DISTRIBUTIONALLY ROBUST DECISION-MAKING UNDER REJECT INFERENCE CONTEXT

Anonymous authors
Paper under double-blind review

ABSTRACT

Reject inference arises when outcomes are observed only after acceptance decisions, yielding selectively labeled data that are shaped by historical policies. While existing methods mainly improve nominal learning under such one-sided feedback, they are often vulnerable to deployment shifts that alter the true conditional success probabilities. We therefore propose a distributionally robust framework for policy learning under the reject inference context. The objective is to find a robust policy that maximizes the worst-case discounted reward, with the ambiguity set characterized as a Kullback–Leibler ball centered at the posterior beliefs. We derive a dynamic programming relation for the worst-case discounted reward and introduce a practical backward approximation algorithm to compute the optimal robust policy. Extensive experiments on synthetic and real lending data show that the proposed method consistently improves robustness and lower-tail performance relative to competitive baselines, at only a modest cost in nominal reward.

1 INTRODUCTION

Reject inference often appears in high-stakes decision settings where outcomes are observed only for individuals who receive an acceptance decision (Smith & Elkan, 2004). In lending, for example, repayment is observed only for approved applicants, while the outcomes of rejected ones remain unknown (Lakkaraju et al., 2017). Similar issues occur in hiring, bail, and healthcare applications.

In such settings, the learner must make sequential decisions based on one-sided and selectively labeled feedback, which makes it particularly vulnerable to distribution shift. When macroeconomic conditions, borrower composition, or local risk patterns evolve over time, the true conditional success probability may easily deviate from the belief formed from previously accepted cases (Si et al., 2023). Because current decisions also determine which future labels become observable, such misspecification is difficult to detect and may propagate through the learning process, causing nominal policies to become overly aggressive and incur substantial downside risk.

Existing work has made important progress on reject inference. One line of research seeks to repair the labeled sample, for example through Bayesian modeling (Smith & Elkan, 2004), self-training (Liao et al., 2022), or reject-aware multi-task learning (Liu et al., 2022). Another line formulates the problem as sequential decision making under endogenous feedback, leading to stochastic decision rules (Kilbertus et al., 2020), structured dynamic programs (Wei, 2021), selective-label active learning (Gao et al., 2024), and exploration-based methods (Keswani et al., 2024). However, these approaches primarily target nominal utility under the observed process, rather than explicitly protecting the learned policy against shifts in the underlying success probabilities.

In this paper, we study robust policy learning for the reject inference problem through a distributionally robust optimization framework. We consider a sequential accept/reject problem with discounted utility and explicitly allow the true conditional success probability to deviate from its posterior estimate. To capture such uncertainty, we construct a Kullback–Leibler ambiguity set around the nominal belief state and optimize the worst-case discounted reward over this set. In the covariate-free case, we formulate the problem in belief space and derive an optimal acceptance policy. We then extend the analysis to the covariate-dependent setting, where each feature value induces its own robust dynamic program with an effective discount factor. Because the resulting value func-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

tions generally do not admit closed-form solutions, we further develop a backward approximation scheme for computing the robust policy.

Empirically, we evaluate the proposed method, Distributionally Robust Reject Inference (DRRI), against myopic, stochastic, selective-label, and UCB-style baselines under both nominal and shifted environments. The results show that although DRRI does not maximize reward in the nominal environment, it is substantially more stable when the environment deteriorates. In a real lending case study based on the FICO challenge dataset, DRRI delivers the strongest downside protection, achieving the best minimum and lower-tail performance across repeated online replays.

2 Robust Sequential Decision-Making under Reject Inference

2.1 Reject Inference and Belief Dynamics

Suppose that at each time $t = 0, 1, \ldots$, an individual arrives with covariates $X_t \in \mathcal{X}$. A binary decision $A_t \in \{0, 1\}$ is then made according to a policy Π, where $A_t = 1$ denotes acceptance and $A_t = 0$ denotes rejection. Let $Y_t \in \{0, 1\}$ denote the binary outcome. The outcome is observed only when the individual is accepted: $Y_t = 1$ denotes success and $Y_t = 0$ denotes failure. We assume $Y_t \perp\!\!\!\perp (A_t = 1) \mid X_t$ and that the covariates are i.i.d. over time with marginal distribution $p_0(x)$. The acceptance probability under policy Π is written as $\Pi(x) = \mathbb{P}(A_t = 1 \mid X_t = x)$.

Each individual contributes a decision-dependent reward. Let $c \in (0, 1)$ denote the cost incurred when an individual is accepted. The reward at time t is therefore $A_t(Y_t - c)$, giving the discounted nominal utility

$$\mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t A_t(Y_t - c)\right] = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t \Pi(X_t)\big(\rho(X_t) - c\big)\right], \tag{1}$$

where $\gamma \in (0, 1)$ is the discount factor and $\rho(x) = \mathbb{P}(Y_t = 1 \mid X_t = x)$ is the conditional success probability. The nominal objective is to choose a policy that maximizes equation 1.

We represent uncertainty about the conditional success probability using a Bayesian belief. For each covariate value $x \in \mathcal{X}$, let $\rho_0(x) \sim \mathrm{Beta}\big(\sigma_0(x), \nu_0(x) - \sigma_0(x)\big)$, where $\sigma_0(x)$ and $\nu_0(x)$ denote the initial pseudo-successes and pseudo-observations for state x. More generally, suppose that at time t, $\rho_t(x) \mid \mathcal{H}_t \sim \mathrm{Beta}\big(\sigma_t(x), \nu_t(x) - \sigma_t(x)\big)$, where $\mathcal{H}_t$ denotes the information available up to time t. By conjugacy, an observed outcome updates the sufficient statistics as $\sigma_{t+1}(x) = \sigma_t(x) + \mathbf{1}_{\{X_t = x\}} A_t Y_t$, $\nu_{t+1}(x) = \nu_t(x) + \mathbf{1}_{\{X_t = x\}} A_t$. The posterior mean is $\mu_t(x) = \frac{\sigma_t(x)}{\nu_t(x)}$. Thus, for each x, the pair $(\mu_t(x), \nu_t(x))$ summarizes the current belief about the success probability and the amount of accumulated information.

Throughout the analysis, we assume that the covariate space is finite, $|\mathcal{X}| < \infty$. The sequential decision problem can therefore be represented by a collection of state-specific belief processes $\{(\mu_t(x), \nu_t(x)) : x \in \mathcal{X}\}$. Accordingly, the policy may be written as $\Pi\big(x; \mu_t(x), \nu_t(x)\big)$, so that the acceptance decision for an individual with covariates x depends on the current belief associated with that state. This state-dependent formulation will be used in the robust analysis below.

2.2 Distributional Ambiguity Under Reject Inference

The belief dynamics in Section 2.1 describe the nominal setting in which the posterior mean $\mu_t(x)$ represents the conditional success probability. In deployment, however, the true success probability may deviate from this belief because of distribution shift, model misspecification, or changes in the underlying environment. Let $q_t(x) \in [0, 1]$ denote the true conditional success probability for covariate state x at time t.

Under reject inference, uncertainty in $q_t(x)$ affects both current reward and future learning. Conditional on acceptance, $q_t(x)$ determines the probability of observing a success or failure, and hence which posterior state is reached after the observation. Thus, robustness and information acquisition are coupled through the selective-feedback mechanism.

To capture deviations from the posterior belief, consider the Kullback–Leibler divergence between Bernoulli parameters $q, p \in [0, 1]$, i.e., $D(q, p) = q \log \frac{q}{p} + (1 - q) \log \frac{1-q}{1-p}$. Given a robustness radius $\eta \geq 0$, define the state-dependent ambiguity set $\mathcal{U}\big(\mu_t(x); \eta\big)$ where $\mathcal{U}\big(\mu_t(x); \eta\big) =$

$\{q_t(x) \in [0,1] : D(q_t(x), \mu_t(x)) \le \eta\}$. When $\eta = 0$, the ambiguity set contains only the nominal value $q_t(x) = \mu_t(x)$. A larger η allows larger deviations from the posterior belief.

For a policy Π, we define its robust value as the worst-case discounted utility over adaptive choices of state-specific success probabilities satisfying the ambiguity constraint:

$$V_\eta^\Pi = \inf_{\substack{q_t(x) \in \mathcal{U}(\mu_t(x);\eta) \\ t \ge 0,\ x \in \mathcal{X}}} \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t \Pi\big(X_t; \mu_t(X_t), \nu_t(X_t)\big)\big(q_t(X_t) - c\big)\right]. \tag{2}$$

The adversary is allowed to choose $q_t(x)$ adaptively based on the information available at time t, subject to the state-wise KL constraint. This stage-wise rectangular formulation permits the worst-case distribution to respond to the current belief state while preserving the sequential structure of the problem. As a result, the optimal robust policy is therefore defined by

$$\Pi_\eta^* \in \arg\max_{\Pi} V_\eta^\Pi. \tag{3}$$

The next section characterizes this optimization problem through a state-dependent Bellman equation. Because individuals with the same covariate value may arrive in different calendar periods, the derivation will first account for the random waiting time between successive visits to each state.

3 Robust Dynamic Programming under Reject Inference

3.1 State-Dependent Robust Bellman Characterization

We now characterize the optimal robust policy directly in the finite covariate space introduced in Section 2.1. The main difficulty is that individuals with the same covariate value x do not necessarily arrive in consecutive periods. Since the covariates are i.i.d. with marginal probability $p_0(x)$, the waiting time between two consecutive arrivals with covariate value x is geometrically distributed. This induces a state-dependent effective discount factor.

For a fixed $x \in \mathcal{X}$, let $K_x \in \{1, 2, \ldots\}$ denote the number of calendar periods between two consecutive arrivals with covariate value x. Then $K_x \sim \mathrm{Geometric}(p_0(x))$ and hence $\bar\gamma(x) := \mathbb{E}\left[\gamma^{K_x}\right] = \frac{\gamma p_0(x)}{1-\gamma(1-p_0(x))}$. Thus, conditional on successive visits to state x, the original calendar-time problem can be represented using the effective discount factor $\bar\gamma(x)$.

Consider a current belief state $(\mu_t(x), \nu_t(x))$. If the current individual with covariate value x is accepted, then the posterior state becomes

$$(\mu_t(x), \nu_t(x)) \longrightarrow \begin{cases} \left(\frac{\mu_t(x)\nu_t(x)+1}{\nu_t(x)+1}, \nu_t(x)+1\right), & \text{observed outcome is a success;} \\ \left(\frac{\mu_t(x)\nu_t(x)}{\nu_t(x)+1}, \nu_t(x)+1\right), & \text{observed outcome is a failure.} \end{cases}$$

Under the robust formulation, the true success probability $q_t(x)$ determines both the immediate reward and the probabilities of these two posterior transitions. In addition, for $x \in \mathcal{X}$, define the robust value function

$$\begin{aligned} &V_\eta^*(x; \mu, \nu) \\ &= \sup_{\Pi} \inf_{\substack{q_t(x) \in \mathcal{U}(\mu_t(x);\eta) \\ t \ge 0}} \mathbb{E}\left[\sum_{t=0}^{\infty} \bar\gamma(x)^t \Pi\big(x; \mu_t(x), \nu_t(x)\big)\big(q_t(x) - c\big) \,\middle|\, (\mu_0(x), \nu_0(x)) = (\mu, \nu)\right]. \end{aligned}$$

The following theorem gives the Bellman characterization of this value function and the corresponding optimal policy.

Theorem 1 (State-Dependent Robust Bellman Characterization)**.** *For every $x \in \mathcal{X}$, $\mu \in [0,1]$, and $\nu > 0$, define*

$$\mathcal{D}(x; \mu, \nu) = \bar\gamma(x)\left[V_\eta^*\left(x; \frac{\mu\nu+1}{\nu+1}, \nu+1\right) - V_\eta^*\left(x; \frac{\mu\nu}{\nu+1}, \nu+1\right)\right] - 1.$$

Then the optimal robust policy is deterministic and satisfies $\Pi_\eta^*(x; \mu, \nu) = \mathbf{1}\{V_\eta^*(x; \mu, \nu) > 0\}$, where $V_\eta^*(x; \mu, \nu)$ is given as follows:

Case $\mathcal{D}(x;\mu,\nu)=0$***:*** $V_\eta^*(x;\mu(x),\nu(x))=\max\left\{\bar\gamma(x)V_\eta^*\left(x;\frac{\mu(x)\nu(x)}{\nu(x)+1},\nu(x)+1\right)-c,0\right\}$.
Case $\mathcal{D}(x;\mu,\nu)\neq 0$***:*** *Define* $q^*(\lambda^*;x,\mu,\nu)$ *as*

$$q^*(\lambda^*;x,\mu,\nu)=\begin{cases}0, & \text{if } \mathcal{D}(x;\mu,\nu)<0,\ \eta\geq-\ln(1-\mu(x)),\\ 1, & \text{if } \mathcal{D}(x;\mu,\nu)>0,\ \eta\geq-\ln\mu(x),\\ \dfrac{\exp\left(\frac{\mathcal{D}(x;\mu,\nu)}{\lambda^*}+\ln\left(\frac{\mu(x)}{1-\mu(x)}\right)\right)}{1+\exp\left(\frac{\mathcal{D}(x;\mu,\nu)}{\lambda^*}+\ln\left(\frac{\mu(x)}{1-\mu(x)}\right)\right)}, & \text{if } \mathcal{D}(x;\mu,\nu)<0,\ \eta<-\ln(1-\mu(x))\\ & \text{or } \mathcal{D}(x;\mu,\nu)>0,\ \eta<-\ln\mu(x)\end{cases},$$

where in the last subcase $\lambda^*>0$ *is the solution of*

$$\eta=q^*(\lambda;x,\mu,\nu)\ln\left(\frac{q^*(\lambda;x,\mu,\nu)}{\mu(x)}\right)+(1-q^*(\lambda;x,\mu,\nu))\ln\left(\frac{1-q^*(\lambda;x,\mu,\nu)}{1-\mu(x)}\right).$$

Then $V_\eta^*(x;\mu(x),\nu(x))$ *satisfies the recursion*

$$V_\eta^*(x;\mu(x),\nu(x))=\max\left\{\begin{array}{l}(q^*(\lambda^*;x,\mu,\nu)-c)\\ \quad+\bar\gamma(x)q^*(\lambda^*;x,\mu,\nu)V_\eta^*\left(x;\frac{\mu(x)\nu(x)+1}{\nu(x)+1},\nu(x)+1\right)\\ \quad+\bar\gamma(x)(1-q^*(\lambda^*;x,\mu,\nu))V_\eta^*\left(x;\frac{\mu(x)\nu(x)}{\nu(x)+1},\nu(x)+1\right)\end{array},0\right\}. \quad (4)$$

The Bellman equation highlights the two effects of accepting an individual. The first is the immediate reward determined by the worst-case success probability $q_\eta^*(x;\mu,\nu)-c$. The second is the value of the information generated by observing the outcome, represented by the two continuation values corresponding to success and failure. Because the same $q_\eta^*(x;\mu,\nu)$ determines both the immediate payoff and the probabilities of the two posterior transitions, the robust policy accounts jointly for reward and future information.

The effective discount factor $\bar\gamma(x)$ further captures the frequency with which information about state x is acquired. States with smaller arrival probability $p_0(x)$ are revisited less frequently and therefore have a smaller effective discount factor. Consequently, the robust Bellman equation is state-dependent even when two states have the same current posterior belief. The resulting formulation provides the basis for the finite-horizon backward approximation developed in Section 3.2.

3.2 Tractable Backward Approximation

The state-dependent Bellman characterization in Section 3.1 is exact but infinite-horizon, while the belief count $\nu_t(x)$ is unbounded. Direct evaluation is therefore impractical. We obtain a tractable approximation by truncating the belief count at a finite level N and solving the resulting recursion backward.

For each $x\in\mathcal{X}$, let $\widetilde{V}_\eta^*(x;\mu(x))$ denote a stationary robust surrogate value that ignores further belief updates. Using the state-specific effective discount factor $\bar\gamma(x)$ from Section 3.1, we define $\widetilde{V}_\eta^*(x;\mu(x))=\frac{\max\left\{\inf_{D(q,\mu(x))\leq\eta}(q-c),0\right\}}{1-\bar\gamma(x)}$. We then use this stationary value as the terminal condition at $\nu(x)=N+1$, i.e., $V_\eta^N(x;\mu(x),N+1)=\widetilde{V}_\eta^*(x;\mu(x))$. Starting from this terminal condition, we compute $V_\eta^N(x;\mu(x),\nu(x))$ backward for $\nu(x)=N,N-1,\ldots,1$ using the Bellman recursion in Section 3.1, with V_η^* replaced by V_η^N. The resulting approximate acceptance policy is $\Pi_\eta^N(x;\mu(x),\nu(x))=\mathbf{1}_{\{V_\eta^N(x;\mu(x),\nu(x))>0\}}$.

For numerical implementation, we discretize the posterior mean on a finite grid $\{\mu_g\}_{g=1}^G$. For a grid point μ_g and belief count $\nu(x)$, accepting an individual with a successful outcome gives the successor mean $\mu_g^+=\frac{\mu_g\nu(x)+1}{\nu(x)+1}$, whereas a failed outcome gives $\mu_g^-=\frac{\mu_g\nu(x)}{\nu(x)+1}$. These successor means need not lie on the grid, so the corresponding continuation values are obtained by interpolation. The backward recursion is then evaluated for every $x\in\mathcal{X}$ and every grid point μ_g, using the state-specific effective discount $\bar\gamma(x)$. This yields the finite-state implementation summarized in Algorithm 1.

When the arrival probabilities $p_0(x)$ are unknown, we replace $\bar\gamma(x)$ by its empirical estimate $\widehat{\bar\gamma}(x)=\frac{\gamma\widehat{p}_0(x)}{1-\gamma(1-\widehat{p}_0(x))}$, where $\widehat{p}_0(x)$ is estimated from the observed covariate arrivals. Thus, the implemented

policy involves two approximations: truncating the infinite-horizon recursion and estimating the state-dependent effective discount factor. We quantify the resulting errors in Section 3.3.

3.3 Approximation and Estimation Guarantees

The backward recursion in Section 3.2 introduces two sources of approximation error: truncation of the infinite-horizon recursion and estimation of the state-dependent effective discount factor. We next quantify both errors. The first result follows from the contraction of the robust Bellman operator, while the second characterizes how uncertainty in the arrival process propagates to the robust value function.

Truncation error. Fix $x \in \mathcal{X}$. Since $\bar{\gamma}(x) < 1$, the state-specific Bellman operator is a contraction with modulus $\bar{\gamma}(x)$. Hence, the infinite-horizon value function is well defined, and errors in the terminal condition decay geometrically as the backward recursion moves away from the truncation boundary. The corresponding results are summarized in Proposition 2.

Proposition 2 (Contraction and truncation error)**.** *For each $x \in \mathcal{X}$, let $V_\eta^*(x; \mu(x), \nu(x))$ be the infinite-horizon robust value function and let $V_\eta^N(x; \mu(x), \nu(x))$ be the truncated value function defined in Section 3.2. Then, for any $\nu(x) \le N+1$, we have*

$$\left\| V_\eta^N(x;\cdot,\nu(x)) - V_\eta^*(x;\cdot,\nu(x)) \right\|_\infty \le \bar{\gamma}(x)^{N+1-\nu(x)} \left\| \widetilde{V}_\eta^*(x;\cdot) - V_\eta^*(x;\cdot,N+1) \right\|_\infty .$$

Since both the terminal surrogate and the optimal value are bounded by $(1-c)/(1-\bar{\gamma}(x))$,

$$\left\| V_\eta^N(x;\cdot,\nu(x)) - V_\eta^*(x;\cdot,\nu(x)) \right\|_\infty \le \frac{(1-c)\bar{\gamma}(x)^{N+1-\nu(x)}}{1-\bar{\gamma}(x)}.$$

Thus, the truncation error decreases geometrically in the number of backward steps between the current belief count and the truncation level. For a prescribed tolerance $\epsilon > 0$, it is sufficient to choose N such that $N \ge \nu(x) - 1 + \left\lceil \frac{\log\left(\frac{1-c}{\epsilon(1-\bar{\gamma}(x))}\right)}{-\log \bar{\gamma}(x)} \right\rceil$. This bound provides a direct criterion for selecting the truncation level, while the sensitivity of the numerical solution to N is examined empirically in the appendix.

Effective-discount estimation. The effective discount factor $\bar{\gamma}(x)$ depends on the state-arrival probability $p_0(x)$, which may be unknown in practice. Importantly, its estimation does not suffer from selective-label bias: the covariate state X_t is observed for every arrival, regardless of the acceptance decision. We therefore estimate $\bar{\gamma}(x)$ from the observed inter-arrival times of state x.

Let $I_j(x)$ denote the number of global arrival periods between the $(j-1)$-st and j-th occurrences of state x, and let m_x be the number of observed inter-arrival intervals for state x. Under the i.i.d. arrival assumption, $I_1(x), \dots, I_{m_x}(x)$ are i.i.d. geometric random variables with parameter $p_0(x)$. Since $\mathbb{E}\left[\gamma^{I_j(x)}\right] = \sum_{i=1}^{\infty} p_0(x)(1-p_0(x))^{i-1}\gamma^i = \bar{\gamma}(x)$, a natural method-of-moments estimator is $\widehat{\bar{\gamma}}(x) = \frac{1}{m_x}\sum_{j=1}^{m_x} \gamma^{I_j(x)}$. It is unbiased, i.e., $\mathbb{E}[\widehat{\bar{\gamma}}(x)] = \bar{\gamma}(x)$, and $0 \le \gamma^{I_j(x)} \le \gamma$. Hoeffding's inequality therefore gives the following concentration guarantee.

Proposition 3 (Estimation error of the effective discount)**.** *For any $x \in \mathcal{X}$ and $\alpha \in (0,1)$, we have $\mathbb{P}\left(|\widehat{\bar{\gamma}}(x) - \bar{\gamma}(x)| \ge \gamma\sqrt{\frac{\log(2/\alpha)}{2m_x}} \right) \le \alpha$. In addition, let $V_{\eta,\beta}^*(x; \mu(x), \nu(x))$ denote the optimal robust value function obtained from the state-specific Bellman recursion when the effective discount factor is fixed at $\beta \in [0, \gamma]$. Then, for any $\beta, \beta' \in [0,\gamma]$, we have $\left\| V_{\eta,\beta}^*(x;\cdot,\nu(x)) - V_{\eta,\beta'}^*(x;\cdot,\nu(x)) \right\|_\infty \le \frac{1-c}{(1-\gamma)^2}|\beta - \beta'|$. Thus, with probability at least $1-\alpha$,*

$$\left\| V_{\eta,\widehat{\bar{\gamma}}(x)}^*(x;\cdot,\nu(x)) - V_{\eta,\bar{\gamma}(x)}^*(x;\cdot,\nu(x)) \right\|_\infty \le \frac{(1-c)\gamma}{(1-\gamma)^2}\sqrt{\frac{\log(2/\alpha)}{2m_x}}.$$

The two results above isolate the two approximation mechanisms in the implemented DRRI policy. Truncation error decreases geometrically with N, whereas estimation error decreases at the standard

$m_x^{-1/2}$ rate and is larger for rarely observed states. More precisely, the total value-function error can be decomposed into

$$\begin{aligned}&\left\|V^N_{\eta,\widehat{\bar{\gamma}}(x)}(x;\cdot,\nu(x)) - V^*_{\eta,\bar{\gamma}(x)}(x;\cdot,\nu(x))\right\|_\infty \\ &\quad\le \left\|V^N_{\eta,\widehat{\bar{\gamma}}(x)}(x;\cdot,\nu(x)) - V^*_{\eta,\widehat{\bar{\gamma}}(x)}(x;\cdot,\nu(x))\right\|_\infty + \left\|V^*_{\eta,\widehat{\bar{\gamma}}(x)}(x;\cdot,\nu(x)) - V^*_{\eta,\bar{\gamma}(x)}(x;\cdot,\nu(x))\right\|_\infty .\end{aligned}$$

The first term is controlled by Proposition 2, while the second is controlled by Proposition 3. Detailed policy-stability results and simultaneous confidence bounds across states are deferred to Appendix B.

4 Simulation Experiments

4.1 Experiment Settings

We consider a selective feedback environment in which each arriving individual belongs to one state on a two-dimensional grid. In our implementation, each state x corresponds to one grid location (x_1, x_2), and arrivals are sampled i.i.d. from a non-uniform feature distribution $p(x)$. For each state x, the success probability $\rho(x) = \mathbb{P}\{Y = 1|X = x\}$ is also heterogeneous and is generated from a nonlinear spatial surface, obtained by combining global trends and local variations on the grid and mapping the result through a sigmoid transformation. Given the sampled state $X_t = x$ at time t, we have $Y_t \sim \mathrm{Bern}(\rho(x))$, implying that the outcome Y_t is drawn from a Bernoulli distribution with success probability $\rho(x)$. The details of the experiment setting are shown in Appendix C.3.

We compare the proposed method against a set of representative baselines, including a myopic decision rule (Greedy), randomized exploration methods (ϵ-Greedy (Langford & Zhang, 2007) and Thompson Sampling (TS) (Osband & Van Roy, 2015)), a stochastic policy-learning method (Consequential Learning (CL)(Kilbertus et al., 2020)), an optimistic selective-label method (R-osl (Bietti et al., 2021)), and a selective-label dynamic-programming method (SL-OPT (Wei, 2021)). Detailed descriptions and implementation settings are provided in Appendix C.2.

4.2 Total Discounted Rewards under Nominal Environment

We first examine the nominal performance before introducing the environmental shift. This experiment is intended to evaluate how effectively each policy balances immediate reward and information acquisition during training in the original environment. In particular, we vary the warm-start size B_0, which denotes the number of initial individuals that are always accepted so that their labels are observed. A larger B_0 provides more initial supervision, but it also incurs training-time cost because these forced acceptances may include unfavorable cases.

We evaluate each policy by its discounted cumulative reward over the training horizon:

$$R_{\mathrm{norm}}(\pi) = \frac{1}{M}\sum_{m=1}^{M}\sum_{t=1}^{T_{\mathrm{train}}} \gamma^{t-1} a_t^{(m)}\big(y_t^{(m)} - c\big),$$

where M is the number of independent runs, T_{train} is the training horizon, $a_t^{(m)} \in \{0, 1\}$ is the decision made by policy π at step t in run m.

Table 1 summarizes R_{norm} of various models with different values of B_0. When computing R_{norm}, we force $a_t^{(m)} = 1$ so that the corresponding labels are always observed for the first B_0 in each run m. We see that SL-OPT achieves the highest discounted reward for every value of B_0, with performance stable across the entire range. This result is consistent with the design of SL-OPT which directly exploits the dynamic-programming structure and is therefore highly effective when the training and deployment environments are aligned. Several other baselines benefit substantially from larger warm-start samples. Greedy, ϵ-Greedy, CL, and R-osl all improve monotonically as B_0 increases, which suggests that they rely more heavily on the quality of early supervised information.

The nominal discounted reward of DRRI is lower than that of SL-OPT and TS. This is expected, since DRRI is not designed to maximize best-case performance under the nominal environment; rather, it sacrifices part of the nominal utility to obtain stronger protection against adverse deployment shifts. At the same time, DRRI is less sensitive to changes in B_0, suggesting that the method does not depend heavily on an initial pool of labeled samples to stabilize its decisions.

Table 1: Discounted total reward R_{norm} of each model under nominal environment (mean $\pm$ std over 100 runs). The best result in each column is highlighted in bold.

model	$B_0 = 0$	$B_0 = 20$	$B_0 = 40$	$B_0 = 60$	$B_0 = 80$	$B_0 = 100$
Greedy	133.90 ± 20.12	135.66 ± 19.48	139.18 ± 19.01	143.82 ± 17.38	148.04 ± 15.41	150.74 ± 14.39
SL-OPT	**164.01 ± 10.44**	**164.12 ± 10.13**	**164.27 ± 10.02**	**164.31 ± 10.03**	**164.33 ± 10.04**	**164.53 ± 9.79**
TS	154.95 ± 9.84	155.56 ± 10.69	156.82 ± 10.35	157.89 ± 9.90	157.94 ± 9.59	159.30 ± 9.60
CL	128.24 ± 8.30	141.03 ± 12.57	149.12 ± 12.07	151.52 ± 11.55	153.40 ± 11.21	155.49 ± 9.71
ϵ-Greedy	137.08 ± 18.48	138.85 ± 17.60	142.14 ± 16.28	145.48 ± 15.60	149.11 ± 13.59	150.91 ± 13.39
R-osl	147.41 ± 16.44	150.39 ± 15.85	153.40 ± 14.68	156.53 ± 12.96	158.39 ± 12.75	159.75 ± 11.89
DRRI	151.17 ± 13.97	151.25 ± 14.11	151.38 ± 13.84	153.10 ± 13.08	154.01 ± 13.08	154.56 ± 12.56

4.3 Robustness Results under Adverse Environments

We next evaluate whether the learned policies remain effective when the deployment environment systematically shifts away from the training environment. The worst-environment experiment provides a distributionally robust assessment under an uncertainty set. It answers whether a policy can maintain utility under globally adverse, model-agnostic distribution shifts.

Starting from the nominal success surface $\rho(x)$, we construct an adverse evaluation environment $\rho_{\text{DRO}}(x; \eta_{eval})$ which equals the KL-based lower envelope of $\rho(x)$, i.e.,

$$\rho_{\text{DRO}}(x; \eta_{eval}) = \inf_{q \in [0,1]} \left\{ q : \mathbb{D}\big(\text{Bern}(q), \text{Bern}(\rho(x))\big) \leq \eta_{eval} \right\}.$$

As η_{eval} increases, the evaluation environment becomes more pessimistic since each conditional success probability is shifted downward in a distributionally robust manner. The evaluation robustness metric, denoted as $R_{\text{DRO}}(\pi; \eta_{eval})$, is the discounted cumulative reward obtained when a policy is deployed online in this worse environment with

$$R_{\text{DRO}}(\pi; \eta_{eval}) = \frac{1}{M} \sum_{m=1}^{M} \sum_{t=1}^{T_{\text{eval}}} \gamma^{t-1} a_t^{(m)} \big(y_t^{(m)} - c\big),$$

where $y_t^{(m)}$ is a realization sampled from $\text{Bern}\big(\rho_{\text{DRO}}(x_t^{(m)}; \eta_{eval})\big)$, $a_t^{(m)} \in \{0, 1\}$ is the decision made by policy π at step t in run m, and T_{eval} is the evaluation horizon. The performance across $M = 100$ independent runs with $B_0 = 0$ is reported in Table 2.

First, all methods exhibit declining performance as η_{eval} increases, which confirms that larger robustness radii indeed correspond to more challenging deployment conditions. Second, under mild or no degradation, CL achieves the highest reward at $\eta_{eval} = 0.00$, TS is best at $\eta_{eval} = 0.01$, and SL-OPT is best at $\eta_{eval} = 0.02$. This pattern is expected, since these methods are optimized for the nominal environment and therefore benefit when the test distribution remains close to training.

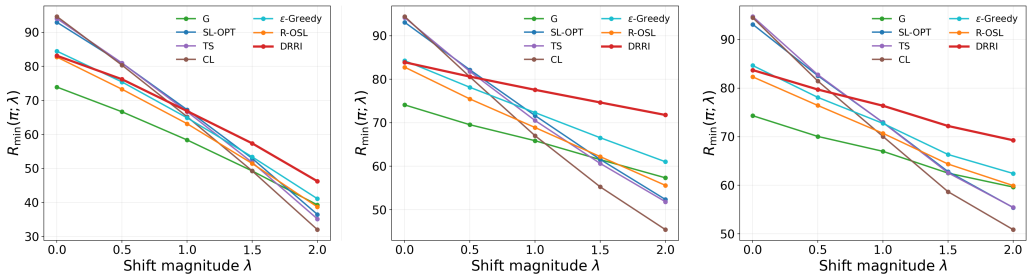
The advantage of DRRI emerges once the degradation becomes sufficiently severe. Beginning at $\eta_{eval} = 0.03$, DRRI becomes the best-performing method and maintains the lead throughout the remaining range of more adverse environments. Although DRRI does not maximize reward in the nominal setting, it degrades substantially more slowly as the environment worsens.

Table 2: Robust performance ($R_{\text{DRO}}(\pi; \eta_{eval})$) of each model under adverse environment (mean $\pm$ std over 100 runs). The best result in each column is highlighted in bold.

model	$\eta_{eval} = 0.00$	$\eta_{eval} = 0.01$	$\eta_{eval} = 0.02$	$\eta_{eval} = 0.03$	$\eta_{eval} = 0.04$	$\eta_{eval} = 0.05$	$\eta_{eval} = 0.06$
Greedy	84.84 ± 11.05	69.87 ± 8.97	62.76 ± 8.06	57.84 ± 7.49	52.97 ± 7.06	49.41 ± 6.85	45.03 ± 6.70
SL-OPT	106.52 ± 3.22	83.46 ± 2.46	**72.64 ± 2.23**	65.19 ± 2.79	58.07 ± 2.99	52.44 ± 3.25	46.04 ± 3.60
TS	108.26 ± 1.72	**83.79 ± 1.89**	72.37 ± 1.84	64.62 ± 2.16	57.16 ± 2.43	51.40 ± 2.71	44.73 ± 2.97
CL	**108.78 ± 1.51**	83.20 ± 1.99	71.35 ± 1.93	63.25 ± 2.47	55.42 ± 2.57	49.40 ± 2.68	42.36 ± 2.87
ϵ-Greedy	96.57 ± 6.57	78.09 ± 5.01	69.35 ± 4.44	63.34 ± 4.23	57.49 ± 4.17	53.08 ± 4.18	47.93 ± 4.47
R-osl	94.68 ± 8.11	76.34 ± 6.50	67.61 ± 5.81	61.55 ± 5.68	55.77 ± 5.30	51.35 ± 5.41	46.20 ± 5.49
DRRI	95.03 ± 6.58	78.72 ± 4.83	70.81 ± 4.14	**65.37 ± 3.87**	**60.29 ± 3.32**	**56.26 ± 3.35**	**51.68 ± 3.29**

4.4 Robustness Results under Perturbation Environment

We next examine robustness under structured perturbations that shift the deployment distribution according to pre-specified patterns rather than uniformly worsening it. By imposing different structured shifts, we examine which environmental changes are most harmful to each policy. Starting



(a) Global case (b) Boundary case (c) Subgroup case

Figure 1: Robust performance ($R_{\min}(\pi;\lambda,s)$) of each model across three cases.

from the nominal success surface $\rho(x)$, we construct a perturbed environment by modifying the log-odds of success as

$$\rho_{\lambda,s}(x) = \sigma(\mathrm{logit}(\rho(x)) - \lambda s(x)),$$

where $s(x)$ specifies the perturbation pattern and $\lambda \geq 0$ controls the shift magnitude. We consider three representative cases of $s(x)$: i) *global* case, ii) *boundary* case, and iii) *subgroup* case. Further elaboration on the construction of $s(x)$ for each case is deferred to Appendix C.4.

To evaluate robustness under structured perturbations, we adopt a worst-trajectory discounted reward metric. For a given perturbation pattern s and shift magnitude λ, we define

$$R_{\min}(\pi;\lambda,s) = \frac{1}{M}\sum_{m=1}^{M} \min_{\ell\in\{1,\dots,L\}} \sum_{t=1}^{T_{\mathrm{eval}}} \gamma^{t-1} a_t^{(m,\ell)} \big(y_t^{(m,\ell)} - c\big),$$

where $y_t^{(m,\ell)}$ is a realization drawn from $\mathrm{Bern}\big(\rho_{\lambda,s}(x_t^{(m,\ell)})\big)$, M is the number of independent training runs, L is total online evaluation trajectories for each run, T_{eval} is the evaluation horizon, $a_t^{(m,\ell)} \in \{0,1\}$ denotes the action taken by policy π at time t on evaluation trajectory ℓ of run m, and $\rho_{\lambda,s}(x)$ is the perturbed success probability under pattern s and perturbation magnitude λ.

Figure 1 shows that all methods deteriorate as the perturbation magnitude λ increases, confirming that larger shifts create more challenging deployment environments. However, the rate of deterioration differs substantially across methods and perturbation types. In the global case, nominally strong baselines such as TS and CL perform well when λ is small, but their rewards decline rapidly as the shift intensifies, whereas DRRI degrades more slowly and becomes the best-performing method in the high-shift regime. A similar pattern appears in the boundary case, where perturbations concentrate near the acceptance threshold and therefore most directly affect decision quality. Here, SL-OPT, TS, and CL suffer sharp performance drops as λ grows, while DRRI remains substantially more stable and achieves the highest reward throughout the moderate- and large-shift regimes. The subgroup case leads to the same overall conclusion. Although the perturbation is localized to a subset of the population, DRRI exhibits the slowest performance deterioration, whereas methods tuned mainly for nominal average performance decline much more sharply.

To further examine the numerical stability and robustness of the proposed method, we conduct sensitivity analyses on the ambiguity set training radius η_{train} in Appendix C.5.1, the cost c in Appendix C.5.2, the discount factor γ in Appendix C.5.3, and the truncation size N in Appendix C.5.4.

5 Case Study of Lending Decisions

We further evaluate the proposed framework on a real lending dataset using the FICO challenge data[1]. The FICO challenge dataset contains anonymized HELOC applications from real homeowners and is therefore well aligned with the rejection-inference setting in lending. Each record corresponds to a credit applicant together with a binary repayment outcome over a 24-month horizon, where *Good* denotes no episode of becoming more than 90 days past due and *Bad* denotes at

[1]FICO. Explainable Machine Learning Challenge, 2018. https://community.fico.com/s/ explainable-machine-learning-challenge?tabset-3158a=2.

Table 3: The performance metric R_{norm} across 1000 runs on the FICO dataset.

model	mean	std	min	$Q = 1\%$	$Q = 5\%$	$Q = 10\%$
Greedy	49.907	14.868	2.889	14.387	23.620	29.756
SL-OPT	57.587	7.861	25.957	37.740	44.832	47.305
TS	49.755	8.072	22.115	30.780	36.718	39.810
CL	33.277	7.136	9.829	17.120	21.532	23.983
ϵ-Greedy	51.528	13.397	11.304	19.659	28.213	33.804
R-osl	61.401	10.468	19.970	33.501	43.576	48.479
DRRI	59.353	8.262	31.061	38.257	45.858	48.762

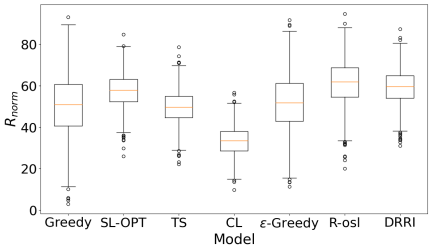


Figure 2: The box plot of R_{norm} across 1000 runs on the FICO dataset.

least one such episode. The dataset consists of 10,459 observations. To match the formulation of our method, we construct discrete borrower states from several important features.

We evaluate all methods in an online replay setting. In each run, the applicant records are presented sequentially according to a randomly generated arrival order, and the policy makes an accept/reject decision at each step. Consistent with the rejection-inference framework, the repayment outcome is revealed only when the applicant is accepted. For a given run r, we measure nominal performance by the discounted cumulative reward R_{norm}. We repeat this procedure for 1000 runs with different random seeds, and summarize the resulting distribution of R_{norm} using the mean, standard deviation, minimum, and lower quantiles.

The empirical results are reported in Table 3 and Figure 2. First, R-osl achieves the highest average reward, indicating strong nominal performance when evaluated purely by the center of the reward distribution. DRRI only attains the second-highest mean reward. However, DRRI provides the strongest downside protection. It achieves the highest minimum reward, 31.061, which is substantially above the corresponding minimum values of all baselines. The same pattern holds for the lower-tail quantiles where DRRI obtains the best 1%, 5%, and 10% quantiles. These improvements are particularly meaningful in lending applications, where the lower tail of the reward distribution reflects unfavorable deployment scenarios and captures the practical risk of poor decision sequences. By contrast, although R-osl delivers the highest mean reward, its minimum and lower quantiles are noticeably weaker than those of DRRI, suggesting that its stronger nominal average is achieved at the cost of greater downside exposure. Figure 2 further supports this conclusion. The box plot shows that DRRI maintains a relatively concentrated reward distribution while shifting the lower tail upward compared with the competing methods.

6 Limitations and Conclusion

This paper studies robust policy learning for the reject-inference problem under environmental shift. We propose a distributionally robust framework that explicitly accounts for uncertainty in the latent success probability and optimizes worst-case discounted utility rather than nominal performance alone. Within this framework, we characterize the resulting robust decision problem in both the covariate-free and covariate-dependent settings. In particular, we show that the optimal robust policy retains a threshold structure, and we develop a practical approximation procedure for computing the associated value function and policy.

Extensive simulation experiments and a real lending case study consistently demonstrate that methods that are strongest under the nominal environment can deteriorate quickly when the environment worsens, whereas the proposed approach sacrifices only a limited amount of nominal reward in exchange for substantially better stability and downside protection. These findings highlight the practical value of robust policy design in high-stakes applications such as lending, where even moderate misspecification can lead to large cumulative losses.

Several directions remain for future research. First, the current framework is limited to absent or relatively small finite covariate spaces; extending it to continuous or high-dimensional covariates is an important next step. Second, incorporating delayed outcome realization would better reflect applications such as lending and would require richer belief updates and policy optimization. Third, alternative ambiguity sets, such as Wasserstein-based formulations, may offer greater flexibility and numerical stability than KL divergence, particularly when μ_t is close to 0 or 1.

AI USE STATEMENT

In this work, we used generative AI tools for assisting with translation, polishing language, and providing feedback on research methodology or experiments. We have not used generative AI tools for developing theoretical models or conceptual frameworks, formulating mathematical claims, providing critical ingredients for proving mathematical claims, assisting in the writing of proofs, cleaning datasets, generating synthetic data sets, interpreting results, and implementing methods, and proposing or refining hypotheses and supporting qualitative and thematic data analysis are not applicable to this work. Additionally, we used generative AI tools for identifing relevant literature and summarizing or analysing existing literature. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work uses publicly available benchmark data and simulated environments for methodological evaluation. It does not involve human-subject experiments, real-time lending decisions, or the use of personally identifiable information. The proposed framework is intended for research purposes and should not be deployed in high-stakes decision systems without appropriate validation, oversight, and consideration of potential fairness and societal impacts.

REPRODUCIBILITY STATEMENT

We provide code, data-processing scripts, experiment configurations, random seeds, and commands for reproducing the reported results, tables, and figures. The supplement gives the proofs, model-selection procedure, data partitions, baseline specifications, and paired inference protocol.

REFERENCES

Alberto Bietti, Alekh Agarwal, and John Langford. A contextual bandit bake-off. *Journal of Machine Learning Research*, 22(133):1–49, 2021.

Ilija Bogunovic, Arpan Losalka, Andreas Krause, and Jonathan Scarlett. Stochastic linear bandits robust to adversarial attacks. In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pp. 991–999, 2021.

Trenton Chang and Jenna Wiens. From biased selective labels to pseudo-labels: An expectation-maximization framework for learning from biased decisions. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, 2024.

Yifang Chen, Chung-Wei Lee, Haipeng Luo, and Chen-Yu Wei. A new algorithm for non-stationary contextual bandits: Efficient, optimal and parameter-free. In *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pp. 696–726, 2019.

Amanda Coston, Ashesh Rambachan, and Alexandra Chouldechova. Characterizing fairness over the set of good models under selective labels. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 2144–2155, 2021.

Qin Ding, Cho-Jui Hsieh, and James Sharpnack. Robust stochastic linear contextual bandits under adversarial attacks. In *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pp. 7111–7123, 2022.

Ruijiang Gao, Mingzhang Yin, and Maytal Saar-Tsechansky. Sel-bald: Deep Bayesian active learning with selective labels. *Advances in Neural Information Processing Systems*, 37:40952–40975, 2024.

Vineet Goyal and Julien Grand-Clément. Robust markov decision processes: Beyond rectangularity. *Mathematics of Operations Research*, 48(1):203–226, 2023.

Garud N. Iyengar. Robust dynamic programming. *Mathematics of Operations Research*, 30(2): 257–280, 2005.

Nathan Kallus, Xiaojie Mao, Kaiwen Wang, and Zhengyuan Zhou. Doubly robust distributionally robust off-policy evaluation and learning. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 10598–10632, 2022.

David L. Kaufman and Andrew J. Schaefer. Robust modified policy iteration. *INFORMS Journal on Computing*, 25(3):396–410, 2013.

Vijay Keswani, Anay Mehrotra, and L. Elisa Celis. Fair classification with partial feedback: An exploration-based data collection approach. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 23547–23576. PMLR, 21–27 Jul 2024.

Niki Kilbertus, Manuel Gomez Rodriguez, Bernhard Schölkopf, Krikamol Muandet, and Isabel Valera. Fair decisions despite imperfect predictions. In *International Conference on Artificial Intelligence and Statistics*, pp. 277–287. PMLR, 2020.

Michael Jong Kim. Robust control of partially observable failing systems. *Operations Research*, 64 (4):999–1014, 2016.

Himabindu Lakkaraju, Jon Kleinberg, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. The selective labels problem: Evaluating algorithmic predictions in the presence of unobservables. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 275–284, 2017.

John Langford and Tong Zhang. The epoch-greedy algorithm for multi-armed bandits with side information. In J. Platt, D. Koller, Y. Singer, and S. Roweis (eds.), *Advances in Neural Information Processing Systems*, volume 20. Curran Associates, Inc., 2007. URL https://proceedings.neurips.cc/paper_files/paper/2007/file/4b04a686b0ad13dce35fa99fa4161c65-Paper.pdf.

Tong Li and Yisha Xiang. Distributionally robust partially observable markov decision process with distance-based ambiguity sets. *IISE Transactions*, 58(4):396–411, 2026.

Jingxian Liao, Wei Wang, Jason Xue, Anthony Lei, Xue Han, and Kun Lu. Combating sampling bias: A self-training method in credit risk models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 12566–12572, 2022.

Shiau Hong Lim, Huan Xu, and Shie Mannor. Reinforcement learning in robust markov decision processes. *Mathematics of Operations Research*, 41(4):1325–1353, 2016.

Qiang Liu, Yingtao Luo, Shu Wu, Zhen Zhang, Xiangnan Yue, Hong Jin, and Liang Wang. Rmt-net: Reject-aware multi-task network for modeling missing-not-at-random data in financial credit scoring. *IEEE Transactions on Knowledge and Data Engineering*, 35(7):7427–7439, 2022.

Haipeng Luo, Chen-Yu Wei, Alekh Agarwal, and John Langford. Efficient contextual bandits in non-stationary worlds. In *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pp. 1739–1776, 2018.

Shie Mannor, Ofir Mebel, and Huan Xu. Robust MDPs with k-rectangular uncertainty. *Mathematics of Operations Research*, 41(4):1484–1509, 2016.

Tong Mu, Yash Chandak, Tatsunori B Hashimoto, and Emma Brunskill. Factored DRO: Factored distributionally robust policies for contextual bandits. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 8318–8331. Curran Associates, Inc., 2022.

Hideaki Nakao, Ruiwei Jiang, and Siqian Shen. Distributionally robust partially observable markov decision process with moment-based ambiguity. *SIAM Journal on Optimization*, 31(1):461–488, 2021.

Arnab Nilim and Laurent El Ghaoui. Robust control of markov decision processes with uncertain transition matrices. *Operations Research*, 53(5):780–798, 2005.

Ian Osband and Benjamin Van Roy. Bootstrapped Thompson sampling and deep exploration. *arXiv preprint arXiv:1507.00300*, 2015.

Takayuki Osogami. Robust partially observable markov decision process. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 106–115. PMLR, 2015.

Miriam Rateike, Ayan Majumdar, Olga Mineeva, Krishna P. Gummadi, and Isabel Valera. Don't throw it away! the utility of unlabeled data in fair decision making. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pp. 1421–1433, 2022.

Soroush Saghafian. Ambiguous partially observable markov decision processes: Structural results and applications. *Journal of Economic Theory*, 178:1–35, 2018.

David Sculley. Practical learning from one-sided feedback. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 609–618, 2007.

Yi Shen, Pan Xu, and Michael Zavlanos. Wasserstein distributionally robust policy evaluation and learning for contextual bandits. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. Featured Certification.

Nian Si, Fan Zhang, Zhengyuan Zhou, and Jose Blanchet. Distributionally robust batch contextual bandits. *Management Science*, 69(10):5772–5793, 2023.

Andrew Smith and Charles Elkan. A Bayesian network framework for reject inference. In *Proceedings of the tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 286–295, 2004.

Jingyuan Wang, Zhimei Ren, Ruohan Zhan, and Zhengyuan Zhou. Distributionally robust policy learning under concept drifts. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 64244–64281, 2025.

Dennis Wei. Decision-making under selective labels: Optimal finite-domain policies and beyond. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 11035–11046, 2021.

Wolfram Wiesemann, Daniel Kuhn, and Berç Rustem. Robust markov decision processes. *Mathematics of Operations Research*, 38(1):153–183, 2013.

Huan Xu and Shie Mannor. Distributionally robust Markov decision processes. *Mathematics of Operations Research*, 37(2):288–300, 2012.

A Related Work

A.1 Reject Inference Problem

The reject inference studies can be divided into several lines. Early work treated the problem as biased sampling or one-sided feedback, proposing Bayesian-network and active-learning style corrections rather than direct policy optimization (Smith & Elkan, 2004; Sculley, 2007). The modern selective-label literature formalizes that observed labels are endogenous to historical decisions. For example, Lakkaraju et al. (2017) study evaluation under unobservables via contraction, while Coston et al. (2021) study fairness over the set of good models under selective labels when selection and outcome are unconfounded conditional on observed features.

A parallel line of work treats the problem primarily as one of prediction repair. For instance, Liao et al. (2022) use self-training to pseudo-label historically rejected applicants in order to mitigate sampling bias, Liu et al. (2022) model reject inference through multi-task learning by coupling approval/rejection and default prediction, and Chang & Wiens (2024) correct biased selective labels through an EM procedure.

More recent work argues that selective labels should be treated as a decision problem, not merely a prediction problem. Kilbertus et al. (2020) show that, under selective labels, learning to predict can be suboptimal for both fairness and utility, and they therefore advocate learning stochastic decision policies directly. In this way, Wei (2021) develops this perspective further in an online setting, where the learner balances current utility against future information and derives structured optimal policies for selective-label environments. Further, related work studies how exploration should be carried out under one-sided feedback, including fairness-constrained online learning (Rateike et al., 2022), exploration-based data collection (Keswani et al., 2024), and active learning when humans may refuse to label queried instances (Gao et al., 2024).

A.2 Robust Policy Learning

Robust policy learning is concerned with the problem that a policy optimized under a nominal training environment may deteriorate substantially after deployment when the environment is misspecified or shifts over time. At a foundational level, this literature originates in robust dynamic programming and robust Markov decision processes, where nominal value maximization is replaced by worst-case optimization over ambiguity sets for the transition law (Iyengar, 2005; Xu & Mannor, 2012).

In the offline-learning regime, robust policy learning has been developed primarily in contextual bandits and off-policy learning. Si et al. (2023) formulate offline policy learning through KL-based ambiguity sets to account for mismatch between the logging and deployment environments. Kallus et al. (2022) strengthen this framework by introducing doubly robust estimators for distributionally robust off-policy evaluation and learning. To reduce conservativeness, Mu et al. (2022) propose a factored DRO framework that separates uncertainty in the covariate distribution from uncertainty in the reward model. More recently, Shen et al. (2024) extend robustness beyond KL-based ambiguity by considering Wasserstein uncertainty sets to better handle support mismatch. Wang et al. (2025) study policy learning under concept drift, where only the conditional outcome model shifts, and derive a doubly robust procedure with statistical guarantees.

In the online-learning regime, the literature typically models robustness through either environmental nonstationarity or adversarial perturbation. In nonstationary contextual bandits, Luo et al. (2018) and Chen et al. (2019) develop algorithms with dynamic-regret guarantees when the reward process changes over time, thereby replacing the usual static benchmark by a time-varying one. A second online line studies adversarially corrupted bandits. Bogunovic et al. (2021) analyze stochastic linear bandits under adversarial reward corruption and provide algorithms with regret guarantees that degrade gracefully with the corruption budget. Ding et al. (2022) extend this perspective to stochastic linear contextual bandits under adaptive attacks on both rewards and contexts, thereby bringing robustness to a setting closer to contextual decision making.

A.3 Robust MDPs and POMDPs

Robust Markov decision processes (MDPs) address the sensitivity of sequential policies to misspecification of transition and reward models. Early studies establish that, under suitable rectangularity assumptions, model uncertainty can be incorporated into a minimax Bellman operator while preserving dynamic consistency and computational tractability (Iyengar, 2005; Nilim & El Ghaoui, 2005). Subsequent work extends this framework in several directions. Distributionally robust MDPs optimize against a family of probability distributions over uncertain model parameters, rather than a deterministic set of parameter realizations (Xu & Mannor, 2012). Data-driven formulations construct uncertainty sets from statistical confidence regions (Wiesemann et al., 2013), while robust modified policy iteration unifies robust value iteration and robust policy iteration (Kaufman & Schaefer, 2013). Other studies investigate how uncertainty sets can be learned or adapted through online interactions (Lim et al., 2016), and how dependence across state-wise transition uncertainties can be introduced without losing all of the tractability provided by rectangularity (Mannor et al., 2016; Goyal & Grand-Clément, 2023).

Partial observability introduces an additional layer of difficulty because policies must be defined over information states or beliefs rather than directly observed physical states. Model uncertainty may also affect both the distribution of future observations and the Bayesian transformation from the current belief to the next belief. Several robust and ambiguous POMDP formulations have been developed to address these challenges. Osogami (2015) studies POMDPs whose uncertain parameters belong to a prescribed region, establishes convexity of the robust value function over the belief space, and develops robust value-iteration and point-based value-iteration algorithms. Kim (2016) considers partially observable failing systems under relative-entropy model misspecification and derives an explicit adversarial response, convexity properties, and robust stopping regions. Saghafian (2018) introduces ambiguous POMDPs that distinguish model ambiguity from the decision maker's attitude toward ambiguity and establishes conditions for structural properties such as convexity, monotonicity, and control-limit policies.

Distributionally robust POMDPs place ambiguity on probability distributions governing the underlying model parameters. Nakao et al. (2021) consider moment-based ambiguity over the joint distribution of transition–observation probabilities. In their formulation, the realized transition–observation parameters can be inferred from side information after an action is taken. They show that the value function remains piecewise linear and convex and develop a heuristic-search value-iteration algorithm. More recently, Li & Xiang (2026) study distance-based ambiguity sets for the joint distributions of rewards and transition–observation probabilities. These distributionally robust formulations differ from classical robust POMDPs in that nature selects a distribution over model parameters rather than a single adverse parameter realization.

B TECHNICAL RESULTS

B.1 JUSTIFICATIONS OF EQUATION 1

Denote $S_T = \sum_{t=0}^{T} \gamma^t A_t(Y_t - c)$. Clearly, $\lim_{T\to\infty} S_T = S_\infty$. In addition, $|S_T| \le \sum_{t=0}^{T} |\gamma^t A_t(Y_t - c)| \le \sum_{t=0}^{\infty} \gamma^t(1+c) \le \frac{1+c}{1-\gamma}$. By the bounded convergence theorem, we have

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t A_t(Y_t - c)\right] &= \mathbb{E}\left[\lim_{T\to\infty} S_T\right] = \lim_{T\to\infty} \mathbb{E}[S_T] \\
&= \lim_{T\to\infty} \sum_{t=0}^{T} \mathbb{E}\big[\mathbb{E}[\gamma^t A_t(Y_t - c)|X_t]\big] \\
&= \lim_{T\to\infty} \sum_{t=0}^{T} \mathbb{E}\big[\gamma^t \mathbb{E}[(Y_t - c)|A_t = 1, X_t]\mathbb{E}[A_t = 1|X_t]\big] \\
&= \lim_{T\to\infty} \sum_{t=0}^{T} \mathbb{E}\Big[\gamma^t \Pi(X_t)\big(\mathbb{E}[Y_t|A_t = 1, X_t] - c\big)\Big] \\
&= \lim_{T\to\infty} \mathbb{E}\Big[\sum_{t=0}^{T} \gamma^t \Pi(X_t)\big(\mathbb{E}[Y_t|X_t] - c\big)\Big] \\
&\overset{*}{=} \mathbb{E}\Big[\lim_{T\to\infty} \sum_{t=0}^{T} \gamma^t \Pi(X_t)\big(\mathbb{E}[Y_t|X_t] - c\big)\Big] \\
&= \mathbb{E}\Big[\sum_{t=0}^{\infty} \gamma^t \Pi(X_t)\big(\mathbb{E}[Y_t|X_t] - c\big)\Big].
\end{aligned}$$

$\overset{*}{=}$ can be justified using the bounded convergence theorem similar to the case when justifying $\mathbb{E}\left[\lim_{T\to\infty} S_T\right] = \lim_{T\to\infty} \mathbb{E}[S_T]$.

B.2 PROOFS OF THEOREMS

We first present two lemmas which will help us prove Theorem 1 in the article.

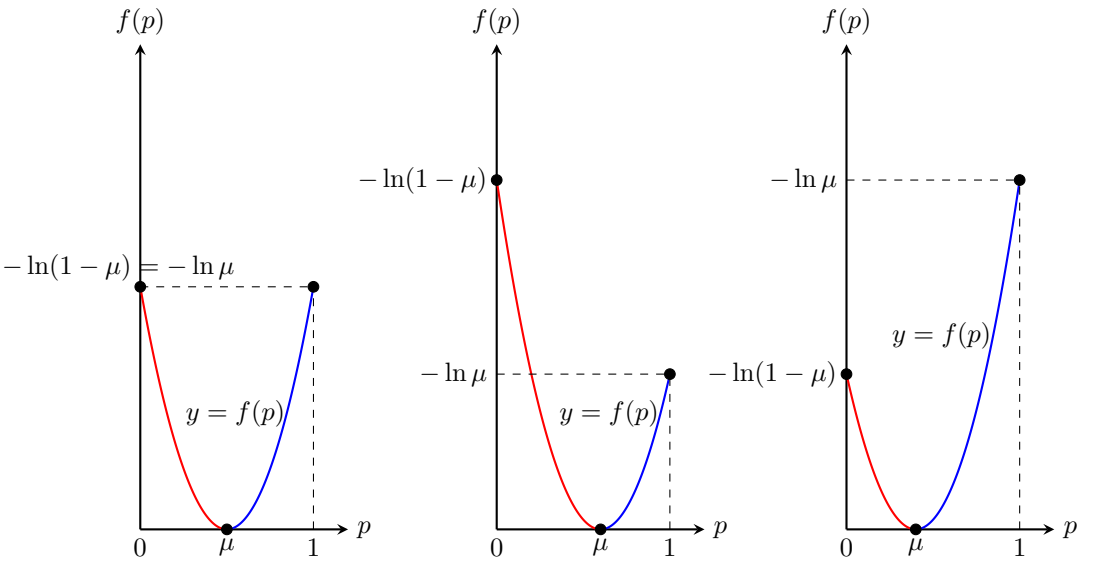
Lemma 4. *Given $\mu \in (0,1)$ and $\eta > 0$. Let $f(p) = p\ln\left(\frac{p}{\mu}\right) + (1-p)\ln\left(\frac{1-p}{1-\mu}\right)$ where $p \in (0,1)$. If $0 < \eta < \max\{-\ln\mu, -\ln(1-\mu)\}$, then there exists $\bar{p}$ such that $\eta = f(\bar{p})$.*

Proof. Note that $f(\mu) = 0$ by direct substitution. In addition, $f'(p) = \ln p + 1 - \ln\mu - \ln(1-p) - 1 + \ln(1-\mu) = \ln\left(\frac{p}{1-p}\right) - \ln\left(\frac{\mu}{1-\mu}\right)$. Applying first derivative test, we have:

- $f(p)$ is decreasing when $p < \mu$ since $f'(p) < 0$;
- $f(p)$ is increasing when $p > \mu$ since $f'(p) > 0$.

Also, simple limit computation gives that $f(p) \to -\ln(1-\mu)$ when $p \to 0$ and $f(p) \to -\ln\mu$ when $p \to 1$. Generally, there are three cases: Case 1: $-\ln(1-\mu) = -\ln\mu$; Case 2: $-\ln(1-\mu) > -\ln\mu$; Case 3: $-\ln(1-\mu) < -\ln\mu$.

Case 1: In this case, we can find two p (say $0 < p_1 < \mu$, $1 > p_2 > \mu$) such that $\eta = f(p_1) = f(p_2)$ when $0 < \eta \le -\ln(1-\mu) = -\ln\mu$ (see Figure 3(a), though the figure is not plotted under exact scaling).



(a) Case 1: $-\ln(1-\mu) = -\ln\mu$ (b) Case 2: $-\ln(1-\mu) > -\ln\mu$ (c) Case 3: $-\ln(1-\mu) < -\ln\mu$

Figure 3: Three possible plots of $f(p)$ against p according to the values of $-\ln(1-\mu)$ and $-\ln\mu$. Case 1: $-\ln(1-\mu) = -\ln\mu$; Case 2: $-\ln(1-\mu) > -\ln\mu$; Case 3: $-\ln(1-\mu) < -\ln\mu$.

Case 2: In this case, we can find a single $p \in (0, \mu)$ such that $\eta = f(p)$ when $-\ln\mu < \eta \leq -\ln(1-\mu)$, and we can find two p (say $0 < p_1 < \mu$, $1 > p_2 > \mu$) such that $\eta = f(p_1) = f(p_2)$ when $0 < \eta \leq -\ln\mu$ (see Figure 3(b), though the figure is not plotted under exact scaling).

Case 3: In this case, we can find a single $p \in (\mu, 1)$ such that $\eta = f(p)$ when $-\ln(1-\mu) < \eta \leq -\ln\mu$, and we can find two p (say $0 < p_1 < \mu$, $1 > p_2 > \mu$) such that $\eta = f(p_1) = f(p_2)$ when $0 < \eta \leq -\ln(1-\mu)$ (see Figure 3(c), though the figure is not plotted under exact scaling).

In any case, we are able to find $\bar{p}$ such that $\eta = f(\bar{p})$. □

Lemma 5. *Let* $M = \ln\left(\frac{\mu}{1-\mu}\right)$ *with* $\mu \in (0,1)$ *and* A, B *are two constants. Suppose that* $0 < \eta < \max\{-\ln\mu, -\ln(1-\mu)\}$ *and consider the following optimization:*

$$\begin{aligned} &\inf_{0\leq q_0\leq 1} && \{(q_0 - c) + Aq_0 + (1-q_0)B\} \\ &\textit{s.t.} && q_0 \ln\left(\frac{q_0}{\mu}\right) + (1-q_0)\ln\left(\frac{(1-q_0)}{(1-\mu)}\right) \leq \eta. \end{aligned} \tag{5}$$

Then we have the following:

Case i) the minimizer of Eq equation 5 can be any q_0 *satisfying* $q_0 \ln\left(\frac{q_0}{\mu}\right) + (1-q_0)\ln\left(\frac{(1-q_0)}{(1-\mu)}\right) \leq \eta$ *when* $B - A - 1 = 0$. *For example, we may choose* $q_0 = \mu$.

Case ii) the minimizer of Eq equation 5 is 0 *when* $B - A - 1 < 0$ *and* $\eta \geq -\ln(1-\mu)$, *and the minimizer of Eq equation 5 is* 1 *when* $B - A - 1 > 0$ *and* $\eta \geq -\ln\mu$.

Case iii) the minimizer of Eq equation 5 is $q_0^*(\lambda^*)$ *such that*

$$q_0^*(\lambda^*) = \frac{\exp\left(\frac{B-A-1}{\lambda^*} + M\right)}{1 + \exp\left(\frac{B-A-1}{\lambda^*} + M\right)}.$$

when $B-A-1<0$, $\eta<-\ln(1-\mu)$ or $B-A-1>0$, $\eta<-\ln\mu$. Here, λ^* is the zero of the following equation:

$$\eta = q_0^*(\lambda)\ln\left(\frac{q_0^*(\lambda)}{\mu}\right)+(1-q_0^*(\lambda))\ln\left(\frac{1-q_0^*(\lambda)}{1-\mu}\right).$$

Proof. The proof can be split into two cases:

Case i): In this case, Eq equation 5 becomes:

$$\begin{aligned}\inf_{0\le q_0\le 1}\quad & B-c\\ \text{s.t.}\quad & q_0\ln\left(\frac{q_0}{\mu}\right)+(1-q_0)\ln\left(\frac{(1-q_0)}{(1-\mu)}\right)\le\eta.\end{aligned}$$

The objective function is $B-c$ which is a constant. Clearly, the optimal solution should be any $q_0\in(0,1)$ such that $q_0\ln\left(\frac{q_0}{\mu}\right)+(1-q_0)\ln\left(\frac{(1-q_0)}{(1-\mu)}\right)\le\eta$ holds. In particular, we can set $q_0=\mu$ since $\mu\in(0,1)$ and $\mu\ln\left(\frac{\mu}{\mu}\right)+(1-\mu)\ln\left(\frac{(1-\mu)}{(1-\mu)}\right)=0\le\eta$.

Case ii): Let us first consider the case $B-A-1<0$, $\eta\ge-\ln(1-\mu)$. If $B-A-1<0$, then $A+1-B>0$ and the objective function in Eq equation 5 becomes $(q_0-c)+Aq_0+(1-q_0)B=(1+A-B)q_0+B-c$, which is an increasing function in q_0. As a result, the objective function attains its infimum when $q_0\to 0$. In addition, we need to check that when $q_0\to 0$, the constrained condition of Eq equation 5 holds. Indeed, note that $q_0\ln\left(\frac{q_0}{\mu}\right)+(1-q_0)\ln\left(\frac{1-q_0}{1-\mu}\right)\to-\ln(1-\mu)\le\eta$ when $q_0\to 0$, which implies that the constraint of Eq equation 5 holds. Therefore, we can set the minimizer as 0. The case for $B-A-1>0$, $\eta\ge-\ln\mu$ can be argued similarly.

Case iii): Let us write $f(q_0)=\{(q_0-c)+Aq_0+(1-q_0)B\}$ and $g(q_0)=q_0\ln\left(\frac{q_0}{\mu}\right)+(1-q_0)\ln\left(\frac{(1-q_0)}{(1-\mu)}\right)-\eta$. Using the Karush–Kuhn–Tucker condition, we have

$$\begin{aligned}\bigtriangledown f(q_0)+\lambda\bigtriangledown g(q_0)&=0\\ \lambda(q_0\ln\left(\frac{q_0}{\mu}\right)+(1-q_0)\ln\left(\frac{(1-q_0)}{(1-\mu)}\right)-\eta)&=0,\ \lambda\ge 0.\end{aligned}$$

Simple calculations give the followings:

$$\begin{aligned}(1+A-B)+\lambda\left(\ln(q_0)+q_0\left(\frac{1}{q_0}\right)-\ln\mu-\ln(1-q_0)+(1-q_0)\left(\frac{-1}{(1-q_0)}\right)+\ln(1-\mu)\right)&=0\\ \lambda(q_0\ln\left(\frac{q_0}{\mu}\right)+(1-q_0)\ln\left(\frac{(1-q_0)}{(1-\mu)}\right)-\eta)&=0.\end{aligned}$$

As a result, we have

$$(1+A-B)+\lambda\left(\ln\left(\frac{q_0}{1-q_0}\right)-\ln\left(\frac{\mu}{1-\mu}\right)\right)=0 \tag{6}$$

$$\lambda(q_0\ln\left(\frac{q_0}{\mu}\right)+(1-q_0)\ln\left(\frac{(1-q_0)}{(1-\mu)}\right)-\eta)=0. \tag{7}$$

Note that $\lambda\neq 0$, otherwise we have $1+A-B=0$ from Eq equation 6. Thus we have: $(q_0\ln\left(\frac{q_0}{\mu}\right)+(1-q_0)\ln\left(\frac{(1-q_0)}{(1-\mu)}\right)-\eta)=0$. Define $M=\ln\left(\frac{\mu}{1-\mu}\right)$. From Eq equation 6, we have

$$\frac{q_0}{1-q_0}=\exp\left(\frac{B-A-1}{\lambda}+M\right)\Rightarrow q_0=\frac{\exp\left(\frac{B-A-1}{\lambda}+M\right)}{1+\exp\left(\frac{B-A-1}{\lambda}+M\right)}.$$

Substituting the above results into Eq equation 7, we have

$$\frac{\exp\left(\frac{B-A-1}{\lambda}+M\right)}{1+\exp\left(\frac{B-A-1}{\lambda}+M\right)}\ln\left(\frac{\frac{\exp\left(\frac{B-A-1}{\lambda}+M\right)}{1+\exp\left(\frac{B-A-1}{\lambda}+M\right)}}{\mu}\right)+\frac{\ln\left(\frac{\frac{1}{1+\exp\left(\frac{B-A-1}{\lambda}+M\right)}}{(1-\mu)}\right)}{1+\exp\left(\frac{B-A-1}{\lambda}+M\right)}=\eta,$$

implying that λ must satisfy

$$\begin{aligned}\eta =& \frac{\exp\left(\frac{B-A-1}{\lambda}+M\right)}{1+\exp\left(\frac{B-A-1}{\lambda}+M\right)} \ln\left(\frac{\exp\left(\frac{B-A-1}{\lambda}+M\right)}{\mu(1+\exp\left(\frac{B-A-1}{\lambda}+M\right))}\right) \\ &+ \frac{1}{1+\exp\left(\frac{B-A-1}{\lambda}+M\right)} \ln\left(\frac{1}{(1-\mu)(1+\exp\left(\frac{B-A-1}{\lambda}+M\right))}\right)\end{aligned} \tag{8}$$

if minimizer of Eq equation 5 exists. According to the given condition $0 < \eta < \max\{-\ln(1-\mu), -\ln\mu\}$, we can find $\bar{p} \in (0,1)$ such that $\eta = f(\bar{p})$, where $f(p) = p\ln\left(\frac{p}{\mu}\right) + (1-p)\ln\left(\frac{(1-p)}{(1-\mu)}\right)$ based on Lemma 4. Now, let us set $p(\lambda) = \frac{\exp\left(\frac{B-A-1}{\lambda}+M\right)}{1+\exp\left(\frac{B-A-1}{\lambda}+M\right)}$. Eq equation 8 would change to: $\eta = f(p(\lambda))$. Thus, our goal is to find $\lambda \geq 0$ such that $p(\lambda) = \bar{p}$. After conducting simple calculations, we have:

$$\bar{p} = \frac{\exp\left(\frac{B-A-1}{\lambda}+M\right)}{1+\exp\left(\frac{B-A-1}{\lambda}+M\right)} \Rightarrow \lambda = \frac{B-A-1}{\ln\left(\frac{\bar{p}}{1-\bar{p}}\right) - \ln\left(\frac{\mu}{1-\mu}\right)}.$$

Let us consider two scenarios.

Scenario 1 $B-A-1 > 0$**,** $\eta < -\ln\mu$**:** In this scenario, $\lambda > 0$ if and only if $1 > \bar{p} > \mu$. According to Figures 3(a) - 3(c), we can find $\bar{p} \in (\mu, 1)$.

Scenario 2 $B-A-1 < 0$**,** $\eta < -\ln(1-\mu)$**:** In this scenario, $\lambda > 0$ if and only if $0 < \bar{p} < \mu$. According to Figures 3(a) - 3(c), we can find $\bar{p} \in (0, \mu)$.

Thus, we can find $\lambda > 0$ such that Eq equation 8 holds. Denote that positive λ satisfying Eq equation 8 as λ^*.

Since $0 \leq q_0 \leq 1$, the target minimizer is $q_0^*(\lambda^*)$ such that

$$q_0^*(\lambda^*) = \frac{\exp\left(\frac{B-A-1}{\lambda^*}+M\right)}{1+\exp\left(\frac{B-A-1}{\lambda^*}+M\right)}, \quad \text{where } \lambda^* \text{ satisfies the following equation:}$$

$$\begin{aligned}\eta =& \frac{\exp\left(\frac{B-A-1}{\lambda^*}+M\right)}{1+\exp\left(\frac{B-A-1}{\lambda^*}+M\right)} \ln\left(\frac{\exp\left(\frac{B-A-1}{\lambda^*}+M\right)}{\mu(1+\exp\left(\frac{B-A-1}{\lambda^*}+M\right))}\right) \\ &+ \frac{1}{1+\exp\left(\frac{B-A-1}{\lambda^*}+M\right)} \ln\left(\frac{1}{(1-\mu)(1+\exp\left(\frac{B-A-1}{\lambda^*}+M\right))}\right).\end{aligned}$$

□

B.3 Proof of Theorem 1

We prove Theorem 1 by conditioning on successive visits to each covariate state $x \in \mathcal{X}$. Fix $x \in \mathcal{X}$ and consider the subsequence of decision times at which $X_t = x$. Since the covariates are i.i.d. with probability $p_0(x)$, the number of time steps between two consecutive visits to x follows a geometric distribution. If $I(x)$ denotes the number of time steps from one visit to x to the next, then

$$\mathbb{P}(I(x) = i) = p_0(x)\big(1-p_0(x)\big)^{i-1}, \qquad i = 1, 2, \ldots.$$

Consequently, the discounted continuation value between two successive visits to x is characterized by the effective discount factor

$$\bar{\gamma}(x) = \mathbb{E}\Big[\gamma^{I(x)}\Big] = \sum_{i=1}^{\infty} \gamma^i p_0(x)\big(1-p_0(x)\big)^{i-1} = \frac{\gamma p_0(x)}{1-\gamma(1-p_0(x))}.$$

For a belief state $(x; \mu(x), \nu(x))$, accepting an individual produces either a success or a failure. The corresponding posterior mean states are $\mu^+(x) = \frac{\mu(x)\nu(x)+1}{\nu(x)+1}$, $\mu^-(x) = \frac{\mu(x)\nu(x)}{\nu(x)+1}$, with

the belief count updated from $\nu(x)$ to $\nu(x)+1$. Accordingly, define the successor belief states $s^+(x;\mu(x),\nu(x)) = \Big(x; \frac{\mu(x)\nu(x)+1}{\nu(x)+1}, \nu(x)+1\Big)$, and $s^-(x;\mu(x),\nu(x)) = \Big(x; \frac{\mu(x)\nu(x)}{\nu(x)+1}, \nu(x)+1\Big)$.

Let $V(x;\mu(x),\nu(x))$ be any bounded continuation-value function. Conditional on accepting and on a success probability $q_t(x)$, the corresponding robust continuation value is

$$q_t(x) - c + \bar{\gamma}(x)\left[q_t(x)V\big(s^+(x;\mu(x),\nu(x))\big) + \big(1-q_t(x)\big)V\big(s^-(x;\mu(x),\nu(x))\big)\right].$$

The adversary chooses $q_t(x)$ from the state-dependent ambiguity set

$$U(\mu(x);\eta) = \{q_t(x) \in [0,1] : D(q_t(x),\mu(x)) \le \eta\}.$$

Thus, the worst-case value of accepting is

$$Q_\eta(x;\mu(x),\nu(x);V) = \inf_{q_t(x)\in U(\mu(x);\eta)} \left\{ \begin{matrix} q_t(x) - c + \bar{\gamma}(x)q_t(x)V(s^+(x;\mu(x),\nu(x))) \\ +\bar{\gamma}(x)(1-q_t(x))V(s^-(x;\mu(x),\nu(x))) \end{matrix} \right\}.$$

Rejecting produces no immediate reward and no posterior update, so its value is zero.

To characterize the minimizing $q_t(x)$, define

$$D_\eta(x;\mu(x),\nu(x);V) = \bar{\gamma}(x)V\big(s^-(x;\mu(x),\nu(x))\big) - \bar{\gamma}(x)V\big(s^+(x;\mu(x),\nu(x))\big) - 1.$$

The objective inside the infimum in equation 9 is affine in $q_t(x)$ and can be written as

$$-c + \bar{\gamma}(x)V\big(s^-(x;\mu(x),\nu(x))\big) - q_t(x)D_\eta(x;\mu(x),\nu(x);V).$$

Hence, when $D_\eta(x;\mu(x),\nu(x);V) < 0$, the adversary minimizes the objective by choosing the largest feasible value of $q_t(x)$, whereas when $D_\eta(x;\mu(x),\nu(x);V) > 0$, it chooses the smallest feasible value.

The extremal feasible values are determined by the Bernoulli KL constraint. In particular,

$$D(0,\mu(x)) = -\log\big(1-\mu(x)\big), \qquad D(1,\mu(x)) = -\log\mu(x).$$

Therefore, if $D_\eta(x;\mu(x),\nu(x);V) < 0$ and $\eta \ge -\log\big(1-\mu(x)\big)$, the worst-case choice is $q_t^*(x) = 0$. Similarly, if $D_\eta(x;\mu(x),\nu(x);V) > 0$ and $\eta \ge -\log\mu(x)$, the worst-case choice is $q_t^*(x) = 1$.

Otherwise, the KL constraint is active and the minimizing $q_t^*(x)$ lies in the interior of $(0,1)$. Introducing a Lagrange multiplier $\lambda > 0$ for the KL constraint, the first-order condition gives

$$\log\frac{q_t^*(x)}{1-q_t^*(x)} = \log\frac{\mu(x)}{1-\mu(x)} + \frac{D_\eta(x;\mu(x),\nu(x);V)}{\lambda}.$$

Equivalently,

$$q_t^*(x;\lambda) = \frac{\exp\Big(D_\eta(x;\mu(x),\nu(x);V)/\lambda + \log\frac{\mu(x)}{1-\mu(x)}\Big)}{1+\exp\Big(D_\eta(x;\mu(x),\nu(x);V)/\lambda + \log\frac{\mu(x)}{1-\mu(x)}\Big)}.$$

The multiplier λ is chosen so that $D(q_t^*(x;\lambda),\mu(x)) = \eta$ whenever the KL constraint is active. This characterizes the worst-case success probability for every belief state.

We can therefore write the robust Bellman operator for state x as $(\mathcal{T}_{\eta,x}V)\big(x;\mu(x),\nu(x)\big) = \max\big\{0,\, Q_\eta\big(x;\mu(x),\nu(x);V\big)\big\}$. The optimal value function is the fixed point of this operator, $V_\eta^* = \mathcal{T}_\eta V_\eta^*$, which gives the Bellman characterization in Theorem 1. In particular,

$$V_\eta^*\big(x;\mu(x),\nu(x)\big) = \max\left\{ \inf_{q_t(x)\in U(\mu(x);\eta)} \left[\begin{matrix} q_t(x) - c + \bar{\gamma}(x)q_t(x)V_\eta^*\big(s^+(x;\mu(x),\nu(x))\big) \\ +\bar{\gamma}(x)(1-q_t(x))V_\eta^*\big(s^-(x;\mu(x),\nu(x))\big) \end{matrix} \right],\ 0 \right\}.$$

Finally, since rejection has value zero, the optimal robust policy accepts precisely when the robust value of acceptance is positive. Hence

$$\Pi_\eta^*\big(x;\mu(x),\nu(x)\big) = \mathbb{I}\left\{V_\eta^*\big(x;\mu(x),\nu(x)\big) > 0\right\},$$

which completes the proof.

B.4 Proof of the Truncation Guarantee

We prove the truncation result stated in Proposition 2. The proof is carried out state-wise for each $x \in \mathcal{X}$, since the state-dependent Bellman recursion is characterized by the effective discount factor $\bar{\gamma}(x)$.

Fix $x \in \mathcal{X}$. For notational clarity within this proof, define the successor belief states

$$s_\nu^+(x;\mu(x)) = \left(\frac{\mu(x)\nu(x)+1}{\nu(x)+1}, \nu(x)+1\right), \qquad s_\nu^-(x;\mu(x)) = \left(\frac{\mu(x)\nu(x)}{\nu(x)+1}, \nu(x)+1\right).$$

For any bounded value function V and any $q \in U(\mu(x);\eta)$, let

$$F_{q,x}(V;\mu(x),\nu(x)) = q - c + \bar{\gamma}(x) q V\left(s_\nu^+(x;\mu(x))\right) + \bar{\gamma}(x)(1-q) V\left(s_\nu^-(x;\mu(x))\right).$$

The state-dependent robust Bellman operator can then be written as

$$(\mathcal{T}_x V)(\mu(x),\nu(x)) = \max\left\{0, \inf_{q \in U(\mu(x);\eta)} F_{q,x}(V;\mu(x),\nu(x))\right\}.$$

For any bounded functions V and W, the difference between the corresponding continuation terms is bounded by

$$\begin{aligned}|F_{q,x}(V;\mu(x),\nu(x)) - F_{q,x}(W;\mu(x),\nu(x))| &\le \bar{\gamma}(x)\left[q\|V-W\|_\infty + (1-q)\|V-W\|_\infty\right] \\ &= \bar{\gamma}(x)\|V-W\|_\infty.\end{aligned}$$

Taking the infimum over q and then the maximum with zero preserves this bound. Hence,

$$\|\mathcal{T}_x V - \mathcal{T}_x W\|_\infty \le \bar{\gamma}(x)\|V-W\|_\infty.$$

Since

$$0 \le \bar{\gamma}(x) = \frac{\gamma p_0(x)}{1-\gamma(1-p_0(x))} < 1,$$

the Bellman operator is a contraction. Therefore, it admits a unique bounded fixed point, which is the state-specific robust value function $V_\eta^*(x;\mu(x),\nu(x))$ characterized in Section 3.1.

We next consider the truncated recursion. Let $V_\eta^N(x;\mu(x),N+1) = \widetilde{V}_\eta^*(x;\mu(x))$ be the terminal condition used in Section 3.2. At each preceding belief level, V_η^N is obtained by applying $\mathcal{T}_x$ once. Thus, for $\nu(x) \le N$,

$$V_\eta^N(x;\cdot,\nu(x)) = \mathcal{T}_x V_\eta^N(x;\cdot,\nu(x)+1),$$

whereas the exact value function satisfies

$$V_\eta^*(x;\cdot,\nu(x)) = \mathcal{T}_x V_\eta^*(x;\cdot,\nu(x)+1).$$

Consequently,

$$\left\|V_\eta^N(x;\cdot,\nu(x)) - V_\eta^*(x;\cdot,\nu(x))\right\|_\infty \le \bar{\gamma}(x)\left\|V_\eta^N(x;\cdot,\nu(x)+1) - V_\eta^*(x;\cdot,\nu(x)+1)\right\|_\infty.$$

Iterating this inequality from $\nu(x)$ to the truncation boundary $N+1$ gives

$$\left\|V_\eta^N(x;\cdot,\nu(x)) - V_\eta^*(x;\cdot,\nu(x))\right\|_\infty \le \bar{\gamma}(x)^{N+1-\nu(x)}\left\|\widetilde{V}_\eta^*(x;\cdot) - V_\eta^*(x;\cdot,N+1)\right\|_\infty.$$

This proves the first bound in Proposition 2.

Finally, the one-period reward is bounded above by $1-c$, so the discounted robust value satisfies

$$0 \le V_\eta^*(x;\mu(x),\nu(x)) \le \sum_{k=0}^{\infty} \bar{\gamma}(x)^k (1-c) = \frac{1-c}{1-\bar{\gamma}(x)}.$$

The terminal surrogate used in Section 3.2 satisfies the same upper bound. Therefore,

$$\left\|\widetilde{V}_\eta^*(x;\cdot) - V_\eta^*(x;\cdot,N+1)\right\|_\infty \le \frac{1-c}{1-\bar{\gamma}(x)},$$

and hence

$$\left\|V_\eta^N(x;\cdot,\nu(x)) - V_\eta^*(x;\cdot,\nu(x))\right\|_\infty \le \frac{(1-c)\bar{\gamma}(x)^{N+1-\nu(x)}}{1-\bar{\gamma}(x)}.$$

This completes the proof.

B.5 Proof of the Effective-Discount Estimation Guarantee

We prove Proposition 3, which quantifies the statistical error from estimating the state-specific effective discount factor and its propagation to the robust value function.

Fix $x \in \mathcal{X}$. Let $I_j(x)$ denote the number of global arrival periods between the $(j-1)$-st and j-th occurrences of state x. Under the i.i.d. covariate-arrival assumption, $I_1(x), I_2(x), \ldots$ are i.i.d. geometric random variables with parameter $p_0(x)$ and support $\{1, 2, \ldots\}$. Therefore,

$$\mathbb{E}\Big[\gamma^{I_j(x)}\Big] = \sum_{i=1}^{\infty} \gamma^i p_0(x)\big(1-p_0(x)\big)^{i-1} = \frac{\gamma p_0(x)}{1-\gamma(1-p_0(x))} = \bar{\gamma}(x).$$

Thus, the empirical estimator used in Section 3.2, $\widehat{\bar{\gamma}}(x) = \frac{1}{m_x}\sum_{j=1}^{m_x} \gamma^{I_j(x)}$, is unbiased, i.e., $\mathbb{E}\big[\widehat{\bar{\gamma}}(x)\big] = \bar{\gamma}(x)$. Moreover, because $0 \le \gamma^{I_j(x)} \le \gamma < 1$, Hoeffding's inequality yields, for every $\varepsilon > 0$, $\mathbb{P}\big(\big|\widehat{\bar{\gamma}}(x) - \bar{\gamma}(x)\big| \ge \varepsilon\big) \le 2\exp\Big(-\frac{2m_x\varepsilon^2}{\gamma^2}\Big)$. Equivalently, with probability at least $1-\alpha$,

$$\big|\widehat{\bar{\gamma}}(x) - \bar{\gamma}(x)\big| \le \gamma\sqrt{\frac{\log(2/\alpha)}{2m_x}}.$$

We next quantify how this estimation error affects the robust value function. For any effective discount factor $\beta \in [0, \gamma]$, define $V^*_{\eta,\beta}(x; \mu(x), \nu(x))$ as the optimal robust value function obtained by replacing $\bar{\gamma}(x)$ in the Bellman recursion with β. Its Bellman operator is

$$(\mathcal{T}_{x,\beta}V)(\mu(x), \nu(x)) = \max\left\{0, \inf_{q \in U(\mu(x);\eta)} \big[q - c + \beta q V(s^+_\nu(x;\mu(x))) + \beta(1-q)V(s^-_\nu(x;\mu(x)))\big]\right\}.$$

For any bounded functions V and W,

$$\|\mathcal{T}_{x,\beta}V - \mathcal{T}_{x,\beta}W\|_\infty \le \beta\|V - W\|_\infty,$$

so $\mathcal{T}_{x,\beta}$ is a contraction for every $\beta \in [0, \gamma]$.

Let $\beta, \beta' \in [0, \gamma]$. For any bounded V, the difference between the two Bellman operators satisfies

$$\begin{aligned}
&\|\mathcal{T}_{x,\beta}V - \mathcal{T}_{x,\beta'}V\|_\infty \\
&\quad\le |\beta - \beta'|\,\big\|qV(s^+_\nu(x;\mu(x))) + (1-q)V(s^-_\nu(x;\mu(x)))\big\|_\infty \\
&\quad\le |\beta - \beta'|\|V\|_\infty.
\end{aligned}$$

The robust value is bounded by $\|V^*_{\eta,\beta}\|_\infty \le \frac{1-c}{1-\beta}$. Using the fixed-point equations for $V^*_{\eta,\beta}$ and $V^*_{\eta,\beta'}$ together with the contraction property gives

$$\big\|V^*_{\eta,\beta} - V^*_{\eta,\beta'}\big\|_\infty \le \frac{|\beta-\beta'|}{1-\beta}\big\|V^*_{\eta,\beta'}\big\|_\infty \le \frac{1-c}{(1-\beta)(1-\beta')}|\beta-\beta'|.$$

Since $\beta, \beta' \le \gamma$,

$$\big\|V^*_{\eta,\beta} - V^*_{\eta,\beta'}\big\|_\infty \le \frac{1-c}{(1-\gamma)^2}|\beta-\beta'|.$$

Taking $\beta = \widehat{\bar{\gamma}}(x)$ and $\beta' = \bar{\gamma}(x)$ gives

$$\Big\|V^*_{\eta,\widehat{\bar{\gamma}}(x)} - V^*_{\eta,\bar{\gamma}(x)}\Big\|_\infty \le \frac{1-c}{(1-\gamma)^2}\,\big|\widehat{\bar{\gamma}}(x) - \bar{\gamma}(x)\big|.$$

Combining this inequality with the concentration bound above yields, with probability at least $1-\alpha$,

$$\Big\|V^*_{\eta,\widehat{\bar{\gamma}}(x)} - V^*_{\eta,\bar{\gamma}(x)}\Big\|_\infty \le \frac{(1-c)\gamma}{(1-\gamma)^2}\sqrt{\frac{\log(2/\alpha)}{2m_x}}.$$

The preceding bound also gives a direct policy-stability condition. Define the robust acceptance advantage under effective discount β by

$$A_{\eta,\beta}(x; \mu(x), \nu(x)) = \inf_{q \in U(\mu(x);\eta)} \big[q - c + \beta q V^*_{\eta,\beta}(x; s^+_\nu(x;\mu(x))) + \beta(1-q)V^*_{\eta,\beta}(x; s^-_\nu(x;\mu(x)))\big].$$

The reject action has value zero, so the optimal action accepts whenever this advantage is positive.

By the same perturbation argument used for the value functions,

$$|A_{\eta,\beta}(x;\mu(x),\nu(x)) - A_{\eta,\beta'}(x;\mu(x),\nu(x))| \le \frac{1-c}{(1-\beta)(1-\beta')}|\beta-\beta'|.$$

In particular,

$$|A_{\eta,\beta}(x;\mu(x),\nu(x)) - A_{\eta,\beta'}(x;\mu(x),\nu(x))| \le \frac{1-c}{(1-\gamma)^2}|\beta-\beta'|.$$

Hence, whenever

$$\left|A_{\eta,\bar{\gamma}(x)}(x;\mu(x),\nu(x))\right| > \frac{1-c}{(1-\gamma)^2}\left|\widehat{\bar{\gamma}}(x)-\bar{\gamma}(x)\right|,$$

the estimated effective discount factor cannot change the sign of the acceptance advantage, and therefore the estimated and oracle policies select the same action at that belief state.

Combining this condition with the concentration inequality gives, with probability at least $1-\alpha$, policy agreement at every state satisfying

$$\left|A_{\eta,\bar{\gamma}(x)}(x;\mu(x),\nu(x))\right| > \frac{(1-c)\gamma}{(1-\gamma)^2}\sqrt{\frac{\log(2/\alpha)}{2m_x}}.$$

Thus, estimation error can affect the policy only near the accept–reject boundary, while states with a sufficiently large decision margin remain stable.

Finally, unlike outcome estimation, estimation of $\bar{\gamma}(x)$ does not suffer from selective-label bias because X_t is observed for every arrival, regardless of whether the individual is accepted. The main statistical difficulty therefore arises for rare states, for which the number m_x of observed inter-arrival intervals is small.

C Experiments Details

C.1 Data Generation Process Details

We construct a finite-state data-generating process (DGP) to evaluate policy learning under selective feedback in a controlled yet heterogeneous environment. The state space is defined on a two-dimensional grid

$$\mathcal{X} = \{(x_1,x_2) : x_1 = 0,\dots,W-1,\ x_2 = 0,\dots,H-1\},$$

where each grid location represents one discrete covariate state. The total number of states is therefore $|\mathcal{X}| = WH$. In the default implementation, $W = H = 5$, yielding 25 finite states. At each time t, an individual arrives with state $X_t \in \mathcal{X}$, and the sequence of arrivals is generated independently from a non-uniform state distribution $p_0(x)$.

To create heterogeneous arrival frequencies across states, we define an unnormalized multimodal surface

$$\begin{aligned}\widetilde{p}_0(x_1,x_2) = 0.20 &+ 1.10\exp\left\{-\frac{(x_1-2.0)^2+(x_2-4.5)^2}{8.0}\right\}\\ &+ 0.95\exp\left\{-\frac{(x_1-7.0)^2+(x_2-7.0)^2}{7.0}\right\}\\ &+ 0.65\exp\left\{-\frac{(x_1-6.5)^2+(x_2-2.0)^2}{5.0}\right\}.\end{aligned}$$

The arrival probability is then obtained by normalization:

$$p_0(x_1,x_2) = \frac{\widetilde{p}_0(x_1,x_2)}{\sum_{(u_1,u_2)\in\mathcal{X}}\widetilde{p}_0(u_1,u_2)}.$$

This construction produces a stable but imbalanced population distribution, with some regions of the state space receiving substantially more arrivals than others. Such heterogeneity is useful for

testing reject-inference methods because sparse states receive limited feedback and are therefore more vulnerable to selective-label bias and belief misspecification.

Given a state $x = (x_1, x_2)$, the nominal success probability is generated from a nonlinear spatial logit surface. Specifically, we define

$$\begin{aligned} g_0(x_1, x_2) = 0.60 - 0.55(x_1 - 4.5) + 0.28(x_2 - 4.5) + 0.95 \sin\left(\frac{x_1 + 1.0}{1.8}\right) \\ - 1.15 \exp\left\{-\frac{(x_1 - 2.6)^2 + (x_2 - 3.2)^2}{1.8}\right\} \\ + 0.80 \exp\left\{-\frac{(x_1 - 7.0)^2 + (x_2 - 7.0)^2}{3.2}\right\} - 0.15 \cos\left(\frac{x_2}{1.7}\right). \end{aligned}$$

The nominal conditional success probability is then

$$\rho_0(x) = \mathrm{clip}\left(\sigma(g_0(x)),\, 0.03,\, 0.97\right),$$

where $\sigma(z) = 1/(1+\exp(-z))$ is the sigmoid function. The clipping operation prevents the success probabilities from becoming too close to 0 or 1, which avoids degenerate Bernoulli outcomes and ensures that each state retains nontrivial uncertainty.

The nominal data stream is generated as follows. At each period t,

$$X_t \sim \mathrm{Categorical}(p_0), \qquad Y_t \mid X_t = x \sim \mathrm{Bernoulli}(\rho_0(x)).$$

The outcome Y_t represents the potential repayment or success indicator. Consistent with the reject-inference setting, however, Y_t is observed only if the policy accepts the individual. If the individual is rejected, the corresponding outcome remains unobserved and cannot be used for belief updating. Thus, the DGP separates the full underlying potential-outcome process from the selectively observed data available to the learner.

C.2 Baseline Descriptions

We compare the proposed method with a set of baselines covering myopic decision rules, stochastic exploration strategies, selective-label learning, and UCB-style exploration. In all experiments, each method maintains state-wise beta–Bernoulli statistics, and $\hat{\mu}_t(x)$ denotes the posterior mean success probability at state x based on the accepted observations collected up to time t. For more detailed descriptions and implementation settings of these baselines, please see Wei (2021).

- **Greedy**: accepts an individual whenever the current posterior mean exceeds the cost threshold, namely $a_t = \mathbb{1}\{\hat{\mu}_t(x_t) > c\}$.
- **ϵ-Greedy** (Langford & Zhang, 2007): takes a random binary action with probability ϵ and otherwise follows the Greedy rule.
- **Thompson Sampling (TS)** (Osband & Van Roy, 2015): draws a latent success rate $\tilde{\rho}_t(x)$ from the current beta posterior of state x and accepts whenever $\tilde{\rho}_t(x_t) > c$.
- **Consequential Learning (CL)** (Kilbertus et al., 2020): maintains a state-wise logit parameter $\theta(x_t)$, initializes it using the empirical success rate from the warm-start sample, and then accepts with probability $\sigma(\theta(x_t))$.
- **R-osl** (Bietti et al., 2021): a selective-label variant of optimistic RegCB. Because rejection yields zero reward, the decision reduces to checking whether an upper confidence bound of the estimated success probability exceeds c. The acceptance rule is $a_t = \mathbb{1}\left\{\hat{\mu}_t(x_t) + \sqrt{C_0/\hat{\nu}_t(x_t)} > c\right\}$, where C_0 controls the confidence width.
- **SL-OPT** (Wei, 2021): treats each state as an independent selective-label problem with an effective discount factor $\bar{\gamma}(x)$. In practice, $\bar{\gamma}(x)$ is estimated from the empirical inter-arrival times of state x, and the policy accepts whenever the corresponding nominal dynamic-programming value is positive.

C.3 Experimental Settings

We provide additional implementation details for both the simulation experiments and the real-data case study. In all experiments, the first B_0 arrivals are always accepted so that their labels are revealed. This warm-start phase provides an initial set of selectively observed outcomes before the policy begins making fully adaptive decisions.

Simulation setting. The synthetic experiments are conducted on a finite two-dimensional state grid. The default grid size is 5×5, yielding 25 discrete states. The arrival distribution $p(x)$ is a normalized multimodal surface formed by combining a constant background with several Gaussian-shaped components, creating a heterogeneous arrival process with both dense and sparse regions. The nominal success probability $\rho(x)$ is generated by a nonlinear logit function that includes linear trends, sinusoidal variation, and local Gaussian perturbations, followed by a sigmoid transformation and clipping to $[0.03, 0.97]$ (see Section C.1 for more details). The default simulation configuration used in the released code is

$$c = 0.65, \qquad \gamma = 0.999, \qquad T_{\text{train}} = 5000, \qquad T_{\text{eval}} = 1000, \qquad \eta_{\text{train}} = 0.01$$

with $B_0 \in \{0, 20, 40, 60, 80, 100\}$ for the nominal training-reward comparison, 100 independent replications, and 30 online evaluation trajectories per replication.

Real-data case study on the FICO dataset. The empirical study uses the FICO Explainable Machine Learning Challenge dataset. Each observation corresponds to an anonymized HELOC applicant together with a binary repayment outcome. We first construct a low-dimensional state representation from the original applicant features. Specifically, we split the data into a training set and a test set, fit an XGBoost classifier on the training set, and use its feature-importance ranking to identify the most informative variables. Starting from the full feature set, we then remove the least important feature one at a time, while monitoring the test-set performance to ensure that the predictive accuracy does not drop significantly. The original full-feature model achieves an AUC-ROC of 0.790 on the test set. Following this screening procedure, we identify two particularly informative variables: *Consolidated version of risk markers* and *Number Satisfactory Trades*.

We then discretize these two features to obtain a finite state space suitable for the proposed online decision framework. To determine the number of bins for each variable, we use supervised tree-based discretization on the training set: for each selected feature, we fit a univariate decision tree to determine label-informative split points, and we choose the smallest number of bins whose resulting test-set performance remains close to that of the original representation. Following this procedure, we partition *Consolidated version of risk markers* into 6 bins and *Number Satisfactory Trades* into 5 bins. Using only these discretized features, the resulting model achieves a test AUC-ROC of 0.777, indicating that the feature-selection and discretization steps retain most of the predictive signal while yielding a tractable finite-state representation for the empirical online decision problem.

The policy implementation uses 30 discrete states. The real-data experiment is conducted as an online replay study. The default empirical configuration is

$$c = 0.5, \qquad \gamma = 0.999, \qquad B_0 = 0, \qquad \eta_{\text{train}} = 0.0005,$$

with 1000 replications. For each method, the final reported empirical metric is the discounted cumulative reward R_{norm} over the full replay stream.

C.4 Adverse Perturbation Design for $R_{\min}$

For the $R_{\min}$ evaluation, we construct adverse environments by perturbing the nominal success probability $\rho_0(x)$ on the logit scale. Specifically, for a given perturbation pattern $s(x)$ and shift magnitude $\lambda \geq 0$, the perturbed success probability is defined as

$$\rho_\lambda(x) = \sigma\big(\text{logit}(\rho_0(x)) - \lambda s(x)\big),$$

where $\sigma(\cdot)$ is the sigmoid function. This construction ensures that the perturbation always acts in a downward direction on the success probability, while allowing different spatial structures through the choice of $s(x)$. We consider three perturbation patterns.

In the *global* case, the perturbation is homogeneous over the entire state space, $s_{\text{global}}(x) = 0.4,\ \forall x$. Thus, all states are shifted downward by the same amount on the logit scale.

In the *boundary* case, the perturbation is concentrated on states whose nominal success probabilities are close to c:

$$s_{\text{boundary}}(x) = \exp\left(-\frac{(\rho_0(x) - c)^2}{2\tau^2}\right), \qquad \tau = 0.10.$$

Hence, states near the decision boundary receive the strongest perturbation, while states far from the boundary are affected much less.

In the *subgroup* case, the perturbation is localized to a fixed rectangular region of the grid:

$$s_{\text{subgroup}}(x_1, x_2) = \mathbb{1}\{3 \le x_1 \le 4,\ 3 \le x_2 \le 4\}.$$

Therefore, only the states within this subgroup are shifted, while all other states remain unchanged.

C.5 Sensitivity Analysis

C.5.1 Ambiguity Set Radius η_{train}

The ambiguity-set radius η_{train} controls the degree of robustness imposed during policy learning and therefore directly determines the trade-off between nominal performance and protection against adverse environments. A larger η_{train} makes the learned policy more conservative by placing greater emphasis on worst-case outcomes within the uncertainty set, whereas a smaller η_{train} yields a policy closer to the nominal solution. Tables 4–5 and Figure 4 show that the proposed method remains stable across a broad range of robustness levels. As expected, increasing η_{train} generally reduces nominal reward, but it also improves resilience under distribution shift.

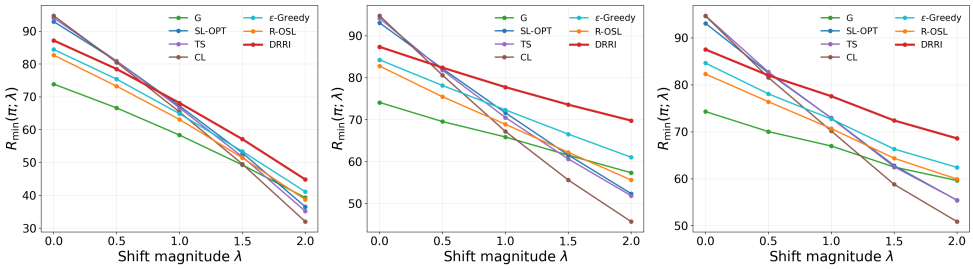
Results on R_{norm} Table 4 shows a clear monotone pattern: as the training ambiguity radius η_{train} increases, the nominal discounted reward R_{norm} decreases consistently across all warm-start sizes B_0. This is expected, since a larger ambiguity set induces a more conservative policy that sacrifices nominal reward in exchange for stronger protection against model misspecification. At the same time, for every fixed η_{train}, R_{norm} still improves as B_0 increases, indicating that additional initial labels remain beneficial even under robust training. Among the tested values, $\eta_{\text{train}} = 0.005$ achieves the highest nominal reward, while larger values such as 0.03 and 0.04 become increasingly over-conservative.

Table 4: Discounted total reward R_{norm} of various ambiguity set radius η_{train} under nominal environment (mean $\pm$ std over 100 runs).

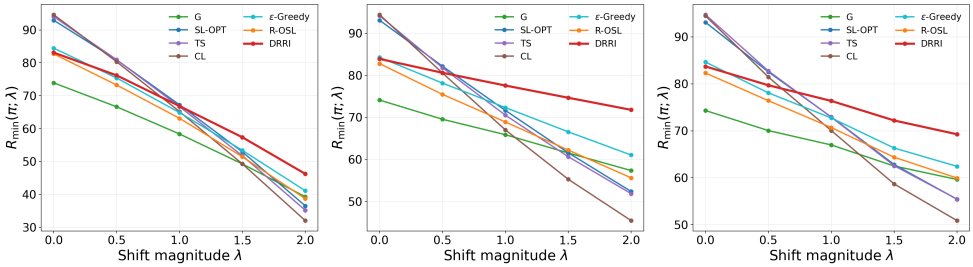
η_{train}	$B_0 = 0$	$B_0 = 20$	$B_0 = 40$	$B_0 = 60$	$B_0 = 80$	$B_0 = 100$
0.005	156.281 ± 13.279	156.622 ± 12.988	156.874 ± 12.979	157.870 ± 12.309	158.060 ± 12.381	158.563 ± 12.061
0.01	151.174 ± 13.971	151.254 ± 14.105	151.377 ± 13.839	153.098 ± 13.077	154.005 ± 13.077	154.556 ± 12.562
0.02	141.282 ± 18.242	142.994 ± 17.114	145.032 ± 16.203	146.650 ± 15.454	148.500 ± 14.478	149.528 ± 13.966
0.03	131.897 ± 20.390	133.724 ± 19.446	135.377 ± 19.190	138.075 ± 18.175	141.151 ± 16.062	142.750 ± 15.374
0.04	123.797 ± 20.840	124.883 ± 20.626	127.751 ± 20.256	131.515 ± 18.951	134.542 ± 17.581	137.017 ± 15.674

Results on R_{DRO} Table 5 further shows that the effect of η_{train} depends on the severity of the adverse environment. When the deployment shift is mild, a smaller ambiguity radius performs best: $\eta_{train} = 0.005$ attains the highest R_{DRO} for $\eta_{eval} = 0.00$ to 0.04. As the evaluation environment becomes more adverse, however, a slightly larger radius becomes preferable: $\eta_{train} = 0.01$ achieves the best performance at $\eta_{eval} = 0.05$ and 0.06. In contrast, larger training radii such as 0.02, 0.03, and 0.04 reduce performance throughout, suggesting that excessive robustness leads to unnecessary conservatism.

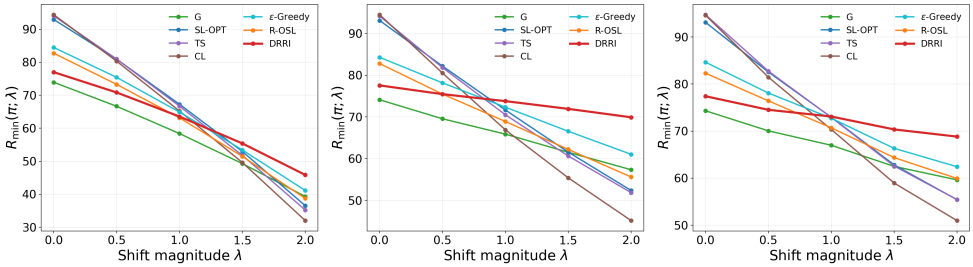
Results on R_{min} Figure 4 leads to a consistent conclusion under structured perturbations. As η_{train} increases, the reward curves of DRRI generally become flatter, indicating slower deterioration as the perturbation magnitude λ grows; however, this comes at the cost of a lower reward level even when the shift is small. Small-to-moderate values, especially $\eta_{train} = 0.005$ and 0.01, deliver the strongest overall performance in the global and subgroup cases, while very large values such as 0.03 and 0.04 appear overly conservative. In the boundary case, larger η_{train} improves stability under severe perturbations, but the loss in low-shift performance is substantial.



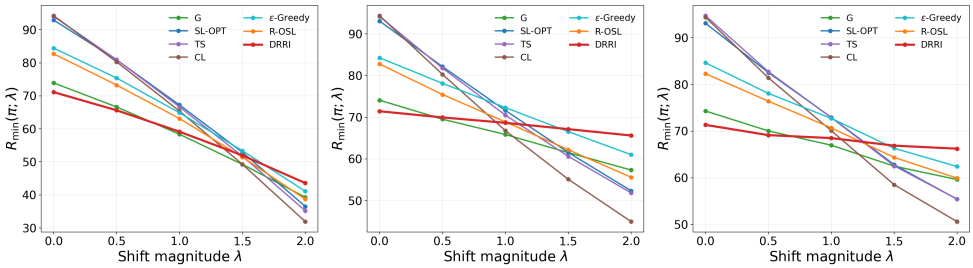
(a) $\eta_{train} = 0.005$ (left: Global case, middle: Boundary case, right: Subgroup case)



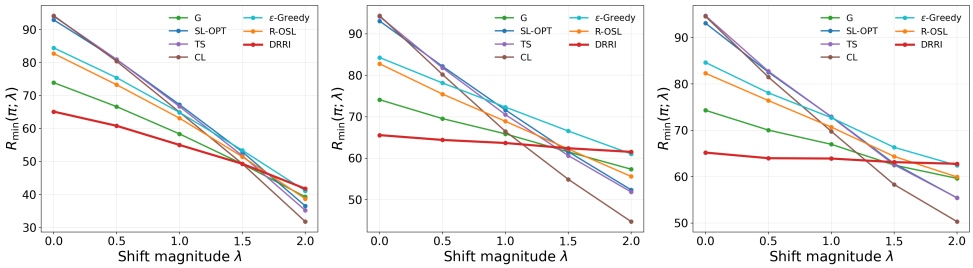
(b) $\eta_{train} = 0.01$ (left: Global case, middle: Boundary case, right: Subgroup case)



(c) $\eta_{train} = 0.02$ (left: Global case, middle: Boundary case, right: Subgroup case)



(d) $\eta_{train} = 0.03$ (left: Global case, middle: Boundary case, right: Subgroup case)



(e) $\eta_{train} = 0.04$ (left: Global case, middle: Boundary case, right: Subgroup case)

Figure 4: Robust performance ($R_{\min}(\pi;\lambda,s)$) of various ambiguity set radius η_{train} across three cases.

Table 5: Robust performance ($R_{\mathrm{DRO}}(\pi; \eta_{eval})$) of various ambiguity set radius η_{train} under different adverse environments (mean $\pm$ std over 100 runs).

η	$\eta_{eval} = 0.00$	$\eta_{eval} = 0.01$	$\eta_{eval} = 0.02$	$\eta_{eval} = 0.03$	$\eta_{eval} = 0.04$	$\eta_{eval} = 0.05$	$\eta_{eval} = 0.06$
0.005	99.362 $\pm$ 5.750	81.166 $\pm$ 4.058	72.370 $\pm$ 3.432	66.373 $\pm$ 3.200	60.705 $\pm$ 2.855	56.248 $\pm$ 2.986	51.209 $\pm$ 3.121
0.01	95.034 $\pm$ 6.579	78.722 $\pm$ 4.826	70.806 $\pm$ 4.138	65.373 $\pm$ 3.868	60.285 $\pm$ 3.316	56.262 $\pm$ 3.352	51.682 $\pm$ 3.288
0.02	87.684 $\pm$ 9.967	73.760 $\pm$ 7.792	66.955 $\pm$ 6.637	62.354 $\pm$ 5.984	57.831 $\pm$ 5.244	54.394 $\pm$ 4.953	50.561 $\pm$ 4.624
0.03	81.096 $\pm$ 11.120	68.690 $\pm$ 8.979	62.668 $\pm$ 7.535	58.545 $\pm$ 7.015	54.487 $\pm$ 6.269	51.493 $\pm$ 5.825	47.912 $\pm$ 5.367
0.04	74.755 $\pm$ 11.233	63.830 $\pm$ 9.175	58.618 $\pm$ 8.003	54.996 $\pm$ 7.439	51.282 $\pm$ 6.869	48.734 $\pm$ 6.438	45.548 $\pm$ 6.047

C.5.2 COST c

The cost parameter c controls the penalty incurred upon acceptance and therefore directly affects the aggressiveness of the learned policy. A larger c makes acceptance less attractive, shifts the decision rule toward greater conservativeness, and reduces the margin available under both nominal and adverse environments. Tables 6 and 7, together with Figure 5, show that the proposed method remains numerically stable across a wide range of cost levels, while the absolute reward level decreases as expected when the acceptance cost becomes larger.

Results on R_{norm} Table 6 shows that the nominal discounted reward decreases substantially as c increases. When $c = 0.35$, the nominal reward is around 453 across different warm-start sizes, whereas it drops to about 151 when $c = 0.65$ and further to about 42 when $c = 0.80$. This pattern is consistent with the economic interpretation of the model where a larger acceptance cost compresses the net gain from successful approvals and makes the policy more selective. At the same time, the effect of the warm-start size B_0 remains limited across all three settings.

Table 6: Discounted total reward R_{norm} of various cost c under nominal environment (mean $\pm$ std over 100 runs).

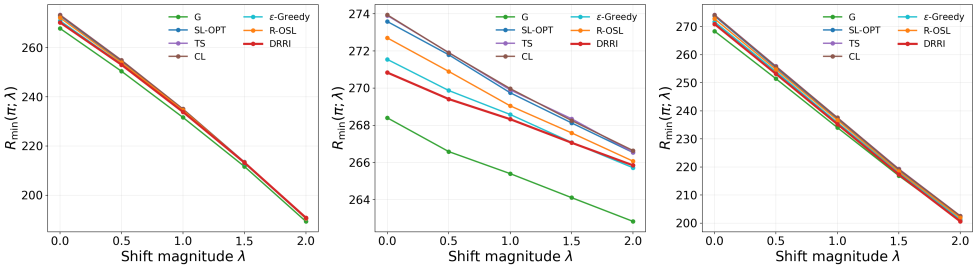
c	$B_0 = 0$	$B_0 = 20$	$B_0 = 40$	$B_0 = 60$	$B_0 = 80$	$B_0 = 100$
0.35	453.741 ± 11.270	453.741 ± 11.270	453.738 ± 11.272	453.738 ± 11.272	453.741 ± 11.274	453.728 ± 11.282
0.50	295.795 ± 16.000	296.019 ± 15.658	296.156 ± 15.549	296.590 ± 15.636	297.563 ± 14.811	298.522 ± 14.127
0.65	151.174 ± 13.971	151.254 ± 14.105	151.377 ± 13.839	153.098 ± 13.077	154.005 ± 13.077	154.556 ± 12.562
0.80	41.810 ± 11.245	41.603 ± 11.307	41.495 ± 11.470	42.386 ± 10.859	42.929 ± 10.476	42.984 ± 10.398

Results on R_{DRO} A similar pattern appears in Table 7 under adverse environments. For all values of c, R_{DRO} declines monotonically as η_{eval} increases, confirming that more severe shifts lead to lower robust utility. More importantly, increasing c shifts the entire reward profile downward. The same ordering holds throughout the full range of η_{eval}. Nevertheless, the degradation remains smooth and stable across η_{eval}, showing that the proposed method preserves robust behavior even when the economic penalty of acceptance becomes substantially larger.

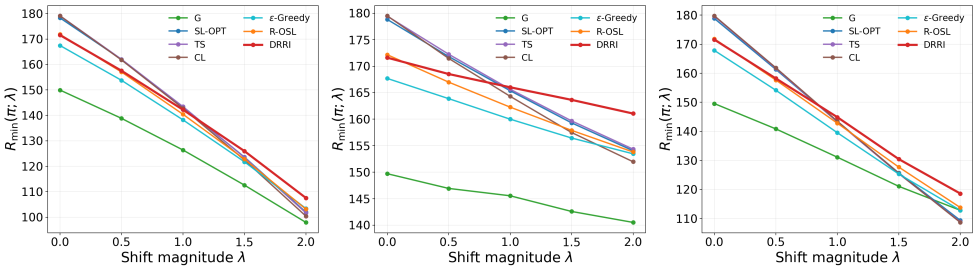
Table 7: Robust performance ($R_{\mathrm{DRO}}(\pi; \eta_{eval})$) of various cost c under different adverse environments (mean $\pm$ std over 100 runs).

c	$\eta_{eval} = 0.00$	$\eta_{eval} = 0.01$	$\eta_{eval} = 0.02$	$\eta_{eval} = 0.03$	$\eta_{eval} = 0.04$	$\eta_{eval} = 0.05$	$\eta_{eval} = 0.06$
0.35	287.280 ± 3.813	256.771 ± 3.130	242.218 ± 2.567	232.072 ± 2.364	222.584 ± 2.036	215.104 ± 1.895	206.919 ± 1.954
0.50	186.489 ± 7.407	161.084 ± 5.234	149.027 ± 4.429	140.766 ± 3.824	132.823 ± 3.319	126.556 ± 3.092	119.566 ± 2.897
0.65	95.034 ± 6.579	78.722 ± 4.826	70.806 ± 4.138	65.373 ± 3.868	60.285 ± 3.316	56.262 ± 3.352	51.682 ± 3.288
0.80	26.918 ± 5.231	19.917 ± 3.992	16.644 ± 3.243	14.269 ± 3.006	11.679 ± 2.548	10.215 ± 2.482	8.177 ± 2.317

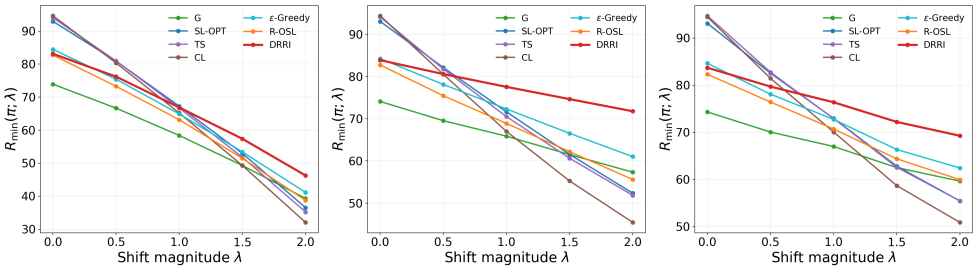
Results on R_{min} Figure 5 reports the robust performance $R_{\mathrm{min}}(\pi; \lambda, s)$ under different cost levels c and shift magnitudes λ across the global, boundary, and subgroup cases. First, for every cost level, R_{min} generally decreases as the shift magnitude λ increases, confirming that stronger adverse shifts lead to worse worst-case performance. Second, the relative advantage of DRRI is most visible at intermediate cost levels, especially in the boundary and subgroup cases, where robustness to local deterioration around the decision threshold is most critical. When $c = 0.35$, the acceptance cost is low, so approving a poor applicant is relatively inexpensive and most methods tend to accept a large fraction of samples; as a result, the curves remain close and the differences across methods are small. When $c = 0.8$, the overall reward level becomes much lower because acceptance is very costly, so most policies reject a large proportion of samples. In this regime, DRRI still shows stronger protection in the global and boundary cases, but in the subgroup case the differences among methods also become small, since the high rejection tendency leaves limited room for policies to separate from one another under a localized shift.



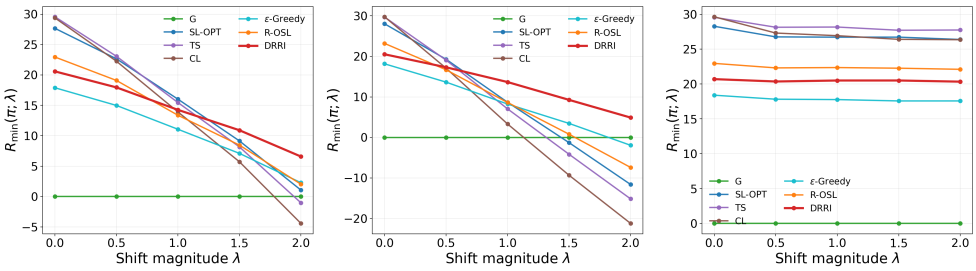
(a) $c = 0.35$ (left: Global case, middle: Boundary case, right: Subgroup case)



(b) $c = 0.5$ (left: Global case, middle: Boundary case, right: Subgroup case)



(c) $c = 0.65$ (left: Global case, middle: Boundary case, right: Subgroup case)



(d) $c = 0.8$ (left: Global case, middle: Boundary case, right: Subgroup case)

Figure 5: Robust performance ($R_{\min}(\pi;\lambda,s)$) of various cost c across three cases.

C.5.3 DISCOUNT FACTOR γ

The discount factor γ controls how strongly future rewards are weighted in policy learning and evaluation, and therefore determines the effective planning horizon of the decision process. Tables 8 and 9, together with the corresponding robustness figure, show that the proposed method remains stable across a broad range of discount factors.

Results on R_{norm} Table 8 shows that the nominal discounted reward increases sharply with the discount factor γ. When $\gamma = 0.99$, the nominal reward remains around 14 to 16 across different warm-start sizes, whereas it rises to roughly 29 to 31 when $\gamma = 0.995$, and further increases to about 151 to 155 when $\gamma = 0.999$. This pattern is fully consistent with the interpretation of γ: a larger discount factor assigns substantially more weight to future rewards and therefore amplifies the cumulative discounted return.

Table 8: Discounted total reward R_{norm} under different discount factors γ in the nominal environment (mean $\pm$ std over 100 runs).

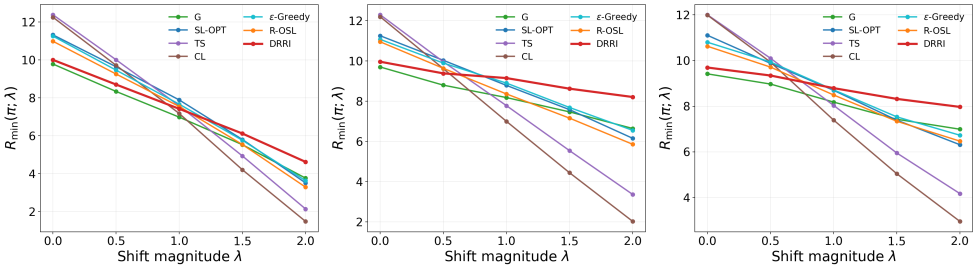
γ	$B_0 = 0$	$B_0 = 20$	$B_0 = 40$	$B_0 = 60$	$B_0 = 80$	$B_0 = 100$
0.99	14.332 ± 3.224	14.558 ± 3.178	14.781 ± 3.073	15.067 ± 3.081	15.344 ± 3.036	15.554 ± 2.936
0.995	29.207 ± 5.093	29.512 ± 4.980	29.897 ± 4.778	30.304 ± 4.692	30.699 ± 4.557	31.016 ± 4.454
0.999	151.174 ± 13.971	151.254 ± 14.105	151.377 ± 13.839	153.098 ± 13.077	154.005 ± 13.077	154.556 ± 12.562

Results on R_{DRO} A similar pattern appears in Table 9 under adverse environments. For every discount factor, R_{DRO} decreases monotonically as η_{eval} increases, confirming that more severe environmental deterioration leads to lower robust utility. More importantly, increasing γ shifts the entire reward profile upward in a systematic way. For example, under the nominal evaluation environment $\eta_{eval} = 0$, the robust reward increases from 13.599 at $\gamma = 0.99$ to 28.028 at $\gamma = 0.995$, and further to 95.034 at $\gamma = 0.999$. The same ordering is preserved across the full range of η_{eval}. This shows that a larger discount factor increases the total value retained from future periods under both nominal and adverse deployment conditions.

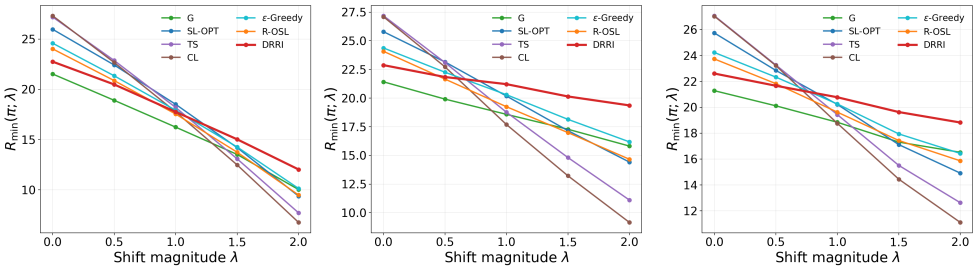
Table 9: Robust performance $\big(R_{\mathrm{DRO}}(\pi; \eta_{eval})\big)$ under different discount factors γ and adverse environments (mean $\pm$ std over 100 runs).

γ	$\eta_{eval} = 0.00$	$\eta_{eval} = 0.01$	$\eta_{eval} = 0.02$	$\eta_{eval} = 0.03$	$\eta_{eval} = 0.04$	$\eta_{eval} = 0.05$	$\eta_{eval} = 0.06$
0.99	13.599 ± 1.659	11.374 ± 1.379	10.266 ± 1.150	9.309 ± 1.050	8.748 ± 0.967	8.186 ± 0.911	7.517 ± 0.877
0.995	28.028 ± 2.935	23.437 ± 2.349	21.083 ± 1.970	19.402 ± 1.837	18.018 ± 1.633	16.869 ± 1.544	15.552 ± 1.539
0.999	95.034 ± 6.579	78.722 ± 4.826	70.806 ± 4.138	65.373 ± 3.868	60.285 ± 3.316	56.262 ± 3.352	51.682 ± 3.288

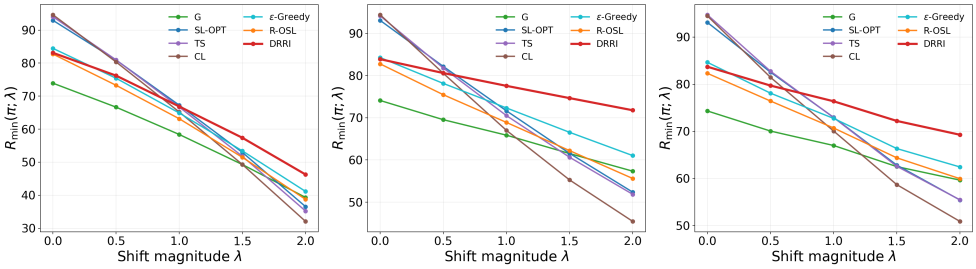
Results on R_{min} The results on R_{min} can be interpreted in a similar manner. As γ increases, the absolute level of worst-trajectory reward is expected to rise across all perturbation patterns, because more future payoffs are preserved in the cumulative objective. At the same time, the qualitative robustness pattern should remain largely unchanged where all methods deteriorate as the shift magnitude λ increases, but the proposed method continues to exhibit the slowest degradation under severe perturbations, especially in the more challenging global and boundary settings.



(a) $\gamma = 0.99$ (left: Global case, middle: Boundary case, right: Subgroup case)



(b) $\gamma = 0.995$ (left: Global case, middle: Boundary case, right: Subgroup case)



(c) $\gamma = 0.999$ (left: Global case, middle: Boundary case, right: Subgroup case)

Figure 6: Robust performance ($R_{\min}(\pi;\lambda,s)$) of various discounted factor γ across three cases.

C.5.4 TRUNCATION SIZE N

The truncation level N controls how deeply the backward dynamic programming recursion is carried out in the numerical implementation and therefore determines how much of the long-run continuation value is explicitly incorporated into policy computation. A larger N allows the approximation to account for more distant future effects, whereas a smaller N places greater reliance on the terminal surrogate used at the truncation boundary. Tables 10 and 11, together with Figure 7, show that the proposed method remains numerically stable across a broad range of truncation levels.

Results on R_{norm} Table 10 shows that the nominal discounted reward is sensitive to the truncation level only when N is very small, but becomes highly stable once N reaches a moderate size. In particular, moving from $N = 10$ to $N = 25$ yields a clear improvement across all warm-start settings, indicating that an overly shallow truncation cannot adequately capture the continuation value in the backward recursion. After that point, however, the results vary only marginally: for $N \geq 25$, the mean rewards are nearly unchanged across all values of B_0, and the differences are very small relative to the corresponding standard deviations. This pattern suggests that the terminal condition mainly affects the far tail of the dynamic program, while its influence on the states that are most relevant for the learned policy diminishes quickly once the truncation depth is sufficiently large.

Table 10: Discounted total reward R_{norm} of different truncation sizes N under nominal environment (mean $\pm$ std over 100 runs).

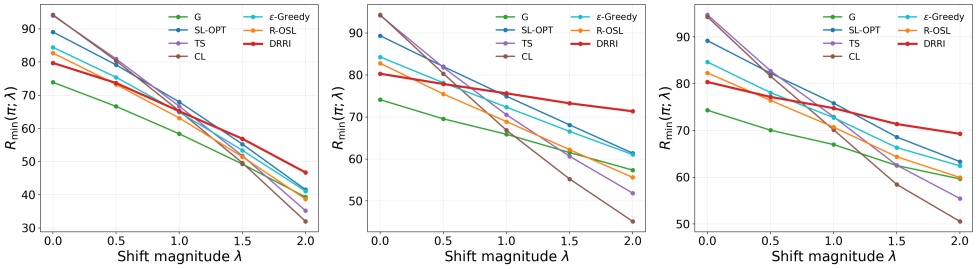
N	$B_0 = 0$	$B_0 = 20$	$B_0 = 40$	$B_0 = 60$	$B_0 = 80$	$B_0 = 100$
10	146.194 ± 14.950	146.356 ± 14.905	146.476 ± 14.565	147.226 ± 14.274	147.730 ± 13.942	148.114 ± 13.830
25	151.227 ± 14.494	151.532 ± 14.202	151.528 ± 14.153	153.013 ± 13.364	153.293 ± 13.405	153.863 ± 13.185
55	151.414 ± 14.243	151.581 ± 14.173	151.715 ± 13.889	153.270 ± 13.157	154.068 ± 13.155	154.839 ± 12.694
110	151.065 ± 13.992	151.301 ± 13.840	151.428 ± 13.704	153.058 ± 12.989	153.977 ± 12.988	154.491 ± 12.524
220	151.174 ± 13.971	151.254 ± 14.105	151.377 ± 13.839	153.098 ± 13.077	154.005 ± 13.077	154.556 ± 12.562
440	151.033 ± 14.258	151.263 ± 14.108	151.416 ± 13.826	153.138 ± 13.061	154.048 ± 13.069	154.581 ± 12.548

Results on R_{DRO} A similar conclusion emerges from Table 11 under adverse environments. When N increases from 10 to 25, the robust reward improves noticeably for every value of η_{eval}, which again suggests that too small a truncation level leads to a less accurate approximation of the robust continuation value. Beyond $N = 25$, however, the performance curves flatten substantially: the mean values of R_{DRO} for $N = 25, 55, 110, 220$, and 440 are extremely close, and the remaining differences are negligible compared with the Monte Carlo variability. There is also a slight downward drift in the mean reward as N becomes very large, which is because the terminal surrogate is mildly optimistic and thus the smaller truncation levels inherit this optimism somewhat more strongly.

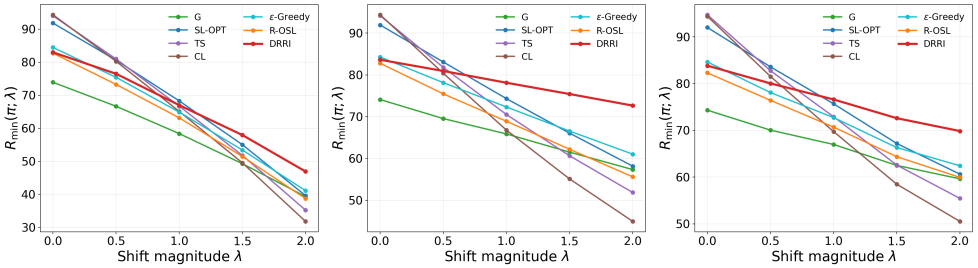
Table 11: Robust performance $\big(R_{\mathrm{DRO}}(\pi; \eta_{eval})\big)$ under different truncation sizes N and adverse environments (mean $\pm$ std over 100 runs).

N	$\eta_{eval} = 0.00$	$\eta_{eval} = 0.01$	$\eta_{eval} = 0.02$	$\eta_{eval} = 0.03$	$\eta_{eval} = 0.04$	$\eta_{eval} = 0.05$	$\eta_{eval} = 0.06$
10	91.054 ± 8.181	76.215 ± 6.003	68.920 ± 5.105	64.003 ± 4.591	59.252 ± 3.887	55.610 ± 3.761	51.523 ± 3.610
25	94.864 ± 6.906	78.828 ± 4.968	70.974 ± 4.153	65.675 ± 3.733	60.589 ± 3.155	56.608 ± 3.181	52.185 ± 3.145
55	95.064 ± 6.745	78.909 ± 4.768	71.028 ± 4.044	65.613 ± 3.656	60.547 ± 3.079	56.559 ± 3.120	52.015 ± 3.075
110	95.055 ± 6.589	78.784 ± 4.848	70.914 ± 4.137	65.496 ± 3.800	60.407 ± 3.264	56.394 ± 3.325	51.840 ± 3.206
220	95.034 ± 6.579	78.722 ± 4.826	70.806 ± 4.138	65.373 ± 3.868	60.285 ± 3.316	56.262 ± 3.352	51.682 ± 3.288
440	94.977 ± 6.741	78.671 ± 4.939	70.745 ± 4.210	65.318 ± 3.927	60.204 ± 3.390	56.184 ± 3.405	51.589 ± 3.360

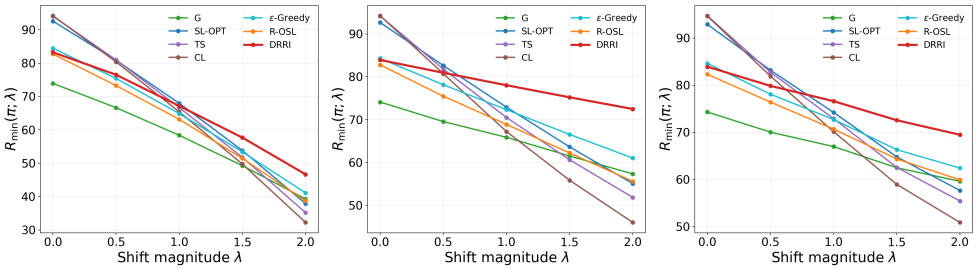
Results on $R_{\min}$ Figure 7 further shows that the worst-trajectory robustness patterns are highly stable across different truncation levels. The relative ordering of methods is largely preserved, the degradation trends as λ increases remain nearly identical, and DRRI continues to exhibit the strongest resistance to severe perturbations.



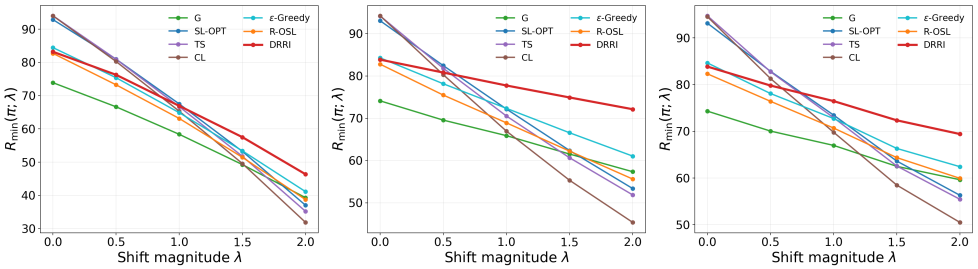
(a) $N = 10$ (left: Global case, middle: Boundary case, right: Subgroup case)



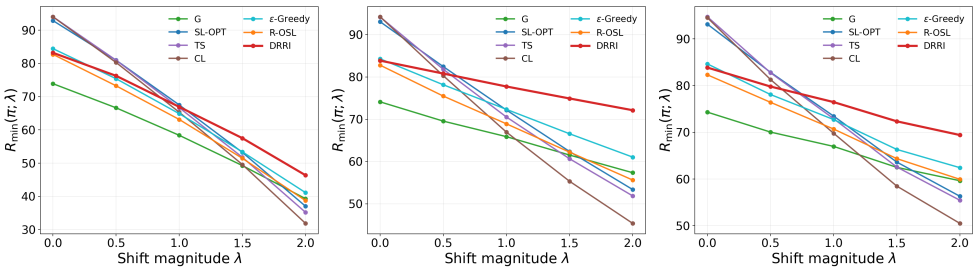
(b) $N = 25$ (left: Global case, middle: Boundary case, right: Subgroup case)



(c) $N = 55$ (left: Global case, middle: Boundary case, right: Subgroup case)



(d) $N = 110$ (left: Global case, middle: Boundary case, right: Subgroup case)



(e) $N = 220$ (left: Global case, middle: Boundary case, right: Subgroup case)

Figure 7: Robust performance ($R_{\min}(\pi;\lambda,s)$) of various truncation sizes N across three cases.

C.6 The Influence of Terminal Condition

The terminal condition specifies the surrogate continuation value imposed at the truncation boundary, that is, the boundary value $\phi(\mu)$ used to initialize the backward recursion through $V_\eta^N(\mu, N+1) = \phi(\mu)$. In this section, we consider four terminal conditions. The first is the proposed ***DRO*** terminal condition $\phi_{\mathrm{DRO}}(\mu)$ with

$$\phi_{\mathrm{DRO}}(\mu) = \frac{\max\{q_{\mathrm{low}}(\mu) - c,\, 0\}}{1-\gamma},$$

where $q_{\mathrm{low}}(\mu) = \inf\{q : \mathbb{D}(q, \mu) \le \eta\}$. Indeed, it is the lower endpoint of the ambiguity set. The second is the *nominal* terminal condition $\phi_{\mathrm{Nominal}}(\mu)$ such that

$$\phi_{\mathrm{Nominal}}(\mu) = \frac{\max\{\mu - c,\, 0\}}{1-\gamma},$$

which uses the nominal posterior mean directly and ignores robustness adjustment at the boundary. The third is the *zero* terminal condition $\phi_{\mathrm{Zero}}(\mu) = 0$. It represents an extremely conservative specification that assigns no residual continuation value beyond the truncation horizon. The fourth is the *optimistic* terminal condition $\phi_{\mathrm{Optimistic}}(\mu)$ with

$$\phi_{\mathrm{Optimistic}}(\mu) = \frac{\max\{q_{\mathrm{high}}(\mu) - c,\, 0\}}{1-\gamma},$$

where $q_{\mathrm{high}}(\mu) = \sup\{q : \mathbb{D}(q, \mu) \le \eta\}$. It is the upper endpoint of the ambiguity set.

Intuitively, more optimistic terminal conditions assign larger tail values and therefore induce more aggressive acceptance, whereas more conservative terminal conditions weaken continuation incentives and lead to more selective decisions.

Results on R_{norm} Table 12 shows that the terminal condition has a substantial impact on nominal performance when the truncation level N is small, but this impact diminishes steadily as N increases. For $N = 10$ and $N = 25$, the optimistic and nominal terminal conditions achieve the highest R_{norm}, the DRO terminal is more conservative, and the zero terminal performs extremely poorly because it suppresses continuation value almost entirely at the truncation boundary. As N becomes larger, however, the differences across terminal conditions shrink markedly. In particular, the zero terminal improves monotonically with N, rising from a severely pessimistic approximation at $N = 10$ to a level close to the other specifications by $N = 440$. At the same time, the optimistic and nominal terminal conditions gradually lose their small-N advantage, and all four specifications converge to very similar nominal rewards once the truncation depth is sufficiently large. This pattern is consistent with the role of the terminal condition in the backward recursion: when N is small, the terminal condition has a strong influence on the computed continuation value, whereas for large N, more of the future effect is evaluated explicitly and the particular terminal choice becomes much less consequential.

Results on R_{DRO} Table 13 shows a clearer distinction across terminal conditions in terms of robustness. For small truncation levels, optimistic and nominal terminal conditions often achieve higher rewards when η_{eval} is close to zero, but their performance deteriorates more rapidly as the evaluation environment becomes more adverse. By contrast, the DRO terminal consistently provides stronger lower-tail protection and becomes preferable as the degree of shift increases. This pattern is most transparent for $N = 10$, 25, and 55, where the optimistic terminal attains the best or near-best performance under no shift but suffers the steepest decline as η_{eval} grows, while the DRO terminal degrades more slowly and retains a clear advantage in the severe-shift regime. The zero terminal condition behaves as an extreme pessimistic boundary and performs especially poorly: for $N = 10$, 25, and 55, the robust reward is essentially zero throughout the entire range of η_{eval}, and even at $N = 110$ it remains far below all other specifications. This outcome is consistent with the fact that the zero boundary assigns no continuation value once the truncation threshold is reached, effectively turning high-ν states into terminal reject regions. As N increases to 440, however, the influence of the terminal boundary weakens substantially, and the DRO, nominal, and optimistic specifications all converge to very similar robust performance.

Table 12: Discounted total reward R_{norm} under different terminal conditions and truncation levels N in the nominal environment (mean $\pm$ std over 100 runs).

N	Terminal condition	$B_0 = 0$	$B_0 = 20$	$B_0 = 40$	$B_0 = 60$	$B_0 = 80$	$B_0 = 100$
10	DRO	146.194 ± 14.950	146.356 ± 14.905	146.476 ± 14.565	147.226 ± 14.274	147.730 ± 13.942	148.114 ± 13.830
10	Nominal	156.971 ± 13.188	157.129 ± 12.944	157.332 ± 12.796	157.652 ± 12.811	157.834 ± 12.887	158.507 ± 12.081
10	Zero	26.245 ± 5.793	26.490 ± 5.778	26.920 ± 5.721	27.519 ± 5.554	28.159 ± 5.656	28.843 ± 5.411
10	Optimistic	162.102 ± 11.760	162.213 ± 11.513	162.546 ± 11.169	162.585 ± 11.233	162.932 ± 10.863	163.239 ± 10.547
25	DRO	151.227 ± 14.494	151.532 ± 14.202	151.528 ± 14.153	153.013 ± 13.364	153.293 ± 13.405	153.863 ± 13.185
25	Nominal	157.425 ± 13.079	157.686 ± 12.644	157.953 ± 12.516	159.046 ± 12.211	159.508 ± 12.347	160.306 ± 11.584
25	Zero	60.240 ± 9.083	61.042 ± 8.923	61.886 ± 8.568	62.792 ± 8.373	63.630 ± 8.099	64.084 ± 7.735
25	Optimistic	161.057 ± 12.260	161.113 ± 12.007	161.341 ± 11.933	161.674 ± 11.862	161.956 ± 11.717	162.477 ± 11.075
55	DRO	151.414 ± 14.243	151.581 ± 14.173	151.715 ± 13.889	153.270 ± 13.157	154.068 ± 13.155	154.839 ± 12.694
55	Nominal	154.164 ± 13.755	154.479 ± 13.435	154.550 ± 13.228	156.082 ± 12.251	156.808 ± 12.584	157.295 ± 12.366
55	Zero	99.681 ± 11.821	100.443 ± 11.962	101.590 ± 11.533	102.956 ± 11.032	104.419 ± 10.395	105.169 ± 9.894
55	Optimistic	157.024 ± 12.823	157.136 ± 12.589	157.120 ± 12.386	158.442 ± 11.917	158.989 ± 11.844	159.663 ± 11.552
110	DRO	151.065 ± 13.992	151.301 ± 13.840	151.428 ± 13.704	153.058 ± 12.989	153.977 ± 12.988	154.491 ± 12.524
110	Nominal	151.884 ± 14.097	152.046 ± 14.030	152.290 ± 13.878	153.853 ± 13.119	154.573 ± 13.092	155.310 ± 12.582
110	Zero	133.249 ± 13.098	133.440 ± 12.971	133.448 ± 12.793	134.783 ± 12.156	135.761 ± 11.924	136.354 ± 11.455
110	Optimistic	152.585 ± 14.051	152.899 ± 13.780	153.017 ± 13.759	154.715 ± 12.880	155.405 ± 12.928	156.046 ± 12.417
220	DRO	151.174 ± 13.971	151.254 ± 14.105	151.377 ± 13.839	153.098 ± 13.077	154.005 ± 13.077	154.556 ± 12.562
220	Nominal	150.993 ± 14.079	151.160 ± 14.007	151.345 ± 13.823	152.990 ± 13.092	153.925 ± 13.093	154.449 ± 12.565
220	Zero	148.531 ± 13.941	148.752 ± 13.794	148.873 ± 13.551	150.492 ± 12.868	151.332 ± 12.633	151.855 ± 12.200
220	Optimistic	151.121 ± 13.942	151.288 ± 13.870	151.413 ± 13.726	153.072 ± 13.011	154.018 ± 12.996	154.562 ± 12.503
440	DRO	151.033 ± 14.258	151.263 ± 14.108	151.416 ± 13.826	153.138 ± 13.061	154.048 ± 13.069	154.581 ± 12.548
440	Nominal	151.164 ± 13.971	151.244 ± 14.105	151.397 ± 13.824	153.119 ± 13.057	154.029 ± 13.066	154.579 ± 12.545
440	Zero	151.033 ± 14.258	151.263 ± 14.108	151.416 ± 13.826	153.138 ± 13.061	154.048 ± 13.069	154.581 ± 12.548
440	Optimistic	151.233 ± 13.882	151.241 ± 14.102	151.394 ± 13.821	153.122 ± 13.051	154.032 ± 13.060	154.582 ± 12.538

Table 13: Robust performance $\left(R_{\mathrm{DRO}}(\pi; \eta_{eval})\right)$ under different terminal conditions and truncation levels N (mean $\pm$ std over 100 runs).

N	Terminal condition	$\eta_{eval} = 0.00$	$\eta_{eval} = 0.01$	$\eta_{eval} = 0.02$	$\eta_{eval} = 0.03$	$\eta_{eval} = 0.04$	$\eta_{eval} = 0.05$	$\eta_{eval} = 0.06$
10	DRO	91.054 ± 8.181	76.215 ± 6.003	68.920 ± 5.105	64.003 ± 4.591	59.252 ± 3.887	55.610 ± 3.761	51.523 ± 3.610
10	Nominal	100.491 ± 5.648	81.192 ± 3.821	71.868 ± 3.393	65.479 ± 3.353	59.402 ± 3.316	54.670 ± 3.554	49.208 ± 3.775
10	Zero	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000
10	Optimistic	104.992 ± 4.221	81.632 ± 3.552	70.687 ± 3.573	63.083 ± 4.364	55.819 ± 4.572	50.115 ± 5.074	43.607 ± 5.442
25	DRO	94.864 ± 6.906	78.828 ± 4.968	70.974 ± 4.153	65.675 ± 3.733	60.589 ± 3.155	56.608 ± 3.181	52.185 ± 3.145
25	Nominal	100.870 ± 5.732	80.966 ± 4.043	71.433 ± 3.485	64.990 ± 3.474	58.694 ± 3.351	53.804 ± 3.677	48.318 ± 3.904
25	Zero	-0.000 ± 0.003	0.001 ± 0.006	0.001 ± 0.009	-0.000 ± 0.002	-0.000 ± 0.005	-0.000 ± 0.004	0.000 ± 0.004
25	Optimistic	104.084 ± 4.769	80.972 ± 3.756	70.031 ± 3.770	62.605 ± 4.266	55.341 ± 4.358	49.768 ± 4.864	43.241 ± 5.249
55	DRO	95.064 ± 6.745	78.909 ± 4.768	71.028 ± 4.044	65.613 ± 3.656	60.547 ± 3.079	56.559 ± 3.120	52.015 ± 3.075
55	Nominal	98.533 ± 6.589	79.596 ± 4.864	70.477 ± 4.309	64.327 ± 4.292	58.363 ± 4.203	53.691 ± 4.412	48.445 ± 4.621
55	Zero	0.007 ± 0.063	0.007 ± 0.065	0.004 ± 0.037	0.006 ± 0.060	0.003 ± 0.031	0.004 ± 0.040	0.004 ± 0.039
55	Optimistic	100.745 ± 5.726	79.809 ± 4.629	69.858 ± 4.263	63.132 ± 4.697	56.550 ± 4.806	51.425 ± 5.267	45.688 ± 5.522
110	DRO	95.055 ± 6.589	78.784 ± 4.848	70.914 ± 4.137	65.496 ± 3.800	60.407 ± 3.264	56.394 ± 3.325	51.840 ± 3.206
110	Nominal	96.240 ± 6.865	78.753 ± 4.906	70.337 ± 4.267	64.551 ± 4.156	59.009 ± 3.958	54.727 ± 4.142	49.785 ± 4.293
110	Zero	6.635 ± 2.074	5.335 ± 1.707	4.589 ± 1.493	3.992 ± 1.425	3.741 ± 1.341	3.291 ± 1.264	2.938 ± 1.196
110	Optimistic	97.070 ± 6.930	78.855 ± 4.926	70.134 ± 4.394	64.160 ± 4.408	58.377 ± 4.247	53.895 ± 4.565	48.772 ± 4.711
220	DRO	95.034 ± 6.579	78.722 ± 4.826	70.806 ± 4.138	65.373 ± 3.868	60.285 ± 3.316	56.262 ± 3.352	51.682 ± 3.288
220	Nominal	95.074 ± 6.728	78.656 ± 4.931	70.631 ± 4.216	65.152 ± 3.992	59.945 ± 3.530	55.879 ± 3.614	51.199 ± 3.620
220	Zero	29.622 ± 4.557	24.312 ± 3.685	21.895 ± 3.406	19.675 ± 3.143	18.467 ± 2.879	16.880 ± 2.826	15.322 ± 2.687
220	Optimistic	95.323 ± 6.550	78.666 ± 4.896	70.589 ± 4.215	65.035 ± 4.065	59.765 ± 3.655	55.657 ± 3.758	50.929 ± 3.805
440	DRO	94.977 ± 6.741	78.671 ± 4.939	70.745 ± 4.210	65.318 ± 3.927	60.204 ± 3.390	56.184 ± 3.405	51.589 ± 3.360
440	Nominal	95.048 ± 6.590	78.722 ± 4.832	70.775 ± 4.139	65.317 ± 3.901	60.198 ± 3.387	56.157 ± 3.430	51.549 ± 3.405
440	Zero	92.960 ± 6.591	76.932 ± 4.893	69.276 ± 4.211	63.940 ± 3.905	58.980 ± 3.472	55.030 ± 3.500	50.535 ± 3.411
440	Optimistic	95.071 ± 6.605	78.726 ± 4.836	70.766 ± 4.140	65.300 ± 3.906	60.169 ± 3.402	56.125 ± 3.441	51.501 ± 3.445

C.7 Sensitivity Analysis on Empirical Dataset

Since the FICO experiment maps high-dimensional HELOC records to a finite online state space, the empirical conclusions should not depend on a single arbitrary feature-selection or binning choice. We therefore add a new preprocessing-sensitivity analysis for the FICO replay study.

The new analysis keeps the online replay protocol and all seven policy implementations unchanged, and varies only the finite-state construction. We consider three alternative preprocessing choices: (i) the same selected variables (x_1, x_5) with coarser quantile bins; (ii) the same selected variables (x_1, x_5) with supervised univariate tree bins; and (iii) a feature-selection variant replacing x_5 with the next retained candidate feature x_{15}, again using supervised univariate tree bins.

As illustrated in Figure 14. Across 1000 replay replications for each preprocessing variant, the qualitative conclusion of Table 3 is unchanged. R-osl continues to achieve the highest nominal mean reward, while DRRI remains very close in mean and consistently improves lower-tail performance. In all three variants, DRRI improves the minimum, 1%, 5%, and 10% reward quantiles relative to R-osl. This is the same trade-off emphasized in the original FICO study: DRRI is not designed to maximize nominal mean reward under the empirical replay distribution; rather, its KL-robust update makes the learned policy more conservative in states where selective labels leave substantial posterior uncertainty, which improves downside performance.

Table 14: Sensitivity of replay performance to alternative FICO preprocessing specifications. For each preprocessing specification, we report the distributional statistics of R-osl and DRRI over 1,000 replay replications. All variants satisfy the predictive-quality and sparsity criteria described above.

Preprocessing	Method	Mean	Std.	Min	1%	5%	10%
(x_1, x_5), coarse quantile	R-osl	62.097	12.959	11.530	23.394	38.788	44.438
	DRRI	61.995	8.287	24.564	38.864	46.569	51.483
(x_1, x_5), supervised tree	R-osl	57.025	14.401	1.097	12.674	27.547	38.428
	DRRI	56.445	8.165	14.651	33.696	42.002	46.602
(x_1, x_{15}), supervised tree	R-osl	60.487	11.629	17.076	28.979	39.238	44.282
	DRRI	59.818	7.820	24.043	36.329	46.158	50.384

These results indicate that the FICO conclusion is not driven by the particular binning rule used to form the original 30 states. The nominal mean ordering is stable: R-osl is slightly higher in average reward, and DRRI is second. The lower-tail ordering is also stable but favors DRRI: the DRRI distribution has a substantially better downside profile in every preprocessing variant. This pattern is consistent with the mechanism of the method. Under reject inference, states that are rejected often receive little label feedback, so aggressive policies can lock in poor early decisions in some replications. DRRI's distributionally robust Bellman update penalizes uncertain success probabilities, which can slightly reduce nominal mean reward but protects against unfavorable replay paths and improves the lower quantiles.

C.8 Runtime and Memory Analysis

The main computational object in DRRI is the homogeneous robust dynamic-programming table. For truncation depth N and belief-grid size G, a single table stores roughly $(N+2)G$ value entries, plus $O(G)$ KL-bound arrays. Thus, the memory for one homogeneous solver is $O(NG)$ and its precomputation time is also approximately linear in NG after the KL bounds have been computed.

In the covariate-dependent finite-state implementation, the online policy additionally maintains posterior trackers for K covariate states, which is $O(K)$ memory. The implementation also lazily caches homogeneous DP solvers for distinct effective discount factors encountered during replay. If C such solvers are cached, the dominant memory term is $O(CNG)$ and the training-time cost is approximately the online replay cost plus the cost of constructing these C solvers. In the worst case, C can grow with the number of distinct state-specific arrival patterns, and therefore with the number of states and the replay horizon; in practice, rounding of the effective discount factors substantially reduces duplicate solver construction.

Table 15: Runtime and memory benchmark for the DRRI simulation implementation. Each row runs one synthetic training replication with DRRI only; K is the number of covariate states, G is the belief-grid size, N is the truncation depth, and C is the number of cached homogeneous DP solvers created during replay.

Sweep	K	G	N	C	Runtime (s)	Peak RSS (MB)
baseline	25	401	220	564	17.42	423.9
state sweep	50	401	220	902	27.64	657.1
state sweep	100	401	220	1352	40.97	967.4
state sweep	200	401	220	2106	64.27	1487.5
grid sweep	25	101	220	564	5.60	133.0
grid sweep	25	201	220	564	9.48	230.0
grid sweep	25	801	220	564	33.00	812.9
N sweep	25	401	55	564	15.76	138.9
N sweep	25	401	110	564	16.15	233.8
N sweep	25	401	440	564	19.52	804.9

To quantify the implementation cost, we ran a one-replication benchmark using the same simulation code with $T = 5000$, $c = 0.65$, $\gamma = 0.999$, $\eta_{\text{train}} = 0.01$, and $B_0 = 80$. We vary one dimension at a time around the default $(K, G, N) = (25, 401, 220)$ and report DRRI training runtime and peak resident memory. The default setting takes 17.4 seconds and 424 MB peak RSS. Increasing the number of states from $K = 25$ to $K = 200$ increases runtime from 17.4 to 64.3 seconds and peak RSS from 424 MB to 1.49 GB, primarily because the number of cached solvers increases from 564 to 2106. Increasing G from 101 to 801 increases runtime from 5.6 to 33.0 seconds and peak RSS from 133 MB to 813 MB. Increasing N from 55 to 440 increases peak RSS from 139 MB to 805 MB, consistent with the linear NG value-table scaling.

These results are now summarized in the revised scalability paragraph. They also motivate the finite-state preprocessing choices in the FICO study: excessively fine state partitions increase the number of cached solvers and can produce rare states that are difficult to update under reject inference. The current implementation is therefore appropriate for small-to-moderate finite state spaces, while larger or continuous covariate spaces would require function approximation or additional solver sharing, which we now state as a limitation and future direction.

D ALGORITHM

In this section, we present the algorithm for the covariate-dependent case (Algorithm 1). In practice, we discretize the belief mean over a grid $\{\mu_g\}_{g=1}^G$ and run a backward recursion over the grid and truncation depth. Since successor means may not fall exactly on grid points, we use interpolation to evaluate continuation values.

The covariate-free setting is a special case of the covariate-dependent formulation: when $\mathcal{X} = \{x\}$ and $p_0(x) = 1$, we have $\mu_t(x) = \mu_t$, $\nu_t(x) = \nu_t$, $V_\eta^*(x; \mu(x), \nu(x)) = V_\eta^*(\mu, \nu)$, and $\bar{\gamma}(x) = \gamma$. Hence the state-dependent Bellman recursion and its backward approximation reduce to the state-independent ones after dropping the redundant index x and the outer loop over $\mathcal{X}$.

Algorithm 1 DRRI for the covariate-dependent case

Require: cost c, discount factor γ, robustness radius η, truncation level N, state space $\mathcal{X}$, state probabilities $\{p_0(x)\}_{x\in\mathcal{X}}$, belief grid $\{\mu_g\}_{g=1}^G$

for all $x \in \mathcal{X}$ **do**
  Compute the effective discount factor $\bar{\gamma}(x) = \frac{\gamma p_0(x)}{1-\gamma(1-p_0(x))}$.
  for each grid point μ_g **do**
    Compute $\bar{q}_{g,x}^*(\lambda^*)$: if $\eta \geq -\ln(1-\mu_g)$, then $\bar{q}_{g,x}^*(\lambda^*) = 0$; otherwise, $\bar{q}_{g,x}^*(\lambda^*) = \frac{\mu_g e^{-1/\lambda^*}}{\mu_g e^{-1/\lambda^*} + (1-\mu_g)}$, where λ^* solves $0 = \bar{q}_{g,x}^*(\lambda) \ln\Big(\frac{\bar{q}_{g,x}^*(\lambda)}{\mu_g}\Big) + \big(1 - \bar{q}_{g,x}^*(\lambda)\big) \ln\Big(\frac{1-\bar{q}_{g,x}^*(\lambda)}{1-\mu_g}\Big) - \eta$.
    Set the terminal value $V_\eta^N(x; \mu_g, N+1) = \frac{\max\{\bar{q}_{g,x}^*(\lambda^*) - c, 0\}}{1-\bar{\gamma}(x)}$.
  end for
  for $\nu = N, N-1, \ldots, 1$ **do**
    for each grid point μ_g **do**
      Compute successor means $\mu_g^+ = \frac{\mu_g \nu + 1}{\nu + 1}$ and $\mu_g^- = \frac{\mu_g \nu}{\nu+1}$.
      Interpolate next-stage values

$$V_\eta^+ = \mathrm{Interp}\big(V_\eta^N(x; \cdot, \nu+1), \mu_g^+\big) \quad \textit{and} \quad V_\eta^- = \mathrm{Interp}\big(V_\eta^N(x; \cdot, \nu+1), \mu_g^-\big).$$

      Define $\mathcal{D}(x; \mu_g, \nu) = \bar{\gamma}(x) V_\eta^- - \bar{\gamma}(x) V_\eta^+ - 1$.
      if $\mathcal{D}(x; \mu_g, \nu) = 0$ **then**
        Update $V_\eta^N(x; \mu_g, \nu) = \max\{0, \bar{\gamma}(x) V_\eta^- - c\}$.
      else
        Update

$$V_\eta^N(x; \mu_g, \nu) = \max\Big\{0,\ \big(q_{g,x}^*(\lambda^*) - c\big) + \bar{\gamma}(x)\Big[q_{g,x}^*(\lambda^*) V_\eta^+ + \big(1 - q_{g,x}^*(\lambda^*)\big) V_\eta^-\Big]\Big\}.$$

        where

$$q_{g,x}^*(\lambda^*) = \begin{cases} 0, & \text{if } \mathcal{D}(x;\mu_g,\nu) < 0 \text{ and } \eta \geq -\ln(1-\mu_g), \\ 1, & \text{if } \mathcal{D}(x;\mu_g,\nu) > 0 \text{ and } \eta \geq -\ln \mu_g, \\ \dfrac{\exp\Big(\dfrac{\mathcal{D}(x;\mu_g,\nu)}{\lambda^*} + \ln\Big(\dfrac{\mu_g}{1-\mu_g}\Big)\Big)}{1 + \exp\Big(\dfrac{\mathcal{D}(x;\mu_g,\nu)}{\lambda^*} + \ln\Big(\dfrac{\mu_g}{1-\mu_g}\Big)\Big)}, & \text{if } \mathcal{D}(x;\mu_g,\nu) < 0,\ \eta < -\ln(1-\mu_g) \\ & \quad \text{or } \mathcal{D}(x;\mu_g,\nu) > 0,\ \eta < -\ln \mu_g, \end{cases}$$

        and $\lambda^* > 0$ is the zero of the following equation

$$\eta = q_{g,x}^*(\lambda) \ln\left(\frac{q_{g,x}^*(\lambda)}{\mu_g}\right) + \big(1 - q_{g,x}^*(\lambda)\big) \ln\left(\frac{1 - q_{g,x}^*(\lambda)}{1-\mu_g}\right),$$

      end if
    end for
  end for
end for
return $\{V_\eta^N(x; \mu_g, \nu)\}_{x\in\mathcal{X},\, \nu=1,\ldots,N+1,\, g=1,\ldots,G}$

Algorithm 2 DRRI for the covariate-dependent case (faster version)

Require: cost c, discount factor γ, robustness radius η, truncation level N, state space $\mathcal{X}$, state probabilities $\{p_0(x)\}_{x\in\mathcal{X}}$, belief grid $\{\mu_g\}_{g=1}^{G}$

for all $x \in \mathcal{X}$ **do**
  Compute the effective discount factor $\bar{\gamma}(x) = \frac{\gamma p_0(x)}{1-\gamma(1-p_0(x))}$.
  for each grid point μ_g **do**
    Set $F_g(q) = \mathbb{D}(q, \mu_g) - \eta$, where $\mathbb{D}(q, \mu_g) = q \ln\Big(\frac{q}{\mu_g}\Big) + (1-q)\ln\Big(\frac{1-q}{1-\mu_g}\Big)$.
    Compute the lower KL interval endpoints $q_{L,g}$ and the upper KL interval endpoints $q_{U,g}$:
    if $-\ln(1-\mu_g) \le \eta$ **then**
      $q_{L,g} = 0$
    else
      Apply Bisection method on $F_g(q) = 0$ to find a zero in $[0, \mu_g]$. Set the zero as $q_{L,g}$.
    end if
    if $-\ln(\mu_g) \le \eta$ **then**
      $q_{U,g} = 1$
    else
      Apply Bisection method on $F_g(q) = 0$ to find a zero in $[\mu_g, 1]$. Set the zero as $q_{U,g}$.
    end if
    Initialize terminal row: $V_\eta^N(x; \mu_g, N+1) = \frac{\max\{q_{L,g}-c, 0\}}{1-\bar{\gamma}(x)}$.
  end for
  for $\nu = N, N-1, \ldots, 1$ **do**
    for each grid point μ_g **do**
      Compute successor means $\mu_g^+ = \frac{\mu_g \nu + 1}{\nu+1}$ and $\mu_g^- = \frac{\mu_g \nu}{\nu+1}$.
      Interpolate next-stage values:

$$V_\eta^+ = \mathrm{Interp}\big(V_\eta^N(x;\cdot,\nu+1), \mu_g^+\big), \qquad V_\eta^- = \mathrm{Interp}\big(V_\eta^N(x;\cdot,\nu+1), \mu_g^-\big).$$

      Compute slope: $s_{g,x} = 1 + \bar{\gamma}(x)\big(V_\eta^+ - V_\eta^-\big)$.
      Choose worst-case endpoint:

$$q_{g,x}^* = \begin{cases} q_{L,g}, & s_{g,x} > 0, \\ q_{U,g}, & s_{g,x} < 0, \\ \mu_g, & s_{g,x} = 0. \end{cases}$$

      Update value:

$$V_\eta^N(x; \mu_g, \nu) = \max\Big\{0,\ \big(q_{g,x}^* - c\big) + \bar{\gamma}(x)\Big[q_{g,x}^* V_\eta^+ + \big(1 - q_{g,x}^*\big)V_\eta^-\Big]\Big\}.$$

    end for
  end for
end for
return $\{V_\eta^N(x; \mu_g, \nu)\}_{x\in\mathcal{X},\, \nu=1,\ldots,N+1,\, g=1,\ldots,G}$

Nevertheless, Algorithm 1 requires the knowledge of $\bar{q}_g^*(\lambda^*)$, which requires computing λ^* and it can be time-consuming. Algorithm 2 is an equivalent version but it can be run faster. Again, we only present the algorithm for the covariate-dependent case.

In addition, computing $\bar{\gamma}(x)$ in Algorithms 1 and 2 requires the knowledge of $p_0(x)$, which is not known naturally. In the case where $p_0(x)$ is not known, we need to estimate $\bar{\gamma}(x)$: Let $I_1(x), \cdots, I_m(x)$ be the inter-arrival times observed at x thus far, the estimate of $\bar{\gamma}(x)$ is $\hat{\bar{\gamma}}(x)$ such that

$$\hat{\bar{\gamma}}(x) = \frac{1}{m}\sum_{i=1}^{m} \gamma^{I_j(x)}.$$

Ultimately, $\bar{\gamma}(x)$ in Algorithms 1 and 2 is replaced by $\hat{\bar{\gamma}}(x)$.

E COMPUTATION INFRASTRUCTURE

All studies and experiments are run on Dell Precision 7920 Workstations with Intel(R) Xeon(R) Gold 6256 CPU at 3.60GHz and three sets of NVIDIA Quadro GV100 GPUs. All models are implemented in Python 3.8. The versions of the main packages of our code are NumPy: 1.19.2, Pandas: 1.1.3, Matplotlib: 3.3.2.