A PRELIMINARIES

A.1 BACKGROUND ON ALGORITHM RECOURSE

Suppose $f : \mathcal{X} \to \mathcal{Y}$ is a classifier that maps features $\boldsymbol{x} \in \mathcal{X} \subset \mathbb{R}^d$ to labels $\mathcal{Y} = \{0, 1\}$, where 0 is the unfavorable outcome and 1 is the favorable outcome. Define $f(\boldsymbol{x}) = g(h(\boldsymbol{x}))$, where h is the scoring function and the activation function is $g(t) = \boldsymbol{I}_{t \geq \eta}$. For ease of illustration, we adopt the setting of loan approval as an example, i.e., $h(\boldsymbol{x}) \geq \eta$ denotes that a loan is granted and $h(\boldsymbol{x}) < \eta$ denotes that it is denied. For an individual $\boldsymbol{x}_0$ that was denied by the loan-granting institution, counterfactual explanation methods could provide the individual with a recourse by identifying which attributes to change for reversing the unfavorable prediction result. Given a cost function $c : \mathbb{R}^d \to \mathbb{R}_+$, the counterfactual explanation $\boldsymbol{x}^{\text{CF}}$ can be found by solving Wachter et al. (Nov. 2017); Ustun et al. (Jan. 2019)

$$\min_{\boldsymbol{x}' \in \mathcal{A}} l(f(\boldsymbol{x}'), 1) + \lambda c(\boldsymbol{x}, \boldsymbol{x}'), \tag{3}$$

where $\mathcal{A}$ is the set of actionable counterfactuals, λ is the trade-off parameter, and l is the loss for invalid recourse. The first term in the objective function guarantees that the prediction result of the counterfactual $\boldsymbol{x}'$ is close to the favorable outcome 1. The second term in the objective function encourages the recourse to have lower cost.

A.2 BACKGROUND ON CONFORMAL PREDICTIVE INFERENCE

Conformal inference framework provides a generic methodology for transforming the outputs of any black box prediction algorithm into a prediction set Gibbs & Candes (Dec. 2021). The algorithms from conformal inference provide a prediction set that has valid marginal coverage $\mathbb{P}(Y_i \in \hat{C}(\boldsymbol{X}_i)) \geq 1 - \alpha$ based on standard properties of quantiles, if the training and test data are exchangeable Cauchois et al. (Aug. 2020); Gibbs & Candes (Dec. 2021).

To produce the prediction set, a conformal predictor uses a nonconformity function, an arbitrary function $s : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, that measures the strangeness of a sample $(\boldsymbol{x}, y)$ Johansson et al. (May 2017). Based on the nonconformity scores of examples with known output labels, and the nonconformity score of a tentatively labels test pattern $(\boldsymbol{x}_{n+1}, \tilde{y})$, a p-value statistic can be calculated to reject the hypothesis that $\tilde{y}$ corresponds with the true label y_{n+1}. Then all labels $\tilde{y} \subset \mathcal{Y}$ that are not rejected at the chosen significance level α constitute the final prediction set, which contains the true label y_{n+1} with a probability of $1 - \alpha$.

In particular, for a given confidence level $(1 - \alpha)$, one can define a confidence set $\hat{C}(\boldsymbol{x})$ based on the validation set $\mathcal{D}_{\text{val}} = \{(\boldsymbol{X}_i, Y_i)\}_{i=1}^n$, i.e.

$$\hat{C}(\boldsymbol{x}) = \{y \in \mathcal{Y} | s(\boldsymbol{x}, y) \leq \hat{\mathcal{Q}}_{n,1-\alpha}\}, \tag{4}$$

where

$$\hat{\mathcal{Q}}_{n,1-\alpha} = \text{Quantile}\left(\left(1 + \frac{1}{n}\right)\alpha; \{s(\boldsymbol{X}_i, Y_i)\}_{i=1}^n\right).$$

Then as long as $\{(\boldsymbol{X}_i, Y_i)\}_{i=1}^{n+1}$ are exchangeable, the confidence set $\hat{C}(\boldsymbol{X}_{n+1})$ satisfies Romano et al. (Dec. 2019)

$$\mathbb{P}(Y_{n+1} \in \hat{C}(\boldsymbol{X}_{n+1})) \geq 1 - \alpha. \tag{5}$$

In the algorithmic recourse scenario, we view the counterfactual sample $(\boldsymbol{x}^{\text{CF}}, y^{\text{CF}})$ as the $(n+1)$-th test sample. Then the nonconformity score $s(\boldsymbol{x}^{\text{CF}}, y^{\text{CF}})$ measures the degree of nonconformity between the counterfactual sample and samples in $\mathcal{D}$. Different from the above-mentioned inference problem, we do not have a prescribed value of α, but have some observed properties on the value of $s(\boldsymbol{x}^{\text{CF}}, y^{\text{CF}})$. Thus, by transforming equation 4, equation 5 and applying them to $(\boldsymbol{x}^{\text{CF}}, y^{\text{CF}})$, we have

$$\mathbb{P}(s(\boldsymbol{x}^{\text{CF}}, y^{\text{CF}}) \leq \hat{\mathcal{Q}}_{n,1-\alpha}) \geq 1 - \alpha, \tag{6}$$

where the value of α can be derived based on the known properties of $s(\boldsymbol{x}^{\text{CF}}, y^{\text{CF}})$. Moreover, equation 6 provides a probability inequality on the value of $s(\boldsymbol{x}^{\text{CF}}, y^{\text{CF}})$, which is useful in measuring the robustness of $\boldsymbol{x}^{\text{CF}}$.

However, the above mentioned results are limited by the exchangeable data assumption. Recently, there are works extending the conformal inference beyond the case of exchangeable data. In particular, a weighted version of conformal inference has been proposed to compute distribution-free prediction intervals for problems in which the test and training covariant distributions differ, but the likelihood ratio between the two distributions is known Tibshirani et al. (Dec. 2019).

Assume that $\{(\boldsymbol{X}_i, Y_i)\}_{i=1}^n \overset{i.i.d.}{\sim} \mathcal{P}$ and the independent test sample $(\boldsymbol{X}_{n+1}, Y_{n+1}) \sim \mathcal{P}'$. Then the likelihood ratio between $\mathcal{P}$ and $\mathcal{P}'$ is defined as

$$v(\boldsymbol{x}, y) = \frac{d\mathcal{P}'}{d\mathcal{P}}(\boldsymbol{x}, y), \tag{7}$$

and $v(\boldsymbol{x}, y)$ is assumed to be known exactly in Tibshirani et al. (Dec. 2019). For any new data sample $(\boldsymbol{x}, y) \in \mathcal{X} \times \mathcal{Y}$ (e.g. the generated counterfactual sample $(\boldsymbol{x}^{\text{CF}}, y^{\text{CF}})$), assign weights to the sample as

$$\begin{aligned} p_i(\boldsymbol{x}, y) &= \frac{v(\boldsymbol{X}_i, Y_i)}{\sum_{j=1}^n v(\boldsymbol{X}_j, Y_j) + v(\boldsymbol{x}, y)}, i = 1, 2, \cdots, n, \\ p_{n+1}(\boldsymbol{x}, y) &= \frac{v(\boldsymbol{x}, y)}{\sum_{j=1}^n v(\boldsymbol{X}_j, Y_j) + v(\boldsymbol{x}, y)}. \end{aligned} \tag{8}$$

Then we have

$$\mathbb{P}'(Y_{n+1} \in \hat{D}(\boldsymbol{X}_{n+1})) \geq 1 - \alpha,$$

where the prediction interval $\hat{D}(\boldsymbol{X}_{n+1})$ is given by

$$\hat{D}(\boldsymbol{X}_{n+1}) = \{y : s(\boldsymbol{X}_{n+1}, y) \leq \hat{S}_{1-\alpha}(y)\},$$

with

$$\hat{S}_{1-\alpha}(y) = \text{Quantile}\left(1 - \alpha, \sum_{i=1}^n p_i(\boldsymbol{X}_{n+1}, y)\delta_{v_{n+1}(S_i)} + p_{n+1}(\boldsymbol{X}_{n+1}, y)\delta_\infty\right),$$

and δ denotes the point mass.

In the algorithm recourse scenario, since the distribution of nonconformity score variable s changes, we can leverage the concept of weighted conformal inference by assigning weights to the validation samples based on their similarity to the counterfactual example to provide probability bounds on the nonconformity score of $(\boldsymbol{x}^{\text{CF}}, y^{\text{CF}})$.

B ADDITIONAL ALGORITHMS

B.1 ALGORITHM FOR FINDING k^* GIVEN A TARGET LEVEL α

Algorithm 3 Procedure of finding k^* given a target level α

Input: training data $\mathcal{D}_{\text{train}}$, calibration data $\mathcal{D}_{\text{calib}}$, recourse $\boldsymbol{x}^{\text{CF}}$, nonconformity score function $s(\cdot,\cdot)$, model shift parameter τ, target level $\alpha \in (0,1)$.

1: Use $\mathcal{D}_{\text{train}}$ to construct the point-wise bounds $\hat{L}(\cdot)$ and $\hat{U}(\cdot)$ for likelihood ratio v.
2: For samples in $\mathcal{D}_{\text{calib}}$, compute $S_i = s(\boldsymbol{X}_i, Y_i), L_i = \hat{L}(S_i), U_i = \hat{U}(S_i)$.
3: For the recourse $\boldsymbol{x}^{\text{CF}}$, compute L^{CF} and U^{CF}.
4: For $1 \le k \le n$, compute $\hat{F}(k)$.
5: Derive $k^* = \min\{k : \hat{F}(k) \ge 1-\alpha\}$.

Output: The value of k^*.

B.2 RECOURSES WITH VARIOUS CHOICES OF λ

In this subsection, we provide an additional algorithm for finding recourses with different robustness levels. The performance of this algorithm can be found in Figure 5.

In state-of-the-art recourse algorithms Wachter et al. (Nov. 2017); Ustun et al. (Jan. 2019); Karimi et al. (Aug. 2020), the low-cost recourse for an adversely predicted sample $\boldsymbol{x}_0$ is found by solving

$$\boldsymbol{x}^{\text{CF}} = \arg\min_{\boldsymbol{x}\in\mathcal{A}} [l(f(\boldsymbol{x}), 1) + \lambda c(\boldsymbol{x}_0, \boldsymbol{x})], \tag{9}$$

where the trade-off parameter λ is considered given. However, we note that as the value of λ changes, the generated recourse $\boldsymbol{x}^{\text{CF}}$ varies, and the recourse cost as well as the recourse invalidation rate also change accordingly. Thus, a natural way to find different recourses with different robustness levels is to vary the value of λ. For each choice of λ (e.g. $\lambda = \lambda_j$), we can generate a recourse $\boldsymbol{x}_j^{\text{CF}}$ by any recourse generating algorithm and derive the corresponding recourse cost $c_j = c(\boldsymbol{x}_0, \boldsymbol{x}_j^{\text{CF}})$. Based on Theorem 12 or Theorem 13, we are able to derive the upper-bound $r_{u,j} = r_u(\boldsymbol{x}_j^{\text{CF}})$ on the recourse invalidation rate, which measures the robustness of $\boldsymbol{x}_j^{\text{CF}}$ to model changes. Then all generated recourses and their corresponding recourse costs as well as robustness metrics (bounds on the recourse invalidation rates), i.e. $\{(\boldsymbol{x}_j^{\text{CF}}, c_j, r_{u,j})\}_j$, can be provided to users. We summarize the procedure in Algorithm 4.

Algorithm 4 Recourses with various choices of λ

Input: negatively predicted sample $\boldsymbol{x}_0$, current model f, model shift parameter τ, maximum trade-off parameter λ_m, increment parameter d_λ of λ.

1: $j = 1$
2: **for** $\lambda = 0 : d_\lambda : \lambda_m$ **do**
3: solve equation 9 by a recourse generating algorithm and the generated recourse is $\boldsymbol{x}_j^{\text{CF}}$;
4: calculate the recourse cost $c_j = c(\boldsymbol{x}_0, \boldsymbol{x}_j^{\text{CF}})$;
5: derive the upper-bound $r_{u,j} = r_u(\boldsymbol{x}_j^{\text{CF}})$ (according to Theorems 12 or 13) on the invalidation rate;
6: $j = j + 1$;
7: **end for**

Output: $\{(\boldsymbol{x}_j^{\text{CF}}, c_j, r_{u,j})\}_j$.

C ADDITIONAL DETAILS ON NUMERICAL RESULTS

All experiments were run on a 2.8 GHz Quad-Core Intel Core i7.

C.1 DETAILS ABOUT THE DATASETS

We conduct our analysis using three real datasets: Criminal justice dataset Lakkaraju et al. (Aug. 2016), Student performance dataset Amrieh et al. (Aug. 2016) and German credit dataset Dua et al. (2017). Each dataset contains two parts, initial data (D_1) and shifted data (D_2).

1. Criminal justice dataset Lakkaraju et al. (Aug. 2016): It contains proprietary data from 1978 (D_1) and 1980 (D_2), with 8395 and 8595 samples, respectively. It includes demographic features such as race, sex, age, time-served, and employment, and a target attribute related to bail decisions. Furthermore, the dataset exhibits an inherent temporal shift, as the data characteristics in 1980 differ from those in 1978.
2. Student performance dataset Amrieh et al. (Aug. 2016): It comprises publicly available data collected from schools in Jordan (D_1) and Kuwait (D_2), with 129 and 122 samples, respectively. The problem of predicting grades is viewed as a binary classification task, with numerical grades transformed into pass and fail. Predictors such as grade, holidays-taken, and class-participation are included, and the dataset demonstrates an inherent geospatial distribution shift as the data characteristics of students vary across countries. The features we use are: "sex", "age", "address", "famsize", "Pstatus", "Medu", "Fedu", "Mjob", "Fjob", "reason", "guardian", "traveltime", "studytime", "failures", "schoolsup", "famsup", "paid", "activities", "nursery", "higher", "internet", "romantic", "famrel", "freetime", "goout", "Dalc", "Walc", "health", "absences".
3. German credit dataset Dua et al. (2017): It contains 900 samples from two versions each. The applicants' loan amount, employment history, and age are used to predict their credit score. Additionally, the data exhibits a data correction-based distribution shift, as the data's characteristics differ due to a change in the data preprocessing step. The features we use are: "duration", "amount", "age", "personal-status-sex".

C.2 CLASSIFICATION MODELS

This subsection outlines the fitting process for the classification models. A standard $4:1$ train-test split was employed for model training and evaluation. Identical architectures were used for all models across the datasets, as shown in Table 3. The model performance is evaluated based on the accuracy as shown in Table 4.

Table 3: Classification models architecture

	LR	NN
Units	[Input dimension, 2]	[Input dimension, 50, 2]
Type	Fully connected	Fully connected
Intermediate activation	NA	ReLu
Last layer activation	Softmax	Softmax

Table 4: Average test accuracy for classification models

	Criminal justice	Student performance	German credit
LR	1.00 ± 0.00	0.92 ± 0.01	0.70 ± 0.01
NN	1.00 ± 0.00	0.95 ± 0.01	0.75 ± 0.02

C.3 IMPLEMENTATION DETAILS

For a given dataset, a particular predictive model (NN or LR), and a specific baseline recourse generating method, to validate the theoretical bounds on the recourse invalidation rate, we

1. train predictive model $\mathcal{M}_1$ on the training fold of D_1;
2. use $\mathcal{M}_1$ to obtain prediction result for each sample in the validation fold of D_1;
3. select samples that have negative prediction results;
4. generate recourses for those negatively-predicted samples based on $\mathcal{M}_1$ by using the specified recourse generating method;
5. derive the updated model $\mathcal{M}_2$ on the shifted data D_2;
6. verify Assumption 3 and derive the value of τ based on $\mathcal{M}_2$;
7. for each recourse, compute bounds on the recourse invalidation rate according to Theorems 12 and 13 (since the bounds are also derived through simulation, we need to run Algorithm 1 when computing the bounds);
8. use $\mathcal{M}_2$ to obtain prediction result for each recourse and evaluate the empirical recourse invalidation rate;
9. compare the empirical invalidation rate and the theoretical bounds.

C.4 ADDITIONAL EXPERIMENTAL RESULTS

In Table 5 and Table 6, we provide empirical invalidation rates of recourses generated by baseline algorithms. We report the averaged empirical invalidation rate as well as its standard deviation.

Table 5: Empirical invalidation rate of recourses under model shifts (ℓ_1 cost)

Algorithm	Dataset	Predictive model	Empirical invalidation rate
CF	Criminal justice	LR	0.69 ± 0.09
		NN	0.48 ± 0.09
	Student performance	LR	0.71 ± 0.09
		NN	0.52 ± 0.09
	German credit	LR	0.46 ± 0.27
		NN	0.53 ± 0.06
AR	Criminal justice	LR	0.84 ± 0.06
		NN	0.35 ± 0.17
	Student performance	LR	0.57 ± 0.14
		NN	0.17 ± 0.10
	German credit	LR	0.47 ± 0.21
		NN	0.69 ± 0.06
MINT	German credit	LR	0.07 ± 0.07
		NN	0.37 ± 0.11

In Table 7, we provide the theoretical and empirical recourse invalidation by the considered baseline algorithms with PFC cost.

The effectiveness of Algorithm 2 (PiRR) is demonstrated by Table 8, which reports the performance of PiRR and four other baseline robust recourse generating methods in terms of invalidation rates before and after the model shift, along with the average cost computed under the PFC cost. Figure 3 compares the performance of PiRR with baseline methods in generating recourse under 3 different prescribed invalidation rates: $0.05, 0.10, 0.15$. Figure 4 investigates the impact of ϵ on the performance of PiRR.

For Algorithm 4, we use the considered three baseline recourse generating methods to generate recourses for negatively-predicted samples in the validation fold. To obtain recourses with different costs and robustness, we vary the value of the trade-off parameter λ. In particular, we choose $\lambda = \{0.1, 0.5, 0.9, 1.3, 1.7\}$. The results are shown in Figure 5.

Table 6: Empirical invalidation rate of recourses under model shifts (PFC cost)

Algorithm	Dataset	Predictive model	Empirical invalidation rate
CF	Criminal justice	LR	0.74 ± 0.11
		NN	0.50 ± 0.13
	Student performance	LR	0.82 ± 0.10
		NN	0.70 ± 0.14
	German credit	LR	0.44 ± 0.33
		NN	0.49 ± 0.12
AR	Criminal justice	LR	0.91 ± 0.05
		NN	0.65 ± 0.17
	Student performance	LR	0.76 ± 0.11
		NN	0.18 ± 0.11
	German credit	LR	0.46 ± 0.27
		NN	0.44 ± 0.15
MINT	German credit	LR	0.05 ± 0.08
		NN	0.36 ± 0.15

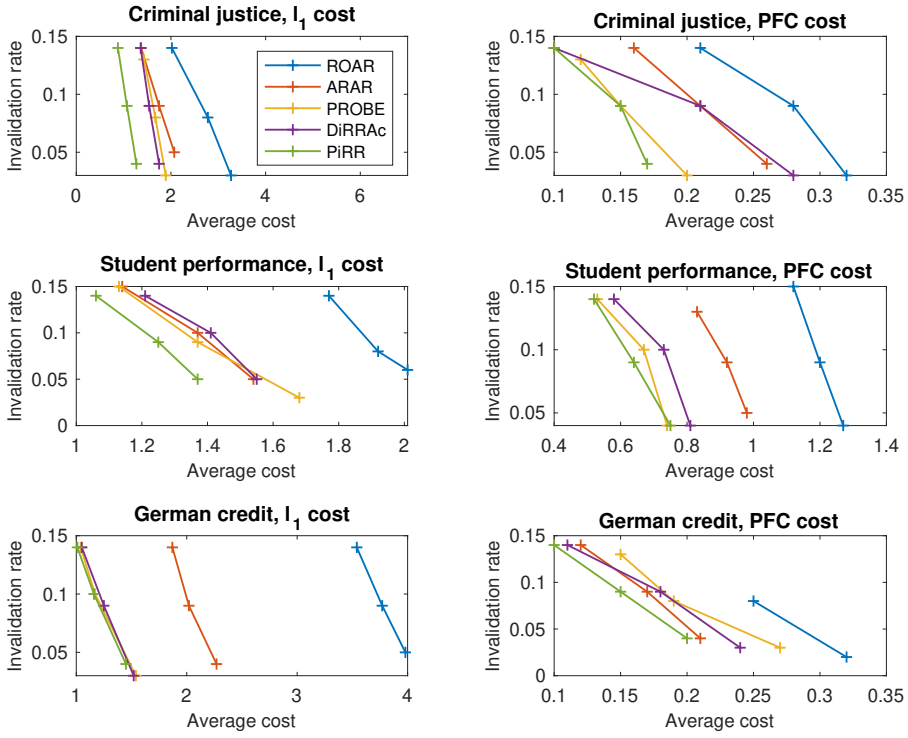

Figure 3: Recourse invalidation rate v.s. recourse cost plot. For any given invalidation rate, PiRR could generate recourses that satisfy the invalidation requirement while maintaining low recourse costs. The average recourse costs of robust recourses generated by PiRR are smaller than other methods under the same invalidation rate constraint.

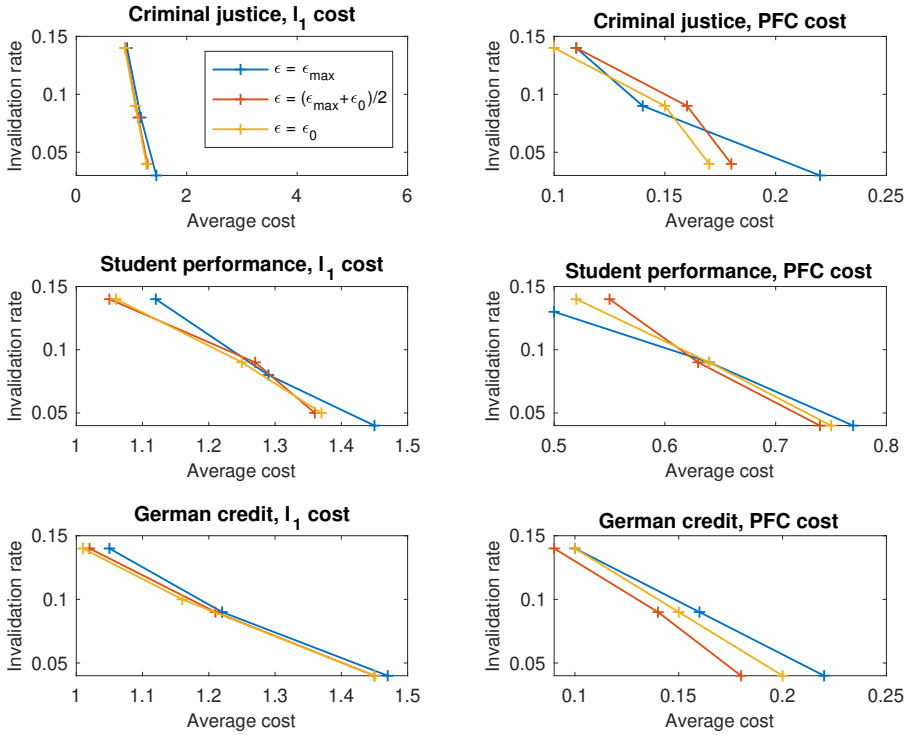

Figure 4: Impact of ϵ on the performance of PiRR, where $\epsilon_0 = \min_{\{s \in \mathcal{S}_{\text{train}}\}} \hat{p}(s), \epsilon_{\max} = \sum_{\{s' \in \mathcal{S}_{\text{train}}: s-\tau \leq s' \leq s+\tau\}} \hat{p}(s')$. PiRR consistently generates recourses that meet the invalidation requirements across different values of ϵ, while maintaining similar overall performance.

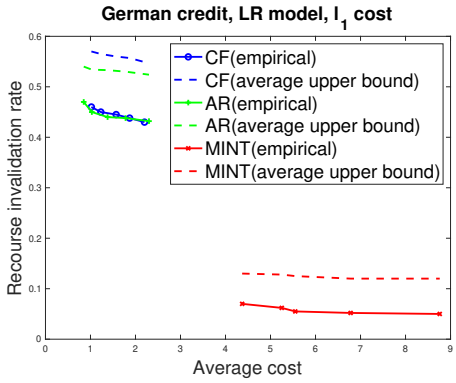

Figure 5: Recourse invalidation rate v.s. recourse cost plot for Algorithm 4. As λ varies, the recourse cost changes, while the recourse invalidation rate changes only slightly. The theoretical bounds on the recourse invalidation rates are valid.

Table 7: Theoretical and empirical recourse invalidation (PFC cost,$\epsilon = \min_{\{s \in \mathcal{S}_{\text{train}}\}} \hat{p}(s)$)

Algorithm	Dataset	Predictive model	Upper-bound in Theorem 12	Upper-bound in Theorem 13	Empirical invalidation rate
CF	Criminal justice	LR	0.85 ± 0.04	0.79 ± 0.05	0.73
		NN	0.67 ± 0.06	0.58 ± 0.10	0.51
	Student performance	LR	0.91 ± 0.08	0.88 ± 0.09	0.82
		NN	0.83 ± 0.08	0.78 ± 0.11	0.69
	German credit	LR	0.64 ± 0.06	0.59 ± 0.06	0.43
		NN	0.66 ± 0.06	0.64 ± 0.07	0.50
AR	Criminal justice	LR	0.93 ± 0.04	0.92 ± 0.04	0.90
		NN	0.77 ± 0.03	0.69 ± 0.04	0.65
	Student performance	LR	0.85 ± 0.08	0.81 ± 0.09	0.76
		NN	0.28 ± 0.04	0.22 ± 0.04	0.19
	German credit	LR	0.64 ± 0.08	0.56 ± 0.11	0.50
		NN	0.59 ± 0.04	0.54 ± 0.05	0.42
MINT	German credit	LR	0.25 ± 0.07	0.16 ± 0.08	0.06
		NN	0.51 ± 0.08	0.45 ± 0.09	0.37

Table 8: The performance of PiRR is compared with other robust recourse generating methods, using LR model and the PFC cost function. In PiRR, the invalidation targets specified by the user are assumed to be 0.10 and 0.05. The results show that PiRR generates recourse solutions that always meet the user's invalidation targets. For the recourse cost, when compared with existing baselines, the recourse solutions generated by PiRR are easier to implement by users.

Dataset	Algorithm	Invalidation rate before shift ($\mathcal{M}_1$)	Invalidation rate after shift ($\mathcal{M}_2$)	Average cost
Criminal justice	ROAR	0.00 ± 0.00	0.02 ± 0.01	0.44 ± 0.12
	ARAR	0.00 ± 0.00	0.02 ± 0.02	0.36 ± 0.10
	PROBE	0.00 ± 0.00	0.02 ± 0.01	$\mathbf{0.25 \pm 0.09}$
	DiRRAc	0.00 ± 0.00	$\mathbf{0.01 \pm 0.02}$	0.28 ± 0.12
	PiRR(0.10)	$\mathbf{0.00 \pm 0.00}$	0.07 ± 0.02	0.16 ± 0.06
	PiRR(0.05)	$\mathbf{0.00 \pm 0.00}$	0.03 ± 0.01	$\mathbf{0.25 \pm 0.08}$
Student performance	ROAR	0.00 ± 0.00	0.09 ± 0.07	1.20 ± 0.10
	ARAR	0.00 ± 0.00	0.06 ± 0.07	0.92 ± 0.09
	PROBE	0.00 ± 0.00	0.04 ± 0.07	0.74 ± 0.10
	DiRRAc	0.00 ± 0.00	0.04 ± 0.06	0.81 ± 0.08
	PiRR(0.10)	$\mathbf{0.00 \pm 0.00}$	0.07 ± 0.02	$\mathbf{0.85 \pm 0.08}$
	PiRR(0.05)	$\mathbf{0.00 \pm 0.00}$	$\mathbf{0.03 \pm 0.02}$	$\mathbf{0.72 \pm 0.10}$
German credit	ROAR	0.00 ± 0.00	$\mathbf{0.00 \pm 0.00}$	0.36 ± 0.08
	ARAR	0.00 ± 0.00	0.02 ± 0.02	0.27 ± 0.06
	PROBE	0.00 ± 0.00	0.02 ± 0.01	$\mathbf{0.27 \pm 0.07}$
	DiRRAc	0.00 ± 0.00	0.01 ± 0.02	0.32 ± 0.08
	PiRR(0.10)	$\mathbf{0.00 \pm 0.00}$	0.07 ± 0.02	0.21 ± 0.06
	PiRR(0.05)	$\mathbf{0.00 \pm 0.00}$	$\mathbf{0.02 \pm 0.02}$	$\mathbf{0.26 \pm 0.07}$

D PROOFS

D.1 PROOF OF LEMMA 5

$$
\begin{aligned}
r_{\text{ivd}}(\boldsymbol{x}^{\text{CF}}) &= \mathbb{P}(f'(\boldsymbol{x}^{\text{CF}}) = 0 | f(\boldsymbol{x}^{\text{CF}}) = 1) \\
&= 1 - \mathbb{P}(h'(\boldsymbol{x}^{\text{CF}}) \geq \eta | h(\boldsymbol{x}^{\text{CF}}) \geq \eta) \\
&\overset{(a)}{=} 1 - \mathbb{P}(h'(\boldsymbol{x}^{\text{CF}}) \geq \eta | h(\boldsymbol{x}^{\text{CF}}) \geq \eta, h'(\boldsymbol{x}^{\text{CF}}) \geq \eta - \tau) \\
&= 1 - \frac{\mathbb{P}(h'(\boldsymbol{x}^{\text{CF}}) \geq \eta, h(\boldsymbol{x}^{\text{CF}}) \geq \eta | h'(\boldsymbol{x}^{\text{CF}}) \geq \eta - \tau)}{\mathbb{P}(h(\boldsymbol{x}^{\text{CF}}) \geq \eta | h'(\boldsymbol{x}^{\text{CF}}) \geq \eta - \tau)} \\
&= 1 - \frac{\mathbb{P}(h'(\boldsymbol{x}^{\text{CF}}) \geq \eta | h'(\boldsymbol{x}^{\text{CF}}) \geq \eta - \tau) \mathbb{P}(h(\boldsymbol{x}^{\text{CF}}) \geq \eta | h'(\boldsymbol{x}^{\text{CF}}) \geq \eta, h'(\boldsymbol{x}^{\text{CF}}) \geq \eta - \tau)}{\mathbb{P}(h(\boldsymbol{x}^{\text{CF}}) \geq \eta | h'(\boldsymbol{x}^{\text{CF}}) \geq \eta - \tau)} \\
&= 1 - \mathbb{P}(h'(\boldsymbol{x}^{\text{CF}}) \geq \eta | h'(\boldsymbol{x}^{\text{CF}}) \geq \eta - \tau) \frac{\mathbb{P}(h(\boldsymbol{x}^{\text{CF}}) \geq \eta | h'(\boldsymbol{x}^{\text{CF}}) \geq \eta, h'(\boldsymbol{x}^{\text{CF}}) \geq \eta - \tau)}{\mathbb{P}(h(\boldsymbol{x}^{\text{CF}}) \geq \eta | h'(\boldsymbol{x}^{\text{CF}}) \geq \eta - \tau)} \\
&\overset{(b)}{=} 1 - \mathbb{P}(h'(\boldsymbol{x}^{\text{CF}}) \geq \eta | h'(\boldsymbol{x}^{\text{CF}}) \geq \eta - \tau) \\
&\quad \cdot \frac{\mathbb{P}(h(\boldsymbol{x}^{\text{CF}}) \geq \eta | h'(\boldsymbol{x}^{\text{CF}}) \geq \eta, h(\boldsymbol{x}^{\text{CF}}) \geq \eta - \tau, h'(\boldsymbol{x}^{\text{CF}}) \geq \eta - \tau)}{\mathbb{P}(h(\boldsymbol{x}^{\text{CF}}) \geq \eta | h'(\boldsymbol{x}^{\text{CF}}) \geq \eta - \tau)} \\
&\leq 1 - \mathbb{P}(h'(\boldsymbol{x}^{\text{CF}}) \geq \eta | h'(\boldsymbol{x}^{\text{CF}}) \geq \eta - \tau) \\
&= 1 - \frac{\mathbb{P}(h'(\boldsymbol{x}^{\text{CF}}) \geq \eta, h'(\boldsymbol{x}^{\text{CF}}) \geq \eta - \tau)}{\mathbb{P}(h'(\boldsymbol{x}^{\text{CF}}) \geq \eta - \tau)} \\
&= 1 - \frac{\mathbb{P}(h'(\boldsymbol{x}^{\text{CF}}) \geq \eta)}{\mathbb{P}(h'(\boldsymbol{x}^{\text{CF}}) \geq \eta - \tau)} \\
&= 1 - \frac{\mathbb{P}(s'(\boldsymbol{x}^{\text{CF}}, 1) \leq 1 - \eta)}{\mathbb{P}(s'(\boldsymbol{x}^{\text{CF}}, 1) \leq 1 - \eta + \tau)},
\end{aligned} \tag{10}
$$

where (a) is true as $h(\boldsymbol{x}^{\text{CF}}) \geq \eta$ implies $h'(\boldsymbol{x}^{\text{CF}}) \geq \eta - \tau$ based on Assumption 3. Similarly, (b) holds because $h'(\boldsymbol{x}^{\text{CF}}) \geq \eta$ implies $h(\boldsymbol{x}^{\text{CF}}) \geq \eta - \tau$ based on Assumption 3.

D.2 PROOF OF LEMMA 7

Under Assumption 3, we have $h(\boldsymbol{x}) - \tau \leq h'(\boldsymbol{x}) \leq h(\boldsymbol{x}) + \tau, \forall \boldsymbol{x} \in \mathcal{X}$, which implies

$$s(\boldsymbol{x}, y = 0) - \tau \leq s'(\boldsymbol{x}, y = 0) \leq s(\boldsymbol{x}, y = 0) + \tau,$$

as well as

$$s(\boldsymbol{x}, y = 1) - \tau = 1 - h(\boldsymbol{x}) - \tau \leq 1 - h'(\boldsymbol{x}) = s'(\boldsymbol{x}, y = 1) \leq 1 - h(\boldsymbol{x}) + \tau = s(\boldsymbol{x}, y = 1) + \tau.$$

Then we have

$$s(\boldsymbol{x}, y) - \tau \leq s'(\boldsymbol{x}, y) \leq s(\boldsymbol{x}, y) + \tau,$$

which indicates that to derive p' from p, only density in the τ-neighborhood of s can be moved to s. Then for any neighborhood of s with radius δ, the cumulative probability under the distribution p' over this neighborhood is always upper-bounded by the cumulative probability under the distribution p over a larger neighborhood of radius $\tau + \delta$ around s. Specifically, we have

$$\int_{s-\delta}^{s+\delta} p'(t)dt \leq \int_{s-\tau-\delta}^{s+\tau+\delta} p(t)d(t).$$

D.3 PROOF OF PROPOSITION 8

In the following, we denote the random variables S_i and realized values $s_i, i = 1, 2, \cdots, n$. For notation simplicity, we denote $s_{n+1} = s'(\boldsymbol{x}^{\text{CF}}, y^{\text{CF}})$. From Tibshirani et al. (Dec. 2019), we know that independent draws are always weighted exchangeable, with weight functions given by likelihood

ratios. Thus, according to Definition 1 and Lemma 2 in Tibshirani et al. (Dec. 2019), we have that random variables $S_1, \cdots, S_{n+1}$ are weighted exchangeable and

$$f(s_1, \cdots, s_{n+1}) = \Pi_{i=1}^{n+1} v_i(s_i) g(s_1, \cdots, s_{n+1}), \tag{11}$$

where f represents the joint pdf, $v_i(s_i) = 1, i = 1, \cdots n, v_{n+1}(s_{n+1}) = v(s'(\boldsymbol{x}^{\text{CF}}, y^{\text{CF}}))$ and g is a permutation-invariant function.

For a set of values $s_1, \cdots, s_{n+1}$ where there might be repeated elements, we denote the unordered set $\boldsymbol{s} = [s_1, \cdots, s_{n+1}]$ and denote an event $E_{\boldsymbol{s}} = \{[S_1, S_2, \cdots, S_{n+1}] = [s_1, s_2, \cdots, s_{n+1}]\}$. Let Π_{n+1} be the set of all permutations of $\{1, \cdots, n+1\}$. Then we have

$$\begin{aligned}
&\mathbb{P}(S_{n+1} = s_i | E_{\boldsymbol{s}}) \\
= \; & \frac{\sum_{\pi \in \Pi_{n+1}: \pi(n+1) = i} f(s_{\pi(1)}, s_{\pi(2)}, \cdots, s_{\pi(n+1)})}{\sum_{\pi \in \Pi_{n+1}} f(s_{\pi(1)}, s_{\pi(2)}, \cdots, s_{\pi(n+1)})} \\
\overset{(a)}{=} \; & \frac{\sum_{\pi \in \Pi_{n+1}: \pi(n+1) = i} \Pi_{i=1}^{n+1} v_i(s_{\pi(i)}) g(s_{\pi(1)}, s_{\pi(2)}, \cdots, s_{\pi(n+1)})}{\sum_{\pi \in \Pi_{n+1}} \Pi_{i=1}^{n+1} v_i(s_{\pi(i)}) g(s_{\pi(1)}, s_{\pi(2)}, \cdots, s_{\pi(n+1)})} \\
= \; & \frac{\sum_{\pi \in \Pi_{n+1}: \pi(n+1) = i} v_{n+1}(s_i) g(s_{\pi(1)}, s_{\pi(2)}, \cdots, s_{\pi(n+1)})}{\sum_{\pi \in \Pi_{n+1}} v_{n+1}(s_{\pi(n+1)}) g(s_{\pi(1)}, s_{\pi(2)}, \cdots, s_{\pi(n+1)})} \\
= \; & \frac{v_{n+1}(s_i)}{\sum_{j=1}^{n+1} v_{n+1}(s_j)} = \frac{v(s_i)}{\sum_{j=1}^{n+1} v(s_j)} := p_i,
\end{aligned}$$

where (a) is due to equation 11.

Then for any unordered set $\boldsymbol{s}$, we have

$$\mathbb{P}(S_{n+1} \leq s_{[k^*]} | E_{\boldsymbol{s}}) = \sum_{i=1}^{n+1} p_i \mathbf{1}_{s_i \leq s_{[k^*]}} = \frac{\sum_{i=1}^{n+1} v(s_i) \mathbf{1}_{s_i \leq s_{[k^*]}}}{\sum_{j=1}^{n+1} v(s_j)}. \tag{12}$$

Recall that $\hat{F}$ is defined as

$$\hat{F}(k) = \frac{\sum_{i=1}^{k} L_{[i]}}{\sum_{i=1}^{k} L_{[i]} + \sum_{i=k+1}^{n} U_{[i]} + U^{\text{CF}}}.$$

Since $k^* = \min\{k : \hat{F}(k) \geq 1 - \alpha\}$, we have

$$S_{[k^*]} = \inf \left\{ s : \frac{\sum_{i=1}^{n} L_i \mathbf{1}_{S_i \leq s}}{\sum_{i=1}^{n} L_i \mathbf{1}_{S_i \leq s} + \sum_{i=1}^{n} U_i \mathbf{1}_{S_i > s} + U^{\text{CF}}} \geq 1 - \alpha \right\},$$

which indicates that

$$\mathbb{E}\left[\frac{\sum_{i=1}^{n} L_i \mathbf{1}_{S_i \leq S_{[k^*]}}}{\sum_{i=1}^{n} L_i \mathbf{1}_{S_i \leq S_{[k^*]}} + \sum_{i=1}^{n} U_i \mathbf{1}_{S_i > S_{[k^*]}} + U^{\text{CF}}} \right] \geq 1 - \alpha. \tag{13}$$

In the meantime, we apply the power property on equation 12 and have

$$\mathbb{P}(S_{n+1} \leq s_{[k^*]}) = \mathbb{E}[\mathbb{P}(S_{n+1} \leq s_{[k^*]}|E_{\boldsymbol{S}})] = \mathbb{E}\left[\frac{\sum_{i=1}^{n+1} v(S_i)\mathbf{1}_{S_i \leq S_{[k^*]}}}{\sum_{j=1}^{n+1} v(S_i)}\right]. \tag{14}$$

By combining equation 13 and equation 14, we have

$$\begin{aligned}
& \mathbb{P}(S_{n+1} \leq s_{[k^*]}) - (1-\alpha) \\
\geq\ & \mathbb{E}\left[\frac{\sum_{i=1}^{n+1} v(S_i)\mathbf{1}_{S_i \leq S_{[k^*]}}}{\sum_{i=1}^{n+1} v(S_i)}\right] - \mathbb{E}\left[\frac{\sum_{i=1}^{n} L_i\mathbf{1}_{S_i \leq S_{[k^*]}}}{\sum_{i=1}^{n} L_i\mathbf{1}_{S_i \leq S_{[k^*]}} + \sum_{i=1}^{n} U_i\mathbf{1}_{S_i > S_{[k^*]}} + U^{\text{CF}}}\right] \\
\geq\ & \mathbb{E}\left[\frac{\sum_{i=1}^{n} v(S_i)\mathbf{1}_{S_i \leq S_{[k^*]}}}{\sum_{i=1}^{n+1} v(S_i)}\right] - \mathbb{E}\left[\frac{\sum_{i=1}^{n} L_i\mathbf{1}_{S_i \leq S_{[k^*]}}}{\sum_{i=1}^{n} L_i\mathbf{1}_{S_i \leq S_{[k^*]}} + \sum_{i=1}^{n} U_i\mathbf{1}_{S_i > S_{[k^*]}} + U^{\text{CF}}}\right] \\
=\ & \mathbb{E}\left[\frac{q(S_1, \cdots, S_{n+1})}{\left[\sum_{i=1}^{n+1} v(S_i)\right]\left[\sum_{i=1}^{n} L_i\mathbf{1}_{S_i \leq S_{[k^*]}} + \sum_{i=1}^{n} U_i\mathbf{1}_{S_i > S_{[k^*]}} + U^{\text{CF}}\right]}\right],
\end{aligned} \tag{15}$$

where

$$\begin{aligned}
& q(S_1, \cdots, S_{n+1}) \\
=\ & \left[\sum_{i=1}^{n} v(S_i)\mathbf{1}_{S_i \leq S_{[k^*]}}\right] \cdot \left[\sum_{i=1}^{n} L_i\mathbf{1}_{S_i \leq S_{[k^*]}}\right] + \left[\sum_{i=1}^{n} v(S_i)\mathbf{1}_{S_i \leq S_{[k^*]}}\right] \cdot \left[\sum_{i=1}^{n} U_i\mathbf{1}_{S_i > S_{[k^*]}}\right] \\
& + \left[\sum_{i=1}^{n} v(S_i)\mathbf{1}_{S_i \leq S_{[k^*]}}\right] U^{\text{CF}} - \left[\sum_{i=1}^{n+1} v(S_i)\right] \cdot \left[\sum_{i=1}^{n} L_i\mathbf{1}_{S_i \leq S_{[k^*]}}\right] \\
=\ & \left[\sum_{i=1}^{n} v(S_i)\mathbf{1}_{S_i \leq S_{[k^*]}}\right] \cdot \left[\sum_{i=1}^{n} L_i\mathbf{1}_{S_i \leq S_{[k^*]}}\right] - \left[\sum_{i=1}^{n+1} v(S_i)\right] \cdot \left[\sum_{i=1}^{n} L_i\mathbf{1}_{S_i \leq S_{[k^*]}}\right] \\
& + \left[\sum_{i=1}^{n} v(S_i)\mathbf{1}_{S_i \leq S_{[k^*]}}\right] \cdot \left[\sum_{i=1}^{n} U_i\mathbf{1}_{S_i > S_{[k^*]}}\right] + \left[\sum_{i=1}^{n} v(S_i)\mathbf{1}_{S_i \leq S_{[k^*]}}\right] U^{\text{CF}} \\
=\ & \left\{\left[\sum_{i=1}^{n} v(S_i)\mathbf{1}_{S_i \leq S_{[k^*]}}\right] \cdot \left[\sum_{i=1}^{n} U_i\mathbf{1}_{S_i > S_{[k^*]}}\right] - \left[\sum_{i=1}^{n} v(S_i)\mathbf{1}_{S_i > S_{[k^*]}}\right] \cdot \left[\sum_{i=1}^{n} L_i\mathbf{1}_{S_i \leq S_{[k^*]}}\right]\right\} \\
& + \left\{U^{\text{CF}}\left[\sum_{i=1}^{n} v(S_i)\mathbf{1}_{S_i \leq S_{[k^*]}}\right] - v(S_{n+1})\left[\sum_{i=1}^{n} L_i\mathbf{1}_{S_i \leq S_{[k^*]}}\right]\right\}
\end{aligned}$$

$$
\begin{aligned}
\overset{(b)}{\geq} \quad & \left\{ \left[\sum_{i=1}^{n} v(S_i) \mathbf{1}_{S_i \leq S_{[k^*]}} \right] \cdot \left[\sum_{i=1}^{n} v(S_i) \mathbf{1}_{S_i > S_{[k^*]}} \right] \right. \\
& \left. - \left[\sum_{i=1}^{n} v(S_i) \mathbf{1}_{S_i > S_{[k^*]}} \right] \cdot \left[\sum_{i=1}^{n} (v(S_i) + \max\{0, L_i - v(S_i)\}) \mathbf{1}_{S_i \leq S_{[k^*]}} \right] \right\} \\
& + \left\{ U^{\text{CF}} \left[\sum_{i=1}^{n} v(S_i) \mathbf{1}_{S_i \leq S_{[k^*]}} \right] - v(S_{n+1}) \left[\sum_{i=1}^{n} (v(S_i) + \max\{0, L_i - v(S_i)\}) \mathbf{1}_{S_i \leq S_{[k^*]}} \right] \right\} \\
= \quad & \left\{ \left[\sum_{i=1}^{n} v(S_i) \mathbf{1}_{S_i \leq S_{[k^*]}} \right] \cdot \left[\sum_{i=1}^{n} v(S_i) \mathbf{1}_{S_i > S_{[k^*]}} \right] \right. \\
& - \left[\sum_{i=1}^{n} v(S_i) \mathbf{1}_{S_i > S_{[k^*]}} \right] \cdot \left[\sum_{i=1}^{n} v(S_i) \mathbf{1}_{S_i \leq S_{[k^*]}} \right] \\
& \left. - \left[\sum_{i=1}^{n} v(S_i) \mathbf{1}_{S_i > S_{[k^*]}} \right] \cdot \left[\sum_{i=1}^{n} \max\{0, L_i - v(S_i)\} \mathbf{1}_{S_i \leq S_{[k^*]}} \right] \right\} \\
& + \left\{ U^{\text{CF}} \left[\sum_{i=1}^{n} v(S_i) \mathbf{1}_{S_i \leq S_{[k^*]}} \right] - v(S_{n+1}) \left[\sum_{i=1}^{n} v(S_i) \mathbf{1}_{S_i \leq S_{[k^*]}} \right] \right. \\
& \left. - v(S_{n+1}) \left[\sum_{i=1}^{n} \max\{0, L_i - v(S_i)\} \mathbf{1}_{S_i \leq S_{[k^*]}} \right] \right\} \\
\geq \quad & - \left[\sum_{i=1}^{n} v(S_i) \mathbf{1}_{S_i > S_{[k^*]}} \right] \cdot \left[\sum_{i=1}^{n} \max\{0, L_i - v(S_i)\} \mathbf{1}_{S_i \leq S_{[k^*]}} \right] \\
& - v(S_{n+1}) \left[\sum_{i=1}^{n} \max\{0, L_i - v(S_i)\} \mathbf{1}_{S_i \leq S_{[k^*]}} \right] \\
\geq \quad & - \left[\sum_{i=1}^{n+1} v(S_i) \right] \cdot \left[\sum_{i=1}^{n} \max\{0, L_i - v(S_i)\} \right], \qquad (16)
\end{aligned}
$$

in which (b) is due to the fact that $\hat{U}(s) \geq v(s)$ almost surely since the formula for $\hat{U}(\cdot)$ is motivated by Lemma 7.

Then we come back to equation 15 and have

$$
\begin{aligned}
& \mathbb{E} \left[\frac{q(S_1, \cdots, S_{n+1})}{\left[\sum_{i=1}^{n+1} v(S_i) \right] \left[\sum_{i=1}^{n} L_i \mathbf{1}_{S_i \leq S_{[k^*]}} + \sum_{i=1}^{n} U_i \mathbf{1}_{S_i > S_{[k^*]}} + U^{\text{CF}} \right]} \right] \\
\geq \quad & \mathbb{E} \left[\frac{- \left[\sum_{i=1}^{n} v(S_i) \right] \cdot \left[\sum_{i=1}^{n} \max\{0, L_i - v(S_i)\} \right]}{\left[\sum_{i=1}^{n+1} v(S_i) \right] \left[\sum_{i=1}^{n} L_i \right]} \right] \\
\geq \quad & - \mathbb{E} \left[\frac{\sum_{i=1}^{n} \max\{0, L_i - v(S_i)\}}{\sum_{i=1}^{n} L_i} \right].
\end{aligned}
$$

By Hölder's Inequality, we have

$$\begin{aligned}
&\mathbb{E}\left[\frac{\sum_{i=1}^{n}\max\{0, L_i - v(S_i)\}}{\sum_{i=1}^{n} L_i}\right] \\
\leq\ & \left\|\frac{1}{n}\sum_{i=1}^{n}\max\{0, L_i - v(S_i)\}\right\|_p \cdot \left\|\frac{n}{\sum_{i=1}^{n} L_i}\right\|_q \\
\overset{(b)}{\leq}\ & \|\max\{0, L_i - v(S_i)\}\|_p \cdot \left\|\frac{n}{\sum_{i=1}^{n} L_i}\right\|_q \\
\leq\ & \|\max\{0, L_i - v(S_i)\}\|_p \cdot \left\|\frac{1}{n}\sum_{i=1}^{n}\frac{1}{L_i}\right\|_q \\
\overset{(c)}{\leq}\ & \|\max\{0, L_i - v(S_i)\}\|_p \cdot \left\|\frac{1}{L_i}\right\|_q,
\end{aligned}$$

where (b) and (c) follow from the Minkowski's inequality. Thus, we have

$$\mathbb{P}(S_{n+1} \leq s_{[k^*]}) - (1-\alpha) \geq -\|\max\{0, L_i - v(S_i)\}\|_p \cdot \left\|\frac{1}{L_i}\right\|_q.$$

Since S_i is a random variable and $L_i = \hat{L}(S_i)$, in general, we have

$$\mathbb{P}(S_{n+1} \leq s_{[k^*]}) - (1-\alpha) \geq -\left\|\frac{1}{\hat{L}(S)}\right\|_q \cdot \left\|\max\{0, \hat{L}(S) - v(S)\}\right\|_p.$$

Proposition 10 can be proved by following similar process.

D.4 Proof of Theorem 12

Given the definition of k_1^*, we have $\mathbb{P}(s'(\boldsymbol{x}^{\text{CF}}, y^{\text{CF}}) \leq s_{[k_1^*]}) \leq \mathbb{P}(s'(\boldsymbol{x}^{\text{CF}}, y^{\text{CF}}) \leq 1-\eta)$. Then if $s'(\boldsymbol{x}^{\text{CF}}, y^{\text{CF}}) \leq 1-\eta$, we have

$$1 - \eta \geq s'(\boldsymbol{x}^{\text{CF}}, y^{\text{CF}} = 1) = 1 - h'(\boldsymbol{x}^{\text{CF}})$$

or

$$1 - \eta \geq s'(\boldsymbol{x}^{\text{CF}}, y^{\text{CF}} = 0) = h'(\boldsymbol{x}^{\text{CF}}). \tag{17}$$

However, according to Assumption 3, we have

$$h'(\boldsymbol{x}^{\text{CF}}) \geq \eta - \tau \overset{(a)}{>} 1 - \eta,$$

where (a) is true because $2\eta - \tau > 1$. Then we see a contradiction in equation 17 and conclude that as long as $s'(\boldsymbol{x}^{\text{CF}}, y^{\text{CF}}) \leq 1-\eta$ and $2\eta - \tau > 1$, we have $y^{\text{CF}} = 1$. Thus, we have

$$\begin{aligned}
r_{\text{ivd}}(\boldsymbol{x}^{\text{CF}}) &\overset{(b)}{\leq} 1 - \frac{\mathbb{P}(h'(\boldsymbol{x}^{\text{CF}}) \geq \eta)}{\mathbb{P}(h'(\boldsymbol{x}^{\text{CF}}) \geq \eta - \tau)} \\
&\leq 1 - \mathbb{P}(h'(\boldsymbol{x}^{\text{CF}}) \geq \eta) \\
&= 1 - \mathbb{P}(s'(\boldsymbol{x}^{\text{CF}}, 1) \leq 1 - \eta) \\
&\leq 1 - \mathbb{P}(s'(\boldsymbol{x}^{\text{CF}}, y^{\text{CF}}) \leq s_{[k_1^*]}) \\
&\leq \alpha_1 + \hat{\Delta}_F,
\end{aligned}$$

where (b) is from equation 10.

Similarly, if $h(\boldsymbol{x}^{\text{CF}}) > 1 - \eta + \tau$, according to Assumption 3, we have $h'(\boldsymbol{x}^{\text{CF}}) \geq h(\boldsymbol{x}^{\text{CF}}) - \tau > 1 - \eta$, which also causes a contradiction in equation 17. Thus, we have $y^{\text{CF}} = 1$ and $r_{\text{ivd}}(\boldsymbol{x}^{\text{CF}}) \leq \alpha_1 + \hat{\Delta}_F$.

D.5 PROOF OF THEOREM 13

$$
\begin{aligned}
r_{\text{ivd}}(\boldsymbol{x}^{\text{CF}}) &\leq 1 - \frac{\mathbb{P}(h'(\boldsymbol{x}^{\text{CF}}) \geq \eta)}{\mathbb{P}(h'(\boldsymbol{x}^{\text{CF}}) \geq \eta - \tau)} \\
&= 1 - \frac{\mathbb{P}(s'(\boldsymbol{x}^{\text{CF}}, 1) \leq 1 - \eta)}{\mathbb{P}(s'(\boldsymbol{x}^{\text{CF}}, y^{\text{CF}}) \leq 1 - \eta + \tau)} \\
&\leq 1 - \frac{\mathbb{P}(s'(\boldsymbol{x}^{\text{CF}}, 1) \leq 1 - \eta)}{\mathbb{P}(s'(\boldsymbol{x}^{\text{CF}}, y^{\text{CF}}) \leq s_{[k_2^*]})} \\
&\leq 1 - \frac{\mathbb{P}(s'(\boldsymbol{x}^{\text{CF}}, y^{\text{CF}}) \leq s_{[k_1^*]})}{\mathbb{P}(s'(\boldsymbol{x}^{\text{CF}}, y^{\text{CF}}) \leq s_{[k_2^*]})} \\
&\leq 1 - \frac{1 - \alpha_1 - \hat{\Delta}_F}{1 - \alpha_2 + \hat{\Delta}_E},
\end{aligned}
$$

where $\alpha_2 = 1 - \hat{E}(k_2^*)$ with $k_2^* = \min\{k_2 : s_{[k_2]} \geq 1 - \eta + \tau\}$.