{i\in\left[\lceil k_n(\widehat{f},\theta)\rceil - 1\right]} s_{(i)}^n(\widehat{f}) + \left(k_n(\widehat{f},\theta) + 1 - \lceil k_n(\widehat{f},\theta)\rceil\right) s_{\left(\lceil k_n(\widehat{f},\theta)\rceil\right)}^n(\widehat{f}) = n\theta. \tag{18d}$$

Since $\widetilde{s}_{(i)}^n(\widetilde{f},R) \ge s_{(i)}^n(\widehat{f}) - \epsilon$, we substitute $s_{(i)}^n(\widehat{f})$ by $\widetilde{s}_{(i)}^n(\widetilde{f},R)$ in equation (18d) and have

$$\sum_{i\in\left[\lceil k_n(\widehat{f},\theta)\rceil - 1\right]} \widetilde{s}_{(i)}^n(\widetilde{f},R) + \left(k_n(\widehat{f},\theta) + 1 - \lceil k_n(\widehat{f},\theta)\rceil\right) \widetilde{s}_{\left(\lceil k_n(\widehat{f},\theta)\rceil\right)}^n(\widetilde{f},R) \ge n\theta - k_n(\widehat{f},\theta)\epsilon. \tag{18e}$$

According to Theorem 1, when Part (i) holds, $k_n(\widetilde{f},\theta)=n$ and $\sum_{i\in[n]} s_i^n(\widetilde{f})<n\theta$; when Part (ii) holds, we have the equation (7). Combining the two cases, we have

$$\sum_{i\in\left[\lceil k_n(\widetilde{f},\theta)\rceil-1\right]} s_{(i)}^n(\widetilde{f})+\left(k_n(\widetilde{f},\theta)+1-\lceil k_n(\widetilde{f},\theta)\rceil\right) s^n_{\left(\lceil k_n(\widetilde{f},\theta)\rceil\right)}(\widetilde{f})\le n\theta. \tag{18f}$$

Using the definition in (18b) and the fact that $\widetilde{s}_{(i)}^n(\widetilde{f},R)\le\kappa$, substituting $s_{(i)}^n(\widetilde{f})$ by $\widetilde{s}_{(i)}^n(\widetilde{f},R)$ in the inequality (18f) yields

$$\sum_{i\in\left[\lceil k_n(\widetilde{f},\theta)\rceil-1\right]} \widetilde{s}_{(i)}^n(\widetilde{f},R)+\left(k_n(\widetilde{f},\theta)+1-\lceil k_n(\widetilde{f},\theta)\rceil\right) \widetilde{s}^n_{\left(\lceil k_n(\widetilde{f},\theta)\rceil\right)}(\widetilde{f},R)\le n\theta+\kappa r_n(R). \tag{18g}$$

where the inequality holds because the differences between $\{\widetilde{s}_i^n(\widehat{f},R)\}_{i\in[n]}$ and $\{s_i^n(\widehat{f})\}_{i\in[n]}$ arise only from indices corresponding to in-sample feature points outside the ball $B_{\mathbf{0}}(R)$. Each such difference is bounded by κ, and the total number of these points is at most $r_n(R)=\sum_{i\in[n]}\mathbb{1}_{\{\hat{\boldsymbol{x}}^i\notin B_{\mathbf{0}}(R)\}}$.

Step II. Comparing the left-hand sides of inequalities (18e) and (18g), we observe that both are partial sums of the sequence $\widetilde{s}_{(i)}^n(\widetilde{f},R)$ in ascending order, up to the point where the sum of the coefficients reaches $k_n(\widehat{f},\theta)$ and $k_n(\widetilde{f},\theta)$, respectively. Subtracting the left-hand side of (18e) from that of (18g), the resulting difference involves terms whose coefficients sum to $(k_n(\widetilde{f},\theta)-k_n(\widehat{f},\theta))$, and each of the corresponding $\widetilde{s}_{(i)}^n(\widetilde{f},R)$ values is at least $\widetilde{s}^n_{(\lceil k_n(\widehat{f},\theta)\rceil)}(\widetilde{f},R)$. Therefore, the difference is bounded below by

$$\left(k_n(\widetilde{f},\theta)-k_n(\widehat{f},\theta)\right)\widetilde{s}^n_{\left(\lceil k_n(\widehat{f},\theta)\rceil\right)}(\widetilde{f},R).$$

Thus, subtracting (18e) from (18g), we arrive at

$$(k_n(\widetilde{f},\theta)-k_n(\widehat{f},\theta))\widetilde{s}^n_{\left(\lceil k_n(\widehat{f},\theta)\rceil\right)}(\widetilde{f},R)\le k_n(\widehat{f},\theta)\epsilon+\kappa r_n(R). \tag{18h}$$

In (18d), using the fact that $s_{(1)}^n(\widehat{f})\le s_{(2)}^n(\widehat{f})\le\cdots\le s^n_{\left(\lceil k_n(\widehat{f},\theta)\rceil\right)}(\widehat{f})$, we have

$$n\theta=\sum_{i\in\left[\lceil k_n(\widehat{f},\theta)\rceil-1\right]} s_{(i)}^n(\widehat{f})+\left(k_n(\widehat{f},\theta)+1-\lceil k_n(\widehat{f},\theta)\rceil\right) s^n_{\left(\lceil k_n(\widehat{f},\theta)\rceil\right)}(\widehat{f})\le k_n(\widehat{f},\theta)s^n_{\left(\lceil k_n(\widehat{f},\theta)\rceil\right)}(\widehat{f}).$$

According to the inequality (18c), we have $k_n(\widehat{f},\theta)<n$. Thus, we can bound $\widetilde{s}^n_{(\lceil k_n(\widehat{f},\theta)\rceil)}(\widetilde{f},R)$ by

$$\widetilde{s}^n_{\left(\lceil k_n(\widehat{f},\theta)\rceil\right)}(\widetilde{f},R)\ge s^n_{\left(\lceil k_n(\widehat{f},\theta)\rceil\right)}(\widehat{f})-\epsilon\ge\frac{n\theta}{k_n(\widehat{f},\theta)}-\epsilon>\theta-\epsilon.$$

Combining this bound with (18h), we have

$$k_n(\widetilde{f},\theta)-k_n(\widehat{f},\theta)\le\frac{k_n(\widehat{f},\theta)\epsilon+\kappa r_n(R)}{\widetilde{s}^n_{\left(\lceil k_n(\widehat{f},\theta)\rceil\right)}(\widetilde{f},R)}<\frac{n\epsilon+\kappa r_n(R)}{\theta-\epsilon}.$$

Therefore, we have

$$v_n^{\mathcal{X}}(\widetilde{f}) - v_n^{\widehat{\mathcal{X}}_n}(\widehat{f}) = \frac{k_n(\widetilde{f},\theta) - k_n(\widehat{f},\theta)}{n} < \frac{\epsilon}{\theta-\epsilon} + \left(\frac{\kappa}{\theta-\epsilon}\right)\frac{r_n(R)}{n}. \tag{18i}$$

Step III. We know that $\lim_{R\to+\infty}\beta(R) = \lim_{R\to+\infty}\mathbb{P}(\widehat{\boldsymbol{x}}^1 \notin B_{\mathbf{0}}(R)) = 0$ by the monotone convergence theorem (Durrett 2010). For any $\eta > 0$, we choose $\epsilon = \frac{\eta\theta}{3+\eta}$ and $R > 0$ such that $\beta(R) \le \frac{\eta\theta}{\kappa(3+\eta)}$. Then according to Lemma 2 and the strong law of large numbers, there exists $n_0 \in \mathbb{N}$ almost surely such that when $n > n_0$, we have both $\frac{r_n(R)}{n} < 2\beta(R)$ and (18a) hold. In this case, the bound in (18i) satisfies

$$\frac{\epsilon}{\theta-\epsilon} + \left(\frac{\kappa}{\theta-\epsilon}\right)\frac{r_n(R)}{n} < \frac{\frac{\eta\theta}{3+\eta}}{\theta - \frac{\eta\theta}{3+\eta}} + \left(\frac{\kappa}{\theta - \frac{\eta\theta}{3+\eta}}\right)\cdot 2\beta(R) \le \frac{\eta}{3} + \frac{\kappa(3+\eta)}{3\theta}\cdot\frac{2\eta\theta}{\kappa(3+\eta)} = \eta.$$

Since this bound does not depend on the specific in-sample policy $\widehat{f}$, and η is an arbitrary positive scalar, we arrive at the conclusion. □

A.10 Proof of Lemma 3

Lemma 3 *Under Assumption 1, let* $\{p_y\}_{y\in\mathcal{Y}}$, $\{\mathbb{P}_y\}_{y\in\mathcal{Y}}$ *be as defined in* (9a). *Then, the optimal UC value is given by*

$$v_{UC} = 1 - \int_{\mathcal{X}} \max_{y\in\mathcal{Y}} \left\{ p_y \frac{\mathrm{d}\mathbb{P}_y}{\mathrm{d}\mathbb{P}_0^{\mathcal{X}}} \right\} \mathrm{d}\mathbb{P}_0^{\mathcal{X}},$$

with an optimal UC policy $f_0(\boldsymbol{x}) = \min\{\arg\max_{y\in\mathcal{Y}}(p_y \frac{\mathrm{d}\mathbb{P}_y}{\mathrm{d}\mathbb{P}_0^{\mathcal{X}}}(\boldsymbol{x}))\}$.

Proof: We set random vectors $(\boldsymbol{X}, Y) \sim \mathbb{P}_0$ in this proof. For any given policy $f \in \mathcal{F}$, by the law of total expectation, we have

$$\mathbb{P}(f(\boldsymbol{X}) \neq Y) = \mathbb{E}_{\boldsymbol{X}}[\mathbb{P}(Y \neq f(\boldsymbol{X})|\boldsymbol{X})].$$

Then we have

$$\mathbb{P}(Y \neq f(\boldsymbol{x})|\boldsymbol{X} = \boldsymbol{x}) = \sum_{y\in\mathcal{Y}, y\neq f(\boldsymbol{x})} p_y \frac{\mathrm{d}\mathbb{P}_y}{\mathrm{d}\mathbb{P}_0^{\mathcal{X}}}(\boldsymbol{x}) = 1 - p_{f(\boldsymbol{x})}\frac{\mathrm{d}\mathbb{P}_{f(\boldsymbol{x})}}{\mathrm{d}\mathbb{P}_0^{\mathcal{X}}}(\boldsymbol{x}) \ge 1 - \max_{y\in\mathcal{Y}}\left\{ p_y \frac{\mathrm{d}\mathbb{P}_y}{\mathrm{d}\mathbb{P}_0^{\mathcal{X}}}(\boldsymbol{x}) \right\}.$$

Here the equality holds if and only if $f(\boldsymbol{x}) \in \arg\max_{y\in\mathcal{Y}}(p_y \frac{\mathrm{d}\mathbb{P}_y}{\mathrm{d}\mathbb{P}_0^{\mathcal{X}}}(\boldsymbol{x}))$. Therefore, we have

$$\mathbb{P}(f(\boldsymbol{X}) \neq Y) = \mathbb{E}_{\boldsymbol{X}}[\mathbb{P}(Y \neq f(\boldsymbol{X})|\boldsymbol{X})] \ge 1 - \mathbb{E}_{\boldsymbol{X}}\left[\max_{y\in\mathcal{Y}}\left\{ p_y \frac{\mathrm{d}\mathbb{P}_y}{\mathrm{d}\mathbb{P}_0^{\mathcal{X}}}(\boldsymbol{X}) \right\}\right] = 1 - \int_{\mathcal{X}} \max_{y\in\mathcal{Y}}\left\{ p_y \frac{\mathrm{d}\mathbb{P}_y}{\mathrm{d}\mathbb{P}_0^{\mathcal{X}}} \right\} \mathrm{d}\mathbb{P}_0^{\mathcal{X}}.$$

Taking the infimum over $f \in \mathcal{F}$, we have

$$v_{UC} \ge 1 - \int_{\mathcal{X}} \max_{y\in\mathcal{Y}} \left\{ p_y \frac{\mathrm{d}\mathbb{P}_y}{\mathrm{d}\mathbb{P}_0^{\mathcal{X}}} \right\} \mathrm{d}\mathbb{P}_0^{\mathcal{X}}.$$

According to (9c), each set U_y consists of feature points for which the label y attains the highest conditional probability. The definition implicitly includes a tie-breaking rule that assigns the

label with the smallest index in the case of multiple labels achieving the maximum. Since the Radon–Nikodym derivatives $\frac{\mathrm{d}\mathbb{P}_y}{\mathrm{d}\mathbb{P}_0^{\mathcal{X}}}$ are measurable for all $y \in \mathcal{Y}$, each set U_y is measurable.

We then define the classification policy $f_0 : \mathcal{X} \to \mathcal{Y}$ such that $f_0(\boldsymbol{x}) = y$ if $\boldsymbol{x} \in U_y$, for all $y \in \mathcal{Y}$. By construction, f_0 is measurable and satisfies $f_0(\boldsymbol{x}) \in \arg\max_{y\in\mathcal{Y}}(p_y \frac{\mathrm{d}\mathbb{P}_y}{\mathrm{d}\mathbb{P}_0^{\mathcal{X}}}(\boldsymbol{x}))$ for each $\boldsymbol{x} \in \mathcal{X}$. Therefore, we have

$$v_{UC} = \mathbb{P}(f_0(\boldsymbol{X}) \neq Y) = 1 - \int_{\mathcal{X}} \max_{y\in\mathcal{Y}} \left\{ p_y \frac{\mathrm{d}\mathbb{P}_y}{\mathrm{d}\mathbb{P}_0^{\mathcal{X}}} \right\} \mathrm{d}\mathbb{P}_0^{\mathcal{X}}.$$

This completes the proof of the optimality of f_0 for the UC problem. □

A.11 Proof of Theorem 3

Theorem 3 *Under Assumption 1, given a $\kappa > 0$ and $\theta > 0$ such that $0 < \theta < \frac{\kappa}{m}$, we have*

$$\liminf_{n\to\infty} v_n^{\mathcal{X}} \geq \liminf_{n\to\infty} v_n^{\widehat{\mathcal{X}}_n} \geq v_{UC}$$

almost surely.

Proof: The first inequality is due to the definition of the in-sample DRUC problem. We mainly prove the second inequality.

Consider the partition introduced in the proof of Lemma 1, where each hypercube has side length $\epsilon = \frac{\theta}{\sqrt{d}}$. We define the index set $\Lambda_\epsilon(R) = \{\boldsymbol{k} \in \mathbb{Z}^d : B_{\mathbf{0}}(R) \cap \mathcal{X} \cap A_{\boldsymbol{k}}^\epsilon \neq \varnothing\}$. For a fixed sample size n and any in-sample policy $\widehat{f}$, the set $T_n(\widehat{f}, \theta)$ is defined as

$$T_n(\widehat{f}, \theta) = \left\{ i \in [n] : s_i^n(\widehat{f}) < \theta \right\},$$

as given in (17).

Recall that $J_n(\widehat{f})$ denotes the set of indices corresponding to misclassified in-sample feature points (see in (6a)). Also, for $i \in J_n(\widehat{f})$, we have $s_i^n(\widehat{f}) = 0$, so $J_n(\widehat{f}) \subseteq T_n(\widehat{f}, \theta)$. By Theorem 1 and Corollary 2, we have

$$v_n^{\widehat{\mathcal{X}}_n}(\widehat{f}) = \frac{k_n(\widehat{f}, \theta)}{n} \geq \frac{|T_n(\widehat{f}, \theta)|}{n}. \tag{19a}$$

For any $R > 0$, consider an arbitrary hypercube $A_{\boldsymbol{k}}^\epsilon$ with index $\boldsymbol{k} \in \Lambda_\epsilon(R)$. We discuss the value of $\widehat{f}$ on in-sample feature points in the hypercube $A_{\boldsymbol{k}}^\epsilon$.

(i) If there exists some $y_{\boldsymbol{k}} \in \mathcal{Y}$, such that $\widehat{f}(\boldsymbol{x}) = y_{\boldsymbol{k}}$ for all $\boldsymbol{x} \in \widehat{\mathcal{X}}_n \cap A_{\boldsymbol{k}}^\epsilon$, then we have

$$\left\{ i \in [n] : \widehat{\boldsymbol{x}}^i \in A_{\boldsymbol{k}}^\epsilon, \widehat{y}^i \neq y_{\boldsymbol{k}} \right\} \subseteq J_n(\widehat{f}) \subseteq T_n(\widehat{f}, \theta).$$

(ii) If $\widehat{f}$ takes more than one value on $\widehat{\mathcal{X}}_n \cap A_{\boldsymbol{k}}^\epsilon$. Then $\widehat{f}^{-1}(\mathcal{Y} \setminus \{y\}) \cap A_{\boldsymbol{k}}^\epsilon \neq \varnothing$ for any $y \in \mathcal{Y}$. In this case, for any $i \in [n]$ and $\widehat{\boldsymbol{x}}^i \in A_{\boldsymbol{k}}^\epsilon$, we have $s_i^n(\widehat{f}) = \min\{\kappa, \mathrm{dist}(\widehat{\boldsymbol{x}}^i, \widehat{f}^{-1}(\mathcal{Y} \setminus \{\widehat{y}^i\}))\} < \sqrt{d}\epsilon = \theta$, since the diameter of the hypercube is $\sqrt{d}\epsilon$. Then, we have $i \in T_n(\widehat{f}, \theta)$ by the definition of $T_n(\widehat{f}, \theta)$. Therefore, we have

$$\left\{ i \in [n] : \widehat{\boldsymbol{x}}^i \in A_{\boldsymbol{k}}^\epsilon \right\} \subseteq T_n(\widehat{f}, \theta).$$

Since one of the these cases must hold for each $\boldsymbol{k} \in \Lambda_\epsilon(R)$, we obtain the following inequality:

$$\left|\left\{i \in T_n(\widehat{f},\theta) : \widehat{\boldsymbol{x}}^i \in A^\epsilon_{\boldsymbol{k}}\right\}\right| \geq \min_{y\in\mathcal{Y}} \left\{\left|\left\{i \in [n] : \widehat{\boldsymbol{x}}^i \in A^\epsilon_{\boldsymbol{k}}, \widehat{y}^i \neq y\right\}\right|\right\}.$$

Dividing both sides by n and summing over all $\boldsymbol{k} \in \Lambda_\epsilon(R)$, we have

$$\sum_{\boldsymbol{k}\in\Lambda_\epsilon(R)} \frac{1}{n}\left|\left\{i \in T_n(\widehat{f},\theta) : \widehat{\boldsymbol{x}}^i \in A^\epsilon_{\boldsymbol{k}}\right\}\right| \geq \sum_{\boldsymbol{k}\in\Lambda_\epsilon(R)} \frac{1}{n}\min_{y\in\mathcal{Y}} \left\{\left|\left\{i \in [n] : \widehat{\boldsymbol{x}}^i \in A^\epsilon_{\boldsymbol{k}}, \widehat{y}^i \neq y\right\}\right|\right\}.$$

Note that the right-hand side does not depend on $\widehat{f}$ and the left-hand side is upper bounded by $v_n^{\widehat{\mathcal{X}}_n}(\widehat{f})$ according to (19a). Therefore, we have

$$v_n^{\widehat{\mathcal{X}}_n} \geq \sum_{\boldsymbol{k}\in\Lambda_\epsilon(R)} \frac{1}{n}\min_{y\in\mathcal{Y}} \left\{\left|\left\{i \in [n] : \widehat{\boldsymbol{x}}^i \in A^\epsilon_{\boldsymbol{k}}, \widehat{y}^i \neq y\right\}\right|\right\}. \tag{19b}$$

By the strong law of large numbers, for each $y \in \mathcal{Y}$, almost surely

$$\lim_{n\to\infty} \frac{1}{n}\left\{\left|\left\{i \in [n] : \widehat{\boldsymbol{x}}^i \in A^\epsilon_{\boldsymbol{k}}, \widehat{y}^i \neq y\right\}\right|\right\} = \lim_{n\to\infty} \frac{1}{n}\sum_{i\in[n]} \mathbb{1}_{\left\{\widehat{\boldsymbol{x}}^i \in A^\epsilon_{\boldsymbol{k}}, \widehat{y}^i \neq y\right\}} = \mathbb{P}\left(\widehat{\boldsymbol{x}}^1 \in A^\epsilon_{\boldsymbol{k}}, \widehat{y}^1 \neq y\right).$$

Since the set $\mathcal{Y}$ is finite, we have

$$\lim_{n\to\infty} \frac{1}{n}\min_{y\in\mathcal{Y}} \left\{\left|\left\{i \in [n] : \widehat{\boldsymbol{x}}^i \in A^\epsilon_{\boldsymbol{k}}, \widehat{y}^i \neq y\right\}\right|\right\} = \min_{y\in\mathcal{Y}} \left\{\mathbb{P}\left(\widehat{\boldsymbol{x}}^1 \in A^\epsilon_{\boldsymbol{k}}, \widehat{y}^1 \neq y\right)\right\}. \tag{19c}$$

Therefore, we can obtain an asymptotic lower bound of $v_n^{\widehat{\mathcal{X}}_n}$ by taking liminf on both sides of inequality (19b). By (19c), the right-hand side of (19b) converges almost surely. Therefore, we can substitute liminf with limit and get:

$$\begin{aligned}
&\liminf_{n\to\infty} v_n^{\widehat{\mathcal{X}}_n} \geq \lim_{n\to\infty} \sum_{\boldsymbol{k}\in\Lambda_\epsilon(R)} \frac{1}{n}\min_{y\in\mathcal{Y}} \left\{\left|\left\{i \in [n] : \widehat{\boldsymbol{x}}^i \in A^\epsilon_{\boldsymbol{k}}, \widehat{y}^i \neq y\right\}\right|\right\} \\
&= \sum_{\boldsymbol{k}\in\Lambda_\epsilon(R)} \left[\lim_{n\to\infty} \frac{1}{n}\min_{y\in\mathcal{Y}} \left\{\left|\left\{i \in [n] : \widehat{\boldsymbol{x}}^i \in A^\epsilon_{\boldsymbol{k}}, \widehat{y}^i \neq y\right\}\right|\right\}\right] = \sum_{\boldsymbol{k}\in\Lambda_\epsilon(R)} \min_{y\in\mathcal{Y}} \left\{\mathbb{P}(\widehat{\boldsymbol{x}}^1 \in A^\epsilon_{\boldsymbol{k}}, \widehat{y}^1 \neq y)\right\}.
\end{aligned} \tag{19d}$$

where the first equality is due to the interchangeability of finite sum and limit, the second one is due to (19c).

We may change the probability into Lebesgue integral by the definition (9a), (9b) of probability measures:

$$\mathbb{P}(\widehat{\boldsymbol{x}}^1 \in A^\epsilon_{\boldsymbol{k}}) = \int_{A^\epsilon_{\boldsymbol{k}}\cap\mathcal{X}} \mathrm{d}\mathbb{P}_0^{\mathcal{X}}, \tag{19e}$$

and

$$\begin{aligned}
&\mathbb{P}(\widehat{\boldsymbol{x}}^1 \in A^\epsilon_{\boldsymbol{k}}, \widehat{y}^1 = y) = \mathbb{P}(\widehat{\boldsymbol{x}}^1 \in A^\epsilon_{\boldsymbol{k}} | \widehat{y}^1 = y)\mathbb{P}(\widehat{y}^1 = y) = p_y \mathbb{P}_y(A^\epsilon_{\boldsymbol{k}}) \\
&= \int_{A^\epsilon_{\boldsymbol{k}}\cap\mathcal{X}} p_y \mathrm{d}\mathbb{P}_y = \int_{A^\epsilon_{\boldsymbol{k}}\cap\mathcal{X}} p_y \frac{\mathrm{d}\mathbb{P}_y}{\mathrm{d}\mathbb{P}_0^{\mathcal{X}}} \mathrm{d}\mathbb{P}_0^{\mathcal{X}} \leq \int_{A^\epsilon_{\boldsymbol{k}}\cap\mathcal{X}} \max_{y\in\mathcal{Y}} \left\{p_y \frac{\mathrm{d}\mathbb{P}_y}{\mathrm{d}\mathbb{P}_0^{\mathcal{X}}}\right\} \mathrm{d}\mathbb{P}_0^{\mathcal{X}}.
\end{aligned} \tag{19f}$$

Since the bound in (19f) holds for any $y \in \mathcal{Y}$, we can apply the maximum over y on the left-hand side of (19f) and the inequality still holds. Thus, we may derive that

$$\begin{aligned}&\min_{y\in\mathcal{Y}}\left\{\mathbb{P}(\widehat{\boldsymbol{x}}^1 \in A^\epsilon_{\boldsymbol{k}}, \widehat{y}^1 \neq y)\right\} = \mathbb{P}(\widehat{\boldsymbol{x}}^1 \in A^\epsilon_{\boldsymbol{k}}) - \max_{y\in\mathcal{Y}}\left\{\mathbb{P}(\widehat{\boldsymbol{x}}^1 \in A^\epsilon_{\boldsymbol{k}}, \widehat{y}^1 = y)\right\}\\ &\geq \int_{A^\epsilon_{\boldsymbol{k}}\cap\mathcal{X}} \mathrm{d}\mathbb{P}_0^{\mathcal{X}} - \int_{A^\epsilon_{\boldsymbol{k}}\cap\mathcal{X}} \max_{y\in\mathcal{Y}}\left\{p_y\frac{\mathrm{d}\mathbb{P}_y}{\mathrm{d}\mathbb{P}_0^{\mathcal{X}}}\right\}\mathrm{d}\mathbb{P}_0^{\mathcal{X}} = \int_{A^\epsilon_{\boldsymbol{k}}\cap\mathcal{X}}\left(1-\max_{y\in\mathcal{Y}}\left\{p_y\frac{\mathrm{d}\mathbb{P}_y}{\mathrm{d}\mathbb{P}_0^{\mathcal{X}}}\right\}\right)\mathrm{d}\mathbb{P}_0^{\mathcal{X}}.\end{aligned}$$

Therefore, we further derive from (19d) that

$$\begin{aligned}&\liminf_{n\to\infty} v_n^{\widehat{\mathcal{X}}_n} \geq \sum_{\boldsymbol{k}\in\Lambda_\epsilon(R)} \min_{y\in\mathcal{Y}}\left\{\mathbb{P}(\widehat{\boldsymbol{x}}^1 \in A^\epsilon_{\boldsymbol{k}}, \widehat{y}^1 \neq y)\right\} \geq \sum_{\boldsymbol{k}\in\Lambda_\epsilon(R)}\int_{A^\epsilon_{\boldsymbol{k}}\cap\mathcal{X}}\left(1-\max_{y\in\mathcal{Y}}\left\{p_y\frac{\mathrm{d}\mathbb{P}_y}{\mathrm{d}\mathbb{P}_0^{\mathcal{X}}}\right\}\right)\mathrm{d}\mathbb{P}_0^{\mathcal{X}}\\ &= \int_{\cup_{\boldsymbol{k}\in\Lambda_\epsilon(R)}(A^\epsilon_{\boldsymbol{k}}\cap\mathcal{X})}\left(1-\max_{y\in\mathcal{Y}}\left\{p_y\frac{\mathrm{d}\mathbb{P}_y}{\mathrm{d}\mathbb{P}_0^{\mathcal{X}}}\right\}\right)\mathrm{d}\mathbb{P}_0^{\mathcal{X}}.\end{aligned} \tag{19g}$$

We note that for any $\boldsymbol{k}\in\mathbb{Z}^d$ such that $A^\epsilon_{\boldsymbol{k}}\cap\mathcal{X}\neq\varnothing$, there exists $R'>0$ such that $A^\epsilon_{\boldsymbol{k}}\subseteq B_{\boldsymbol{0}}(R)$ for any $R\geq R'$, hence $\boldsymbol{k}\in\Lambda_\epsilon(R)$. Therefore, we have

$$\lim_{R\to+\infty}\bigcup_{\boldsymbol{k}\in\Lambda_\epsilon(R)}(A^\epsilon_{\boldsymbol{k}}\cap\mathcal{X}) = \mathcal{X}.$$

We then let R go to infinity in (19g). According to the monotone convergence theorem, we have

$$\begin{aligned}&\liminf_{n\to\infty} v_n^{\widehat{\mathcal{X}}_n} \geq \lim_{R\to+\infty}\int_{\cup_{\boldsymbol{k}\in\Lambda_\epsilon(R)}(A^\epsilon_{\boldsymbol{k}}\cap\mathcal{X})}\left(1-\max_{y\in\mathcal{Y}}\left\{p_y\frac{\mathrm{d}\mathbb{P}_y}{\mathrm{d}\mathbb{P}_0^{\mathcal{X}}}\right\}\right)\mathrm{d}\mathbb{P}_0^{\mathcal{X}}\\ &= \int_{\mathcal{X}}\left(1-\max_{y\in\mathcal{Y}}\left\{p_y\frac{\mathrm{d}\mathbb{P}_y}{\mathrm{d}\mathbb{P}_0^{\mathcal{X}}}\right\}\right)\mathrm{d}\mathbb{P}_0^{\mathcal{X}} = 1-\int_{\mathcal{X}}\left(\max_{y\in\mathcal{Y}}\left\{p_y\frac{\mathrm{d}\mathbb{P}_y}{\mathrm{d}\mathbb{P}_0^{\mathcal{X}}}\right\}\right)\mathrm{d}\mathbb{P}_0^{\mathcal{X}} = v_{UC},\end{aligned}$$

where the last equality is due to Lemma 3. This completes the proof. □

A.12 Proof of Lemma 4

Lemma 4 *Under Assumptions 1 and 2, we have* $\lim_{r\to 0_+}\gamma(r)=0$.

Proof: For any $0<r<r'$, if $\boldsymbol{x}\in U_y$ and $B_{\boldsymbol{x}}(r')\cap\mathcal{X}\subseteq U_y$, we have $B_{\boldsymbol{x}}(r)\cap\mathcal{X}\subseteq B_{\boldsymbol{x}}(r')\cap\mathcal{X}\subseteq U_y$, and $V_y(r')\subseteq V_y(r)$. Thus, $\{V_y(r)\}_{r>0}$ is inclusion-wise non-increasing in r. Hence, we have $\{U_y\setminus V_y(r)\}_{r>0}$ is inclusion-wise non-decreasing in r. Therefore, by the monotone convergence theorem (Durrett 2010), we have

$$\lim_{r\to 0_+}\mathbb{P}_0^{\mathcal{X}}(U_y\setminus V_y(r)) = \mathbb{P}_0^{\mathcal{X}}\left(\lim_{r\to 0_+}(U_y\setminus V_y(r))\right) = \mathbb{P}_0^{\mathcal{X}}\left(\bigcap_{r>0}(U_y\setminus V_y(r))\right).$$

Note that by the definition (10) of $V_y(r)$, we have $U_y\setminus V_y(r) = \{\boldsymbol{x}\in U_y : B_{\boldsymbol{x}}(r)\cap\mathcal{X}\nsubseteq U_y\}$. So we obtain that $\cap_{r>0}(U_y\setminus V_y(r)) = \{\boldsymbol{x}\in U_y : \forall r>0, B_{\boldsymbol{x}}(r)\cap\mathcal{X}\nsubseteq U_y\}\subseteq \mathrm{bd}^{\mathcal{X}}(U_y)$. By Assumption 2, we have

$$\lim_{r\to 0_+}\mathbb{P}_0^{\mathcal{X}}(U_y\setminus V_y(r)) = \mathbb{P}_0^{\mathcal{X}}\left(\bigcap_{r>0}(U_y\setminus V_y(r))\right)\leq \mathbb{P}_0^{\mathcal{X}}(\mathrm{bd}^{\mathcal{X}}(U_y)) = 0.$$

Taking the summation over $y\in\mathcal{Y}$, by the definition (11) of $\gamma(r)$, we have $\lim_{r\to 0_+}\gamma(r)=0$. □

A.13 Proof of Theorem 4

Theorem 4 *Under Assumptions 1 and 2, given any $\kappa \in \mathbb{R}_{++}$ and $\epsilon \in \mathbb{R}_{++}$, there exists a $\bar{\theta}$ such that for any $\theta \in (0, \bar{\theta}]$, we have*

$$\limsup_{n\to\infty} v_n^{\widehat{\mathcal{X}}_n} \le \limsup_{n\to\infty} v_n^{\mathcal{X}} \le v_{UC} + \epsilon.$$

almost surely.

Proof: The first inequality is a consequence of the definition of the in-sample DRUC problem. We mainly prove the second inequality.

By Lemma 4, we can pick $r \le \kappa$ such that $\gamma(r) \le \frac{\epsilon}{2}$. Let $\bar{\theta} = \frac{\epsilon r}{2}$, for any $0 < \theta \le \bar{\theta}$, let us consider the optimal UC policy f_0, defined as $f_0(\boldsymbol{x}) = y$ for each $\boldsymbol{x} \in U_y$ and $y \in \mathcal{Y}$, which aligns with the definition in the proof of Lemma 3.

By Theorem 1, we have

$$v_n^{\mathcal{X}}(f_0) = \frac{k_n(f_0, \theta)}{n}.$$

According to (6a), we have

$$J_n(f_0) = \left\{ i \in [n] : f_0(\widehat{\boldsymbol{x}}^i) \neq \widehat{y}^i \right\}.$$

And for $i \in J_n(f_0)$, we have $s_i^n(f_0) = 0$. By the strong law of large numbers, we have

$$\lim_{n\to\infty} \frac{|J_n(f_0)|}{n} = \lim_{n\to\infty} \frac{1}{n} \left| \left\{ i \in [n] : f_0(\widehat{\boldsymbol{x}}^i) \neq \widehat{y}^i \right\} \right| = \lim_{n\to\infty} \sum_{i\in[n]} \frac{1}{n} \mathbb{1}_{\{f_0(\widehat{\boldsymbol{x}}^1) \neq \widehat{y}^1\}} = \mathbb{P}(f_0(\widehat{\boldsymbol{x}}^1) \neq \widehat{y}^1) = v_{UC}, \tag{20a}$$

where the last equation follows from the proof of Lemma 3.

Let us denote

$$M_n(r) = \left\{ i \in I_n(f_0) : \widehat{\boldsymbol{x}}^i \in \bigcup_{y\in\mathcal{Y}} (U_y \setminus V_y(r)) \right\},$$

where the set $I_n(f_0)$ is as defined in (6a). By the strong law of large numbers and the definition of $\gamma(r)$, we have almost surely,

$$\limsup_{n\to\infty} \frac{|M_n(r)|}{n} \le \gamma(r).$$

For each $i \in I_n(f_0)$, if $\widehat{\boldsymbol{x}}^i \in \bigcup_{y\in\mathcal{Y}} V_y(r)$, then we must have $\widehat{\boldsymbol{x}}^i \in V_{\widehat{y}^i}(r)$ since $f_0(\widehat{\boldsymbol{x}}^i) = \widehat{y}^i$. We then have $s_i^n(f_0) \ge r$, since by definition of $V_{\widehat{y}^i}(r)$ that $B_{\widehat{\boldsymbol{x}}^i}(r) \cap \mathcal{X} \subseteq U_{\widehat{y}^i}$, we have $\operatorname{dist}(\widehat{\boldsymbol{x}}^i, f_0^{-1}(\mathcal{Y} \setminus \{\widehat{y}^i\})) \ge r$. Therefore, we can obtain a lower bound on each of $\{s_i^n(f_0)\}_{i\in[n]}$:

$$s_i^n(f_0) \begin{cases} = 0, & \text{if } i \in J_n(f_0); \\ \ge 0, & \text{if } i \in M_n(r); \\ \ge r, & \text{if } i \in I_n(f_0) \setminus M_n(r). \end{cases}$$

By Corollary 1, we can derive an upper bound on $k_n(f_0,\theta)$ by substituting $\{s_i^n(f_0)\}_{i\in[n]}$ with appropriate lower bounds. In particular, according to Theorem 1, this upper bound can be expressed as

$$k_n(f_0,\theta) \le |J_n(f_0)| + |M_n(r)| + \frac{n\theta}{r}.$$

Dividing both sides by n, we then obtain an asymptotic upper bound of $v_n^{\mathcal{X}}(f_0) = k_n(f_0,\theta)/n$ by taking limsup on both sides:

$$\limsup_{n\to\infty} \frac{k_n(f_0,\theta)}{n} \le \left(\lim_{n\to\infty} \frac{|J_n(f_0)|}{n}\right) + \left(\limsup_{n\to\infty} \frac{|M_n(r)|}{n}\right) + \frac{\theta}{r} \le v_{UC} + \frac{\epsilon}{2} + \frac{\epsilon}{2} = v_{UC} + \epsilon,$$

where the second inequality is due to the choice of parameters such that $\gamma(r) \le \frac{\epsilon}{2}$ and $\theta \le \frac{\epsilon r}{2}$. By Theorem 1, we conclude that $\limsup_{n\to\infty} v_n^{\mathcal{X}}(f_0) \le v_{UC} + \epsilon$. This completes the proof. □

A.14 Proof of Proposition 3

Proposition 3 *Under Assumptions 1 and 3, suppose that the feature space is $\mathcal{X} = \mathbb{R}^d$, the marginal distribution $\mathbb{P}_0^{\mathcal{X}}$ of the true distribution $\mathbb{P}_0$ has a continuous density function g, and there exists a positive function $\phi : \mathbb{R}^+ \to \mathbb{R}^+$ such that for any $R > 0$ and any $\boldsymbol{x} \in B_{\mathbf{0}}(R)$, we have $g(\boldsymbol{x}) \ge \phi(R)$. Then, under the conditions of Theorem 2, for any $\eta > 0$, if*

$$n \ge \frac{-\kappa^2(3+\eta)^2 \ln\frac{\alpha}{2}}{2\eta^2\theta^2} \vee \frac{\ln \frac{2(\pi d)^{d/2}(R+\kappa)^d(3+\eta)^d}{\Gamma\left(\frac{d}{2}+1\right)\alpha(\eta\theta)^d}}{\phi(R+\kappa)\left(\frac{\eta\theta}{(3+\eta)\sqrt{d}}\right)^d}, \text{ where } R = c\sqrt{-\ln\frac{\eta\theta}{2\kappa(3+\eta)}},$$

then we obtain the following finite-sample guarantee: $\mathbb{P}(v_n^{\mathcal{X}} - v_n^{\widehat{\mathcal{X}}_n} \ge \eta) \le \alpha$.

Proof: According to the proof of Theorem 2, we need to bound two probabilities $\mathbb{P}(\frac{r_n(R)}{n} \ge 2\beta(R))$ and $\mathbb{P}(\bigcup_{i\in[n]:\widehat{\boldsymbol{x}}^i\in B_{\mathbf{0}}(R)}\{s_i^n(\widetilde{f}) < s_i^n(\widehat{f}) - \epsilon\})$.

Step I. Let $\epsilon = \frac{\eta\theta}{3+\eta}$. That is, $\frac{\epsilon}{\theta-\epsilon} = \frac{\eta}{3}$. Take R such that $\beta(R) = \mathbb{P}(\widehat{\boldsymbol{x}}^1 \notin B_{\mathbf{0}}(R)) \le \frac{\eta\theta}{\kappa(3+\eta)}$. By Assumption 3, let us pick

$$R = c\sqrt{-\ln\frac{\eta\theta}{2\kappa(3+\eta)}}.$$

Recall that $r_n(R) = \sum_{i\in[n]} \mathbb{1}_{\{\widehat{\boldsymbol{x}}^i \notin B_{\mathbf{0}}(R)\}}$. By Hoeffding's inequality, we know that

$$\mathbb{P}\left(\frac{r_n(R)}{n} \ge 2\beta(R)\right) = \mathbb{P}\left(\frac{\sum_{i\in[n]} \mathbb{1}_{\{\widehat{\boldsymbol{x}}^i \notin B_{\mathbf{0}}(R)\}}}{n} \ge 2\beta(R)\right) \le \exp(-2n\beta(R)^2).$$

If the sample size satisfies

$$n \ge \frac{-\ln\frac{\alpha}{2}}{2\beta(R)^2} = \frac{-\kappa^2(3+\eta)^2\ln\frac{\alpha}{2}}{2\eta^2\theta^2},$$

then we have

$$\mathbb{P}\left(\frac{r_n(R)}{n}\frac{\kappa}{\theta-\epsilon} \ge \frac{2\eta}{3}\right) = \mathbb{P}\left(\frac{r_n(R)}{n} \ge 2\beta(R)\right) \le \frac{\alpha}{2}.$$

Step II. Consider the contrapositive of Lemma 2. Specifically, suppose that for some $i \in [n]$ with $\widehat{\boldsymbol{x}}^i \in B_{\mathbf{0}}(R)$, the inequality $s_i^n(\widetilde{f}) < s_i^n(\widehat{f}) - \epsilon$ holds. Then, there must exist an index $\boldsymbol{k} \in \bar{\Lambda}_{\epsilon/\sqrt{d}}(R+\kappa-\epsilon)$ such that $\widehat{\boldsymbol{x}}^i \notin \mathrm{cl}(A_{\boldsymbol{k}}^{\epsilon/\sqrt{d}})$ for all $i \in [n]$. We aim to bound the probability of this event. Note that every cube indexed by $\boldsymbol{k} \in \bar{\Lambda}_{\epsilon/\sqrt{d}}(R+\kappa-\epsilon)$ is contained in the ball $B_{\mathbf{0}}(R+\kappa)$. Therefore, we have

$$\begin{aligned}
&\mathbb{P}\left(\bigcup_{i\in[n]:\,\widehat{\boldsymbol{x}}^i\in B_{\mathbf{0}}(R)}\left\{s_i^n(\widetilde{f}) < s_i^n(\widehat{f})-\epsilon\right\}\right) \le \mathbb{P}\left(\bigcup_{\boldsymbol{k}\in\bar{\Lambda}_{\epsilon/\sqrt{d}}(R+\kappa-\epsilon)}\left\{\widehat{\boldsymbol{x}}^i \notin \mathrm{cl}(A_{\boldsymbol{k}}^{\epsilon/\sqrt{d}}),\, \forall i\in[n]\right\}\right)\\
&\le \sum_{\boldsymbol{k}\in\bar{\Lambda}_{\epsilon/\sqrt{d}}(R+\kappa-\epsilon)} \mathbb{P}\left(\widehat{\boldsymbol{x}}^i \notin \mathrm{cl}(A_{\boldsymbol{k}}^{\epsilon/\sqrt{d}}),\, \forall i\in[n]\right) = \sum_{\boldsymbol{k}\in\bar{\Lambda}_{\epsilon/\sqrt{d}}(R+\kappa-\epsilon)} \mathbb{P}\left(\widehat{\boldsymbol{x}}^1 \notin \mathrm{cl}(A_{\boldsymbol{k}}^{\epsilon/\sqrt{d}})\right)^n,
\end{aligned}$$

where the last inequality follows from the union bound and the i.i.d. assumption of the data points. Each hypercube $\mathrm{cl}(A_{\boldsymbol{k}}^{\epsilon/\sqrt{d}})$ with $\boldsymbol{k} \in \bar{\Lambda}_{\epsilon/\sqrt{d}}(R+\kappa-\epsilon)$ is a subset of $B_{\mathbf{0}}(R+\kappa)$. Thus, we have

$$\mathbb{P}\left(\widehat{\boldsymbol{x}}^1 \in \mathrm{cl}(A_{\boldsymbol{k}}^{\epsilon/\sqrt{d}})\right) \ge \phi(R+\kappa)\left(\frac{\epsilon}{\sqrt{d}}\right)^d.$$

We next bound the cardinality of $\bar{\Lambda}_{\epsilon/\sqrt{d}}(R+\kappa-\epsilon)$ by the ratio of the volume of the ball $B_{\mathbf{0}}(R+\kappa)$ and that of each hypercube, i.e.,

$$\left|\bar{\Lambda}_{\epsilon/\sqrt{d}}(R+\kappa-\epsilon)\right| \le \frac{\frac{\pi^{d/2}}{\Gamma\left(\frac{d}{2}+1\right)}(R+\kappa)^d}{\left(\frac{\epsilon}{\sqrt{d}}\right)^d}.$$

Therefore, we have

$$\begin{aligned}
&\mathbb{P}\left(\bigcup_{i\in I_n(\widehat{f}):\,\widehat{\boldsymbol{x}}^i\in B_{\mathbf{0}}(R)}\left\{s_i^n(\widetilde{f}) < s_i^n(\widehat{f})-\epsilon\right\}\right) \le \sum_{\boldsymbol{k}\in\bar{\Lambda}_{\epsilon/\sqrt{d}}(R+\kappa-\epsilon)} \mathbb{P}\left(\widehat{\boldsymbol{x}}^1 \notin \mathrm{cl}(A_{\boldsymbol{k}}^{\epsilon/\sqrt{d}})\right)^n\\
&\le \left|\bar{\Lambda}_{\epsilon/\sqrt{d}}(R+\kappa-\epsilon)\right|\left(1-\phi(R+\kappa)\left(\frac{\epsilon}{\sqrt{d}}\right)^d\right)^n \le \frac{(\pi d)^{d/2}(R+\kappa)^d}{\Gamma\left(\frac{d}{2}+1\right)\epsilon^d}\left(1-\phi(R+\kappa)\left(\frac{\epsilon}{\sqrt{d}}\right)^d\right)^n,
\end{aligned}$$

Thus, when the sample size satisfies

$$n \ge \frac{\ln\frac{2(\pi d)^{d/2}(R+\kappa)^d}{\Gamma\left(\frac{d}{2}+1\right)\alpha\epsilon^d}}{\phi(R+\kappa)\left(\frac{\epsilon}{\sqrt{d}}\right)^d} \ge \frac{\ln\frac{2(\pi d)^{d/2}(R+\kappa)^d}{\Gamma\left(\frac{d}{2}+1\right)\alpha\epsilon^d}}{-\ln\left(1-\phi(R+\kappa)\left(\frac{\epsilon}{\sqrt{d}}\right)^d\right)},$$

we have

$$\mathbb{P}\left(\bigcup_{i\in[n]:\,\widehat{\boldsymbol{x}}^i\in B_{\mathbf{0}}(R)}\left\{s_i^n(\widetilde{f}) < s_i^n(\widehat{f})-\epsilon\right\}\right) \le \frac{\alpha}{2}.$$

Step III. Combining the two steps and using the union bound, when

$$n \ge \frac{-\kappa^2(3+\eta)^2\ln\frac{\alpha}{2}}{2\eta^2\theta^2} \vee \frac{\ln\frac{2(\pi d)^{d/2}(R+\kappa)^d(3+\eta)^d}{\Gamma\left(\frac{d}{2}+1\right)\alpha(\eta\theta)^d}}{\phi(R+\kappa)\left(\frac{\eta\theta}{(3+\eta)\sqrt{d}}\right)^d},$$

we have

$$\mathbb{P}(v_n^{\mathcal{X}} - v_n^{\widehat{\mathcal{X}}_n} \geq \eta) \leq \mathbb{P}\left(\frac{r_n(R)}{n}\frac{\kappa}{\theta-\epsilon} \geq \frac{2\eta}{3}\right) + \mathbb{P}\left(\bigcup_{i\in[n]:\,\widehat{\boldsymbol{x}}^i \in B_{\mathbf{0}}(R)} \left\{s_i^n(\widetilde{f}) < s_i^n(\widehat{f}) - \epsilon\right\}\right) \leq \alpha.$$

This completes our proof. □

A.15 Proof of Proposition 4

Proposition 4 *Under Assumption 1,suppose that for each $y \in \mathcal{Y}$, the conditional distribution of $(\boldsymbol{X}, Y) \sim \mathbb{P}_0$ given $Y = y$ is supported on a hyper-rectangle $H_y = \prod_{j=1}^d (a_{y,j}, b_{y,j})$. Let the marginal distribution $\mathbb{P}_0^{\mathcal{X}}$ have a continuous density function g, and suppose that the side lengths of all hyper-rectangles are uniformly bounded, i.e., $L \leq b_{y,j} - a_{y,j} \leq U$ for all $j \in [d]$ and $y \in \mathcal{Y}$. Furthermore, suppose there exists a constant $\zeta > 0$ such that $g(\boldsymbol{x}) \geq \zeta$ for all $\boldsymbol{x} \in \bigcup_{y\in\mathcal{Y}} H_y$. Then, under the conditions of Theorem 2, given any $\eta > 0$, if*

$$n \geq \frac{\ln\frac{m}{\alpha} + d\ln\frac{(1+\eta)U\sqrt{d}}{\eta\theta}}{\zeta\left(\frac{\eta\theta L}{L\sqrt{d}+\eta(L\sqrt{d}+\theta)}\right)^d},$$

then we have $\mathbb{P}(v_n^{\mathcal{X}} - v_n^{\widehat{\mathcal{X}}_n} \geq \eta) \leq \alpha$.

Proof: Different from the proof of Proposition 3, due to the bounded support, we only need to bound the probability $\mathbb{P}(\bigcup_{i\in[n]:\,\widehat{\boldsymbol{x}}^i \in B_{\mathbf{0}}(R)} \{s_i^n(\widetilde{f}) < s_i^n(\widehat{f}) - \epsilon\})$.

Take $\epsilon = \frac{\eta\theta}{1+\eta}$. Then we have $\frac{\epsilon}{\theta-\epsilon} = \eta$. Next, we introduce an alternative partitioning scheme for the feature space, which is conceptually similar but different from the approach used in Lemma 1. For each $y \in \mathcal{Y}$ and $j \in [d]$, define

$$z_{y,j}^{\epsilon} = \frac{b_{y,j} - a_{y,j}}{\left\lceil \frac{b_{y,j}-a_{y,j}}{\epsilon/\sqrt{d}} \right\rceil},$$

which serves as the adjusted side length of the partition cells along the j-th dimension. Based on this, we define the partition elements

$$A_{k_{y,1},k_{y,2},\cdots,k_{y,d}}^{rec,\epsilon,y} = \prod_{j=1}^d \left[a_{y,j} + (k_{y,j}-1)z_{y,j}, a_{y,j} + k_{y,j}z_{y,j}\right),$$

where the multi-index $(k_{y,1}, k_{y,2}, \ldots, k_{y,d})$ belongs to the index set

$$\bar{\Lambda}_{rec,\epsilon,y} = \prod_{j=1}^d \left\{1, \cdots, \left\lceil \frac{b_{y,j} - a_{y,j}}{\epsilon/\sqrt{d}} \right\rceil\right\},$$

for each $y \in \mathcal{Y}$.

Consider the contrapositive of Lemma 2. If the inequality $s_i^n(\widetilde{f}) < s_i^n(\widehat{f}) - \epsilon$ holds for some $i \in [n]$, there must exist some $(k_{y,j})_{j\in[d]} \in \bar{\Lambda}_{rec,\epsilon,y}$ for some $y \in \mathcal{Y}$ such that $\widehat{\boldsymbol{x}}^i \notin \mathrm{cl}(A_{\boldsymbol{k}}^{rec,\epsilon,y})$ for all $i \in [n]$. We then have

$$\begin{aligned}&\mathbb{P}\left(\bigcup_{i\in I_n(\widehat{f})}\left\{s_i^n(\widetilde{f}) < s_i^n(\widehat{f}) - \epsilon\right\}\right) \le \mathbb{P}\left(\bigcup_{y\in\mathcal{Y}}\bigcup_{\boldsymbol{k}\in\bar{\Lambda}_{rec,\epsilon,y}}\left\{\widehat{\boldsymbol{x}}^i \notin \mathrm{cl}(A_{\boldsymbol{k}}^{rec,\epsilon,y}),\, \forall i\in[n]\right\}\right)\\ \le &\sum_{y\in\mathcal{Y}}\sum_{\boldsymbol{k}\in\bar{\Lambda}_{rec,\epsilon,y}} \mathbb{P}\left(\widehat{\boldsymbol{x}}^i \notin \mathrm{cl}(A_{\boldsymbol{k}}^{rec,\epsilon,y}),\, \forall i\in[n]\right) = \sum_{y\in\mathcal{Y}}\sum_{\boldsymbol{k}\in\bar{\Lambda}_{rec,\epsilon,y}} \mathbb{P}\left(\widehat{\boldsymbol{x}}^1 \notin \mathrm{cl}(A_{\boldsymbol{k}}^{rec,\epsilon,y})\right)^n.\end{aligned}$$

We note that for any $y \in \mathcal{Y}$, we may bound the number of hypercubes as:

$$|\bar{\Lambda}_{rec,\epsilon,y}| = \prod_{j=1}^d \left\lceil \frac{b_{y,j}-a_{y,j}}{\epsilon/\sqrt{d}} \right\rceil \le \prod_{j=1}^d \left(\frac{b_{y,j}-a_{y,j}}{\epsilon/\sqrt{d}} + 1\right) \le \left(\frac{U\sqrt{d}}{\epsilon} + 1\right)^d.$$

For any $y \in \mathcal{Y}$ and $j \in [d]$, we can obtain a lower bound on the side length of the hypercube:

$$z_{y,j}^\epsilon = \frac{b_{y,j}-a_{y,j}}{\lceil \frac{b_{y,j}-a_{y,j}}{\epsilon/\sqrt{d}} \rceil} \ge \frac{b_{y,j}-a_{y,j}}{\frac{b_{y,j}-a_{y,j}}{\epsilon/\sqrt{d}} + 1} = \frac{1}{\frac{1}{\epsilon/\sqrt{d}} + \frac{1}{b_{y,j}-a_{y,j}}} \ge \frac{1}{\frac{1}{\epsilon/\sqrt{d}} + \frac{1}{L}} = \frac{\eta\theta L}{L\sqrt{d} + \eta(L\sqrt{d}+\theta)}.$$

Therefore, we bound the probability that a feature point lies in a hypercube as:

$$\mathbb{P}\left(\widehat{\boldsymbol{x}}^1 \in \mathrm{cl}(A_{\boldsymbol{k}}^{rec,\epsilon,y})\right) \ge \zeta \prod_{j=1}^d z_{y,j}^\epsilon \ge \zeta\left(\frac{\eta\theta L}{L\sqrt{d} + \eta(L\sqrt{d}+\theta)}\right)^d.$$

Therefore, we have:

$$\begin{aligned}&\mathbb{P}\left(\bigcup_{i\in I_n(\widehat{f})}\left\{s_i^n(\widetilde{f}) < s_i^n(\widehat{f}) - \epsilon\right\}\right) \le \sum_{y\in\mathcal{Y}}\sum_{\boldsymbol{k}\in\bar{\Lambda}_{rec,\epsilon,y}} \mathbb{P}\left(\widehat{\boldsymbol{x}}^1 \notin \mathrm{cl}(A_{\boldsymbol{k}}^{rec,\epsilon,y})\right)^n\\ \le &m\left(\frac{U\sqrt{d}}{\epsilon}+1\right)^d\left(1-\zeta\left(\frac{\eta\theta L}{L\sqrt{d}+\eta(L\sqrt{d}+\theta)}\right)^d\right)^n.\end{aligned}$$

Thus, when

$$n \ge \frac{\ln\frac{m}{\alpha} + d\ln\left(\frac{(1+\eta)U\sqrt{d}}{\eta\theta}+1\right)}{\zeta\left(\frac{\eta\theta L}{L\sqrt{d}+\eta(L\sqrt{d}+\theta)}\right)^d},$$

we have

$$\mathbb{P}(v_n^{\mathcal{X}} - v_n^{\widehat{\mathcal{X}}_n} \ge \eta) \le \mathbb{P}\left(\bigcup_{i\in I_n(\widehat{f})}\left\{s_i^n(\widetilde{f}) < s_i^n(\widehat{f}) - \epsilon\right\}\right) \le \alpha.$$

This completes our proof. □

A.16 Proof of Proposition 5

Proposition 5 *Under Assumptions 1, 3, and the conditions of Theorem 3, given any $\eta > 0$, when the sample size satisfies*

$$n \geq \frac{2}{\eta^2}\left(\frac{(\pi d)^{d/2}(c\sqrt{-\ln\frac{\eta}{4}}+\theta)^d}{\Gamma\left(\frac{d}{2}+1\right)\theta^d}\ln m - \ln\alpha\right),$$

then we have $\mathbb{P}(v_n^{\widehat{\mathcal{X}}_n} \leq v_{UC} - \eta) \leq \alpha$.

Proof: We select R such that $\beta(R) = \mathbb{P}(\widehat{\boldsymbol{x}}^1 \notin B_{\mathbf{0}}(R)) \leq \frac{\eta}{2}$. Under Assumption 3, it suffices to set $R = c\sqrt{-\ln\frac{\eta}{4}}$. Thus, according to Lemma 3, we have

$$\begin{aligned} &v_{UC} - \sum_{\boldsymbol{k}\in\Lambda_\epsilon(R)}\left[\int_{A_{\boldsymbol{k}}^\epsilon\cap\mathcal{X}}\left(1-\max_{y\in\mathcal{Y}}\left\{p_y\frac{\mathrm{d}\mathbb{P}_y}{\mathrm{d}\mathbb{P}_0^{\mathcal{X}}}\right\}\right)\mathrm{d}\mathbb{P}_0^{\mathcal{X}}\right] \\ &\leq \int_{\mathcal{X}\setminus\left(\bigcup_{\boldsymbol{k}\in\Lambda_\epsilon(R)}A_{\boldsymbol{k}}^\epsilon\right)}\left(1-\max_{y\in\mathcal{Y}}\left\{p_y\frac{\mathrm{d}\mathbb{P}_y}{\mathrm{d}\mathbb{P}_0^{\mathcal{X}}}\right\}\right)\mathrm{d}\mathbb{P}_0^{\mathcal{X}} \leq \int_{\mathcal{X}\setminus B_{\mathbf{0}}(R)}\mathrm{d}\mathbb{P}_0^{\mathcal{X}} = \beta(R) = \frac{\eta}{2}, \end{aligned} \tag{21a}$$

where the second inequality is because $1-\max_{y\in\mathcal{Y}}\{p_y\frac{\mathrm{d}\mathbb{P}_y}{\mathrm{d}\mathbb{P}_0^{\mathcal{X}}}\} \leq 1$ and $B_{\mathbf{0}}(R) \subseteq \bigcup_{\boldsymbol{k}\in\Lambda_\epsilon(R)} A_{\boldsymbol{k}}^\epsilon$. In the proof of Theorem 3, we show the inequalities

$$\int_{A_{\boldsymbol{k}}^\epsilon\cap\mathcal{X}}\left(1-\max_{y\in\mathcal{Y}}\left\{p_y\frac{\mathrm{d}\mathbb{P}_y}{\mathrm{d}\mathbb{P}_0^{\mathcal{X}}}\right\}\right)\mathrm{d}\mathbb{P}_0^{\mathcal{X}} \leq \min_{y\in\mathcal{Y}}\left\{\mathbb{P}(\widehat{\boldsymbol{x}}^1\in A_{\boldsymbol{k}}^\epsilon, \widehat{y}^1\neq y)\right\},$$

$$v_n^{\widehat{\mathcal{X}}_n} \geq \sum_{\boldsymbol{k}\in\Lambda_\epsilon(R)}\frac{1}{n}\min_{y\in\mathcal{Y}}\left\{\left|\left\{i\in[n]:\widehat{\boldsymbol{x}}^i\in A_{\boldsymbol{k}}^\epsilon,\widehat{y}^i\neq y\right\}\right|\right\}.$$

Therefore, together with the inequality in (21a), the probability $\mathbb{P}(v_n^{\widehat{\mathcal{X}}_n} \leq v_{UC}-\eta)$ must be less than or equal to the probability

$$\mathbb{P}\left(\sum_{\boldsymbol{k}\in\Lambda_\epsilon(R)}\frac{1}{n}\min_{y\in\mathcal{Y}}\left\{\left|\left\{i\in[n]:\widehat{\boldsymbol{x}}^i\in A_{\boldsymbol{k}}^\epsilon,\widehat{y}^i\neq y\right\}\right|\right\} \leq \sum_{\boldsymbol{k}\in\Lambda_\epsilon(R)}\min_{y\in\mathcal{Y}}\left\{\mathbb{P}(\widehat{\boldsymbol{x}}^1\in A_{\boldsymbol{k}}^\epsilon,\widehat{y}^1\neq y)\right\}-\frac{\eta}{2}\right). \tag{21b}$$

We simplify the proof by introducing the following random variable for each $\boldsymbol{k}\in\mathbb{Z}^d$ and $y\in\mathcal{Y}$

$$S_{\boldsymbol{k},y} = \left|\left\{i\in[n]:\widehat{\boldsymbol{x}}^i\in A_{\boldsymbol{k}}^\epsilon,\widehat{y}^i\neq y\right\}\right| = \sum_{i\in[n]}\mathbb{1}_{\{\widehat{\boldsymbol{x}}^i\in A_{\boldsymbol{k}}^\epsilon,\widehat{y}^i\neq y\}},$$

and the constant

$$q_{\boldsymbol{k},y} = \mathbb{P}(\widehat{\boldsymbol{x}}^1\in A_{\boldsymbol{k}}^\epsilon,\widehat{y}^1\neq y).$$

Thus, we equivalently represent (21b) as

$$\mathbb{P}\left(\sum_{\boldsymbol{k}\in\Lambda_\epsilon(R)}\min_{y\in\mathcal{Y}}\frac{S_{\boldsymbol{k},y}}{n} \leq \sum_{\boldsymbol{k}\in\Lambda_\epsilon(R)}\min_{y\in\mathcal{Y}}q_{\boldsymbol{k},y}-\frac{\eta}{2}\right).$$

Next, we transform the summation of minima into the minimum of all possible summations. Then, by applying the union bound, we obtain

$$\left\{\sum_{\boldsymbol{k}\in\Lambda_\epsilon(R)}\min_{y\in\mathcal{Y}}\frac{S_{\boldsymbol{k},y}}{n}\le\sum_{\boldsymbol{k}\in\Lambda_\epsilon(R)}\min_{y\in\mathcal{Y}}q_{\boldsymbol{k},y}-\frac{\eta}{2}\right\}=\left\{\min_{\boldsymbol{y}\in\mathcal{Y}^{\Lambda_\epsilon(R)}}\sum_{\boldsymbol{k}\in\Lambda_\epsilon(R)}\frac{S_{\boldsymbol{k},y_{\boldsymbol{k}}}}{n}\le\sum_{\boldsymbol{k}\in\Lambda_\epsilon(R)}\min_{y\in\mathcal{Y}}q_{\boldsymbol{k},y}-\frac{\eta}{2}\right\}$$
$$=\bigcup_{\boldsymbol{y}\in\mathcal{Y}^{\Lambda_\epsilon(R)}}\left\{\sum_{\boldsymbol{k}\in\Lambda_\epsilon(R)}\frac{S_{\boldsymbol{k},y_{\boldsymbol{k}}}}{n}\le\sum_{\boldsymbol{k}\in\Lambda_\epsilon(R)}\min_{y\in\mathcal{Y}}q_{\boldsymbol{k},y}-\frac{\eta}{2}\right\}\subseteq\bigcup_{\boldsymbol{y}\in\mathcal{Y}^{\Lambda_\epsilon(R)}}\left\{\sum_{\boldsymbol{k}\in\Lambda_\epsilon(R)}\frac{S_{\boldsymbol{k},y_{\boldsymbol{k}}}}{n}\le\sum_{\boldsymbol{k}\in\Lambda_\epsilon(R)}q_{\boldsymbol{k},y_{\boldsymbol{k}}}-\frac{\eta}{2}\right\}.$$

Therefore,

$$\mathbb{P}\left(\sum_{\boldsymbol{k}\in\Lambda_\epsilon(R)}\min_{y\in\mathcal{Y}}\frac{S_{\boldsymbol{k},y}}{n}\le\sum_{\boldsymbol{k}\in\Lambda_\epsilon(R)}\min_{y\in\mathcal{Y}}q_{\boldsymbol{k},y}-\frac{\eta}{2}\right)\le\sum_{\boldsymbol{y}\in\mathcal{Y}^{\Lambda_\epsilon(R)}}\mathbb{P}\left(\sum_{\boldsymbol{k}\in\Lambda_\epsilon(R)}\frac{S_{\boldsymbol{k},y_{\boldsymbol{k}}}}{n}\le\sum_{\boldsymbol{k}\in\Lambda_\epsilon(R)}q_{\boldsymbol{k},y_{\boldsymbol{k}}}-\frac{\eta}{2}\right).$$

Note that

$$\sum_{\boldsymbol{k}\in\Lambda_\epsilon(R)}\frac{S_{\boldsymbol{k},y_{\boldsymbol{k}}}}{n}=\frac{1}{n}\sum_{i\in[n]}\left(\sum_{\boldsymbol{k}\in\Lambda_\epsilon(R)}\mathbb{1}_{\{\widehat{\boldsymbol{x}}^i\in A^\epsilon_{\boldsymbol{k}},\widehat{y}^i\neq y_{\boldsymbol{k}}\}}\right),$$

where

$$0\le\sum_{\boldsymbol{k}\in\Lambda_\epsilon(R)}\mathbb{1}_{\{\widehat{\boldsymbol{x}}^i\in A^\epsilon_{\boldsymbol{k}},\widehat{y}^i\neq y_{\boldsymbol{k}}\}}\le 1,$$

and

$$\mathbb{E}\left[\sum_{\boldsymbol{k}\in\Lambda_\epsilon(R)}\mathbb{1}_{\{\widehat{\boldsymbol{x}}^i\in A^\epsilon_{\boldsymbol{k}},\widehat{y}^i\neq y_{\boldsymbol{k}}\}}\right]=\sum_{\boldsymbol{k}\in\Lambda_\epsilon(R)}\mathbb{P}\left(\widehat{\boldsymbol{x}}^i\in A^\epsilon_{\boldsymbol{k}},\widehat{y}^i\neq y_{\boldsymbol{k}}\right)=\sum_{\boldsymbol{k}\in\Lambda_\epsilon(R)}q_{\boldsymbol{k},y_{\boldsymbol{k}}}$$

By Hoeffding's inequality, we have

$$\mathbb{P}\left(\sum_{\boldsymbol{k}\in\Lambda_\epsilon(R)}\frac{S_{\boldsymbol{k},y_{\boldsymbol{k}}}}{n}\le\sum_{\boldsymbol{k}\in\Lambda_\epsilon(R)}q_{\boldsymbol{k},y_{\boldsymbol{k}}}-\frac{\eta}{2}\right)\le\exp\left(-\frac{n\eta^2}{2}\right)$$

According to the proof of Proposition 3, we know that

$$|\Lambda_\epsilon(R)|\le\frac{\pi^{d/2}(R+\sqrt{d}\epsilon)^d}{\Gamma\left(\frac{d}{2}+1\right)\epsilon^d}.$$

So when

$$n\ge\frac{2}{\eta^2}\left(\frac{\pi^{d/2}(R+\sqrt{d}\epsilon)^d}{\Gamma\left(\frac{d}{2}+1\right)\epsilon^d}\ln m-\ln\alpha\right),$$

we have

$$\mathbb{P}\left(\sum_{\boldsymbol{k}\in\Lambda_\epsilon(R)}\min_{y\in\mathcal{Y}}\frac{S_{\boldsymbol{k},y}}{n}\le\sum_{\boldsymbol{k}\in\Lambda_\epsilon(R)}\min_{y\in\mathcal{Y}}q_{\boldsymbol{k},y}-\frac{\eta}{2}\right)\le m^{|\Lambda_\epsilon(R)|}\exp\left(-\frac{n\eta^2}{2}\right)\le\alpha.$$

By Theorem 3, if we let $\epsilon=\frac{\theta}{\sqrt{d}}$, then we obtain the sample size

$$n\ge\frac{2}{\eta^2}\left(\frac{(\pi d)^{d/2}(c\sqrt{-\ln\frac{\eta}{4}}+\theta)^d}{\Gamma\left(\frac{d}{2}+1\right)\theta^d}\ln m-\ln\alpha\right).$$

This completes our proof. □

A.17 Proof of Proposition 6

Proposition 6 *Under Assumptions 1, 2 and the conditions of Theorem 4, given any $\eta \in (0,1)$, if $\gamma(r) \leq \frac{\eta}{3}$, $\theta = \frac{\eta r}{3}$, and $n \geq \frac{-9\ln\alpha}{2\eta^2}$, then we have $\mathbb{P}(v_n^{\mathcal{X}} \geq v_{UC} + \eta) \leq \alpha$.*

Proof: Choose r such that $\gamma(r) \leq \frac{\eta}{3}$ and let $\theta = \frac{\eta r}{3}$. We have shown in the proof of Theorem 4 that

$$v_n^{\mathcal{X}}(f_0) = \frac{|J_n(f_0)|}{n} + \frac{|M_n(r)|}{n} + \frac{\theta}{r}.$$

Since f_0 is a feasible policy, we have

$$\mathbb{P}\left(v_n^{\mathcal{X}} \geq v_{UC} + \eta\right) \leq \mathbb{P}\left(v_n^{\mathcal{X}}(f_0) \geq v_{UC} + \eta\right) \leq \mathbb{P}\left(\frac{|J_n(f_0)|}{n} + \frac{|M_n(r)|}{n} \geq v_{UC} + \gamma(r) + \frac{\eta}{3}\right).$$

We notice that

$$|J_n(f_0)| + |M_n(r)| = \sum_{i\in[n]} \left(\mathbb{1}_{\{f_0(\widehat{\boldsymbol{x}}^i) \neq \widehat{y}^i\}} + \mathbb{1}_{\{f_0(\widehat{\boldsymbol{x}}^i) = \widehat{y}^i, \widehat{\boldsymbol{x}}^i \in \cup_{y\in\mathcal{Y}}(U_y \setminus V_y(r))\}}\right).$$

By Lemma 3 and equation (11), we know

$$\mathbb{P}(f_0(\widehat{\boldsymbol{x}}^1) \neq \widehat{y}^1) = v_{UC},\ \mathbb{P}(f_0(\widehat{\boldsymbol{x}}^1) = \widehat{y}^1, \widehat{\boldsymbol{x}}^1 \in \cup_{y\in\mathcal{Y}}(U_y \setminus V_y(r))) \leq \mathbb{P}(\widehat{\boldsymbol{x}}^1 \in \cup_{y\in\mathcal{Y}}(U_y \setminus V_y(r))) = \gamma(r).$$

By Hoeffding's Inequality, we have

$$\mathbb{P}\left(\frac{1}{n}\sum_{i\in[n]} \left(\mathbb{1}_{\{f_0(\widehat{\boldsymbol{x}}^i) \neq \widehat{y}^i\}} + \mathbb{1}_{\{f_0(\widehat{\boldsymbol{x}}^i) = \widehat{y}^i, \widehat{\boldsymbol{x}}^i \in \cup_{y\in\mathcal{Y}}(U_y \setminus V_y(r))\}}\right) \geq v_{UC} + \gamma(r) + \frac{\eta}{3}\right) \leq \exp\left(-\frac{2n\eta^2}{9}\right)$$

Therefore, we have $\mathbb{P}(v_n^{\widehat{\mathcal{X}}_n} \geq v_{UC} + \eta) \leq \alpha$ whenever $n \geq \frac{-9\ln\alpha}{2\eta^2}$. This completes our proof. □

A.18 Proof of Proposition 7

Proposition 7 *Suppose that the feature space $\mathcal{X} = \mathbb{R}^d$, and the conditional random variable $(\boldsymbol{X}|Y=y)_{(\boldsymbol{X},Y)\sim\mathbb{P}_0}$ follows a multivariate normal distribution $\mathcal{N}(\boldsymbol{\mu}_y, \boldsymbol{\Sigma})$ for each $y \in \mathcal{Y}$, where the mean vectors $\{\boldsymbol{\mu}_y\}_{y\in\mathcal{Y}}$ are distinct. Denote the smallest eigenvalue of Σ by σ_d^2. Then we have*

(i) the quantity $\gamma(r) \leq \frac{m(m-1)r}{\sqrt{2\pi}\sigma_d}$;

(ii) the parameter θ in Proposition 6 can be chosen as $\theta = \frac{\eta^2\sqrt{2\pi}\sigma_d}{9m(m-1)}$; and

(iii) the least sample size n needed to guarantee that

$$\mathbb{P}(v_{UC} - \eta \leq v_n^{\widehat{\mathcal{X}}_n} \leq v_n^{\mathcal{X}} \leq v_{UC} + \eta) \geq 1 - 2\alpha,$$

is $n = \mathcal{O}(C^d d^{d/2}(d + \ln(\eta^{-1}))^{d/2}\eta^{-2d-2}m^{2d+1}\ln m + \eta^{-2}\ln(\alpha^{-1}))$, where C is a constant that does not depend on m, η, α and d.

Proof: Let g_y denote the continuous density function of the conditional distribution of the feature values for each label y. From the definition (9c) of the set U_y, it follows that

$$\mathrm{bd}^{\mathcal{X}}(U_y) = \mathrm{bd}(U_y) \subseteq \bigcup_{y' \in \mathcal{Y}, y' \neq y} \{\boldsymbol{x} \in \mathbb{R}^d : p_y g_y(\boldsymbol{x}) = p_{y'} g_{y'}(\boldsymbol{x})\}.$$

Therefore,

$$\begin{aligned} U_y \setminus V_y(r) \subseteq & \{\boldsymbol{x} \in \mathbb{R}^d : \mathrm{dist}(\boldsymbol{x}, \mathrm{bd}(U_y)) < r\} \\ \subseteq & \{\boldsymbol{x} \in \mathbb{R}^d : \exists y' \in \mathcal{Y}, y' \neq y, \mathrm{dist}(\boldsymbol{x}, \{\boldsymbol{x}' \in \mathbb{R}^d : p_y g_y(\boldsymbol{x}') = p_{y'} g_{y'}(\boldsymbol{x}')\}) < r\}. \end{aligned}$$

Taking the union over $y \in \mathcal{Y}$, we have

$$\bigcup_{y \in \mathcal{Y}} (U_y \setminus V_y(r)) \subseteq \bigcup_{y, y' \in \mathcal{Y}, y < y'} \{\boldsymbol{x} \in \mathbb{R}^d : \mathrm{dist}(\boldsymbol{x}, \{\boldsymbol{x}' \in \mathbb{R}^d : p_y g_y(\boldsymbol{x}') = p_{y'} g_{y'}(\boldsymbol{x}')\}) < r\}.$$

We denote $\{\boldsymbol{x} \in \mathbb{R}^d : \mathrm{dist}(\boldsymbol{x}, \{\boldsymbol{x}' \in \mathbb{R}^d : p_y g_y(\boldsymbol{x}') = p_{y'} g_{y'}(\boldsymbol{x}')\}) < r\}$ by $Q_{y,y'}(r)$. Then,

$$\gamma(r) = \mathbb{P}_0^{\mathcal{X}} \left(\bigcup_{y \in \mathcal{Y}} (U_y \setminus V_y(r)) \right) \leq \mathbb{P}_0^{\mathcal{X}} \left(\bigcup_{y, y' \in \mathcal{Y}, y < y'} Q_{y,y'}(r) \right).$$

For $y, y' \in \mathcal{Y}$ such that $y < y'$ we have

$$\begin{aligned} & p_y g_y(\boldsymbol{x}) = p_{y'} g_{y'}(\boldsymbol{x}) \\ \Leftrightarrow & \frac{p_y \exp\left(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_y)^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_y)\right)}{\sqrt{(2\pi)^d \det \boldsymbol{\Sigma}}} = \frac{p_{y'} \exp\left(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_{y'})^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_{y'})\right)}{\sqrt{(2\pi)^d \det \boldsymbol{\Sigma}}} \\ \Leftrightarrow & (\boldsymbol{x} - \boldsymbol{\mu}_y)^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_y) - (\boldsymbol{x} - \boldsymbol{\mu}_{y'})^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_{y'}) = 2 \ln \left(\frac{p_y}{p_{y'}} \right) \\ \Leftrightarrow & (\boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_{y'} - \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_y)^\top \boldsymbol{x} = \ln \left(\frac{p_y}{p_{y'}} \right) + \frac{\boldsymbol{\mu}_{y'}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_{y'} - \boldsymbol{\mu}_y^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_y}{2}. \end{aligned}$$

Let

$$\boldsymbol{h}_{y,y'} = \frac{1}{\|\boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_{y'} - \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_y\|} (\boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_{y'} - \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_y).$$

Then we have

$$p_y g_y(\boldsymbol{x}) = p_{y'} g_{y'}(\boldsymbol{x}) \Leftrightarrow \boldsymbol{h}_{y,y'}^\top \boldsymbol{x} = \frac{\ln \left(\frac{p_y}{p_{y'}} \right) + \frac{\boldsymbol{\mu}_{y'}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_{y'} - \boldsymbol{\mu}_y^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_y}{2}}{\|\boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_{y'} - \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_y\|}.$$

Denote the right-hand side by the constant $C_{y,y'}$, then we know that $\{\boldsymbol{x}' \in \mathbb{R}^d : p_y g_y(\boldsymbol{x}') = p_{y'} g_{y'}(\boldsymbol{x}')\}$ is exactly the hyperplane $\{\boldsymbol{x}' \in \mathbb{R}^d : \boldsymbol{h}_{y,y'}^\top \boldsymbol{x}' = C_{y,y'}\}$. Thus, we have

$$Q_{y,y'}(r) = \{\boldsymbol{x} \in \mathbb{R}^d : C_{y,y'} - r < \boldsymbol{h}_{y,y'}^\top \boldsymbol{x} < C_{y,y'} + r\}.$$

Since the marginal distribution $(\boldsymbol{X} | Y = y'')_{(\boldsymbol{X},Y) \sim \mathbb{P}_0}$ follows the multivariate normal distribution $\mathcal{N}(\boldsymbol{\mu}_{y''}, \boldsymbol{\Sigma})$, we have $(\boldsymbol{h}_{y,y'}^\top \boldsymbol{X} | Y = y'')_{(\boldsymbol{X},Y) \sim \mathbb{P}_0}$ follows the normal distribution $\mathcal{N}(\boldsymbol{h}_{y,y'}^\top \boldsymbol{\mu}_{y''}, \boldsymbol{h}_{y,y'}^\top \boldsymbol{\Sigma} \boldsymbol{h}_{y,y'})$. Therefore, we have

$$\mathbb{P}_{y''}\left(Q_{y,y'}(r)\right) = \mathbb{P}_{(\boldsymbol{X},Y) \sim \mathbb{P}_0}(\boldsymbol{X} \in Q_{y,y'}(r) | Y = y'') \leq \frac{2r}{\sqrt{2\pi \boldsymbol{h}_{y,y'}^\top \boldsymbol{\Sigma} \boldsymbol{h}_{y,y'}}} \leq \frac{2r}{\sqrt{2\pi} \sigma_d},$$

where the first inequality is because the Gaussian density is maximized at the mean, and the second one is due to the definition of σ_d. Taking expectation over y'', we obtain the probability bound as

$$\mathbb{P}_0^{\mathcal{X}}\left(Q_{y,y'}(r)\right) = \sum_{y''\in\mathcal{Y}} p_{y''}\mathbb{P}_{y''}\left(Q_{y,y'}(r)\right) \le \frac{2r}{\sqrt{2\pi}\sigma_d}.$$

Then we can bound $\gamma(r)$ by

$$\gamma(r) \le \sum_{y,y'\in\mathcal{Y},y<y'} \mathbb{P}_0^{\mathcal{X}}\left(Q_{y,y'}(r)\right) \le \frac{m(m-1)r}{\sqrt{2\pi}\sigma_d}$$

By Proposition 6, we may take r such that

$$\frac{m(m-1)r}{\sqrt{2\pi}\sigma_d} = \frac{\eta}{3},$$

and take $\theta = \frac{\eta r}{3}$. Thus, we choose θ to be

$$\theta = \frac{\eta^2\sqrt{2\pi}\sigma_d}{9m(m-1)}.$$

Under this setting, instead of considering hypercubes that intersect with $B_{\mathbf{0}}(R)$ as in Proposition 5, we consider hypercubes that intersect with any one of ellipsoids

$$E_y(R) = \{\boldsymbol{x}\in\mathbb{R}^d : (\boldsymbol{x}-\boldsymbol{\mu}_y)^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_y) \le R^2\},$$

where R makes

$$\mathbb{P}_{\boldsymbol{X}\sim\mathcal{N}(\mathbf{0},\boldsymbol{\Sigma})}(\boldsymbol{X}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{X} > R^2) \le \frac{\eta}{2}.$$

We know that when $\boldsymbol{X}\sim\mathcal{N}(\mathbf{0},\boldsymbol{\Sigma})$, $\boldsymbol{X}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{X}\sim\chi^2(d)$. By the Laurent-Massart bound (Lemma 1, Laurent and Massart (2000)), we may choose

$$R = \sqrt{d + 2\sqrt{d\ln(2\eta^{-1})} + 2\ln(2\eta^{-1})} \sim O\left(\sqrt{d+\ln(\eta^{-1})}\right).$$

The volume of $E_y(R)$ is then bounded by $(2R\sigma_d)^d = \mathcal{O}(C_1^d(d+\ln(\eta^{-1}))^{d/2})$, where C_1 is a constant that does not depend on m, η, α and d. Recall that we choose $\theta\sim\Theta(\eta^2m^{-2})$. Similar to the proof of Proposition 5, the number of hypercubes involved can be bounded by

$$m\frac{\mathcal{O}(C_1^d(d+\ln(\eta^{-1}))^{d/2})}{(\Theta(\eta^2m^{-2})/\sqrt{d})^d} = \mathcal{O}(C^d d^{d/2}(d+\ln(\eta^{-1}))^{d/2}\eta^{-2d}m^{2d+1}),$$

where C is a constant that does not depend on m, η, α and d. Next, following the proof of Proposition 5, we then obtain the least sample size n as

$$n = \mathcal{O}(C^d d^{d/2}(d+\ln(\eta^{-1}))^{d/2}\eta^{-2d-2}m^{2d+1}\ln m + \eta^{-2}\ln(\alpha^{-1})).$$

This completes the proof. □

A.19 Proof of Proposition 8

Proposition 8 *Suppose that for every $y \in \mathcal{Y}$, the underlying true distribution $(\boldsymbol{X}, Y) \sim \mathbb{P}_0$ conditioned on $Y = y$ is a uniform distribution supported on a hyper-rectangle $H_y = \prod_{j=1}^{d} [a_{y,j}, b_{y,j}]$. Suppose that there is a uniform bound $L \leq b_{y,j} - a_{y,j} \leq U$ on the side length for all $j \in [d]$ and $y \in \mathcal{Y}$. Then when $r \leq \frac{\ln 2}{2d} L$, we have*

(i) the quantity $\gamma(r) \leq \frac{8md}{L} r$;

(ii) the parameter θ in Proposition 6 can be chosen as

$$\theta = \frac{\eta^2 L}{72md};$$

(iii) the least sample size n needed to guarantee that

$$\mathbb{P}(v_{UC} - \eta \leq v_n^{\widehat{\mathcal{X}}_n} \leq v_n^{\mathcal{X}} \leq v_{UC} + \eta) \geq 1 - 2\alpha,$$

is $n = \mathcal{O}(C^d d^{3d/2} \eta^{-2d-2} m^{d+1} \ln m + \eta^{-2} \ln(\alpha^{-1}))$, where C is a constant that does not depend on m, η, α, and d.

Proof: Without loss of generality, we assume that volumes of hyper-rectangles $\{H_y\}_{y \in \mathcal{Y}}$ are non-decreasing concerning y. Then, since the conditional distributions are uniform, the conditional densities are non-increasing. Then for any $y \in \mathcal{Y}$, by definition (9c), we have $U_y = H_y \setminus (\cup_{y' \in [y-1]} H_{y'})$. Therefore, we must have that

$$U_y \setminus V_y(r) \subseteq \bigcup_{y' \in [y]} \{\boldsymbol{x} \in U_y : \operatorname{dist}(\boldsymbol{x}, \operatorname{bd}(H_{y'})) < r\}.$$

Hence, we have

$$\bigcup_{y \in \mathcal{Y}} (U_y \setminus V_y(r)) \subseteq \bigcup_{y \in \mathcal{Y}} \left\{ \boldsymbol{x} \in \bigcup_{y' \in \mathcal{Y} \setminus [y-1]} U_{y'} : \operatorname{dist}(\boldsymbol{x}, \operatorname{bd}(H_y)) < r \right\}. \tag{22a}$$

Note that for any $y \in \mathcal{Y}$, we have

$$\left\{ \boldsymbol{x} \in \bigcup_{y' \in \mathcal{Y} \setminus [y-1]} U_{y'} : \operatorname{dist}(\boldsymbol{x}, \operatorname{bd}(H_y)) < r \right\} \subseteq \prod_{j \in [d]} (a_{y,j} - r, b_{y,j} + r) \setminus \prod_{j \in [d]} [a_{y,j} + r, b_{y,j} - r],$$

Here we note that $r \leq \frac{\ln 2}{2d} L < \frac{1}{2} L$, hence $a_{y,j} + r < b_{y,j} - r$. Denote $b_{y,j} - a_{y,j}$ by $z_{y,j}$, then the volume of the right-hand side is:

$$\prod_{j \in [d]} (z_{y,j} + 2r) - \prod_{j \in [d]} (z_{y,j} - 2r).$$

It is noted that on the set $\cup_{y' \in \mathcal{Y} \setminus [y-1]} U_{y'}$, the density corresponding to the probability measure $\mathbb{P}_0^{\mathcal{X}}$ must be less than the conditional density of y, i.e., $(\prod_{j \in [d]} z_{y,j})^{-1}$. The reason is that on this set

(i.e., $\cup_{y'\in\mathcal{Y}\setminus[y-1]}U_{y'}$), all labels that appear have a conditional density less than that of the label y. This leads to

$$\begin{aligned}&\mathbb{P}_0^{\mathcal{X}}\left(\left\{\boldsymbol{x}\in\bigcup_{y'\in\mathcal{Y}\setminus[y-1]}U_{y'}:\operatorname{dist}(\boldsymbol{x},\operatorname{bd}(H_y))<r\right\}\right)\\ \leq&\frac{1}{\prod_{j\in[d]}z_{y,j}}\left[\prod_{j\in[d]}(z_{y,j}+2r)-\prod_{j\in[d]}(z_{y,j}-2r)\right]=\prod_{j\in[d]}\left(1+\frac{2r}{z_{y,j}}\right)-\prod_{j\in[d]}\left(1-\frac{2r}{z_{y,j}}\right)\end{aligned}\tag{22b}$$

We see that the right-hand side above is monotone decreasing in $z_{y,j}$ for $j\in[d]$. Therefore, we have

$$\prod_{j\in[d]}\left(1+\frac{2r}{z_{y,j}}\right)-\prod_{j\in[d]}\left(1-\frac{2r}{z_{y,j}}\right)\leq\left(1+\frac{2r}{L}\right)^d-\left(1-\frac{2r}{L}\right)^d.\tag{22c}$$

Since $r\leq\frac{\ln 2}{2d}L$, we may convert the right-hand side to integration and have

$$\begin{aligned}&\left(1+\frac{2r}{L}\right)^d-\left(1-\frac{2r}{L}\right)^d=\int_{-r}^{r}\frac{2d}{L}\left(1+\frac{2x}{L}\right)^{d-1}\mathrm{d}x\leq\int_{-r}^{r}\frac{2d}{L}\left(1+\frac{\ln 2}{d}\right)^{d-1}\mathrm{d}x\\ \leq&\int_{-r}^{r}\frac{2d}{L}\left(1+\frac{\ln 2}{d}\right)^{d-1}\mathrm{d}x=\int_{-r}^{r}\frac{2d}{L}\exp\left((d-1)\ln\left(1+\frac{\ln 2}{d}\right)\right)\mathrm{d}x\end{aligned}\tag{22d}$$

We know that $ln(1+t)\leq t$ for any $t>-1$. So we may further derive

$$\int_{-r}^{r}\frac{2d}{L}\exp\left((d-1)\ln\left(1+\frac{\ln 2}{d}\right)\right)\mathrm{d}x\leq\int_{-r}^{r}\frac{2d}{L}\exp\left(d\cdot\frac{\ln 2}{d}\right)\mathrm{d}x=\int_{-r}^{r}\frac{4d}{L}\mathrm{d}x=\frac{8dr}{L}.\tag{22e}$$

By the definition (11), the inclusion (22a) and the inequalities (22b)-(22e), we have

$$\gamma(r)=\mathbb{P}_0^{\mathcal{X}}\left(\bigcup_{y\in\mathcal{Y}}(U_y\setminus V_y(r))\right)\leq\sum_{y\in\mathcal{Y}}\mathbb{P}_0^{\mathcal{X}}\left(\left\{\boldsymbol{x}\in\bigcup_{y'\in\mathcal{Y}\setminus[y-1]}U_{y'}:\operatorname{dist}(\boldsymbol{x},\operatorname{bd}(H_y))<r\right\}\right)\leq\frac{8md}{L}r.$$

By Proposition 6, we may take r such that

$$\gamma(r)\leq\frac{\eta}{3}.$$

To achieve this, we choose

$$r=\min\left\{\frac{\eta L}{24md},\frac{L\ln 2}{2d}\right\}=\frac{\eta L}{24md}.$$

Here, the second equation is because $\eta<1$ and $m\geq 2$. Thus, we choose θ to be

$$\theta=\frac{\eta r}{3}=\frac{\eta^2 L}{72md}.$$

Under this setting, the feature space is bounded. Following Proposition 5, we consider all hypercubes that intersect with the feature space, which is the union of the hyper-rectangles H_y. Similar to the proof of Proposition 5, the number of hypercubes involved can be bounded by

$$m\frac{U^d}{(\Theta(\eta^2m^{-1}d^{-1})/\sqrt{d})^d}=\mathcal{O}(C^d d^{3d/2}\eta^{-2d}m^{d+1}),$$

where C is a constant that does not depend on m, η, α and d. Similar to the proof of Proposition 5, we then obtain the least sample size n as

$$n = \mathcal{O}(C^d d^{3d/2} \eta^{-2d-2} m^{d+1} \ln m + \eta^{-2} \ln(\alpha^{-1})).$$

This completes the proof. □

A.20 Proof of Theorem 5

Theorem 5 *Given any $\kappa \in \mathbb{R}_{++}, \theta \in \mathbb{R}_{++}$, $n \in \mathbb{Z}^+$, and an in-sample policy $\widehat{f} \in \widehat{\mathcal{F}}_n$, for each $i \in [n]$ and $y \in [m]$, define*

$$w_{iy} = \begin{cases} 1, & \text{if } \widehat{f}(\widehat{\boldsymbol{x}}^i) = y; \\ 0, & \text{if } \widehat{f}(\widehat{\boldsymbol{x}}^i) \neq y. \end{cases}$$

Then, the in-sample DRUC value $\widehat{v}_n(\widehat{f})$ can be computed by solving the following linear program:

$$\begin{array}{rll} \widehat{v}_n(\widehat{f}) = \max & \displaystyle\sum_{i,j\in[n]} (1 - w_{j\widehat{y}^i}) r_{ij} + \sum_{i\in[n]} r_{i0} & \\ \text{s.t.} & \displaystyle\sum_{i\in[n]} \kappa r_{i0} + \sum_{i,j\in[n]} d_{ij} r_{ij} \leq \theta, & \\ & r_{i0} + \displaystyle\sum_{j\in[n]} r_{ij} \leq \frac{1}{n}, & \forall i \in [n], \\ & r_{i0} \geq 0, r_{ij} \geq 0, & \forall i, j \in [n]. \end{array} \tag{12}$$

Proof: Recall the definition (6b) of $s_i^n(\widehat{f})$ for every $i \in [n]$:

$$s_i^n(\widehat{f}) = \min\{\kappa, \operatorname{dist}(\widehat{\boldsymbol{x}}^i, \widehat{f}^{-1}(\mathcal{Y} \setminus \{\widehat{y}^i\}))\},$$

which implies that either there exists $j \in [n]$ such that $\widehat{f}(\widehat{\boldsymbol{x}}^j) \neq \widehat{y}^i$ (i.e., $w_{j\widehat{y}^i} = 0$) and $s_i^n(\widehat{f}) = d_{ij}$; or $s_i^n(\widehat{f}) = \kappa$. We can then define an index $a_i \in [n] \cup \{0\}$ for each $i \in [n]$:

$$a_i = \begin{cases} 0, & \text{if } s_i^n(\widehat{f}) = \kappa; \\ j, & \text{if } s_i^n(\widehat{f}) = d_{ij} < \kappa, w_{j\widehat{y}^i} = 0. \end{cases} \tag{23a}$$

Now we consider a feasible solution $(r^*_{ij})_{i\in[n], j\in[n]\cup\{0\}}$ of (12) such that for all $i \in [n]$, $r^*_{ij} = 0$ for $j \neq a_i$. For this solution, the constraints of (12) are then reduced to:

$$\begin{cases} \displaystyle\sum_{i\in[n]} r^*_{ia_i} s_i^n(\widehat{f}) \leq \theta; \\ 0 \leq r^*_{ia_i} \leq \dfrac{1}{n}, & \forall i \in [n]. \end{cases} \tag{23b}$$

Next, we discuss two cases in which we define the values $\{r^*_{ia_i}\}_{i\in[n]}$ and prove that they form an optimal solution to the linear program (12), achieving the objective value $v_n^{\widehat{\mathcal{X}}_n}(\widehat{f})$. By Theorem 1, there are two cases:

(i) If $\sum_{i\in[n]} s_i^n(\widehat{f}) < n\theta$, we have $v_n^{\widehat{\mathcal{X}}_n}(\widehat{f}) = 1$. We then simply take $r^*_{ia_i} = \frac{1}{n}$ for all $i \in [n]$. This solution satisfies (23b), so it is feasible. It has objective 1:

$$\sum_{i,j\in[n]} (1 - w_{j\widehat{y}^i}) r^*_{ij} + \sum_{i\in[n]} r^*_{i0} = \sum_{i\in[n]} r^*_{ia_i} = 1.$$

Here the first equation comes from the fact that $r^*_{ij} = 0$ for $j \neq a_i$ and our definition (23a) of a_i: for $a_i > 0$, $w_{a_i\widehat{y}^i} = 0$.

Moreover, for any feasible solution $(r_{ij})_{i\in[n],j\in[n]\cup\{0\}}$, the objective value must satisfy

$$\sum_{i,j\in[n]} (1 - w_{j\widehat{y}^i}) r_{ij} + \sum_{i\in[n]} r_{i0} \leq \sum_{i\in[n]} \left(r_{i0} + \sum_{j\in[n]} r_{ij} \right) \leq 1.$$

Thus, the optimal value is equal to 1 in this case, which equals $v_n^{\widehat{\mathcal{X}}_n}(\widehat{f})$.

(ii) If $\sum_{i\in[n]} s_i^n(\widehat{f}) \geq n\theta$, we prove the result by giving a pair of primal and dual feasible solutions, which satisfy complementary slackness and have objective value that equals to $v_n^{\widehat{\mathcal{X}}_n}(\widehat{f})$. By part (ii) of Theorem 1, we have

$$\sum_{i\in[\lceil k_n(\widehat{f},\theta)\rceil - 1]} \frac{1}{n} s^n_{(i)}(\widehat{f}) + \frac{1}{n} \left(k_n(\widehat{f},\theta) + 1 - \lceil k_n(\widehat{f},\theta)\rceil \right) s^n_{(\lceil k_n(\widehat{f},\theta)\rceil)}(\widehat{f}) = \theta. \tag{23c}$$

We let $r^*_{ia_i}$ be the corresponding (in the original index) coefficients of $s_i^n(\widehat{f})$ in (23c). Then we know that $(r^*_{ij})_{i\in[n],j\in[n]\cup\{0\}}$ is a primal feasible solution, since it satisfies the equivalent constraint (23b). It has an objective

$$\sum_{i,j\in[n]} (1 - w_{j\widehat{y}^i}) r^*_{ij} + \sum_{i\in[n]} r^*_{i0} = \sum_{i\in[n]} r^*_{ia_i} = \frac{k_n(\widehat{f},\theta)}{n} = v_n^{\widehat{\mathcal{X}}_n}(\widehat{f}),$$

where the first equality follows from the fact that $r^*_{ij} = 0$ for $j \neq a_i$ and our definition (23a) of a_i: for $a_i > 0$, $w_{a_i\widehat{y}^i} = 0$.

On the other hand, we note that the dual problem of (12) is:

$$\begin{array}{ll} \min & \theta\alpha + \displaystyle\sum_{i\in[n]} \frac{1}{n}\lambda_i \\ \text{s.t.} & d_{ij}\alpha + \lambda_i \geq 1 - w_{j\widehat{y}^i}, \ \forall i,j \in [n]; \\ & \kappa\alpha + \lambda_i \geq 1, \qquad\qquad \forall i \in [n]; \\ & \alpha \geq 0; \\ & \lambda_i \geq 0, \qquad\qquad \forall i \in [n]. \end{array} \tag{23d}$$

We construct a dual solution:

$$\alpha^* = \frac{1}{s^n_{(\lceil k_n(\widehat{f},\theta)\rceil)}(\widehat{f})}, \ \lambda_i^* = \left(1 - \frac{s_i^n(\widehat{f})}{s^n_{(\lceil k_n(\widehat{f},\theta)\rceil)}(\widehat{f})} \right)^+, \forall i \in [n]. \tag{23e}$$

By the definition

$$s_i^n(\widehat{f}) = \min\{\kappa, \mathrm{dist}(\widehat{\boldsymbol{x}}^i, \widehat{f}^{-1}(\mathcal{Y} \setminus \{\widehat{y}^i\}))\},$$

we know that $s_i^n(\widehat{f}) \le d_{ij}$ for any $\widehat{f}(\widehat{\boldsymbol{x}}^j) \neq \widehat{y}^i$ (or equivalently, $w_{j\widehat{y}^i} = 0$) and $s_i^n(\widehat{f}) \le \kappa$. Thus, for $w_{j\widehat{y}^i} = 0$, we have

$$d_{ij}\alpha^* + \lambda_i^* \ge \frac{s_i^n(\widehat{f})}{s^n_{(\lceil k_n(\widehat{f},\theta)\rceil)}(\widehat{f})} + \left(1 - \frac{s_i^n(\widehat{f})}{s^n_{(\lceil k_n(\widehat{f},\theta)\rceil)}(\widehat{f})}\right) = 1.$$

And similarly,

$$\kappa\alpha^* + \lambda_i^* \ge \frac{s_i^n(\widehat{f})}{s^n_{(\lceil k_n(\widehat{f},\theta)\rceil)}(\widehat{f})} + \left(1 - \frac{s_i^n(\widehat{f})}{s^n_{(\lceil k_n(\widehat{f},\theta)\rceil)}(\widehat{f})}\right) = 1.$$

The first constraint in (23d) is trivial when $w_{j\widehat{y}^i} = 1$. So the solution we propose is dual feasible.

Now we prove that complementary slackness holds for $(r_{ij}^*)_{i\in[n],j\in[n]\cup\{0\}}, \alpha^*, (\lambda_i^*)_{i\in[n]}$. In (23c), note that only $i \in [n]$ that makes $s_i^n(\widehat{f}) \le s^n_{(\lceil k_n(\widehat{f},\theta)\rceil)}(\widehat{f})$ can have nonzero coefficient. Therefore, by the definition of a_i, any nonzero $r^*_{ia_i}$ with $a_i > 0$ must satisfy $w_{a_i\widehat{y}^i} = 0$, $d_{ia_i} = s_i^n(\widehat{f})$ and $s_i^n(\widehat{f}) \le s^n_{(\lceil k_n(\widehat{f},\theta)\rceil)}(\widehat{f})$, so

$$d_{ia_i}\alpha^* + \lambda_i^* = \frac{d_{ia_i}}{s^n_{(\lceil k_n(\widehat{f},\theta)\rceil)}(\widehat{f})} + 1 - \frac{s_i^n(\widehat{f})}{s^n_{(\lceil k_n(\widehat{f},\theta)\rceil)}(\widehat{f})} = 1 = 1 - w_{a_i\widehat{y}^i}.$$

That is, the dual constraints corresponding to the primal nonzero entry $r^*_{ia_i}$ are binding. Similarly, the constraint corresponding to nonzero entry $r^*_{ia_i}$ when $a_i = 0$ is binding:

$$\kappa\alpha^* + \lambda_i^* = 1.$$

Additionally, the first constraint of (12) must be binding under this case by (23c). For $\lambda_i^* > 0$, by the definition (23e), we must have $s_i^n(\widehat{f}) < s^n_{(\lceil k_n(\widehat{f},\theta)\rceil)}(\widehat{f})$. Thus, its coefficient is $\frac{1}{n}$ in the summation on the left-hand side of (23c); i.e., $r^*_{ia_i} = \frac{1}{n}$. This implies that the primal constraint corresponding to λ_i^*:

$$r_{i0}^* + \sum_{j\in[n]} r_{ij}^* \le \frac{1}{n},$$

is binding. Therefore, we have complementary slackness, which allows us to determine the optimality of the pair of solutions.

Combining these two cases, the optimal value of (12) is exactly $v_n^{\widehat{\mathcal{X}}_n}(\widehat{f})$. □

A.21 Proof of Corollary 4

Corollary 4 *The optimal in-sample DRUC value $v_n^{\widehat{\mathcal{X}}_n}$ is equal to the optimal value of the following MILP formulation:*

$$\begin{array}{lll} \min\ \theta\alpha + \displaystyle\sum_{i\in[n]} \frac{1}{n}\lambda_i & & \\ \text{s.t.} & d_{ij}\alpha + \lambda_i + w_{j\widehat{y}^i} \geq 1, & \forall i,j \in [n], \\ & \kappa\alpha + \lambda_i \geq 1, & \forall i \in [n], \\ & \sum_{y\in[m]} w_{iy} = 1, & \forall i \in [n], \\ & \alpha \geq 0, \lambda_i \geq 0, w_{iy} \in \{0,1\}, & \forall i \in [n], y \in [m]. \end{array} \tag{13}$$

Proof: The dual problem of (12) is:

$$\begin{array}{lll} v_n^{\widehat{\mathcal{X}}_n}(\widehat{f}) = \min\ \theta\alpha + \displaystyle\sum_{i\in[n]} \frac{1}{n}\lambda_i & & \\ \text{s.t.} & d_{ij}\alpha + \lambda_i \geq 1 - w_{j\widehat{y}^i}, & \forall i,j \in [n], \\ & \kappa\alpha + \lambda_i \geq 1, & \forall i \in [n], \\ & \alpha \geq 0, \lambda_i \geq 0, & \forall i \in [n]. \end{array}$$

To obtain $v_n^{\widehat{\mathcal{X}}_n}$, we take the minimum of $v_n^{\widehat{\mathcal{X}}_n}(\widehat{f})$ over all $\widehat{f} \in \widehat{\mathcal{F}}_n$, which corresponds to minimizing over all possible assignment values of $\{w_{iy}\}_{i\in[n],y\in[m]}$. Therefore, by considering all possible values of $\{w_{iy}\}_{i\in[n],y\in[m]}$, we conclude that the optimal value of the MILP (13) is exactly $v_n^{\widehat{\mathcal{X}}_n}$. □

A.22 Proof of Theorem 6

Theorem 6 *Let $\{\widetilde{w}_{iy}\}_{i\in[n],y\in[m]}$ be the solution obtained through MaxLin, and its corresponding objective value in* (13) *be $\widetilde{v}_n$. Then we have $v_n^{\widehat{\mathcal{X}}_n} \leq \widetilde{v}_n \leq 2v_n^{\widehat{\mathcal{X}}_n}$.*

Proof: It is clear that $v_n^{\widehat{\mathcal{X}}_n} \leq \widetilde{v}_n$ since $v_n^{\widehat{\mathcal{X}}_n}$ is the optimal value. Let $\alpha^*, \{\lambda_i^*\}_{i\in[n]}, \{w_{iy}^*\}_{i\in[n],y\in[m]}$ denote an optimal LP relaxation solution. Let us construct a solution to the original problem as $\widetilde{\alpha} = 2\alpha^*, \widetilde{\lambda}_i = 2\lambda_i^*$ for all $i \in [n]$, and $\{\widetilde{w}_{iy}\}_{i\in[n],y\in[m]}$ defined in (14). Note that for $y \neq \min(\arg\max_{y'\in[m]} w_{iy'}^*)$, we must have $w_{iy}^* \leq \frac{1}{2}$.

We next check the feasibility of $\widetilde{\alpha}, \{\widetilde{\lambda}_i\}_{i\in[n]}, \{\widetilde{w}_{iy}\}_{i\in[n],y\in[m]}$ of the formulation (13). If $\widetilde{w}_{j,\widehat{y}^i} = 0$ for some $i \in [n]$ and $j \in [n]$, we know that $w_{j\widehat{y}^i}^* \leq \frac{1}{2}$ by our definition of $\widetilde{w}_{iy}$. Therefore,

$$d_{ij}\widetilde{\alpha} + \widetilde{\lambda}_i + \widetilde{w}_{j,\widehat{y}^i} = 2(d_{ij}\alpha + \lambda_i) \geq 2(1 - w_{j\widehat{y}^i}) \geq 2 \times \frac{1}{2} = 1 = (1 - \widetilde{w}_{j\widehat{y}^i}).$$

All remaining constraints in (13) (including the first constraint with $\widetilde{w}_{j,\widehat{y}^i} = 1$) are satisfied for the constructed solution. Therefore, it is a feasible solution of the MILP (13) that has twice the objective value as that of the LP relaxation. This argument applies to any solution of the LP relaxation, including the optimal one. We also know that the optimal value of the LP relaxation must be less than or equal to the MILP optimal value $v_n^{\widehat{\mathcal{X}}_n}$. Therefore, we have $\widetilde{v}_n \leq 2v_n^{\widehat{\mathcal{X}}_n}$. □