MINIMAX OPTIMALITY FOR FACTORED OPTIMAL-TRANSPORT DISTRIBUTIONALLY-ROBUST POLICY LEARNING IN CONTEXTUAL BANDITS

Anonymous authors
Paper under double-blind review

ABSTRACT

Offline contextual-bandit policies may be deployed under simultaneous shifts in the context distribution and reward laws. We study policy learning for a factored distributionally robust optimization (DRO) objective that models these shifts through separate ambiguity sets. Optimal transport (OT) ambiguity sets are particularly suitable for this purpose, allowing the transport cost to encode application-specific severity of a shift. We develop a pessimistic OT-DRO learner based on sharp confidence bounds for robust rewards and analyze context estimation using a distribution-dependent concentration bound. Requiring knowledge of neither the behavior policy nor the context distribution, our learner achieves a regret of order $DC^\star_{\mathrm{rob}}\sqrt{\log(|\mathcal{X}||\mathcal{A}|/\delta)/n}$. Our robust optimal-policy coverage coefficient $C^\star_{\mathrm{rob}}$ is a natural generalization of the single-policy coverage coefficient in prior non-robust pessimism work, capturing the additional statistical difficulty caused by worst-case context shift. A matching lower bound shows that this rate, including its dependence on $C^\star_{\mathrm{rob}}$, is minimax-optimal up to constants.

1 INTRODUCTION

Offline contextual bandit learning seeks to optimize a decision policy using logged data collected by a possibly different behavior policy (Dudík et al., 2011; Swaminathan & Joachims, 2015). It is central to applications in which online experimentation is costly or unsafe, such as recommendation systems, advertising, and personalized treatment (Li et al., 2010; Bottou et al., 2013; Tewari & Murphy, 2017). Standard methods typically evaluate policies under the assumption that the deployment environment coincides with the environment that generated the logged data. This assumption is often unrealistic: the population of contexts may change over time, and the reward distribution associated with a given context–action cell may also shift. A policy optimized for the nominal distribution can therefore perform poorly when the context or reward distribution shifts after deployment.

Distributionally robust optimization (DRO) addresses this problem by optimizing the worst-case policy value over a neighborhood of the nominal environment. Early approaches place a single divergence ball around the joint data-generating distribution (Si et al., 2020; Kallus et al., 2022). Because context shift and conditional reward shift may occur at different scales, factored DRO assigns separate ambiguity sets to the context distribution and the conditional reward laws. Existing factored formulations use Kullback-Leibler (KL) divergence (Mu et al., 2022) or Wasserstein distance (Shen et al., 2024) to formulate the ambiguity sets. We study factored DRO with general optimal-transport (OT) costs, which we call factored OT-DRO. Unlike KL divergence, OT can accommodate distributions with different supports. General transport costs also allow the severity of a shift to be specified more flexibly than the squared-distance cost used by Shen et al. (2024). For example, two changes in patient state could have the same squared distance but very different clinical consequences; a chosen transport cost could assign them different costs.

Existing learning guarantees for factored DRO contextual bandits, however, require uniform coverage of all context–action cells (Mu et al., 2022; Shen et al., 2024). This requirement is restrictive in offline learning because the behavior policy may rarely select actions that are irrelevant to an optimal policy. In the nominal setting, pessimistic policy learning avoids paying for such irrelevant actions by constructing lower confidence bounds and requiring sufficient coverage only along an optimal

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.

policy (Rashidinejad et al., 2021; Li et al., 2022; Jin et al., 2025; Wang et al., 2024). This suggests that robust policy learning should likewise adapt to the coverage of an optimal robust policy.

Extending this principle to factored OT-DRO is nontrivial. Under context shift, the adversary can place additional mass on contexts that are uncommon in the logged data, so nominal optimal-policy coverage alone does not capture the resulting statistical difficulty. An appropriate coverage condition must account simultaneously for the behavior policy, an optimal robust policy, and every context distribution permitted by the ambiguity set. This raises the central question of this work: *Can factored OT-DRO policies be learned at the minimax-optimal rate without uniform action coverage, paying only for coverage along an optimal robust policy?* We show that the answer is affirmative.

1.1 Problem Setup

To state our contributions and compare their guarantees precisely with prior work, we first introduce the logged-data model, the factored OT-DRO objective, and the notation used throughout the paper. We then summarize our contributions and discuss related work. Throughout this paper, let $\Delta(\Omega)$ denote the set of probability distributions over the set Ω. Denote $\mathcal{X}$ the finite context set, $\mathcal{A}$ the finite action space, and $\mathcal{Y} = [0, D]$ the reward range with known constant $D < \infty$.

Assumption 1 (Logged contextual-bandit model)**.** *We observe offline logged bandit data $\mathcal{D}_n = \{(X_i, A_i, Y_i)\}_{i=1}^n$ consisting of i.i.d. triples generated by*

$$X_i \sim P^0, \qquad A_i \mid X_i \sim \pi^0_{X_i}, \qquad Y_i \mid X_i, A_i \sim Q^0_{X_i,A_i},$$

or equivalently by the joint law $P^0(x)\pi^0_x(a)Q^0_{x,a}(y)$. Here $P^0 \in \Delta(\mathcal{X})$ is the nominal context distribution, $\pi^0_x \in \Delta(\mathcal{A})$ is the behavior policy conditional on context x, and $Q^0_{x,a} \in \Delta(\mathcal{Y})$ is the nominal reward distribution conditional on context x and action a. This data generating model rules out unobserved confounding: conditional on X_i, the logging action is drawn from $\pi^0_{X_i}$, and the reward law depends on the sample only through (X_i, A_i).

Let Π be a nonempty closed class of stochastic policies $\pi : \mathcal{X} \to \Delta(\mathcal{A})$. Our goal is to learn a policy $\pi^\star \in \Pi$ with optimal factored OT-DRO value. The following subsection defines this value.

1.2 Factored OT-DRO Objective

Assumption 2 (Transport costs)**.** *For $\mathsf{S} \in \{\mathcal{X}, \mathcal{Y}\}$, let $c_\mathsf{S} : \mathsf{S}^2 \to [0, \infty)$ be a nonzero lower semicontinuous cost satisfying $c_\mathsf{S}(s, s) = 0$ and $C_\mathsf{S} := \sup_{s,t\in\mathsf{S}} c_\mathsf{S}(s, t) \in (0, \infty)$.*

The transport cost c_S defines the severity of a shift in the space S. For example, in a clinical application, $c_\mathcal{X}(x, x')$ may be small if patient states x and x' are clinically similar, while $c_\mathcal{X}(x, x')$ may be large if they are substantially different. Similarly, $c_\mathcal{Y}(y, y')$ may be small if rewards y and y' are close in magnitude, while it may be large if they differ substantially.

Let $\mu, \nu \in \Delta(\mathsf{S})$ and $\Gamma(\mu, \nu)$ denote the set of all couplings between μ and ν. Define the transport discrepancy and the corresponding uncertainty set by

$$\mathcal{T}_\mathsf{S}(\mu, \nu) := \inf_{\gamma\in\Gamma(\mu,\nu)} \int c_\mathsf{S}(s, t)\, d\gamma(s, t), \qquad \mathcal{U}_\mathsf{S}(\mu, \epsilon) := \{\nu \in \Delta(\mathsf{S}) : \mathcal{T}_\mathsf{S}(\mu, \nu) \le \epsilon\}.$$

Instead of optimizing the nominal policy value $\mathbb{E}_{x\sim P^0}[\mathbb{E}_{a\sim\pi_x}[\mathbb{E}_{y\sim Q^0_{x,a}}[y]]]$, we assume that the deployment distributions of the context and reward lie in the factored OT ambiguity sets $\mathcal{U}_\mathcal{X}(P^0, \epsilon_\mathcal{X})$ and $\mathcal{U}_\mathcal{Y}(Q^0_{x,a}, \epsilon_\mathcal{Y})$ for given radii $\epsilon_\mathcal{X} > 0$ and $\epsilon_\mathcal{Y} > 0$. The factored OT-DRO value of a policy π is then defined as the worst-case expected reward over these ambiguity sets.

Definition 1 (Factored OT-DRO value)**.** *With given $\epsilon_\mathcal{X} > 0$ and $\epsilon_\mathcal{Y} > 0$, the factored OT-DRO value of policy $\pi \in \Pi$ is defined as*

$$V(\pi) = \inf_{P\in\mathcal{U}_\mathcal{X}(P^0,\epsilon_\mathcal{X})} \mathbb{E}_{x\sim P}\Big[\mathbb{E}_{a\sim\pi_x}[q(x, a)]\Big],$$

$$\text{where } q(x, a) := \inf_{Q_{x,a}\in\mathcal{U}_\mathcal{Y}(Q^0_{x,a},\epsilon_\mathcal{Y})} \mathbb{E}_{y\sim Q_{x,a}}[y]$$

is the robust reward for context-action cell (x, a).

The factored objective separates into an inner reward-distribution DRO problem for each context-action cell and an outer context-distribution DRO problem. The next two lemmas give dual representations of these problems. Their proofs follow standard arguments and are included in Appendix A for the reader's convenience.

Lemma 1 (Reward-level OT duality). *For every* $(x, a) \in \mathcal{X} \times \mathcal{A}$,

$$q(x,a) = \sup_{\lambda \in [0, D/\epsilon_{\mathcal{Y}}]} \left\{ -\lambda \epsilon_{\mathcal{Y}} + \mathbb{E}_{y \sim Q^0_{x,a}} \left[\inf_{z \in \mathcal{Y}} \{ z + \lambda c_{\mathcal{Y}}(y, z) \} \right] \right\}.$$

Lemma 2 (Context-level OT duality). *For every policy* $\pi \in \Pi$,

$$V(\pi) = \sup_{\eta \in [0, D/\epsilon_{\mathcal{X}}]} \left\{ -\eta \epsilon_{\mathcal{X}} + \mathbb{E}_{x \sim P^0} \left[\inf_{z \in \mathcal{X}} \left\{ \mathbb{E}_{a \sim \pi_z}[q(z,a)] + \eta c_{\mathcal{X}}(x, z) \right\} \right] \right\}.$$

1.3 Contributions

Sharp confidence bounds for robust rewards. Subsection 2.1 derives a finite-sample confidence bound β_Y for the robust reward $q(x, a)$ in each context–action cell. The key observation is that the OT dual integrands $s_\lambda(y) = \inf_{z \in \mathcal{Y}} \{z + \lambda c_{\mathcal{Y}}(y, z)\}$ are pointwise ordered as λ varies. This structure lets us control the entire dual family without discretizing either the dual parameter or the reward support. Consequently, the bound applies to arbitrary bounded reward distributions, including continuous ones, and removes the reward-support-cardinality dependence in the bound of Shen et al. (2024). Remark 1 shows that it is also strictly sharper.

Uniform confidence bounds for robust policy values. Subsection 2.2 propagates the reward bounds through the context-DRO problem and combines them with a distribution-dependent context-estimation bound, yielding a lower confidence bound β_X for $V(\pi)$ simultaneously for every policy $\pi \in \Pi$. Retaining the dependence on P^0, rather than replacing it with a bound in terms of $|\mathcal{X}|$, allows the context-estimation term in the regret analysis to be absorbed into $C^\star_{\mathrm{rob}}$.

Pessimistic learning for factored OT-DRO. Subsection 2.3 develops a pessimistic policy learning procedure (Algorithm 1) that maximizes the pessimistic value $\underline{V}(\pi; \delta) = \widehat{V}(\pi) - \beta_X(\delta)$ over the policy class Π. The algorithm requires knowledge of neither the true context distribution P^0 nor the true behavior policy π^0. For the unrestricted policy class, a maximizing policy is obtained by choosing $\widehat{\pi}(x) \in \arg\max_{a \in \mathcal{A}} \underline{q}(x, a) \forall x$, making the policy optimization computationally simple.

Minimax optimal regret rate. In Section 3, Theorem 1 gives a regret bound for the learned policy in terms of the robust optimal-policy coverage coefficient $C^\star_{\mathrm{rob}}$ (Assumption 3). Remark 2 shows that this coefficient is a natural generalization of the single-policy coverage coefficient of Li et al. (2022) to the factored OT-DRO setting by accounting for the worst-case context shift. Theorem 2 in Section 4 shows that our regret rate, including its dependence on $C^\star_{\mathrm{rob}}$, is minimax-optimal.

Numerical results. Section 5 presents experiments on synthetic and real-world datasets. The results validate the theoretical guarantees and demonstrate practical performance of our approach.

1.4 Related work

Distributionally Robust Offline Contextual Bandits. Table 1 compares the policy-learning guarantees of existing distributionally robust approaches. Si et al. (2020) study a joint KL-DRO objective using self-normalized inverse propensity weighting. They establish asymptotic $n^{-1/2}$ convergence, assuming that the behavior policy is known. Kallus et al. (2022) retain the joint KL-DRO objective and develop a doubly robust learner that permits an estimated behavior policy. Their finite-sample regret contains the product $\rho_{\pi^0} \rho_f$ of the behavior-policy and transformed-outcome regression errors. Consequently, the $n^{-1/2}$ rate is retained when both nuisance estimators converge faster than $n^{-1/4}$.

Mu et al. (2022) introduce separate KL ambiguity sets for context and reward shifts. Their cellwise plug-in analysis requires uniform context and action coverage. The resulting learning rate additionally depends on the reward-layer KL radius $\epsilon_{\mathcal{Y}}$ and a positive lower bound on its optimal dual

variable $\lambda^\star$. Shen et al. (2024) replace the two KL ambiguity sets by factored squared-Wasserstein balls. Their Theorem 5.1 contains a term depending on $|\Xi|$, where Ξ is the finite, known support of a latent observation ξ and the reward is represented as $Y = y_{x,a}(\xi)$. In our direct-reward model, one takes $\xi = Y$ and $y_{x,a}(\xi) = \xi$, so $\Xi = \mathcal{Y}$. In Table 1, $|\mathcal{Y}|$ refers there to the cardinality of a finite discrete reward support, and their guarantee does not directly cover continuous reward distributions.

Our work follows the factored approaches but accommodates OT ambiguity sets with general transport cost. Moreover, our bound replaces the worst-cell coverage factor $1/\sqrt{\underline{P}^0 \underline{\pi}^0}$ with the robust optimal-policy coverage coefficient $C^\star_{\mathrm{rob}}$. It has no dependence on optimal dual-variable bounds or the reward-support cardinality, and requires no assumed convergence rates for nuisance estimators. The lower bound shows that the dependence on $C^\star_{\mathrm{rob}}$ and $\log(|\mathcal{X}||\mathcal{A}|/\delta)$ is minimax-optimal.

Work	**Regret Bound**
Si et al. (2020) Joint KL-DRO	Asymptotic convergence at $n^{-1/2}$
Kallus et al. (2022) Joint KL-DRO	$O\left(\frac{\overline{\lambda} e^{D/\underline{\lambda}}}{\underline{\pi}^0 \sqrt{n}} \left[\kappa(\Pi) + \frac{D\overline{\lambda}}{\underline{\lambda}^2} + \sqrt{\log(1/\delta)}\right] + \frac{\overline{\lambda} e^{D/\underline{\lambda}}}{(\underline{\pi}^0)^2} \rho_{\pi^0} \rho_f\right)$
Mu et al. (2022) Factored KL-DRO	$O\left(\frac{D^3 e^{D/\underline{\lambda}}}{\epsilon_{\mathcal{Y}}^2 \underline{\lambda}^2 \underline{P}^0 \underline{\pi}^0} \sqrt{\frac{\log n \, \log(\lvert\mathcal{X}\rvert\lvert\mathcal{A}\rvert/\delta)}{n}}\right)$
Shen et al. (2024) Factored Wasserstein-DRO	$O\left(D \sqrt{\frac{\lvert\mathcal{Y}\rvert + \log(\lvert\mathcal{X}\rvert\lvert\mathcal{A}\rvert/\delta)}{n \underline{P}^0 \underline{\pi}^0}}\right)$
This work	
Factored OT-DRO (Theorem 1)	$O\left(D C^\star_{\mathrm{rob}} \sqrt{\frac{\log(\lvert\mathcal{X}\rvert\lvert\mathcal{A}\rvert/\delta)}{n}}\right)$
Lower bound (Theorem 2)	$\Omega\left(D C^\star_{\mathrm{rob}} \sqrt{\frac{\log(\lvert\mathcal{X}\rvert\lvert\mathcal{A}\rvert/\delta)}{n}}\right)$

Table 1: Regret bounds for offline contextual bandit learning under distributional shift.[1]

Pessimistic Offline Contextual Bandits. In the absence of environmental distribution shift, pessimistic offline learning can replace uniform coverage by coverage of an optimal policy. Direct approaches construct lower confidence bounds by estimating conditional rewards. In the tabular setting, Rashidinejad et al. (2021) obtain matching upper and lower bounds governed by a single-policy density-ratio coefficient. For linear reward models, Li et al. (2022) derive a sharper instance-dependent regret bound that aggregates inverse behavior probabilities across contexts. These methods rely on tabular or linear outcome structure.

Propensity-weighted approaches instead construct pessimistic estimates of policy-level performance. For general policy classes, Jin et al. (2025) establish minimax guarantees depending on the overlap of the optimal policy, covering both fixed and adaptively collected logging data, while Wang et al. (2024) obtain analogous coverage-adaptive guarantees with an oracle-efficient learner. Recent work by Wan et al. (2026) extends this approach to Lipschitz risk functionals using propensity-weighted estimators of the policy-induced reward distribution. These methods assume that the logging propensities are known.

All these works evaluate policies under a nominal environment, without adversarial shifts in either the conditional reward laws or the context distribution. Our method extends the pessimistic direct approach to the factored OT-DRO objective, which accounts separately for shifts in the context distribution and the conditional reward distributions. Correspondingly, $C^\star_{\mathrm{rob}}$ extends the single-policy coverage principle by incorporating the worst-case context distribution in the ambiguity set. In the no-context-shift specialization $\epsilon_{\mathcal{X}} = 0$, it reduces to the corresponding non-robust coverage scaling; see Remark 2.

[1]Written in our notation, with $\underline{\pi}^0 = \min_{x,a} \pi^0_x(a)$, $\underline{P}^0 = \min_x P^0(x)$, $\overline{\lambda} = \sup \lambda^\star$, $\underline{\lambda} = \inf \lambda^\star$. Moreover, $\kappa(\Pi)$ is the Hamming entropy integral, ρ_{π^0} is the L_2 behavior-policy estimation rate, and ρ_f is the uniform L_2 estimation rate of the transformed-outcome regressions.

2 PESSIMISTIC POLICY LEARNING FOR FACTORED OT-DRO

The learning procedure follows three steps. First, it estimates the robust reward $\widehat{q}(x,a)$ in every observed cell, subtracts its confidence penalty $\beta_Y(N_{x,a};\delta)$, and clips at zero to obtain the pessimistic reward $\underline{q}(x,a)$. Second, it evaluates policies through the outer context-DRO problem using these pessimistic rewards. Third, it selects the policy with the largest pessimistic empirical value. A context-distribution penalty $\beta_X(\delta)$ turns this objective into a valid policy-value lower confidence bound. Because this penalty is common to every policy, it provides the lower-confidence interpretation without affecting policy selection. The next three subsections develop these steps in order.

Empirical distributions. All estimators use the full logged dataset $\mathcal{D}_n$. For each cell (x,a), define the sample count $N_{x,a} := \sum_{i=1}^n \mathbf{1}\{X_i = x, A_i = a\}$. The empirical context distribution is denoted by $\widehat{P}(x) := \frac{1}{n}\sum_{i=1}^n \mathbf{1}\{X_i = x\}$.

2.1 ROBUST REWARD LOWER CONFIDENCE BOUNDS

For $N_{x,a} > 0$, define the empirical robust reward by

$$\widehat{q}(x,a) := \sup_{\lambda \in [0, D/\epsilon_{\mathcal{Y}}]} \left\{ -\lambda \epsilon_{\mathcal{Y}} + \frac{1}{N_{x,a}} \sum_{i: X_i = x, A_i = a} \inf_{z \in \mathcal{Y}} \{ z + \lambda c_{\mathcal{Y}}(Y_i, z) \} \right\}. \tag{1}$$

If $N_{x,a} = 0$, define $\widehat{q}(x,a) := 0$.

For $m \in \{0, 1, \ldots, n\}$, define the cellwise reward confidence radius

$$\beta_Y(m;\delta) := \begin{cases} D, & m = 0, \\ D \min\left\{ 1, \sqrt{\frac{\pi}{2m}} + \sqrt{\frac{\log(2|\mathcal{X}||\mathcal{A}|/\delta)}{2m}} \right\}, & m \geq 1. \end{cases} \tag{2}$$

Now define the pessimistic reward estimate

$$\underline{q}(x,a) := \max\{0,\ \widehat{q}(x,a) - \beta_Y(N_{x,a};\delta)\}.$$

Thus $\underline{q}(x,a) \in [0, D]$ for every cell; in particular, an empty cell has $\widehat{q}(x,a) = \underline{q}(x,a) = 0$. The next proposition shows that $\underline{q}$ is a valid lower confidence bound for the robust reward q.

Proposition 1 (Reward-cell concentration)**.** *Under Assumptions 1 and 2, with probability at least $1 - \delta/2$ over the draw of the logged data $\mathcal{D}_n$, simultaneously for every $(x,a) \in \mathcal{X} \times \mathcal{A}$,*

$$|\widehat{q}(x,a) - q(x,a)| \leq \beta_Y(N_{x,a};\delta).$$

On the same event, $0 \leq \underline{q}(x,a) \leq q(x,a)\ \forall (x,a)$, and $q(x,a) - \underline{q}(x,a) \leq 2\beta_Y(N_{x,a};\delta)\ \forall (x,a)$.

Remark 1 (Comparison with Shen et al. (2024))**.** *For each context–action cell, Shen et al. (2024) use the plug-in empirical conditional law and control the difference between the empirical and population Wasserstein-robust costs by the total-variation error of that law on a finite, discrete reward support $\mathcal{Y}$; see their Lemma A.4. Applied to a cell containing m observations, that route gives a confidence radius of*

$$D\left\{ \sqrt{\frac{2|\mathcal{Y}|}{m}} + \sqrt{\frac{4\log(2|\mathcal{X}||\mathcal{A}|/\delta)}{m}} \right\},$$

which becomes a dependence on the reward-support cardinality. Our proof of Proposition 1 instead controls the OT dual integrand $s_\lambda(y) = \inf_{z \in \mathcal{Y}}\{z + \lambda c_{\mathcal{Y}}(y,z)\}$ directly. Since $s_\lambda(y)$ increases in λ pointwise in y, its pointwise ordering, a layer-cake representation, and the Dvoretzky–Kiefer–Wolfowitz inequality yield the radius

$$D\left\{ \sqrt{\frac{\pi}{2m}} + \sqrt{\frac{\log(2|\mathcal{X}||\mathcal{A}|/\delta)}{2m}} \right\}$$

before truncation at D; see our detailed proof in Appendix B.1. Thus our bound has no reward-support-cardinality term. It therefore applies to arbitrary bounded reward laws, including continuous ones, and is strictly sharper. This improvement comes from the ordered-class analysis.

2.2 From Reward Bounds to Policy-Value Bounds

With the empirical context distribution $\widehat{P}$ and the pessimistic reward table $\underline{q}$, the empirical context-robust value based on pessimistic rewards is defined by

$$\begin{aligned}\widehat{V}(\pi) &:= \inf_{P\in\mathcal{U}_{\mathcal{X}}(\widehat{P},\epsilon_{\mathcal{X}})} \mathbb{E}_{X\sim P}\left[\sum_{a\in\mathcal{A}} \pi_X(a)\underline{q}(X,a)\right] \\ &= \sup_{\eta\in[0,D/\epsilon_{\mathcal{X}}]}\left\{-\eta\epsilon_{\mathcal{X}} + \mathbb{E}_{X\sim\widehat{P}}\left[\min_{z\in\mathcal{X}}\left\{\sum_{a\in\mathcal{A}}\pi_z(a)\underline{q}(z,a) + \eta c_{\mathcal{X}}(X,z)\right\}\right]\right\}.\end{aligned} \tag{3}$$

Define the distribution-dependent context penalty

$$\beta_X(\delta) := D\min\left\{1, \frac{1}{2\sqrt{n}}\left[\sum_{x\in\mathcal{X}}\sqrt{P^0(x)(1-P^0(x))} + \sqrt{2\log(2/\delta)}\right]\right\}. \tag{4}$$

The pessimistic factored OT-DRO value estimate of policy π is

$$\underline{V}(\pi;\delta) := \widehat{V}(\pi) - \beta_X(\delta).$$

The following proposition shows that $\underline{V}(\pi;\delta)$ is a valid lower confidence bound for the factored OT-DRO value $V(\pi)$. Its proof is given in Appendix B.3.

Proposition 2 (Policy-value lower confidence bound)**.** *Under Assumption 1 and 2, with probability at least $1-\delta$ over the draw of the logged data $\mathcal{D}_n$, simultaneously for every $\pi\in\Pi$, $\underline{V}(\pi;\delta)\le V(\pi)$. Consequently, on the same event,*

$$V(\pi) - \underline{V}(\pi;\delta) \le 2\sup_{P\in\mathcal{U}_{\mathcal{X}}(P^0,\epsilon_{\mathcal{X}})}\mathbb{E}_{X\sim P}\left[\sum_{a\in\mathcal{A}}\pi_X(a)\beta_Y(N_{X,a};\delta)\right] + 2\beta_X(\delta) \qquad \forall\pi\in\Pi.$$

2.3 Pessimistic Policy Learning

Pessimistic Policy Learning maximizes the pessimistic value $\underline{V}(\pi;\delta)$ over the policy class Π. Because the context penalty $\beta_X(\delta)$ is independent of π, this is equivalent to maximizing the empirical value $\widehat{V}(\pi)$: $\arg\max_{\pi\in\Pi}\underline{V}(\pi;\delta) = \arg\max_{\pi\in\Pi}\widehat{V}(\pi)$. Consequently, $\beta_X(\delta)$ need not be computed. There is no need to replace the sharp P^0-dependent penalty by a looser computable one: the learner maximizes $\widehat{V}$, while $\underline{V}$ remains a valid policy-value lower confidence bound in the analysis. Algorithm 1 summarizes the resulting procedure.

Algorithm 1 Pessimistic Factored OT-DRO Offline Policy Learning

1: Input: data $\mathcal{D}_n$, policy class Π, costs $c_{\mathcal{Y}}, c_{\mathcal{X}}$, radii $\epsilon_{\mathcal{Y}}, \epsilon_{\mathcal{X}}$, reward bound D, confidence δ.
2: Compute $\widehat{P}$ and the cell counts $N_{x,a}$.
3: **for** each $(x,a)\in\mathcal{X}\times\mathcal{A}$ **do**
4: Compute $\widehat{q}(x,a)$ by (1) and $\beta_Y(N_{x,a};\delta)$ by (2).
5: Compute $\underline{q}(x,a) = \max\{0, \widehat{q}(x,a) - \beta_Y(N_{x,a};\delta)\}$.
6: **end for**
7: Solve $\max_{\pi\in\Pi}\widehat{V}(\pi)$, with $\widehat{V}(\pi)$ defined by (3).
8: Output any $\widehat{\pi}_\delta\in\arg\max_{\pi\in\Pi}\widehat{V}(\pi)$; equivalently, $\widehat{\pi}_\delta\in\arg\max_{\pi\in\Pi}\underline{V}(\pi;\delta)$.

Lexicographic tie-breaking with unclipped pessimistic rewards. In Algorithm 1, when the primary optimization has multiple solutions, we may break ties using the corresponding unclipped scores $\widetilde{q}(x,a) := \widehat{q}(x,a) - \beta_Y(N_{x,a};\delta)$. Specifically, define

$$\widetilde{V}(\pi) := \inf_{P\in\mathcal{U}_{\mathcal{X}}(\widehat{P},\epsilon_{\mathcal{X}})}\mathbb{E}_{X\sim P}\left[\sum_{a\in\mathcal{A}}\pi_X(a)\widetilde{q}(X,a)\right], \qquad \widehat{\Pi}_\delta := \arg\max_{\pi\in\Pi}\widehat{V}(\pi),$$

and select $\widehat{\pi}_\delta\in\arg\max_{\pi\in\widehat{\Pi}_\delta}\widetilde{V}(\pi)$. This secondary criterion retains information discarded by clipping: when many pessimistic rewards $\underline{q}$ equal zero, it favors, among the policies tied under the clipped objective, those with larger values of $\widehat{q}(x,a) - \beta_Y(N_{x,a};\delta)$. Since the selected policy remains an element of $\widehat{\Pi}_\delta$, this tie-breaking rule does not alter any of the theoretical guarantees.

Optimization over the unrestricted policy class. For the class of all stochastic policies, $\Pi = \Delta(\mathcal{A})^{|\mathcal{X}|}$, the optimization in Algorithm 1 separates across contexts. An optimal policy can therefore be chosen deterministically according to $\widehat{\pi}_\delta(x) \in \arg\max_{a\in\mathcal{A}} \underline{q}(x,a), \forall x \in \mathcal{X}$. Indeed, this choice maximizes the pessimistic reward at every context and hence its expectation under every context distribution in the ambiguity set. Moreover, one may choose $\widehat{\pi}_\delta(x) \in \arg\max_{a\in\mathcal{A}} \widetilde{q}(x,a), \forall x \in \mathcal{X}$. Because $\underline{q}(x,a) = \max\{0, \widetilde{q}(x,a)\}$, every maximizer of $\widetilde{q}(x,\cdot)$ is also a maximizer of $\underline{q}(x,\cdot)$. This implements the lexicographic tie-breaking described above.

Optimization over a structured policy class. When Π is described by linear constraints on the policy probabilities, the policy-learning step is a linear program; other structured classes may be handled using an appropriate optimization oracle. Exact optimization is not required for the statistical guarantee: if the returned policy satisfies $\widehat{V}(\widehat{\pi}_\delta) \geq \sup_{\pi\in\Pi} \widehat{V}(\pi) - \tau$ for some $\tau \geq 0$, then all subsequent regret upper bounds hold with an additional additive term τ.

Regret in terms of random cell counts. Let the true optimal policy be $\pi^\star \in \arg\max_{\pi\in\Pi} V(\pi)$ and the selected policy be $\widehat{\pi}_\delta \in \arg\max_{\pi\in\Pi} \underline{V}(\pi;\delta)$. On the event of Proposition 2, we have $V(\widehat{\pi}_\delta) \geq \underline{V}(\widehat{\pi}_\delta;\delta) \geq \underline{V}(\pi^\star;\delta)$. Applying the gap bound from Proposition 2 gives a regret bound in terms of the random cell counts:

$$V(\pi^\star) - V(\widehat{\pi}_\delta) \leq 2 \sup_{P\in\mathcal{U}_\mathcal{X}(P^0,\epsilon_\mathcal{X})} \mathbb{E}_{X\sim P}\left[\sum_{a\in\mathcal{A}} \pi^\star_X(a)\beta_Y(N_{X,a};\delta)\right] + 2\beta_X(\delta).$$

The next section converts this into a deterministic regret rate under a coverage condition.

3 Regret Under Robust Optimal-Policy Coverage

The preceding bound depends on random cell counts. We now convert it into a deterministic rate under a robust optimal-policy coverage condition. Assumption 3 generalizes the single-policy coverage condition to the factored OT-DRO setting. Theorem 1 then gives an explicit regret bound in terms of this robust optimal-policy coverage coefficient.

Assumption 3 (Robust optimal-policy coverage)**.**

$$C^\star_{\mathrm{rob}} := \sup_{P\in\mathcal{U}_\mathcal{X}(P^0,\epsilon_\mathcal{X})} \sum_{x:P(x)>0} \frac{P(x)}{\sqrt{P^0(x)}} \sum_{a:\pi^\star_x(a)>0} \frac{\pi^\star_x(a)}{\sqrt{\pi^0_x(a)}} < \infty.$$

Remark 2 (Recovery of non-robust single-policy coverage)**.** *Consider the no-context-shift specialization $\epsilon_\mathcal{X} = 0$ with $\mathcal{U}_\mathcal{X}(P^0,0) = \{P^0\}$; the singleton condition holds, for example, when $c_\mathcal{X}(x,z) > 0$ for all $x \neq z$. Then*

$$\begin{aligned} C^\star_{\mathrm{rob}}\big|_{\epsilon_\mathcal{X}=0} &= \sum_{x:P^0(x)>0} \sqrt{P^0(x)} \sum_{a:\pi^\star_x(a)>0} \frac{\pi^\star_x(a)}{\sqrt{\pi^0_x(a)}} \\ &= \sum_{x:P^0(x)>0} \sqrt{\frac{P^0(x)}{\pi^0_x(\pi^\star(x))}} \quad \textit{for a deterministic optimal policy.} \end{aligned}$$

This is exactly the instance-dependent coverage term in Corollary 1 of Li et al. (2022). Meanwhile, the density-ratio single-policy concentrability coefficient of Rashidinejad et al. (2021) gives the coarser comparison

$$C^\star_{\mathrm{rob}}\big|_{\epsilon_\mathcal{X}=0} \leq \sqrt{|\mathcal{X}|C^\star_1}, \qquad C^\star_1 := \max_{x:P^0(x)>0} \frac{1}{\pi^0_x(\pi^\star(x))},$$

resulting in regret bounds of order $\tilde{O}(\sqrt{|\mathcal{X}|C^\star_1/n})$ in their Theorem 4 (upper bound) and Theorem 5 (lower bound). Similarly, the stochastic-policy pseudo-loss of Wang et al. (2024) also gives

$$C^\star_{\mathrm{rob}}\big|_{\epsilon_\mathcal{X}=0} \leq \sqrt{|\mathcal{X}|C^\star_2}, \qquad C^\star_2 := \mathbb{E}_{X\sim P^0}\left[\sum_{a:\pi^\star_X(a)>0} \frac{\pi^\star_X(a)}{\pi^0_X(a)}\right].$$

Hence the no-context-shift specialization recovers the familiar single-policy coverage scaling. Note that the factor $\sqrt{|\mathcal{X}|}$ in this comparison comes from replacing the instance-dependent sum by the coarser worst-case coefficient $C^\star_1$; it is not an additional cost of our robust bound.

Robust optimal-policy coverage is stronger Because $P^0 \in \mathcal{U}_{\mathcal{X}}(P^0, \epsilon_{\mathcal{X}})$, the robust coverage coefficient is at least as large as the corresponding coefficient evaluated under the nominal context distribution. Moreover, since $\epsilon_{\mathcal{X}} > 0$ and $c_{\mathcal{X}}$ is bounded, every context $x \in \mathcal{X}$ receives positive mass under some $P \in \mathcal{U}_{\mathcal{X}}(P^0, \epsilon_{\mathcal{X}})$: one can transport an arbitrarily small positive amount of mass from any context in the support of P^0 to x while remaining within the transport budget. Consequently, finiteness of $C^\star_{\mathrm{rob}}$ is equivalent to

$$P^0(x) > 0 \quad \text{for every } x \in \mathcal{X}, \qquad \text{and} \qquad \pi^0_x(a) > 0 \quad \text{for every } (x, a) \text{ such that } \pi^\star_x(a) > 0.$$

Thus, unlike the no-context-shift setting, which requires coverage only at contexts in the support of P^0, positive-radius context robustness requires the logged data distribution to have full context support. Nevertheless, it requires action coverage only along the optimal robust policy and imposes no coverage requirement on actions that $\pi^\star$ never uses. The mismatch between the contexts emphasized by an adversarial deploying distribution and those observed under the logging distribution is what makes robust optimal-policy coverage stronger than its non-robust counterpart.

Using the robust optimal-policy coverage coefficient, we can convert the random-cell-count regret bound into a deterministic rate in the following theorem. Its proof is given in Appendix C.

Theorem 1 (Regret under robust optimal-policy coverage)**.** *Let $\delta \in (0, 1)$ and suppose Assumption 1-3 holds. Then the learner run with confidence $\delta/2$ satisfies, with probability at least $1 - \delta$,*

$$V(\pi^\star) - V(\widehat{\pi}_{\delta/2}) \le \left(2\sqrt{\frac{2}{1 - \log 2}} + 1 + \sqrt{2}\right) D C^\star_{\mathrm{rob}} \sqrt{\frac{\log(4|\mathcal{X}||\mathcal{A}|/\delta)}{n}}.$$

In our analysis, the context-estimation error, controlled by the distribution-dependent radius $\beta_X(\delta)$ in (4), is absorbed into the robust optimal-policy coverage coefficient:

$$\sum_{x \in \mathcal{X}} \sqrt{P^0(x)(1 - P^0(x))} \le \sum_{x \in \mathcal{X}} \sqrt{P^0(x)} \le C^\star_{\mathrm{rob}}.$$

By comparison, Theorem 5.1 of Shen et al. (2024), using the total-variation bound in their equation (A.3), incurs a context-estimation term of order $D(\sqrt{\log(1/\delta)/n} + \sqrt{|\mathcal{X}|/n})$. Under their uniform context-coverage assumption, $|\mathcal{X}| \le 1/\underline{P}^0$, so the second term can be absorbed into the overall $1/\sqrt{\underline{P}^0}$ dependence.

4 Minimax Lower Bound

We next show that the leading dependence on robust optimal-policy coverage and the full context–action cardinality logarithm are unavoidable. The construction has one anchor and $|\mathcal{X}| - 1$ hard contexts, and it indexes the hard family jointly by a context and an action. Both ambiguity radii are strictly positive: the context adversary transports mass from the anchor toward the hard contexts, while the reward adversary lowers every Bernoulli reward law within a squared-Wasserstein ball. The proof is given in Appendix D.

Theorem 2 (Lower bound)**.** *Fix $D, \kappa > 0$, finite sets $\mathcal{X}$ and $\mathcal{A}$ with $|\mathcal{X}| \ge 3$ and $|\mathcal{A}| \ge 3$, $0 < \epsilon_{\mathcal{X}} < \kappa/2$, and $0 < \epsilon_{\mathcal{Y}} \le D^2/2$. Write*

$$r := |\mathcal{X}| - 1, \qquad L := |\mathcal{A}| - 1, \qquad \rho := \frac{\epsilon_{\mathcal{X}}}{\kappa}, \qquad w := \frac{1}{2r} + \rho.$$

Fix $0 < \mu \le \frac{w^2}{1 + Lw^2}$. For every $\delta \in (0, 1/4]$ satisfying $n\mu \ge \frac{7r}{8} \log\left(\frac{4|\mathcal{X}||\mathcal{A}|}{\delta}\right)$, and every learner $\widehat{\pi}$ that outputs a policy in the supplied policy class, there exists a $|\mathcal{X}|$-context, $|\mathcal{A}|$-action instance with both context and reward shifts active and

$$C^\star_{\mathrm{rob}} = w\sqrt{\frac{2r}{\mu}} + \left(\frac{1}{2} - \frac{1}{2r}\right)\sqrt{\frac{2r}{1 - L\mu}} + \left(\frac{1}{2} - \rho\right)\sqrt{2},$$

such that, under its logged-data law $\mathbb{P}^n$,

$$\mathbb{P}^n\left(V(\pi^\star) - V(\widehat{\pi}) \ge \frac{\sqrt{7}}{512} D C^\star_{\mathrm{rob}} \sqrt{\frac{\log(4|\mathcal{X}||\mathcal{A}|/\delta)}{n}}\right) > \delta.$$

Theorem 2 matches the rate of Theorem 1, up to universal constants. Thus the reward scale, robust optimal-policy coverage, context–action cardinality logarithm, confidence dependence, and $n^{-1/2}$ rate are all unavoidable in the worst case.

5 Numerical Experiments

We examine the robust-reward confidence bound and the predicted regret dependence through synthetic experiments, then evaluate algorithm 1 on real logged-bandit data.

Robust-reward confidence bound. Appendix E.1 evaluates the bound in Proposition 1 on 58 random and structured reward distributions, using quadratic transport cost and 20 sample sizes between 10 and 10^6. No bound violation occurs across 1.16 million replications. The empirical 95th-percentile estimation errors exhibit approximately $m^{-1/2}$ decay and remain below the theoretical radius throughout (Figure 1). The radius is at least 2.79 times these empirical quantiles, supporting the confidence guarantee on the tested distributions while demonstrating conservative constants.

Dependence on sample size and coverage. Appendix E.2 studies Bernoulli instances with 4 contexts, 9 actions, and an exactly prescribed coverage coefficient $C^\star_{\mathrm{rob}}$. At each of 425 combinations of n and $C^\star_{\mathrm{rob}}$, we select a reward gap with large estimated regret and evaluate it using independent simulations. The regret curves exhibit a small-sample plateau followed by decay (Figures 2–4). For $n \geq 2^{22}$, a joint log–log fit gives exponents -0.5004 for n and 1.0021 for $C^\star_{\mathrm{rob}}$. These results support the $C^\star_{\mathrm{rob}}/\sqrt{n}$ dependence predicted by Theorem 1 in the large-sample regime of this family.

Real-data application. We use the Open Bandit Dataset (OBD), collected on the fashion e-commerce platform ZOZOTOWN (Saito et al., 2020). The uniform-random logs for the *men*, *women*, and *all* campaigns contain approximately 0.45–1.37 million rows per campaign; recommended items are actions, and binary clicks are rewards. We treat each logged item as a separate bandit observation and split each log by rows into 60% training, 20% validation, and 20% test data. Appendix E.3 provides the full protocol.

We implement Algorithm 1 over the unrestricted policy class. To adapt the ambiguity radius to the low click-through rate (CTR), we set $\epsilon_{\mathcal{Y}} = \rho\,\widehat{\mathrm{CTR}}_{\mathrm{train}}$ where ρ is a relative-radius multiplier. Motivated by the conservatism observed above, we introduce a confidence scale $c \geq 0$ and select $\widehat{\pi}_c(x) \in \arg\max_{a\in\mathcal{A}} \{\widehat{q}(x,a) - c\,\beta_Y(N_{x,a}; 0.1)\}$. Using the unclipped scores implements the tie-breaking described in Section 2.3. For this policy class, selection is independent of $\epsilon_{\mathcal{X}}$, and β_X need not be computed. The choice $c = 1$ retains the theoretical confidence radius; choosing $c < 1$ reduces conservatism. We tune (ρ, c) over a 36-point grid using validation self-normalized inverse propensity scoring. The selected policy is then evaluated on the held-out test split.

Results. Table 2 compares our learner with factored W-DRO (Shen et al., 2024), the IPW Learner (Saito et al., 2020), POEM (Swaminathan & Joachims, 2015), and joint KL-DRO (Si et al., 2020). Our learner attains the largest test SNIPS point estimate in every campaign.

Campaign	Behavior policy	Factored W-DRO	IPW Learner	POEM	Joint KL-DRO	Ours
men	0.544%	2.129%	1.087%	1.570%	0.614%	**2.155%**
women	0.483%	1.763%	1.410%	1.315%	0.936%	**1.796%**
all	0.327%	1.734%	1.436%	1.086%	0.837%	**1.817%**

Table 2: Test policy-value estimates (%) on OBD. Learned policies are evaluated by SNIPS.

6 Conclusion

We established minimax-optimal regret guarantees for pessimistic policy learning under factored OT-DRO. The guarantees require action coverage only along an optimal robust policy, and the robust optimal-policy coverage coefficient captures how admissible context shifts amplify limited coverage in the logged data. Our analysis combines sharp confidence bounds for robust rewards with a distribution-dependent context-estimation bound. The real-data experiment illustrates the learner's performance on logged recommendations, but it does not isolate the benefit of our weaker action-coverage requirement. It does not test performance under a measured deployment shift. Because this work focuses on the theoretical guarantees, we leave experiments with controlled coverage and meaningful distribution shifts, as well as extensions beyond finite context spaces, to future work.

AI USE STATEMENT

In this work, we used generative AI tools to assist with implementing the experimental evaluation, including coding the three experiments reported in the paper. We also used generative AI tools to polish writing and to identify related work. We did not use generative AI tools to develop the theoretical model or conceptual framework, formulate mathematical claims, provide ingredients for or write mathematical proofs, propose or refine hypotheses, design the research methodology or experiments, generate or analyze datasets, or interpret experimental results.

All AI-generated experimental code was reviewed, verified, and tested for correctness by the authors. AI-identified related work was independently reviewed and validated by the authors. We reviewed all AI-assisted work and take responsibility for the final content of this work, including text, claims, code, and other artifacts produced with the aid of generative AI.

REFERENCES

Léon Bottou, Jonas Peters, Joaquin Quiñonero-Candela, Denis X. Charles, D. Max Chickering, Elon Portugaly, Dipankar Ray, Patrice Simard, and Ed Snelson. Counterfactual reasoning and learning systems: The example of computational advertising. *Journal of Machine Learning Research*, 14 (101):3207–3260, 2013. URL http://jmlr.org/papers/v14/bottou13a.html.

Miroslav Dudík, John Langford, and Lihong Li. Doubly robust policy evaluation and learning. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, ICML'11, pp. 1097–1104, Madison, WI, USA, 2011. Omnipress. ISBN 9781450306195.

Ying Jin, Zhimei Ren, Zhuoran Yang, and Zhaoran Wang. Policy learning "without" overlap: Pessimism and generalized empirical Bernstein's inequality. *The Annals of Statistics*, 53(4):1483–1512, 2025. doi: 10.1214/25-AOS2511. URL https://arxiv.org/abs/2212.09900.

Nathan Kallus, Xiaojie Mao, Kaiwen Wang, and Zhengyuan Zhou. Doubly robust distributionally robust off-policy evaluation and learning. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 10598–10632. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/kallus22a.html.

Gene Li, Cong Ma, and Nati Srebro. Pessimism for offline linear contextual bandits using \ell_p confidence sets. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 20974–20987. Curran Associates, Inc., 2022. doi: 10.52202/068431-1525. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/8443219a991f068c34d9491ad68ffa94-Paper-Conference.pdf.

Lihong Li, Wei Chu, John Langford, and Robert E. Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, pp. 661–670, New York, NY, USA, 2010. Association for Computing Machinery. ISBN 9781605587998. doi: 10.1145/1772690.1772758. URL https://doi.org/10.1145/1772690.1772758.

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. doi: 10.52202/068431-0605. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/375cc00f0a2297af84304d0c0cd3b7ed-Paper-Conference.pdf.

Paria Rashidinejad, Banghua Zhu, Cong Ma, Jiantao Jiao, and Stuart Russell. Bridging offline reinforcement learning and imitation learning: A tale of pessimism. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 11702–11716. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/60ce36723c17bbac504f2ef4c8a46995-Paper.pdf.

Yuta Saito, Shunsuke Aihara, Megumi Matsutani, and Yusuke Narita. A large-scale open dataset for bandit algorithms. *CoRR*, abs/2008.07146, 2020. URL https://arxiv.org/abs/2008.07146.

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. URL https://openreview.net/forum?id=NmpjDHWIvg. Featured Certification.

Nian Si, Fan Zhang, Zhengyuan Zhou, and Jose Blanchet. Distributionally robust policy evaluation and learning in offline contextual bandits. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 8884–8894. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/si20a.html.

Adith Swaminathan and Thorsten Joachims. Counterfactual risk minimization: Learning from logged bandit feedback. In Francis Bach and David Blei (eds.), *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 814–823, Lille, France, 07–09 Jul 2015. PMLR. URL https://proceedings.mlr.press/v37/swaminathan15.html.

Ambuj Tewari and Susan A. Murphy. From ads to interventions: Contextual bandits in mobile health. In *Mobile Health - Sensors, Analytic Methods, and Applications*, 2017. URL https://api.semanticscholar.org/CorpusID:18778220.

Yilong Wan, Yuqiang Li, and Xianyi Wu. Pessimistic risk-aware policy learning in contextual bandits, 2026. URL https://arxiv.org/abs/2605.15620.

Lequn Wang, Akshay Krishnamurthy, and Alex Slivkins. Oracle-efficient pessimism: Offline policy optimization in contextual bandits. In Sanjoy Dasgupta, Stephan Mandt, and Yingzhen Li (eds.), *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pp. 766–774. PMLR, 02–04 May 2024. URL https://proceedings.mlr.press/v238/wang24a.html.

APPENDIX

This appendix provides detailed proofs of the main results and further information on the numerical experiments. The appendix is organized as follows.

CONTENTS

Item	Page
Appendix A: OT Duality and Deterministic Properties	12
A.1 Reward-level OT duality (Lemma 1)	12
A.2 Reward dual integrand ordering and regularity (Lemma 3)	14
A.3 Context-level OT duality (Lemma 2)	15
A.4 Monotonicity and Lipschitzness of the context robust value (Lemma 4)	16
Appendix B: Finite-Sample Confidence Bounds	18
B.1 Reward-cell concentration (Proposition 1)	18
B.2 Total-variation context concentration (Lemma 5)	20
B.3 Policy-value lower confidence bound (Proposition 2)	21
Appendix C: Policy-Learning Guarantees	23
C.1 Count-to-confidence conversion (Lemma 6)	23
C.2 Regret under robust optimal-policy coverage (Theorem 1)	24
Appendix D: Proof of the Minimax Lower Bound (Theorem 2)	26
D.1 Step 1: Hard instances	26
D.2 Step 2: Robust regret and coverage	27
D.3 Step 3: Relative entropy	28
D.4 Step 4: Fano and binary testing	28
Appendix E: Experiment Details	31
E.1 Empirical Behavior of the Robust-Reward Confidence Radius	31
E.2 Dependence of Regret on n and C^*_{rob}	33
E.3 Real-Data Application: Off-Policy Learning on the Open Bandit Dataset	36

A OT DUALITY AND DETERMINISTIC PROPERTIES

We derive the reward- and context-level dual representations by expressing each transport constraint through couplings and introducing a Lagrange multiplier. We then establish two properties used in the statistical analysis: ordering and regularity of the reward dual integrands, and monotonicity and Lipschitzness of the context-robust value. The proofs use standard arguments and are included for the reader's convenience.

A.1 REWARD-LEVEL OT DUALITY (LEMMA 1)

Proof. Fix a context-action pair (x, a) and write $c := c_{\mathcal{Y}}$. Recall that $q(x, a) = \inf_{Q:\mathcal{T}_{\mathcal{Y}}(Q^0_{x,a},Q)\le\epsilon_{\mathcal{Y}}} \mathbb{E}_{Z\sim Q}[Z]$. We first write the problem using couplings. Let

$$\mathfrak{C}(Q^0_{x,a}) := \{\gamma \in \Delta(\mathcal{Y}\times\mathcal{Y}) : \gamma_1 = Q^0_{x,a}\},$$

where γ_1 is the first marginal. If $\gamma \in \mathfrak{C}(Q^0_{x,a})$, then its second marginal γ_2 is a candidate reward distribution. Conversely, every candidate reward distribution Q and every coupling in $\Gamma(Q^0_{x,a}, Q)$ give an element of $\mathfrak{C}(Q^0_{x,a})$.

Because $\mathcal{Y} = [0, D]$ is compact and c is lower semicontinuous and bounded below, the optimal transport infimum defining $\mathcal{T}_{\mathcal{Y}}(Q^0_{x,a}, Q)$ is attained. Hence $\mathcal{T}_{\mathcal{Y}}(Q^0_{x,a}, Q) \le \epsilon_{\mathcal{Y}}$ is equivalent to the existence of a coupling $\gamma \in \Gamma(Q^0_{x,a}, Q)$ satisfying $\int c(y,z)\, d\gamma(y,z) \le \epsilon_{\mathcal{Y}}$. Therefore, we have

$$q(x,a) = \inf_{\gamma \in \mathfrak{C}(Q^0_{x,a}) : \int c(y,z) d\gamma(y,z) \le \epsilon_{\mathcal{Y}}} \int z\, d\gamma(y,z).$$

For brevity define

$$A(\gamma) := \int z\, d\gamma(y,z), \qquad B(\gamma) := \int c(y,z)\, d\gamma(y,z).$$

Then

$$q(x,a) = \inf_{\gamma \in \mathfrak{C}(Q^0_{x,a}) : B(\gamma) \le \epsilon_{\mathcal{Y}}} A(\gamma).$$

For $\lambda \ge 0$, define the Lagrangian $L(\gamma, \lambda) := A(\gamma) + \lambda(B(\gamma) - \epsilon_{\mathcal{Y}})$. If γ is feasible, then $B(\gamma) \le \epsilon_{\mathcal{Y}}$, so $L(\gamma, \lambda) \le A(\gamma)$. For every $\lambda \ge 0$, we have

$$\inf_{\gamma \in \mathfrak{C}(Q^0_{x,a})} L(\gamma, \lambda) \le \inf_{\gamma \in \mathfrak{C}(Q^0_{x,a}) : B(\gamma) \le \epsilon_{\mathcal{Y}}} A(\gamma) = q(x,a).$$

Taking the supremum over $\lambda \ge 0$ gives weak duality:

$$\sup_{\lambda \ge 0} \inf_{\gamma \in \mathfrak{C}(Q^0_{x,a})} L(\gamma, \lambda) \le q(x,a).$$

We now prove equality. Since $\mathcal{Y} \times \mathcal{Y}$ is compact, $\mathfrak{C}(Q^0_{x,a})$ is weakly compact. The map A is continuous because z is bounded and continuous. The map B is lower semicontinuous because c is lower semicontinuous and bounded below. Since $0 \le c \le C_{\mathcal{Y}}$, every coupling satisfies $0 \le B(\gamma) \le C_{\mathcal{Y}}$.

Define the value function

$$w(t) := \inf_{\gamma \in \mathfrak{C}(Q^0_{x,a}) : B(\gamma) \le t} A(\gamma), \qquad t \in [0, C_{\mathcal{Y}}].$$

The diagonal coupling has cost zero because $c(y,y) = 0$, so $w(t)$ is finite for all $t \in [0, C_{\mathcal{Y}}]$. The function w is nonincreasing because a larger budget gives a larger feasible set. It is also convex: if γ_1 and γ_2 are feasible for budgets t_1 and t_2, then $\theta\gamma_1 + (1-\theta)\gamma_2$ is feasible for budget $\theta t_1 + (1-\theta)t_2$, and the objective value is the same convex combination. Taking near-minimizers and letting the approximation error vanish gives convexity.

If $\epsilon_{\mathcal{Y}} \ge C_{\mathcal{Y}}$, then the transport constraint is inactive. The primal value is $\inf_{\gamma \in \mathfrak{C}(Q^0_{x,a})} A(\gamma)$, and this is exactly the dual value at $\lambda = 0$. Now suppose $0 < \epsilon_{\mathcal{Y}} < C_{\mathcal{Y}}$. Since w is finite and convex on an interval, it has a subgradient at $\epsilon_{\mathcal{Y}}$. Let $s \in \partial w(\epsilon_{\mathcal{Y}})$. Since w is nonincreasing, we may take $s \le 0$. Set $\lambda^\star := -s \ge 0$. The subgradient inequality gives, for all $t \in [0, C_{\mathcal{Y}}]$,

$$w(t) \ge w(\epsilon_{\mathcal{Y}}) + s(t - \epsilon_{\mathcal{Y}}) = w(\epsilon_{\mathcal{Y}}) - \lambda^\star(t - \epsilon_{\mathcal{Y}}).$$

For any $\gamma \in \mathfrak{C}(Q^0_{x,a})$, set $t = B(\gamma)$. Since $A(\gamma) \ge w(B(\gamma))$,

$$A(\gamma) + \lambda^\star(B(\gamma) - \epsilon_{\mathcal{Y}}) \ge w(\epsilon_{\mathcal{Y}}).$$

Taking the infimum over $\gamma \in \mathfrak{C}(Q^0_{x,a})$ yields

$$\inf_{\gamma \in \mathfrak{C}(Q^0_{x,a})} L(\gamma, \lambda^\star) \ge w(\epsilon_{\mathcal{Y}}) = q(x,a).$$

Combining this with weak duality gives

$$q(x,a) = \sup_{\lambda \ge 0} \inf_{\gamma \in \mathfrak{C}(Q^0_{x,a})} \left\{ \int z\, d\gamma + \lambda \left(\int c(y,z)\, d\gamma(y,z) - \epsilon_{\mathcal{Y}} \right) \right\}.$$

It remains to evaluate the inner infimum. For fixed $\lambda \geq 0$, one has

$$\inf_{\gamma \in \mathfrak{C}(Q^0_{x,a})} L(\gamma, \lambda) = -\lambda \epsilon_{\mathcal{Y}} + \inf_{\gamma : \gamma_1 = Q^0_{x,a}} \int \{z + \lambda c(y,z)\}\, d\gamma(y,z).$$

Disintegrate $\gamma(dy, dz) = Q^0_{x,a}(dy) K_y(dz)$, where K_y is a Markov kernel on $\mathcal{Y}$. Then

$$\int \{z + \lambda c(y,z)\}\, d\gamma(y,z) = \int \left[\int \{z + \lambda c(y,z)\} K_y(dz) \right] Q^0_{x,a}(dy).$$

For each fixed y, every probability measure K_y has average at least $\inf_{z \in \mathcal{Y}} \{z + \lambda c(y,z)\}$. Therefore

$$\inf_{\gamma : \gamma_1 = Q^0_{x,a}} \int \{z + \lambda c(y,z)\}\, d\gamma(y,z) \geq \mathbb{E}_{Y \sim Q^0_{x,a}} \left[\inf_{z \in \mathcal{Y}} \{z + \lambda c(Y,z)\} \right].$$

Conversely, define $f_\lambda(y,z) := z + \lambda c(y,z)$. This function is lower semicontinuous, and hence Borel measurable, on the compact metric space $\mathcal{Y} \times \mathcal{Y}$. Its value function $s_\lambda(y) := \min_{z \in \mathcal{Y}} f_\lambda(y,z)$ is lower semicontinuous, and the argmin correspondence has nonempty compact values and a Borel graph. The Arsenin–Kunugui measurable-selection theorem therefore provides a Borel map $z_\lambda : \mathcal{Y} \to \mathcal{Y}$ satisfying

$$z_\lambda(y) \in \arg\min_{z \in \mathcal{Y}} f_\lambda(y,z) \qquad \forall y \in \mathcal{Y}.$$

Taking $K_y = \delta_{z_\lambda(y)}$ gives the reverse inequality, and hence

$$\inf_{\gamma : \gamma_1 = Q^0_{x,a}} \int \{z + \lambda c(y,z)\}\, d\gamma(y,z) = \mathbb{E}_{Y \sim Q^0_{x,a}} \left[\inf_{z \in \mathcal{Y}} \{z + \lambda c(Y,z)\} \right].$$

Thus

$$q(x,a) = \sup_{\lambda \geq 0} \left\{ -\lambda \epsilon_{\mathcal{Y}} + \mathbb{E}_{Y \sim Q^0_{x,a}} \left[\inf_{z \in \mathcal{Y}} \{z + \lambda c(Y,z)\} \right] \right\}.$$

Finally, we restrict the dual domain. For every $y \in \mathcal{Y}$, we have $0 \leq \inf_{z \in \mathcal{Y}} \{z + \lambda c(y,z)\} \leq D$. The lower bound follows from $z \geq 0$ and $c \geq 0$. The upper bound follows by choosing $z = y$ and using $c(y,y) = 0$. Hence the dual objective is at most $-\lambda \epsilon_{\mathcal{Y}} + D$. At $\lambda = 0$, the objective equals 0. If $\lambda > D/\epsilon_{\mathcal{Y}}$, then $-\lambda \epsilon_{\mathcal{Y}} + D < 0$, so such λ cannot improve over $\lambda = 0$. Therefore the supremum may be restricted to $[0, D/\epsilon_{\mathcal{Y}}]$. □

A.2 Reward dual integrand ordering and regularity (Lemma 3)

Lemma 3 (Reward dual integrand ordering and regularity)**.** *For $\lambda \in [0, D/\epsilon_{\mathcal{Y}}]$ and $y \in \mathcal{Y}$, let $s_\lambda(y) := \inf_{z \in \mathcal{Y}} \{z + \lambda c_{\mathcal{Y}}(y,z)\}$. Then:*

1. *For every $y \in \mathcal{Y}$ and $0 \leq \lambda \leq \lambda' \leq D/\epsilon_{\mathcal{Y}}$,*
$$0 \leq s_\lambda(y) \leq s_{\lambda'}(y) \leq y \leq D.$$
2. *For every $y \in \mathcal{Y}$ and $\lambda, \lambda' \in [0, D/\epsilon_{\mathcal{Y}}]$,*
$$|s_\lambda(y) - s_{\lambda'}(y)| \leq C_{\mathcal{Y}} |\lambda - \lambda'|.$$
3. *For every λ, $y \mapsto s_\lambda(y)$ is Borel measurable, and the class $\{s_\lambda : \lambda \in [0, D/\epsilon_{\mathcal{Y}}]\}$ is pointwise separable by rational values of λ.*

Proof. Fix $y \in \mathcal{Y}$. The bounds $0 \leq s_\lambda(y) \leq y \leq D$ follow from $z \in [0, D]$, $c_{\mathcal{Y}} \geq 0$, and the admissible choice $z = y$. If $\lambda \leq \lambda'$, then for every $z \in \mathcal{Y}$,

$$z + \lambda c_{\mathcal{Y}}(y,z) \leq z + \lambda' c_{\mathcal{Y}}(y,z),$$

because the cost is nonnegative. Taking the infimum over z on both sides gives $s_\lambda(y) \leq s_{\lambda'}(y)$. This ordering uses no structural property of the cost beyond its standing nonnegativity assumption.

For Lipschitzness, define $f_\lambda(z) := z + \lambda c_{\mathcal{Y}}(y,z)$. For every $z \in \mathcal{Y}$, one has

$$|f_\lambda(z) - f_{\lambda'}(z)| = |\lambda - \lambda'| c_{\mathcal{Y}}(y,z) \leq C_{\mathcal{Y}} |\lambda - \lambda'|.$$

Thus

$$f_\lambda(z) \le f_{\lambda'}(z) + C_{\mathcal{Y}}|\lambda - \lambda'| \qquad \forall z \in \mathcal{Y}.$$

By monotonicity of the infimum, we have

$$\begin{aligned} s_\lambda(y) &= \inf_{z\in\mathcal{Y}} f_\lambda(z) \\ &\le \inf_{z\in\mathcal{Y}} \{f_{\lambda'}(z) + C_{\mathcal{Y}}|\lambda - \lambda'|\} \\ &= s_{\lambda'}(y) + C_{\mathcal{Y}}|\lambda - \lambda'|. \end{aligned}$$

Interchanging λ and λ' gives

$$|s_\lambda(y) - s_{\lambda'}(y)| \le C_{\mathcal{Y}}|\lambda - \lambda'|.$$

This proves the Lipschitz inequality and, in particular, continuity in λ.

For fixed λ, the map $(y, z) \mapsto z + \lambda c_{\mathcal{Y}}(y, z)$ is lower semicontinuous on $\mathcal{Y} \times \mathcal{Y}$. Since the minimization is over compact $\mathcal{Y}$, the resulting value function $y \mapsto s_\lambda(y)$ is lower semicontinuous and hence Borel measurable. Finally, continuity in λ implies that, for every y and every λ, $s_{\lambda_k}(y) \to s_\lambda(y)$ along any rational sequence $\lambda_k \to \lambda$. Thus the countable rationally indexed subclass is pointwise dense, which supplies the claimed separability. □

A.3 Context-level OT duality (Lemma 2)

Proof. It suffices to prove the following generic identity for an arbitrary center $R \in \Delta(\mathcal{X})$ and an arbitrary real-valued function $h : \mathcal{X} \to \mathbb{R}$. We prove

$$\inf_{P\in\Delta(\mathcal{X}):\mathcal{T}_{\mathcal{X}}(R,P)\le\epsilon_{\mathcal{X}}} \mathbb{E}_{Z\sim P}[h(Z)] = \sup_{\eta\ge 0} \left\{ -\eta\epsilon_{\mathcal{X}} + \mathbb{E}_{X\sim R}\left[\min_{z\in\mathcal{X}}\{h(z) + \eta c_{\mathcal{X}}(X, z)\}\right]\right\}.$$

Because $\mathcal{X}$ is finite, write the optimization directly in coupling variables. Let γ_{xz} be the mass transported from x to z. The first marginal constraint is

$$\sum_{z\in\mathcal{X}} \gamma_{xz} = R(x), \qquad \forall x \in \mathcal{X}.$$

The second marginal is free. Therefore the primal problem is

$$\begin{aligned} \min_{\gamma\ge 0} \quad & \sum_{x,z\in\mathcal{X}} \gamma_{xz} h(z) \\ \text{subject to} \quad & \sum_{z\in\mathcal{X}} \gamma_{xz} = R(x), \qquad \forall x \in \mathcal{X}, \\ & \sum_{x,z\in\mathcal{X}} \gamma_{xz} c_{\mathcal{X}}(x, z) \le \epsilon_{\mathcal{X}}. \end{aligned}$$

The feasible set is nonempty because the diagonal coupling $\gamma_{xx} = R(x)$ has transport cost zero.

Define

$$A_X(\gamma) := \sum_{x,z} \gamma_{xz} h(z), \qquad B_X(\gamma) := \sum_{x,z} \gamma_{xz} c_{\mathcal{X}}(x, z).$$

For $\eta \ge 0$, define

$$L_X(\gamma, \eta) := A_X(\gamma) + \eta(B_X(\gamma) - \epsilon_{\mathcal{X}}).$$

For every feasible γ, we have $L_X(\gamma, \eta) \le A_X(\gamma)$, which gives weak duality.

To prove equality, define

$$w_X(t) := \min_{\gamma\ge 0:\sum_z \gamma_{xz}=R(x),\ B_X(\gamma)\le t} A_X(\gamma), \qquad t \in [0, C_{\mathcal{X}}].$$

The diagonal coupling makes $w_X(t)$ finite for all $t \in [0, C_{\mathcal{X}}]$. The function w_X is nonincreasing and convex by the same convex-combination argument used in the reward-level proof. If $\epsilon_{\mathcal{X}} \ge C_{\mathcal{X}}$, the transport constraint is inactive and $\eta = 0$ gives equality. Otherwise $0 < \epsilon_{\mathcal{X}} < C_{\mathcal{X}}$; choose

$s \in \partial w_X(\epsilon_{\mathcal{X}})$ with $s \le 0$ and set $\eta^\star = -s \ge 0$. The subgradient inequality implies, for every row-coupling γ with first marginal R, we have

$$A_X(\gamma) + \eta^\star(B_X(\gamma) - \epsilon_{\mathcal{X}}) \ge w_X(\epsilon_{\mathcal{X}}).$$

Taking the infimum over all row-couplings and combining with weak duality proves strong duality:

$$w_X(\epsilon_{\mathcal{X}}) = \sup_{\eta \ge 0} \inf_{\gamma : \sum_z \gamma_{xz} = R(x)} L_X(\gamma, \eta).$$

It remains to evaluate the inner infimum. For fixed $\eta \ge 0$,

$$\inf_{\gamma : \sum_z \gamma_{xz} = R(x)} L_X(\gamma, \eta) = -\eta\epsilon_{\mathcal{X}} + \inf_{\gamma : \sum_z \gamma_{xz} = R(x)} \sum_{x,z} \gamma_{xz}\{h(z) + \eta c_{\mathcal{X}}(x, z)\}.$$

The row constraints separate across x. For each fixed x, the row has mass $R(x)$ and can put all of that mass on a minimizer of $z \mapsto h(z) + \eta c_{\mathcal{X}}(x, z)$. Since $\mathcal{X}$ is finite, the minimum exists. Therefore

$$\inf_{\gamma : \sum_z \gamma_{xz} = R(x)} L_X(\gamma, \eta) = -\eta\epsilon_{\mathcal{X}} + \sum_{x \in \mathcal{X}} R(x) \min_{z \in \mathcal{X}} \{h(z) + \eta c_{\mathcal{X}}(x, z)\}.$$

This is the stated dual formula.

It remains to justify the bounded dual intervals used in the manuscript. Suppose first that $h : \mathcal{X} \to [0, D]$. For every $x \in \mathcal{X}$,

$$0 \le \min_{z \in \mathcal{X}} \{h(z) + \eta c_{\mathcal{X}}(x, z)\} \le h(x) + \eta c_{\mathcal{X}}(x, x) = h(x) \le D.$$

Hence the dual objective is at most $-\eta\epsilon_{\mathcal{X}} + D$, while its value at $\eta = 0$ is $\min_z h(z) \ge 0$. No $\eta > D/\epsilon_{\mathcal{X}}$ can improve over $\eta = 0$. Applying this with $R = P^0$ and $h(z) = \sum_a \pi_z(a) q(z, a)$ proves Lemma 2.

For the empirical pessimistic value, clipping gives $h_{\pi, \underline{q}} : \mathcal{X} \to [0, D]$. The same argument therefore shows that no $\eta > D/\epsilon_{\mathcal{X}}$ can improve over $\eta = 0$, proving that the interval in (3) is valid. □

A.4 Monotonicity and Lipschitzness of the context robust value (Lemma 4)

Lemma 4 (Monotonicity and Lipschitzness of the context robust value)**.** *For a center $R \in \Delta(\mathcal{X})$ and a function $f : \mathcal{X} \to \mathbb{R}$, define*

$$\Phi_R(f) := \inf_{P : \mathcal{T}_{\mathcal{X}}(R, P) \le \epsilon_{\mathcal{X}}} \mathbb{E}_{X \sim P}[f(X)].$$

Let $h, g : \mathcal{X} \to \mathbb{R}$. Then:

1. *If $h \le g$, then $\Phi_R(h) \le \Phi_R(g)$.*
2. *$|\Phi_R(h) - \Phi_R(g)| \le \|h - g\|_\infty$.*
3. *If $h \le g$, then*

$$0 \le \Phi_R(g) - \Phi_R(h) \le \sup_{P : \mathcal{T}_{\mathcal{X}}(R, P) \le \epsilon_{\mathcal{X}}} \mathbb{E}_{X \sim P}[g(X) - h(X)].$$

Proof. Let

$$\mathcal{U}_X(R) := \{P \in \Delta(\mathcal{X}) : \mathcal{T}_{\mathcal{X}}(R, P) \le \epsilon_{\mathcal{X}}\}.$$

This feasible set is identical for h and g.

If $h(x) \le g(x)$ for every x, then for every $P \in \mathcal{U}_X(R)$,

$$\mathbb{E}_P[h(X)] \le \mathbb{E}_P[g(X)].$$

Taking the infimum over the same feasible set gives the monotonicity inequality.

For Lipschitzness, let $r = \|h - g\|_\infty$. Then $h \le g + r$ and $g \le h + r$. By monotonicity and by the fact that adding a constant shifts every feasible expectation by that constant,

$$\inf_{P \in \mathcal{U}_X(R)} \mathbb{E}_P[h(X)] \le \inf_{P \in \mathcal{U}_X(R)} \mathbb{E}_P[g(X)] + r,$$

and symmetrically,

$$\inf_{P\in\mathcal{U}_X(R)} \mathbb{E}_P[g(X)] \le \inf_{P\in\mathcal{U}_X(R)} \mathbb{E}_P[h(X)] + r.$$

Combining the two inequalities gives the Lipschitz bound.

For the sharper one-sided bound, assume $h \le g$ and set $\Delta(x) = g(x) - h(x) \ge 0$. Monotonicity gives the lower bound. Since $\mathcal{X}$ is finite, the infimum over $\mathcal{U}_X(R)$ is attained. Let P_h be a minimizer for h. Then

$$\begin{aligned}\inf_{P\in\mathcal{U}_X(R)} \mathbb{E}_P[g(X)] - \inf_{P\in\mathcal{U}_X(R)} \mathbb{E}_P[h(X)] &\le \mathbb{E}_{P_h}[g(X)] - \mathbb{E}_{P_h}[h(X)] \\ &= \mathbb{E}_{P_h}[\Delta(X)] \\ &\le \sup_{P\in\mathcal{U}_X(R)} \mathbb{E}_P[\Delta(X)].\end{aligned}$$

This proves the sharper bound. □

B FINITE-SAMPLE CONFIDENCE BOUNDS

The confidence analysis has three parts. First, pointwise ordering of the reward dual integrands yields a cellwise bound without discretizing the dual parameter or introducing reward-support-cardinality dependence. Second, a total-variation bound controls the effect of replacing the nominal context distribution by its empirical estimate, uniformly over policies and bounded reward tables. Combining these results gives a lower confidence bound for every policy's robust value.

B.1 REWARD-CELL CONCENTRATION (PROPOSITION 1)

The key step is to control the reward dual uniformly over its continuous parameter λ without splitting λ into a finite grid. Pointwise ordering of the dual integrands permits a layer-cake representation and a Dvoretzky–Kiefer–Wolfowitz bound; bounded differences then turn the resulting mean bound into a high-probability bound.

Proof of Proposition 1. Conditional on the context–action samples $\mathcal{F}_{XA} := \sigma\big((X_i, A_i)_{i=1}^n\big)$, every count $N_{x,a}$ is fixed. Also, within each nonempty cell (x, a), the observed rewards are independent draws from $Q^0_{x,a}$.

Recall the reward dual integrand $s_\lambda(y) := \inf_{z\in\mathcal{Y}}\{z + \lambda c_{\mathcal{Y}}(y, z)\}$, with $\lambda \in [0, D/\epsilon_{\mathcal{Y}}]$ and $y \in \mathcal{Y}$. Lemma 3 shows that these integrands are Borel measurable in y, bounded in $[0, D]$, continuous in λ, and nondecreasing in λ pointwise in y. These are the only properties of the integrands used below.

We now prove concentration directly for these integrands. Fix a nonempty cell (x, a), let $m = N_{x,a}$, and relabel its rewards as $Y_1, \dots, Y_m$. In this subsection, write

$$P_m f := \frac{1}{m}\sum_{i=1}^m f(Y_i), \qquad Pf := \mathbb{E}_{Y\sim Q^0_{x,a}}[f(Y)], \qquad \Delta_m := \sup_{0\le\lambda\le D/\epsilon_{\mathcal{Y}}} |(P_m - P)s_\lambda|.$$

Step 1: reduce the supremum to countably many values of λ.

Let $\Lambda_{\mathbb{Q}} := ([0, D/\epsilon_{\mathcal{Y}}] \cap \mathbb{Q}) \cup \{D/\epsilon_{\mathcal{Y}}\}$ be a countable dense subset of the dual interval $[0, D/\epsilon_{\mathcal{Y}}]$. Pointwise continuity of $s_\lambda(y)$ implies that $P_m s_\lambda$ is continuous in λ. The bound $0 \le s_\lambda \le D$ and dominated convergence imply that Ps_λ is also continuous in λ. Hence, Δ_m can be expressed as:

$$\Delta_m = \sup_{\lambda\in\Lambda_{\mathbb{Q}}} |(P_m - P)s_\lambda|.$$

Each term on the right hand side is measurable, so Δ_m is measurable.

Step 2: bound the mean of Δ_m by layer cake and DKW.

For every $v \in [0, D]$, we can write v as a layer-cake identity: $v = \int_0^D \mathbf{1}\{r < v\}\, dr$. Applying this identity to s_λ yields

$$(P_m - P)s_\lambda = \int_0^D (P_m - P)\mathbf{1}\{r < s_\lambda\}\, dr.$$

Taking absolute values and applying the triangle inequality, for each $\lambda \in \Lambda_{\mathbb{Q}}$ we have

$$|(P_m - P)s_\lambda| \le \int_0^D |(P_m - P)\mathbf{1}\{r < s_\lambda\}|\, dr.$$

Taking the supremum over λ and moving it inside the integral gives

$$\Delta_m \le \int_0^D \sup_{\lambda\in\Lambda_{\mathbb{Q}}} |(P_m - P)\mathbf{1}\{r < s_\lambda\}|\, dr. \tag{5}$$

Fix $r \in [0, D]$ and enumerate $\Lambda_{\mathbb{Q}}$ as $\lambda_1, \lambda_2, \dots$. Because $s_\lambda(y)$ is nondecreasing in λ, the sets $\{y : r < s_{\lambda_j}(y)\}$ are nested. We encode this family by the measurable function

$$H_r(y) := \sum_{j\ge1} 2^{-j}\mathbf{1}\{r < s_{\lambda_j}(y)\}.$$

For a fixed j, suppose $r < s_{\lambda_j}(y)$ but $r \geq s_{\lambda_j}(y')$. Nestedness implies that, for every k, we have $\mathbf{1}\{r < s_{\lambda_k}(y)\} \geq \mathbf{1}\{r < s_{\lambda_k}(y')\}$. Consequently,

$$H_r(y) \geq H_r(y') + 2^{-j}.$$

Hence every y satisfying $r < s_{\lambda_j}(y)$ has an H_r value at least 2^{-j} larger than every y' that does not satisfy it. We can therefore choose $c_j \in \mathbb{R}$ between these two ranges so that

$$\{y : r < s_{\lambda_j}(y)\} = \{y : H_r(y) > c_j\}.$$

If the set on the left is empty or equals $\mathcal{Y}$, choose c_j above or below the range of H_r, respectively.

Conditional on $\mathcal{F}_{XA}$, the variables $H_r(Y_1), \ldots, H_r(Y_m)$ are i.i.d. Each event $r < s_{\lambda_j}(Y)$ can be written as $H_r(Y) > c_j$ for some c_j. The Dvoretzky–Kiefer–Wolfowitz–Massart inequality controls the empirical probabilities of all such thresholds simultaneously. Thus, for every $t > 0$,

$$\mathbb{P}\left(\sup_{\lambda \in \Lambda_{\mathbb{Q}}} |(P_m - P)\mathbf{1}\{r < s_\lambda\}| > t \;\middle|\; \mathcal{F}_{XA} \right) \leq 2e^{-2mt^2}.$$

We use this tail bound to control the expected supremum at the fixed level r. Since the expectation of a nonnegative random variable equals the integral of its tail probabilities,

$$\mathbb{E}\left[\sup_{\lambda \in \Lambda_{\mathbb{Q}}} |(P_m - P)\mathbf{1}\{r < s_\lambda\}| \;\middle|\; \mathcal{F}_{XA} \right] \leq \int_0^\infty 2e^{-2mt^2}\, dt = \sqrt{\frac{\pi}{2m}}.$$

This bound holds for every $r \in [0, D]$, so we can integrate it over r. Each term in the supremum is jointly measurable in the rewards and r, and the supremum is over a countable set. The integrand in (5) is therefore measurable and nonnegative. Taking conditional expectations and applying Tonelli's theorem gives

$$\mathbb{E}[\Delta_m \mid \mathcal{F}_{XA}] \leq \int_0^D \mathbb{E}\left[\sup_{\lambda \in \Lambda_{\mathbb{Q}}} |(P_m - P)\mathbf{1}\{r < s_\lambda\}| \;\middle|\; \mathcal{F}_{XA} \right] dr \leq D\sqrt{\frac{\pi}{2m}}. \tag{6}$$

Step 3: turn the mean bound into a high-probability bound.

Replace one observation Y_i by another reward, and write P'_m for the resulting empirical average. Since $0 \leq s_\lambda \leq D$, for every λ, $|P_m s_\lambda - P'_m s_\lambda| \leq \frac{D}{m}$. Let $\Delta'_m := \sup_{\lambda \in \Lambda_{\mathbb{Q}}} |(P'_m - P)s_\lambda|$. The triangle inequality gives $|\Delta_m - \Delta'_m| \leq D/m$.

Conditional on $\mathcal{F}_{XA}$, the rewards are independent. McDiarmid's inequality consequently yields

$$\mathbb{P}\left(\Delta_m - \mathbb{E}[\Delta_m \mid \mathcal{F}_{XA}] > t \;\middle|\; \mathcal{F}_{XA} \right) \leq \exp\left(-\frac{2mt^2}{D^2} \right).$$

For $u \in (0, 1)$, take $t = D\sqrt{\log(1/u)/(2m)}$ and use (6). With conditional probability at least $1 - u$,

$$\Delta_m \leq D\left\{ \sqrt{\frac{\pi}{2m}} + \sqrt{\frac{\log(1/u)}{2m}} \right\}. \tag{7}$$

Step 4: make the reward bound simultaneous over cells.

For a nonempty cell, the empirical and population dual objectives contain the same term $-\lambda\epsilon_{\mathcal{Y}}$. Their dual representations and the inequality $|\sup F - \sup G| \leq \sup |F - G|$ give

$$\begin{aligned} |\widehat{q}(x,a) - q(x,a)| &= \left| \sup_{\lambda \in [0, D/\epsilon_{\mathcal{Y}}]} \{P_m s_\lambda - \lambda\epsilon_{\mathcal{Y}}\} - \sup_{\lambda \in [0, D/\epsilon_{\mathcal{Y}}]} \{P s_\lambda - \lambda\epsilon_{\mathcal{Y}}\} \right| \\ &\leq \sup_{\lambda \in [0, D/\epsilon_{\mathcal{Y}}]} |(P_m - P)s_\lambda| = \Delta_m. \end{aligned}$$

Apply (7) in each nonempty cell with $m = N_{x,a}$ and $u = \delta/(2|\mathcal{X}||\mathcal{A}|)$. With conditional probability at least $1 - \delta/(2|\mathcal{X}||\mathcal{A}|)$,

$$|\widehat{q}(x,a) - q(x,a)| \leq D\left\{ \sqrt{\frac{\pi}{2N_{x,a}}} + \sqrt{\frac{\log(2|\mathcal{X}||\mathcal{A}|/\delta)}{2N_{x,a}}} \right\}.$$

Both robust rewards lie in $[0, D]$, so their difference is also at most D. Taking the smaller bound gives $|\widehat{q}(x,a) - q(x,a)| \le \beta_Y(N_{x,a}; \delta)$. For an empty cell, $\widehat{q}(x,a) = 0$ and $q(x,a) \in [0, D]$, so the same bound holds deterministically because $\beta_Y(0; \delta) = D$.

There are at most $|\mathcal{X}||\mathcal{A}|$ nonempty cells. A union bound gives conditional probability at least $1 - \delta/2$ that the bound holds in every cell. Averaging over the context–action sample gives the same unconditional probability.

On this simultaneous event, fix any cell and write $\beta = \beta_Y(N_{x,a}; \delta)$. Since $\widehat{q}(x,a) - \beta \le q(x,a)$ and $q(x,a) \ge 0$, the clipped lower confidence bound satisfies

$$0 \le \underline{q}(x,a) = \max\{0, \widehat{q}(x,a) - \beta\} \le q(x,a).$$

Also, $\underline{q}(x,a) \ge \widehat{q}(x,a) - \beta$, so

$$0 \le q(x,a) - \underline{q}(x,a) \le q(x,a) - \widehat{q}(x,a) + \beta \le 2\beta.$$

Finally, $\widehat{q}(x,a) \in [0, D]$ and $\beta \ge 0$ imply $\underline{q}(x,a) \in [0, D]$ for every sample, with $\underline{q}(x,a) = 0$ in an empty cell. This proves all assertions. □

B.2 Total-variation context concentration (Lemma 5)

Lemma 5 (Total-variation context concentration)**.** *For any reward table $r : \mathcal{X} \times \mathcal{A} \to [0, D]$, define*

$$V_r(\pi) := \inf_{P : \mathcal{T}_{\mathcal{X}}(P^0, P) \le \epsilon_{\mathcal{X}}} \mathbb{E}_{X \sim P}\left[\sum_a \pi_X(a) r(X, a)\right]$$

and

$$\widehat{V}_r(\pi) := \inf_{P : \mathcal{T}_{\mathcal{X}}(\widehat{P}, P) \le \epsilon_{\mathcal{X}}} \mathbb{E}_{X \sim P}\left[\sum_a \pi_X(a) r(X, a)\right].$$

Deterministically,

$$\left|\widehat{V}_r(\pi) - V_r(\pi)\right| \le \frac{D}{2}\|\widehat{P} - P^0\|_1 \qquad \forall \pi \in \Pi, \quad \forall r : \mathcal{X} \times \mathcal{A} \to [0, D].$$

With probability at least $1 - \delta/2$,

$$\left|\widehat{V}_r(\pi) - V_r(\pi)\right| \le \beta_X(\delta) \qquad \forall \pi \in \Pi, \quad \forall r : \mathcal{X} \times \mathcal{A} \to [0, D].$$

The high-probability event is the same for all r and π. Hence both bounds also apply when the reward table and policy are chosen using the logged sample.

Proof. For every $x \in \mathcal{X}$, Jensen's inequality gives

$$\mathbb{E}\left[|\widehat{P}(x) - P^0(x)|\right] \le \sqrt{\frac{P^0(x)(1 - P^0(x))}{n}}.$$

Consequently,

$$\mathbb{E}\left[\|\widehat{P} - P^0\|_1\right] \le \frac{1}{\sqrt{n}} \sum_{x \in \mathcal{X}} \sqrt{P^0(x)(1 - P^0(x))}.$$

Replacing one context observation changes $\|\widehat{P} - P^0\|_1$ by at most $2/n$. McDiarmid's inequality therefore implies that, with probability at least $1 - \delta/2$,

$$\|\widehat{P} - P^0\|_1 \le \frac{1}{\sqrt{n}}\left[\sum_{x \in \mathcal{X}} \sqrt{P^0(x)(1 - P^0(x))} + \sqrt{2\log(2/\delta)}\right]. \tag{8}$$

Fix a policy $\pi \in \Pi$ and a table $r : \mathcal{X} \times \mathcal{A} \to [0, D]$. Define

$$h_{\pi,r}(z) := \sum_{a \in \mathcal{A}} \pi_z(a) r(z, a).$$

Because π_z is a probability distribution, we have $h_{\pi,r}(z) \in [0, D]$. The context dual may therefore be restricted to $\eta \in [0, D/\epsilon_{\mathcal{X}}]$. For η in this interval, define

$$g_{\eta,\pi,r}(x) := \min_{z \in \mathcal{X}} \{h_{\pi,r}(z) + \eta c_{\mathcal{X}}(x, z)\}.$$

Then $0 \le g_{\eta,\pi,r}(x) \le D$ for every $x \in \mathcal{X}$. The lower bound follows from $c_{\mathcal{X}} \ge 0$, while the upper bound follows by choosing $z = x$ and using $c_{\mathcal{X}}(x, x) = 0$. Hence, for every η,

$$\left|\mathbb{E}_{\widehat{P}}[g_{\eta,\pi,r}(X)] - \mathbb{E}_{P^0}[g_{\eta,\pi,r}(X)]\right| \le \frac{D}{2}\|\widehat{P} - P^0\|_1.$$

The dual formula for the context robust value gives

$$\widehat{V}_r(\pi) = \sup_{\eta \in [0, D/\epsilon_{\mathcal{X}}]} \left\{-\eta\epsilon_{\mathcal{X}} + \mathbb{E}_{\widehat{P}}[g_{\eta,\pi,r}(X)]\right\},$$
$$V_r(\pi) = \sup_{\eta \in [0, D/\epsilon_{\mathcal{X}}]} \left\{-\eta\epsilon_{\mathcal{X}} + \mathbb{E}_{P^0}[g_{\eta,\pi,r}(X)]\right\}.$$

Using $|\sup f - \sup g| \le \sup |f - g|$, we obtain, without invoking any concentration event,

$$\left|\widehat{V}_r(\pi) - V_r(\pi)\right| \le \frac{D}{2}\|\widehat{P} - P^0\|_1.$$

Moreover, both $\widehat{V}_r(\pi)$ and $V_r(\pi)$ lie in $[0, D]$, so

$$\left|\widehat{V}_r(\pi) - V_r(\pi)\right| \le D.$$

Combining these two deterministic bounds with (8) yields, with probability at least $1 - \delta/2$,

$$\left|\widehat{V}_r(\pi) - V_r(\pi)\right| \le D \min\left\{1, \frac{1}{2\sqrt{n}}\left[\sum_{x \in \mathcal{X}} \sqrt{P^0(x)(1 - P^0(x))} + \sqrt{2\log(2/\delta)}\right]\right\} = \beta_X(\delta).$$

The concentration event does not depend on r, π, or η. Therefore, the bound holds simultaneously over all stated reward tables and policies, including data-dependent reward tables. □

B.3 Policy-value lower confidence bound (Proposition 2)

Proof. Let $\mathcal{E}_Y$ be the event of Proposition 1 and let $\mathcal{E}_X$ be the event of Lemma 5. They have probabilities at least $1 - \delta/2$ each, so

$$\mathbb{P}(\mathcal{E}_Y \cap \mathcal{E}_X) \ge 1 - \delta.$$

Work on this intersection.

For every (x, a), Proposition 1 gives $0 \le \underline{q}(x, a) \le q(x, a)$, so $\underline{q} : \mathcal{X} \times \mathcal{A} \to [0, D]$. Therefore, for every policy π and context x,

$$h_{\pi,\underline{q}}(x) := \sum_a \pi_x(a)\underline{q}(x, a) \le \sum_a \pi_x(a)q(x, a) =: h_{\pi,q}(x).$$

By Lemma 5 applied to $r = \underline{q}$ and then by monotonicity from Lemma 4,

$$\widehat{V}_{\underline{q}}(\pi) - \beta_X(\delta) \le V_{\underline{q}}(\pi) \le V_q(\pi) = V(\pi).$$

The left-hand side is exactly $\underline{V}(\pi; \delta)$, hence $\underline{V}(\pi; \delta) \le V(\pi)$ for every $\pi \in \Pi$.

For the gap, write

$$h_\pi(x) := h_{\pi,q}(x), \qquad \underline{h}_\pi(x) := h_{\pi,\underline{q}}(x).$$

Then

$$V(\pi) - \underline{V}(\pi; \delta) = V_q(\pi) - \widehat{V}(\pi) + \beta_X(\delta)$$
$$= \{V_q(\pi) - V_{\underline{q}}(\pi)\} + \{V_{\underline{q}}(\pi) - \widehat{V}(\pi)\} + \beta_X(\delta).$$

Lemma 5 bounds the second brace by

$$\beta_X(\delta).$$

For the first brace, $\underline{h}_\pi \le h_\pi$, so the sharper part of Lemma 4 gives

$$V_q(\pi) - V_{\underline{q}}(\pi) \le \sup_{P:\mathcal{T}_{\mathcal{X}}(P^0,P)\le\epsilon_{\mathcal{X}}} \mathbb{E}_{X\sim P}[h_\pi(X) - \underline{h}_\pi(X)]$$
$$= \sup_{P:\mathcal{T}_{\mathcal{X}}(P^0,P)\le\epsilon_{\mathcal{X}}} \mathbb{E}_{X\sim P}\left[\sum_a \pi_X(a)\{q(X,a) - \underline{q}(X,a)\}\right].$$

Proposition 1 gives $q(x,a) - \underline{q}(x,a) \le 2\beta_Y(N_{x,a};\delta)$ for all cells. Substituting this bound gives

$$V(\pi) - \underline{V}(\pi;\delta) \le 2 \sup_{P:\mathcal{T}_{\mathcal{X}}(P^0,P)\le\epsilon_{\mathcal{X}}} \mathbb{E}_{X\sim P}\left[\sum_a \pi_X(a)\beta_Y(N_{X,a};\delta)\right] + 2\beta_X(\delta).$$

□

C Policy-Learning Guarantees

We convert the confidence bound involving random cell counts into an explicit regret rate. A binomial concentration argument bounds each cell's reward penalty in terms of its nominal sampling probability, including cells with small expected counts. We then aggregate these bounds along an optimal robust policy using $C^{\star}_{\mathrm{rob}}$ and absorb the context-estimation term into the same coverage coefficient.

For (x, a), define the nominal cell probability

$$p^0_{x,a} := P^0(x)\pi^0_x(a).$$

C.1 Count-to-confidence conversion (Lemma 6)

Lemma 6 (Count-to-confidence conversion)**.** *Fix $\delta \in (0, 1)$. With probability at least $1 - \delta/2$, simultaneously for every (x, a) with $p^0_{x,a} > 0$,*

$$\beta_Y(N_{x,a}; \delta/2) \le \sqrt{\frac{2}{1 - \log 2}} \frac{D}{\sqrt{np^0_{x,a}}} \sqrt{\log\left(\frac{4|\mathcal{X}||\mathcal{A}|}{\delta}\right)}.$$

Moreover, on the same event,

$$N_{x,a} \ge \frac{1}{2} np^0_{x,a} \qquad \textit{for every } (x, a) \textit{ such that} \qquad np^0_{x,a} \ge \frac{2}{1 - \log 2} \log\left(\frac{2|\mathcal{X}||\mathcal{A}|}{\delta}\right).$$

Proof. For a fixed cell (x, a), each logged sample falls in this cell with probability

$$p^0_{x,a} = P^0(x)\pi^0_x(a),$$

independently across samples. Hence

$$N_{x,a} \sim \mathrm{Binomial}(n, p^0_{x,a}).$$

The multiplicative Chernoff bound gives, for a binomial variable N with mean μ,

$$\mathbb{P}(N \le (1 - t)\mu) \le \left(\frac{e^{-t}}{(1 - t)^{1-t}}\right)^{\mu}, \qquad 0 < t < 1.$$

Taking $t = 1/2$ gives

$$\mathbb{P}\left(N_{x,a} < \frac{1}{2} np^0_{x,a}\right) \le \exp\left(-\frac{1 - \log 2}{2} np^0_{x,a}\right).$$

For every cell satisfying the condition in the lemma,

$$\exp\left(-\frac{1 - \log 2}{2} np^0_{x,a}\right) \le \frac{\delta}{2|\mathcal{X}||\mathcal{A}|}.$$

A union bound over the at most $|\mathcal{X}||\mathcal{A}|$ cells shows that the displayed count lower bound holds simultaneously with probability at least $1 - \delta/2$. Work on this event and fix a cell with $p^0_{x,a} > 0$.

If

$$np^0_{x,a} \ge \frac{2}{1 - \log 2} \log\left(\frac{2|\mathcal{X}||\mathcal{A}|}{\delta}\right),$$

then the count lower bound and the definition of β_Y give

$$\begin{aligned}
\beta_Y(N_{x,a}; \delta/2) &\le \frac{D}{\sqrt{np^0_{x,a}}} \left\{\sqrt{\pi} + \sqrt{\log\left(\frac{4|\mathcal{X}||\mathcal{A}|}{\delta}\right)}\right\} \\
&\le \sqrt{\frac{2}{1 - \log 2}} \frac{D}{\sqrt{np^0_{x,a}}} \sqrt{\log\left(\frac{4|\mathcal{X}||\mathcal{A}|}{\delta}\right)}.
\end{aligned}$$

The last step uses $\log(4|\mathcal{X}||\mathcal{A}|/\delta) \geq \log 4$ and the elementary inequality

$$\sqrt{\pi} + \sqrt{u} \leq \sqrt{\frac{2}{1-\log 2}}\sqrt{u}, \qquad u \geq \log 4.$$

If instead

$$np^0_{x,a} < \frac{2}{1-\log 2} \log\left(\frac{2|\mathcal{X}||\mathcal{A}|}{\delta}\right),$$

then the truncation $\beta_Y \leq D$ gives

$$\begin{aligned}\beta_Y(N_{x,a};\delta/2) &\leq D \\ &\leq \sqrt{\frac{2}{1-\log 2}}\,\frac{D}{\sqrt{np^0_{x,a}}}\sqrt{\log\left(\frac{4|\mathcal{X}||\mathcal{A}|}{\delta}\right)}.\end{aligned}$$

Thus the count-to-confidence bound holds for every supported cell. □

C.2 Regret under robust optimal-policy coverage (Theorem 1)

Proof. Lemma 6 gives an event of probability at least $1-\delta/2$. Proposition 2, applied with confidence parameter $\delta/2$, gives another event of probability at least $1-\delta/2$. On this event, optimality of $\widehat{\pi}_{\delta/2}$ and the lower-confidence property imply

$$\begin{aligned}V(\pi^\star) - V(\widehat{\pi}_{\delta/2}) &\leq V(\pi^\star) - \underline{V}(\widehat{\pi}_{\delta/2};\delta/2) \\ &\leq V(\pi^\star) - \underline{V}(\pi^\star;\delta/2) \\ &\leq 2 \sup_{P:\mathcal{T}_{\mathcal{X}}(P^0,P)\leq\epsilon_{\mathcal{X}}} \mathbb{E}_{X\sim P}\left[\sum_a \pi^\star_X(a)\beta_Y(N_{X,a};\delta/2)\right] + 2\beta_X(\delta/2).\end{aligned}$$

Their intersection has probability at least $1-\delta$. Work on this intersection. Finiteness of $C^\star_{\mathrm{rob}}$ ensures that cells with $\pi^\star_x(a) > 0$ and $p^0_{x,a} = 0$ have zero mass under every admissible context distribution; otherwise the coverage coefficient would be infinite. Therefore

$$\begin{aligned}&\sup_{P:\mathcal{T}_{\mathcal{X}}(P^0,P)\leq\epsilon_{\mathcal{X}}} \mathbb{E}_{X\sim P}\left[\sum_a \pi^\star_X(a)\beta_Y(N_{X,a};\delta/2)\right] \\ &\leq \sqrt{\frac{2}{1-\log 2}}\,D\sqrt{\frac{\log(4|\mathcal{X}||\mathcal{A}|/\delta)}{n}} \sup_{P:\mathcal{T}_{\mathcal{X}}(P^0,P)\leq\epsilon_{\mathcal{X}}} \mathbb{E}_{X\sim P}\left[\sum_a \pi^\star_X(a)(p^0_{X,a})^{-1/2}\right] \\ &= \sqrt{\frac{2}{1-\log 2}}\,DC^\star_{\mathrm{rob}}\sqrt{\frac{\log(4|\mathcal{X}||\mathcal{A}|/\delta)}{n}}.\end{aligned}$$

It remains to absorb context estimation into the same coverage term. Since P^0 itself belongs to the context ambiguity set and $C^\star_{\mathrm{rob}} < \infty$,

$$\begin{aligned}C^\star_{\mathrm{rob}} &\geq \sum_{x:P^0(x)>0} \sqrt{P^0(x)} \sum_{a\in\mathcal{A}} \frac{\pi^\star_x(a)}{\sqrt{\pi^0_x(a)}} \\ &\geq \sum_{x\in\mathcal{X}} \sqrt{P^0(x)} \geq 1.\end{aligned}$$

The second inequality uses $\pi^0_x(a) \leq 1$ on every cell used by $\pi^\star$; finiteness of the coverage coefficient excludes an uncovered such cell whenever $P^0(x) > 0$. Hence

$$\sum_{x\in\mathcal{X}} \sqrt{P^0(x)(1-P^0(x))} \leq C^\star_{\mathrm{rob}}.$$

By the definition of β_X,

$$2\beta_X(\delta/2) \leq \frac{D}{\sqrt{n}}\left[\sum_{x\in\mathcal{X}} \sqrt{P^0(x)(1-P^0(x))} + \sqrt{2\log(4/\delta)}\right].$$

Since $\log(4|\mathcal{X}||\mathcal{A}|/\delta) \geq \log 4 > 1$ and $\log(4/\delta) \leq \log(4|\mathcal{X}||\mathcal{A}|/\delta)$, the context term is at most

$$(1+\sqrt{2})DC^{\star}_{\mathrm{rob}}\sqrt{\frac{\log(4|\mathcal{X}||\mathcal{A}|/\delta)}{n}}.$$

Adding this to the reward-estimation term proves the theorem. □

D Proof of the Minimax Lower Bound (Theorem 2)

This appendix proves Theorem 2. The anchor supplies high-value context mass that the context adversary can move to a hard context, making context shift active and introducing the factor $w = 1/(2r) + \rho$ into both robust regret and coverage. The hard family is indexed jointly by one of the $|\mathcal{X}| - 1$ hard contexts and one of the $|\mathcal{A}| - 1$ competing actions. A Fano argument over these $(|\mathcal{X}| - 1)(|\mathcal{A}| - 1)$ instances supplies the context–action cardinality logarithm, while a binary subfamily supplies the confidence dependence.

D.1 Step 1: Hard instances

Let

$$\sigma := \frac{\epsilon_{\mathcal{Y}}}{D^2} \in (0, 1/2], \qquad r := |\mathcal{X}| - 1, \qquad L := |\mathcal{A}| - 1, \qquad \rho := \frac{\epsilon_{\mathcal{X}}}{\kappa}.$$

Label the context and action sets as

$$\mathcal{X} = \{x_1, \dots, x_r, x_{\mathrm{anc}}\}, \qquad \mathcal{A} = \{0, 1, \dots, L\},$$

where x_{anc} is the anchor context. The nominal context distribution and context transport cost are

$$P^0(x_t) = \frac{1}{2r}, \quad t \in \{1, \dots, r\}, \qquad P^0(x_{\mathrm{anc}}) = \frac{1}{2}, \qquad c_{\mathcal{X}}(x, z) = \kappa \mathbf{1}\{x \neq z\}.$$

Write

$$b := 1 - L\mu.$$

At each hard context, the behavior policy chooses every competing action with probability μ and action 0 with probability b:

$$\pi^0_{x_t}(a) = \mu, \quad a \in \{1, \dots, L\}, \qquad \pi^0_{x_t}(0) = b.$$

At the anchor it always chooses action 0. The condition

$$\mu \leq \frac{w^2}{1 + Lw^2}, \qquad w := \frac{1}{2r} + \rho,$$

implies

$$\mu \leq w^2 b < b, \tag{9}$$

so the behavior policy is well defined.

The policy class is

$$\Pi = \{\pi_{t,j} : t \in \{1, \dots, r\},\ j \in \{1, \dots, L\}\},$$

where

$$\pi_{t,j}(x_t) = j, \qquad \pi_{t,j}(x) = 0 \quad \text{for } x \neq x_t.$$

Thus every policy chooses one competing action at one hard context and chooses action 0 everywhere else.

For $\theta \in [0, 1]$, let

$$Q_\theta = (1 - \theta)\delta_0 + \theta\delta_D, \qquad \theta_0 := \frac{1 + \sigma}{2}.$$

Fix

$$0 < h \leq \frac{1 - \sigma}{4}.$$

For every $(t, j) \in \{1, \dots, r\} \times \{1, \dots, L\}$, define an instance $\mathsf{M}_{t,j}$ as follows. At every hard context, action 0 has the common reward law

$$Q^0_{x_s,0} = Q_{\theta_0}, \qquad s \in \{1, \dots, r\}.$$

The single elevated competing cell has reward law

$$Q^0_{x_t,j} = Q_{\theta_0 + h},$$

and every other competing cell has reward law

$$Q^0_{x_s,a} = Q_{\theta_0 - h}, \qquad (s, a) \neq (t, j), \quad a \in \{1, \dots, L\}.$$

At the anchor, set $Q^0_{x_{\mathrm{anc}},0} = Q_1$; reward laws of unused anchor actions may be set to Q_0. The instances have the same context law, behavior policy, policy class, costs, and ambiguity radii. They differ only in which of the rL competing cells is elevated.

D.2 Step 2: Robust regret and coverage

For any coupling (Y, Z) with $Y \sim Q_\theta$, $Z \in [0, D]$, and $\mathbb{E}[(Y - Z)^2] \le \epsilon_{\mathcal{Y}}$, nonnegativity of Z and Cauchy–Schwarz give

$$\begin{aligned} D\theta - \mathbb{E}[Z] &= \mathbb{E}[Y - Z] \\ &\le \mathbb{E}[(Y - Z)\mathbf{1}\{Y = D\}] \\ &\le \sqrt{\theta \epsilon_{\mathcal{Y}}}. \end{aligned}$$

When $\theta > \sigma$, equality is attained by coupling $Z = 0$ when $Y = 0$ and $Z = D - \sqrt{\epsilon_{\mathcal{Y}}/\theta}$ when $Y = D$. Consequently,

$$q(\theta) := \inf_{Q: W_2^2(Q_\theta, Q) \le \epsilon_{\mathcal{Y}}} \mathbb{E}_{Y \sim Q}[Y] = D\big(\theta - \sqrt{\sigma\theta}\big). \tag{10}$$

The restriction on h places $\theta_0 - h, \theta_0, \theta_0 + h$ in

$$\left[\frac{1 + 3\sigma}{4}, \frac{3 + \sigma}{4}\right] \subset (\sigma, 1).$$

On this interval,

$$q'(\theta) = D\left(1 - \frac{\sqrt{\sigma}}{2\sqrt{\theta}}\right) \ge \frac{D}{2}.$$

Writing

$$q_- := q(\theta_0 - h), \qquad q_0 := q(\theta_0), \qquad q_+ := q(\theta_0 + h), \qquad q_{\mathrm{s}} := q(1),$$

the mean-value theorem gives

$$q_+ - q_0 \ge \frac{Dh}{2}, \qquad q_0 - q_- \ge \frac{Dh}{2}. \tag{11}$$

Moreover, $q_{\mathrm{s}} > q_+ > q_0 > q_-$.

For the discrete context cost, the diagonal mass of any coupling is at most $\sum_x \min\{P^0(x), P(x)\}$, and this bound is attainable by coupling the remaining excess and deficit mass off the diagonal. Thus

$$\begin{aligned} \mathcal{T}_{\mathcal{X}}(P^0, P) &= \kappa\left(1 - \sum_{x \in \mathcal{X}} \min\{P^0(x), P(x)\}\right) \\ &= \kappa \|P - P^0\|_{\mathrm{TV}}. \end{aligned}$$

For a linear objective, the positive and negative parts of $P - P^0$ have the same mass, equal to $\|P - P^0\|_{\mathrm{TV}}$. Hence each unit of mass decreases the objective most by moving it from a largest coefficient to a smallest coefficient (and increases it most by reversing this move). Here the anchor contains $1/2 > \rho$ mass and every hard context can receive $\rho < 1/2$ mass, so moving the full TV budget in this way is feasible. Under $\mathsf{M}_{t,j}$, the policy $\pi_{t,j}$ has robust reward q_+ at x_t, q_0 at the other hard contexts, and q_{s} at the anchor. Since $r \ge 2$, the context adversary moves ρ mass from the anchor to a hard context with reward q_0. Therefore

$$V_{t,j}(\pi_{t,j}) = \frac{q_+}{2r} + \left(\frac{1}{2} - \frac{1}{2r} + \rho\right) q_0 + \left(\frac{1}{2} - \rho\right) q_{\mathrm{s}}.$$

Every other policy $\pi_{s,k} \ne \pi_{t,j}$ receives reward q_- at its selected competing cell and q_0 at every other hard context. Its context adversary moves ρ anchor mass to that low-reward context, so

$$V_{t,j}(\pi_{s,k}) = \left(\frac{1}{2r} + \rho\right) q_- + \left(\frac{1}{2} - \frac{1}{2r}\right) q_0 + \left(\frac{1}{2} - \rho\right) q_{\mathrm{s}}.$$

It follows from (11) that $\pi_{t,j}$ is uniquely optimal and every wrong policy incurs robust regret at least

$$\begin{aligned} V_{t,j}(\pi_{t,j}) - V_{t,j}(\pi_{s,k}) &= \frac{q_+ - q_-}{2r} + \rho(q_0 - q_-) \\ &\ge \frac{Dh}{2r} + \frac{\rho D h}{2} \\ &= \frac{wDh}{2} + \frac{Dh}{4r} \\ &\ge \frac{wDh}{2}. \end{aligned} \tag{12}$$

Both shifts are active: the context adversary strictly lowers the optimal policy value by moving ρ mass away from the anchor, and (10) is strictly smaller than the nominal mean for every relevant reward law.

For the optimal policy $\pi_{t,j}$, the coverage weights of the selected context–action cells are

$$\sqrt{\frac{2r}{\mu}} \quad \text{at } x_t, \qquad \sqrt{\frac{2r}{b}} \quad \text{at every other hard context,} \qquad \sqrt{2} \quad \text{at the anchor.}$$

By (9), the first weight is the largest and the anchor weight is the smallest. The coverage adversary consequently moves ρ mass from the anchor to x_t, giving every instance the common coefficient

$$C^{\star}_{\mathrm{rob}} = w\sqrt{\frac{2r}{\mu}} + \left(\frac{1}{2} - \frac{1}{2r}\right)\sqrt{\frac{2r}{b}} + \left(\frac{1}{2} - \rho\right)\sqrt{2}. \tag{13}$$

Let $H := \sqrt{2r/\mu}$. Since $\sqrt{\mu/b} \le w$,

$$\left(\frac{1}{2} - \frac{1}{2r}\right)\sqrt{\frac{2r}{b}} \le \frac{wH}{2}.$$

Also, $\sqrt{\mu} \le w\sqrt{b} \le w$, and hence

$$\left(\frac{1}{2} - \rho\right)\sqrt{2} \le \frac{\sqrt{\mu}}{2\sqrt{r}}H \le \frac{wH}{2}.$$

Thus

$$C^{\star}_{\mathrm{rob}} \le 2w\sqrt{\frac{2r}{\mu}}, \qquad \text{and hence} \qquad w\sqrt{\frac{r}{\mu}} \ge \frac{C^{\star}_{\mathrm{rob}}}{2\sqrt{2}}. \tag{14}$$

D.3 Step 3: Relative entropy

Let $\mathbb{P}^n_{t,j}$ be the logged-data law under $\mathsf{M}_{t,j}$. Any two distinct instances differ only at their two elevated-cell indices. Each such cell has nominal probability $\mu/(2r)$. The chain rule for relative entropy therefore gives

$$\mathrm{KL}(\mathbb{P}^n_{t,j} \| \mathbb{P}^n_{s,k}) = \frac{n\mu}{2r}\Big\{ \mathrm{kl}(\theta_0 + h, \theta_0 - h) + \mathrm{kl}(\theta_0 - h, \theta_0 + h) \Big\}.$$

For Bernoulli distributions,

$$\mathrm{kl}(u, v) \le \frac{(u - v)^2}{v(1 - v)}.$$

Every parameter belongs to $[(1 + 3\sigma)/4, (3 + \sigma)/4]$, on which

$$v(1 - v) \ge \frac{(3 + \sigma)(1 - \sigma)}{16}.$$

Since the parameters differ by $2h$,

$$\mathrm{KL}(\mathbb{P}^n_{t,j} \| \mathbb{P}^n_{s,k}) \le \frac{64 n\mu h^2}{r(3 + \sigma)(1 - \sigma)}. \tag{15}$$

D.4 Step 4: Fano and binary testing

Let

$$K := rL = (|\mathcal{X}| - 1)(|\mathcal{A}| - 1).$$

Because $|\mathcal{X}|, |\mathcal{A}| \ge 3$, we have $K \ge 4$. We treat separately the cardinality-dominated and confidence-dominated regimes. Randomized learners are included by adjoining their independent random seed to the logged data, which does not change any relative entropy.

Suppose first that

$$\log K \ge \log(1/\delta).$$

Set

$$h := \sqrt{\frac{(3+\sigma)(1-\sigma) r \log K}{512 n \mu}}.$$

Since

$$\frac{3+\sigma}{1-\sigma} \le 7, \qquad \log K \le \log\left(\frac{4|\mathcal{X}||\mathcal{A}|}{\delta}\right),$$

the sample-size condition in the theorem implies $h \le (1-\sigma)/4$. Equation (15) gives

$$\max_{(t,j)\neq(s,k)} \mathrm{KL}(\mathbb{P}^n_{t,j} \| \mathbb{P}^n_{s,k}) \le \frac{1}{8} \log K.$$

Fano's inequality under the uniform prior on the K instances yields

$$\begin{aligned} \frac{1}{K} \sum_{t=1}^{r} \sum_{j=1}^{L} \mathbb{P}^n_{t,j}(\widehat{\pi} \neq \pi_{t,j}) &\ge 1 - \frac{1}{8} - \frac{\log 2}{\log K} \\ &\ge \frac{3}{8}. \end{aligned}$$

Hence, in at least one instance, the learner chooses a wrong policy with probability at least $3/8 > \delta$. On this event, (12) and (14) give

$$\begin{aligned} V_{t,j}(\pi_{t,j}) - V_{t,j}(\widehat{\pi}) &\ge \frac{D\sqrt{(3+\sigma)(1-\sigma)}}{128} C^\star_{\mathrm{rob}} \sqrt{\frac{\log K}{n}} \\ &\ge \frac{\sqrt{7}}{256} D C^\star_{\mathrm{rob}} \sqrt{\frac{\log K}{n}}, \end{aligned}$$

where the last inequality uses $\sigma \le 1/2$.

Suppose now that

$$\log K < \log(1/\delta).$$

Use any two distinct instances and set

$$h := \sqrt{\frac{(3+\sigma)(1-\sigma) r \,\mathrm{kl}(1-\delta, \delta)}{128 n \mu}}.$$

For $\delta \in (0, 1/4]$,

$$\frac{1}{4} \log \frac{1}{\delta} \le \mathrm{kl}(1-\delta, \delta) \le \log \frac{1}{\delta}.$$

The theorem's sample-size condition, the upper bound above, and $\log(1/\delta) \le \log(4|\mathcal{X}||\mathcal{A}|/\delta)$ imply $h \le (1-\sigma)/4$, while (15) gives

$$\mathrm{KL}(\mathbb{P}^n_{t,j} \| \mathbb{P}^n_{s,k}) \le \frac{1}{2} \mathrm{kl}(1-\delta, \delta).$$

If the learner chose the correct policy with probability at least $1-\delta$ in both instances, data processing applied to $\mathbf{1}\{\widehat{\pi} = \pi_{t,j}\}$ would give

$$\mathrm{KL}(\mathbb{P}^n_{t,j} \| \mathbb{P}^n_{s,k}) \ge \mathrm{kl}(1-\delta, \delta),$$

a contradiction. Thus, in at least one of the two instances, the learner chooses a wrong policy with probability greater than δ. On this event,

$$\begin{aligned} V_{t,j}(\pi_{t,j}) - V_{t,j}(\widehat{\pi}) &\ge \frac{D\sqrt{(3+\sigma)(1-\sigma)}}{64} C^\star_{\mathrm{rob}} \sqrt{\frac{\mathrm{kl}(1-\delta, \delta)}{n}} \\ &\ge \frac{\sqrt{7}}{256} D C^\star_{\mathrm{rob}} \sqrt{\frac{\log(1/\delta)}{n}}. \end{aligned}$$

In either regime, some instance has error probability greater than δ and regret on the error event at least

$$\frac{\sqrt{7}}{256} D C^\star_{\mathrm{rob}} \sqrt{\frac{\max\{\log K, \log(1/\delta)\}}{n}}.$$

Since $r, L \geq 2$,

$$|\mathcal{X}| = r + 1 \leq \frac{3}{2}r, \qquad |\mathcal{A}| = L + 1 \leq \frac{3}{2}L.$$

Consequently,

$$4|\mathcal{X}||\mathcal{A}| \leq 9K \leq K^3,$$

where the last inequality uses $K \geq 4$. Hence

$$\log\left(\frac{4|\mathcal{X}||\mathcal{A}|}{\delta}\right) \leq 4\max\left\{\log K, \log\frac{1}{\delta}\right\}.$$

Substitution proves Theorem 2.

Finally, taking $\delta = 1/4$ gives an error probability greater than $1/4$. Since regret is nonnegative, the same construction gives, whenever

$$n\mu \geq \frac{7r}{8}\log(16|\mathcal{X}||\mathcal{A}|),$$

an instance satisfying

$$\mathbb{E}\left[V(\pi^\star) - V(\widehat{\pi})\right] \geq \frac{\sqrt{7}}{2048}DC^\star_{\mathrm{rob}}\sqrt{\frac{\log(16|\mathcal{X}||\mathcal{A}|)}{n}}.$$

E EXPERIMENT DETAILS

E.1 EMPIRICAL BEHAVIOR OF THE ROBUST-REWARD CONFIDENCE RADIUS

We examine the finite-sample confidence radius in Proposition 1 across a collection of random and structured reward distributions. The experiment isolates reward estimation within a single context–action cell, setting $|\mathcal{X}||\mathcal{A}| = 1$, and compares the largest empirical error quantile across the tested distributions with the theoretical radius.

Reward distributions and ambiguity set. We take $\mathcal{Y} = [0, 1]$, so $D = 1$, and use the quadratic transport cost $c_{\mathcal{Y}}(y, z) = |y - z|^2$ with ambiguity radius $\epsilon_{\mathcal{Y}} = 0.04$. We evaluate $J = 58$ nominal reward distributions: 50 randomly generated distributions and 8 structured cases. For each random instance, we first draw its support size uniformly from $\{2, \ldots, 99\}$. Conditional on this size, we draw the support points independently from the uniform distribution on $[0, 1]$ and their probability vector from $\mathrm{Dirichlet}(1, \ldots, 1)$.

The structured cases broaden the test to include balanced and highly unbalanced probabilities, concentrated supports, and degenerate rewards:

Instance	Support	Probabilities
Balanced Bernoulli	$(0, 1)$	$(0.5, 0.5)$
Rare high reward	$(0, 1)$	$(0.99, 0.01)$
Rare low reward	$(0, 1)$	$(0.01, 0.99)$
Support near zero	$(0, 0.01, 0.05)$	$(0.1, 0.2, 0.7)$
Support near one	$(0.95, 0.99, 1)$	$(0.2, 0.3, 0.5)$
Near-zero robust reward	$(0, 1)$	$(0.959, 0.041)$
Deterministic zero	$\{0\}$	(1)
Deterministic one	$\{1\}$	(1)

Although the nominal distributions have finite supports, the adversary may transport rewards anywhere in $[0, 1]$.

Computation of robust rewards. For instance j, let Q_j^0 denote its nominal distribution and q_j its true robust reward. We compute both q_j and its empirical estimate using the continuous quadratic-cost problem. For any nominal distribution Q, if $\mathbb{E}_Q[Y^2] \le \epsilon_{\mathcal{Y}}$, transporting all rewards to zero is feasible and the robust reward equals zero. Otherwise, the optimal transport maps

$$Y \longmapsto \max\{0, Y - t_Q\},$$

where t_Q satisfies

$$\mathbb{E}_Q\big[\min\{Y, t_Q\}^2\big] = \epsilon_{\mathcal{Y}}.$$

The robust reward is therefore $\mathbb{E}_Q[\max\{0, Y - t_Q\}]$. We solve the scalar equation by bisection, using the known probabilities for q_j and the empirical frequencies for $\widehat{q}_{j,m}$. This computation is equivalent to the reward dual formulation and requires no discretization of the adversarial reward space. As a numerical check, we compare the batched implementation with an independent scalar dual optimizer on 232 population and empirical distributions; the maximum absolute discrepancy is 1.28×10^{-14}.

Monte Carlo design. We use 20 geometrically spaced sample sizes between 10 and 10^6, rounded to integers. For every instance–sample-size pair, we generate $R = 1000$ independent replications, giving $58 \times 20 \times 1000 = 1{,}160{,}000$ replications in total. For computational efficiency, we sample multinomial reward counts directly. These counts have exactly the same distribution as those obtained from m independent draws from the corresponding nominal distribution.

We set $\delta = 0.1$. With $D = 1$ and $|\mathcal{X}||\mathcal{A}| = 1$, the theoretical confidence radius is

$$\beta_Y(m; 0.1) = \min\left\{1, \sqrt{\frac{\pi}{2m}} + \sqrt{\frac{\log 20}{2m}}\right\}.$$

Proposition 1 gives, for each fixed instance and sample size,

$$\Pr(|\widehat{q}_{j,m} - q_j| \le \beta_Y(m; 0.1)) \ge 1 - \frac{\delta}{2} = 0.95.$$

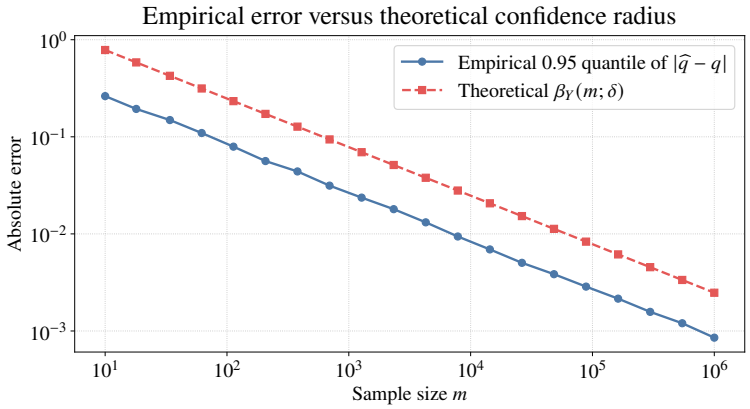

Figure 1: Empirical error versus theoretical confidence radius. At each sample size, the solid curve shows the maximum across 58 nominal reward distributions of the empirical 0.95 quantile of $|\widehat{q}_{j,m} - q_j|$, computed from 1000 replications per distribution. The dashed curve shows $\beta_Y(m; 0.1)$. Both axes use logarithmic scales.

We therefore compute the empirical 0.95 absolute-error quantile separately for each instance,

$$\widehat{e}_{j,0.95}(m) := \mathrm{Quantile}_{0.95}\left(\left\{|\widehat{q}_{j,m}^{(r)} - q_j|\right\}_{r=1}^{R}\right),$$

and plot

$$\widehat{e}_{\max,0.95}(m) := \max_{1\le j\le J} \widehat{e}_{j,0.95}(m)$$

against $\beta_Y(m; 0.1)$. Taking the maximum prevents averaging across instances from obscuring a distribution with a larger error quantile. The 58 instances are separate experiments, rather than 58 context–action cells, so their number does not enter the cell-count factor in β_Y.

Results. Figure 1 shows that the largest empirical 0.95 error quantile remains below the theoretical radius at every tested sample size. Across all 1160 instance–sample-size pairs,

$$\max_{j,m} \frac{\widehat{e}_{j,0.95}(m)}{\beta_Y(m; 0.1)} = 0.358228.$$

Moreover, no individual replication violates $|\widehat{q}_{j,m} - q_j| \le \beta_Y(m; 0.1)$. The balanced Bernoulli distribution attains the largest empirical error quantile among the tested instances at every sample size. At $m = 10$, this quantile is approximately 0.262536, compared with a theoretical radius of 0.783355. At $m = 10^6$, the corresponding values are approximately 0.000854 and 0.002477.

The two curves exhibit similar decay on logarithmic axes. A log–log regression using the reported empirical quantiles gives a slope of approximately -0.498, close to the $m^{-1/2}$ dependence of the theoretical radius.

Interpretation. These results provide numerical evidence consistent with the confidence guarantee across the tested random and structured distributions, while illustrating its $m^{-1/2}$ dependence. The theoretical radius is at least approximately 2.79 times the empirical 0.95 error quantile for every tested pair, indicating conservative constants in this experiment. Such conservatism is compatible with a distribution-free bound that controls the reward dual integrands uniformly.

The experiment does not establish universal validity or worst-case tightness. The reported maximum is taken over a finite collection of empirical quantiles and is subject to Monte Carlo variability. In particular, it is not the 0.95 quantile of a joint maximum error, nor does the plot assert simultaneous 95% coverage across all instances and sample sizes. The varying nominal support sizes also provide a check across different discrete distributions, but this experiment does not isolate the effect of support cardinality or establish the same behavior for every transport cost and ambiguity radius.

E.2 Dependence of Regret on n and C^*_{rob}

We examine the dependence of Algorithm 1's robust regret on the sample size n and the robust optimal-policy coverage coefficient $C^\star_{\mathrm{rob}}$. With the remaining parameters fixed, the upper bound in Theorem 1 scales as $C^\star_{\mathrm{rob}}/\sqrt{n}$. Our experiment examines when this scaling is observed over a controlled family of instances.

Instance construction. We use $|\mathcal{X}| = 4$ uniformly distributed contexts, $|\mathcal{A}| = 9$ actions indexed by $0, \ldots, 8$, and reward range $\mathcal{Y} = [0, D]$ with $D = 1$. The context ambiguity set is the total-variation ball $\{P : \|P - P^0\|_{\mathrm{TV}} \le \rho\}$ with $\rho = 0.2$. For rewards, we use squared transport cost $c_{\mathcal{Y}}(y, z) = (y - z)^2$ and radius $\epsilon_{\mathcal{Y}} = 0.05D^2 = 0.05$. A nominal reward $Y \sim \mathrm{Bernoulli}(\theta)$, has robust mean

$$q(\theta) = D \max\{\theta - \sqrt{\epsilon_{\mathcal{Y}}\theta}, 0\}.$$

At every context, action 0 has deterministic reward zero, actions $2, \ldots, 8$ have Bernoulli parameter $\theta_0 = 0.65$, and action 1 has parameter θ_1 chosen so that $q(\theta_1) = q(\theta_0) + \Delta$, for a positive robust reward gap Δ. Thus, action 1 is optimal at every context and $V(\pi^\star) = q(\theta_1)$.

To prescribe a coverage coefficient C, we set

$$\mu = \frac{|\mathcal{X}|}{C^2}, \qquad \pi^0_x(a) = \mu \quad (a = 1, \ldots, 8), \qquad \pi^0_x(0) = 1 - 8\mu.$$

These probabilities are valid throughout the sweep. Since $P^0(x)\pi^0_x(1) = C^{-2}$ at every context, the coverage coefficient is exactly

$$C^\star_{\mathrm{rob}} = \sup_{P:\|P-P^0\|_{\mathrm{TV}}\le\rho} \sum_x \frac{P(x)}{\sqrt{P^0(x)\pi^0_x(1)}} = C.$$

Increasing C therefore reduces the expected number of observations in each optimal-action cell to n/C^2, while preserving the context distribution, action set, and ambiguity radii.

Learner and regret evaluation. We fix $\delta = 0.1$ and use the confidence split from Theorem 1:

$$\widehat{\pi}_{\delta/2}(x) \in \arg\max_{a\in\mathcal{A}} \{\widehat{q}(x, a) - \beta_Y(N_{x,a}; \delta/2)\}.$$

Here $\widehat{q}(x, a) = q(\widehat{\theta}_{x,a})$ is computed from the empirical Bernoulli frequency. The confidence penalty has multiplier one, and the pessimistic scores are not clipped. Empty cells receive score $-D$, and exact ties are resolved by the smallest action index. For each learned policy, we compute its population robust regret as

$$R = q(\theta_1) - \min_{P:\|P-P^0\|_{\mathrm{TV}}\le\rho} \sum_x P(x) q\big(x, \widehat{\pi}_{\delta/2}(x)\big).$$

The minimization uses the known instance parameters and is solved exactly; there is no additional Monte Carlo error from policy-value evaluation. Logged datasets are simulated through multinomial cell counts and conditional binomial success counts. These are exact sufficient statistics for n independent logged observations and allow large sample sizes without storing every observation.

Sweep and independent gap selection. We evaluate all 425 combinations of 25 geometrically spaced sample sizes from 2^{10} to 2^{30}, rounded to integers, and 17 geometrically spaced coverage coefficients from 16 to 128. At each combination, we search the same fixed grid of 61 geometrically spaced gaps in $[10^{-5}D, 0.18D]$. Using 1,000 training-dataset repetitions per candidate gap, we select the gap with the largest estimated mean regret. We then estimate regret at this selected gap using 5,000 fresh, independent repetitions, giving 2,125,000 evaluation repetitions in total. Within each search, cell counts and decoy-action observations are shared across candidate gaps; each candidate retains the correct logged-data marginal distribution.

This procedure approximates the maximum expected regret over the gap grid, rather than following a single fixed reward distribution as n changes. The target exponents do not enter the gap grid or selection rule, and independent evaluation avoids reusing the selection data to report regret. The upper gap endpoint is selected in 111 of the 425 configurations; the lower endpoint is never selected. Consequently, the small-sample results also reflect the finite gap range, and should not be interpreted as the exact worst case over all reward distributions. Error bars are the evaluation mean plus or minus 1.96 Monte Carlo standard errors, conditional on the selected gap; they do not quantify uncertainty in identifying the maximizing gap.

Small-sample behavior. The relevant expected cell count is $m = n/(C^\star_{\mathrm{rob}})^2$. At $n = 2^{10}$, it ranges from 4 for $C^\star_{\mathrm{rob}} = 16$ to 0.0625 for $C^\star_{\mathrm{rob}} = 128$. In the latter case, an optimal-action cell is empty with probability $(1 - 128^{-2})^{1024} \approx 0.9394$. For our fixed confidence level and problem dimensions, the penalty is

$$\beta_Y(N; \delta/2) = D \min\left\{1, \frac{k}{\sqrt{N}}\right\}, \qquad k = \sqrt{\frac{\pi}{2}} + \sqrt{\frac{1}{2}\log\frac{4|\mathcal{X}||\mathcal{A}|}{\delta}} \approx 3.1602, \qquad N \geq 1.$$

Thus, the penalty saturates at D for counts $N \leq 9$. Poorly observed actions can have lower pessimistic scores than the frequently observed zero-reward action, resulting in substantial regret. Increasing n eventually reduces the penalties enough for the learner to select among the positive-reward actions. This transition need not follow a single power law.

Boundedness also prevents an exact $C^\star_{\mathrm{rob}}/\sqrt{n}$ relationship throughout the full range. On the event of the theoretical guarantee, bounded regret and Theorem 1 together imply a bound of the form

$$R \leq \min\left\{D,\ KDC^\star_{\mathrm{rob}}\sqrt{\frac{\log(4|\mathcal{X}||\mathcal{A}|/\delta)}{n}}\right\},$$

where K is a theorem-dependent constant. This is a high-probability upper bound, not an equality for the mean regret plotted here.

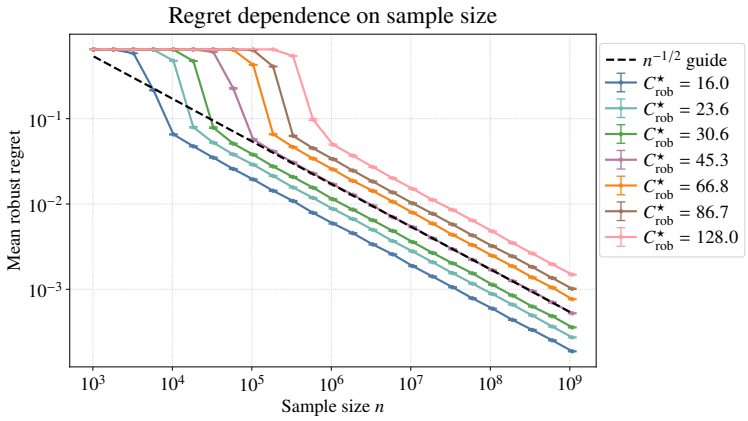

Figure 2: Mean population robust regret versus sample size for seven coverage coefficients. Each point uses independent evaluation data after gap selection. Error bars show 1.96 Monte Carlo standard errors; the dashed line is an $n^{-1/2}$ scaling guide.

Results and interpretation. Figures 2 and 3 show seven evenly spaced slices of the sweep on logarithmic axes. At $n = 2^{10}$, mean regret ranges from 0.6464 to 0.6497 across coverage coefficients, close to the largest optimal robust value in the tested family, $q(0.65) + 0.18D \approx 0.6497D$. The mean fraction of contexts at which the learner chooses action 0 ranges from 0.9719 to 1. The sample-size curves exhibit a plateau followed by a sharp decrease, with the transition occurring at larger sample sizes for larger coverage coefficients. Beyond this transition, the curves approach the $n^{-1/2}$ reference slope; the dependence on coverage becomes approximately linear when sample sizes are sufficiently large.

The reference lines indicate scaling only, not theoretical upper bounds. The $n^{-1/2}$ guide is anchored at the largest sample size of the middle displayed coverage slice; the linear coverage guide is anchored at the smallest coverage coefficient of the largest displayed sample size. To quantify the distinction between the full range and the large-sample behavior, we fit

$$\log \overline{R} = b_0 + b_n \log n + b_C \log C^\star_{\mathrm{rob}}$$

using two explicitly specified sets of configurations. A joint fit over all 425 points gives $b_n = -0.5708$ and $b_C = 1.0256$, with an R^2 of 0.96057 on the logarithmic scale. These global coefficients mix the plateau, transition, and subsequent decay and are not estimates of a single large-sample power law.

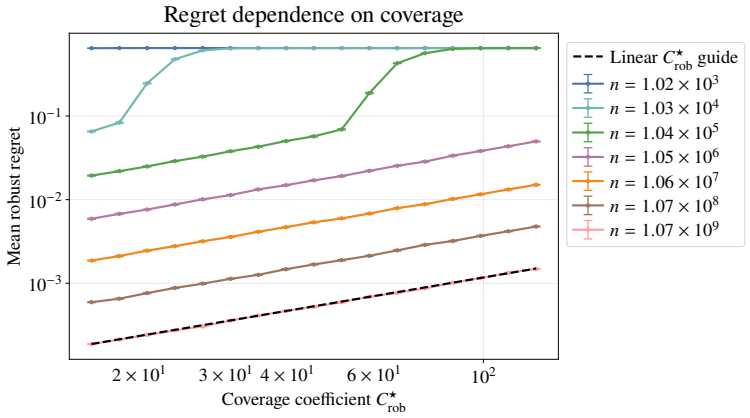

Figure 3: Mean population robust regret versus the exact coverage coefficient $C^\star_{\mathrm{rob}}$ for seven sample sizes. Error bars show 1.96 Monte Carlo standard errors; the dashed line is a linear coverage scaling guide anchored to the largest sample size. Both axes are logarithmic.

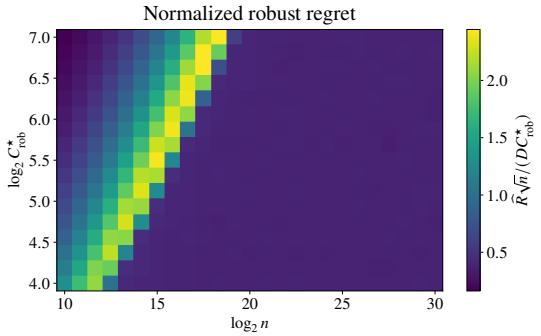

Figure 4: Normalized mean regret $\overline{R}\sqrt{n}/(DC^\star_{\mathrm{rob}})$ across all 425 configurations of the sweep.

When applying the lower cutoff $n \geq 2^{22}$, the resulting exponents are $b_n = -0.5004$ and $b_C = 1.0021$, with $R^2 = 0.99988$. None of these 170 configurations selects a gap-grid endpoint. Moreover, their expected optimal-action cell counts exceed 256, so they are well above the count range in which the confidence penalty saturates. These diagnostics support interpreting this restricted fit as a description of the observed large-sample regime; they do not define a universal threshold for such behavior.

Figure 4 provides a joint check using all configurations. Over the full range, the normalized mean regret $\overline{R}\sqrt{n}/(DC^\star_{\mathrm{rob}})$ varies from 0.1624 to 2.4448, making the departure from a uniform power law visible. On the restricted range $n \geq 2^{22}$, it lies between 0.3585 and 0.3910, with mean 0.3815 and coefficient of variation approximately 1.13%. Thus, in this large-sample portion of the sweep, the evaluated family exhibits approximately

$$\overline{R} \approx 0.382 \, \frac{DC^\star_{\mathrm{rob}}}{\sqrt{n}}.$$

The expanded sweep therefore shows both the finite-sample transition and subsequent behavior consistent with the predicted sample-size and coverage exponents on this family. The full curves are retained so that the restricted fit is not mistaken for a relationship holding at every sample size. These observations do not establish a universal lower bound, require this rate on every fixed instance, or validate the constants and failure probability of the theoretical high-probability bound. The logarithmic factor is constant in this experiment because $|\mathcal{X}|$, $|\mathcal{A}|$, and δ are fixed.

E.3 Real-Data Application: Off-Policy Learning on the Open Bandit Dataset

The two preceding experiments verify the theorem's guarantees against a known ground truth. We complement them with an application of Algorithm 1 to a real, publicly available logged-bandit dataset: the **Open Bandit Dataset** (OBD), collected on the large-scale fashion e-commerce platform ZOZOTOWN and released together with an open-source pipeline for reproducible off-policy evaluation (Saito et al., 2020). Unlike the synthetic instances above, OBD carries no known ground-truth context distribution P^0 or reward laws $\{Q^0_{x,a}\}_{x,a}$, so this experiment cannot verify the regret rate of Theorem 1 against the robust optimal-policy coverage coefficient $C^\star_{\mathrm{rob}}$: there is no ground truth against which to certify the learned policy's regret, as there was in Appendix E.2. We instead evaluate the off-policy value of the policy that Algorithm 1 learns from logged data.

Dataset. OBD logs recommendation interactions collected under a live production system and is partitioned into three *campaigns*, corresponding to different sections of the platform's catalog: *all* (the union campaign, 80 items), *men* (34 items), and *women* (46 items). For each campaign, OBD provides logs collected under a **uniform-random** policy. Every logged round records the item recommended, its position among the three items displayed simultaneously to the user, a binary click indicator, an anonymized user context, and the propensity score of the item actually recommended, computed directly by the logging system rather than estimated post hoc. We use the dataset's full-scale release. Each campaign–policy log is split at the row level into training, validation, and test subsets, in proportions $60\%/20\%/20\%$.

Campaign	Logging policy	Rows	$\lvert\mathcal{X}\rvert$	$\lvert\mathcal{A}\rvert$	CTR (train/val/test)
men	random	452,949	345	34	$0.490\%/0.546\%/0.544\%$
women	random	864,585	351	46	$0.478\%/0.490\%/0.484\%$
all	random	1,374,327	404	80	$0.353\%/0.349\%/0.327\%$

Table 3: Dataset scale per campaign and logging policy: row counts, the resulting discretized context and action cardinalities, and the click-through rate (CTR) in each split.

Context, action, and reward. We map each logged round to an i.i.d. draw (x, a, y) from the paper's contextual-bandit model as follows.

Action. The action $a \in \mathcal{A}$ is the item recommended by the logging policy; $|\mathcal{A}| \in \{34, 46, 80\}$ depending on the campaign (Table 3).

Context. OBD logs four anonymized, categorical `user_feature` fields per round, whose semantics are not disclosed by the dataset. We one-hot encode these four fields, dropping one reference level per field, and take the set of *distinct* resulting rows as the finite context set $\mathcal{X}$ that the paper's theory requires; this discretization is not given directly by the raw data and yields $|\mathcal{X}|$ between 345 and 404 depending on campaign (Table 3). OBD additionally logs a *position* field, since three items are displayed simultaneously per round. We do not fold position into x, nor otherwise model the resulting three-slot slate structure; every logged row is instead treated as an independent draw from a flat, non-slate contextual bandit.

Reward. The reward $y \in \{0, 1\}$ is the logged click indicator. The click-through rate (CTR) is low across all campaigns, between approximately 0.33% and 0.55% (Table 3); this low reward rate is the dominant source of statistical difficulty throughout the experiment, both for estimating $\underline{q}(x, a)$ from the training split and for off-policy evaluation on the test split.

Learner. We run Algorithm 1 without computing $\beta_{\mathcal{X}}$ or using $\epsilon_{\mathcal{X}}$. For the unrestricted policy class $\Pi = \Delta(\mathcal{A})^{|\mathcal{X}|}$ used throughout this experiment, Section 2.3 shows that the per-context greedy rule $\widehat{\pi}(x) \in \arg\max_a \tilde{q}(x, a)$ is exactly optimal for *every* $\epsilon_{\mathcal{X}}, c_{\mathcal{X}}$, so no linear program and no dual variable η is required to select the policy; only the pessimistic reward table $\tilde{q}(x, a)$, computed independently for each cell $(x, a) \in \mathcal{X} \times \mathcal{A}$ from the training split, is needed.

Closed form of the robust reward. Write $N_{x,a}$ for the number of training rounds logged with context x and action a, and $\widehat{\theta}_{x,a} \in [0, 1]$ for the empirical click frequency among them (the empirical CTR of cell (x, a)). Because the reward is Bernoulli within each cell, the reward-level dual of Lemma 1

reduces to the same closed form used in Appendix E.2: a nominal reward $Y \sim \text{Bernoulli}(\theta)$, has robust value $q(\theta) = D \max\{\theta - \sqrt{\epsilon_{\mathcal{Y}}\theta},\, 0\}$, and the point estimate is $\widehat{q}(x,a) = q(\widehat{\theta}_{x,a})$.

Relative reward-ambiguity radius. We parameterize the reward-ambiguity radius relative to the training set's overall click-through rate rather than as an absolute constant, $\epsilon_{\mathcal{Y}} = \rho \cdot \widehat{\text{CTR}}_{\text{train}}$, for a dimensionless multiplier $\rho > 0$.

Confidence radius. With $\delta = 0.1$ fixed throughout and $n_{\text{cells}} = |\mathcal{X}||\mathcal{A}|$ the total number of context–action cells for the campaign at hand (Table 3), Proposition 1's confidence radius is

$$\beta_Y(N_{x,a};\delta) = D \min\left\{1,\ \sqrt{\frac{\pi}{2N_{x,a}}} + \sqrt{\frac{\log(2\, n_{\text{cells}}/\delta)}{2N_{x,a}}}\right\},$$

with the convention $\beta_Y(0;\delta) = D$. Since experiment E.1 shows that β_Y may be too conservative, we introduce a tunable, dimensionless scale $c \geq 0$ multiplying the confidence radius, giving the signed pessimistic score

$$\widetilde{q}_c(x,a) := \widehat{q}(x,a) - c \cdot \beta_Y(N_{x,a};\delta).$$

The value $c = 1$ recovers exactly the radius that Proposition 1 (and, in turn, Theorem 1) is proved for; $c < 1$ trades away the formal high-probability coverage guarantee for a less conservative learner.

Hyperparameter grid and selection. Two hyperparameters are swept, both on the validation split described above: the relative ambiguity radius $\rho \in \{0.01, 0.03, 0.1, 0.3, 0.6, 1.0\}$ and the confidence-radius scale $c \in \{1.0, 0.5, 0.25, 0.1, 0.05, 0.0\}$, giving a common $6 \times 6 = 36$-point grid for every campaign. For each ρ, the reward table $\widehat{q}(\cdot,\cdot)$ is recomputed once at $\epsilon_{\mathcal{Y}} = \rho \cdot \widehat{\text{CTR}}_{\text{train}}$; the inner sweep over c then only rescales β_Y and re-selects the greedy policy, reusing the same $\widehat{q}$ and cell counts $N_{x,a}$. Each grid point's policy is scored by a self-normalized off-policy value estimate on the validation split (defined below), and any grid point matched by fewer than 30 validation rounds that actually resulted in a click is discarded as unreliable before selecting the best; the winning (ρ, c) pair is then evaluated exactly once on the held-out test split.

Off-policy value estimators. Every deterministic policy π considered in this experiment — the greedy policy $\widehat{\pi}_c$ of Algorithm 1 at any grid point, and every baseline evaluated against it below — selects a single action $\pi(x) \in \mathcal{A}$ for each context $x \in \mathcal{X}$. We evaluate such a policy on a held-out split $S = \{(x_i, a_i, y_i, \text{pscore}_i)\}_{i=1}^n$, drawn from the same log as training but disjoint from it (either the validation split, for hyperparameter selection, or the test split, for the single final report), using *inverse propensity scoring (IPS)* and its *self-normalized* variant (*SNIPS*), built from OBD's own logged propensity $\text{pscore}_i = \pi^0(a_i \mid x_i)$:

$$\widehat{V}_{\text{IPS}}(\pi) = \frac{1}{n}\sum_{i=1}^{n} \frac{\mathbf{1}\{a_i = \pi(x_i)\}}{\text{pscore}_i}\, y_i, \qquad \widehat{V}_{\text{SNIPS}}(\pi) = \frac{\sum_{i=1}^n \mathbf{1}\{a_i = \pi(x_i)\}\, y_i/\text{pscore}_i}{\sum_{i=1}^n \mathbf{1}\{a_i = \pi(x_i)\}/\text{pscore}_i}.$$

Both estimators are unbiased (IPS) or asymptotically unbiased and typically lower-variance (SNIPS) for $V(\pi) = \mathbb{E}_{x\sim P^0}[y \mid x, \pi(x)]$ under the standard overlap condition $\text{pscore}_i > 0$ whenever $\pi(x_i)$ has positive probability under $P^0(\cdot \mid x_i)$ — satisfied here since π^0 places positive mass on every action throughout logging. We use SNIPS as the selection and headline criterion below, reporting IPS alongside it as a secondary check, since SNIPS is invariant to the arbitrary normalization of pscore_i and empirically more stable.

We deliberately avoid an unweighted matching estimator that averages y_i over rows on which the logged action happens to match $\pi(x_i)$. Beyond being biased toward whichever actions the logging policy favors most, such an estimator is a form of rejection sampling: it discards every row with $a_i \neq \pi(x_i)$ from the average without correcting for the fact that a matching row was already less likely to occur for actions π^0 under-explores. This is excluded by the same experiment-design constraints, carried over from the synthetic experiments' design, that govern the rest of this appendix.

Baselines. We compare Algorithm 1 with the following methods, all trained on the same campaign-specific random-log split:

- **Random logging policy:** the observed test CTR provides a reference for the policy that collected the data (Saito et al., 2020).

- **Factored W-DRO:** the plug-in learner of Shen et al. (2024) chooses actions using empirical Wasserstein-robust rewards, without our statistical confidence penalty.
- **IPW Learner:** the Open Bandit Pipeline learner fits a classifier using click rewards weighted by inverse logged propensities (Saito et al., 2020).
- **POEM:** the stochastic linear-softmax policy of Swaminathan & Joachims (2015) minimizes clipped counterfactual risk with an empirical-variance penalty. We use batch optimization of that objective.
- **Joint KL-DRO:** the stable, self-normalized learner of Si et al. (2020) optimizes worst-case value in a KL neighborhood of the joint environment distribution. We use a depth-three decision-tree policy class.

Results. Table 4 reports coverage of the empirical context–action table and the hyperparameters chosen for Algorithm 1. Table 5 compares test SNIPS. Algorithm 1 has the largest point estimate in each campaign, with a narrow margin over factored W-DRO. POEM is below the IPW Learner on women and all, but above it on men. The joint KL-DRO tree has the lowest learned-policy estimate in each campaign. These are nominal off-policy values, not estimates of worst-case deployment value or robust regret.

Campaign	Occupied cells	Mean count	Selected ρ	Selected c
men	7,263/11,730	37.4	1.0	0.25
women	9,071/16,146	57.2	0.01	0.10
all	18,049/32,320	45.7	0.03	0.25

Table 4: Training-cell occupancy, mean count among occupied cells, and selected hyperparameters for Algorithm 1 under uniform-random logging.

Campaign	Behavior policy	Factored W-DRO	IPW Learner	POEM	Joint KL-DRO	Ours
men	0.544%	2.129%	1.087%	1.570%	0.614%	**2.155%**
women	0.483%	1.763%	1.410%	1.315%	0.936%	**1.796%**
all	0.327%	1.734%	1.436%	1.086%	0.837%	**1.817%**

Table 5: Test-split SNIPS under uniform-random logging. POEM is stochastic; the other learned policies are deterministic. Bold marks the largest estimate in each row.

POEM selects $(M, \lambda) = (34, 5), (46, 0), (40, 20)$ for men, women, and all, respectively; its test IPS values are 1.586%, 1.274%, and 1.105%. The selected M equals the inverse random-logging propensity for men and women, so clipping is inactive there, and $\lambda = 0$ removes the variance penalty for women. For joint KL-DRO, every tested KL radius yields the same tree policy on these binary-click data; the test estimates in Table 5 therefore do not identify a preferred radius.

The joint KL-DRO test SNIPS estimates rely on only 16, 34, and 28 matching clicked rows for men, women, and all. For POEM, the expected numbers of logged clicked actions chosen under stochastic policy sampling are 42.3, 47.9, and 38.0; these are not deterministic match counts. Such sparse click support limits precision, so the table does not establish a statistically significant ranking.