Counterfactual Online Conformal Prediction Under Adaptive Logging

Xinyu Qiao[1] **Yichen Lin**[1] **Kaihong Ji**[1] **Xue Wang**[2] **Tao Yao**[1]*
[1]Shanghai Jiao Tong University, Shanghai, China
[2]Meta
[1]{qxinyu25,linyichen,jkh040607,taoyao}@sjtu.edu.cn
[2]xuewang0725@meta.com

Abstract

Online conformal prediction can fail when predictions shape actions and actions determine which outcomes enter calibration. Standard adaptive methods may retain marginal coverage while systematically miscovering the counterfactual outcomes of rarely selected actions. This paper formalizes the failure through counterfactual coverage and introduces Propensity-Weighted Online Conformal Prediction, an inverse-propensity-weighted recursion that debiases calibration. A doubly robust variant further reduces nuisance bias to the product of outcome-model and propensity errors. Under positivity and known logging probabilities, both methods attain the minimax expected coverage rate for nondegenerate calibration. Experiments on synthetic decision tasks, Open Bandit data, and financial rebalancing demonstrate counterfactual coverage gains under selection-biased logging, with prediction-set widths close to those of the oracle in the reported sharpness study.

1 Introduction

Modern decision systems use conformal prediction to guide decisions in clinical trials, content recommendation, and inventory pricing [15, 25, 17, 2]. These decisions also shape subsequent calibration data: a logging policy selects which action is taken, and only that action's outcome enters calibration. Standard online conformal methods [9, 10, 1] preserve average coverage on the realized stream, but the rarely selected actions, often exactly the alternatives a downstream rule must compare, may remain miscalibrated. Context-dependent selection, rather than rarity alone, drives this bias and can also affect frequently selected actions.

The relevant target is therefore not the marginal coverage of the played action but the *counterfactual coverage* of every action: would the prediction set have covered the outcome had that action been chosen [15, 25, 37]? We show that this gap is structural under endogenous logging. Logged-quantile per-action calibration targets the action-conditional quantile of the *logged* score distribution rather than the counterfactual one. In our construction, the two differ by an amount that does not vanish with the horizon.

Our correction, Propensity-Weighted Online Conformal Prediction (PW-OCP), replaces the observed error indicator with an inverse-propensity-weighted estimator of counterfactual miscoverage, realigning the recursion's fixed point with the counterfactual target. A doubly robust variant further reduces nuisance bias to the product of outcome-model and propensity errors. Under positivity and with known propensities, PW-OCP and DR-OCP attain the minimax expected counterfactual coverage rate for nondegenerate calibration. We also bound downstream robust pseudo-regret through a decomposition involving prediction-set accuracy, logged coverage, and exploration.

*Corresponding author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

Contributions. Our contributions are fourfold:

(1). We formulate *counterfactual coverage* for online CP under endogenous logging and prove that per-action calibrators tracking the logged quantile suffer a counterfactual gap of order $\Omega(\rho)$, where ρ quantifies the action-induced shift in the outcome law.

(2). We introduce PW-OCP and DR-OCP, proving high-probability counterfactual coverage bounds and a matching minimax expected rate for nondegenerate calibration with known propensities.

(3). We bound robust pseudo-regret through sharpness, logged coverage, and exploration, and excess oracle robust risk through a bounded prediction-set error. The decompositions allow adaptive logging, with explicit rates under stationary logging and local regularity.

(4). Across seven benchmark settings covering synthetic tasks, Open Bandit data, and financial rebalancing, we evaluate counterfactual coverage, propensity robustness, and prediction-set quality. PW-OCP and DR-OCP substantially reduce coverage gaps in the selection-biased synthetic settings, while gains vary across benchmarks.

2 Problem Formulation

Let $\mathcal{X}$ denote the context space, $\mathcal{A}$ a finite action space with $|\mathcal{A}| = K$, and $\mathcal{Y}$ the outcome space. The interaction lasts T rounds. Write $\mathcal{H}_{t-1} = (X_s, a_s, Y_s)_{s<t}$ for the observed history before round t. The logging policy may depend on the observed history and need not be stationary. Within this sequential sampling model, we impose the causal and positivity conditions below.

Nonconformity scores and prediction sets. Fix a nonconformity score $s : \mathcal{X} \times \mathcal{Y} \times \mathcal{A} \to \mathbb{R}_{\geq 0}$ [27]; standard choices include residual scores $s(x, y; a) = |y - \hat{\mu}(x, a)|$ and scores used in conformalized quantile regression [21]. Given a threshold $\hat{q}_t(a)$, the prediction set for action a at round t is

$$\mathcal{C}_t(a) = \{y \in \mathcal{Y} : s(X_t, y; a) \leq \hat{q}_t(a)\}. \tag{1}$$

The learner maintains one threshold $\hat{q}_t(a)$ per action at target miscoverage level $\alpha \in (0, 1)$.

Sequential interaction. At each round t:

1. the environment reveals X_t;
2. the learner outputs $\{\mathcal{C}_t(a)\}_{a\in\mathcal{A}}$;
3. the logger draws $a_t \sim \pi_t(\cdot \mid X_t, \mathcal{H}_{t-1})$, where $\pi_t(a \mid X_t, \mathcal{H}_{t-1}) := \mathbb{P}(a_t = a \mid X_t, \mathcal{H}_{t-1})$;
4. the environment reveals $Y_t \sim P(\cdot \mid X_t, a_t)$.

The classical online conformal setting corresponds to $P(\cdot \mid X_t, a_t) = P(\cdot \mid X_t)$ [9].

2.1 Potential Outcomes and Structural Assumptions

For each $a \in \mathcal{A}$, let $Y_t(a) \in \mathcal{Y}$ denote the potential outcome under the intervention $a_t = a$ [22], so that $Y_t = Y_t(a_t)$. The environment specifies a joint law over $(X_t, \{Y_t(a)\}_{a\in\mathcal{A}})$. The environment is nonanticipating and may depend on the full past. Conditional on that past, X_t, and the current potential outcomes, the logger uses fresh randomization with probabilities $\pi_t(\cdot \mid X_t, \mathcal{H}_{t-1})$. Thus neither temporal independence of outcomes nor equality of observed-history and full-history outcome kernels is needed for PW-OCP.

Assumption 2.1 (No unmeasured confounders)**.** $\{Y_t(a)\}_{a\in\mathcal{A}} \perp\!\!\!\perp a_t \mid X_t, \mathcal{H}_{t-1}$.

This is sequential ignorability for contextual bandits. It holds whenever every variable used by the logger is included in $(X_t, \mathcal{H}_{t-1})$, e.g., randomized trials and ε-greedy production logging.

Assumption 2.2 (Positivity)**.** $\pi_t(a \mid X_t, \mathcal{H}_{t-1}) \geq \pi_{\min} > 0$ for all t, $a \in \mathcal{A}$, and $X_t \in \mathcal{X}$.

Remark 2.3 (Design role of $\pi_{\min}$). $\pi_{\min}$ is a known design constant in randomized assignment, A/B allocation, and fairness-constrained exposure. The regime $\pi_{\min} \to 0$ (e.g., UCB, Thompson sampling without forced exploration) is excluded; Theorem 5.4 quantifies the information cost of rare observations for nondegenerate calibration. The floor trades exploration cost against effective sample size $T\pi_{\min}$. In our tested setting, $\pi_{\min} = 0.01$ is a practical reference, not a universal threshold.

2.2 Coverage Targets

Definition 2.4 (Coverage targets)**.** 1. **Marginal coverage.** The realized marginal coverage is

$$\mathrm{MCov}_T := \frac{1}{T}\sum_{t=1}^{T} \mathbf{1}\{Y_t \in \mathcal{C}_t(a_t)\}. \tag{2}$$

2. **Action-conditional coverage.** Let $N_T(a) = \sum_{t=1}^{T} \mathbf{1}\{a_t = a\}$. For each action $a \in \mathcal{A}$ with $N_T(a) > 0$, the action-conditional coverage is

$$\mathrm{ACov}_T(a) := \frac{\sum_{t=1}^{T} \mathbf{1}\{a_t = a, Y_t(a) \in \mathcal{C}_t(a)\}}{N_T(a)}. \tag{3}$$

   If $N_T(a) = 0$, this diagnostic is undefined and omitted.

3. **Counterfactual coverage.** For each action $a \in \mathcal{A}$, the counterfactual coverage is

$$\mathrm{CCov}_T(a) := \frac{1}{T}\sum_{t=1}^{T} \mathbf{1}\{Y_t(a) \in \mathcal{C}_t(a)\}. \tag{4}$$

Both MCov_T and $\mathrm{ACov}_T(a)$ for sampled actions are observable. $\mathrm{ACov}_T(a)$ averages over rounds with $a_t = a$ and therefore reflects the policy-induced context distribution, which can differ from the marginal context distribution relevant to $\mathrm{CCov}_T(a)$. $\mathrm{CCov}_T(a)$ averages counterfactual coverage indicators over all rounds and is the sequential analogue of batch counterfactual coverage [15].

The three quantities are not deterministically ordered. When $\rho > 0$ (Definition 2.5), Theorem 3.2 shows that control of MCov_T or $\mathrm{ACov}_T(a)$ need not imply control of $\mathrm{CCov}_T(a)$.

2.3 Endogeneity Coefficient

Definition 2.5 (Endogeneity coefficient)**.**

$$\rho := \sup_{x\in\mathcal{X}} \max_{a\neq a'} D_{\mathrm{TV}}\big(P(\cdot \mid x, a),\, P(\cdot \mid x, a')\big) \;\in\; [0, 1], \tag{5}$$

where $D_{\mathrm{TV}}(P, Q) = \frac{1}{2}\int |p(y) - q(y)|\, dy$ is the total-variation distance.

The case $\rho = 0$ means that the outcome law is action-invariant conditional on the context. It does not exclude context selection by the logger or imply exchangeability over time. For $\rho > 0$, the logged scores are drawn from a policy-dependent mixture rather than the per-action laws required for counterfactual calibration. Calibrating $\{\mathcal{C}_t(a)\}_{a\in\mathcal{A}}$ then becomes a causal estimation problem under selection-biased sampling.

3 Failure of Standard Online CP Under Endogenous Logging

Standard online conformal methods calibrate the recursion against the realized logged error stream. This controls the logged empirical coverage MCov_T (Theorem 3.1). However, Theorem 3.2 shows that calibrators tracking logged per-action quantiles can retain a counterfactual coverage gap that does not vanish with T.

3.1 Logged Empirical Coverage Survives

Theorem 3.1 (Long-run empirical coverage robustness, [9])**.** *Fix $\alpha \in (0, 1)$, step size $\gamma > 0$, and initial value $\alpha_1 \in [0, 1]$. Let $\hat{Q}_t : \mathbb{R} \to \mathbb{R} \cup \{\pm\infty\}$ be a nondecreasing $(\mathcal{H}_{t-1}, X_t)$-measurable score-quantile map satisfying the ACI boundary convention*

$$\hat{Q}_t(u) = -\infty \quad \textit{for } u \le 0, \qquad \hat{Q}_t(u) = +\infty \quad \textit{for } u \ge 1. \tag{6}$$

Set $\hat{q}_t := \hat{Q}_t(1-\alpha_t)$, $\mathcal{C}_t(a) := \{y \in \mathcal{Y} : s(X_t, y; a) \le \hat{q}_t\}$, $\mathrm{err}_t := \mathbf{1}\{Y_t \notin \mathcal{C}_t(a_t)\}$, and run the ACI update

$$\alpha_{t+1} = \alpha_t + \gamma(\alpha - \mathrm{err}_t). \tag{7}$$

Then, for every sample path $(X_t, a_t, Y_t)_{t=1}^T$,

$$|\mathrm{MCov}_T - (1-\alpha)| \le \frac{\max\{\alpha_1,\, 1-\alpha_1\} + \gamma}{\gamma T}. \tag{8}$$

The bound (8) follows the pathwise telescoping argument of ACI. It controls the logged empirical average MCov_T, not a per-round marginal-coverage guarantee, and is agnostic to how $\hat{Q}_t$ is constructed: global, per-arm, or IPW-reweighted quantiles all satisfy the same telescoping identity. This robustness does not ensure counterfactual validity. Because MCov_T averages over the realized action sequence, (8) can hold even when α_t leaves $[0,1]$ and (6) forces $\mathcal{C}_t$ to degenerate to $\mathcal{Y}$ or $\emptyset$ on a non-vanishing fraction of rounds. The bound thus imposes no constraint on rarely selected actions. The proof is in Appendix D.1.

3.2 Counterfactual Coverage Breaks

Theorem 3.2 (Impossibility for logged-quantile per-action calibration)**.** *Fix* $\alpha \in (0,1)$. *There exist absolute constants* $\rho_0, L > 0$ *and a constant* $c(\alpha) > 0$ *such that, for every* $\rho \in (0, \rho_0]$, *there is a stationary two-context, two-action Gaussian instance satisfying Assumptions 2.1–2.2 with endogeneity coefficient* ρ *on which the following holds. Let* $\{\hat{q}_t(a)\}_{t,a}$ *be any sequence of predictable per-action thresholds under the absolute-residual score, and let* $q_a^{\log}$ *denote the* $(1-\alpha)$ *quantile of the score distribution conditional on* $a_t = a$. *Let* $F_{\log}^{(a)}$ *denote the CDF of this logged score distribution, and define the tracking error*

$$\varepsilon_T := \max_{a \in \mathcal{A}} \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T} |F_{\log}^{(a)}(\hat{q}_t(a)) - (1-\alpha)|\right]. \tag{9}$$

Then

$$\mathbb{E}\left[\max_{a \in \mathcal{A}} |\mathrm{CCov}_T(a) - (1-\alpha)|\right] \;\ge\; c(\alpha)\,\rho - L\,\varepsilon_T. \tag{10}$$

In particular, $\varepsilon_T = O(T^{-1/2})$ *implies* $\mathbb{E}[\max_a |\mathrm{CCov}_T(a) - (1-\alpha)|] \ge c(\alpha)\rho - O(T^{-1/2})$.

Theorem 3.2 pinpoints the failure source. The per-action update tracks the logged quantile $q_a^{\log}$ rather than the counterfactual one. In the constructed instance, context-dependent logging and action-dependent outcome laws make the logged and counterfactual quantiles differ by $\Theta(\rho)$, producing a corresponding gap in $\mathrm{CCov}_T(a)$.

The bound (10) is non-vacuous in the regime $\rho \gg \varepsilon_T$.

The result applies to per-action calibrators through their logged-CDF tracking error. A persistent $\Omega(\rho)$ gap follows when this error is small relative to ρ, including for smoothed or aggregated methods whenever they satisfy the same tracking property. The proof, which uses a two-context Gaussian variance-shift construction, is in Appendix D.2. For a concrete example, at $\pi_{\min} = 0.001$, ACI's worst-action gap is 0.4785: a nominal 90% set covers only about 42% of that action's counterfactual outcomes.

4 Propensity-Weighted Online Conformal Prediction

Theorem 3.2 traces standard online CP failure to the calibration target rather than the conformal quantile. PW-OCP corrects the target by replacing the logged error indicator err_t in the ACI recursion (7) with an inverse-propensity-weighted estimator of counterfactual miscoverage [12, 31].

4.1 The IPW Error Signal

Definition 4.1 (IPW error signal)**.**

$$Z_t(a) \;:=\; \frac{\mathbf{1}\{a_t = a\}}{\pi_t(a \mid X_t, \mathcal{H}_{t-1})} \cdot \mathbf{1}\{Y_t \notin \mathcal{C}_t(a)\}. \tag{11}$$

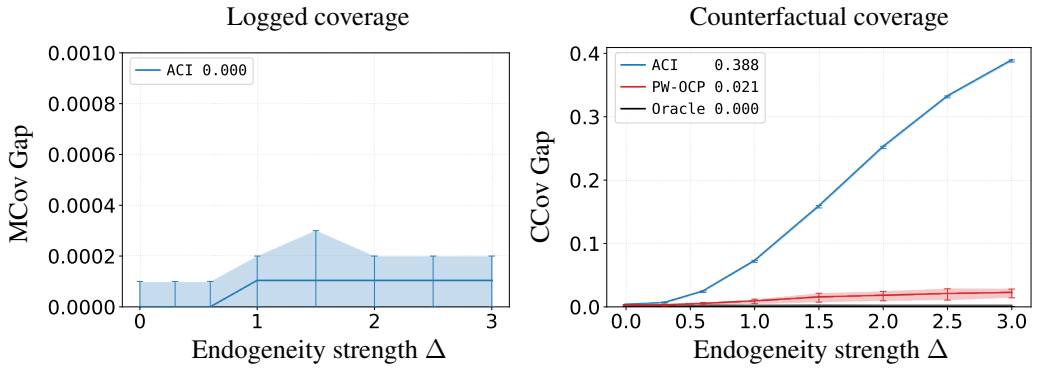

Figure 1: **Logged coverage can hide counterfactual failure.** Both panels sweep the mean-shift strength $\Delta \in [0, 3]$ at $T = 20{,}000$ and $\alpha = 0.1$. The left panel shows that ACI's logged marginal gap stays essentially zero on a 10^{-3} scale. The right panel shows that ACI's counterfactual coverage gap increases with Δ, whereas PW-OCP remains close to Oracle. This illustrates the failure mode described by Theorem 3.2 and the effect of propensity weighting. Bands are 95% confidence intervals over 10 seeds.

Lemma 4.2 (Unbiasedness)**.** *Fix $a \in \mathcal{A}$ and assume $\mathcal{C}_t(a)$ is $\sigma(\mathcal{H}_{t-1}, X_t)$-measurable. Under Assumptions 2.1–2.2,*

$$\mathbb{E}[Z_t(a) \mid X_t, \mathcal{H}_{t-1}] \;=\; \mathbb{P}\big(Y_t(a) \notin \mathcal{C}_t(a) \mid X_t, \mathcal{H}_{t-1}\big). \tag{12}$$

The propensity weight cancels the logger's selection probability under no unmeasured confounding (Assumption 2.1), so $Z_t(a)$ is an unbiased estimator of the round-t counterfactual miscoverage probability. The proof is given in Appendix D.3.

4.2 Algorithm

PW-OCP maintains one miscoverage index $\alpha_t(a) \in \mathbb{R}$ per action. Let $\mathcal{S}_t(a) := \{s(X_\tau, Y_\tau; a) : \tau < t,\ a_\tau = a\}$ be the logged score set for action a, and let $\hat{Q}_t^{(a)}$ denote its empirical quantile under the boundary convention (6). If $\mathcal{S}_t(a) = \emptyset$, set $\hat{Q}_t^{(a)}(u) = +\infty$ for $0 < u < 1$ and retain (6) otherwise. The threshold and recursion are

$$\hat{q}_t(a) := \hat{Q}_t^{(a)}\big(1 - \alpha_t(a)\big), \qquad \alpha_{t+1}(a) := \alpha_t(a) + \gamma\big(\alpha - Z_t(a)\big), \tag{13}$$

where $Z_t(a)$ is the IPW signal in Definition 4.1, evaluated at the prediction set $\mathcal{C}_t(a) := \{y : s(X_t, y; a) \le \hat{q}_t(a)\}$. Each threshold is computed from logged scores for the corresponding action, whereas the recursion is driven by the unbiased IPW estimator of counterfactual miscoverage from Lemma 4.2. Scores are stored in sorted per-action buffers, with $O(n_a)$ insertion for n_a stored scores. Algorithm 1 summarizes the online procedure. The DR-OCP variant, the Online COPP baseline, and nuisance-estimation details are deferred to Appendix B.

4.3 Doubly Robust Calibration

The conditional variance of the IPW signal $Z_t(a)$ is bounded in terms of the inverse propensity floor, so the signal can be noisy when $\pi_{\min}$ is small. It also requires the true logging propensity. Augmenting the IPW signal with an outcome model yields a doubly robust signal whose conditional bias is second order in the nuisance errors. Let

$$\mu_t^*(a) := \mathbb{P}\big(Y_t(a) \notin \mathcal{C}_t(a) \mid X_t, \mathcal{H}_{t-1}\big) \tag{14}$$

denote the round-t counterfactual miscoverage probability, which is the target identified by (12). Let $\hat{\mu}_t(a, X_t) \in [0, 1]$ be a $\sigma(\mathcal{H}_{t-1}, X_t)$-measurable estimator of $\mu_t^*(a)$, and let $\hat{\pi}_t(a \mid X_t, \mathcal{H}_{t-1}) \in [\hat{\pi}_{\min}, 1]$ be a $\sigma(\mathcal{H}_{t-1}, X_t)$-measurable propensity estimator with floor $\hat{\pi}_{\min} > 0$.

Definition 4.3 (DR error signal)**.**

$$Z_t^{\mathrm{DR}}(a) := \hat{\mu}_t(a, X_t) + \frac{\mathbf{1}\{a_t = a\}}{\hat{\pi}_t(a \mid X_t, \mathcal{H}_{t-1})}\big(\mathbf{1}\{Y_t \notin \mathcal{C}_t(a)\} - \hat{\mu}_t(a, X_t)\big). \tag{15}$$

Algorithm 1 PW-OCP: Propensity-Weighted Online Conformal Prediction

Input: target miscoverage α, step size γ, logging propensities π_t
Initialize $\alpha_1(a) \leftarrow \alpha$ and $\mathcal{S}_1(a) \leftarrow \emptyset$ for all $a \in \mathcal{A}$
for $t = 1, \ldots, T$ **do**
  // Construct action-wise prediction sets
  Observe X_t
  for each $a \in \mathcal{A}$ **do**
    $\hat{q}_t(a) \leftarrow \hat{Q}_t^{(a)}(1 - \alpha_t(a))$
    $\mathcal{C}_t(a) \leftarrow \{y : s(X_t, y; a) \le \hat{q}_t(a)\}$
  end for
  // Observe logged feedback
  Draw $a_t \sim \pi_t(\cdot \mid X_t, \mathcal{H}_{t-1})$ and observe Y_t
  $\mathcal{S}_{t+1}(a_t) \leftarrow \mathcal{S}_t(a_t) \cup \{s(X_t, Y_t; a_t)\}$
  $\mathcal{S}_{t+1}(a) \leftarrow \mathcal{S}_t(a)$ for all $a \neq a_t$
  // Propensity-weighted calibration
  for each $a \in \mathcal{A}$ **do**
    $p_t(a) \leftarrow \pi_t(a \mid X_t, \mathcal{H}_{t-1})$
    $Z_t(a) \leftarrow \frac{\mathbf{1}\{a_t = a\}}{p_t(a)} \cdot \mathbf{1}\{s(X_t, Y_t; a_t) > \hat{q}_t(a_t)\}$
    $\alpha_{t+1}(a) \leftarrow \alpha_t(a) + \gamma(\alpha - Z_t(a))$
  end for
end for

DR-OCP replaces $Z_t(a)$ in (13) with $Z_t^{\mathrm{DR}}(a)$. The construction reduces to PW-OCP when $\hat{\mu}_t \equiv 0$ and $\hat{\pi}_t \equiv \pi_t$. Theorem 5.2 shows that the residual conditional bias equals $(\hat{\mu}_t - \mu_t^*)(\hat{\pi}_t - \pi_t)/\hat{\pi}_t$. Algorithm 2 in Appendix B.1 gives the implementation.

5 Theoretical Analysis

This section establishes counterfactual coverage guarantees for PW-OCP and DR-OCP and a matching information-theoretic lower bound for nondegenerate calibration. All proofs are in Appendix E.

5.1 Coverage Guarantees

Theorem 5.1 (Counterfactual coverage of PW-OCP)**.** *Under Assumptions 2.1–2.2, Algorithm 1 (recursion (13)) satisfies, for every $a \in \mathcal{A}$ and every sample path, the telescoping identity*

$$\frac{1}{T}\sum_{t=1}^{T} Z_t(a) - \alpha \;=\; \frac{\alpha_1(a) - \alpha_{T+1}(a)}{\gamma T}. \tag{16}$$

With $K = |\mathcal{A}|$ and $\gamma = c_\gamma\sqrt{\pi_{\min}/T}$ for fixed $c_\gamma > 0$, the maximal envelope in Lemma D.1 gives, with probability at least $1 - \delta$,

$$\max_a |\mathrm{CCov}_T(a) - (1-\alpha)| \le C_{c_\gamma}\left\{\sqrt{\frac{\log(2K/\delta)}{T\pi_{\min}}} + \frac{\log(2K/\delta)}{T\pi_{\min}}\right\}. \tag{17}$$

The IPW update restores the counterfactual error signal. The price of observing each action with probability at least $\pi_{\min}$ enters as $(T\pi_{\min})^{-1/2}$. The telescoping identity (16) is exact and requires no clipping or projection of the recursion. The proof is in Appendix E.1.

Theorem 5.2 (Counterfactual coverage of DR-OCP)**.** *Run Algorithm 2 (DR signal (15)) with $\sigma(\mathcal{H}_{t-1}, X_t)$-measurable nuisance estimates $\hat{\mu}_t(a, X_t) \in [0,1]$ and $\hat{\pi}_t(a \mid X_t, \mathcal{H}_{t-1}) \in [\hat{\pi}_{\min}, 1]$, $\hat{\pi}_{\min} > 0$, and suppose there is a deterministic constant $\kappa \ge 1$ such that, almost surely,*

$$\sup_{t,a} \frac{\pi_t(a \mid X_t, \mathcal{H}_{t-1})}{\hat{\pi}_t(a \mid X_t, \mathcal{H}_{t-1})} \le \kappa. \tag{18}$$

When propensities are estimated, assume that $\mathbb{P}(Y_t(a) \notin \mathcal{C}_t(a) \mid \mathcal{F}_{t-1}, X_t) = \mu_t^*(a)$, *where* $\mathcal{F}_{t-1}$ *also contains past potential outcomes. This condition is unnecessary when* $\hat{\pi}_t \equiv \pi_t$. *Define the realized nuisance RMSEs against* $\mu_t^*(a)$ *in (14) and* $\pi_t(a \mid X_t, \mathcal{H}_{t-1})$,

$$\epsilon_{\mu,a} := \sqrt{\frac{1}{T}\sum_{t=1}^{T}\big(\hat{\mu}_t(a, X_t) - \mu_t^*(a)\big)^2}, \quad \epsilon_{\pi,a} := \sqrt{\frac{1}{T}\sum_{t=1}^{T}\big(\hat{\pi}_t(a|X_t,\mathcal{H}_{t-1}) - \pi_t(a|X_t,\mathcal{H}_{t-1})\big)^2}. \tag{19}$$

Then, under Assumptions 2.1–2.2, the unprojected recursion satisfies, with probability at least $1-\delta$,

$$\max_a |\mathrm{CCov}_T(a) - (1-\alpha)| \le \frac{1}{\gamma T} + 2\max_a \frac{\epsilon_{\mu,a}\epsilon_{\pi,a}}{\hat{\pi}_{\min}} + C\left\{\sqrt{\frac{\kappa \log(2K/\delta)}{T\hat{\pi}_{\min}}} + \frac{\log(2K/\delta)}{T\hat{\pi}_{\min}}\right\}. \tag{20}$$

The constant C *is absolute. If* $\hat{\pi}_t \equiv \pi_t$ *and* $\gamma = c_\gamma\sqrt{\pi_{\min}/T}$, *bounds (17) and (EC.58) also hold for DR-OCP with any predictable* $\hat{\mu}_t \in [0,1]$.

The product term $\epsilon_{\mu,a}\epsilon_{\pi,a}/\hat{\pi}_{\min}$ is the standard doubly robust bias and vanishes when either nuisance is consistent. When $\hat{\pi}_t \equiv \pi_t$, the DR signal equals the PW-OCP signal plus a control variate that can reduce variance. Setting $\hat{\mu}_t \equiv \mu_t^*$ removes the first-order bias from propensity estimation. See Appendix E.2.

Remark 5.3 (Scope of the DR guarantee). Theorem 5.2 gives a joint high-probability bound expressed in terms of the realized nuisance errors $\epsilon_{\mu,a}, \epsilon_{\pi,a}$, following the standard semiparametric convention [e.g., 15, 3]. The online nuisance protocol used in the experiments is in Appendix B.3. We do not claim a general cross-fitting rate for arbitrary adaptive logging. Establishing such rates under endogenous online data remains open.

5.2 Information-Theoretic Lower Bound

The comparison class excludes persistently uninformative thresholds through an expectation condition.

Theorem 5.4 (Minimax lower bound for nondegenerate calibration)**.** *Let* $\mathcal{P}$ *be the sequential model class of Section 2, including contextual and history-dependent logging. Its submodel* $\mathcal{P}_G$ *has independent Gaussian potential outcomes* $Y_t(a) \sim \mathcal{N}(0, \sigma_a^2)$ *across rounds and actions,* $\sigma_a \in [1, 5/4]$, *score* $s(y;a) = |y|$, *and context-free loggers* $\pi_t(a \mid \mathcal{H}_{t-1}) \ge \pi_{\min}$ *that may depend on the observed history. Fix* $\alpha \in (0,1)$, *let* q_α *be the* $(1-\alpha)$ *quantile of* $|\mathcal{N}(0,1)|$, *and set* $J_\alpha = [q_\alpha/2, 2q_\alpha]$. *Let* $\mathfrak{A}_{\mathrm{nd}}$ *be the predictable calibrators based on logged data satisfying*

$$\inf_{(P,\pi)\in\mathcal{P}_G} \min_a \mathbb{E}_{P,\pi} L_T(a) \ge \frac{2}{3}, \qquad L_T(a) := \frac{1}{T}\sum_{t=1}^{T} \mathbf{1}\{\hat{q}_t(a) \in J_\alpha\}. \tag{21}$$

For $K \ge 4$, $T \ge 1$, *and* $\pi_{\min} \in (0, 1/K]$,

$$\inf_{\mathcal{M}\in\mathfrak{A}_{\mathrm{nd}}} \sup_{(P,\pi)\in\mathcal{P}} \mathbb{E}\max_a |\mathrm{CCov}_T(a) - (1-\alpha)| \ge c(\alpha)\min\left\{1, \sqrt{\frac{\log K}{T\pi_{\min}}}\right\}. \tag{22}$$

The same lower bound already holds with the supremum restricted to $\mathcal{P}_G$.

Condition (21) allows infinite thresholds. Its interval and fraction define the comparison class, rather than algorithm inputs. The condition is imposed only on $\mathcal{P}_G$. Appendix E.3 extends the result to any fixed compact interval $[l,u] \subset (0,\infty)$ with $l < u$ and any fixed expected fraction in $(0,1)$.

Corollary 5.5 (Minimax rate of PW-OCP and DR-OCP)**.** *Algorithm 1 and DR-OCP with known propensities and any predictable* $\hat{\mu}_t \in [0,1]$, *both using* $\gamma = c_\gamma\sqrt{\pi_{\min}/T}$ *for fixed* $c_\gamma > 0$, *have expected worst-action coverage error* $O(\min\{1, \sqrt{\log(2K)/(T\pi_{\min})}\})$ *uniformly over* $\mathcal{P}$. *Both belong to* $\mathfrak{A}_{\mathrm{nd}}$ *whenever*

$$T\pi_{\min} \ge C_{\alpha,c_\gamma}. \tag{23}$$

Consequently, their worst-case expected coverage rate matches (22) up to constants in this regime. No threshold truncation or change to empty-buffer predictions is needed. The coverage upper bound itself holds for all T.

Sharpness and decisions. Coverage alone is not a useful endpoint if the learner attains it by returning uninformative sets. Appendix C makes this distinction formal. Under stationary logging and local score-density regularity, Theorem C.3 shows that projected PW-OCP also controls the average threshold error relative to the oracle counterfactual quantile after an initial sampling period, at rate $(T\pi_{\min})^{-1/4}$ up to logarithmic factors. Corollary C.7 shows that logged-quantile calibrators retain an $\Omega(\rho)$ threshold error on the impossibility instance in expectation when their tracking error is $o(\rho)$, while Theorem C.5 decomposes robust pseudo-regret into sharpness, logged coverage, and exploration terms. Proposition C.9 extends these sharpness guarantees to unprojected PW-OCP and DR-OCP with known propensities. Proposition C.10 gives an adaptive-logging bound on excess oracle robust risk and a stationary specialization using these sharpness rates.

6 Experiments

We evaluate counterfactual coverage, prediction-set quality, and all-action excess loss under selection-biased logging. PW-OCP and DR-OCP use the coverage-motivated step size $\gamma^\star \asymp \sqrt{\pi_{\min}/T}$ from Theorem 5.1. Full setup details and pseudocode are in Appendices F and B.

Methods and benchmarks. Baselines are ACI [9], FACI [10], AgACI [36], Conformal PID [1], SAOCP [4], NEx-CP [3], ECI [33], and Online COPP [25, 37] (Algorithm 3). Our methods are PW-OCP and DR-OCP; Oracle calibrates directly on counterfactual outcomes and is shown only as an unattainable reference. Synthetic benchmarks cover mean shift, variance shift, multi-arm scaling, propensity misspecification, convergence rate, and prediction-set quality. We construct semi-synthetic benchmarks from ZOZOTOWN Open Bandit Women/Men data [23] and evaluate financial rebalancing on DJIA returns with momentum-biased logging.

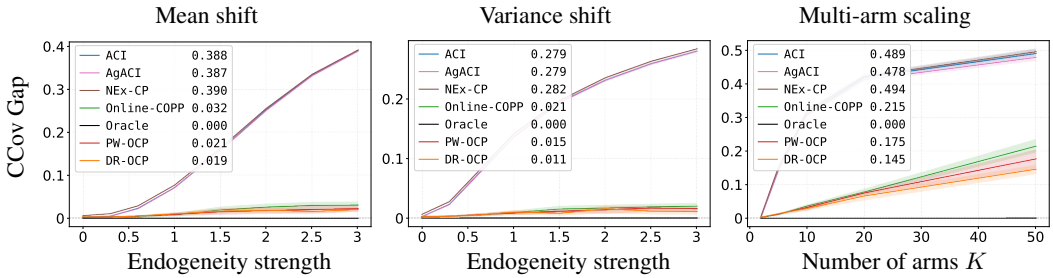

Figure 2: **Counterfactual coverage gap across three synthetic stress tests.** Mean-shift and variance-shift simulations sweep endogeneity at $K = 2$: in the first panel Δ is a mean shift, and in the second it is a standard-deviation increment. The multi-arm panel sweeps $K \in \{2, 5, 10, 20, 50\}$. Standard online CP baselines retain substantial counterfactual gaps, whereas PW-OCP and DR-OCP remain closer to Oracle across these settings. Bands are 95% confidence intervals over 10 seeds.

Table 1: **Master results across all benchmarks.** Columns report counterfactual coverage gaps (lower is better) and all-action excess-loss reduction in the mean-shift experiment at $\Delta = 3$, normalized so ACI is 0% and Oracle is 100% (higher is better). Entries report means with 95% confidence intervals over 10 seeds. Red and blue mark the best and second-best non-oracle methods.

Method	Mean shift	Variance shift	Multi-arm	Propensity error	OBD Women	OBD Men	DJIA	Loss vs ACI
ACI [9]	0.252 ± 0.002	0.230 ± 0.002	0.418 ± 0.009	0.252 ± 0.004	0.023 ± 0.005	0.025 ± 0.005	0.024 ± 0.003	—
FACI [10]	0.275 ± 0.001	0.250 ± 0.002	0.503 ± 0.009	0.275 ± 0.004	0.041 ± 0.006	0.043 ± 0.005	0.055 ± 0.007	-7.0% ± 1.1%
AgACI [36]	0.250 ± 0.001	0.230 ± 0.002	0.411 ± 0.007	0.251 ± 0.004	0.027 ± 0.005	0.033 ± 0.008	0.024 ± 0.003	+0.5% ± 1.1%
Conformal PID [1]	0.290 ± 0.002	0.261 ± 0.002	0.689 ± 0.003	0.251 ± 0.004	0.245 ± 0.005	0.262 ± 0.011	0.024 ± 0.004	-20.5% ± 1.2%
SAOCP [4]	0.269 ± 0.004	0.245 ± 0.002	0.450 ± 0.008	0.270 ± 0.005	0.042 ± 0.003	0.047 ± 0.006	0.074 ± 0.004	-4.5% ± 1.4%
NEx-CP [3]	0.255 ± 0.002	0.234 ± 0.002	0.422 ± 0.007	0.257 ± 0.003	0.025 ± 0.005	0.028 ± 0.005	0.024 ± 0.007	-1.0% ± 1.1%
ECI [33]	0.369 ± 0.001	0.327 ± 0.001	0.607 ± 0.004	0.369 ± 0.003	0.122 ± 0.005	0.121 ± 0.006	0.027 ± 0.005	-35.0% ± 1.0%
Online COPP [25, 37]	0.027 ± 0.008	0.018 ± 0.005	0.079 ± 0.010	0.256 ± 0.005	0.043 ± 0.010	0.052 ± 0.011	0.039 ± 0.009	+94.9% ± 0.9%
PW-OCP (ours)	0.017 ± 0.007	0.013 ± 0.006	0.074 ± 0.007	0.252 ± 0.002	0.023 ± 0.003	0.028 ± 0.005	0.019 ± 0.004	+91.1% ± 1.8%
DR-OCP (ours)	0.018 ± 0.006	0.014 ± 0.006	0.065 ± 0.012	0.030 ± 0.008	0.021 ± 0.003	0.024 ± 0.003	0.019 ± 0.004	+93.5% ± 1.9%
Oracle	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.007 ± 0.000	+100.0%

Counterfactual coverage. Figure 2 and Table 1 show that exogenous online CP baselines meet their logged targets while failing counterfactual coverage. With mean-shift magnitude $\Delta = 2$ and $K = 2$

actions, baseline counterfactual gaps cluster in $[0.25, 0.37]$, whereas PW-OCP reduces the gap to 0.017 ± 0.007 and DR-OCP to 0.018 ± 0.006. In the variance-shift simulation, PW-OCP and DR-OCP achieve gaps near 0.01 against ≥ 0.23 for standard baselines. At $K = 20$ arms with $\pi_{\min} = 0.005$, ACI has gap 0.418 ± 0.009, compared with 0.074 ± 0.007 for PW-OCP and 0.065 ± 0.012 for DR-OCP. The larger residual gap is consistent with the $1/\sqrt{T\pi_{\min}}$ rate in Theorem 5.1. At smaller exploration floors, Online COPP can outperform DR-OCP in finite samples, as shown in Appendix F.4.

Propensity robustness. This experiment isolates the contribution of the doubly robust augmentation by feeding the learner increasingly biased logging probabilities. The propensity-misspecification column of Table 1 reports the extreme case (uniform propensity model): PW-OCP collapses to the ACI gap, while DR-OCP retains gap 0.030 ± 0.008. Figure 3 sweeps the misspecification level ξ from 0 (correct) to 1 (uniform); DR-OCP remains comparatively stable, consistent with the product-form nuisance-error bound in Theorem 5.2.

AgACI does not use propensities and is unchanged by this corruption.

Prediction-set quality, rates, and excess loss. Figure 3 shows decreasing counterfactual coverage gaps and finite-threshold errors as T grows. Table 2 shows that PW-OCP and DR-OCP achieve near-oracle mean widths. Appendix Figure 5 independently varies T and $\pi_{\min}$ and confirms the effective-sample-size dependence predicted by Theorem 5.1. The last column of Table 1 reports all-action excess-loss reduction under a shared logger, as defined in Appendix F. Online COPP gives the largest reduction, while PW-OCP and DR-OCP also achieve large loss reductions in this mean-shift experiment.

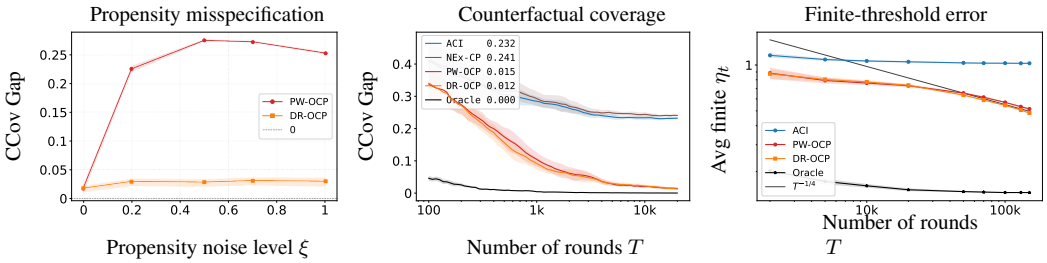

Figure 3: **Robustness, rates, and prediction-set quality. Left**: propensity misspecification is induced by giving the learner $\hat{\pi} = (1-\xi)\pi + \xi/K$, where $\xi = 0$ corresponds to the true propensities and $\xi = 1$ to uniform propensities. PW-OCP degrades as the supplied propensities become less accurate, while DR-OCP remains comparatively stable, consistent with the product-form nuisance-error bound in Theorem 5.2. **Middle**: as the horizon T grows, the counterfactual coverage gaps of PW-OCP and DR-OCP decrease toward the Oracle gap, whereas logged-calibration baselines retain persistent bias. **Right**: the average maximum oracle-quantile error over finite thresholds decreases with T. Rounds with no finite thresholds are excluded. The $T^{-1/4}$ curve provides a reference slope.

PW-OCP and DR-OCP produce intervals with near-oracle widths despite logged overcoverage, which also occurs for Oracle. ACI's narrower sets instead under-cover counterfactual outcomes. When prediction sets determine a softmin logger, Appendix F.6 shows a similar counterfactual coverage advantage for PW-OCP and DR-OCP in the heteroscedastic setting.

Table 2: **Logged coverage and interval efficiency.** $K = 5$, nominal coverage 0.90, and evaluation after $1{,}000$ rounds. Entries are means over 20 seeds, with 95% confidence intervals for relative width and counterfactual gap. Width, gap, and interval-score excess use finite sets. All PW-OCP and DR-OCP prediction sets in the evaluation period are finite.

Method	Logged coverage	Width / oracle	CCov gap	Score excess
ACI [9]	0.900	71.9% ± 0.3%	0.2330 ± 0.0032	1.77
NEx-CP [3]	0.895	71.5% ± 0.3%	0.2393 ± 0.0032	1.99
Online COPP [25, 37]	0.972	99.6% ± 0.4%	0.0155 ± 0.0021	0.03
PW-OCP (ours)	0.969	100.6% ± 0.4%	0.0248 ± 0.0033	0.28
DR-OCP (ours)	0.970	101.6% ± 0.4%	0.0213 ± 0.0027	0.29
Oracle	0.973	100.0%	0.00033 ± 0.00004	0.00

Open Bandit and financial benchmarks. On Open Bandit and DJIA, counterfactual gaps are smaller than in the synthetic stress tests, and several baselines already achieve near-nominal counterfactual coverage. PW-OCP and DR-OCP nonetheless remain among the best non-oracle methods across Women, Men, and financial rebalancing. Appendix Figure 7 shows counterfactual coverage for these tasks. Logged $\mathrm{MCov}, \mathrm{ACov}$ diagnostics for the synthetic benchmarks are provided in Appendix Figure 4. When selection bias is small, weighting can incur a variance cost. With uniform supplied propensities, the OBD Men gap rises from 0.0270 to 0.2284 for PW-OCP but only from 0.0255 to 0.0372 for DR-OCP.

7 Related Work

Online CP under distribution shift. Adaptive conformal inference [9, 10] and related methods, including conformal PID [1], SAOCP [4], AgACI [36], ECI [33], and beyond-exchangeable CP [3], deliver marginal coverage under exogenous, learner-independent distribution shift by tracking the realized error stream. None of these methods detects the selection bias created when prediction sets drive the logging policy: the stream they calibrate on is itself policy-biased. Our work provides counterfactual coverage for every action, including those rarely selected, through an IPW recursion that decouples the calibration target from the logged context distribution.

Partial-feedback CP. Recent work on partial-feedback conformal prediction [8, 32, 34] studies calibration with restricted or selectively observed outcome feedback. Our focus is the distinction between calibration on the selected context distribution and coverage under the marginal context distribution. Context-dependent logging can separate these targets even when $\rho = 0$. PW-OCP and its doubly robust extension correct the selection bias through the recorded propensities.

Counterfactual, batch, and decision-theoretic CP. Counterfactual coverage has been established for batch treatment effects [15] and for off-policy conformal prediction from logged batch data [25, 37], but these methods require a fixed offline sample and do not adapt as the logging policy evolves. Decision-theoretic conformal methods place prediction sets inside downstream optimization [14, 16, 17, 13, 35, 38, 2]. These works assume calibration data are exogenous and do not analyze the feedback loop in which predictions shape the data. Building on IPW [12] and doubly robust estimation [5], we establish sequential counterfactual coverage under adaptive logging, together with an impossibility result and a lower bound for nondegenerate calibration matching the expected coverage rate. Related work on conformal e-prediction [26, 28, 30, 7, 20], adversarial coverage objectives [19, 24], feedback-loop CP [29], and performative distribution shift [18] addresses other forms of robustness, but none isolates action-induced selection bias on counterfactual coverage.

8 Discussion and Limitations

This paper identifies endogenous logging as an obstacle for online conformal prediction. Marginal online validity can survive adaptive logging, yet logged-quantile per-action calibration retains a persistent $\Omega(\rho)$ counterfactual gap. PW-OCP removes this gap, with a coverage rate matching the minimax benchmark for nondegenerate calibration, and DR-OCP reduces nuisance sensitivity. The experiments show the same separation in finite samples: methods that appear calibrated on logged outcomes can remain unreliable on rarely selected actions, whereas propensity-weighted calibration reduces counterfactual gaps in the selection-biased settings. The decision decompositions characterize how prediction-set accuracy can support downstream decisions.

Limitations and open directions. The framework addresses contextual bandits rather than reinforcement learning: it identifies one-step potential outcomes at the realized history, not hypothetical action-sequence trajectories. Our coverage guarantees rely on positivity, and maintaining an exploration floor may be costly. Explicit sharpness and decision rates require stationary logging and local score-density regularity. The decision decompositions permit adaptive logging. The DR-OCP bound with estimated propensities depends on the realized nuisance errors. Exact pivotal normalization can invalidate the construction of Theorem 3.2 even when $\rho > 0$. Whether an analogous impossibility result holds for general normalized scores remains open. Open problems include diagnostics for ρ, adaptive cross-fitting under endogenous streams, tighter coverage–regret lower bounds beyond Theorem C.4, adaptive exploration budgets, delay-adapted step sizes, and anytime counterfactual CP via conformal e-processes [26, 28, 30, 7, 20].

Acknowledgments and Disclosure of Funding

This research was supported by the National Science Foundation (NSF) under grant DMS 1820702 and the National Natural Science Foundation of China (NSFC) under grants W2441021, 72371172, 72342023, and 71929101.

References

[1] Anastasios N Angelopoulos, Emmanuel J Candès, and Ryan J Tibshirani. Conformal PID control for time series prediction. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

[2] Yilun Bao, Yachong Hu, Haojie Ren, and Changliang Zou. Optimal model selection for conformalized robust optimization. *arXiv preprint arXiv:2507.04716*, 2025.

[3] Rina Foygel Barber, Emmanuel J Candès, Aaditya Ramdas, and Ryan J Tibshirani. Conformal prediction beyond exchangeability. In *The Annals of Statistics*, volume 51, pages 816–845, 2023.

[4] Aadyot Bhatnagar, Hao Wang, Caiming Xiong, and Yu Bai. Improved online conformal prediction via strongly adaptive online learning. In *International Conference on Machine Learning (ICML)*, 2023.

[5] Miroslav Dudík, John Langford, and Lihong Li. Doubly robust policy evaluation and learning. *arXiv preprint arXiv:1103.4601*, 2011.

[6] David A Freedman. On tail probabilities for martingales. *The Annals of Probability*, 3(1):100–118, 1975.

[7] Etienne Gauthier, Francis Bach, and Michael I Jordan. E-values expand the scope of conformal prediction. *arXiv preprint arXiv:2503.13050*, 2025.

[8] Haosen Ge, Hamsa Bastani, and Osbert Bastani. Stochastic online conformal prediction with semi-bandit feedback. In *International Conference on Machine Learning (ICML)*, 2025.

[9] Isaac Gibbs and Emmanuel Candès. Adaptive conformal inference under distribution shift. *Advances in Neural Information Processing Systems*, 34:1660–1672, 2021.

[10] Isaac Gibbs and Emmanuel Candès. Conformal inference for online prediction with arbitrary distribution shifts. *Journal of Machine Learning Research*, 25(162):1–36, 2024.

[11] Tilmann Gneiting, Fadoua Balabdaoui, and Adrian E Raftery. Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society Series B*, 69(2):243–268, 2007.

[12] Daniel G Horvitz and Donovan J Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47(260):663–685, 1952.

[13] Yang Hu, Jieren Tan, Changliang Zou, Yajie Bao, and Haojie Ren. Conformal robustness control: A new strategy for robust decision. In *International Conference on Learning Representations (ICLR)*, 2026.

[14] Shayan Kiyani, George Pappas, Aaron Roth, and Hamed Hassani. Decision theoretic foundations for conformal prediction: Optimal uncertainty quantification for risk-averse agents. In *International Conference on Machine Learning (ICML)*, 2025.

[15] Lihua Lei and Emmanuel J Candès. Conformal inference of counterfactuals and individual treatment effects. *Journal of the Royal Statistical Society Series B*, 83(5):911–938, 2021.

[16] Jordan Lekeufack, Anastasios N Angelopoulos, Andrea V Bajcsy, and Michael I Jordan. Conformal decision theory: Safe autonomous decisions from imperfect predictions. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 11668–11675, 2024.

[17] Yash P Patel, Sahana Rayan, and Ambuj Tewari. Conformal contextual robust optimization. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 2485–2493, 2023.

[18] Juan Perdomo, Tijana Zrnic, Celestine Mendler-Dünner, and Moritz Hardt. Performative prediction. In *International Conference on Machine Learning (ICML)*, 2020.

[19] Ramya Ramalingam, Shayan Kiyani, and Aaron Roth. The relationship between no-regret learning and online conformal prediction. In *International Conference on Machine Learning (ICML)*, 2025.

[20] Aaditya Ramdas and Ruodu Wang. Hypothesis testing with e-values. *Foundations and Trends in Statistics*, 1(1–2):1–390, 2025.

[21] Yaniv Romano, Evan Patterson, and Emmanuel Candès. Conformalized quantile regression. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

[22] Donald B Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688–701, 1974.

[23] Yuta Saito, Shunsuke Aihara, Megumi Matsutani, and Yusuke Narita. A large-scale open dataset for bandit algorithms. *arXiv preprint arXiv:2008.07146*, 2020.

[24] Yusuf Sale and Aaditya Ramdas. Online selective conformal prediction: Errors and solutions. *Transactions on Machine Learning Research (TMLR)*, 2025.

[25] Muhammad Faaiz Taufiq, Jean-François Ton, Rob Cornish, Yee Whye Teh, and Arnaud Doucet. Conformal off-policy prediction in contextual bandits. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

[26] Vladimir Vovk. Conformal e-prediction. *Pattern Recognition*, 166:111674, 2025.

[27] Vladimir Vovk, Alex Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, 2005.

[28] Vladimir Vovk and Ruodu Wang. Conformal e-prediction in the presence of confounding. *arXiv preprint arXiv:2603.11134*, 2026.

[29] Han Wang and Chao Ning. Conformal prediction in the loop: A feedback-based uncertainty model for trajectory optimization. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

[30] Hongjian Wang, Sanjit Dandapanthula, and Aaditya Ramdas. Anytime-valid FDR control with the stopped e-BH procedure. *Statistics & Probability Letters*, 219:110512, 2025.

[31] Yu-Xiang Wang, Alekh Agarwal, and Miroslav Dudík. Optimal and adaptive off-policy evaluation in contextual bandits. In *International Conference on Machine Learning (ICML)*, pages 3589–3597, 2017.

[32] Zhou Wang and Xingye Qiao. Generalized prediction set with bandit feedback. *Transactions on Machine Learning Research (TMLR)*, 2025.

[33] Junxi Wu, Dongjian Hu, Yajie Bao, Shu-tao Xia, and Changliang Zou. Error-quantified conformal inference for time series. In *The Thirteenth International Conference on Learning Representations*, 2025.

[34] Junyoung Yang, Kyungmin Kim, and Sangdon Park. Online conformal prediction with adversarial semi-bandit feedback via regret minimization. In *International Conference on Learning Representations (ICLR)*, 2026.

[35] Christopher Yeh, Nicolas Christianson, Adam Wierman, and Yisong Yue. Conformal risk training. *arXiv preprint arXiv:2510.08748*, 2025.

[36] Margaux Zaffran, Olivier Féron, Yannig Goude, Julie Josse, and Aymeric Dieuleveut. Adaptive conformal predictions for time series. In *International Conference on Machine Learning (ICML)*, pages 25834–25866, 2022.

[37] Yingying Zhang, Chengchun Shi, and Shikai Luo. Conformal off-policy prediction. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 2751–2768, 2023.

[38] Wenbin Zhou and Shixiang Zhu. Calibrating decision robustness via inverse conformal risk control. *arXiv preprint arXiv:2510.07750*, 2025.

A Broader impacts

Reliable counterfactual coverage benefits decision systems where rarely selected actions matter, such as clinical trials, dynamic pricing, and recommendation. Without validation, miscalibrated propensities or vanishing exploration produce biased prediction sets on underrepresented contexts. In deployment, practitioners should audit propensities, prefer the doubly robust variant when propensity estimates are uncertain, maintain an exploration floor, and monitor counterfactual coverage using propensity-weighted estimates or additional outcome collection.

B Additional Algorithms and Implementation

B.1 Doubly Robust Online Conformal Prediction

DR-OCP is the operational version of Definition 4.3. It uses the same score buffers and action-wise threshold construction as PW-OCP, but fits nuisance models on past data and evaluates them at the current context before observing Y_t. After the played outcome is revealed, each action receives an imputed miscoverage estimate $\hat{\mu}_t(a, X_t)$ plus an inverse-propensity residual when that action was actually played. This step removes first-order sensitivity to either the propensity or outcome model, as shown in Theorem 5.2.

Algorithm 2 DR-OCP: Doubly Robust Online Conformal Prediction

Input: target α, step size γ, predictable estimators $\hat{\mu}_t, \hat{\pi}_t$
Initialize $\alpha_1(a) \leftarrow \alpha$ and $\mathcal{S}_1(a) \leftarrow \emptyset$ for all $a \in \mathcal{A}$
for $t = 1, \ldots, T$ **do**
  // Construct action-wise prediction sets
  Observe X_t and compute $\hat{\mu}_t(a, X_t)$, $\hat{\pi}_t(a \mid X_t, \mathcal{H}_{t-1})$
  for each $a \in \mathcal{A}$ **do**
    $\hat{q}_t(a) \leftarrow \hat{Q}_t^{(a)}(1 - \alpha_t(a))$
    $\mathcal{C}_t(a) \leftarrow \{y : s(X_t, y; a) \le \hat{q}_t(a)\}$
  end for
  // Observe logged feedback
  Draw $a_t \sim \pi_t(\cdot \mid X_t, \mathcal{H}_{t-1})$ and observe Y_t
  $E_t \leftarrow \mathbf{1}\{s(X_t, Y_t; a_t) > \hat{q}_t(a_t)\}$
  $\mathcal{S}_{t+1}(a_t) \leftarrow \mathcal{S}_t(a_t) \cup \{s(X_t, Y_t; a_t)\}$
  $\mathcal{S}_{t+1}(a) \leftarrow \mathcal{S}_t(a)$ for all $a \neq a_t$
  // Doubly robust calibration
  for each $a \in \mathcal{A}$ **do**
    $\hat{p}_t(a) \leftarrow \hat{\pi}_t(a \mid X_t, \mathcal{H}_{t-1})$
    $R_t(a) \leftarrow E_t - \hat{\mu}_t(a, X_t)$
    $Z_t^{\mathrm{DR}}(a) \leftarrow \hat{\mu}_t(a, X_t) + \frac{\mathbf{1}\{a_t = a\}}{\hat{p}_t(a)} R_t(a)$
    $\alpha_{t+1}(a) \leftarrow \alpha_t(a) + \gamma(\alpha - Z_t^{\mathrm{DR}}(a))$
  end for
end for

B.2 Online COPP Baseline

Online COPP is a natural online adaptation of batch off-policy conformal prediction [25, 37]: an IPW empirical quantile computed from a sliding window of the $W = 500$ most recent logged scores for each action. Unlike PW-OCP, it does not maintain an adaptive miscoverage index, which isolates the contribution of the online stochastic-approximation recursion in the comparison.

B.3 Nuisance Estimation for DR-OCP

Theorem 5.2 bounds coverage in terms of the realized nuisance RMSEs in (19), leaving the choice of estimators to the implementer. The experiments use the following protocol, which ensures that $\hat{\mu}_t$ and $\hat{\pi}_t$ are both measurable with respect to $\sigma(\mathcal{H}_{t-1}, X_t)$.

Algorithm 3 Online COPP baseline

Input: target miscoverage α, window size W, logging propensities π_t
Initialize score buffers $\mathcal{S}_1(a) \leftarrow \emptyset$ and weight buffers $\mathcal{W}_1(a) \leftarrow \emptyset$
for $t = 1, \ldots, T$ **do**
  // Weighted off-policy quantile
  Observe X_t
  for each $a \in \mathcal{A}$ **do**
    $\hat{q}_t(a) \leftarrow$ weighted empirical $(1-\alpha)$ quantile of $\mathcal{S}_t(a)$ using $\mathcal{W}_t(a)$
    $\mathcal{C}_t(a) \leftarrow \{y : s(X_t, y; a) \le \hat{q}_t(a)\}$
  end for
  // Observe logged feedback and update buffers
  Draw $a_t \sim \pi_t(\cdot \mid X_t, \mathcal{H}_{t-1})$ and observe Y_t
  Append $s(X_t, Y_t; a_t)$ to $\mathcal{S}_t(a_t)$
  Append $1/\pi_t(a_t \mid X_t, \mathcal{H}_{t-1})$ to $\mathcal{W}_t(a_t)$
  Keep the most recent W score–weight pairs for action a_t
end for

1. **Fit on past data only.** Before round t, fit $\hat{\mu}_t$ and $\hat{\pi}_t$ using $\mathcal{H}_{t-1}$ alone; the current outcome Y_t is never used.
2. **Propensity estimator.** When the logging rule is recorded, set $\hat{\pi}_t = \pi_t$. For tabular contexts, estimate the probabilities from past context-action counts with add-one smoothing, then project the vector onto $\{p : p_a \ge \hat{\pi}_{\min}, \sum_a p_a = 1\}$. Update counts only after observing the current action.
3. **Outcome model.** Regress the past miscoverage indicators $\mathbf{1}\{s(X_\tau, Y_\tau; a) > \hat{q}_\tau(a)\}$ on (X_τ, a) using only rounds with $a_\tau = a$, with pooled or hierarchical smoothing when an arm has few observations.
4. **Block updates.** Refit the nuisance models in time blocks rather than at every round; this stabilizes the residual term in $Z_t^{\mathrm{DR}}(a)$ from (15).

The protocol is an implementation device. The DR coverage bound (20) remains valid for any predictable nuisance sequence through the explicit product term $\max_a \epsilon_{\mu,a}\epsilon_{\pi,a}/\hat{\pi}_{\min}$.

With balanced logging and $T = 20{,}000$, median prediction-and-update time at $K = 10$ and 50 is $0.011/0.035$ ms for DR-OCP, $0.034/0.017$ ms for Online COPP, and $0.064/0.015$ ms for NEx-CP, over 3 seeds and 3 repetitions. The latter two benefit from shorter per-action buffers as K increases.

C Additional Theoretical Results

This appendix presents complementary results on prediction-set sharpness (Theorem C.3 and Corollary C.7), the tradeoff between coverage and regret (Theorem C.4), and the connection between prediction sets and downstream decisions (Theorem C.5 and Corollary C.6).

C.1 Prediction Set Sharpness

Coverage alone does not guarantee useful prediction sets, since the trivial set $\mathcal{Y}$ achieves perfect coverage [11]. We measure efficiency by the time-averaged threshold error relative to the oracle counterfactual quantile:

$$\eta_t := \max_{a \in \mathcal{A}} |\hat{q}_t(a) - q^*(a)|, \qquad q^*(a) := (F_{\mathrm{true}}^{(a)})^{-1}(1-\alpha), \tag{EC.1}$$

where $F_{\mathrm{true}}^{(a)}$ is the CDF of the score $s(X_t, Y_t(a); a)$ under the marginal context distribution. Bounding η_t requires stationary logging and density regularity (Assumption C.1).

Assumption C.1 (Stationary logging and quantile regularity)**. (a) Stationarity.** The sequence $(X_t, \{Y_t(a)\}_{a \in \mathcal{A}})_{t \ge 1}$ is i.i.d., and the logging policy is time-invariant: $\pi_t(a \mid X_t, \mathcal{H}_{t-1}) = \pi(a \mid X_t) \ge \pi_{\min}$ almost surely.

(b) Score densities. For each $a \in \mathcal{A}$, let $F^{(a)}_{\text{true}}$ and $F^{(a)}_{\text{bias}}$ denote the CDFs of $s(X_t, Y_t(a); a)$ under the marginal context distribution and conditional on $a_t = a$, respectively, with continuous densities $f^{(a)}_{\text{true}}$ and $f^{(a)}_{\text{bias}}$. Let $\alpha^\star(a) := 1 - F^{(a)}_{\text{bias}}(q^*(a))$ denote the population fixed point of the PW-OCP recursion under stationary logging.

(c) Local regularity. For each a, there exist a compact interval $\mathcal{I}_a \subset (0,1)$ with $\alpha^\star(a) \in \text{int}(\mathcal{I}_a)$ and a radius $\varepsilon_q > 0$ such that, defining

$$\mathcal{Q}_a := \{(F^{(a)}_{\text{bias}})^{-1}(1-\beta) : \beta \in \mathcal{I}_a\}, \qquad \widetilde{\mathcal{Q}}_a := \{q : \text{dist}(q, \mathcal{Q}_a) \le \varepsilon_q\}, \tag{EC.2}$$

the logged-score density satisfies $0 < f_{\min} \le f^{(a)}_{\text{bias}} \le f_{\max} < \infty$ on $\widetilde{\mathcal{Q}}_a$, and the density ratio is bounded:

$$0 < \lambda_{\min} \le \frac{f^{(a)}_{\text{true}}(q)}{f^{(a)}_{\text{bias}}(q)} \le \lambda_{\max} < \infty, \qquad q \in \widetilde{\mathcal{Q}}_a. \tag{EC.3}$$

Remark C.2 (Scope of Assumption C.1). Assumption C.1 is used by the sharpness results below, including Theorem C.3 and Proposition C.9, and by the decision bounds in Section C.2 that invoke these rates. The basic coverage guarantees in Theorems 5.1 and 5.2 allow history-dependent logging. The minimax comparison in Theorem 5.4 ranges over the full sequential model, with condition (21) imposed only on its Gaussian submodel. For sharpness, stationarity covers fixed randomized policies and time-invariant contextual logging. Extending these sharpness results to drifting policies requires additional control.

Theorem C.3 (Sharpness rate of projected PW-OCP)**.** *Run PW-OCP with each index projected onto $\mathcal{I}_a$ after the update and $\gamma = c_\gamma\sqrt{\pi_{\min}/T}$ for any fixed $c_\gamma > 0$. Under Assumptions 2.1–C.1, let $0 < \delta < 1/2$, $\Lambda = \log(2KT/\delta)$, and $t_0 = 1 + \lceil C_0\Lambda/\pi_{\min}\rceil$. With probability at least $1-\delta$,*

$$\frac{1}{T}\sum_{t=t_0}^{T} \eta_t \le C(T\pi_{\min})^{-1/4}\sqrt{\Lambda}. \tag{EC.4}$$

An empty sum is zero. The constants depend on the local regularity parameters, uniform index-interval margins, and c_γ, but not on $K, T, \pi_{\min}, \delta$. No score-tail assumption or change to empty-buffer predictions is required.

The startup time t_0 is an analysis quantity, not an algorithm input. If $t_0 \le T/2$, the average over rounds $t_0, \dots, T$ satisfies the same rate up to a factor of two. A full-horizon raw threshold bound additionally includes $T^{-1}\sum_{t<t_0}\eta_t$, which may be infinite under an empty-buffer convention. The decision analysis below instead bounds startup loss directly. The exponent $1/4$ comes from stochastic-approximation fluctuations, while the local quantile map is Lipschitz.

Theorem C.4 (Worst-case coverage–regret lower bound)**.** *Fix $\alpha \in (0,1)$. There exist constants $\delta_0(\alpha) > 0$ and $c(\alpha) > 0$ such that, for every target accuracy $\delta \in (0, \delta_0]$, the following holds. There exist two context-free two-arm instances P_0, P_1 with squared loss $\ell(a,y) = y^2$ such that, for any predictable threshold predictor using score $s(y;a) = |y|$, coupled with any decision rule that observes only the outcome of the played action and satisfies*

$$\min_{i\in\{0,1\}} \min_{a\in\{1,2\}} \mathbb{E}_i L_T(a) \ge \frac{2}{3}, \tag{EC.5}$$

with $L_T(a)$ defined in (21) for this coupled action process, the condition

$$\sup_{P\in\{P_0,P_1\}} \mathbb{E}_P\Big[\max_a |\text{CCov}_T(a) - (1-\alpha)|\Big] \le \delta \tag{EC.6}$$

implies

$$\delta \cdot \sup_{P\in\{P_0,P_1\}} \mathbb{E}_P[\text{Regret}_T] \ge c(\alpha), \tag{EC.7}$$

where $\text{Regret}_T := \sum_{t=1}^{T}\big(\ell(a_t, Y_t) - \min_{a\in\mathcal{A}}\mathbb{E}_P[\ell(a, Y_t(a))]\big)$ is the realized excess loss whose expectation is the standard pseudo-regret against the best fixed arm under P.

Theorem C.4 gives a worst-case regret lower bound proportional to $1/\delta$ under the same per-action nondegeneracy condition, quantifying the exploration cost of more accurate counterfactual calibration.

C.2 From Prediction Sets to Decisions

The following decomposition bounds decision regret using sharpness, logged coverage, and exploration. It permits adaptive logging, while the sharpness rates below use stationary logging.

Theorem C.5 (Sharpness–coverage decomposition)**.** *The loss may depend on the revealed context, with this dependence suppressed in $\ell(a, y)$ below. Its Lipschitz and range bounds are uniform in the context. Let $d_t(a) := \sup_{y \in \mathcal{C}_t(a)} \ell(a, y)$ be the worst-case loss inside $\mathcal{C}_t(a)$ and let $\bar{a}_t \in \arg\min_{a \in \mathcal{A}} d_t(a)$ be the robust exploitation action (ties broken arbitrarily). The learner plays $a_t = \bar{a}_t$ with probability $1 - \varepsilon$ and explores uniformly with probability ε; the played action is also the logged action. Assume:*

(i) $\ell(a, \cdot)$ *is L-Lipschitz on* $\mathcal{Y}$ *for* $L > 0$.

(ii) $\mathcal{C}_t(a)$ *is non-empty,* $d_t(a)$ *and* $d_t^*(a)$ *are finite, and* $\ell(a, y) - d_t(a) \le B$ *for all* $y \notin \mathcal{C}_t(a)$.

(iii) $d_t(a) - d_t(\bar{a}_t) \le \Delta_{\max}$ *for every action* a.

Let $\mathcal{C}_t^*(a)$ be the oracle prediction sets at level $1 - \alpha$, $d_t^*(a) := \sup_{y \in \mathcal{C}_t^*(a)} \ell(a, y)$, $a_t^* \in \arg\min_a d_t^*(a)$, and define the set discrepancy

$$\eta_t^{\text{set}} := \max_{a \in \mathcal{A}} d_H(\mathcal{C}_t(a), \mathcal{C}_t^*(a)), \tag{EC.8}$$

where d_H is the Hausdorff distance on $\mathcal{Y}$. The robust pseudo-regret

$$\widetilde{R}_T := \sum_{t=1}^{T} \big[\ell(a_t, Y_t(a_t)) - d_t^*(a_t^*)\big] \tag{EC.9}$$

satisfies

$$\mathbb{E}[\widetilde{R}_T] \;\le\; \underbrace{L \sum_{t=1}^{T} \mathbb{E}[\eta_t^{\text{set}}]}_{\text{sharpness}} + \underbrace{\alpha BT + BT\, \mathbb{E}|\mathrm{MCov}_T - (1-\alpha)|}_{\text{logged coverage}} + \underbrace{\varepsilon T \Delta_{\max}}_{\text{exploration}}. \tag{EC.10}$$

For a loss in $[0, M]$ and nonempty sets, conditions (ii) and (iii) hold with $B = \Delta_{\max} = M$, and the sharpness term in (EC.10) can be replaced by $\sum_t \mathbb{E} \min\{M, L\eta_t^{\text{set}}\}$. The capped sharpness term remains finite even when empty buffers yield infinite thresholds. A bounded loss on another interval uses the width of that interval. For unbounded losses such as squared loss under Gaussian outcomes, the theorem applies post-truncation, with the residual tail error controlled separately. The decomposition (EC.10) isolates three sources of decision loss. The exploration term is fixed by design and the logged-coverage term is controlled by any method satisfying Theorem 3.1. The sharpness term is where counterfactual calibration enters, since biased calibration produces persistently wide or narrow sets on rarely selected actions.

The bridge uses logged coverage for the realized action and counterfactual calibration through sharpness. For absolute-residual sets with the same centers, Hausdorff distance equals threshold distance, so

$$c_H \max_a |\hat{q}_t(a) - q^*(a)| \le \eta_t^{\text{set}} \le C_H \max_a |\hat{q}_t(a) - q^*(a)| \tag{EC.11}$$

holds with $c_H = C_H = 1$. Other score sets require the stated comparison.

For a stationary logger, write $w_a = \mathbb{P}(a_t = a)$ and define its oracle logged miscoverage

$$b^* := \sum_a w_a \{1 - F_{\text{bias}}^{(a)}(q^*(a))\}. \tag{EC.12}$$

This is a property of the instance, not an assumed coverage-transfer inequality. On the local quantile event in Theorem C.3, the existing logged-density bound gives

$$\sum_a w_a \{1 - F_{\text{bias}}^{(a)}(\hat{q}_t(a))\} \le b^* + f_{\max}\eta_t, \qquad t \ge t_0. \tag{EC.13}$$

No bound on densities conditional on individual contexts is needed. On the Gaussian construction of Theorem 3.2, $b^* \le \alpha$ follows directly from the logger and outcome laws.

Corollary C.6 (Comparison of decision regret)**.** *Assume Theorem C.5 and* (EC.11).

(a) Logged-quantile calibrators. *Define the bounded set error* $\bar{\eta}_t^{\mathrm{set}} := \min\{1, \eta_t^{\mathrm{set}}\}$. *On the instance of Theorem 3.2, any predictable thresholds with tracking error* ε_T *in* (9) *satisfy*

$$\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}\bar{\eta}_t^{\mathrm{set}}\right] \geq c'(\alpha)\rho - C_{\mathrm{tr}}\varepsilon_T. \tag{EC.14}$$

If the decision rule additionally satisfies

$$\mathbb{E}[\widetilde{R}_T] \geq c_{\mathrm{dec}} L \sum_t \mathbb{E}[\bar{\eta}_t^{\mathrm{set}}] - C_{\mathrm{dec}}/\gamma, \tag{EC.15}$$

then $\mathbb{E}[\widetilde{R}_T] \geq c_{\mathrm{dec}} LT\{c'(\alpha)\rho - C_{\mathrm{tr}}\varepsilon_T\} - C_{\mathrm{dec}}/\gamma$. *The constants are independent of* T *and* ρ *and depend only on the Gaussian family, the set comparison, and the stated decision condition. The cap is used only in the analysis and leaves predictions unchanged. Condition* (EC.15) *must be verified for the particular decision problem and does not follow from calibration bias alone. If* $\varepsilon_T = o(\rho)$ *and* $1/\gamma = o(\rho T)$, *it yields an* $\Omega(\rho T)$ *regret lower bound.*

(b) PW-OCP. *Consider the bounded-loss case* $\ell \in [0, M]$ *of Theorem C.5, with* $B = \Delta_{\max} = M$. *If the actual logger also satisfies Assumption C.1, projected PW-OCP with the step size of Theorem C.3 satisfies*

$$\mathbb{E}[\widetilde{R}_T] \leq Bb^*T + \varepsilon T\Delta_{\max} + O\left(T^{3/4}\pi_{\min}^{-1/4}\sqrt{\log(2KT)}\right). \tag{EC.16}$$

The implicit constant depends on the loss and local regularity parameters. On the Gaussian construction, $Bb^*T \leq \alpha BT$, *recovering that leading term without an additional oracle assumption. The bound includes all startup rounds and permits infinite empty-buffer thresholds.*

The stationary specialization in part (b) requires the robust decision rule and logger to describe the same action process. With deterministic tie-breaking and $\varepsilon < 1$, this requires the exploitation action to be a fixed function of the current context almost surely. A policy whose exploitation action changes with the calibration history instead uses the adaptive decomposition in Theorem C.5, without the stationary sharpness rate.

Corollary C.7 (Sharpness comparison)**.** *Let* t_0 *be as in Theorem C.3 and* $b_T = \min\{T, t_0 - 1\}/T$. *On the instance of Theorem 3.2, logged-quantile thresholds satisfy*

$$\mathbb{E}\left[\frac{1}{T}\sum_{t=t_0}^{T}\eta_t\right] \geq (1 - b_T)c_\eta(\alpha)\rho - C_{\mathrm{tr}}\varepsilon_T. \tag{EC.17}$$

Projected PW-OCP under Assumption C.1, with any fixed step-size constant $c_\gamma > 0$, *satisfies with probability at least* $1 - \delta$

$$\frac{1}{T}\sum_{t=t_0}^{T}\eta_t \leq C(T\pi_{\min})^{-1/4}\sqrt{\log(2KT/\delta)}. \tag{EC.18}$$

When $T\pi_{\min}/\log(2KT/\delta) \to \infty$, the startup fraction vanishes. Logged-quantile tracking with $\varepsilon_T = o(\rho)$ then retains a persistent threshold error, while PW-OCP has the stated high-probability rate. The regret lower bound additionally requires (EC.15). The bounded-loss upper bound uses localization directly, rather than asserting an unconditional moment bound for raw threshold error.

Corollary C.8 (Full-horizon counterfactual probability error)**.** *Under the assumptions and projected PW-OCP recursion of Theorem C.3, define*

$$\mathcal{D}_T := \frac{1}{T}\sum_{t=1}^{T}\max_a \left|F_{\mathrm{true}}^{(a)}(\hat{q}_t(a)) - (1 - \alpha)\right|. \tag{EC.19}$$

Then, for every $T \geq 1$,

$$\mathbb{E}[\mathcal{D}_T] \leq C(T\pi_{\min})^{-1/4}\sqrt{\log(2KT)}. \tag{EC.20}$$

The constant has the same permitted dependence as in Theorem C.3. This quantity averages absolute probability errors before any cancellation across rounds. It includes infinite-threshold predictions and averages over the current context distribution, rather than conditioning on each context value.

Proposition C.9 (Sharpness without projection)**.** *Under Assumption C.1, run unprojected PW-OCP or DR-OCP with known propensities and any predictable $\hat{\mu}_t \in [0,1]$, with $\alpha_1(a) = \alpha$ and $\gamma = c_\gamma\sqrt{\pi_{\min}/T}$. Write $N = T\pi_{\min}$ and $\Lambda = \log(2KT/\delta)$, where $0 < \delta < 1/2$. There are constants $A, C_0, B, C > 0$ such that, if $N \geq A\Lambda^2$, then with probability at least $1-\delta$,*

$$\frac{1}{T}\sum_{t=t_1}^{T}\eta_t \leq CN^{-1/4}\sqrt{\Lambda}, \qquad t_1 = 1 + \left\lceil \frac{C_0\Lambda}{\pi_{\min}} \right\rceil + \left\lceil \frac{B}{\gamma} \right\rceil. \tag{EC.21}$$

For every $T \geq 1$, the full-horizon probability error in (EC.19) satisfies

$$\mathbb{E}\mathcal{D}_T \leq C\min\{1, N^{-1/4}\sqrt{\log(2KT)}\}. \tag{EC.22}$$

Constants have the same permitted dependence as in Theorem C.3. The time t_1 is used only in the analysis. The algorithm retains its original buffers, initialization, and infinite-threshold convention.

Proposition C.10 (Excess oracle robust risk)**.** *Let the loss be L-Lipschitz for $L > 0$ and lie in $[0, M]$, uniformly in the context, and let the oracle sets be nonempty. Use the robust exploitation action $\bar{a}_t$ and uniform exploration probability ε from Theorem C.5. If a learned set is empty, allow any exploitation action and set $\eta_t^{\text{set}} = +\infty$. Define*

$$R_T^* := \sum_{t=1}^{T}\{d_t^*(a_t) - d_t^*(a_t^*)\}. \tag{EC.23}$$

(a) Adaptive logging. *For any logging process generated by this rule,*

$$\mathbb{E}R_T^* \leq \sum_{t=1}^{T}\mathbb{E}\min\{M, 2L\eta_t^{\text{set}}\} + \varepsilon MT. \tag{EC.24}$$

(b) Stationary specialization. *If this same process satisfies Assumption C.1, and the upper set comparison in (EC.11) holds whenever the thresholds lie in the local quantile neighborhoods, projected PW-OCP from Theorem C.3 and either unprojected method from Proposition C.9 satisfy*

$$\mathbb{E}R_T^* \leq \varepsilon MT + CT\min\{1, (T\pi_{\min})^{-1/4}\sqrt{\log(2KT)}\}. \tag{EC.25}$$

Here C depends only on the loss, set comparison, and local regularity parameters and on c_γ.

Proof. When all learned sets are nonempty, Lipschitz continuity and optimality of $\bar{a}_t$ give

$$\begin{aligned} d_t^*(\bar{a}_t) - d_t^*(a_t^*) &\leq |d_t^*(\bar{a}_t) - d_t(\bar{a}_t)| + |d_t(a_t^*) - d_t^*(a_t^*)| \\ &\leq 2L\eta_t^{\text{set}}. \end{aligned} \tag{EC.26}$$

The same gap is always at most M. Exploration contributes at most εM per round, proving (EC.24), including empty-set rounds. For (EC.25), use the localization event of the relevant sharpness proof with $\delta_T = (2KT)^{-2}$. On this event, post-startup set errors are at most $C_H\eta_t$. Startup rounds and the event's complement contribute at most $Mt_1 + MT\delta_T$, with t_0 in place of t_1 for projection. These terms are absorbed by $CT(T\pi_{\min})^{-1/4}\sqrt{\log(2KT)}$. Outside the effective-sample regime, this rate expression is bounded below by a positive constant, so $R_T^* \leq MT$ suffices. □

This criterion measures decisions using the oracle prediction sets. It is distinct from the realized-loss pseudo-regret $\widetilde{R}_T$, and requires no decision-margin or coverage-transfer condition. Part (a) permits adaptive logging. Part (b) has the same stationarity restriction as Corollary C.6(b), and does not give this rate for a policy that changes its exploitation action with the calibration history.

D Proofs for Section 3

D.1 Proof of Theorem 3.1

Proof. The argument is the pathwise empirical-coverage bound of adaptive conformal inference [9], adapted to the logged sequence. The boundary convention implies that if $\alpha_t < 0$, then $\hat{q}_t = +\infty$,

$\mathcal{C}_t(a_t) = \mathcal{Y}$, and hence $\mathrm{err}_t = 0$; if $\alpha_t > 1$, then $\hat{q}_t = -\infty$, $\mathcal{C}_t(a_t) = \emptyset$, and hence $\mathrm{err}_t = 1$. Since $\alpha_1 \in [0, 1]$ and each update has magnitude at most γ, the standard ACI envelope argument gives

$$\alpha_t \in [-\gamma, 1 + \gamma] \qquad \text{for all } t. \tag{EC.27}$$

Telescoping the recursion yields

$$\alpha_{T+1} = \alpha_1 + \sum_{t=1}^{T} \gamma(\alpha - \mathrm{err}_t). \tag{EC.28}$$

Since $\mathrm{MCov}_T = 1 - T^{-1} \sum_{t=1}^{T} \mathrm{err}_t$, we have

$$\mathrm{MCov}_T - (1 - \alpha) = \frac{1}{T} \sum_{t=1}^{T} (\alpha - \mathrm{err}_t) = \frac{\alpha_{T+1} - \alpha_1}{\gamma T}. \tag{EC.29}$$

Using (EC.27),

$$|\alpha_{T+1} - \alpha_1| \le \max\{\alpha_1, 1 - \alpha_1\} + \gamma, \tag{EC.30}$$

which proves the claim. The proof is pathwise and uses no exchangeability, exogeneity, stationarity, or distributional assumption. □

D.2 Proof of Theorem 3.2

Proof. We give the construction for a small variance shift $\Delta > 0$ and convert it to the endogeneity coefficient at the end. Let $\mathcal{A} = \{0, 1\}$ and $\mathcal{X} = \{L, H\}$. Across rounds, draw X_t i.i.d. uniformly on $\mathcal{X}$, draw the potential outcomes conditionally independently given X_t, and draw the logged action from

$$\pi(0 \mid L) = \pi(1 \mid H) = \frac{3}{4}, \qquad \pi(1 \mid L) = \pi(0 \mid H) = \frac{1}{4}. \tag{EC.31}$$

Thus Assumption 2.2 holds with $\pi_{\min} = 1/4$, and Assumption 2.1 holds because the logging draw depends only on X_t. Let $r = 1 + \Delta$ and set

$$X = L: \quad Y(0) \sim \mathcal{N}(5, 1), \qquad Y(1) \sim \mathcal{N}(5, r^2), \tag{EC.32}$$

$$X = H: \quad Y(0) \sim \mathcal{N}(5, r^2), \qquad Y(1) \sim \mathcal{N}(5, 1). \tag{EC.33}$$

Use the standard absolute-residual score $s(x, y; a) = |y - 5|$.

Let $G \sim \mathcal{N}(0, 1)$, let $F(q) = \mathbb{P}(|G| \le q)$, and let $f(q) = F'(q) = 2\phi(q)$ for $q > 0$. Write $p = 1 - \alpha$ and $q_0 = F^{-1}(p)$. We analyze arm 0; arm 1 is symmetric. Conditional on $a_t = 0$, context L has probability $3/4$ and context H has probability $1/4$. Hence the logged and counterfactual score CDFs for arm 0 are

$$F_{\log}(q; \Delta) = \frac{3}{4} F(q) + \frac{1}{4} F\left(\frac{q}{r}\right), \qquad F_{\mathrm{cf}}(q; \Delta) = \frac{1}{2} F(q) + \frac{1}{2} F\left(\frac{q}{r}\right). \tag{EC.34}$$

Let $q_\Delta = q_0^{\log}$ be the unique solution of $F_{\log}(q_\Delta; \Delta) = p$. Since $\partial_q F_{\log}(q_0; 0) = f(q_0) > 0$, the implicit-function theorem gives

$$q_\Delta = q_0 + \frac{q_0}{4} \Delta + O(\Delta^2). \tag{EC.35}$$

More directly, for $0 < \Delta \le 1/4$, $q_0 \le q_\Delta \le r q_0$ and

$$p - F_{\mathrm{cf}}(q_\Delta; \Delta) = \tfrac{1}{4} \{F(q_\Delta) - F(q_\Delta / r)\} \ge c_0(\alpha) \Delta. \tag{EC.36}$$

The lower bound follows from the positive minimum of $q f(q/s)/s^2$ for $q \in [q_0, 5q_0/4]$ and $s \in [1, 5/4]$.

It remains to pass from the population logged quantile to the time-averaged random thresholds. Let $F_{0,x}$ be the score CDF for arm 0 conditional on $X_t = x$. Predictability permits dependence on the current context, so

$$\mathbb{E}[\mathrm{CCov}_T(0)] = \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}[F_{0, X_t}(\hat{q}_t(0))], \qquad \mathbb{E}[F_{0, X_t}(q_\Delta)] = F_{\mathrm{cf}}(q_\Delta; \Delta). \tag{EC.37}$$

Each conditional score density is at most four times the logged mixture density, since both context weights in that mixture are at least $1/4$. Integrating between q and q_Δ gives

$$|F_{0,x}(q) - F_{0,x}(q_\Delta)| \le 4|F_{\log}(q;\Delta) - p|. \tag{EC.38}$$

The inequality extends to infinite thresholds by limits. Hence, with $L = 4$,

$$|\mathbb{E}[\mathrm{CCov}_T(0)] - F_{\mathrm{cf}}(q_\Delta;\Delta)| \le 4\mathbb{E}\left[\frac{1}{T}\sum_t |F_{\log}(\hat{q}_t(0);\Delta) - p|\right] \le L\varepsilon_T. \tag{EC.39}$$

This tracking error is bounded by one, including empty-buffer predictions. On this Gaussian family, a uniform upper density bound also makes it no larger than a constant times the mean absolute quantile-tracking error.

By Jensen's inequality and the definition of the maximum over actions,

$$\begin{aligned}\mathbb{E}\left[\max_{a\in\mathcal{A}} |\mathrm{CCov}_T(a) - p|\right] &\ge \mathbb{E}[|\mathrm{CCov}_T(0) - p|] \\ &\ge |\mathbb{E}[\mathrm{CCov}_T(0)] - p| \\ &\ge c_0(\alpha)\Delta - L\varepsilon_T.\end{aligned} \tag{EC.40}$$

Finally, the endogeneity coefficient of the construction is

$$\rho(\Delta) = D_{\mathrm{TV}}(\mathcal{N}(5,1), \mathcal{N}(5, r^2)). \tag{EC.41}$$

A direct Gaussian calculation gives

$$\rho(\Delta) = 2\Phi(d_\Delta) - 2\Phi(d_\Delta/r), \qquad d_\Delta = r\sqrt{\frac{2\log r}{r^2 - 1}}, \tag{EC.42}$$

and hence $\rho(\Delta) = 2\phi(1)\Delta + O(\Delta^2)$ as $\Delta \downarrow 0$. The displayed formula is continuous and strictly increasing for $\Delta \in [0, 1/4]$, and $\rho(\Delta)/\Delta$ has positive finite bounds there by continuous extension at zero. Taking $\rho_0 = \rho(1/4) > 0$, independently of α, every $\rho \in (0, \rho_0]$ is therefore attained by the construction. Absorbing the proportionality constant into $c(\alpha)$ gives

$$\mathbb{E}\left[\max_{a\in\mathcal{A}} |\mathrm{CCov}_T(a) - (1-\alpha)|\right] \ge c(\alpha)\rho - L\varepsilon_T. \tag{EC.43}$$

The stated $O(T^{-1/2})$ form follows when $\varepsilon_T = O(T^{-1/2})$. □

D.3 Proof of Lemma 4.2

Proof. Let $\mathcal{F}^X_{t-1} := \sigma(\mathcal{H}_{t-1}, X_t)$. By predictability, $\mathcal{C}_t(a)$ and $\pi_t(a \mid X_t, \mathcal{H}_{t-1})$ are fixed conditional on $\mathcal{F}^X_{t-1}$, and positivity ensures that the inverse-propensity weight is well defined. By consistency,

$$\mathbf{1}\{a_t = a\}\mathbf{1}\{Y_t \notin \mathcal{C}_t(a)\} = \mathbf{1}\{a_t = a\}\mathbf{1}\{Y_t(a) \notin \mathcal{C}_t(a)\}. \tag{EC.44}$$

Therefore

$$\begin{aligned}\mathbb{E}[Z_t(a) \mid \mathcal{F}^X_{t-1}] &= \mathbb{E}\left[\frac{\mathbf{1}\{a_t = a\}}{\pi_t(a \mid X_t, \mathcal{H}_{t-1})}\mathbf{1}\{Y_t(a) \notin \mathcal{C}_t(a)\} \,\middle|\, \mathcal{F}^X_{t-1}\right] \\ &= \frac{\mathbb{P}(a_t = a \mid \mathcal{F}^X_{t-1})}{\pi_t(a \mid X_t, \mathcal{H}_{t-1})}\mathbb{P}\big(Y_t(a) \notin \mathcal{C}_t(a) \mid \mathcal{F}^X_{t-1}, a_t = a\big).\end{aligned} \tag{EC.45}$$

Assumption 2.1, together with the $\mathcal{F}^X_{t-1}$-measurability of $\mathcal{C}_t(a)$, gives

$$\mathbb{P}\big(Y_t(a) \notin \mathcal{C}_t(a) \mid \mathcal{F}^X_{t-1}, a_t = a\big) = \mathbb{P}\big(Y_t(a) \notin \mathcal{C}_t(a) \mid \mathcal{F}^X_{t-1}\big). \tag{EC.46}$$

The definition of the logging policy gives

$$\mathbb{P}(a_t = a \mid \mathcal{F}^X_{t-1}) = \pi_t(a \mid X_t, \mathcal{H}_{t-1}), \tag{EC.47}$$

so the propensity factor cancels, yielding the claim. □

D.4 Envelope for the Weighted Recursion

The IPW recursion lacks the deterministic envelope of standard ACI. A direct comparison with the realized counterfactual error gives a maximal martingale bound without a union bound over time.

Lemma D.1 (High-probability envelope for α_t)**.** *Under Assumptions 2.1–2.2 and the sequential randomization convention of Section 2, Algorithm 1 satisfies, with probability at least* $1-\delta$*,*

$$\max_{a}\max_{t\le T+1}|\alpha_t(a)-\alpha_1(a)| \le 1+\frac{\gamma}{\pi_{\min}}+C\gamma\left\{\sqrt{\frac{T\log(2K/\delta)}{\pi_{\min}}}+\frac{\log(2K/\delta)}{\pi_{\min}}\right\}. \tag{EC.48}$$

Proof. Let $E_t(a)=\mathbf{1}\{Y_t(a)\notin\mathcal{C}_t(a)\}$, $I_t(a)=\mathbf{1}\{a_t=a\}$, and $p_t(a)=\pi_t(a\mid X_t,\mathcal{H}_{t-1})$. Define

$$U_t(a)=Z_t(a)-E_t(a)=\left(\frac{I_t(a)}{p_t(a)}-1\right)E_t(a). \tag{EC.49}$$

Use a filtration containing the full past, including past potential outcomes and previously used algorithmic randomness. Conditional also on X_t and the current potential outcomes, fresh assignment gives

$$\mathbb{E}[U_t(a)\mid\mathcal{F}_{t-1},X_t,\{Y_t(b)\}_b]=0, \tag{EC.50}$$

$$\mathbb{E}[U_t(a)^2\mid\mathcal{F}_{t-1},X_t,\{Y_t(b)\}_b]=(p_t(a)^{-1}-1)E_t(a)\le\pi_{\min}^{-1}. \tag{EC.51}$$

Thus $U_t(a)$ is a martingale difference by iterated conditioning, and $|U_t(a)|\le\pi_{\min}^{-1}$. The maximal form of Freedman's inequality [6], with a union bound only over signs and actions, gives

$$S_T:=\max_{a}\max_{u\le T}\left|\sum_{t=1}^{u}U_t(a)\right|\le C\left\{\sqrt{\frac{T\log(2K/\delta)}{\pi_{\min}}}+\frac{\log(2K/\delta)}{\pi_{\min}}\right\}. \tag{EC.52}$$

Pathwise, the update is $\alpha_{t+1}(a)=\alpha_t(a)+\gamma\{\alpha-E_t(a)\}-\gamma U_t(a)$. Outside $[0,1]$, its first drift term points inward by the quantile boundary convention. Each actual step has magnitude at most $\gamma/\pi_{\min}$. During any excursion above 1, the initial overshoot is at most $\gamma/\pi_{\min}$ and the subsequent outward displacement is at most $2\gamma S_T$. The corresponding argument below 0 is identical. Since $\alpha_1(a)\in[0,1]$,

$$\max_{a,t\le T+1}|\alpha_t(a)-\alpha_1(a)|\le 1+\gamma/\pi_{\min}+2\gamma S_T. \tag{EC.53}$$

This proves the claim without an assumption on the temporal outcome law. □

E Proofs for the Theoretical Analysis

E.1 Proof of Theorem 5.1

Proof. Summing the update proves (16). Since $Z_t(a)=E_t(a)+U_t(a)$, the same identity gives

$$\mathrm{CCov}_T(a)-(1-\alpha)=\frac{\alpha_{T+1}(a)-\alpha_1(a)}{\gamma T}+\frac{1}{T}\sum_{t=1}^{T}U_t(a). \tag{EC.54}$$

Equation (EC.52) first gives

$$\max_{a\in\mathcal{A}}|\mathrm{CCov}_T(a)-(1-\alpha)| \;\le\; \max_{a\in\mathcal{A}}\frac{|\alpha_1(a)-\alpha_{T+1}(a)|}{\gamma T}+C\left\{\sqrt{\frac{\log(2K/\delta)}{T\pi_{\min}}}+\frac{\log(2K/\delta)}{T\pi_{\min}}\right\}, \tag{EC.55}$$

Combining this with (EC.53) yields the pathwise bound

$$G_T:=\max_{a}|\mathrm{CCov}_T(a)-(1-\alpha)|\le\frac{1}{\gamma T}+\frac{1}{T\pi_{\min}}+\frac{3S_T}{T}, \tag{EC.56}$$

and hence (17) for the stated step size. Integrating the exponential tail in (EC.52) gives

$$\mathbb{E}S_T \le C\left\{\sqrt{\frac{T\log(2K)}{\pi_{\min}}}+\frac{\log(2K)}{\pi_{\min}}\right\}. \tag{EC.57}$$

If $T\pi_{\min}\ge\log(2K)$, the linear term is absorbed by the square-root term after division by T. Otherwise use $G_T\le 1$. This proves

$$\mathbb{E}\max_a|\mathrm{CCov}_T(a)-(1-\alpha)|\le C_{c_\gamma}\min\left\{1,\sqrt{\frac{\log(2K)}{T\pi_{\min}}}\right\}. \tag{EC.58}$$

These are finite-horizon maximal bounds for the chosen T, rather than a simultaneous guarantee over all horizons. □

E.2 Proof of Theorem 5.2

Proof. Use the full-history filtration from Lemma D.1. Abbreviate the estimated propensity and outcome mean by $\hat{p}_t(a)$ and $\hat{\mu}_t(a)$. The direct error difference is

$$D_t(a):=Z_t^{\mathrm{DR}}(a)-E_t(a)=\left(\frac{I_t(a)}{\hat{p}_t(a)}-1\right)\{E_t(a)-\hat{\mu}_t(a)\}. \tag{EC.59}$$

For estimated propensities, the stated conditional-mean assumption implies

$$b_t(a):=\mathbb{E}[D_t(a)\mid\mathcal{F}_{t-1},X_t]=\{\hat{\mu}_t(a)-\mu_t^*(a)\}\left(1-\frac{p_t(a)}{\hat{p}_t(a)}\right). \tag{EC.60}$$

Then $M_t(a)=D_t(a)-b_t(a)$ is a martingale difference. Fresh randomization and $|E_t(a)-\hat{\mu}_t(a)|\le 1$ give

$$\mathbb{E}[D_t(a)^2\mid\mathcal{F}_{t-1},X_t]\le\frac{p_t(a)}{\hat{p}_t(a)^2}-\frac{2p_t(a)}{\hat{p}_t(a)}+1\le\frac{2\kappa}{\hat{\pi}_{\min}}. \tag{EC.61}$$

Consequently $|M_t(a)|\le 2/\hat{\pi}_{\min}$ and its conditional variance is at most $2\kappa/\hat{\pi}_{\min}$. Maximal Freedman gives, with probability at least $1-\delta$,

$$S_T^{\mathrm{DR}}:=\max_a\max_{u\le T}\left|\sum_{t=1}^{u}M_t(a)\right|\le C\left\{\sqrt{\frac{\kappa T\log(2K/\delta)}{\hat{\pi}_{\min}}}+\frac{\log(2K/\delta)}{\hat{\pi}_{\min}}\right\}. \tag{EC.62}$$

The exact update is

$$\alpha_{t+1}(a)=\alpha_t(a)+\gamma\{\alpha-E_t(a)\}-\gamma b_t(a)-\gamma M_t(a). \tag{EC.63}$$

Its actual one-step magnitude is at most $\gamma/\hat{\pi}_{\min}$. Applying the excursion argument of Lemma D.1, now allowing the total bias displacement, gives

$$\max_{t\le T+1}|\alpha_t(a)-\alpha_1(a)|\le 1+\frac{\gamma}{\hat{\pi}_{\min}}+2\gamma S_T^{\mathrm{DR}}+\gamma\sum_{t=1}^{T}|b_t(a)|. \tag{EC.64}$$

Telescoping against $E_t(a)$ therefore yields

$$G_T^{\mathrm{DR}}\le\frac{1}{\gamma T}+\frac{1}{T\hat{\pi}_{\min}}+\frac{3S_T^{\mathrm{DR}}}{T}+2\max_a\frac{1}{T}\sum_{t=1}^{T}|b_t(a)|. \tag{EC.65}$$

Cauchy–Schwarz gives

$$\frac{1}{T}\sum_{t=1}^{T}|b_t(a)|\le\frac{\epsilon_{\mu,a}\epsilon_{\pi,a}}{\hat{\pi}_{\min}}. \tag{EC.66}$$

This proves (20), with $1/(T\hat{\pi}_{\min})$ absorbed into its logarithmic term.

If $\hat{p}_t = p_t$, conditioning on the current potential outcomes already gives $\mathbb{E}[D_t(a) \mid \mathcal{F}_{t-1}, X_t, \{Y_t(b)\}_b] = 0$. Thus $b_t = 0$ without any outcome-kernel restriction, and the same maximal bound and tail integration prove (17) and (EC.58) for DR-OCP.

For a projected update define $r_t(a) = \alpha_{t+1}(a) - \{\alpha_t(a) + \gamma(\alpha - Z_t^{\mathrm{DR}}(a))\}$. Its separate exact identity is

$$\frac{1}{T}\sum_t Z_t^{\mathrm{DR}}(a) - \alpha = \frac{\alpha_1(a) - \alpha_{T+1}(a) + \sum_t r_t(a)}{\gamma T}. \tag{EC.67}$$

It implies a coverage bound with the absolute value of this endpoint-plus-residual term, plus S_T^{DR}/T and the product bias in (EC.66). A bounded projection interval alone does not control the accumulated residual or justify the unprojected envelope. □

Variance of the augmentation. For the unprojected recursion with known propensities, let $\bar{\mu}_t(a) = \mathbb{E}[E_t(a) \mid \mathcal{F}_{t-1}, X_t]$. Fresh randomization gives the exact identity

$$\mathbb{E}[D_t(a)^2 \mid \mathcal{F}_{t-1}, X_t] = (p_t(a)^{-1} - 1)\{\bar{\mu}_t(a)(1 - \bar{\mu}_t(a)) + (\hat{\mu}_t(a) - \bar{\mu}_t(a))^2\}. \tag{EC.68}$$

At fixed thresholds, the corresponding PW-OCP error difference $Z_t(a) - E_t(a)$ has conditional second moment $(p_t(a)^{-1} - 1)\bar{\mu}_t(a)$. Thus the DR error difference has no larger conditional variance whenever $(\hat{\mu}_t(a) - \bar{\mu}_t(a))^2 \le \bar{\mu}_t(a)^2$. The oracle choice $\hat{\mu}_t = \bar{\mu}_t$ minimizes this variance when $\bar{\mu}_t$ is measurable from the observed history and current context. Validity with known propensities does not require this accuracy condition.

Put $V_T(a) = \sum_{t=1}^T \mathbb{E}[D_t(a)^2 \mid \mathcal{F}_{t-1}]$ and $x = \log(2K/\delta)$. For any deterministic $v \ge 0$, maximal Freedman and the same excursion argument give

$$\mathbb{P}\left(G_T^{\mathrm{DR}} > \frac{1}{\gamma T} + \frac{1}{T\pi_{\min}} + \frac{C}{T}\left\{\sqrt{vx} + \frac{x}{\pi_{\min}}\right\}, \max_a V_T(a) \le v\right) \le \delta. \tag{EC.69}$$

Indeed, D_t is a martingale difference with $|D_t| \le 1/\pi_{\min}$, and the pathwise envelope gives $G_T^{\mathrm{DR}} \le 1/(\gamma T) + 1/(T\pi_{\min}) + 3\max_{a,u\le T} |\sum_{t\le u} D_t(a)|/T$. Apply the variance-budget form of Freedman's inequality and union over actions and signs. This is a joint-event bound with a fixed budget v, rather than conditioning on a future variance event. It quantifies a possible reduction in the stochastic term without changing the recursion or its worst-case rate.

E.3 Proof of Theorem 5.4 and Corollary 5.5

Proof of Theorem 5.4. Write $p = 1 - \alpha$, $F(q) = \mathbb{P}(|\mathcal{N}(0,1)| \le q)$, $J = J_\alpha$, and $N = T\pi_{\min}$. It suffices to work inside $\mathcal{P}_G$ with the fixed logger $\pi(a) = \pi_{\min}$ for $a < K$ and $\pi(K) = 1 - (K-1)\pi_{\min}$. Let Q_v denote the complete logged-data law, including an independent algorithmic seed, when arm $v < K$ has scale $1 + \Delta$ and all other arms have scale 1. Let Q_0 be the law when all scales equal 1. Here $0 < \Delta \le 1/4$.

Separation in expectation. Positivity of the half-normal density on compact subsets of $(0, \infty)$ gives

$$F(q) - F(q/(1+\Delta)) \ge d_\alpha \Delta \quad \text{for every } q \in J, \tag{EC.70}$$

for a constant $d_\alpha > 0$. This follows by integrating $qF'(q/s)/s^2$ over $s \in [1, 1+\Delta]$. The difference is nonnegative outside J, including infinite thresholds. Define $A_a = T^{-1}\sum_t F_a(\hat{q}_t(a))$ and $M_a = T^{-1}\sum_t F(\hat{q}_t(a))$. Under Q_v, $A_a = M_a$ for $a \ne v$ and $M_v - A_v \ge d_\alpha \Delta L_T(v)$. Since $0 \le L_T(v) \le 1$ and $\mathbb{E}_v L_T(v) \ge 2/3$,

$$\mathbb{P}_v(L_T(v) \ge 1/3) \ge 1/2. \tag{EC.71}$$

On this event, $\max_a |A_a - p| < d_\alpha\Delta/12$ makes $\arg\max_{a<K} M_a$ identify v. Therefore

$$\mathbb{P}_v(\text{testing error}) \le \frac{1}{2} + \frac{12}{d_\alpha \Delta}\mathbb{E}_v \max_a |A_a - p|. \tag{EC.72}$$

The Gaussian KL divergence is at most $C\Delta^2$, so the logged-data KL chain rule gives $\mathrm{KL}(Q_v \| Q_0) \le CN\Delta^2$. For $K \ge 257$, choose

$$\Delta = c_\alpha \min\left\{1, \sqrt{\frac{\log K}{N}}\right\} \tag{EC.73}$$

small enough that the mutual information under a uniform choice of $v < K$ is at most $\log(K-1)/8$. Since $\log 2/\log(K-1) \le 1/8$, Fano's inequality gives testing error at least $3/4$ and hence predictable coverage risk at least a constant times Δ.

For $4 \le K < 257$, use Q_0, Q_1. On the common part $Q_0 \wedge Q_1$ of their densities, the triangle inequality gives

$$\mathbb{E}_0|M_1 - p| + \mathbb{E}_1|A_1 - p| \ge d_\alpha \Delta \int L_T(1)\, d(Q_0 \wedge Q_1) \ge d_\alpha \Delta\{2/3 - \mathrm{TV}(Q_0, Q_1)\}. \quad \text{(EC.74)}$$

Choose $\Delta = c_\alpha \min\{1, N^{-1/2}\}$ so that Pinsker's inequality makes the total variation at most $1/3$. The bounded range of $\log K$ then gives the same rate, after adjusting constants. In both cases,

$$\sup_{(P,\pi)\in\mathcal{P}_G} \mathbb{E} \max_a |A_a - p| \ge c_1 r, \qquad r = \min\left\{1, \sqrt{\frac{\log K}{N}}\right\}. \quad \text{(EC.75)}$$

Passage to realized coverage. Write $s_T = \min\{1, \sqrt{\log K/T}\}$. Fresh independent potential outcomes make $\mathbf{1}\{Y_t(a) \in \mathcal{C}_t(a)\} - F_a(\hat{q}_t(a))$ a bounded martingale difference. Integrating Azuma's inequality over actions gives $\mathbb{E}\max_a |\mathrm{CCov}_T(a) - A_a| \le C_1 s_T$. Thus the worst-case realized risk R satisfies $R \ge c_1 r - C_1 s_T$.

A second bound controls the hidden-outcome fluctuation. Under Q_0, let

$$H_a = \sum_{t=1}^{T} \mathbf{1}\{a_t \ne a,\ \hat{q}_t(a) \in J\}, \qquad a < K. \quad \text{(EC.76)}$$

Fresh fixed-probability assignment gives $\mathbb{E}H_a = (1-\pi_{\min})T\mathbb{E}L_T(a) \ge T/2$. Since $H_a \le T$, $\mathbb{P}(H_a \ge T/4) \ge 1/3$. Consequently, if $B = \sum_{a<K} \mathbf{1}\{H_a \ge T/4\}$, then $\mathbb{E}B \ge (K-1)/3$ and

$$\mathbb{P}(B \ge (K-1)/6) \ge 1/5. \quad \text{(EC.77)}$$

On this event there are at least $M = \max\{1, \lfloor (K-1)/6 \rfloor\}$ arms with at least $m = \lceil T/4 \rceil$ hidden outcomes at informative thresholds. Conditional on the complete log and the algorithmic seed, these outcomes are independent, and their coverage probabilities lie in $[b, 1-b]$ for a constant $b = b(\alpha) > 0$.

For mutually independent Bernoulli variables B_{ji} with success probabilities $p_{ji} \in [b, 1-b]$, let $S_j = \sum_{i=1}^{m} B_{ji}$. For arbitrary offsets z_j,

$$\mathbb{E} \max_{j \le M} |S_j - z_j| \ge c_b \min\{m, \sqrt{m \log(M+1)}\}. \quad \text{(EC.78)}$$

To prove this, introduce an independent copy S' of S and write $f(x) = \max_j |x_j|$. Each summand of $S_j - S'_j$ equals either sign with probability $p_{ji}(1-p_{ji}) \ge v := b(1-b)$. For a convex function, increasing the common probability of the two signs increases its expectation, since $f(x+e_j)+f(x-e_j) \ge 2f(x)$ for each coordinate vector e_j. Applying this fact one summand at a time, together with the triangle inequality, gives $\mathbb{E}f(S-z) \ge \frac{1}{2}\mathbb{E}f(S-S') \ge \frac{1}{2}\mathbb{E}f(W)$, where the W_j are independent sums of m i.i.d. variables taking $1, -1, 0$ with probabilities $v, v, 1-2v$.

For a sum V_n of $n \ge 1$ such variables, $\mathbb{E}V_n^2 = 2vn$ and $\mathbb{E}V_n^4 \le 2vn + 12v^2n^2$. Symmetry and the Paley–Zygmund inequality imply

$\mathbb{P}(V_n \ge \sqrt{vn}) \ge \frac{(vn)^2}{2(2vn+12v^2n^2)} \ge \beta_v := \frac{v}{4+24v} > 0$. For any $1 \le k \le m$, partition each sum W_j into k independent blocks of lengths at least $\lfloor m/k \rfloor \ge m/(2k)$. With probability at least β_v^k, every block of length n has sum at least $\sqrt{vn}$. Hence $\mathbb{P}\Big(W_j \ge \sqrt{vmk/2}\Big) \ge \beta_v^k$. Put $C_v = \log(1/\beta_v)$ and choose $k = \min\Big\{m, \max\Big\{1, \Big\lfloor \frac{\log(M+1)}{2C_v} \Big\rfloor\Big\}\Big\}$. Then $k \ge (4C_v)^{-1}\min\{m, \log(M+1)\}$. If $k = 1$, a single sum reaches the displayed threshold with probability at least β_v. If $k \ge 2$, then $\beta_v^k \ge (M+1)^{-1/2}$, so independence gives $\mathbb{P}\Big(\max_{j\le M} W_j \ge \sqrt{vmk/2}\Big) \ge 1-(1-\beta_v^k)^M \ge 1-e^{-1/\sqrt{2}}$. Combining these bounds with symmetrization proves (EC.78) for all $m, M \ge 1$, including the saturated regime $\log(M+1) \ge m$.

Condition further on the remaining hidden outcomes, which together with the revealed outcomes contribute only fixed offsets, and apply (EC.78) to the selected sums. Since $\log(M+1)$ is bounded below by an absolute positive multiple of $\log K$ for $K \geq 4$, averaging over the event above gives $R \geq c_2 s_T$. Finally,

$$R \geq \max\{c_2 s_T, c_1 r - C_1 s_T\} \geq \frac{c_1 c_2}{C_1 + c_2}\, r. \tag{EC.79}$$

All hard instances lie in $\mathcal{P}_G \subset \mathcal{P}$, proving both assertions. □

Lemma E.1 (Nondegeneracy under adaptive logging)**.** *On $\mathcal{P}_G$, run Algorithm 1 or DR-OCP with known propensities and any predictable $\hat{\mu}_t \in [0,1]$, using $\gamma = c_\gamma \sqrt{\pi_{\min}/T}$ for fixed $c_\gamma > 0$. Put $\nu_T(a) = 1 - L_T(a)$. For each action a and $0 < \delta < 1/2$,*

$$T\pi_{\min} \geq C_{\alpha,c_\gamma} \log(2/\delta) \quad \Longrightarrow \quad \mathbb{P}(\nu_T(a) > 1/4) \leq \delta. \tag{EC.80}$$

The constant is uniform over the Gaussian instance class and history-dependent loggers. The algorithm retains its original quantile rule and empty-buffer predictions.

Proof. Fix an action and suppress its index. Write $p_0 = \pi_{\min}$, $N = Tp_0$, $L_\delta = \log(6/\delta)$, and $F_a(q) = \mathbb{P}(|\mathcal{N}(0,\sigma_a^2)| \leq q)$. Define

$$e_t = \alpha_t - \alpha, \qquad u_t = \Pi_{[0,1]}(\alpha_t) - \alpha, \tag{EC.81}$$

$$\zeta_t = F_a(\hat{q}_t) - \{1 - \Pi_{[0,1]}(\alpha_t)\}. \tag{EC.82}$$

Here $\Pi_{[0,1]}$ describes the quantile boundary convention in the analysis, not a projection of the algorithm. For a nonempty buffer of size n, $|\zeta_t|$ is at most its uniform empirical-CDF error D_n plus an empirical-rank correction $1/n$. For an empty buffer, $|\zeta_t| \leq 1$.

A fixed-accuracy empirical-CDF event. The logger selects an action before seeing its current outcome. Under the context-free Gaussian model, successive observed scores of this action form an i.i.d. stream from F_a, even when the selection probabilities depend on past observed scores. We use a uniform event on this stream, without conditioning on its random sample count. Fix a small $\epsilon > 0$, to be chosen using only α. By DKW and a union bound over all $n \geq n_0$,

$$\mathbb{P}(\exists n \geq n_0 : D_n > \epsilon) \leq \frac{2e^{-2n_0\epsilon^2}}{1 - e^{-2\epsilon^2}} \leq \delta/3 \tag{EC.83}$$

for $n_0 \geq C_\epsilon L_\delta$, chosen also so that $1/n_0 \leq \epsilon$. Let N_m count observations of this action through round m. For $\lambda > 0$, iterated conditioning gives

$$\mathbb{E}[e^{-\lambda N_m}] \leq (1 - p_0 + p_0 e^{-\lambda})^m, \qquad \mathbb{P}(N_m < mp_0/2) \leq e^{-mp_0/8}. \tag{EC.84}$$

Choose $m = \lceil C_\epsilon L_\delta / p_0 \rceil$ large enough that $mp_0/2 \geq n_0$ and the last probability is at most $\delta/3$. If $m \leq T$, monotonicity of the count then gives

$$\frac{1}{T}\sum_{t=1}^{T} \zeta_t^2 \leq \frac{m}{T} + 4\epsilon^2 \leq \frac{C_\epsilon L_\delta}{N} + 4\epsilon^2 \tag{EC.85}$$

with probability at least $1 - 2\delta/3$. The sample-size condition below ensures $m \leq T$.

Controlling the index without restricting its range. Use the nonnegative, 1-Lipschitz function

$$V(e) = \begin{cases} e^2/2, & |e| \leq 1, \\ |e| - 1/2, & |e| > 1, \end{cases} \qquad g(e) = V'(e) = \max\{-1, \min\{e, 1\}\}. \tag{EC.86}$$

Its derivative is 1-Lipschitz. Put $a_0 = \min\{\alpha, 1-\alpha\} > 0$. For every real e, $g(e)\{\Pi_{[0,1]}(\alpha + e) - \alpha\} \geq a_0 g(e)^2$. With $W_t = \alpha - Z_t(a)$ for PW-OCP, or $W_t = \alpha - Z_t^{\mathrm{DR}}(a)$ for DR-OCP with known propensities, unbiasedness gives

$$\mathbb{E}[W_t \mid \mathcal{H}_{t-1}] = -u_t + \zeta_t, \tag{EC.87}$$

$$|W_t| \leq 1/p_0, \qquad \mathbb{E}[W_t^2 \mid \mathcal{H}_{t-1}] \leq C/p_0. \tag{EC.88}$$

Smoothness and Young's inequality yield

$$\mathbb{E}[V(e_{t+1}) - V(e_t) \mid \mathcal{H}_{t-1}] \le -\frac{a_0\gamma}{2} g(e_t)^2 + \frac{\gamma}{2a_0}\zeta_t^2 + \frac{C\gamma^2}{p_0}. \tag{EC.89}$$

The centered increments of $V(e_t)$ have magnitude at most $2\gamma/p_0$ and conditional variance at most $C\gamma^2/p_0$. Freedman's inequality controls their sum by $C\gamma\{\sqrt{TL_\delta/p_0} + L_\delta/p_0\}$ with failure probability at most $\delta/3$. Summing (EC.89), using $e_1 = 0$ and $V \ge 0$, gives

$$\frac{1}{T}\sum_{t=1}^{T} g(e_t)^2 \le C_\alpha \left\{ \frac{1}{T}\sum_{t=1}^{T} \zeta_t^2 + \frac{\gamma}{p_0} + \sqrt{\frac{L_\delta}{N}} + \frac{L_\delta}{N} \right\}. \tag{EC.90}$$

Each oracle quantile $\sigma_a q_\alpha$ lies strictly inside J_α. Compactness of the scale range gives

$$h_\alpha := \inf_{\sigma\in[1,5/4]} \min\{p - F_\sigma(q_\alpha/2),\ F_\sigma(2q_\alpha) - p\} > 0, \qquad p = 1-\alpha. \tag{EC.91}$$

If $\hat{q}_t \notin J_\alpha$, then $|F_a(\hat{q}_t) - p| \ge h_\alpha$. Since $F_a(\hat{q}_t) - p = -u_t + \zeta_t$ and $u_t^2 \le g(e_t)^2$,

$$\nu_T(a) \le \frac{2}{h_\alpha^2 T}\sum_{t=1}^{T}\{g(e_t)^2 + \zeta_t^2\} \le C_\alpha \left\{ 4\epsilon^2 + \frac{C_\epsilon L_\delta}{N} + \frac{c_\gamma}{\sqrt{N}} + \sqrt{\frac{L_\delta}{N}} \right\} \tag{EC.92}$$

on an event of probability at least $1-\delta$. Choose ϵ so that the first term is at most $1/8$, and then $N \ge C_{\alpha,c_\gamma} L_\delta$ so that the remaining terms sum to at most $1/8$. This also ensures $m \le T$. Absorbing the fixed ratio between L_δ and $\log(2/\delta)$ proves (EC.80). These choices are proof constants, not algorithm parameters. □

Proof of Corollary 5.5. Theorems 5.1 and 5.2 give (EC.58) uniformly over $\mathcal{P}$, without threshold restrictions. On $\mathcal{P}_G$, Lemma E.1 with $\delta = 0.05$ gives, separately for every action,

$$\mathbb{E}L_T(a) \ge (1 - 0.05)\cdot\frac{3}{4} = \frac{57}{80} > \frac{2}{3} \tag{EC.93}$$

whenever (23) holds. Thus both algorithms belong to $\mathfrak{A}_{\mathrm{nd}}$, uniformly over adaptive context-free loggers on the regular submodel. Combining their full-model upper bounds with Theorem 5.4 proves the minimax rate up to constants. This is a worst-case comparison over $\mathcal{P}$, not an assertion of instancewise optimality on every outcome law. □

Other informative intervals and fractions. Fix $J = [l, u]$ with $0 < l < u < \infty$ and $0 < \vartheta < 1$. Replace (21) by

$$\inf_{(P,\pi)\in\mathcal{P}_G} \min_a \mathbb{E}_{P,\pi}\left[\frac{1}{T}\sum_t \mathbf{1}\{\hat{q}_t(a) \in J\}\right] \ge \vartheta. \tag{EC.94}$$

Theorem 5.4 retains the rate in (22), with constant $c(\alpha, J, \vartheta) > 0$. Thus neither the interval J_α nor the fraction $2/3$ is essential to this lower bound.

To verify the extension, set $L_T^J(a) = T^{-1}\sum_t \mathbf{1}\{\hat{q}_t(a) \in J\}$ and $d_J = \inf_{q\in J,\, s\in[1,5/4]} qF'(q/s)/s^2 > 0$. Then the scale separation is at least $d_J\Delta L_T^J(a)$, and $\mathbb{P}(L_T^J(a) \ge \vartheta/2) \ge \vartheta/(2-\vartheta) \ge \vartheta/2$. Writing $R_{\mathrm{pred}} = \sup \mathbb{E}\max_a |A_a - p|$, the same decoder has error at most $1 - \vartheta/2 + 8R_{\mathrm{pred}}/(d_J\vartheta\Delta)$. For $K \ge K_\vartheta$, choose K_ϑ so that $\log 2/\log(K-1) \le \vartheta/8$, and choose $\Delta = c_\vartheta \min\{1, \sqrt{\log K/N}\} \le 1/4$ so that the normalized KL term is at most $\vartheta/8$. Fano then gives error at least $1 - \vartheta/4$, hence $R_{\mathrm{pred}} \ge d_J\vartheta^2\Delta/32$. For $4 \le K < K_\vartheta$, the common-measure argument gives separation at least $d_J\Delta(\vartheta - \mathrm{TV})$. Choosing $\Delta = c_\vartheta \min\{1, N^{-1/2}\}$ to ensure $\mathrm{TV} \le \vartheta/2$ proves the same rate, since $\log K$ is bounded in terms of ϑ.

For the realized-coverage step, fixed assignment gives $\mathbb{E}H_a \ge hT$ with $h = 3\vartheta/4$. Hence $\mathbb{P}(H_a \ge hT/2) \ge h/2$. With probability at least $h/4$, at least $h(K-1)/4$ arms meet this count. Apply (EC.78) to $m = \lceil hT/2\rceil$ hidden outcomes in each of $M = \max\{1, \lfloor h(K-1)/4\rfloor\}$ such arms.

On J, their Bernoulli probabilities are bounded away from zero and one by constants depending on J. This gives $R \geq c_{J,\vartheta} s_T$. Combining it with $R \geq c_{\alpha,J,\vartheta} r - C_1 s_T$ completes the extension.

If $[q_\alpha, (5/4)q_\alpha] \subset \mathrm{int}(J)$, both original algorithms also satisfy (EC.94) for $N \geq C_{\alpha,c_\gamma,J,\vartheta}$. In Lemma E.1, replace h_α by the positive uniform CDF margin of J. Choose the empirical-CDF tolerance and then N so that $\mathbb{P}(1 - L_T^J(a) > \tau) \leq \delta$, with $\tau = \delta = (1-\vartheta)/2$. Then $\mathbb{E}L_T^J(a) \geq (1-\delta)(1-\tau) > \vartheta$. The constants depend on the interval's separation and CDF margins and on ϑ and c_γ, but not on $K, T, \pi_{\min}$. These choices remain comparison criteria and proof constants, not algorithm inputs.

E.4 Proof of Theorem C.3

Proof. Write $p_0 = \pi_{\min}$, $N = Tp_0$, and $\Lambda = \log(2KT/\delta)$. For an arm, abbreviate the logged and counterfactual score CDFs by F_b, F_c, and put $Q_b = F_b^{-1}$. The fixed point is $\alpha^* = 1 - F_b(q^*)$, and

$$h(\beta) = F_c(Q_b(1-\beta)) - (1-\alpha), \qquad -\lambda_{\max} \leq h'(\beta) \leq -\lambda_{\min} < 0. \tag{EC.95}$$

The projected indices belong to $\mathcal{I}_a$ from round two onward, regardless of the initial empty-buffer prediction.

Localizing the empirical quantiles. Let $t_0 = 1 + \lceil C_0 \Lambda / p_0 \rceil$. The claim is immediate if $t_0 > T$. Under the fixed logger, logged scores at their arrival times are i.i.d. from F_b. Chernoff's inequality for action counts and DKW uniformly over arms and sample sizes imply, with probability at least $1 - \delta/2$,

$$\max_a |r_t(a)| \leq C\sqrt{\frac{\Lambda}{tp_0}}, \quad t_0 \leq t \leq T, \qquad r_t(a) = \hat{q}_t(a) - Q_b^{(a)}(1 - \alpha_t(a)). \tag{EC.96}$$

Choose C_0 using the local density lower bound, ε_q, and the distance of the index intervals from 0 and 1. The empirical quantiles then lie in $\widetilde{\mathcal{Q}}_a$, uniformly over their data-dependent indices. This follows by inverting the uniform CDF bound locally, with any empirical-rank correction absorbed into its constant. In particular, $\eta_t \leq D_q := 1/f_{\min} + \varepsilon_q$ for $t \geq t_0$ on this event. This argument uses no global score-tail bound.

Controlling the recursion on time blocks. First suppose $\gamma\lambda_{\max} \leq 1$. The noiseless projected map contracts toward α^* by at most $1 - \gamma\lambda_{\min}$. Put

$$e_t(a) = \alpha_t(a) - \alpha^*(a), \qquad m_t(a) = \mathbb{E}[Z_t(a) \mid \mathcal{H}_{t-1}], \tag{EC.97}$$

$$M_t(a) = m_t(a) - Z_t(a), \qquad \xi_t(a) = F_c^{(a)}(\hat{q}_t(a)) - F_c^{(a)}(Q_b^{(a)}(1 - \alpha_t(a))). \tag{EC.98}$$

The threshold depends on past scores. Hence $M_t(a)$ is a martingale difference, with magnitude and conditional variance at most C/p_0. On the quantile event, $|\xi_t(a)| \leq C|r_t(a)|$, and

$$\alpha_{t+1}(a) = \Pi_{\mathcal{I}_a}\big(\alpha_t(a) + \gamma\{h_a(\alpha_t(a)) + \xi_t(a) + M_t(a)\}\big). \tag{EC.99}$$

Partition $\{t_0, \ldots, T\}$ into blocks of length $m = \lceil \log 2/(\lambda_{\min}\gamma) \rceil$, with a possibly shorter last block. Freedman's maximal inequality and a union bound over arms and blocks give, with probability at least $1 - \delta/2$,

$$\max_a \max_{1 \leq u \leq m} \left| \sum_{i=s}^{s+u-1} M_i(a) \right| \leq C\left\{ \sqrt{\frac{m\Lambda}{p_0}} + \frac{\Lambda}{p_0} \right\} \tag{EC.100}$$

for every block start s, with sums stopping at T.

Compare the noisy trajectory on a block with the noiseless projected trajectory started at the same index. Projection secants lie in $[0,1]$, and the secants of $\beta + \gamma h(\beta)$ lie in $[0, 1 - \gamma\lambda_{\min}]$. Expanding the difference gives a noise term $\gamma \sum_i w_i M_i$, where $0 \leq w_i \leq 1$ and w_i are nondecreasing in i. Abel summation bounds its magnitude by twice the displayed maximal partial sum times γ. The weights may depend on the whole block, since this comparison is pathwise. The bias contribution is at most $C \sup_{i \text{ in block}} |r_i|$.

Writing E_j for the largest index error at block start s_j and $\nu = C\{\sqrt{\gamma\Lambda/p_0} + \gamma\Lambda/p_0\}$ yields, for full blocks,

$$E_{j+1} \leq \tfrac{1}{2}E_j + C\sqrt{\frac{\Lambda}{s_j p_0}} + \nu, \qquad \max_{a,\, t \text{ in block } j} |e_t(a)| \leq E_j + C\sqrt{\frac{\Lambda}{s_j p_0}} + \nu. \tag{EC.101}$$

Summing the geometric recursion, and accounting for the first block by $O(m/T)$, gives

$$\frac{1}{T}\sum_{t=t_0}^{T}\max_a|e_t(a)| \le C\left\{\frac{1}{\gamma T}+\sqrt{\frac{\Lambda}{Tp_0}}+\sqrt{\frac{\gamma\Lambda}{p_0}}+\frac{\gamma\Lambda}{p_0}\right\}. \tag{EC.102}$$

Local inverse-CDF regularity gives $|\hat{q}_t(a)-q^*(a)| \le |r_t(a)|+|e_t(a)|/f_{\min}$. Substitution of $\gamma = c_\gamma\sqrt{p_0/T}$ therefore proves (EC.4) when $N \ge \Lambda^2$ and $\gamma\lambda_{\max} \le 1$.

The remaining regimes. If $N < \Lambda^2$, the claimed right-hand side is at least a positive constant, and the localized bound $\eta_t \le D_q$ suffices. If $\gamma\lambda_{\max} > 1$, then $N < (c_\gamma\lambda_{\max})^2$, so the same localized bound proves the claim after adjusting C. Thus neither an effective-sample restriction nor a density-dependent choice of c_γ is needed. The two failure probabilities sum to at most δ. □

E.5 Proof of Theorem C.4

Proof. Under P_0, both arms have law $\mathcal{N}(0,1)$. Under P_1, arm 1 has the same law and arm 2 has law $\mathcal{N}(0,(1+\Delta)^2)$. Potential outcomes are independent across rounds and arms. Write $p = 1-\alpha$ and $F(q) = \mathbb{P}(|\mathcal{N}(0,1)| \le q)$. For $0 < \Delta \le 1/4$, equation (EC.70) supplies the uniform separation on J_α.

First the coverage premise controls the horizon needed for testing. Let H_a count unobserved outcomes of arm a whose thresholds lie in J_α. Since only one outcome is observed each round,

$$H_1+H_2 \ge T\{L_T(1)+L_T(2)-1\}, \qquad \mathbb{E}_i(H_1+H_2) \ge T/3. \tag{EC.103}$$

Also $H_1+H_2 \le T$. Hence $H_1+H_2 \ge T/6$ with probability at least $1/5$, and then one arm has at least $\lceil T/12\rceil$ such outcomes. Conditional on the complete log and the algorithmic seed, hidden outcomes remain independent even under the adaptive decision rule. Their coverage probabilities on J_α are bounded away from zero and one. The single-arm case of (EC.78), allowing arbitrary offsets, therefore gives

$$\sup_{i\in\{0,1\}}\mathbb{E}_i\max_a|\mathrm{CCov}_T(a)-p| \ge c_0(\alpha)/\sqrt{T}. \tag{EC.104}$$

The accuracy premise implies $T^{-1/2} \le \delta/c_0(\alpha)$.

Let Q_0, Q_1 be the complete logged-data laws, and put

$$M = \frac{1}{T}\sum_t F(\hat{q}_t(2)), \qquad M^\Delta = \frac{1}{T}\sum_t F(\hat{q}_t(2)/(1+\Delta)). \tag{EC.105}$$

The realized-minus-predictable coverage martingale has expected absolute value at most $1/(2\sqrt{T})$. Thus, for a constant $C_1 = C_1(\alpha)$,

$$\mathbb{E}_0|M-p| \le C_1\delta, \qquad \mathbb{E}_1|M^\Delta-p| \le C_1\delta. \tag{EC.106}$$

On every logged history, $M - M^\Delta \ge d_\alpha\Delta L_T(2)$. Integrating the triangle inequality over the common part of Q_0, Q_1 gives

$$2C_1\delta \ge d_\alpha\Delta\{2/3-\mathrm{TV}(Q_0,Q_1)\}. \tag{EC.107}$$

Choose $\Delta = 12C_1\delta/d_\alpha$ and $\delta_0(\alpha)$ small enough that $\Delta \le 1/4$. Then $\mathrm{TV}(Q_0,Q_1) \ge 1/2$. Pinsker's inequality and the adaptive KL chain rule yield

$$c \le \mathrm{KL}(Q_1\|Q_0) = \mathbb{E}_1[N_2]\,\mathrm{KL}(\mathcal{N}(0,(1+\Delta)^2)\|\mathcal{N}(0,1)) \le C\Delta^2\mathbb{E}_1[N_2]. \tag{EC.108}$$

Each pull of arm 2 under P_1 has excess expected squared loss at least 2Δ. Therefore

$$\mathbb{E}_1[\mathrm{Regret}_T] \ge 2\Delta\mathbb{E}_1[N_2] \ge c/\Delta = c'(\alpha)/\delta. \tag{EC.109}$$

This proves the implication without a separate horizon condition. It does not assert attainability when the accuracy premise is infeasible. □

E.6 Proof of Theorem C.5

Proof. For every action a, the Lipschitz assumption gives

$$|d_t(a) - d_t^*(a)| \le L\, d_H(\mathcal{C}_t(a), \mathcal{C}_t^*(a)) \le L\eta_t^{\mathrm{set}}. \tag{EC.110}$$

On exploitation rounds, $a_t = \bar{a}_t$ and $\bar{a}_t$ minimizes $d_t(a)$, so

$$d_t(a_t) = d_t(\bar{a}_t) \le d_t(a_t^*) \le d_t^*(a_t^*) + L\eta_t^{\mathrm{set}}, \tag{EC.111}$$

where the second inequality uses (EC.110). If $Y_t(a_t) \in \mathcal{C}_t(a_t)$, then $\ell(a_t, Y_t(a_t)) \le d_t(a_t)$. If the outcome is outside the set, assumption (ii) adds at most B. Therefore

$$\ell(a_t, Y_t(a_t)) - d_t^*(a_t^*) \le L\eta_t^{\mathrm{set}} + B\mathbf{1}\{Y_t(a_t) \notin \mathcal{C}_t(a_t)\} \tag{EC.112}$$

on exploitation rounds. On exploration rounds, assumption (iii) gives $d_t(a_t) \le d_t(\bar{a}_t) + \Delta_{\max}$, so the same inequality holds with this additional term. Summing pathwise, with $I_t^{\exp}$ denoting the exploration indicator, gives

$$\widetilde{R}_T \le L\sum_{t=1}^{T} \eta_t^{\mathrm{set}} + B\sum_{t=1}^{T} \mathbf{1}\{Y_t(a_t) \notin \mathcal{C}_t(a_t)\} + \Delta_{\max}\sum_{t=1}^{T} I_t^{\exp}. \tag{EC.113}$$

The middle sum equals $T(1 - \mathrm{MCov}_T)$. Taking expectations and using $\mathbb{E}[\sum_t I_t^{\exp}] = \varepsilon T$ yields

$$\begin{aligned}\mathbb{E}[\widetilde{R}_T] &\le L\sum_{t=1}^{T} \mathbb{E}[\eta_t^{\mathrm{set}}] + BT(1 - \mathbb{E}[\mathrm{MCov}_T]) + \varepsilon T\Delta_{\max} \\ &\le L\sum_{t=1}^{T} \mathbb{E}[\eta_t^{\mathrm{set}}] + \alpha BT + BT\,\mathbb{E}|\mathrm{MCov}_T - (1-\alpha)| + \varepsilon T\Delta_{\max}.\end{aligned} \tag{EC.114}$$

This proves (EC.10). If $\ell \in [0, M]$, the perturbation $d_t(a_t^*) - d_t^*(a_t^*)$ is also at most M. Replacing its Lipschitz bound by $\min\{M, L\eta_t^{\mathrm{set}}\}$ throughout the argument proves the bounded-loss refinement, including infinite set distances. □

E.7 Proof of Corollary C.6

Proof. The Gaussian construction has $q_a^{\log} = q_0 + q_0\Delta/4 + O(\Delta^2)$ and $q_a^* = q_0 + q_0\Delta/2 + O(\Delta^2)$. Thus $|F_{\log}^{(a)}(q_a^*) - (1-\alpha)| \ge c_0(\alpha)\rho$. The Gaussian logged densities have a uniform upper bound L_0, so for every threshold, including infinite values,

$$L_0|\hat{q}_t(a) - q_a^*| \ge c_0(\alpha)\rho - |F_{\log}^{(a)}(\hat{q}_t(a)) - (1-\alpha)|. \tag{EC.115}$$

Choose c_0 smaller if necessary so that $c_0\rho \le 1$, and put $e_t(a) = |F_{\log}^{(a)}(\hat{q}_t(a)) - (1-\alpha)|$ and $\kappa = \min\{1, c_H/L_0\}$. The lower set comparison gives, pathwise,

$$\bar{\eta}_t^{\mathrm{set}} \ge \kappa\{c_0\rho - e_t(a)\}. \tag{EC.116}$$

Indeed, the right-hand side is nonpositive when $e_t(a) \ge c_0\rho$. Otherwise $0 < c_0\rho - e_t(a) \le 1$, and $\min\{1, (c_H/L_0)(c_0\rho - e_t(a))\} \ge \kappa(c_0\rho - e_t(a))$. This also covers infinite thresholds. Average and take expectations to obtain (EC.14) with $c' = \kappa c_0$ and $C_{\mathrm{tr}} = \kappa$. Substitution into the separately assumed (EC.15) proves part (a). Any fixed positive cap gives the same order with constants depending on that cap.

For part (b), first suppose $N = T\pi_{\min} \ge \log^2(2KT)$. Apply the localization proof of Theorem C.3 with failure probability $\delta_T = (2KT)^{-2}$. Its event E has probability at least $1 - \delta_T$, startup length $t_0 = O(\log(2KT)/\pi_{\min})$, and

$$\sum_{t=t_0}^{T} \eta_t \le CTr_T \quad \text{on } E, \qquad r_T = N^{-1/4}\sqrt{\log(2KT)}. \tag{EC.117}$$

For bounded losses, the robust-risk perturbation in the proof of Theorem C.5 is at most $\min\{M, LC_H\eta_t\}$. This bounds both infinite-threshold startup rounds and the complement of E without a score moment assumption. Its total expectation is at most

$$M\min\{T, t_0 - 1\} + LC_HCTr_T + MT\delta_T. \tag{EC.118}$$

Here the minimum is a bound on loss in the proof, not truncation of the algorithm's predictions.

Under the stationary logger, the conditional mean logged miscoverage before observing the current context is

$$b_t = \sum_a w_a\{1 - F^{(a)}_{\text{bias}}(\hat{q}_t(a))\}, \qquad \mathbb{E}[\mathbf{1}\{Y_t(a_t) \notin \mathcal{C}_t(a_t)\}] = \mathbb{E}[b_t]. \tag{EC.119}$$

On E and for $t \geq t_0$, the already assumed marginal logged-density bound gives $b_t \leq b^* + f_{\max}\eta_t$. Elsewhere use $b_t \leq 1$. Consequently,

$$\sum_t \mathbb{E}[b_t] \leq Tb^* + t_0 + Cf_{\max}Tr_T + T\delta_T. \tag{EC.120}$$

Insert these bounds into the signed miscoverage decomposition in Appendix E.6, retaining the exploration term. The startup and failure terms are absorbed into $O(Tr_T)$. If $N < \log^2(2KT)$, then $r_T > 1$ and the direct bound $\widetilde{R}_T \leq MT$ proves the same result. There is thus no effective-sample requirement in the corollary.

For the Gaussian construction, with $r = 1 + \Delta$ and F the unit-scale absolute-residual CDF, oracle logged coverage is

$$1 - b^* = \tfrac{3}{4}F(q^*) + \tfrac{1}{4}F(q^*/r) = 1 - \alpha + \tfrac{1}{4}\{F(q^*) - F(q^*/r)\} \geq 1 - \alpha. \tag{EC.121}$$

This proves the stated αBT specialization from the model itself. The fixed logger can also coincide with a robust decision rule: take $a^0(L) = 0$, $a^0(H) = 1$, loss $\ell(x, a, y) = A\mathbf{1}\{a \neq a^0(x)\} + \min\{|y - 5|, B_0\}$ with $A > B_0 > 0$, and uniform exploration probability $1/2$. Robust exploitation always selects $a^0(x)$, giving the required probabilities $3/4$ and $1/4$. This example verifies compatibility of the upper-bound assumptions, not the separate margin in part (a). □

E.8 Proof of Corollary C.7

Proof. Summing the CDF-based lower bound in Appendix E.7 over $t_0, \ldots, T$ gives

$$\begin{aligned}\frac{1}{T}\sum_{t=t_0}^{T}\mathbb{E}[\eta_t] &\geq (1 - b_T)c_\eta(\alpha)\rho - \frac{1}{L_0 T}\sum_{t=t_0}^{T}\mathbb{E}|F^{(a)}_{\log}(\hat{q}_t(a)) - (1-\alpha)| \\ &\geq (1 - b_T)c_\eta(\alpha)\rho - C_{\text{tr}}\varepsilon_T.\end{aligned} \tag{EC.122}$$

The PW-OCP statement is Theorem C.3. This comparison uses the same post-startup threshold quantity for both methods. It does not infer a raw-threshold expectation bound from a high-probability event. For decision regret, bounded loss controls the complement of that event directly, as shown in Appendix E.7. □

E.9 Proof of Corollary C.8

Proof. Set $N = T\pi_{\min}$, $L_T = \log(2KT)$, and $r_T = N^{-1/4}\sqrt{L_T}$. If $N < L_T^2$, then $r_T > 1$ and the result follows from $\mathcal{D}_T \leq 1$. Otherwise apply the proof of Theorem C.3 with $\delta_T = (2KT)^{-2}$. On its joint localization and recursion event E, all empirical thresholds from t_0 onward lie in $\widetilde{\mathcal{Q}}_a$ and their average error is at most Cr_T. Since $q^*(a) \in \mathcal{Q}_a$ and $f^{(a)}_{\text{true}} \leq \lambda_{\max}f_{\max}$ on this interval,

$$\max_a |F^{(a)}_{\text{true}}(\hat{q}_t(a)) - (1-\alpha)| \leq \lambda_{\max}f_{\max}\eta_t, \qquad t \geq t_0. \tag{EC.123}$$

On startup rounds and on E^c, the probability error is at most one, irrespective of threshold magnitude. Therefore

$$\mathbb{E}[\mathcal{D}_T] \leq \frac{\min\{T, t_0 - 1\}}{T} + Cr_T + \delta_T \leq C'r_T, \tag{EC.124}$$

using $t_0 - 1 = O(L_T/\pi_{\min})$ and $N \geq L_T^2$. The argument also covers $t_0 > T$ by the same startup bound. No moment bound for the raw thresholds is used. □

E.10 Proof of Proposition C.9

Proof. Write $p_0 = \pi_{\min}$, $p = 1-\alpha$, $\beta_t(a) = \alpha_t(a)$, $\beta_a^* = \alpha^\star(a)$, and $I_a = \mathcal{I}_a$. Use $F_b^{(a)}, F_c^{(a)}$ for the logged and counterfactual score CDFs and $q_a^* = q^*(a)$.

Suppress the action index when this creates no ambiguity. Define $h(\beta) = F_c(F_b^{-1}(1-\beta)) - p, \qquad \beta \in I_a$. Then $h(\beta^*) = 0$ and $-\lambda_{\max} \le h'(\beta) \le -\lambda_{\min}$. Choose a fixed $r > 0$ small enough that $J_{0,a} = [\beta_a^* - r, \beta_a^* + r], \qquad J_{1,a} = [\beta_a^* - 2r, \beta_a^* + 2r] \subset I_a$ for every action. This is an analysis choice based on the existing margins.

A uniform empirical-quantile event. Put $t_0 = 1 + \lceil C_0\Lambda/p_0 \rceil$. Under fixed contextual logging, each action's observed scores, in arrival order, are independent draws from its fixed logged law $F_b^{(a)}$. Chernoff bounds for action counts and DKW bounds for these arrival streams, unioned over actions, rounds, and buffer sizes at most T, give an event E_Q of probability at least $1 - \delta/3$ on which, for $t \ge t_0$,

$$n_t(a) \ge ctp_0, \qquad \sup_{\beta \in J_{1,a}} |\hat{Q}_t^{(a)}(1-\beta) - (F_b^{(a)})^{-1}(1-\beta)| \le C\sqrt{\frac{\Lambda}{tp_0}}. \tag{EC.125}$$

The supremum is over all indices, so it also covers the data-dependent index. Choosing C_0 sufficiently large makes the CDF error and the rank correction smaller than the fixed local margins. Local inversion then proves (EC.125) and keeps these empirical quantiles in the prescribed density neighborhood. No tail bound outside that neighborhood is used.

For all real β, put $H_t^{(a)}(\beta) = F_c^{(a)}(\hat{Q}_t^{(a)}(1-\beta)) - p$. The map $H_t^{(a)}$ is nonincreasing. Equation (EC.125) at the two endpoints of $J_{0,a}$ implies, on E_Q,

$$H_t^{(a)}(\beta) \le -b \text{ if } \beta \ge \beta_a^* + r, \qquad H_t^{(a)}(\beta) \ge b \text{ if } \beta \le \beta_a^* - r, \tag{EC.126}$$

where $b = \lambda_{\min} r/2$. Monotonicity extends these inequalities to indices outside the local interval, including those producing infinite thresholds.

Unconditional martingale events. Let $\mathcal{H}_{t-1}$ be the observed history. For this proof, write $Z_t(a)$ for the IPW or DR signal used by the respective algorithm. In either case, $M_t(a) = \mathbb{E}[Z_t(a) \mid \mathcal{H}_{t-1}] - Z_t(a)$ is a martingale difference, with magnitude at most C/p_0 and conditional second moment at most C/p_0. This follows from IPW unbiasedness, or DR unbiasedness with true propensities, after averaging over the independent current context. Thus

$$\beta_{t+1}(a) = \beta_t(a) + \gamma\{H_t^{(a)}(\beta_t(a)) + M_t(a)\}. \tag{EC.127}$$

Each actual step has magnitude at most γ/p_0, and $|H_t^{(a)}| \le 1$.

Maximal Freedman up to $t_0 - 1$, unioned over actions, gives an event of probability at least $1 - \delta/3$ on which $\max_{a,u<t_0} |\beta_{u+1}(a) - \alpha| \le \gamma t_0 + C\gamma\left\{\sqrt{\frac{t_0\Lambda}{p_0}} + \frac{\Lambda}{p_0}\right\} \le C\Lambda/\sqrt{N}$. Increase A so that $N \ge A\Lambda^2$ makes this at most one. In particular, $\beta_{t_0}(a) \in [-1, 2]$ for every action.

Choose $B \ge \max\{4/b, \log 2/\lambda_{\min}\}$ and partition the updates from t_0 into deterministic blocks of length $m = \lceil B/\gamma \rceil$. Maximal Freedman on each block, with a union bound over arms and blocks, gives another event of probability at least $1 - \delta/3$ on which every block starting at s satisfies

$$\gamma \max_a \max_{1 \le u \le m} \left| \sum_{i=s}^{s+u-1} M_i(a) \right| \le \epsilon_N := C_B \left\{ \sqrt{\frac{\gamma\Lambda}{p_0}} + \frac{\gamma\Lambda}{p_0} \right\}. \tag{EC.128}$$

Sums stop at T. Enlarging A ensures $\epsilon_N \le r/16$, $\gamma/p_0 \le r/16$, and $\gamma\lambda_{\max} \le 1$. It also ensures $t_0 + m \le T$ after enlarging A by another constant if needed. All these events were constructed for the original, unstopped process. We now reason pathwise on their intersection, which has probability at least $1 - \delta$.

Entrance and retention without projection. Each action enters $J_{0,a}$ during the first full block. Otherwise it stays on one side throughout, since a step smaller than $2r$ cannot jump over that interval. Equation (EC.126) then accumulates an inward displacement of at least $b\gamma m \ge 4$. The noise displacement is at most ϵ_N. This contradicts remaining on the same side when the block starts in $[-1, 2]$ and $J_{0,a} \subset (0, 1)$.

After an entrance into $J_{0,a}$, the index cannot exit $J_{1,a}$. For example, consider a candidate upper exit and the last preceding index at most $\beta_a^* + r$. The next index is at most $\beta_a^* + r + \gamma/p_0$. Until the candidate exit, drift is nonpositive and is at most $-b$ whenever an update begins above this boundary. Decompose this excursion into complete deterministic blocks and at most two partial blocks. Each complete block has net displacement at most $-b\gamma m + \epsilon_N \le 0$. The two partial blocks contribute at most $4\epsilon_N$ outward noise in total, by differences of the partial sums in (EC.128). The whole outward displacement is therefore at most

$\gamma/p_0 + 4\epsilon_N \le 5r/16 < r$, contradicting the upper exit. The lower exit is identical. Hence all actions remain in $J_{1,a}$ from the deterministic time $t_1 = t_0 + m$ onward.

Local contraction. For $t \ge t_1$, write $H_t^{(a)}(\beta_t(a)) = h_a(\beta_t(a)) + \xi_t(a)$, $|\xi_t(a)| \le C\sqrt{\Lambda/(tp_0)}$. The noiseless map $\beta \mapsto \beta + \gamma h_a(\beta)$ preserves $J_{1,a}$ and contracts toward β_a^* by a factor at most $1 - \gamma\lambda_{\min}$. Compare the noisy trajectory on each block with the noiseless trajectory starting at the same index. Both trajectories stay in $J_{1,a}$ on the event established above. Expanding their difference gives products of secants in $[0, 1 - \gamma\lambda_{\min}]$. The weights on successive noise increments are nondecreasing and bounded by one. Abel summation bounds the noise contribution by twice the maximal partial sum in (EC.128). This argument is pathwise and does not require the weights to be predictable. The bias contribution is at most $C \sup_{t \text{ in block}} \sqrt{\Lambda/(tp_0)}$.

Let $E_j = \max_a |\beta_{s_j}(a) - \beta_a^*|$ at successive block starts $s_j \ge t_1$. Complete blocks satisfy

$E_{j+1} \le \frac{1}{2}E_j + C\sqrt{\frac{\Lambda}{s_j p_0}} + 2\epsilon_N$, and the largest error within the block is at most $E_j + C\sqrt{\Lambda/(s_j p_0)} + 2\epsilon_N$. Summing the geometric recursion gives

$$\frac{1}{T}\sum_{t=t_1}^{T} \max_a |\beta_t(a) - \beta_a^*| \le C\left\{\frac{1}{\gamma T} + \sqrt{\frac{\Lambda}{N}} + \sqrt{\frac{\gamma\Lambda}{p_0}} + \frac{\gamma\Lambda}{p_0}\right\}. \tag{EC.129}$$

For the decreasing bias sequence, the first block contributes at most Cm/T and the later blocks are bounded by the integral of $C\sqrt{\Lambda/(tp_0)}$. This also handles a shorter final block. Local inverse-CDF regularity implies $|\hat{q}_t(a) - q_a^*| \le C\sqrt{\Lambda/(tp_0)} + |\beta_t(a) - \beta_a^*|/f_{\min}$. Substituting $\gamma = c_\gamma\sqrt{p_0/T}$ into (EC.129) and using $N \ge A\Lambda^2$ proves (EC.21).

Full-horizon probability error. Write $L = \log(2KT)$ and take $\delta = (2KT)^{-2}$, so $\Lambda = 3L$. If $N \ge 9AL^2$, the preceding event controls post-t_1 probability errors by the raw threshold error times the local density bound $\lambda_{\max} f_{\max}$. The initial fraction is

$\frac{t_1 - 1}{T} \le C\{L/N + N^{-1/2}\}$. Every probability error is at most one, including startup predictions and the complement of the event. Therefore $\mathbb{E}\left[\frac{1}{T}\sum_t \max_a |F_c^{(a)}(\hat{q}_t(a)) - p|\right] \le C\{N^{-1/4}\sqrt{L} + L/N + N^{-1/2}\} + (2KT)^{-2} \le CN^{-1/4}\sqrt{L}$. If $N < 9AL^2$, the rate expression is bounded below by a positive constant depending only on A, and the unit upper bound suffices after adjusting C. Combining with that same unit bound proves (EC.22) for every horizon. □

F Experimental Details

F.1 Benchmark Design

Common settings. Unless otherwise stated, all experiments use $\alpha = 0.1$, $T = 20{,}000$, and 10 random seeds. Baselines retain their default step size $\gamma = 0.005$. PW-OCP and DR-OCP use the coverage-motivated step size $\gamma^\star = \sqrt{\pi_{\min}/T}$ from Theorem 5.1, with $\pi_{\min}$ taken from the logging policy of each benchmark. Figures show means with error bars or bands spanning ± 1.96 standard errors across seeds.

DR-OCP fits the contextual outcome model $\hat{\mu}_t(a, X_t)$ in (15) as an exponential moving average (decay 0.01) of the per-arm-per-context miscoverage indicator. Both PW-OCP and DR-OCP clip the iterate $\alpha_t(a)$ to $[-0.5, 1.5]$ for numerical stability. This implementation choice is distinct from the unprojected coverage analysis. The nonconformity score is $s(x, y; a) = |y|$ for the mean-shift simulation and $s(x, y; a) = |y - \hat{\mu}(x, a)|$ for the variance-shift and multi-arm simulations.

Baseline implementation notes. ACI uses the original update [9]. Conformal PID implements the proportional+tan-saturated-integrator update of [1] (Scorecaster omitted in pure-bandit settings). SAOCP follows [4] with per-arm scale auto-tuned from the first observed score, matching the salesforce/online_conformal reference. AgACI uses BOA with pinball loss over a bank of $k = 8$ logarithmically spaced step sizes [36]. NEx-CP follows [3] with the recommended geometric decay weight. FACI follows [10] with $k = 8$ alpha-streams and sigma-mixing. ECI follows the quantile-space update of [33]; we expose it through the alpha-index parameterization with the corresponding sign-corrected error-quantification term. Online COPP uses the sliding-window inverse-propensity-weighted quantile in Algorithm 3.

All-action excess loss. The last column of Table 1 uses the mean-shift experiment at $\Delta = 3$ with a shared logger. For method m, its cumulative excess loss is

$$\mathcal{E}_m = 5 \sum_{t=1}^{T} \sum_{a \in \mathcal{A}} \left[(|Y_t(a)| - \hat{q}_t^m(a))_+ - (|Y_t(a)| - \hat{q}_t^{\text{Oracle}}(a))_+ \right]. \tag{EC.130}$$

The reported reduction is normalized as $100(1 - \overline{\mathcal{E}}_m / \overline{\mathcal{E}}_{\text{ACI}})\%$, with averages over seeds. This diagnostic evaluates threshold exceedances across all actions under the same logged data. It is distinct from the decision criteria $\widetilde{R}_T$ and R_T^*. Since increasing a threshold can only decrease its exceedance loss, this column is interpreted jointly with coverage and prediction-set width.

Mean-shift simulation. Two actions and two contexts, $\mathcal{X} = \{L, H\}$. Here Δ denotes the mean-shift magnitude. For $X = L$, $Y(0) \sim \mathcal{N}(0, 1)$ and $Y(1) \sim \mathcal{N}(\Delta, 1)$; the roles swap for $X = H$. The logger mixes a softmax distribution over negative absolute conditional means at temperature 0.3 with uniform exploration of probability $\varepsilon = 0.05$, giving the floor $\pi_{\min} = \varepsilon/2 = 0.025$.

Variance-shift simulation. Two actions and two contexts with conditional mean fixed at $\mu = 5$ throughout. Here Δ denotes the standard-deviation increment. In context L the standard deviations are $(\sigma_0, \sigma_1) = (1, 1 + \Delta)$; in context H they swap to $(1 + \Delta, 1)$. The residual score $|y - \hat{\mu}(x, a)|$ is paired with a risk-averse logger that applies softmax to negative conditional standard deviations with temperature 0.3 and mixes in uniform exploration with probability $\varepsilon = 0.05$.

Multi-arm scaling. Symmetric setup with K contexts and K actions. Context k favors action k (mean 0), while every other action has mean-shift magnitude $\Delta = 2$. We sweep $K \in \{2, 5, 10, 20, 50\}$. The logger uses the same softmax rule as in the mean-shift simulation, with uniform exploration probability $\varepsilon_K = \max\{0.05, 2/K\}$ and probability floor $\pi_{\min} = \varepsilon_K / K$. At $K = 20$, this yields $\varepsilon_K = 0.1$ and $\pi_{\min} = 0.005$.

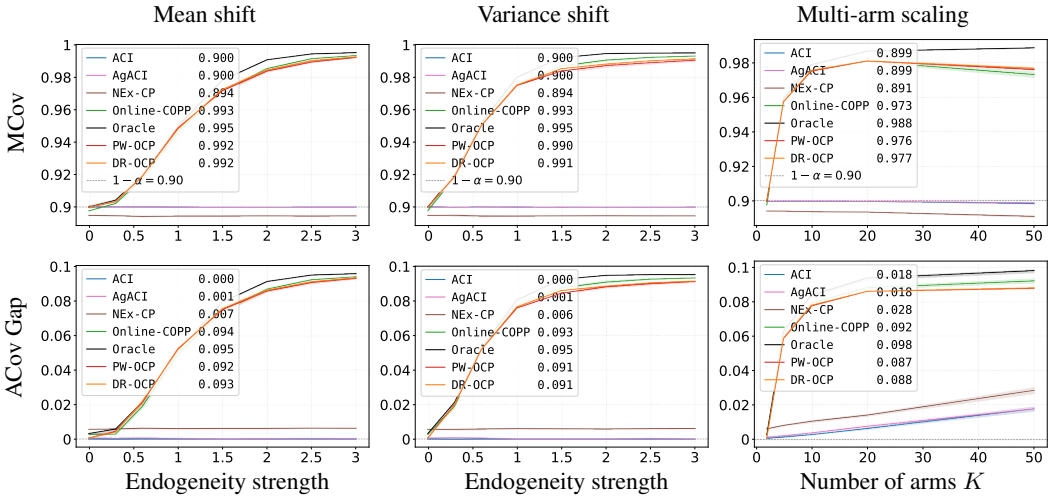

Figure 4: **Logged coverage diagnostics across synthetic settings.** The top row reports marginal coverage MCov_T and the bottom row reports action-conditional coverage gaps for the same mean-shift, variance-shift, and multi-arm settings as Figure 2. These logged diagnostics can appear well calibrated even when the corresponding counterfactual gaps remain large, illustrating why logged metrics alone are insufficient for counterfactual decision-making.

Propensity misspecification. The learner uses a biased propensity model

$$\hat{\pi} = (1 - \xi)\pi + \xi/K, \qquad \xi \in [0, 1], \tag{EC.131}$$

while data are generated under the true logging policy π. The master table reports the severe case $\xi = 1$ (uniform propensity model); Figure 3 sweeps ξ.

Rate verification and prediction-set quality. For the convergence-rate study, $K = 10$ is fixed while T ranges over $\{2{,}000, 5{,}000, 10{,}000, 20{,}000, 50{,}000\}$ and $\pi_{\min}$ ranges over $\{0.003, 0.005, 0.01, 0.02, 0.05\}$. The prediction-set quality study reports the running counterfactual coverage gap and finite-threshold error. At each round, we take the maximum oracle-quantile error over actions with finite thresholds, then average over rounds with at least one finite threshold. As T increases from 20,000 to 150,000, the mean proportion of non-finite action–round predictions decreases from 10.49% to 0.73% for PW-OCP and from 9.35% to 0.52% for DR-OCP.

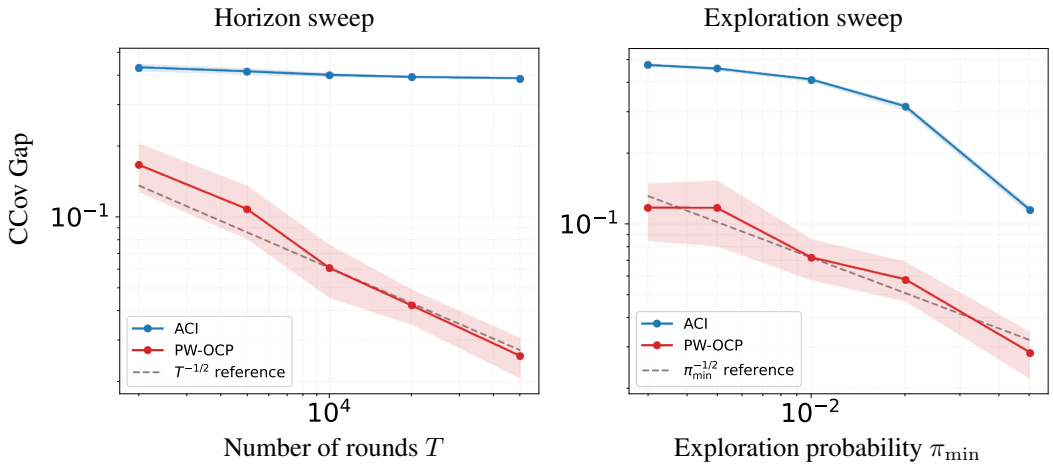

Figure 5: **Coverage-gap scaling for PW-OCP.** With $K = 10$, the left panel varies $T \in \{2{,}000, 5{,}000, 10{,}000, 20{,}000, 50{,}000\}$ at fixed $\pi_{\min} = 0.01$, and the right panel varies $\pi_{\min} \in \{0.003, 0.005, 0.01, 0.02, 0.05\}$ at fixed $T = 8{,}000$. Curves compare ACI and PW-OCP, with 95% confidence bands over 10 seeds. Dashed lines indicate inverse-square-root reference slopes.

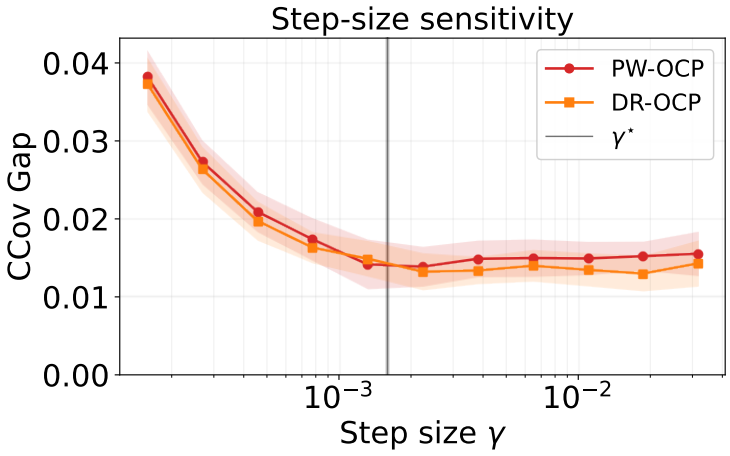

Figure 6: **Step-size sensitivity.** Counterfactual coverage gap is plotted against γ for $K = 5$, $T = 20{,}000$, and $\pi_{\min} = 0.05$. The dashed vertical line marks $\gamma^\star = \sqrt{\pi_{\min}/T} \approx 1.58 \times 10^{-3}$. PW-OCP performs best near this scale, and both methods remain stable over a broad neighborhood of $\gamma^\star$. Points show means, and error bars indicate 95% confidence intervals over 20 seeds.

F.2 Open Bandit and Financial Logging

The Open Bandit experiments use the ZOZOTOWN Open Bandit Dataset [23]. For both Women and Men campaigns, we keep the top 15 items by Bernoulli Thompson sampling frequency and aggregate user features into four contexts. We estimate item-context click rates from random-policy data and use these estimates to construct continuous potential rewards $Y(a) \sim \mathcal{N}(\mu(x,a), \sigma(x,a)^2)$ while preserving observed heterogeneity. The DJIA experiment uses a momentum-biased logging policy over AAPL, MSFT, BA, and KO across $T = 2{,}500$ trading days.

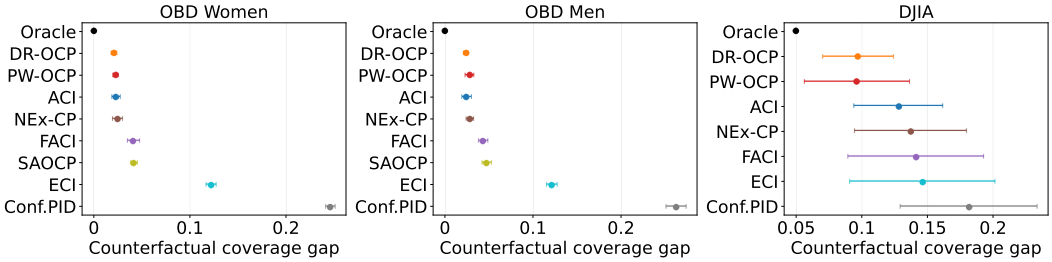

Figure 7: **Counterfactual coverage on Open Bandit and DJIA tasks.** The three panels report the same counterfactual coverage-gap metric as Table 1 for OBD Women, OBD Men, and DJIA portfolio rebalancing. Absolute counterfactual gaps are smaller than in the synthetic stress tests, and PW-OCP and DR-OCP remain among the best non-oracle methods across the three benchmarks.

F.3 Prediction-set efficiency

The experiments below use 20 seeds, $\alpha = 0.1$, and $T = 20{,}000$. PW-OCP and DR-OCP use $\gamma = \sqrt{\pi_{\min}/T}$ and indices in $[0.001, 0.999]$. Online COPP uses $W = 500$. Finite-set gaps average miscoverage over finite-threshold rounds for each action before taking the maximum deviation from α. Interval-score excess is relative to Oracle. Full-set rates are computed over action-round pairs.

Table 2 uses five actions, mean shift $\Delta = 2$, and exploration mixture $\varepsilon = 0.25$. Full-set rates after the first $1{,}000$ rounds are below 0.001. ACI under-covers the three context-selected actions, whereas PW-OCP remains near oracle width on every action. A symmetric control isolates frequency alone: the frequent action's width changes by -0.4% of oracle width under skewed allocation, with normal-approximation 95% CI $[-1.0\%, 0.2\%]$, within the prespecified $\pm 5\%$ margin.

Table 3: Per-action pull shares, oracle-relative widths, and finite-set gaps. The first three actions are selected according to context, and the last two through uniform exploration.

Action	Pull share	PW width	PW gap	ACI width	ACI gap
a_1	0.50	95.3%	0.030	64.8%	0.186
a_2	0.20	100.2%	0.016	71.7%	0.222
a_3	0.20	99.9%	0.018	71.5%	0.223
a_4	0.05	101.3%	0.009	100.9%	0.009
a_5	0.05	101.0%	0.007	100.0%	0.007

F.4 Robustness to limited exploration and imperfect feedback

Exploration. With $K = 10$, a larger floor improves counterfactual calibration at an increased logging cost. In Table 4, ESS is the minimum action-wise importance-weight effective sample size and logging regret is mean excess loss relative to the context-wise best action. At $\pi_{\min} = 0.001$, increasing T from $5{,}000$ to $80{,}000$ reduces DR-OCP's finite-set gap from 0.361 to 0.144.

Table 4: Exploration tradeoff. Gaps are finite-set means with 95% Student-t confidence intervals.

$\pi_{\min}$	DR-OCP gap	Online COPP gap	Min-action ESS	Logging regret
0.001	0.226 ± 0.011	0.149 ± 0.018	40	0.040
0.01	0.099 ± 0.007	0.048 ± 0.008	222	0.199
0.05	0.020 ± 0.003	0.022 ± 0.003	1062	0.910

Nuisance misspecification. The outcome model groups ten contexts into b categories, while the supplied propensities are $(1-\lambda)\pi + \lambda/K$. The largest gaps occur when both nuisances are inaccurate.

Table 5: Finite-set gaps under nuisance misspecification, with 95% Student-t confidence intervals.

Method / nuisance	$\lambda = 0$	$\lambda = 1$
DR-OCP, $b = 1$	0.0652 ± 0.0064	0.3056 ± 0.0047
DR-OCP, $b = 2$	0.0574 ± 0.0060	0.1313 ± 0.0050
DR-OCP, $b = 10$	0.0575 ± 0.0048	0.0584 ± 0.0051
DR-OCP, analytic nuisance	0.0555 ± 0.0047	0.0114 ± 0.0013
Online COPP	0.0360 ± 0.0040	0.3158 ± 0.0062

Nonstationarity and delay. A persistent 0.75 mean shift after $10{,}000$ rounds raises DR-OCP's whole-horizon gap from 0.0535 to 0.0673 and Online COPP's from 0.0349 to 0.0499. Over the first 250 rounds after the shift, the gaps are 0.287 for DR-OCP and 0.264 for Online COPP. Online COPP also handles long delays better. Delayed updates use prediction-time thresholds, propensities, and nuisance estimates. The oracle recursion's failure shows that stale feedback matters even without propensity weighting.

Table 6: Finite-set gaps under delayed feedback. Full-set rates at $d = 1{,}000$ include cold start.

Method	$d = 0$	$d = 200$	$d = 1{,}000$	Full sets
DR-OCP	0.0575	0.0748	0.1374	5.0%
Online COPP	0.0360	0.0362	0.0377	5.0%
Oracle recursion	0.0003	0.0020	0.2392	68.3%

Open Bandit propensity corruption. The semi-synthetic experiment uses 15 actions, context-specific logging frequencies with floor 0.02, and a score center fitted on a separate split. Corruption changes only the supplied propensities. AgACI remains unchanged, while DR-OCP is more stable than PW-OCP.

Table 7: Open Bandit finite-set gaps under supplied propensities $(1-\lambda)\pi + \lambda/K$.

Campaign	λ	ACI	AgACI	Online COPP	PW-OCP	DR-OCP
Men	0.0	0.0257	0.0258	0.0371	0.0270	0.0255
Men	1.0	0.0257	0.0258	0.0369	0.2284	0.0372
Women	0.0	0.0231	0.0229	0.0325	0.0238	0.0216
Women	1.0	0.0231	0.0229	0.0332	0.2023	0.0246

F.5 Score normalization

All variants share the true center $\mu_0 = 5$ and vary only the scale in $|y - \mu_0|/\widehat{\sigma}(x, a)$. Exact normalization makes the score pivotal and removes the present impossibility construction without changing outcome-level ρ. Scale estimates fitted on past data progressively improve ACI's counterfactual calibration, with structured shrinkage pooling diagonal and off-diagonal context-action groups.

Table 8: Scale-estimator RMSE and whole-horizon counterfactual gaps. True-scale RMSE is zero by construction.

Scale choice	Scale RMSE	ACI	PW-OCP	DR-OCP
Raw residual	—	0.2827	0.0400	0.0289
True scale	0	0.0069	0.0137	0.0142
EMA scale	0.373	0.1161	0.0164	0.0149
Shrinkage scale	0.156	0.0350	0.0145	0.0129
Structured shrinkage	0.027	0.0078	0.0136	0.0127

F.6 Prediction-dependent logging

The logger ranks actions by predicted loss plus a weighted conformal half-width:

$$r_t(a) = \widehat{m}_t(X_t, a) + 0.20\widehat{q}_t(a), \quad \pi_t(a|X_t, \mathcal{H}_{t-1}) = \frac{\varepsilon}{K} + (1-\varepsilon)\frac{\exp[-r_t(a)/\tau]}{\sum_b \exp[-r_t(b)/\tau]}. \tag{EC.132}$$

We use $K = 5$, $\varepsilon = 0.20$, and $\tau = 0.05$. The loss center is fitted from past feedback, and the exact sampling probabilities are recorded. The homoscedastic design has standard deviation 0.50 for all actions. The heteroscedastic design uses 0.25 for the context-wise best action and 0.75 otherwise. The same policy formula is used for both designs, with trajectories determined by each method's prediction sets.

Table 9: Finite-set counterfactual gaps when prediction sets determine logging probabilities.

Outcome scale	ACI	PW-OCP	DR-OCP
Homoscedastic	0.0084	0.0169	0.0145
Heteroscedastic	0.2895	0.0382	0.0302

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction state the paper's main contributions: the counterfactual-coverage formulation, the impossibility result, PW-OCP/DR-OCP with coverage guarantees, and decision implications. The scope is explicitly limited to one-step contextual decision settings with positive logging propensities.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Section 8 and the scope remarks discuss the main limitations, including positivity, one-step rather than full-RL counterfactuals, nuisance-estimation assumptions for DR-OCP, and stationarity/density regularity for sharpness and regret results. The paper also comments on computational scaling through the action-wise online updates.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: Each theorem states or references its assumptions, with the core coverage and lower-bound results relying on no-unmeasured-confounding and positivity, and the sharpness/regret results adding stationarity and quantile regularity. Complete proofs are provided in the appendix, including the identification lemma, martingale bounds, impossibility construction, and minimax lower bound.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Appendix F specifies the experimental settings, including hyperparameters, horizons, arm counts, context construction, propensity floors, score functions, seed counts, and the ZOZOTOWN/DJIA preprocessing protocols. The algorithms and baselines are also described in Appendix B and the implementation notes.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.

(c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: An anonymized code archive is included in the supplementary material.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Appendix F specifies the online experimental design, shared streams within each seed, fixed hyperparameters, and DR-OCP outcome-model construction. No train/test split is used because the evaluation is sequential online calibration.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

Justification: Quantitative results report means with uncertainty over 10–20 random seeds, and the paper states the error-bar convention for tables and figures. These seeds capture outcome noise and algorithmic randomness from the logging policy.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: All reported experiments are lightweight CPU experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: The work is theoretical and methodological, uses only public or synthetic benchmark data, and does not involve human subjects, scraped personal data, deception, or high-risk model release.

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

Justification: The paper discusses positive impacts such as more reliable action-wise uncertainty quantification under selection-biased logging, which can support safer decision-making for rarely selected actions. It also discusses negative risks from misuse under violated assumptions, especially silent positivity violations in deployed logging systems.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper releases neither pretrained models nor scraped datasets.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The paper cites the ZOZOTOWN Open Bandit Dataset and uses it according to its released terms, and DJIA data are accessed through standard public-data tooling without redistributing raw prices. Standard scientific Python dependencies are documented in the supplementary code release.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The released code archive includes documentation for the proposed algorithms, baselines, configuration files, and reproduction commands. Per-setting configuration files record the hyperparameters and seeds used in the reported experiments.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing or research with human subjects.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: No human subjects research is conducted; IRB review is therefore not applicable.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: LLMs are not a component of the proposed methods or experimental pipeline. Any use during manuscript preparation was limited to writing or editing assistance and did not affect the scientific content, methodology, or originality of the work.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.