J. Appl. Numer. Optim. 7 (2025), No. 3, pp. 333-358
Available online at http://jano.biemdas.com
https://doi.org/10.23952/jano.7.2025.3.04

ROBUST CONTEXTUAL BANDIT METHOD FOR OPTIMAL LOAN OFFERING

HANSHENG SUN[1], ROY KWON[2,*]

[1]*Internal Ratings Management, Bank of Nova Scotia, Canada*
[2]*Department of Mechanical and Industrial Engineering, University of Toronto, Canada*

Abstract. This paper proposes a Group-DRO enhanced doubly-robust contextual bandit approach to designing optimal policies for loan product offerings. This approach is particularly suited to high-stakes decision-making such as lending decisions, where one must leverage historical data (with inherent biases and uncertainties) to design future policies. By using doubly-robust estimation, we make efficient use of the data and mitigate bias from unknown logging propensities. By incorporating distributional robustness with group-based ambiguity sets, we ensure that the learned policy is insulated against worst-case shifts in each subgroup, thereby protecting the overall performance from crashing if, say, economic conditions change that strongly impact a minority group. By adding fairness constraints such as demographic parity or equal opportunity, we can align the policy with ethical and regulatory standards, ensuring that no group is left behind or unfairly treated by the automated decision process. We present empirical evidence on a small business credit card portfolio, demonstrating significant improvements over standard methods. This proposed framework contributes a step toward responsible AI in finance.

Keywords. Contextual bandit; Distributionally robust optimization; Doubly robust estimation; Fair lending decisions; Group-based ambiguity set.

2020 Mathematics Subject Classification. 90Bxx, 91-XX.

1. INTRODUCTION

Optimizing credit policy for loan offerings often involves balancing risk and profit while ensuring fairness. A contextual bandit approach can leverage historical data of credit offers and outcomes to learn a better offering policy. However, standard off-policy evaluation and learning methods that rely solely on logged data can be fragile when the deployment conditions differ from historical data. Recently, Kallus et al. [9] proposed a doubly robust distributionally robust (DR-DR) framework for off-policy evaluation and learning, combining doubly-robust estimation with distributional robustness to guard against such shifts. In parallel, fairness concerns arise because sensitive groups (e.g. defined by race or sex of the applicants) may experience disparate outcomes under a learned policy. Sagawa et al. [12] introduced Group-DRO, a distributionally robust optimization method that ensures strong performance on worst-case groups by minimizing the maximum loss among pre-defined groups. In this chapter, we refine the DR-DR contextual bandit framework for the application of a bank's line of credit offer policy,

*Corresponding author.
E-mail address: rkwon@mie.utoronto.ca (R. Kwon).
Received 6 June 2025; Accepted 10 September 2025; Published online 24 November 2025.

integrating a Group-DRO-style modification to enforce robust and fair performance across sensitive subgroups. Our approach restricts distributional uncertainty to within sensitive groups and uses a regularized objective to capture fairness considerations. We aim to maximize overall risk-adjusted return while maintaining high performance for each sensitive group. Including fairness regularization such as demographic parity or equal opportunity into the learning process may prevent discriminatory outcomes.

We provide a detailed formulation of the method, an algorithm with pseudocode, and a discussion of implementation details including logging policy estimation, fairness metrics, and evaluation protocols. We demonstrate the effectiveness of our proposed method with empirical experiments on a small business credit card portfolio.

In determining the grouping of the customers for Group-DRO policy learning, a practical approach is to rely on business knowledge; alternatively, we have proposed a quantitative approach using robust metric learning through a distributionally robust modification of the widely studied Large Margin Nearest Neighbor (LMNN) approach [13] ensuring that the learned metric is stable under moderate data perturbation, which we briefly discuss in the Appendix section.

2. Background: Contextual Bandits

A contextual bandit (CB) problem is a simplified reinforcement learning (RL) task. In full RL, an action in state influences not only the immediate reward but also the next state. By contrast, the contextual bandit formulation assumes an action in a given context only affects the immediate reward and does not influence subsequent contexts.

Formally, let:

- $\mathscr{X}$ be the **context (state) space**, with random variable $X \in \mathscr{X}$.
- $\mathscr{A}$ be a finite **action set** (also called arms): $\mathscr{A} = \{1, 2, \ldots, k\}$.
- $R : \mathscr{X} \times \mathscr{A} \to \mathbb{R}$ be a **reward function**, where $R(X, A)$ is the random reward obtained when action A is taken in context X.
- A policy $\pi : \mathscr{X} \to \mathscr{A}$ maps context vectors $x \in \mathscr{X}$ to actions $A \in \mathscr{A}$.

The policy value function $Q(\pi)$ is defined by $Q(\pi) = \mathbb{E}_P\big[R(X, \pi(X))\big]$, where P is the underlying joint distribution of $(X, R(X,A))_{A \in \mathscr{A}}$. In typical offline or batch contextual bandit data, we observe n samples $\{(x_i, a_i, r_i)\}_{i=1}^n$, where

- $x_i \in \mathscr{X}$ is the observed context for sample i,
- $a_i \in \mathscr{A}$ is the action taken by some logging policy (often unknown),
- $r_i = R(x_i, a_i)$ is the observed reward under action a_i at context x_i.

Note that for each i, we only observe r_i for the chosen action a_i; the rewards for other actions in $\mathscr{A} \setminus \{A_i\}$ are not observed (partial feedback). Specifically, in determining loan approval policy, the bank is unable to assess the performance of the rejected applicants. Furthermore, the logging loan approval policy may be unspecified or a mixture of expert rules, making it challenging to correct for selection bias. Additionally, data collected under heterogeneous economic conditions complicates the learning process, especially if we want a robust policy that generalizes to new or shifting environments.

Although contextual bandit approaches are often discussed in an online setting, where the agent selects actions based on each new or existing customer's context, balancing exploitation (choosing the best-known loan offer) and exploration (trying alternative offers to refine

estimates), in financial setting, large-scale randomization can be restricted by regulatory and fairness considerations, so offline or batched contextual bandit is the focus of our discussion.

3. DR-DR Batched Contextual Bandit Off-policy Evaluation

In a contextual bandit, at each decision point we observe a context (feature vector) X (e.g. business attributes), choose an action A (e.g. offer credit or not), and receive an outcome reward R (e.g. return on regulatory capital from repayment/default). The goal in our setting is to learn a policy $\pi(A|X)$ that decides when to offer credit to maximize long-term return. Since experimentation is against regulation, it is necessary use historical logged data from a past policy (logging/behavior policy) to evaluate and learn a new policy without deploying it (off-policy). Off-policy evaluation (OPE) estimates the performance of π using logged data $(x_i, a_i, r_i)_{i=1}^n$ collected under a possibly different logging policy π_b. A basic OPE estimator is inverse propensity scoring (IPS), which reweights outcomes by the ratio $\pi(a_i|x_i)/\pi_b(a_i|x_i)$. However, IPS can be high-variance and is biased if π_b is unknown or recorded poorly. A more advanced estimator is doubly robust (DR) OPE, which combines IPS with an outcome model's direct estimation (DE) to remain consistent if either the propensity model or outcome model is correct, concretely,

We have contexts $X \in \mathscr{X}$, actions $A \in \{1, \ldots, k\}$, and reward $R(X, A)$. Historical data

$$\{(x_i, a_i, r_i)\}_{i=1}^n$$

is collected by an unknown or partially known logging policy π_b. The goal is to evaluate a given new policy π and eventually learn an optimal policy that maximizes expected risk-adjusted return.

Combine inverse propensity scoring (IPS) and direct estimation (DE) to form the doubly robust (DR) estimator [4] for the policy value:

$$\hat{V}(\pi) = \frac{1}{n} \sum_{i=1}^{n} \Big[\frac{(r_i - \hat{\rho}(x_i, a_i))\, \pi(a_i|x_i)}{\hat{\pi}_b(a_i|x_i)} + \sum_{j=1}^{k} \pi(a_j|x_i)\, \hat{\rho}(x_i, a_j) \Big],$$

where $\hat{\rho}$ estimates the reward model and $\hat{\pi}_b$ models the logging policy. This helps reduce bias or variance as long as one of $\hat{\rho}$ or $\hat{\pi}_b$ is accurate.

Despite its advantages, conventional off-policy evaluation assumes that the test distribution is the same as historical. In practice, distribution shift may occur: the covariate distribution of small businesses seeking credit in the future, or their behavioral patterns, may differ from the historical data. Standard OPE is sensitive to such shifts, which can lead to overestimating performance if the policy exploits areas with little reliable data. To address this, distributionally robust OPE techniques consider an ambiguity set of plausible test distributions around the empirical distribution and evaluate a policy on the worst-case distribution in that set. By optimizing a policy for its worst-case performance, we can ensure more reliable returns under unforeseen changes. Recently, [9] introduced a distributionally robust DR contextual bandit framework for off-policy evaluation and off-policy learning, aiming to reduce the impact of model misspecification and distributional shifts for ambiguity set characterized by KL-divergence. The distributionally robust framework defines an ambiguity set around the empirical distribution of the context–action–reward tuples. The goal is to guard against the worst-case distribution in this ambiguity set while maintaining the doubly robust property. Formally, let $\hat{P}_n$ denote the

empirical distribution. For a given target policy π, ambiguity set $\mathscr{U}(\hat{P}_n, \delta)$, which we refer to as $\mathscr{U}(\delta)$ to simplify the notation.

The distributionally robust doubly robust policy value is then defined as:

$$V_\delta(\pi) = \inf_{P_1 \in \mathscr{U}(\delta)} \mathbb{E}_{P_1}[R(\pi(X))]. \tag{3.1}$$

That is, $V_\delta(\pi)$ evaluates policy π under the worst-case distribution P_1 within the ambiguity set $\mathscr{U}$, which we choose to be the KL -divergence ball of radius δ around the empirical distribution $\hat{P}_n$, ensuring robustness to possible shifts in the environment. The policy evaluation algorithm is detailed in the Appendix.

Policy learning aims to find a near-optimal robust policy $\hat{\pi} \in \Pi$ with small regret in worst-case policy value $\mathscr{R}_\delta(\pi) := V_\delta(\pi^*) - V_\delta(\pi)$, where $\pi^* \in argmax_{\pi \in \Pi} V_\delta(\pi)$

This approach aims to yield a policy that is robust to both model misspecification and local distributional shifts. While this method addresses robustness, guarding against arbitrary shifts sometimes yields overly conservative results and leads to unrealistic policy may impact the performance.

4. GROUP-DRO-DR OFF-POLICY EVALUATION

In practice, banks usually divide their customers into difference risk tiers or based on other risk-based grouping criteria, as a result, it may be useful to consider a specific type of DRO set-up.

Group-DRO is a special case of distributional robustness focusing on predefined sensitive groups, it guards against worst-case reweighting of groups. The policy π hedges against both uncertainty in group composition and distributional shifts:

- Uncertainty in group composition: the weights w_g shift across groups.
- Distributional shifts within groups: distributional shifts within each group.

Group specific robust value, for each group g, the distributionally robust policy value is:

$$V_{g,\delta_g}(\pi) = \inf_{\mathbb{P}_{1,g} \in \mathscr{U}(\delta_g)} \mathbb{E}_{\mathbb{P}_{1,g}}[R(\pi(X))],$$

where:

$$\mathscr{U}(\delta_g) = \left\{ \mathbb{P}_{1,g} : \mathbb{P}_{1,g} \ll \mathbb{P}_{0,g}, D_{KL}(\mathbb{P}_{1,g} \| \mathbb{P}_{0,g}) \leq \delta_g \right\}.$$

With $P_{1,g} << P_{0,g}$, the worst-case distribution $P_{1,g}$ is absolutely continuous with respect to the nominal distribution $P_{0,g}$, i.e. $P_{1,g}$ does not assign positive probability to events that $P_{0,g}$ consider impossible (probability 0), ensures $\frac{dP_{1,g}}{dP_{0,g}}$ (density ratio) exists for defining KL divergence and dual form, ensuring that the perturbed distribution has the same support as the empirical. By duality, this becomes $V_{g,\delta_g}(\pi) = \max_{\alpha_g > 0} \left[-\alpha_g \log W_g(\pi, \alpha_g) - \alpha_g \delta_g \right]$, with

$$W_g(\pi, \alpha_g) = \mathbb{E}_{\mathbb{P}_{0,g}} \left[\exp\left(-\frac{R(\pi(X))}{\alpha_g} \right) \right].$$

The overall robust value is

$$V_{robust}(\pi) = \inf_{w \in \mathscr{U}_w} \sum_{g=1}^{G} w_g V_{g,\delta_g}(\pi),$$

where

$$\mathscr{U}_w = \left\{ w : \sum_{g=1}^{G} w_g = 1, w_g \geq 0, D_{KL}(w \| w_0) \leq \delta_w \right\}.$$

Dual reformulation is given by the following.

For $c_g = V_{g,\delta_g}(\pi)$, the infimum over w is:

$$\inf_{w \in \mathscr{U}_w} \sum_{g=1}^{G} w_g c_g = \max_{\beta > 0} \left[-\beta \log \left(\sum_{g=1}^{G} w_{0,g} \exp\left(\frac{c_g}{\beta} \right) \right) - \beta \delta_w \right].$$

Thus

$$V_{robust}(\pi) = \max_{\beta > 0} \left[-\beta \log \left(\sum_{g=1}^{G} w_{0,g} \exp\left(\frac{V_{g,\delta_g}(\pi)}{\beta} \right) \right) - \beta \delta_w \right].$$

For joint maximization, we define

$$\phi_g(\pi, \alpha_g) = -\alpha_g \log W_g(\pi, \alpha_g) - \alpha_g \delta_g,$$

so $V_{g,\delta_g}(\pi) = \max_{\alpha_g > 0} \phi_g(\pi, \alpha_g)$. Substitute and perform joint maximization:

$$V_{robust}(\pi) = \max_{\beta > 0, \{\alpha_g > 0\}_{g=1}^{G}} \left[-\beta \log \left(\sum_{g=1}^{G} w_{0,g} \exp\left(\frac{-\alpha_g \log W_g(\pi, \alpha_g) - \alpha_g \delta_g}{\beta} \right) \right) - \beta \delta_w \right].$$

Estimate $W_g(\pi, \alpha_g)$ using group-specific doubly robust estimators, solving moment conditions for $\widehat{\alpha}_g$ and $\widehat{\beta}$. The steps for policy evaluation is detailed in the Appendix section in Algorithm 3.

For policy learning, we propose the Continuum Doubly Robust Group-DROP (CDR-Group-DROP) Algorithm 4 detailed in the Appendix section.

Another important aspect of our formulation is explicitly incorporating fair lending consideration, where we expand the objective function to incorporate fairness penalty, such as, a demographic parity constraint can ensure the offer rate to each group is at least some fraction of the population or equal across groups. A more commonly adopted approach is to mandate equal opportunity: for customers who would repay (outcome R positive), the probability of being offered credit is equal across groups, which we will discuss further in the subsequent section.

5. GROUP-DRO DR CONTEXTUAL BANDIT FOR SMALL BUSINESS LOAN OFFERING

In this section, we apply the proposed policy evaluation and policy learning algorithms to the credit limit offering problem, where the goal is to assign credit limits to small business applications while maximizing a robust policy value and ensuring fairness across demographic subgroups.

The state space X is defined by the features of small business applications, including risk category and majority/minority status of the owners. The action space A is the discrete credit limit options. The policy π_θ is a softmax regression model that assigns the probabilities to each credit limit based on application features. The reward R is the risk-adjusted return of receiving a specific credit limit. The objective is to maximize a robust policy value while ensuring equal opportunity within risk groups across majority and minority subgroups.

The policy is

$$\pi_\theta(a|x) = \frac{\exp((W^{(2)}z + b^{(2)})_a)}{\sum_{a'=1}^{|A|}\exp((W^{(2)}z + b^{(2)})_{a'})},$$

where $z = \sigma(W^{(1)}x + b^{(1)})$, $\sigma(x) = \max(0,x)$, and $\theta = \{W^{(1)}, b^{(1)}, W^{(2)}, b^{(2)}\}$ are the parameters for the neural network policy class with one hidden layer.

Groups are defined by risk category G and majority/minority status. The group-specific robust value is

$$V_{g,\delta_g}(\pi) = \max_{\alpha_g>0}\left[-\alpha_g \log W_g(\pi,\alpha_g) - \alpha_g\delta_g\right],$$

where $W_g(\pi,\alpha_g) = \mathbb{E}_{\mathbb{P}_{0,g}}[\exp(-R(\pi(S))/\alpha_g)]$, and $\mathscr{U}(\delta_g)$ is a KL-divergence uncertainty set.

The robust policy value is:

$$V_{robust}(\pi) = \max_{\beta>0,\{\alpha_g>0\}}\left[-\beta\log\left(\sum_{g=1}^{G} w_{0,g}\exp\left(\frac{\phi_g(\pi,\alpha_g)}{\beta}\right)\right) - \beta\delta_w\right]$$

with $\phi_g(\pi,\alpha_g) = -\alpha_g \log W_g(\pi,\alpha_g) - \alpha_g\delta_g$, and $\mathscr{U}_w$ defined via KL-divergence.

The doubly robust estimator $\widehat{V}^{DR}_{robust}(\pi_\theta)$ is computed by:

- Estimating $\widehat{\pi}_{0,g}$ and $\widehat{f}_{0,g}(s,a;\alpha_g)$ for each group using random forests.
- Computing $\widehat{W}^{DR}_g(\pi,\alpha_g)$ with cross-fitting.
- Optimizing the dual variables β and $\{\alpha_g\}$.

The fairness penalty is

$$P(\theta) = \sum_{k=1}^{K}\sum_{a\in A}\left|\widehat{\mathbb{E}}[\pi_\theta(a|X)\mid g(X)=k, m=0] - \widehat{\mathbb{E}}[\pi_\theta(a|X)\mid g(X)=k, m=1]\right|.$$

The objective is

$$\max_\theta\left[\widehat{V}^{DR}_{robust}(\pi_\theta) - \lambda P(\theta)\right].$$

6. Application on Adjudication of Revolving Loans

In this section, we apply our proposed approach to a real-world North American small business line of credit dataset and compare the empirical result with some industry standard benchmark approaches.

6.1. **Business Background.**

We focus our discussion on the business application of evaluation and determination line of credit offers to new customers. The goal of the bank is to maximize its portfolio risk-adjusted return. The method discussed is applicable to other scenarios in making business decisions under incomplete data, where the logging policy may create systematic bias in historical data.

Revolving credits such as personal line-of-credits or credit cards are offered by a bank to meet its customers' on-going demand for funds. A customer can draw on the loan facility, and the interest is only paid on the actual withdrawn outstanding amount. Bank's primary interest is to maximize risk adjusted return of the portfolio, which can be achieved by imposing an optimal policy in the credit offer to their customers.

A widely adopted approach as described in Haimowitz [8] follows a two-step approach by first constructing a risk-return matrix that partitions the bank's customer base, then determine

an appropriate credit limit policy for each cell in the risk-return matrix either through quantitative analysis or subjectively using business judgment. This industry standard approach has a few drawbacks: most notably the performance of the trained policy is highly dependent on the representativeness of the training data, it may have high expected reward under the static environment but quickly deteriorate when customers' behavior changes or risk increases under a different future environment; in addition, the there is no fair lending consideration explicitly incorporated in the policy learning process. Our proposed distributional robust method can effectively mitigate the shortcomings of the industry standard approach. In consideration of fair lending, we specifically incorporate the sensitive attribute indicating whether the business owner belongs to a visible minority group. To enforce fairness explicitly in the policy learning process, we introduce an equal opportunity constraint, requiring that, for applicants with comparable credit risk ratings, membership in a minority group does not reduce the likelihood of receiving a favorable credit offer. This constraint directly embeds fairness into policy optimization, providing stronger guarantees against bias compared to the industry-standard approach of simply omitting sensitive attributes such as gender or race. Merely excluding sensitive features can inadvertently perpetuate bias through correlated or confounding attributes, whereas our explicit constraint approach effectively mitigates such risks.

6.2. **Data Description.** A small business line of credit portfolio with a sample of 16000 observations from 9500 customers from September 2014 to September 2018 is selected for our analysis. There are 5 actions of the bank can take for the customer: offering a credit limit with normal distribution averaged at $\{5000, 20000, 40000, 60000, 80000\}$ with standard deviation of 2000 to reflect the flexibility usually exercised by the bank's staff in actual loan offering. The customer's total credit limit before the adjustment and its subsequent performance data (unavailable at the time of action) including monthly utilization rates, outstanding balances, credit ratings and the corresponding long-run PD are collected, from which we can calculate the risk-adjusted-return of the customers as the observed empirical reward associated with policy. Specifically, the total accumulated spending and the outstanding balance are used to calculate the present value of the cumulative revenue generated by the customer over a 24-month period. The credit risk rating of the customers is assigned by the bank and can be mapped to the long-run probability of default to estimate customers' regulatory capital. We assume that the credit risk ratings of the customers are assessed periodically and available. The risk adjusted return is defined as the return on regulatory capital as the reward of the customer, which are defined in Equation 7.5 in Appendix.

The context vector consists of 15 raw input features listed in Table 1. The geographical area of service is a feature that categorizes the population density on the geographical location that the customer operating in. Business of operation categorizes the type of business that the customers are operating, including: restaurants, retailers, farmers, professional service, etc.

6.3. **Training and Evaluation Procedure.** To comprehensively evaluate the performance of the out-of-sample learned policy, we select customers from a subset of postal codes and years to train the policy and evaluate its performance on samples from other postal codes and years. We repeat this process by selecting 3 sets of training periods and postal codes and then compare the average out-of-sample performance.

TABLE 1. List of raw input features

Geographical Area of Service	Business of Operation	Total Asset
Total Revenue	Total Operating Income	Total Debt
Total Long Term Debt	Total Current Liability	Total Inventory
Account Payable	Cost of Sales	Number of Years in Operation
Owner's Credit Bureau Score	Owner's Recent Year Delinquency	Number of Employee

TABLE 2. Service Area and Business of Operation Composition Top 8

Service Area	**Proportion**	**Business of Operation**	**Proportion**
High Density Commercial with Residential	0.289	Restaurant	0.232
High Density Commercial	0.163	Retail	0.203
High Density Residential	0.147	Professional Service	0.177
Rural	0.097	Farm	0.121
Medium Density Commercial	0.094	Construction	0.075
Industrial	0.079	Manufacturing	0.050
Medium Density Residential	0.070	Health Care	0.050
Low Density Commercial	0.051	Oil Gas Service	0.030

TABLE 3. Credit Rating corresponding PD and Companies' Median Key Financial Metrics (in thousand USD)

Customer Credit Rating	**Corresponding PD**	**Total Asset**	**Total Revenue**	**Total Debt**
BB+	0.163%	4798.3	554.26	2294.2
BB	0.264%	3691.5	495.68	2352.1
BB-	0.426%	3342.4	396.40	2391.5
B+	0.689%	3295.03	362.67	2981.4
B	0.849%	3011.08	231.29	1913.3
B-	1.3282 %	2960.15	170.14	2099.9
CCC+	2.5597 %	2858.39	74.621	2848.5
CCC	5.48%	1677.06	32.370	2314.0

Customers are divided into four groups according to their credit risk profiles. Each group is further split into two subgroups based on the minority status of the owners' only for setting the equal opportunity constraint.

The grouping of customers based on risk profiles can be roughly described as follows:

- Group 1 (Low Risk): Customers who are on time with payments pay all of monthly balance. These customers have moderate spending activity, have high credit ratings with BB- and above, and lowest risk. Many Professional Service providers belong to this group. This group of customers offer high risk adjusted return due to low credit risk.
- Group 2 (Moderate Risk): Customers who pay on time usually pay most of the monthly balance. These customers have high spending and receive mixed credit risk ratings. This group includes many franchise owners, restaurants, and retailers in high-density

commercial and residential locations. The group has the highest profitability for the bank.

- Group 3 (Medium Risk): Customers who are risky. These customers usually pay a portion of the monthly balance, with high spending, and could become delinquent during economic downturn. The group spans a wide range of operations, is profitable for the bank under normal condition.
- Group 4 (High Risk): Customers who are highly risky with low credit risk ratings and only paying a portion of their monthly balances. These customers are not profitable due to high credit risk.

Intuitively, a bank would be able to maximize its risk-adjusted return by introducing a policy that assigns high credit limit to Group 1, Group 2 and low limit to Group 4. The grouping of the customers are assessed at initiation; the group composition usually changes in the subsequence years. Within each group, the customers' behavior also varies.

Estimating logging policy $\hat{\pi}_b$ of the training data can be modeled as multi-class classification. Estimating the reward function $\hat{R}_g(\pi(X))$ can be modeled using least square regression. We use popular state-of-the-art XGBoost [2] package, which is based on Gradient Boosting algorithm [6] for both multi-class classification and regression. The hyper-parameters are selected with 5-fold cross-validation with parameter max-depth: [3, 4, 5] , learning rate: [0.1, 0.3, 0.5], early-stopping round: [5, 10], regularization parameter λ: [50, 75, 100] and max-iterations: 1000.

For policy learning, we implement a simple neural network with a linear hidden layers with hyperbolic tangent activation function, and soft-max loss function objective.

6.4. **Experiment.** In summary, the application involves a contextual bandit problem with four groups ($G = 4$) based on credit risk ratings of the small business customers, each group can be divided into two sub-groups based on minority status of the owner ($m = 2$)

- State Space: $\mathscr{X}$ contains feature vectors obtained from transformations of the raw features from Table 1.
- Action Space: $\mathscr{A} = \{5000, 20000, 40000, 60000, 80000\}$, average credit limit offered to the customer.
- Behavior Policy: each group g fit a gradient-boosted classifier to predict $\hat{pi}_{b,g(X)}$.
- Rewards Estimator: For group g for a gradient-boosted regressor to predict the rewards.
- Group Weights: Nominal weights $w_0 = (0.32, 0.38, 0.21, 0.09)$ corresponding to average group compositions over history.

The size of the uncertainty set was determined through grid-search for the following grid:

- Group Uncertainty Radii: $\delta_g \in \{0.01, 0.05, 0.1, 0.15\}$ for all groups.
- Weight Uncertainty Radius: $\delta_w \in \{0.03, 0.05, 0.7\}$.

For larger problems, one can consider the following approach for determining the size of the uncertainty sets. For δ_g (Group-Specific Radius): we start with a chi-squared KL balls: $\delta_g = \frac{\chi^2_{d,1-\alpha}}{2n_g}$, where n_g is samples per group, d is number of features, $\alpha = 0.05$ (confidence), and χ^2 is the chi-squared quantile. Set larger δ_g if the group has larger variance. For δ_w (Weight Radius): Since weights are over G groups, use a simpler bound:$\delta_w = \frac{\log(1/\alpha)}{2N}$, where N is total samples. Scale by a factor of 0.5, 1.5, 2 etc. and select based on the performance on the validation set.

TABLE 4. Comparison of risk adjusted returns

	Mean	Std.	min	10th percentile	20th percentile	30th percentile
$\hat{\pi}_{B1}$	1	0.059	0.673	0.737	0.781	0.852
$\hat{\pi}_{B2}$	1.044	0.061	0.642	0.760	0.795	0.851
$\hat{\pi}_{DR}$	1.051	0.058	0.657	0.773	0.801	0.855

For policy evaluation, we adopt the Algorithm3 (LDR^2-Group-DROPE) detailed in the Appendix and policy learning using Algorithm 4 CDR^2-Group-DROPL).

6.5. **Benchmark approaches.** We compare our proposed approach with industry standard benchmark approaches:

The first benchmark approach is the "argmax policy", where we directly use weighted least square regression to predict the reward given X_i and A_i, with importance weight $\frac{1}{\hat{(}\pi_b(A_i|X_i))}$. This approach reduces the policy learning problem to a standard regression problem. The optimal policy is the action associated with the highest predicted reward. We learn a separate regressor for each action. We use the popular state-of-the-art XGBoost [2] to train the regressors. The hyper-parameters are selected using 5-fold cross validation, with max-depth: [3, 4, 5] , learning rate: [0.1, 0.3, 0.5], max-iterations: 1000 and early-stopping round: 10, which means we stop if we experience 10 of rounds without improvements.The predictor $\hat{\pi}_b$ follows a approach as discussed previously.

The second benchmark approach is to use the importance weighted multi-class classification, where we convert each observation $(X_i, \frac{R_i}{\hat{\pi}_b(A_i|X_i)})$ and $\frac{R_i}{\hat{\pi}_b(A_i|X_i)}$ is the cost of not predicting label A_i on input X_i. Specifically in 3 Steps. Step 1, convert each observation: Historical data is (x_i, a_i, r_i), where x_i is context, a_i is logged action, r_i is reward. To turn this into classification: For each sample i, create pseudo-samples for all possible actions (not just the observed a_i). For the observed action a_i, create a positive example with label $= 1$ (or weighted by r_i if rewards are used as pseudo-labels). For unobserved actions $a \neq a_i$, create negative examples with label $= 0$ (or pseudo-label based on estimated rewards). It expands the dataset from N samples to $N \times A$, where A is number of actions, allowing the classifier to learn probabilities $\pi(a|x)$ across all actions. Without this, classification would only learn the logging policy, not a new one. Step 2, importance weighting (Inverse Propensity Scoring - IPS): To handle off-policy bias (data is from logging policy π_b, not the target policy), weight each pseudo-sample by the inverse propensity: $w_i = \frac{1}{\hat{\pi}_b(a_i|x_i)}$ for observed actions (high weight if action was rare), and 0 or clipped for unobserved. The loss is weighted cross-entropy: Minimize $\sum_i w_i \cdot \ell(\hat{\pi}(a_i|x_i), y_i)$, where y_i is the pseudo-label. Cap weights at 10 to prevent variance explosion. Step 3 model training : Use gradient boosted decision tree algorithms XGboost and output Softmax probabilities $\hat{\pi}(a|x)$ for each action. For inference: Select action $\arg\max_a \hat{\pi}(a|x)$. The XGboost hyper-parameters are selected using 5-fold cross validation, with max-depth: [4, 5, 6], learning rate: [0.3, 0.5, 0.7], max-iterations: 1000 and early-stopping round: 10. The predictor $\hat{\pi}_b$ follows an approach as discussed previously.

6.6. **Experimental Result and Interpretation without fairness penalty.** The proposed policy has higher out-of-sample average risk adjusted returns compared to the benchmark approaches, as presented in Table 4, scaled by the return of the first benchmark policy.

TABLE 5. Comparison of average credit limit for small sample groups

	CL. with increasing risk		CL. with increasing spending	
Policy	Mean	Std.	Mean	Std.
$\hat{Q}(\hat{\pi}_{B1})$	31422	9048	42511	12107
$\hat{Q}(\hat{\pi}_{B2})$	32297	9665	45724	13531
$\hat{Q}(\hat{\pi}_{DRO})$	28613	7954	49123	12587

The average policy return for $\hat{\pi}_{DR}$ is significantly higher than that for $\hat{\pi}_{B1}$, with a p-value less than 0.001. Although the average policy return for $\hat{\pi}_{DR}$ is not significantly higher than that for $\hat{\pi}_{B2}$, a p-value of approximately 0.065 suggests the proposed policy often leads higher average risk adjusted returns.

Our method avoids picking the policy that performs well for the highly represented groups but bad for the atypical group. Any changes in economic or competitive environment may lead to changes in the customer risk profile, especially during recession, some customers from moderately delinquent group may become highly delinquent. Our proposed policy assigns a lower average credit limit than other benchmark methods for the customers that experience increase in credit risk during the out-of-sample testing period. In addition, although outstanding balance is often associated with high credit risk, there is an uncommon group of customers with high utilization rate and low credit risk usually gives high risk adjusted return. Our proposed policy assigns a higher average credit limit than the benchmark methods for the customers that show increase in spending while maintaining comparable or lower credit risk during the out-of-sample testing period.

The policies are evaluated under the following assumptions, if the credit limit assigned by the new policy is lower than data collection policy and transaction amount and outstanding balance is above the new policy assigned credit limit then the spending amount and outstanding balance are capped at the new credit limit. If the transaction amount and the outstanding balance is below the new policy assigned credit limit, then the full amount is used, however, if the facility is fully drawn and the new credit limit is higher, we apply the average utilization rate of the risk grade as a proxy to the incremental limit to estimate the incremental balance and use the customer's average repayment rate as a proxy to estimate incremental outstanding amount.

6.7. **Experimental Result and Interpretation including fair-lending penalty.** The average credit limit received for the Caucasian and Minority owners is comparable for the low-risk (Group 1) customers. However, for the higher risk groups (Group 2, 3 and 4) the minority owners receives notably lower limit if there is no fair-lending consideration.

The refined Group-DRO approach yields policies that are more stable to composition shifts in, say, the proportion of high-risk or high-return segments. We observe an improvement in out-of-sample risk-adjusted returns by increasing the average credit limit offered to the low to mild risk group. Equal opportunity penalty ensures the minority group receives better offers; however, leads to small reduction in the overall return of the portfolio by increasing the average credit limit of the high risk group.

TABLE 6. Comparison of avg. credit limit to Caucasian vs. Minority group

	Low Risk Group		Moderate Risk Group		Medium Risk Group		High Risk Group	
Policy	Cau.	Minor.	Cau.	Minor.	Cau.	Minor.	Cau.	Minor.
$\hat{Q}(\hat{\pi}_{B1})$	56524	55521	46973	41603	36421	32053	21234	15131
$\hat{Q}(\hat{\pi}_{B2})$	55259	54536	46782	43500	36323	32921	22162	14323
$\hat{Q}(\hat{\pi}_{DRO})$	56125	55756	49943	47851	35583	33681	21242	18100

The proposed method can be extended to other revolving credit portfolio management applications such as determining repayment term and interest rate offer, or more generally any business applications involving making policy decisions under a dynamic environment based on historical observational data.

7. APPENDIX

7.1. **Distributionally Robust DR Policy Evaluation.** To compute the distributionally robust policy value for a specific group, we follow the approach outlined in recent paper [9] where the author extended the distributionally robust policy evaluation method initially proposed in [10] to a doubly-robust (DR) estimator:

For ambiguity set given by KL divergence with size δ:

$$\mathscr{U}(\delta) = \{P_1 : P_1 \ll P_0 \text{ and } D_{\mathrm{KL}}(P_1 \,\|\, P_0) \le \delta\}.$$

Under regularity conditions, unconfoundedness and strong overlapping [10], for a policy π, and size of the ambiguity set δ, the distributionally robust value $V_\delta(\pi)$ is given by Equation 3.1.

$$V_\delta(\pi) := \inf_{P_1 \in \mathscr{U}(\delta)} \mathbb{E}_{P_1}[R(\pi(X))].$$

Although infinite-dimensional infimum is intractable, it is equivalent to solving supremum over a dual variable α (Lemma 1 in [10]). Specifically by the following steps:

Step 1: Reformulate the Infimum as a Constrained Optimization.

The problem is

$$\inf_{P_{1,g}} \mathbb{E}_{P_{1,g}}[R(\pi(X))] \quad \text{subject to} \quad D_{\mathrm{KL}}(P_{1,g} \| P_{0,g}) \le \delta_g, \quad P_{1,g} \ll P_{0,g}.$$

Let $\rho = \frac{dP_{1,g}}{dP_{0,g}}$ (density ratio, $\rho > 0$, $\mathbb{E}_{P_{0,g}}[\rho] = 1$). Change the expectation to be over $P_{0,g}$:

$$\mathbb{E}_{P_{1,g}}[R(\pi(X))] = \mathbb{E}_{P_{0,g}}[\rho \cdot R(\pi(X))].$$

The KL constraint becomes $D_{\mathrm{KL}}(P_{1,g} \| P_{0,g}) = \mathbb{E}_{P_{0,g}}[\rho \log \rho] \le \delta_g$. So the problem is

$$\inf_{\rho > 0} \mathbb{E}_{P_{0,g}}[\rho R(\pi(X))] \quad \text{subject to} \quad \mathbb{E}_{P_{0,g}}[\rho \log \rho] \le \delta_g, \quad \mathbb{E}_{P_{0,g}}[\rho] = 1.$$

Step 2: Form the Lagrangian.

Introduce Lagrange multipliers $\alpha_g \ge 0$ for the KL inequality and μ for the normalization equality:

$$\mathscr{L}(\rho, \alpha_g, \mu) = \mathbb{E}_{P_{0,g}}[\rho R(\pi(X))] + \alpha_g \left(\mathbb{E}_{P_{0,g}}[\rho \log \rho] - \delta_g\right) + \mu \left(\mathbb{E}_{P_{0,g}}[\rho] - 1\right).$$

The dual function is $g(\alpha_g,\mu)=\inf_{\rho>0}\mathscr{L}(\rho,\alpha_g,\mu)$, and the dual problem is $\sup_{\alpha_g\geq 0,\mu} g(\alpha_g,\mu)$.

Step 3: Solve the Inner Infimum over ρ.

For fixed α_g,μ, the infimum is pointwise over the functional. The integrand is

$$\rho R(\pi(X))+\alpha_g\rho\log\rho+\mu\rho+\text{constants}.$$

To minimize, we take the functional derivative w.r.t. ρ and set to zero

$$R(\pi(X))+\alpha_g(\log\rho+1)+\mu=0$$

$$\implies \log\rho=-\frac{R(\pi(X))+\mu+\alpha_g}{\alpha_g} \implies \rho^*=\exp\left(-\frac{R(\pi(X))+\mu+\alpha_g}{\alpha_g}\right).$$

Step 4: Enforce Normalization and Plug Back.

The normalization $\mathbb{E}_{P_{0,g}}[\rho^*]=1$ gives

$$\mathbb{E}_{P_{0,g}}\left[\exp\left(-\frac{R(\pi(X))}{\alpha_g}\right)\right]\cdot\exp\left(-\frac{\mu+\alpha_g}{\alpha_g}\right)=1 \implies \exp\left(-\frac{\mu+\alpha_g}{\alpha_g}\right)=\frac{1}{W_g(\pi,\alpha_g)},$$

where

$$W_g(\pi,\alpha_g)=\mathbb{E}_{P_{0,g}}\left[\exp\left(-\frac{R(\pi(X))}{\alpha_g}\right)\right].$$

Thus $\mu=-\alpha_g-\alpha_g\log W_g(\pi,\alpha_g)$. Plugging ρ^* into the Lagrangian and simplifying (the terms cancel appropriately), the dual function becomes

$$g(\alpha_g,\mu)=-\alpha_g\delta_g-\alpha_g\log W_g(\pi,\alpha_g).$$

Step 5: Dual Problem.

Maximizing over $\alpha_g>0$ (note μ is eliminated, and $\alpha_g=0$ recovers the nominal case), we obtain $V_{g,\delta_g}(\pi)=\max_{\alpha_g>0}\left[-\alpha_g\log W_g(\pi,\alpha_g)-\alpha_g\delta_g\right]$, where

$$W(\pi,\alpha):=\mathbb{E}\left[\exp\left(-\frac{R(\pi(X))}{\alpha}\right)\right].$$

The function $\phi(\pi,\alpha)$ is strictly concave and attains its maximum at a unique value $\alpha^*\in(0,1/\delta]$. Kallus et. al. proposed [9] to estimate $V_\delta(\pi)$ in a doubly robust way, this, however, requires estimating a continuum of regression functions parameterized by the dual variable α, to overcome this they proposed to case the estimation of $\alpha^*(\pi)$ and $V_\delta(\pi)$ into a joint moment estimation problem, then develop a localized doubly robust algorithm. For a strictly concave function $\phi(\pi,\alpha)$ observe that α^* is the unique root to $\frac{\partial\phi(\pi,\alpha)}{\partial\alpha}=0$

$$-\log W_0(\pi,\alpha^*)-\frac{W_1(\pi,\alpha^*)}{\alpha^* W_0(\pi,\alpha^*)}-\delta=0,$$

where $W_j(\pi,\alpha):=\mathbb{E}[R(\pi(X))^j exp(-R(\pi(X))/\alpha)]$. We also have $V_\delta(\pi)=-\alpha^*\log W_0^*-\alpha^*\delta$. Therefore, estimation of α^* and $V_\delta(\pi)$ is equivalent to estimating the root of the following moment equation with parameter $\theta=[\alpha,W_0,W_1,V_\delta]^T$

$$\mathbb{E}\left[U(R(\pi(S));\alpha)+V(\theta)\right]=\mathbf{0},$$

where

$$\psi(z;\theta,\eta_1(z;\alpha),\eta_2(z))=\frac{\pi(a|s)}{\eta_2(s,a)}\left(U(r;\alpha)-\eta_1(s,a;\alpha)\right)+\mathbb{E}_{a\sim\pi(s)}\left[\eta_1(s,a;\alpha)\right]+V(\theta),$$

and

$$U(r;\alpha)=\begin{bmatrix}\exp(-r/\alpha)\\ r\exp(-r/\alpha)\\ 0\\ 0\end{bmatrix},\quad V(\theta)=\begin{bmatrix}-W_0\\ -W_1\\ -\delta-\log W_0-\frac{W_1}{\alpha W_0}\\ -V_\delta-\alpha\log W_0-\alpha\delta\end{bmatrix}$$

Specifically, doubly robust distributionally robust off-policy evaluation algorithm (*LDR*^{2}*OPE* algorithm from [9]) is detailed in Algorithm 1. This method combines propensity scores and regression functions in a doubly robust framework, ensuring robustness to both environment shifts and estimation errors in nuisance functions. By employing a localization technique, it avoids the computational burden of fitting a continuum of regression models, achieving $\sqrt{N}$-consistency and asymptotic efficiency under weak product rate conditions. The algorithm operates as follows. First, the dataset $\mathscr{D}$ is divided into K folds to enable cross-fitting, reducing overfitting. For each fold k, out-of-fold data $\mathscr{D}[\mathscr{I}_k^C]$ is used to train three estimators: $\hat{\pi}_b^{(k)}$ for the behavior policy π_b, an initial dual variable estimate $\hat{\alpha}_{\text{init}}^{(k)}$, and regression functions $\hat{f}_j^{(k)}$ for $j=0,1$. These estimators inform the construction of $\hat{W}_j(\alpha)$, which approximates the moment functions. The algorithm then solves the estimated moment equation to obtain $\hat{\alpha}$, the optimal dual variable. Finally, the distributionally robust value $\hat{V}_\delta$ is computed using $\hat{\alpha}$ and $\hat{W}_0(\hat{\alpha})$, providing a worst-case performance estimate over an uncertainty set of radius δ.

Algorithm 1 Localized Doubly Robust DRO Policy Evaluation

Input: Data $\mathscr{D}$, policy π, uncertainty set radius δ.
Randomly split $\mathscr{D}$ into K even folds, with indices $\mathscr{I}_k$ for the k-th fold.
for $k=1$ to K **do**
  Using $\mathscr{D}[\mathscr{I}_k^C]$, train $\hat{\pi}_b^{(k)}$ to fit the behavior policy π_b.
  Randomly split $\mathscr{I}_k^C$ into two halves $\mathscr{J}_1$, $\mathscr{J}_2$.
  $\hat{\alpha}_{\text{init}}^{(k)} \leftarrow \text{InitialEstimate}(\mathscr{D}[\mathscr{J}_1],\delta,\pi)$.
  Using $\mathscr{D}[\mathscr{J}_2]$, train $\hat{f}_j^{(k)}$ to fit $f_j(\cdot;\hat{\alpha}_{\text{init}}^{(k)})$, for $j=0,1$.
end for
Find $\hat{\alpha}>0$ that solves the moment equation:

$$-\log\left(\hat{W}_0(\alpha)\right)-\frac{\hat{W}_1(\alpha)}{\alpha\cdot\hat{W}_0(\alpha)}-\delta=0,$$

where

$$\hat{W}_j(\alpha):=\frac{1}{N}\sum_{k=1}^{K}\sum_{i\in\mathscr{I}_k}\hat{W}_j^{(i,k)}(\alpha)$$

and

$$\hat{W}_j^{(i,k)}(\alpha):=\sum_{a\in\mathscr{A}}\pi(a|x_i)\hat{f}_j^{(k)}(x_i,a)+\frac{\pi(a_i|x_i)}{\hat{\pi}_0^{(k)}(a_i|x_i)}\left(r_i^j\exp\left(-\frac{r_i}{\alpha}\right)-\hat{f}_j^{(k)}(x_i,a_i)\right)$$

Calculate the estimated value $\hat{V}_\delta \leftarrow -\hat{\alpha}\log\hat{W}_0(\hat{\alpha})-\hat{\alpha}\delta$.
Return: $\hat{\theta}^{\text{LDR}^2\text{OPE}}=\left(\hat{\alpha},\hat{W}_0(\hat{\alpha}),\hat{W}_1(\hat{\alpha}),\hat{V}_\delta\right)$.

Algorithm 2 Doubly Robust Distributionally Robust Off-Policy Evaluation

Input: Logged data $\{(x_i, a_i, r_i)\}_{i=1}^n$, target policy π, behavior policy π_b, uncertainty size δ
Output: Estimate of worst-case expected reward $V_\delta(\pi)$
Split the data into K folds for cross-fitting.
for $k = 1$ to K **do**
  Train on all data excluding fold k:
    Outcome regression model $\hat{r}(x, a)$
    Propensity score model $\hat{\pi}_0(a \mid x)$
  for each (x_i, a_i, r_i) in fold k **do**
    Compute importance weight:

$$w_i = \frac{\pi(a_i \mid x_i)}{\hat{\pi}_0(a_i \mid x_i)}.$$

    Compute doubly robust estimate:

$$\hat{r}_i^{DR} = \hat{r}(x_i, \pi(x_i)) + w_i(r_i - \hat{r}(x_i, a_i)).$$

  end for
end for
Aggregate all $\hat{r}_i^{DR}$ estimates.
Define:

$$\phi(\alpha) = -\alpha \log\left(\frac{1}{n}\sum_{i=1}^{n} \exp\left(-\frac{\hat{r}_i^{DR}}{\alpha}\right)\right) - \alpha\delta.$$

Solve for the optimal value:

$$\alpha^* = \arg\max_{\alpha>0} \phi(\alpha).$$

Compute:

$$V_\delta(\pi) = \phi(\alpha^*).$$

Return: $V_\delta(\pi)$

7.2. **Localized Doubly Robust Group DROPE (LDR^2-Group-DROPE).** The Localized Doubly Robust Group DROPE (LDR^2-Group-DROPE) method estimates the robust policy value $V_{robust}(\pi)$, defined as:

$$V_{robust}(\pi) = \inf_{w \in \mathscr{U}_w} \sum_{g=1}^{G} w_g V_{g,\delta_g}(\pi),$$

where

$$V_{g,\delta_g}(\pi) = \inf_{\mathbb{P}_{1,g} \in \mathscr{U}(\delta_g)} \mathbb{E}_{\mathbb{P}_{1,g}}[R(\pi(S))]$$

is the group-specific distributionally robust value, $\mathscr{U}(\delta_g)$ is the uncertainty set for group g with radius δ_g, and $\mathscr{U}_w$ is the ambiguity set for the group weights with radius δ_w. This approach extends the LDR^{2}OPE framework[9] by localizing around initial estimates of the dual variables α_g, making computation feasible.

Algorithm 3 Localized Doubly Robust Group DROPE (LDR^2-Group-DROPE)

Input: Data $\mathscr{D} = \bigcup_{g=1}^{G} \mathscr{D}_g$, policy π, group uncertainty radii $\{\delta_g\}_{g=1}^{G}$, weight uncertainty radius δ_w, nominal weights w_0.
Randomly split each $\mathscr{D}_g$ into K even folds, with indices $\mathscr{I}_{k,g}$ for group g, fold k.
for $g = 1$ to G **do**
  for $k = 1$ to K **do**
    Using $\mathscr{D}_g[\mathscr{I}_{k,g}^C]$, train $\widehat{\pi}_{0,g}^{(k)}$ to estimate the behavior policy $\pi_{0,g}$.
    Randomly split $\mathscr{I}_{k,g}^C$ into two halves $\mathscr{J}_{1,g}$ and $\mathscr{J}_{2,g}$.
    Compute $\widehat{\alpha}_{g,\text{init}}^{(k)} \leftarrow \text{InitialEstimate}(\mathscr{D}_g[\mathscr{J}_{1,g}], \delta_g, \pi)$ (e.g., cross-fitted SNIPS).
    Using $\mathscr{D}_g[\mathscr{J}_{2,g}]$, train $\widehat{f}_{0,g}^{(k)}$ and $\widehat{f}_{1,g}^{(k)}$ to estimate:

$$f_{0,g}(s,a;\widehat{\alpha}_{g,\text{init}}^{(k)}) = \mathbb{E}_{\mathbb{P}_{0,g}}[\exp(-R/\widehat{\alpha}_{g,\text{init}}^{(k)}) \mid X = x, A = a],$$

$$f_{1,g}(s,a;\widehat{\alpha}_{g,\text{init}}^{(k)}) = \mathbb{E}_{\mathbb{P}_{0,g}}[R\exp(-R/\widehat{\alpha}_{g,\text{init}}^{(k)}) \mid X = x, A = a].$$

  end for
end for
for $g = 1$ to G **do**
  Define estimated moments as functions of α_g:

$$\widehat{W}_{0,g}(\alpha_g) = \frac{1}{N_g}\sum_{k=1}^{K}\sum_{i\in\mathscr{I}_{k,g}}\left[\sum_{a\in\mathscr{A}}\pi(a \mid x_i)\widehat{f}_{0,g}^{(k)}(x_i,a) + \frac{\pi(a_i \mid x_i)}{\widehat{\pi}_{0,g}^{(k)}(a_i \mid x_i)}\left(\exp(-r_i/\alpha_g) - \widehat{f}_{0,g}^{(k)}(x_i,a_i)\right)\right],$$

$$\widehat{W}_{1,g}(\alpha_g) = \frac{1}{N_g}\sum_{k=1}^{K}\sum_{i\in\mathscr{I}_{k,g}}\left[\sum_{a\in\mathscr{A}}\pi(a \mid x_i)\widehat{f}_{1,g}^{(k)}(x_i,a) + \frac{\pi(a_i \mid x_i)}{\widehat{\pi}_{0,g}^{(k)}(a_i \mid x_i)}\left(r_i\exp(-r_i/\alpha_g) - \widehat{f}_{1,g}^{(k)}(x_i,a_i)\right)\right].$$

  Solve for $\widehat{\alpha}_g > 0$ such that:

$$-\log\widehat{W}_{0,g}(\widehat{\alpha}_g) - \frac{\widehat{W}_{1,g}(\widehat{\alpha}_g)}{\widehat{\alpha}_g\widehat{W}_{0,g}(\widehat{\alpha}_g)} - \delta_g = 0.$$

  Compute the group-specific robust value:

$$\widehat{V}_{g,\delta_g} = -\widehat{\alpha}_g\log\widehat{W}_{0,g}(\widehat{\alpha}_g) - \widehat{\alpha}_g\delta_g.$$

end for
Define the dual objective:

$$\widehat{f}(\beta) = -\beta\log\left(\sum_{g=1}^{G} w_{0,g}\exp\left(\frac{\widehat{V}_{g,\delta_g}}{\beta}\right)\right) - \beta\delta_w.$$

Find $\widehat{\beta} = \arg\max_{\beta>0}\widehat{f}(\beta)$ using numerical optimization.
Compute the robust policy value:

$$\widehat{V}_{robust} = \widehat{f}(\widehat{\beta}).$$

Return: $\widehat{V}_{robust}$.

Under standard product rate conditions on the estimation errors of the nuisance functions $(\widehat{\pi}_{0,g}, \widehat{f}_{0,g}, \widehat{f}_{1,g})$ and assuming sufficiently accurate initial estimates $\widehat{\alpha}_{g,\text{init}}$, the estimator $\widehat{V}_{robust}$ is $\sqrt{N}$-consistent and asymptotically normal, inheriting the semiparametric efficiency of the LDR^{2}OPE method [9].

7.3. **Continuum Doubly Robust Group-DROPL (CDR^2-Group-DROPL).** The Continuum Doubly Robust Group-DROPL (CDR^2-Group-DROPL) extends the CDR^{2}OPL framework to the group distributionally robust optimization (group-DRO) setting for policy learning. The objective is to optimize a policy π that maximizes the robust policy value:

$$V_{robust}(\pi) = \inf_{w \in \mathscr{U}_w} \sum_{g=1}^{G} w_g V_{g,\delta_g}(\pi),$$

where

$$V_{g,\delta_g}(\pi) = \inf_{\mathbb{P}_{1,g} \in \mathscr{U}(\delta_g)} \mathbb{E}_{\mathbb{P}_{1,g}}[R(\pi(S))]$$

represents the group-specific distributionally robust value, $\mathscr{U}(\delta_g)$ is the uncertainty set for group g with radius δ_g, and $\mathscr{U}_w$ is the ambiguity set for the group weights with radius δ_w. Leveraging the dual formulation, this can be expressed as:

$$V_{robust}(\pi) = \max_{\beta>0,\{\alpha_g>0\}} \left[-\beta \log \left(\sum_{g=1}^{G} w_{0,g} \exp\left(\frac{\phi_g(\pi, \alpha_g)}{\beta} \right) \right) - \beta \delta_w \right],$$

where

$$\phi_g(\pi, \alpha_g) = -\alpha_g \log W_g(\pi, \alpha_g) - \alpha_g \delta_g$$

and

$$W_g(\pi, \alpha_g) = \mathbb{E}_{\mathbb{P}_{0,g}}[\exp(-R(\pi(S))/\alpha_g)]$$

is the moment-generating function under the nominal distribution $\mathbb{P}_{0,g}$ for group g, with $w_{0,g}$ as the nominal group weights.

Algorithm: CDR^2-Group-DROPL

To learn the optimal policy, we seek:

$$\widehat{\pi} \in \arg\max_{\pi \in \Pi} \widehat{V}^{DR}_{robust}(\pi),$$

where $\widehat{V}^{DR}_{robust}(\pi)$ is a doubly robust estimator of $V_{robust}(\pi)$. The algorithm estimates group-specific nuisance functions and optimizes the policy iteratively, handling the continuum of dual variables α_g using local weighting techniques (e.g., random forests). The detailed steps are as follows:

Algorithm 4 CDR^2-Group-DROPL

Input: Data $\mathscr{D} = \bigcup_{g=1}^{G} \mathscr{D}_g$, policy class Π, uncertainty radii $\{\delta_g\}_{g=1}^{G}$, δ_w, nominal weights $w_0 = \{w_{0,g}\}_{g=1}^{G}$.
Randomly split each $\mathscr{D}_g$ into K even folds, denoted by indices $\mathscr{I}_{k,g}$, with complement $\mathscr{I}_{k,g}^{C}$.
for $g = 1$ to G **do**
  for $k = 1$ to K **do**
    Using $\mathscr{D}_g[\mathscr{I}_{k,g}^{C}]$, train $\widehat{\pi}_{0,g}^{(k)}$ to estimate the behavior policy $\pi_{0,g}$.
    Using $\mathscr{D}_g[\mathscr{I}_{k,g}^{C}]$, train $\widehat{f}_{0,g}^{(k)}(s,a;\alpha_g)$ for $\alpha_g \in (0, \bar{\alpha}_g]$ via random forests, where:

$$f_{0,g}(s,a;\alpha_g) = \mathbb{E}_{\mathbb{P}_{0,g}}[\exp(-R/\alpha_g) \mid X = x, A = a].$$

  end for
end for
Initialize $\widehat{\pi} \in \Pi$.
while not converged **do**
  for $g = 1$ to G **do**
    Compute the doubly robust estimator:

$$\widehat{W}_g^{DR}(\widehat{\pi}, \alpha_g) = \frac{1}{N_g} \sum_{k=1}^{K} \sum_{i \in \mathscr{I}_{k,g}} \left[\frac{\widehat{\pi}(a_i \mid x_i)}{\widehat{\pi}_{0,g}^{(k)}(a_i \mid x_i)} \left(\exp(-r_i/\alpha_g) - \widehat{f}_{0,g}^{(k)}(x_i, a_i; \alpha_g) \right) + \sum_{a \in \mathscr{A}} \widehat{\pi}(a \mid x_i) \widehat{f}_{0,g}^{(k)}(x_i, a; \alpha_g) \right],$$

    where $\mathscr{D}_g = \{(x_i, a_i, r_i)\}_{i=1}^{N_g}$.
    Optimize $\widehat{\alpha}_g \leftarrow \arg\max_{\alpha_g > 0} \left[-\alpha_g \log \widehat{W}_g^{DR}(\widehat{\pi}, \alpha_g) - \alpha_g \delta_g \right]$.
    Compute the group-specific robust value:

$$\widehat{V}_{g,\delta_g}^{DR}(\widehat{\pi}) = -\widehat{\alpha}_g \log \widehat{W}_g^{DR}(\widehat{\pi}, \widehat{\alpha}_g) - \widehat{\alpha}_g \delta_g.$$

  end for
  Compute the overall robust value estimator:

$$\widehat{V}_{robust}^{DR}(\widehat{\pi}) = \max_{\beta > 0} \left[-\beta \log \left(\sum_{g=1}^{G} w_{0,g} \exp\left(\frac{\widehat{V}_{g,\delta_g}^{DR}(\widehat{\pi})}{\beta} \right) \right) - \beta \delta_w \right].$$

  Update $\widehat{\pi}$ using policy gradient or another optimization method to maximize $\widehat{V}_{robust}^{DR}(\pi)$.
end while
Return: $\widehat{\pi}$.

The regret of the learned policy, defined as

$$\mathscr{R}_{robust}(\widehat{\pi}) = V_{robust}(\pi^*) - V_{robust}(\widehat{\pi}),$$

where $\pi^* = \arg\max_{\pi \in \Pi} V_{robust}(\pi)$ is analyzed under standard estimation assumptions. If the estimation errors for $\widehat{\pi}_{0,g}$ and $\widehat{f}_{0,g}$ satisfy

$$\text{Rate}_{\pi_{0,g}}(N_g, \beta/K) \cdot \text{Rate}_{f,g}^{c}(N_g, \beta/K) = o(N_g^{-1/2}),$$

then the regret is $\mathcal{O}(N^{-1/2})$, where $N = \sum_{g=1}^{G} N_g$. This extends the guarantees from the single-group CDR^{2}OPL setting to the group-DRO context.

7.4. **Simulation: Impact of Uncertainty Set Size on Robust Policy Evaluation and Learning.** This section summarizes the simulation setup and results for evaluating the impact of uncertainty set size on the performance of the Localized Doubly Robust Group DROPE (LDR^2-Group-DROPE) and Continuum Doubly Robust Group DROPL (CDR^2-Group-DROPL) algorithms, compared to non-robust baselines.

The simulation involves a contextual bandit problem with two groups ($G = 2$), designed to mimic the data-generating process from Section 5 of [9], adapted for group-specific distributions.

- State Space: $\mathscr{X} = [-1,1]^2$, with states $X \sim \text{Unif}([-1,1]^2)$.
- Action Space: $\mathscr{A} = \{0,1,2,3,4\}$.
- Behavior Policy: For group $g \in \{1,2\}$, the behavior policy is a softmax policy:
$$\pi_{0,g}(a \mid x) \propto \exp(2x^\top \beta_{a,g}),$$
where $\beta_{a,g} = (\text{Re}(\zeta_a), \text{Im}(\zeta_a))$, $\zeta_a = \exp(2a\pi i/5)$, and $\beta_{a,2} = \beta_{a,1} + (0.1, 0.1)$ for group 2.
- Rewards: For group g, potential outcomes are:
$$R(a \mid X = x, g) \sim \mathscr{N}(x^\top \beta_{a,g}, \sigma_g^2),$$
with $\sigma_1 = 0.1$, $\sigma_2 = 0.15$.
- Sample Size: $N_g = 500$ per group, total $N = 1000$.
- Group Weights: Nominal weights $w_0 = (0.5, 0.5)$.
- Distributional Shift: In the test environment, reward means are shifted by 0.2 for group 1 and 0.3 for group 2.

The experiment tests the effect of varying uncertainty set sizes:

- Group Uncertainty Radii: $\delta_g \in \{0.01, 0.05, 0.1, 0.2\}$ for both groups ($\delta_1 = \delta_2 = \delta$).
- Weight Uncertainty Radius: $\delta_w \in \{0.01, 0.05, 0.1\}$, paired as
$$(\delta, \delta_w) \in \{(0.01, 0.01), (0.05, 0.05), (0.1, 0.05), (0.2, 0.1)\}.$$
- Algorithms:
  - LDR^2-Group-DROPE: Estimates the robust policy value $V_{robust}(\pi)$.
  - Non-Robust OPE: Standard doubly robust off-policy evaluation (CFDR).
  - CDR^2-Group-DROPL: Learns a robust policy $\widehat{\pi}$.
  - Non-Robust OPL: Standard CFDR-based off-policy learning.
- Metrics:
  - Evaluation: Mean Squared Error (MSE) of $\widehat{V}_{robust}$ compared to the true $V_{robust}(\pi)$.
  - Learning: Regret $\mathscr{R}_{robust}(\widehat{\pi}) = V_{robust}(\pi^*) - V_{robust}(\widehat{\pi})$.
- Runs: 50 simulations per (δ, δ_w) combination.

The target policy for evaluation is $\pi(x) = \arg\max_a x^\top \beta_{a,1}$, and the policy class for learning consists of linear policies $\pi_\theta(x) = \arg\max_a \theta^\top \phi(x,a)$.

The results, averaged over 50 simulations, are presented in Table 7. The robust methods outperform non-robust baselines across all uncertainty set sizes, with performance improving as δ increases.

The proposed method significantly outperforms the baseline under the simulated distributional shift. LDR^2-Group-DROPE MSE decreases from 0.0038 to 0.0020 as δ increases from 0.01 to 0.2, indicating that larger uncertainty sets better capture the distributional shifts (reward mean shifts of 0.2 and 0.3). The MSE is significantly lower than that of non-robust OPE, which remains high (0.0172–0.0201) due to unmodeled shifts. CDR^2-Group-DROPL Regret decreases from 0.0124 to 0.0085 with increasing δ, showing improved robustness of learned policies. Non-robust OPL regret is higher (0.0269–0.0305), as it fails to account for shifts. Effect of δ_w: Increasing δ_w (e.g., from 0.01 to 0.1) has a modest impact, as group weights are stable ($w_0 = (0.5, 0.5)$). Larger δ_w slightly enhances performance when paired with larger δ.

TABLE 7. Performance Metrics for Different Uncertainty Set Sizes

δ	δ_w	MSE (Evaluation)		Regret (Learning)	
		LDR^2	Non-Robust	CDR^2	Non-Robust
0.01	0.01	0.0038	0.0201	0.0124	0.0305
0.05	0.05	0.0025	0.0187	0.0098	0.0283
0.1	0.05	0.0021	0.0179	0.0087	0.0276
0.2	0.1	0.0020	0.0172	0.0085	0.0269

7.5. **Detailed Definition of Risk Adjusted Return.** The empirical risk adjusted return of a customer depends on the joint distribution of $\mathbf{v} = [v^{pd}, v^{trans}, v^{prin}]$, where v^{pd} is the estimated PD of the customer over a future period; v^{trans} is the transaction amount of the customer; v^{prin} is the outstanding balance owed by the customer.

Both v_{it}^{trans} and v_{it}^{prin} are observed empirically. The bank receives a commission rate η^{trans} for the transaction amount and charges a interest rate η^{prin} for the outstanding balance. Typically, v_{it}^{pd} is estimated by the bank internally from a known separate process and refreshed at least annually. Estimating v_{it}^{pd} is out of scope for this work, we treat it as empirically observed. This is a reasonable assumption as the credit limits assigned by the bank are not considered to have causal influence on a customer's credit risk.

$$R_i = \left[\sum_{t=1}^{T} \gamma^t * \left[[v_{it}^{pd} * \eta^{lgd} * A_i] \right. \right.$$
$$\left. \left. + [1 - v_{it}^{pd}] * [-v_{it}^{trans} * \eta^{trans} - min(A_i, v_{it}^{prin}) * \eta^{prin}] + \eta^{equity} * K(v_{it}^{pd}) * A_i \right] \right]$$

A customer will fully utilize its credit limit at default i.e. the exposure at default equals to A_i; and LGD η^{lgd} is a known, which is a good assumption for an unsecured revolving lending products. The expected credit loss is approximately $\eta^{lgd} * v_{it}^{pd} * A_{it}$. And γ is a known discount factor for future cash flow.

Expected regulatory capital (RC), $K(v_{it}^{pd})$ for a customer with 1-year forward probability of default (PD) at time t can be calculated analytically, provided by Basel Committee, where Φ is normal cumulative distribution function, and ρ is a known constant describing the correlation

of a customer risk with the latent systematic factor. The RC formula essentially captures the 99.9% credit Value-at-Risk under model assumptions

$$K(v^{pd}) = [\Phi\big((1-C)^{-\frac{1}{2}}\Phi^{-1}(v^{pd}) + (\frac{C}{(1-C)})^{\frac{1}{2}}\Phi^{-1}(0.999)\big) - (v^{pd})],$$

where C is a correlation parameter given by the regulators and Φ is normal cumulative distribution function. The risk weighted asset (RWA) equals to $12.5 * K(v^{pd}) * \gamma^{lgd} * A_i$. Banks expect a minimum percentage return on the risk weighted asset, as it needs to maintain a stable Common Tier-I capital (CET1) ratio at least 4.5%, CET1 ratio = Common Tier 1 Equity / Risk Weighted Asset, that is, the tail credit loss of the customer has to be covered by the shareholders' equity of the bank. Since the bank can raise additional equity through issuing new shares or sell assets, we treat it as a soft constraint, with the expected long run return on equity as the cost of raising equity, while maintaining the CET1 ratio. This can be written as a penalty term $\eta^{equity} * K(v_{it}^{pd}) * A_i$ under the simplified assumption both η^{lgd} and expected return on shareholder's equity are known. We choose $\eta^{equity} = 0.04$ in our experiment.

7.6. **Distributionally Robust Large Margin Nearest Neighbor (DR-LMNN).** Robust metric learning aims to learn an embedding $f_\theta : \mathscr{X} \to \mathscr{Z}$ that clusters similar contexts (borrowers) together and remains stable to outliers or moderate distribution shifts. Incorporating a robust embedding can:

- Reduce noise in the context features,
- Promote better generalization of reward information,
- Safeguard decisions under partial feedback and non-stationary conditions.

In this subsection, we introduce a distributionally robust modification of the widely studied Large Margin Nearest Neighbor (LMNN) metric learning approach [13]. Standard LMNN focuses on learning a linear transformation of the feature space that keeps target neighbors (points of the same class or "similar" group) close, while separating points of different classes (or dissimilar group) by a large margin. However, standard LMNN is sensitive to outliers or shifts in the data distribution. We therefore propose a min-max formulation that equips LMNN with distributional robustness, ensuring the learned metric is stable under moderate data perturbations.

Let $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$ be a labeled dataset, with $\mathbf{x}_i \in \mathbb{R}^d$ and $y_i \in \{1, \ldots, C\}$ denoting the class label (or cluster ID). We wish to learn a linear transformation $\mathbf{L} \in \mathbb{R}^{D \times d}$ such that distances between same-class points are small and distances between different-class points are large, under the transformed metric:

$$d_{\mathbf{L}}(\mathbf{x}_i, \mathbf{x}_j) \;=\; \|\mathbf{L}\mathbf{x}_i - \mathbf{L}\mathbf{x}_j\| \,.$$

Typically, LMNN introduces a set of target neighbors $\mathscr{N}_i$ for each point $\mathbf{x}_i$: these are k nearest neighbors of $\mathbf{x}_i$ sharing the same label y_i in the original space (or an iteratively updated space). The standard LMNN objective is:

$$\min_{\mathbf{L}\succeq\mathbf{0}} \sum_i \sum_{j\in\mathscr{N}_i} \|\mathbf{L}(\mathbf{x}_i - \mathbf{x}_j)\|^2 \;+\; \lambda \sum_i \sum_{j\in\mathscr{N}_i} \sum_{l: y_l \neq y_i} \big[1 + \|\mathbf{L}(\mathbf{x}_i - \mathbf{x}_j)\|^2 - \|\mathbf{L}(\mathbf{x}_i - \mathbf{x}_l)\|^2\big]_+,$$

where the first term encourages same-class neighbors to be close, while the second term (the *margin* or *impulse* term) enforces a large margin between different-class points.

While LMNN has been successful in various metric learning tasks, it has a known limitation:

- **Sensitivity to Outliers and Small Distribution Shifts.** A few incorrectly labeled points or moderate changes in the data distribution can adversely affect the learned metric $\mathbf{L}$.
- **Non-Stationary or Adversarial Data.** In many real-world scenarios, data may shift over time or be deliberately perturbed by an attacker. Standard LMNN is not inherently designed to handle such shifts.

DR-LMNN. To address these issues, we adopt a distributionally robust optimization perspective. Rather than minimizing the empirical objective alone, we protect against a worst-case perturbation of the empirical distribution within some uncertainty set $\mathscr{U}_\rho(P_0)$ of radius ρ.

Let $\hat{P}_n$ denote the empirical distribution over the set of triplets $\{(\mathbf{x}_i, \mathbf{x}_j, \mathbf{x}_l)\}$ relevant to LMNN (i.e., (i,j) for target neighbors and (i,j,l) for impostors). A standard approach is to construct an *uncertainty set* $\mathscr{U}_\rho(\hat{P}_n)$ around $\hat{P}_n$, often using a Wasserstein or f-divergence ball. Concretely:

$$\min_{\mathbf{L}\succeq\mathbf{0}} \quad \max_{Q\in\mathscr{U}_\rho(\hat{P}_n)} \quad \mathbb{E}_{z\sim Q}\Big[\mathscr{L}_{\mathrm{LMNN}}(\mathbf{L};z)\Big], \tag{7.1}$$

where z indexes either pairwise or triplet structures $(\mathbf{x}_i, \mathbf{x}_j, \mathbf{x}_l)$, and $\mathscr{L}_{\mathrm{LMNN}}(\mathbf{L};z)$ corresponds to the LMNN loss:

$$\mathscr{L}_{\mathrm{LMNN}}(\mathbf{L};(\mathbf{x}_i,\mathbf{x}_j,\mathbf{x}_l)) \;=\; \|\mathbf{L}(\mathbf{x}_i-\mathbf{x}_j)\|^2 \;+\; \lambda\left[1+\|\mathbf{L}(\mathbf{x}_i-\mathbf{x}_j)\|^2-\|\mathbf{L}(\mathbf{x}_i-\mathbf{x}_l)\|^2\right]_+.$$

(Here, $(\mathbf{x}_i, \mathbf{x}_j)$ is a same-class pair and $\mathbf{x}_l$ an impostor with $y_l \neq y_i$.)

The inner maximization over $Q \in \mathscr{U}_\rho(\hat{P}_n)$ forces $\mathbf{L}$ to be *robust* against local perturbations of the distribution. If an adversary (or natural shift) reweighs data points in some small neighborhood around $\hat{P}_n$, the resulting $\mathbf{L}$ will still maintain good margins.

A key design choice is how to define $\mathscr{U}_\rho(\hat{P}_n)$. Two common options are Wasserstein Ball [1] and f-Divergence Ball [11]:

- **Wasserstein Ball**:
$$\mathscr{U}_\rho(\hat{P}_n) = \big\{Q \,\big|\, W_c\big(\hat{P}_n, Q\big) \le \rho\big\},$$
where W_c is the Wasserstein distance with cost function $c(\cdot,\cdot)$.
- **f-Divergence Ball**:
$$\mathscr{U}_\rho(\hat{P}_n) = \big\{Q \,\big|\, D_f\big(Q \,\|\, \hat{P}_n\big) \le \rho\big\},$$
for some convex f, e.g. KL or χ^2 divergence.

  In our problem, we assume a Wasserstein-1 ball uncertainty set.

Either choice provides a worst-case distribution around $\hat{P}_n$; the difference lies in how they measure "distance" in sample space.

Solving the DR-LMNN Objective. Equation (7.1) is a *min–max* problem:

$$\min_{\mathbf{L}\succeq\mathbf{0}} \quad \max_{Q\in\mathscr{U}_\rho(\hat{P}_n)} \mathbb{E}_{z\sim Q}\big[\mathscr{L}_{\mathrm{LMNN}}(\mathbf{L};z)\big].$$

In large-scale settings, one can adopt mini-batch techniques with gradient descent on $\mathbf{L}$ and gradient ascent on distributional parameters that parameterize Q. We present an approximate stochastic min–max procedure for solving the DR-LMNN problem. We assume a simple **Wasserstein-**1 uncertainty set

$$\mathscr{U}_\rho(\hat{P}_n) \;=\; \big\{Q : W_1\big(Q, \hat{P}_n\big) \le \rho\big\},$$

where W_1 denotes the 1-Wasserstein (earth-mover) distance, and $\rho > 0$ is the radius. Recall the DR-LMNN objective:

$$\min_{\mathbf{L}\succeq 0} \quad \max_{Q\in\mathscr{U}_\rho(\hat{P}_n)} \quad \mathbb{E}_{z\sim Q}\Big[\mathscr{L}_{\mathrm{LMNN}}(\mathbf{L};z)\Big],$$

where z indexes triplets (x_i,x_j,x_l) for LMNN, and

$$\mathscr{L}_{\mathrm{LMNN}}(\mathbf{L};z) \;=\; \|\mathbf{L}(x_i-x_j)\|^2 \;+\; \lambda\left[1+\|\mathbf{L}(x_i-x_j)\|^2-\|\mathbf{L}(x_i-x_l)\|^2\right]_+.$$

To handle the inner maximization over Q in the Wasserstein ball, we adopt an *approximate* (stochastic) method that alternates between:

- Estimating adversarial distributions/samples to maximize the LMNN loss,
- Updating $\mathbf{L}$ to minimize the resulting worst-case objective.

One key practical consideration, defining target neighbors in contextual bandit can be nuanced. If we only rely on reward bins, we may merge different actions or mask important distinctions. Incorporating additional domain knowledge.

7.7. **Approximate Stochastic Min–Max Algorithm for DRLMNN.** Algorithm 5 summarizes a typical procedure. The method is inspired by adversarial training in deep learning [11] and stochastic mirror descent in robust optimization [5].

7.8. **Algorithm: DR-LMNN with Smooth Surrogate and Mini-Batch Adversarial Reweighting.** Below is an outline of the distributionally robust LMNN procedure for moderately large data, using a smooth margin surrogate and mini-batch adversarial reweighting under a Wasserstein-1 uncertainty set.

Notes on the Algorithm.

- **Smooth Surrogate:** Replacing the raw hinge margin with a smooth convex envelope (e.g. $\phi(t)=\beta\ln(1+e^{t/\beta})$ [7]) provides well-defined gradients and avoids subgradient instability.
- **Factorization**: We learn $\mathbf{L}$ directly, removing the explicit $\mathbf{M}\succeq 0$ constraint. This is a *local* approach but widely used in metric learning.
- **Mini-Batch Adversarial Reweighting**: We approximate $\max_{Q\in\mathscr{U}_\rho}$ by local reweighting in each mini-batch, which is far cheaper than a full LP or global OT across all N triplets. We typically keep J small (e.g. 1–5).
- **Computational Feasibility**: The method scales roughly as standard mini-batch gradient descent plus a small additional overhead for each adversarial loop.

The mini-batch $\mathscr{B}_k$ approximates sampling from the empirical distribution $\hat{P}_n$. The inner loop tries to reweight the mini-batch to maximize the LMNN loss, subject to a constraint that p^* remains within a Wasserstein radius ρ of the uniform distribution (or of some reference distribution) on $\mathscr{B}_k$. In practice, one can implement this via Entropic Optimal Transport or **Sinkhorn** approximations to handle the reweighting if we have pairwise distances among data points in $\mathscr{B}_k$. The Metric Update step is just a gradient descent on $\mathbf{L}$ given the "adversarially inflated" loss $\widehat{L}_k$. By sampling mini-batches, we avoid computing exact worst-case distributions across *all* n triplets at once, reducing computational overhead. Approximate min–max can significantly mitigate outlier effects and better handle moderate domain shifts and improve the robustness over standard LMNN.

Algorithm 5 DR-LMNN (Smooth Surrogate + Mini-Batch Adversarial)

(1) **Input:**
- **Data:** $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$ with labels $y_i \in \{1, \dots, C\}$.
- **Triplets** (or pairs) for LMNN: $z_u = (\mathbf{x}_i, \mathbf{x}_j, \mathbf{x}_l)$ enumerated or sampled; $\hat{P}_n = \frac{1}{N}\sum_{u=1}^N \delta_{z_u}$ is the empirical distribution.
- **Surrogate margin loss**:

$$L_{\text{sur}}(\mathbf{L}; z) \;=\; \|\mathbf{L}(\mathbf{x}_i - \mathbf{x}_j)\|^2 \;+\; \lambda\, \phi\big(1 + \|\mathbf{L}(\mathbf{x}_i - \mathbf{x}_j)\|^2 - \|\mathbf{L}(\mathbf{x}_i - \mathbf{x}_l)\|^2\big),$$

  where $\phi(\cdot)$ is smooth (e.g. log-sum-exp).
- **Metric factor**: $\mathbf{M} = \mathbf{L}^\top \mathbf{L}$, $\mathbf{L} \in \mathbb{R}^{r \times d}$.
- **Uncertainty set**: $\mathscr{U}_\rho(\hat{P}_n)$ (Wasserstein-1 of radius ρ).
- **Hyperparams**: mini-batch size B, adversarial steps J, step sizes $\{\eta_k\}$, total outer iterations K.

(2) **Initialize:**
- $\mathbf{L}^{(0)}$ (e.g. identity-like or small random).

(3) **For** $k = 1$ to K (outer iterations):

(a) **Sample mini-batch** $\mathscr{B}_k$ of size B from $\hat{P}_n$ (randomly choose B triplets).

(b) Let $p^{(0)}$ be uniform on $\mathscr{B}_k$, i.e. $p_u^{(0)} = 1/B$.

(c) **Adversarial loop** (J steps):

(i) For $j = 1, \dots, J$:
- Compute gradient-like direction

$$g\big(p^{(j-1)}, \mathbf{L}^{(k-1)}\big) \;=\; \nabla_p \sum_{z \in \mathscr{B}_k} p_z L_{\text{sur}}\big(\mathbf{L}^{(k-1)}; z\big).$$

- Update $p^{(j)}$ by ascending in $g(\cdot)$ under constraints:

$$\sum_{z \in \mathscr{B}_k} p_z = 1, \;\; p_z \geq 0, \;\; W(p^{(j)}, p^{(0)}) \leq \rho'$$

  (where W could be the Wasserstein-1 distance restricted to the mini-batch; we can solve via entropic OT or projected gradient).

(ii) Let $p^* \leftarrow p^{(J)}$.

(d) **Compute adversarial mini-batch loss**:

$$\widehat{L}_k \;=\; \sum_{z \in \mathscr{B}_k} p_z^* L_{\text{sur}}\big(\mathbf{L}^{(k-1)}; z\big).$$

(e) **Metric update**:

$$\mathbf{L}^{(k)} \;\leftarrow\; \mathbf{L}^{(k-1)} \;-\; \eta_k \nabla_{\mathbf{L}} \big[\widehat{L}_k\big].$$

(Optionally project or regularize $\mathbf{L}^{(k)}$ if needed.)

(4) **End For**

(5) **Return:** $\mathbf{L}^{(K)}$ (and hence $\mathbf{M} = \mathbf{L}^{(K)\top}\mathbf{L}^{(K)}$) as the final robust LMNN metric.

An approximate stochastic min–max optimizer is a practical way to implement DR-LMNN under Wasserstein uncertainty. By iteratively sampling mini-batches, finding approximate adversarial reweightings within radius ρ, and updating the metric $\mathbf{L}$, we achieve a balance between robustness and computational feasibility. Although inner maximization and hyperparameter tuning add complexity beyond standard LMNN, the resultant metric is less vulnerable to

outliers and moderate distribution shifts, thereby improving out-of-distribution performance in real-world tasks like loan risk assessment.

7.9. **DR-LMNN-Enhanced DR-DR.** We propose integrating DR-LMNN with the DR-DR contextual bandit method to achieve a robust and more equitable off-policy solution in contextual bandit settings. Once we embed each context $\mathbf{x}_i$ into $\tilde{\mathbf{x}}_i$, we replace (X_i, A_i, R_i) by $(\tilde{X}_i, A_i, R_i)$ and proceed to construct the empirical measure $\hat{P}_n$ in this new space. That is, define $\hat{P}_n^{(\mathrm{LMNN})} = \frac{1}{n}\sum_{i=1}^{n} \delta_{(\tilde{X}_i, A_i, R_i)}$, where δ denotes the Dirac measure.

In the DR-DR framework, we now consider an ambiguity set $\mathscr{U}\left(\hat{P}_n^{(\mathrm{LMNN})}, \rho\right)$ around this measure. The divergence used could be computed in the $\tilde{X}$-space:

$$\mathscr{U}\left(\hat{P}_n^{(\mathrm{LMNN})}, \rho\right) \;=\; \left\{ Q : D\big(Q, \hat{P}_n^{(\mathrm{LMNN})}\big) \;\le\; \rho \right\},$$

where $D(\cdot,\cdot)$ is defined with respect to distances in the transformed feature space (e.g., a Wasserstein distance that uses $d_{\mathbf{M}}$). The DR-DR objective then becomes

$$\max_{Q \in \mathscr{U}\left(\hat{P}_n^{(\mathrm{LMNN})}, \rho\right)} \Big[\mathrm{DR}(Q, \pi) \Big].$$

In consumer finance, systematically modifying representation spaces must comply with regulations. One must ensure that transformations do not violate the protected-group's rights of accessing credit or lead to opaque decision rules. Ideally, domain experts should help interpret the learned distances. [3] and [14] study fairness-aware transformations, discussing how data representation can mitigate discrimination. Our work synthesizes these ideas by embedding LMNN into the DR-DR paradigm, highlighting potential avenues for improved fairness and robustness. In selecting the ambiguity set, we present a refined Group-DRO approach that better aligns with real-world finance application leads interpretable policies.

Acknowledgments

We would like to thank the editor and reviewers for their careful and thoughtful comments.

REFERENCES

[1] J. Blanchet, K. Murthy, V. Nguyen, Statistical analysis of wasserstein distributionally robust estimators, Emerging Optimization Methods and Modeling Techniques with Applications, 2021. DOI: 10.1287/educ.2021.0233

[2] T. Chen, C. Guestrin, XGBoost: A scalable tree boosting system, In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016.

[3] F. Calmon, D. Wei, B. Vinzamuri, K. N. Ramamurthy, K. Varshney, Optimized pre-processing for discrimination prevention, In: Advances in Neural Information Processing Systems (NIPS), pp. 3992-4001, 2017.

[4] M. Dudik, D. Erhan, J. Langford, L. Li, Doubly robust policy evaluation and optimization, Statistical Sci. 29 (2014), 485–511.

[5] J. Duchi, P. Glynn, H. Namkoong, Statistics of robust optimization: a generalized empirical likelihood approach, Math. Oper. Res. ematics of Operations Research, 46 (2021), 946-969.

[6] J. H. Friedman, Greedy function approximation: A gradient boosting machine, Ann. Statist. 29 (2001), 1189-1232.

[7] A. Genevay, M. Cuturi, G. Peyré, F. Bach, Stochastic optimization for large-scale optimal transport, In: 30th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, 2016.

[8] I. Haimowitz, H. Schwarz, Clustering and prediction for credit line optimization, AAAI Technical Report WS-97-07, 1997.

[9] N. Kallus, X. Mao, K. Wang, Z. Zhou, Doubly robust distributionally robust off policy evaluation and learning, In: Proceedings of the 39th International Conference on Machine Learning, 2022.

[10] N. Si, F. Zhang, Z. Zhou, J. Blanchet, Distributionally robust batch contextual bandits, Manag. Sci. 69 (2023), 5772-5793.

[11] A. Sinha, H. Namkoong, J. Duchi, Certifying some distributional robustness with principled adversarial training, In: International Conference on Learning Representations (ICLR), 2018.

[12] S. Sagawa, P. W. Koh, T. Hashimoto, P. Liang, Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case regularization, arXiv:1911.08731, 2020.

[13] K. Weinberger, L. Saul, Distance metric learning for large margin nearest neighbor classification, J. Mach. Learn. Res. 10 (2009), 207-244.

[14] R. Zemel, Y. Wu, K. Swersky, T. Pitassi, C. Dwork, Learning fair representations, In: International Conference on Machine Learning (ICML), pp. 325-333, 2013.