DISAGREEMENT-ADAPTIVE SAFE POLICY UPDATING UNDER PROTECTED SUBGROUP CONSTRAINTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Safe policy improvement from offline data seeks to improve upon an incumbent policy while satisfying prespecified safety constraints. Existing work has studied high-confidence policy improvement under multiple fairness or safety constraints and disagreement-aware policy comparison, while it remains unclear whether candidate–incumbent disagreement can sharpen finite-sample certification across protected subgroups. To address this gap, we study safe policy updating when candidate policies may differ from the incumbent only on a small fraction of each evaluation population. We show that direct estimation of candidate–incumbent value differences yields simultaneous confidence bounds whose leading width scales as $\sqrt{p/n}$, where p is the disagreement probability and n is the audit sample size, and establish a protected-subgroup lower bound together with a local minimax result showing that this $\sqrt{p/n}$ dependence cannot be uniformly improved near a binding subgroup constraint. We then develop DiSPU, a **di**sagreement-adaptive framework for **s**afe **p**olicy **u**pdating (DiSPU) that incorporates these confidence bounds into a constrained deployment optimization, and prove high-probability safety and disagreement-dependent oracle-regret guarantees. Extensive experimental results confirm that exploiting policy disagreement reduces certification conservatism and enables larger certified improvements while maintaining protected-subgroup safety.

1 INTRODUCTION

Offline policy learning uses historical data to learn decision rules that map individual characteristics to actions with the goal of improving outcomes upon deployment (Dudík et al., 2011; Swaminathan & Joachims, 2015; Joachims et al., 2018; Kallus, 2018; Kitagawa & Tetenov, 2018; Athey & Wager, 2021). In many deployment settings, however, an incumbent policy is already in use, so a newly learned policy must be evaluated relative to that incumbent before deployment. This setting is commonly studied as safe policy improvement, where a candidate update is deployed only when the available data provide sufficient evidence that it satisfies a prespecified performance guarantee relative to the incumbent (Thomas et al., 2015; Laroche et al., 2019; Karampatziakis et al., 2021). Reliable baseline-relative certification is therefore essential for deploying learned policy updates from offline data.

Baseline-relative certification should, at a minimum, ensure that a policy update does not reduce expected outcomes in the target population. However, population-average improvement can coexist with substantial losses for particular subgroups, which is problematic in high-stakes applications such as healthcare and resource allocation (Ustun et al., 2019; Jia et al., 2026). Safe deployment under protected-subgroup requirements therefore calls for simultaneous guarantees of overall non-degradation and prespecified limits on subgroup-specific degradation relative to the incumbent. Prior work has incorporated high-probability fairness constraints into offline policy learning (Metevier et al., 2019) and multiple baseline-relative safety constraints into safe policy improvement (Thomas et al., 2021). In finite samples, however, these guarantees rely on uncertain off-policy comparisons between the candidate and incumbent, and wide uncertainty can rule out beneficial updates even when the population constraints are satisfied.

A natural way to sharpen these comparisons is to exploit policy disagreement. When a candidate policy and the incumbent prescribe the same action for an individual, the corresponding contribu-

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

tion to their policy value difference is zero. Direct estimation of the policy value difference can therefore yield more precise comparisons when the candidate differs from the incumbent only on a small fraction of the population (Laroche et al., 2019; Cho et al., 2025; Sawarni et al., 2026). Prior work has used this property to improve the precision of policy comparisons or certify overall policy improvement (Narang et al., 2024; Cho et al., 2025). Protected-subgroup constraints require several baseline-relative comparisons to be certified simultaneously, with potentially different disagreement probabilities across the overall population and protected subgroups. Existing disagreement-based methods do not exploit this subgroup-specific disagreement to sharpen simultaneous finite-sample certification and the resulting constrained deployment. This motivates us to study the following question:

> *How can candidate–incumbent disagreement be exploited to certify safe policy improvement under prespecified protected-subgroup constraints?*

In this paper, we introduce DiSPU, a **dis**agreement-adaptive framework for **s**afe **p**olicy **u**pdating under protected-subgroup constraints. DiSPU directly estimates candidate–incumbent value differences using paired doubly robust scores and constructs simultaneous lower confidence bounds that adapt to policy disagreement. These bounds are incorporated into a constrained optimization that selects a mixture over the incumbent and candidate policies while enforcing overall and subgroup safety requirements. We establish finite-sample safety and disagreement-dependent oracle-relative performance guarantees, together with a protected-subgroup lower bound and a local minimax result showing that the $\sqrt{p/n}$ dependence is statistically unavoidable near a binding subgroup constraint.

Our **contributions** are threefold: (i) We develop a disagreement-adaptive approach to safe policy improvement under multiple protected-subgroup constraints, combining direct candidate–incumbent value-difference estimation with confidence-adjusted deployment optimization. (ii) We derive simultaneous confidence bounds with leading width $\sqrt{p/n}$, where p is the probability that the candidate and incumbent disagree, and use them to obtain safety and regret guarantees. A local minimax lower bound shows that this rate cannot be improved near a binding subgroup constraint. (iii) Extensive experimental results show that direct paired certification can reduce conservatism and increase certified gains while maintaining protected-subgroup safety.

2 Related Work

Safe offline policy improvement. Safe offline policy improvement aims to improve upon a baseline policy from logged data while controlling performance degradation. Existing approaches either certify candidate policies statistically, using high-confidence tests or confidence sequences (Thomas et al., 2015; Karampatziakis et al., 2021), or incorporate safety directly into policy optimization, through robust baseline-relative objectives, baseline bootstrapping, or pessimistic value estimation (Petrik et al., 2016; Laroche et al., 2019; Cheng et al., 2022). However, these methods primarily focus on population-level performance guarantees and do not explicitly address safety requirements across protected subgroups.

Policy learning under fairness and subgroup constraints. Learning under fairness and subgroup constraints aims to control how performance or outcomes vary across subpopulations. In classification, subgroup auditing and subgroup-specific guarantees can reveal or reduce disparities hidden in aggregate comparisons (Kearns et al., 2018; Ustun et al., 2019). These ideas have been extended to policy learning with causal fairness constraints (Nabi et al., 2019), to offline decision making with multiple baseline-relative fairness or safety constraints (Metevier et al., 2019; Thomas et al., 2021), and to Bayesian safe policy learning with distributional constraints on baseline-relative conditional risk (Jia et al., 2026). However, none of these methods exploits localized disagreement between the candidate and the incumbent to sharpen finite-sample certification across protected subgroups.

Policy difference estimation. Exploiting policy disagreement can sharpen statistical guarantees for policy comparison. In online and reinforcement learning, disagreement has been used to characterize instance-dependent sample complexity and to reduce the cost of comparing policies by estimating differences relative to a reference policy (Li et al., 2022; Narang et al., 2024). In offline evaluation, Cho et al. (2025) construct doubly robust estimators of value differences between threshold policies, where observations on which the policies agree contribute zero. Sawarni et al. (2026)

instead use disagreement among nearly optimal policies to identify uncertain regions and achieve safe policy improvement by reverting to the baseline on these regions. However, these methods do not use localized policy differences to obtain simultaneous finite-sample safety guarantees across protected subgroups.

3 PROBLEM SETUP

Offline policy learning. We first introduce the standard policy learning setup (Athey & Wager, 2021). Let $O_i = (X_i, A_i, Y_i)$, $i = 1, \ldots, n$, be i.i.d. observations, where $X_i \in \mathcal{X}$ denotes pre-treatment covariates, $A_i \in \mathcal{A} = \{0, \ldots, K-1\}$ is the logged action, and $Y_i \in \mathbb{R}$ is the observed outcome, with larger values preferred. For each $a \in \mathcal{A}$, let $Y_i(a)$ denote the potential outcome under action a. To identify policy values from the observed data, we impose the following standard causal assumptions.

Assumption 1 (Causal identification)**.** *We impose the standard identification conditions: (i)* consistency*:* $Y = Y(A)$*; (ii)* unconfoundedness*:* $\{Y(a) : a \in \mathcal{A}\} \perp A \mid X$*; and (iii)* overlap*: letting* $e(a \mid x) := P(A = a \mid X = x)$ *denote the logging propensity, there exists* $\kappa > 0$ *such that* $e(a \mid x) \geq \kappa$ *for every* $a \in \mathcal{A}$ *and almost every* x*. We additionally assume no interference across units.*

We consider deterministic policies, i.e., measurable maps $d : \mathcal{X} \to \mathcal{A}$. The value of a policy d is $V(d) := E[Y(d(X))]$. Under Assumption 1, $V(d)$ is identified from the observed-data distribution. Our main finite-sample analysis assumes that the logging propensities are recorded and hence known. Appendix B considers estimated propensities.

Evaluation populations and candidate policies. We evaluate policy updates on an overall deployment population and a collection of prespecified protected subgroups. Let $s = 0$ index the overall deployment population, let $\mathcal{H} = \{1, \ldots, H\}$ index the protected subgroups, and define $\mathcal{S} = \{0\} \cup \mathcal{H}$. For each $s \in \mathcal{S}$, let P_s denote the corresponding evaluation distribution and let E_s denote expectation under P_s. When subgroup s is defined by an observed membership indicator G_s, the usual protected-subgroup setting has $P_s = P_0(\cdot \mid G_s = 1)$. We observe an audit sample $\mathcal{D}_s^{\mathrm{aud}} = \{O_{is} = (X_{is}, A_{is}, Y_{is})\}_{i=1}^{n_s}$, with $n_s \geq 2$, whose observations are i.i.d. under P_s. The protected subgroups need not be disjoint, and we do not require independence across the audit samples indexed by s. Let $Y_s(a)$ denote the potential outcome under action a for a generic unit indexed by s, and let $e_s(a \mid x) := P_s(A_s = a \mid X_s = x)$. Assumption 1 holds under every P_s. For a policy d, define its value under P_s as $V_s(d) := E_s[Y_s(d(X_s))]$.

Let $b : \mathcal{X} \to \mathcal{A}$ denote the incumbent policy. An arbitrary training procedure, using data independent of the audit samples, produces a finite collection of candidate policies $\{d_1, \ldots, d_M\}$. We set $d_0 = b$. For each $s \in \mathcal{S}$ and $j \in \{0, \ldots, M\}$, define the policy value difference between candidate d_j and the incumbent, together with their action disagreement probability, by $D_{sj} := V_s(d_j) - V_s(b)$ and $p_{sj} := P_s(d_j(X_s) \neq b(X_s))$. The quantity D_{sj} is the policy value difference to be certified under P_s, while p_{sj} is the disagreement probability between candidate d_j and the incumbent under P_s.

Deployment under protected-subgroup constraints. Following prior work on conservative and high-confidence policy improvement (Kakade & Langford, 2002; Thomas et al., 2015), we do not restrict the update to replacing the incumbent with a single candidate policy. Instead, we allow the deployment policy to mix the incumbent and the candidate policies. This formulation provides additional flexibility to trade off overall improvement against protected-subgroup constraints, while retaining the incumbent as a feasible fallback. Define $\Delta_M := \left\{\theta \in [0,1]^{M+1} : \sum_{j=0}^{M} \theta_j = 1\right\}$, where Δ_M is the simplex of deployment weights over the incumbent and candidate policies. The induced stochastic policy is $\pi_\theta(a \mid x) := \sum_{j=0}^{M} \theta_j \mathbf{1}\{d_j(x) = a\}$. Its policy value difference from the incumbent under P_s is $D_s(\theta) := \sum_{j=0}^{M} \theta_j D_{sj}$.

For each protected subgroup $s \in \mathcal{H}$, let $\Delta_s \geq 0$ denote the maximum allowable degradation relative to the incumbent. If the population quantities D_{sj} were known, the oracle deployment would solve $W^* := \max_{\theta \in \Delta_M} D_0(\theta)$ subject to $D_s(\theta) \geq -\Delta_s$ for all $s \in \mathcal{H}$. The choice $\theta = (1, 0, \ldots, 0)$ deploys the incumbent policy for every unit and satisfies $D_s(\theta) = 0$ for all $s \in \mathcal{S}$. Hence, the oracle optimization problem is always feasible and $W^* \geq 0$.

Definition 1 (High-confidence safe update)**.** *Fix $\alpha \in (0,1)$. A data-dependent deployment $\widehat{\theta} \in \Delta_M$ is a $(1-\alpha)$-safe update if* $\Pr\Big(D_0(\widehat{\theta}) \geq 0$ *and* $D_s(\widehat{\theta}) \geq -\Delta_s$ *for all* $s \in \mathcal{H}\Big) \geq 1-\alpha$.

Our goal is to construct a $(1-\alpha)$-safe deployment whose improvement $D_0(\widehat{\theta})$ is as close as possible to the oracle value W^*.

4 Disagreement-Adaptive Safe Policy Updating

In this section, we develop a disagreement-adaptive safe policy updating procedure. We directly estimate candidate–incumbent policy value differences from independent audit data, construct high-confidence lower bounds, and incorporate these bounds into a protected-subgroup-constrained optimization to determine the deployment mixture over the incumbent and candidate policies.

Let $\mathcal{D}^{\mathrm{tr}}$ denote the training sample, which is independent of the audit samples $\{\mathcal{D}_s^{\mathrm{aud}} : s \in \mathcal{S}\}$ defined in Section 3. The training data $\mathcal{D}^{\mathrm{tr}}$ are used to construct the candidate policy library and fit the outcome regression estimators. The audit samples are reserved for certification and are not used to fit either the candidate policies or the outcome regressions. We clip $\widehat{m}_s(x,a)$ to $[-B,B]$ and impose the following bounded-outcome assumption.

Assumption 2 (Bounded outcomes)**.** *There exists $B < \infty$ such that $|Y_s(a)| \leq B$ almost surely for all $s \in \mathcal{S}$ and $a \in \mathcal{A}$.*

For each candidate policy d_j and each $s \in \mathcal{S}$, we directly estimate D_{sj}, the policy value difference from the incumbent, using a paired doubly robust score. The score directly contrasts the candidate and incumbent on the same audit observations and thereby avoids separate policy-value estimation.

$$Z_{sj}(O) = \sum_{a \in \mathcal{A}} \big[\mathbf{1}\{d_j(X)=a\} - \mathbf{1}\{b(X)=a\}\big] \left[\widehat{m}_s(X,a) + \frac{\mathbf{1}\{A=a\}}{e_s(a \mid X)}\{Y - \widehat{m}_s(X,a)\}\right]. \tag{1}$$

The paired score compares the candidate and incumbent on the same audit observation. Conditional on the training data, it satisfies $\mathbb{E}_s[Z_{sj}(O) \mid \mathcal{D}^{\mathrm{tr}}] = D_{sj}$ when the logging propensity is known, without requiring $\widehat{m}_s$ to equal the true outcome regression. It also satisfies $Z_{sj}(O) = 0$ whenever $d_j(X) = b(X)$, so observations in the policy agreement region make no contribution to either the estimated value difference or its empirical variability. Given the audit sample $\mathcal{D}_s^{\mathrm{aud}}$, we estimate D_{sj} by the empirical mean of the paired scores and quantify its sampling variability using the sample variance, as defined below.

$$\widehat{D}_{sj} := \frac{1}{n_s}\sum_{i=1}^{n_s} Z_{sj}(O_{is}), \quad \widehat{\sigma}_{sj}^2 := \frac{1}{n_s - 1}\sum_{i=1}^{n_s}\Big(Z_{sj}(O_{is}) - \widehat{D}_{sj}\Big)^2. \tag{2}$$

A point estimate $\widehat{D}_{sj}$ alone is insufficient for the high-confidence guarantee in Definition 1, as it does not account for finite-sample estimation uncertainty (Thomas et al., 2015; Metevier et al., 2019). We therefore construct simultaneous lower confidence bounds for the policy value differences. Specifically, for each $s \in \mathcal{S}$ and $j \geq 1$, we define the empirical Bernstein radius

$$c_{sj} := \sqrt{\frac{2\widehat{\sigma}_{sj}^2 \ell_\alpha}{n_s}} + \frac{14 L \ell_\alpha}{3(n_s - 1)}. \tag{3}$$

Here, $\ell_\alpha := \log(8N/\alpha)$ is the corresponding logarithmic factor, $N := M|\mathcal{S}|$ is the number of policy value differences controlled simultaneously, and $L := 2B(1+\kappa^{-1})$ is the uniform bound on the absolute value of the paired doubly robust score. The motivation for this radius and its finite-sample advantages are established in Theorem 1 of Section 5. Using this confidence radius, we then define the lower confidence bound $\underline{D}_{sj} := \widehat{D}_{sj} - c_{sj}$ for each $s \in \mathcal{S}$ and $j \geq 1$. This quantity provides a high-confidence lower bound on the policy value difference D_{sj} and will be used directly in the subsequent policy update optimization. Given the lower confidence bounds $\{\underline{D}_{sj}\}$, we select the

Algorithm 1 Disagreement-Adaptive Safe Policy Updating

Input: Training logs $\mathcal{D}^{\mathrm{tr}}$; audit logs $\{\mathcal{D}_s^{\mathrm{aud}} : s \in \mathcal{S}\}$ independent of the training data; incumbent policy b; subgroup tolerances $\{\Delta_s\}_{s\in\mathcal{H}}$; failure level α.

Use $\mathcal{D}^{\mathrm{tr}}$ to construct candidate policies $\{d_1, \ldots, d_M\}$ and clipped outcome regressions $\{\widehat{m}_s\}_{s\in\mathcal{S}}$; set $d_0 = b$.
for $s \in \mathcal{S}$ **do**
  for $j = 1, \ldots, M$ **do**
    Compute the paired scores $Z_{sj}(O_{is})$ using equation 1.
    Compute $\widehat{D}_{sj}$ and $\widehat{\sigma}_{sj}^2$ using equation 2.
    Compute the empirical Bernstein radius c_{sj} using equation 3.
    Set $\underline{D}_{sj} \leftarrow \widehat{D}_{sj} - c_{sj}$.
  end for
end for
Solve equation 4 to obtain $\widehat{\theta} = (\widehat{\theta}_0, \ldots, \widehat{\theta}_M)$.
Construct the deployment policy $\pi_{\widehat{\theta}}$ from $\widehat{\theta}$ and $\{d_0, d_1, \ldots, d_M\}$.

Output: Rollout weights $\widehat{\theta}$ and the resulting safe stochastic policy $\pi_{\widehat{\theta}}$.

deployment weights $\theta \in \Delta_M$ by solving the following confidence-adjusted linear program:

$$\widehat{\theta} \in \arg\max_{\theta\in\Delta_M} \sum_{j=0}^{M} \theta_j \underline{D}_{0j} \quad \text{s.t.} \quad \sum_{j=0}^{M} \theta_j \underline{D}_{sj} \geq -\Delta_s, \qquad \forall s \in \mathcal{H}. \tag{4}$$

The objective maximizes the lower confidence bound on overall policy improvement, while each protected-subgroup constraint requires the corresponding lower confidence bound to remain above the allowable degradation level $-\Delta_s$. The incumbent is always feasible since its value difference and confidence radius are both zero, providing a safe fallback when no candidate can be certified. Allowing mixtures over the policy library accommodates heterogeneous trade-offs across protected subgroups. Algorithm 1 summarizes the resulting procedure.

5 Theoretical Analysis

We now establish the finite-sample guarantees for our disagreement-adaptive safe policy updating procedure. We first show that the confidence radius for each policy value difference D_{sj} adapts to the disagreement probability p_{sj} between the candidate policy d_j and the incumbent policy b, yielding a $\sqrt{p_{sj}/n_s}$ rate. We then establish a protected-subgroup lower bound and a local minimax result at the same $\sqrt{p/n}$ scale near a binding subgroup constraint. Finally, we propagate these confidence bounds through the deployment optimization to obtain high-probability safety and an oracle-relative regret guarantee. Detailed proofs of all theoretical results are provided in Appendix A.

5.1 Disagreement-Adaptive Finite-Sample Confidence Bounds

We first establish simultaneous finite-sample validity of the empirical Bernstein radii used by our procedure. Conditioning on the realized training data, the following result shows that $\widehat{D}_{sj}$ concentrates uniformly around D_{sj} over all evaluation populations and candidate policies.

Theorem 1 (Simultaneous finite-sample bounds)**.** *Suppose Assumptions 1-2 hold. For any fixed realization of the training data $\mathcal{D}^{\mathrm{tr}}$, with probability at least $1 - \alpha/2$ over the audit data,*

$$\left|\widehat{D}_{sj} - D_{sj}\right| \leq c_{sj}, \qquad \forall\, s \in \mathcal{S},\ j \in \{0, \ldots, M\}, \tag{5}$$

where c_{sj} is given explicitly in equation 3.

Theorem 1 establishes simultaneous finite-sample validity, but does not yet make explicit how the resulting confidence radius depends on the action disagreement between d_j and b. We next show that the empirical variance appearing in c_{sj} is itself controlled by the fraction of audit observations on which the two policies disagree.

Proposition 1 (Empirical disagreement controls the confidence radius). *For any realized audit sample, any $s \in \mathcal{S}$, and any $j \in \{1, \dots, M\}$, $\widehat{\sigma}_{sj}^2 \le \frac{n_s}{n_s-1} L^2 \widehat{p}_{sj}$. Consequently, the empirical Bernstein radius defined in equation 3 satisfies*

$$c_{sj} \le L\sqrt{\frac{2\widehat{p}_{sj}\ell_\alpha}{n_s-1}} + \frac{14L\ell_\alpha}{3(n_s-1)}.$$

Proposition 1 applies to the realized confidence radii used in the deployment optimization. If a candidate changes the incumbent action on only a small fraction of the audit sample, the paired score has correspondingly low empirical variance. Corollary 1 converts this empirical-disagreement bound into a bound in terms of the population disagreement probability p_{sj}.

Corollary 1 (Population-disagreement rate). *For $j \ge 1$, define $r_{sj} := L\left(\sqrt{\frac{3p_{sj}\ell_\alpha}{n_s-1}} + \frac{7\ell_\alpha}{n_s-1}\right)$ and set $r_{s0} := 0$. Under the conditions of Theorem 1, with probability at least $1-\alpha$, simultaneously for all $s \in \mathcal{S}$ and $j \in \{0, \dots, M\}$, $\left|\widehat{D}_{sj} - D_{sj}\right| \le c_{sj} \le r_{sj}$. Consequently, the confidence radius used by our procedure has the finite-sample scaling*

$$c_{sj} = O\left(\sqrt{\frac{p_{sj}\log(M|\mathcal{S}|/\alpha)}{n_s}} + \frac{\log(M|\mathcal{S}|/\alpha)}{n_s}\right). \tag{6}$$

The leading confidence width is $\sqrt{p_{sj}/n_s}$, so certification becomes tighter when the candidate differs from the incumbent on a smaller fraction of the population. The same radius enters the overall objective and each subgroup constraint, giving candidate-specific and subgroup-specific uncertainty adjustments.

5.2 A Protected-Subgroup Minimax Lower Bound

We next establish that the $\sqrt{p/n}$ dependence is a local statistical limit near a binding protected-subgroup constraint. The key difficulty is that two data-generating processes can induce nearly indistinguishable audit distributions while placing the candidate policy on opposite sides of the subgroup safety boundary. Consequently, any procedure that is required to be uniformly safe must remain conservative on at least one of the two instances.

We consider a binary-action subclass with one protected subgroup and zero subgroup tolerance, where $P_1 = P_0(\cdot \mid G = 1)$. We construct two statistically close instances that share the same positive overall gain but have opposite protected-subgroup value differences at the $\sqrt{p/n}$ scale.

Theorem 2 (Protected-subgroup lower bound near a binding constraint). *Fix $\alpha \le 1/8$, $p \in (0, 1]$, and n such that $np \ge 1$. For any $\Gamma \in (0, p/16]$, there exist two data-generating processes $\mathbb{P}_-$ and $\mathbb{P}_+$ satisfying Assumptions 1–2, with one protected subgroup satisfying $P_1 = P_0(\cdot \mid G = 1)$, such that $p_{01} = p_{11} = p$, $D_{0,-}(d) = D_{0,+}(d) = \Gamma$, and $D_{1,-}(d) = -\delta_n$ while $D_{1,+}(d) = \delta_n$, where $\delta_n = (1/32)\sqrt{p/n}$. Consider any possibly randomized deployment rule that observes n audit observations from each of P_0 and P_1 and outputs a candidate weight $\widehat{\lambda} \in [0, 1]$. If the rule is $(1-\alpha)$-safe under both $\mathbb{P}_-$ and $\mathbb{P}_+$ for the zero-tolerance protected-subgroup constraint, then under $\mathbb{P}_+$, $\Pr_+(\widehat{\lambda} = 0) \ge 5/8$ and $\mathbb{E}_+[W_+^* - D_{0,+}(\widehat{\lambda})] \ge (5/8)\Gamma$.*

Theorem 2 shows that when the protected-subgroup value difference is of order $\sqrt{p/n}$, its sign cannot be reliably determined from the audit data. Consequently, any uniformly safe rule must revert to the incumbent with nonvanishing probability, even when the candidate is safe and improves the overall value. Since the confidence radii of DiSPU scale as $\sqrt{p/n}$ by Corollary 1, this lower bound shows that its disagreement dependence cannot be improved beyond logarithmic factors. Corollary 4 (see Appendix A.4) further gives the corresponding $\Theta(\sqrt{p/n})$ local minimax regret at this boundary scale.

5.3 Protected Safety and Achievable Deployment Value

We next propagate the confidence bounds through the protected-subgroup-constrained deployment problem. Define the confidence-adjusted population problem $\max_{\theta\in\Delta_M} \sum_{j=0}^{M} \theta_j(D_{0j} - 2c_{0j})$ subject to $\sum_{j=0}^{M} \theta_j(D_{sj} - 2c_{sj}) \geq -\Delta_s$ for all $s \in \mathcal{H}$. The next result shows that the empirical deployment is safe under the true policy values and achieves at least the value of this benchmark.

Theorem 3 (Protected safety and achievable deployment value)**.** *On the simultaneous confidence event of Theorem 1, the deployment weights $\widehat{\theta}$ returned by the confidence-adjusted linear program satisfy $D_s(\widehat{\theta}) \geq -\Delta_s$ for all $s \in \mathcal{H}$ and $D_0(\widehat{\theta}) \geq 0$. Moreover,*

$$D_0(\widehat{\theta}) \geq \max_{\theta\in\Delta_M} \left\{ \sum_{j=0}^{M} \theta_j(D_{0j} - 2c_{0j}) : \sum_{j=0}^{M} \theta_j(D_{sj} - 2c_{sj}) \geq -\Delta_s,\ \forall s \in \mathcal{H} \right\}.$$

Theorem 3 guarantees safety under the true policy values and lower bounds the achieved overall gain by the confidence-adjusted population benchmark. Corollary 2 compares this benchmark with the oracle value W^*.

Corollary 2 (Oracle regret away from the safety boundary)**.** *Let θ^* be an optimal solution to the oracle problem in Section 3, and define $C_s(\theta) := \sum_{j=0}^{M} \theta_j c_{sj}$. If $D_s(\theta^*) + \Delta_s \geq 2C_s(\theta^*)$ for all $s \in \mathcal{H}$, then, on the simultaneous confidence event of Theorem 1, $W^* - D_0(\widehat{\theta}) \leq 2C_0(\theta^*)$. More generally, suppose that $\Delta_s > 0$ for every $s \in \mathcal{H}$, and define $\rho^* := \min\{1,\ \min_{s\in\mathcal{H}} \frac{\Delta_s}{\Delta_s + 2C_s(\theta^*)}\}$. Then, on the same event, $W^* - D_0(\widehat{\theta}) \leq (1-\rho^*)W^* + 2\rho^* C_0(\theta^*)$.*

Substituting $c_{sj} \leq r_{sj}$ from Corollary 1 expresses the oracle-regret bound in terms of the disagreement probabilities. This makes the dependence on policy disagreement explicit and shows how smaller disagreement tightens the achievable deployment guarantee. The resulting bound is stated in Corollary 3.

Corollary 3 (Disagreement-dependent regret)**.** *Let θ^* be an optimal solution to the oracle problem in Section 3, and let r_{sj} be defined as in Corollary 1. If $D_s(\theta^*) + \Delta_s \geq 2\sum_{j=0}^{M} \theta_j^* r_{sj}$ for all $s \in \mathcal{H}$, then, with probability at least $1-\alpha$,*

$$W^* - D_0(\widehat{\theta}) \leq 2\sum_{j=0}^{M} \theta_j^* r_{0j} = 2L \sum_{j=1}^{M} \theta_j^* \left(\sqrt{\frac{3p_{0j}\ell_\alpha}{n_0 - 1}} + \frac{7\ell_\alpha}{n_0 - 1} \right).$$

More generally, suppose that $\Delta_s > 0$, and define $\bar{\rho} := \min\{1,\ \min_{s\in\mathcal{H}} \frac{\Delta_s}{\Delta_s + 2\sum_{j=0}^{M} \theta_j^* r_{sj}}\}$. Then, with probability at least $1-\alpha$,

$$W^* - D_0(\widehat{\theta}) \leq (1-\bar{\rho})W^* + 2\bar{\rho} \sum_{j=0}^{M} \theta_j^* r_{0j}.$$

When the oracle solution is sufficiently far from the subgroup safety boundaries, the regret is governed by $\sqrt{p_{0j}/n_0}$, so candidates that change the incumbent's action on only a small fraction of the population incur smaller regret. When the subgroup constraints are close to binding, the same radii determine how far the oracle deployment must be shrunk toward the incumbent.

6 Experiments

Evaluation benchmarks. We evaluate our method on synthetic data and a semi-synthetic benchmark based on real-world ACS 2018 data. We report the overall gain relative to the incumbent, $D_0(\hat{\theta}) = \sum_{j=0}^{M} \hat{\theta}_j[V_0(d_j) - V_0(b)]$, and the worst-group safety margin, $M(\hat{\theta}, \Delta) = \min_{s\in\mathcal{H}}\{\sum_{j=0}^{M} \hat{\theta}_j[V_s(d_j) - V_s(b)] + \Delta\}$, where we set $\Delta_s = \Delta$ for all protected subgroups. Both metrics are computed using the true policy values under the data-generating mechanism. A nonnegative margin indicates that all subgroup constraints are satisfied. **Baselines.** We select RobinHood (Metevier et al., 2019) and HCPI (Thomas et al., 2015) as baselines and adapt both to our

finite candidate library. RobinHood selects candidate policies using inflated Student-t confidence bounds and then independently audits them. HCPI likewise freezes candidate policies before auditing but uses uninflated Student-t bounds during selection. For each baseline, we evaluate both its original estimators and our paired doubly robust estimator. RobinHood originally uses ordinary importance sampling (IS), while HCPI uses weighted importance sampling (WIS) for selection utility and ordinary IS for safety bounds. All variants share the same candidate library, logged data, and subgroup constraints. **Candidate Policy Library.** We construct $M = 16$ candidate policies using T-learners with different machine learning models and disagreement budgets. The library is frozen before auditing, with construction details provided in Appendix C.2. Throughout, we set the failure probability to $\alpha = 0.05$.

6.1 Results on Simulated Datasets

We construct a synthetic environment with heterogeneous treatment effects and six overlapping protected subgroups, including pairwise intersections. The outcome mechanism creates localized opportunities for policy improvement while allowing updates that benefit the overall population to harm particular subgroups. The full data-generating process is provided in Appendix C.1. We evaluate overall gain and worst-group safety margin across audit sample sizes $n_{\text{aud}} \in \{10{,}000, 25{,}000, 50{,}000, 100{,}000, 200{,}000\}$ and subgroup tolerances $\Delta \in \{0, 0.01, \ldots, 0.06\}$, keeping the training sample size fixed at $50{,}000$. This design examines how the methods respond to increasing evidence for certification and progressively relaxed subgroup constraints.

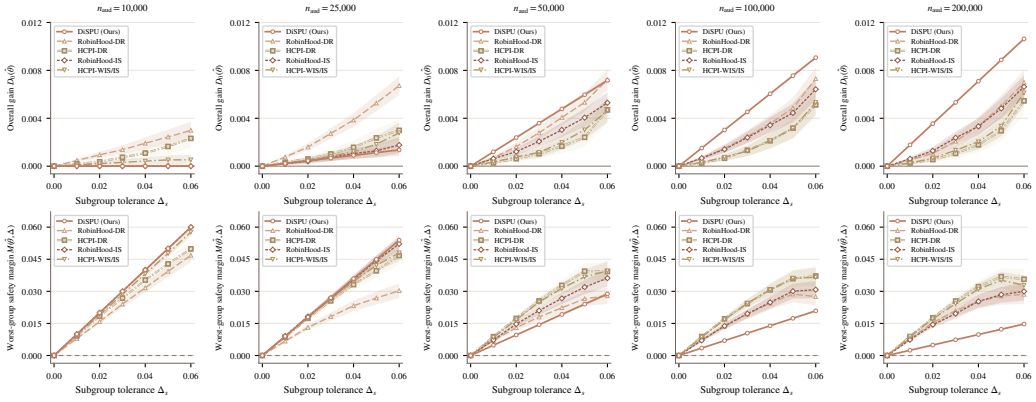

Figure 1: Overall gain (top) and worst-group safety margin (bottom) on synthetic data across subgroup tolerances and audit sample sizes. Columns show increasing audit sample sizes from $10{,}000$ to $200{,}000$. Dashed horizontal lines mark the safety boundary.

Figure 1 shows that DiSPU's performance depends on the amount of audit data available. With limited audit data, DiSPU yields negligible gains and remains more conservative than the baselines. As the audit sample grows, its gains increase substantially, eventually exceeding those of both baselines across positive subgroup tolerances. Relaxing the subgroup constraints generally allows larger gains, although DiSPU benefits little from this relaxation when audit data are scarce. The reported worst-group safety margins remain nonnegative for all methods. With more audit data, DiSPU achieves larger gains while operating closer to the safety boundary, indicating less conservative deployment without negative margins in the results.

6.2 Ablation Studies

To isolate the contributions of paired difference estimation and confidence-based certification, we compare DiSPU with two ablations: Separate-DR-EB-LP bounds candidate and incumbent values separately instead of directly bounding their paired difference, removing the benefit of paired uncertainty cancellation. Empirical-DR-LP retains the paired DR difference estimates but removes their confidence radii from both the objective and subgroup constraints, thereby testing the role of uncertainty adjustment in safe deployment. We use the same synthetic environment, fixing both the

training and audit sample sizes at $50{,}000$ and varying $\Delta \in \{0, 0.01, \ldots, 0.06\}$. All variants are evaluated in the same environment as in Section 6.1.

Figure 2 highlights the complementary roles of paired estimation and confidence adjustment. Separate-DR-EB-LP yields negligible gains despite relaxed subgroup tolerances, with large safety margins indicating overly conservative deployment. DiSPU achieves increasing gains while retaining positive reported safety margins under positive tolerances, illustrating the benefit of directly bounding paired policy differences. Empirical-DR-LP attains the largest gains but exhibits negative worst-group safety margins across several tolerances, indicating subgroup constraint violations. Together, these results show that paired estimation reduces conservatism, while confidence adjustment helps prevent the additional gains from coming at the expense of subgroup safety.

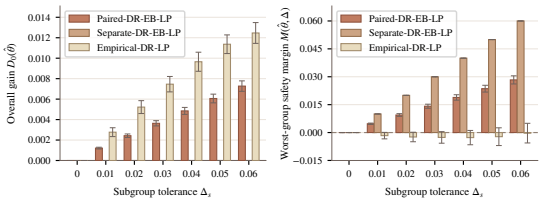

Figure 2: Ablation results: overall gain (left) and worst-group safety margin (right) across subgroup tolerances. Paired-DR-EB-LP is ours. The dashed line marks the safety boundary.

6.3 Results on Real-World Datasets

We evaluate DiSPU on ACSIncome, a real-world dataset derived from the 2018 American Community Survey. We use its demographic and socioeconomic covariates to construct a policy-updating benchmark with simulated intervention outcomes. Detailed data processing and outcome generation are provided in Appendix C.1. We use $40{,}000$ samples for model fitting, $10{,}000$ for candidate construction and selection, and $50{,}000$ for independent auditing, varying the subgroup tolerance over $\Delta \in \{0, 0.02, 0.04, 0.06, 0.08\}$.

Table 1: Performance ($\times 10^{-2}$) on ACSIncome across protected-subgroup tolerances Δ.

	$\Delta = 0$		$\Delta = 0.02$		$\Delta = 0.04$		$\Delta = 0.06$		$\Delta = 0.08$	
Method	Gain ↑	Margin	Gain ↑	Margin	Gain ↑	Margin	Gain ↑	Margin	Gain ↑	Margin
DiSPU (Ours)	0.00 ± 0.00	0.00 ± 0.00	**0.33 ± 0.03**	0.84 ± 0.10	**0.66 ± 0.06**	1.68 ± 0.19	**0.99 ± 0.09**	2.52 ± 0.29	**1.32 ± 0.12**	3.36 ± 0.39
RobinHood-DR	0.00 ± 0.00	0.00 ± 0.00	0.13 ± 0.24	1.67 ± 0.62	0.29 ± 0.49	3.15 ± 1.41	0.58 ± 0.79	4.25 ± 2.35	1.03 ± 1.03	4.73 ± 3.24
HCPI-DR	0.00 ± 0.00	0.00 ± 0.00	0.01 ± 0.08	1.96 ± 0.25	0.02 ± 0.17	3.96 ± 0.38	0.04 ± 0.31	5.89 ± 0.80	0.33 ± 0.83	7.13 ± 2.19

Table 1 shows that DiSPU consistently achieves the largest overall gain for every positive subgroup tolerance while maintaining nonnegative worst-group safety margins. As the constraints are relaxed, DiSPU translates the additional safety budget into steadily improved performance and operates closer to the constraint boundary than the baselines, indicating less conservative deployment. Its gains also exhibit substantially lower variability across repetitions. Under the strict zero-tolerance requirement, all methods revert to the incumbent. Overall, disagreement-adaptive paired certification enables larger and more stable improvements while preserving protected-subgroup safety.

7 Conclusion

We study safe policy updating under protected-subgroup constraints using direct value-difference estimation between each candidate and the incumbent. DiSPU exploits the zero contribution from policy-agreement regions, yielding confidence radii with leading $\sqrt{p/n}$ dependence and a corresponding constrained deployment rule. We prove finite-sample subgroup safety and disagreement-dependent regret guarantees. A protected-subgroup lower bound and local minimax result establish a $\Theta(\sqrt{p/n})$ regret rate near a binding subgroup constraint, showing that the $\sqrt{p/n}$ dependence cannot be uniformly improved in this local setting. Experiments on synthetic and semi-synthetic data show higher gains from paired certification while keeping the reported worst-group margins nonnegative. Our analysis focuses on a fixed candidate library and a single policy update. Extending the framework to sequential policy updating is an important direction for future work, as repeated deployment and re-estimation introduce additional dependence across update stages that is not captured by the current analysis.

AI USE STATEMENT

Generative AI tools were used solely for language polishing. The authors retain full responsibility for the correctness of the theoretical derivations, experimental design, and analysis of the results.

REFERENCES

Susan Athey and Stefan Wager. Policy learning with observational data. *Econometrica*, 89(1): 133–161, 2021.

Ching-An Cheng, Tengyang Xie, Nan Jiang, and Alekh Agarwal. Adversarially trained actor critic for offline reinforcement learning. In *International Conference on Machine Learning*, pp. 3852–3878. PMLR, 2022.

Brian Cho, Ana-Roxana Pop, Kyra Gan, Sam Corbett-Davies, Israel Nir, Ariel Evnine, and Nathan Kallus. Cspi-mt: Calibrated safe policy improvement with multiple testing for threshold policies. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 1*, pp. 165–176, 2025.

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

Zeyang Jia, Eli Ben-Michael, and Kosuke Imai. Bayesian safe policy learning with chance constrained optimization: Application to military security assessment during the vietnam war. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 189(3):1448–1472, 2026.

Thorsten Joachims, Adith Swaminathan, and Maarten De Rijke. Deep learning with logged bandit feedback. In *International Conference on Learning Representations*, 2018.

Sham M. Kakade and John Langford. Approximately optimal approximate reinforcement learning. In *Proceedings of the 19th International Conference on Machine Learning*, pp. 267–274, 2002.

Nathan Kallus. Balanced policy evaluation and learning. *Advances in neural information processing systems*, 31, 2018.

Nikos Karampatziakis, Paul Mineiro, and Aaditya Ramdas. Off-policy confidence sequences. In *International Conference on Machine Learning*, pp. 5301–5310. PMLR, 2021.

Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In *International conference on machine learning*, pp. 2564–2572. PMLR, 2018.

Toru Kitagawa and Aleksey Tetenov. Who should be treated? empirical welfare maximization methods for treatment choice. *Econometrica*, 86(2):591–616, 2018.

Romain Laroche, Paul Trichelair, and Remi Tachet Des Combes. Safe policy improvement with baseline bootstrapping. In *International conference on machine learning*, pp. 3652–3661. PMLR, 2019.

Zhaoqi Li, Lillian Ratliff, Kevin G Jamieson, Lalit Jain, et al. Instance-optimal pac algorithms for contextual bandits. *Advances in Neural Information Processing Systems*, 35:37590–37603, 2022.

Andreas Maurer and Massimiliano Pontil. Empirical bernstein bounds and sample variance penalization. *arXiv preprint arXiv:0907.3740*, 2009.

Blossom Metevier, Stephen Giguere, Sarah Brockman, Ari Kobren, Yuriy Brun, Emma Brunskill, and Philip S Thomas. Offline contextual bandits with high probability fairness guarantees. *Advances in neural information processing systems*, 32, 2019.

Razieh Nabi, Daniel Malinsky, and Ilya Shpitser. Learning optimal fair policies. In *International conference on machine learning*, pp. 4674–4682. PMLR, 2019.

Adhyyan Narang, Andrew Wagenmaker, Lillian J Ratliff, and Kevin Jamieson. Sample complexity reduction via policy difference estimation in tabular reinforcement learning. *Advances in Neural Information Processing Systems*, 37:22772–22826, 2024.

Marek Petrik, Mohammad Ghavamzadeh, and Yinlam Chow. Safe policy improvement by minimizing robust baseline regret. In *Advances in Neural Information Processing Systems*, pp. 2306–2314. Curran Associates, Inc., 2016.

Ayush Sawarni, Jikai Jin, Justin Whitehouse, and Vasilis Syrgkanis. Policy learning with abstention. In *Proceedings of The 29th International Conference on Artificial Intelligence and Statistics*, volume 300 of *Proceedings of Machine Learning Research*, pp. 4465–4473. PMLR, 2026.

Adith Swaminathan and Thorsten Joachims. Counterfactual risk minimization: Learning from logged bandit feedback. In *International conference on machine learning*, pp. 814–823. PMLR, 2015.

Philip Thomas, Georgios Theocharous, and Mohammad Ghavamzadeh. High confidence policy improvement. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 2380–2388. PMLR, 2015.

Philip S Thomas, Joelle Pineau, Romain Laroche, et al. Multi-objective spibb: Seldonian offline policy improvement with safety constraints in finite mdps. *Advances in Neural Information Processing Systems*, 34:2004–2017, 2021.

Berk Ustun, Yang Liu, and David Parkes. Fairness without harm: Decoupled classifiers with preference guarantees. In *International Conference on Machine Learning*, pp. 6373–6382. PMLR, 2019.

A Detailed Proofs

A.1 Proof of Theorem 1

Proof. Let $\mathcal{F}_{\mathrm{tr}}$ denote the σ-field generated by the training data. Conditional on $\mathcal{F}_{\mathrm{tr}}$, the incumbent policy b, the candidate policies $d_1, \dots, d_M$, and the clipped regression functions $\widehat{m}_s$ are fixed. Moreover, because the audit data are independent of the training data, for each fixed $s \in \mathcal{S}$ the observations in $\mathcal{D}_s^{\mathrm{aud}}$ remain i.i.d. from P_s. We first identify the conditional mean of the paired score and then establish a uniform almost-sure bound.

Fix $s \in \mathcal{S}$ and $j \in \{1, \dots, M\}$, and write

$$m_s(x,a) := \mathbb{E}_s[Y \mid X = x, A = a], \qquad h_{sj}(x,a) := \mathbf{1}\{d_j(x) = a\} - \mathbf{1}\{b(x) = a\}.$$

Since the propensity score $e_s(a \mid x)$ is the true logging propensity, for every $a \in \mathcal{A}$,

$$\begin{aligned}
&\mathbb{E}_s\left[\frac{\mathbf{1}\{A = a\}}{e_s(a \mid X)}\{Y - \widehat{m}_s(X,a)\} \,\middle|\, X, \mathcal{F}_{\mathrm{tr}}\right] \\
&\quad = \frac{\mathbb{P}_s(A = a \mid X)}{e_s(a \mid X)}\{m_s(X,a) - \widehat{m}_s(X,a)\} \\
&\quad = m_s(X,a) - \widehat{m}_s(X,a). \qquad (7)
\end{aligned}$$

Substituting equation 7 into the paired score in equation 1 gives

$$\begin{aligned}
\mathbb{E}_s[Z_{sj}(O) \mid X, \mathcal{F}_{\mathrm{tr}}] &= \sum_{a \in \mathcal{A}} h_{sj}(X,a) m_s(X,a) \\
&= m_s\{X, d_j(X)\} - m_s\{X, b(X)\}. \qquad (8)
\end{aligned}$$

Under Assumption 1, $m_s(x,a) = \mathbb{E}_s[Y_s(a) \mid X_s = x]$ almost surely. Hence, integrating equation 8 with respect to X yields

$$\mathbb{E}_s[Z_{sj}(O) \mid \mathcal{F}_{\mathrm{tr}}] = V_s(d_j) - V_s(b) = D_{sj}. \qquad (9)$$

In particular, the equality in equation 9 does not require $\widehat{m}_s$ to equal the true outcome regression.

We next bound the score. The definition in equation 1 can be rewritten as

$$Z_{sj}(O) = \widehat{m}_s\{X, d_j(X)\} - \widehat{m}_s\{X, b(X)\} + \frac{h_{sj}(X, A)}{e_s(A \mid X)}\{Y - \widehat{m}_s(X, A)\}. \tag{10}$$

If $d_j(X) = b(X)$, then both terms in equation 10 vanish. Otherwise, clipping gives

$$|\widehat{m}_s\{X, d_j(X)\} - \widehat{m}_s\{X, b(X)\}| \le 2B,$$

while Assumptions 1 and 2 imply

$$|h_{sj}(X, A)| \le 1, \qquad |Y - \widehat{m}_s(X, A)| \le 2B, \qquad e_s(A \mid X) \ge \kappa.$$

Therefore

$$|Z_{sj}(O)| \le 2B + \frac{2B}{\kappa} = L \qquad \text{almost surely.} \tag{11}$$

It remains to apply an empirical Bernstein inequality with the constants used in the main text. Let $U_1, \ldots, U_n$ be i.i.d. random variables in $[-L, L]$, $n \ge 2$, with sample mean and sample variance

$$\overline{U} := \frac{1}{n}\sum_{i=1}^{n} U_i, \qquad S_U^2 := \frac{1}{n-1}\sum_{i=1}^{n}(U_i - \overline{U})^2.$$

For $L > 0$, define $W_i = (U_i + L)/(2L) \in [0, 1]$. Then

$$S_W^2 = \frac{1}{n-1}\sum_{i=1}^{n}(W_i - \overline{W})^2 = \frac{S_U^2}{4L^2}.$$

Applying the one-sided empirical Bernstein inequality of Maurer & Pontil (2009, Theorem 4) to $W_1, \ldots, W_n$ with failure probability $\delta/2$, and again to $1 - W_1, \ldots, 1 - W_n$, gives by a union bound

$$|\overline{U} - \mathbb{E}[U_1]| \le \sqrt{\frac{2S_U^2 \log(4/\delta)}{n}} + \frac{14L \log(4/\delta)}{3(n-1)} \tag{12}$$

with probability at least $1 - \delta$. The case $L = 0$ is immediate, since then all paired scores are identically zero.

We now apply equation 12 conditionally on $\mathcal{F}_{\mathrm{tr}}$. For a fixed (s, j) with $j \ge 1$, take $U_i = Z_{sj}(O_{is})$ and $n = n_s$. By equation 9 and equation 11, the conditional mean is D_{sj}, the sample mean is $\widehat{D}_{sj}$, and the sample variance is $\widehat{\sigma}_{sj}^2$. Let

$$N := M|\mathcal{S}|, \qquad \delta := \frac{\alpha}{2N}.$$

Then

$$\log\frac{4}{\delta} = \log\frac{8N}{\alpha} = \ell_\alpha,$$

and therefore

$$\mathbb{P}\Big(|\widehat{D}_{sj} - D_{sj}| > c_{sj} \,\Big|\, \mathcal{F}_{\mathrm{tr}}\Big) \le \frac{\alpha}{2N}. \tag{13}$$

A union bound over the N population–candidate pairs yields

$$\mathbb{P}\Big(\exists\, s \in \mathcal{S},\ j \in \{1, \ldots, M\} : |\widehat{D}_{sj} - D_{sj}| > c_{sj} \,\Big|\, \mathcal{F}_{\mathrm{tr}}\Big) \le \sum_{s\in\mathcal{S}}\sum_{j=1}^{M} \mathbb{P}\Big(|\widehat{D}_{sj} - D_{sj}| > c_{sj} \,\Big|\, \mathcal{F}_{\mathrm{tr}}\Big) \le \frac{\alpha}{2}. \tag{14}$$

No independence across populations or across candidate policies is needed in equation 14. For $j = 0$, since $d_0 = b$, we have $D_{s0} = \widehat{D}_{s0} = c_{s0} = 0$ deterministically. Consequently, with

$$\mathcal{E}_D := \bigcap_{s\in\mathcal{S}}\bigcap_{j=0}^{M}\Big\{|\widehat{D}_{sj} - D_{sj}| \le c_{sj}\Big\},$$

we have

$$\mathbb{P}(\mathcal{E}_D \mid \mathcal{F}_{\mathrm{tr}}) \ge 1 - \frac{\alpha}{2} \qquad \text{almost surely.}$$

Finally, taking expectations with respect to the training data,

$$\mathbb{P}(\mathcal{E}_D) = \mathbb{E}[\mathbb{P}(\mathcal{E}_D \mid \mathcal{F}_{\mathrm{tr}})] \ge 1 - \frac{\alpha}{2},$$

which proves the theorem. □

A.2 Proof of Proposition 1

Proof. Fix $s \in \mathcal{S}$ and $j \in \{1, \dots, M\}$, and consider an arbitrary realization of the corresponding audit sample. The argument is entirely deterministic. For brevity, write

$$Z_i := Z_{sj}(O_{is}), \qquad I_i := \mathbf{1}\{d_j(X_{is}) \neq b(X_{is})\}, \qquad i = 1, \dots, n_s.$$

Then

$$\widehat{D}_{sj} = \frac{1}{n_s} \sum_{i=1}^{n_s} Z_i, \qquad \widehat{p}_{sj} = \frac{1}{n_s} \sum_{i=1}^{n_s} I_i.$$

The paired construction implies a pointwise relation between the score and the disagreement indicator. If $I_i = 0$, the candidate and incumbent choose the same action and every policy-difference coefficient in equation 1 is zero; hence $Z_i = 0$. If $I_i = 1$, the same boundedness argument as in equation 11 gives $|Z_i| \leq L$. Thus, for every i,

$$Z_i^2 \leq L^2 I_i. \tag{15}$$

Summing equation 15 over the audit sample,

$$\sum_{i=1}^{n_s} Z_i^2 \leq L^2 \sum_{i=1}^{n_s} I_i = n_s L^2 \widehat{p}_{sj}. \tag{16}$$

Next, expand the sample variance:

$$\begin{aligned}
(n_s - 1)\widehat{\sigma}_{sj}^2 &= \sum_{i=1}^{n_s} (Z_i - \widehat{D}_{sj})^2 \\
&= \sum_{i=1}^{n_s} Z_i^2 - 2\widehat{D}_{sj} \sum_{i=1}^{n_s} Z_i + n_s \widehat{D}_{sj}^2 \\
&= \sum_{i=1}^{n_s} Z_i^2 - n_s \widehat{D}_{sj}^2 \\
&\leq \sum_{i=1}^{n_s} Z_i^2 \leq n_s L^2 \widehat{p}_{sj}.
\end{aligned} \tag{17}$$

Since $n_s \geq 2$, dividing equation 17 by $n_s - 1$ yields

$$\widehat{\sigma}_{sj}^2 \leq \frac{n_s}{n_s - 1} L^2 \widehat{p}_{sj}. \tag{18}$$

Finally, substitute equation 18 into equation 3. Because the square-root function is monotone,

$$\begin{aligned}
c_{sj} &= \sqrt{\frac{2\widehat{\sigma}_{sj}^2 \ell_\alpha}{n_s}} + \frac{14 L \ell_\alpha}{3(n_s - 1)} \\
&\leq \sqrt{\frac{2\ell_\alpha}{n_s} \frac{n_s}{n_s - 1} L^2 \widehat{p}_{sj}} + \frac{14 L \ell_\alpha}{3(n_s - 1)} \\
&= L \sqrt{\frac{2\widehat{p}_{sj} \ell_\alpha}{n_s - 1}} + \frac{14 L \ell_\alpha}{3(n_s - 1)}.
\end{aligned} \tag{19}$$

This proves both claims. □

A.3 Proof of Corollary 1

Proof. As before, condition on $\mathcal{F}_{\mathrm{tr}}$, so that the candidate policies and incumbent are fixed. For every $s \in \mathcal{S}$ and $j \in \{1, \dots, M\}$, define

$$I_{isj} := \mathbf{1}\{d_j(X_{is}) \neq b(X_{is})\}, \qquad i = 1, \dots, n_s.$$

Conditional on $\mathcal{F}_{\mathrm{tr}}$, these variables are i.i.d. Bernoulli with mean p_{sj}. Hence

$$\widehat{p}_{sj} = \frac{1}{n_s}\sum_{i=1}^{n_s} I_{isj}, \qquad \mathrm{Var}(I_{isj} \mid \mathcal{F}_{\mathrm{tr}}) = p_{sj}(1-p_{sj}) \le p_{sj}.$$

We first control the empirical disagreement probability. Since $|I_{isj} - p_{sj}| \le 1$, Bernstein's inequality implies that, for every $t > 0$,

$$\mathbb{P}\left(\sum_{i=1}^{n_s}(I_{isj} - p_{sj}) > t \,\middle|\, \mathcal{F}_{\mathrm{tr}}\right) \le \exp\left\{-\frac{t^2}{2(n_s p_{sj} + t/3)}\right\}. \tag{20}$$

For $u > 0$, set

$$t_u := \sqrt{2n_s p_{sj} u} + \frac{2u}{3}.$$

A direct calculation gives

$$t_u^2 - 2u\left(n_s p_{sj} + \frac{t_u}{3}\right) = \frac{2u}{3}\sqrt{2n_s p_{sj} u} \ge 0.$$

Substituting $t = t_u$ into equation 20 therefore yields

$$\mathbb{P}\left(\widehat{p}_{sj} > p_{sj} + \sqrt{\frac{2p_{sj}u}{n_s}} + \frac{2u}{3n_s} \,\middle|\, \mathcal{F}_{\mathrm{tr}}\right) \le e^{-u}. \tag{21}$$

Using $2ab \le a^2 + b^2$ with $a = \sqrt{p_{sj}/2}$ and $b = \sqrt{u/n_s}$,

$$\sqrt{\frac{2p_{sj}u}{n_s}} \le \frac{p_{sj}}{2} + \frac{u}{n_s}.$$

Hence, with conditional probability at least $1 - e^{-u}$,

$$\widehat{p}_{sj} \le \frac{3}{2}p_{sj} + \frac{5u}{3n_s}. \tag{22}$$

Let $N = M|\mathcal{S}|$ and choose

$$u := \log\frac{2N}{\alpha} \le \log\frac{8N}{\alpha} = \ell_\alpha.$$

Define

$$\mathcal{E}_p := \bigcap_{s\in\mathcal{S}}\bigcap_{j=1}^{M}\left\{\widehat{p}_{sj} \le \frac{3}{2}p_{sj} + \frac{5\ell_\alpha}{3n_s}\right\}.$$

By equation 22 and a union bound,

$$\mathbb{P}(\mathcal{E}_p^c \mid \mathcal{F}_{\mathrm{tr}}) \le Ne^{-u} = \frac{\alpha}{2}. \tag{23}$$

On $\mathcal{E}_p$, Proposition 1 gives, simultaneously over all s and $j \ge 1$,

$$\begin{aligned}
c_{sj} &\le L\sqrt{\frac{2\widehat{p}_{sj}\ell_\alpha}{n_s - 1}} + \frac{14L\ell_\alpha}{3(n_s-1)} \\
&\le L\sqrt{\frac{2\ell_\alpha}{n_s-1}\left(\frac{3}{2}p_{sj} + \frac{5\ell_\alpha}{3n_s}\right)} + \frac{14L\ell_\alpha}{3(n_s-1)} \\
&\le L\sqrt{\frac{3p_{sj}\ell_\alpha}{n_s-1}} + L\sqrt{\frac{10}{3}}\,\frac{\ell_\alpha}{\sqrt{n_s(n_s-1)}} + \frac{14L\ell_\alpha}{3(n_s-1)} \\
&\le L\left(\sqrt{\frac{3p_{sj}\ell_\alpha}{n_s-1}} + \left(\sqrt{\frac{10}{3}} + \frac{14}{3}\right)\frac{\ell_\alpha}{n_s-1}\right) \\
&\le L\left(\sqrt{\frac{3p_{sj}\ell_\alpha}{n_s-1}} + \frac{7\ell_\alpha}{n_s-1}\right) = r_{sj}.
\end{aligned} \tag{24}$$

Here we used $\sqrt{x+y} \le \sqrt{x} + \sqrt{y}$, $\sqrt{n_s(n_s-1)} \ge n_s - 1$, and $\sqrt{10/3} + 14/3 < 7$.

Let $\mathcal{E}_D$ denote the simultaneous confidence event in the proof of Theorem 1. We have

$$\mathbb{P}(\mathcal{E}_D^c \mid \mathcal{F}_{\mathrm{tr}}) \le \frac{\alpha}{2}$$

almost surely, while equation 23 gives the same bound for $\mathcal{E}_p^c$. Therefore,

$$\mathbb{P}(\mathcal{E}_D \cap \mathcal{E}_p \mid \mathcal{F}_{\mathrm{tr}}) \ge 1 - \alpha.$$

On this intersection,

$$|\widehat{D}_{sj} - D_{sj}| \le c_{sj} \le r_{sj}$$

simultaneously for all $s \in \mathcal{S}$ and $j \ge 1$. For $j = 0$, $D_{s0} = \widehat{D}_{s0} = c_{s0} = r_{s0} = 0$ deterministically, so the same statement extends to the baseline. Removing the conditioning by the tower property yields the claimed probability $1 - \alpha$.

Finally, because $n_s - 1 \ge n_s/2$ for $n_s \ge 2$,

$$r_{sj} \le L\sqrt{\frac{6p_{sj}\ell_\alpha}{n_s}} + \frac{14L\ell_\alpha}{n_s},$$

which gives

$$c_{sj} = O\left(L\sqrt{\frac{p_{sj}\ell_\alpha}{n_s}} + L\frac{\ell_\alpha}{n_s}\right)$$

on the same event. Since $\ell_\alpha = \log(8M|\mathcal{S}|/\alpha)$, this is exactly the rate stated in the corollary. □

A.4 Proof of Theorem 2

Proof. Fix $p \in (0, 1]$, n with $np \ge 1$, and $\Gamma \in (0, p/16]$. We construct two protected-subgroup data-generating processes indexed by $\nu \in \{-1, +1\}$. In both processes, the overall population contains a covariate vector $X = (G, U)$, where $G \sim \mathrm{Bernoulli}(1/2)$ and $U \sim \mathrm{Bernoulli}(p)$ are independent. The protected subgroup is $G = 1$, so its evaluation law is exactly

$$P_{1,\nu} = P_{0,\nu}(\cdot \mid G = 1).$$

The incumbent and candidate policies are

$$b(g, u) = 0, \qquad d(g, u) = u,$$

so disagreement occurs exactly when $U = 1$. Consequently,

$$p_{01} = P_{0,\nu}\{d(X) \ne b(X)\} = p, \qquad p_{11} = P_{1,\nu}\{d(X) \ne b(X)\} = p.$$

Set

$$h := \frac{1}{32\sqrt{np}}.$$

Since $np \ge 1$, we have $h \le 1/32$. Conditional on (G, U), generate the potential outcomes by

$$Y(0) \sim \mathrm{Bernoulli}(1/2)$$

and

$$Y(1) \sim \mathrm{Bernoulli}\left(\frac{1}{2} + U\left[\nu h\,\mathbf{1}\{G = 1\} + \left(\frac{2\Gamma}{p} - \nu h\right)\mathbf{1}\{G = 0\}\right]\right),$$

with the two potential outcomes conditionally independent. Draw $A \sim \mathrm{Bernoulli}(1/2)$ independently of the potential outcomes given (G, U) and set $Y = Y(A)$. Because $\Gamma \le p/16$ and $h \le 1/32$, all Bernoulli parameters lie in $[15/32, 21/32]$. Thus both processes satisfy consistency, unconfoundedness, known logging propensities equal to $1/2$, overlap, and bounded outcomes with $B = 1$.

Under the protected-subgroup law $P_{1,\nu}$, the treatment effect on the disagreement region $U = 1$ equals νh. Therefore

$$D_{1,\nu}(d) = p\nu h = \frac{\nu}{32}\sqrt{\frac{p}{n}}. \tag{25}$$

Under the overall law $P_{0,\nu}$, averaging over G gives

$$D_{0,\nu}(d) = p\left[\frac{1}{2}\nu h + \frac{1}{2}\left(\frac{2\Gamma}{p} - \nu h\right)\right] = \Gamma. \tag{26}$$

Hence the two processes have the same positive overall gain and opposite protected-subgroup value differences.

We next bound the statistical distance between the complete audit experiments. Let the updating rule observe n i.i.d. observations from $P_{0,\nu}$ and n i.i.d. observations from $P_{1,\nu}$. We take these two audit samples to be independent in the lower-bound construction. This is a valid special case of the sampling model in Section 3.

For Bernoulli distributions, the inequality

$$\mathrm{kl}(u, v) \le \frac{(u-v)^2}{v(1-v)} \tag{27}$$

holds for $u, v \in (0, 1)$. In every conditional outcome distribution that differs between the two processes, the two Bernoulli parameters differ by $2h$, while the second parameter lies in $[15/32, 5/8]$. Therefore $v(1-v) \ge 15/64$, and

$$\mathrm{kl}(u, v) \le 64h^2. \tag{28}$$

Under the protected-subgroup law, the observed outcome distribution differs only when $U = 1$ and $A = 1$, an event with probability $p/2$. Hence

$$\mathrm{KL}(P_{1,-} \| P_{1,+}) \le 32ph^2. \tag{29}$$

Under the overall law, conditional outcome distributions differ only when $U = 1$ and $A = 1$. Splitting this event by $G \in \{0, 1\}$ and using $P_0(G = g) = 1/2$ gives

$$\mathrm{KL}(P_{0,-} \| P_{0,+}) \le \frac{p}{4}(64h^2) + \frac{p}{4}(64h^2) = 32ph^2. \tag{30}$$

By additivity of Kullback-Leibler divergence over the two independent audit samples,

$$\begin{aligned} &\mathrm{KL}\big(P_{0,-}^{\otimes n} \otimes P_{1,-}^{\otimes n} \,\big\|\, P_{0,+}^{\otimes n} \otimes P_{1,+}^{\otimes n}\big) \\ &\quad \le 32nph^2 + 32nph^2 = 64nph^2 = \frac{1}{16}. \end{aligned} \tag{31}$$

Pinsker's inequality yields

$$\mathrm{TV}\big(P_{0,-}^{\otimes n} \otimes P_{1,-}^{\otimes n}, P_{0,+}^{\otimes n} \otimes P_{1,+}^{\otimes n}\big) \le \sqrt{\frac{1}{32}} < \frac{1}{4}. \tag{32}$$

Represent a possibly randomized deployment rule by a Markov kernel from the complete audit input to $\widehat{\lambda} \in [0, 1]$, and let Q_- and Q_+ denote the induced output laws. A common Markov kernel cannot increase total variation, so

$$\mathrm{TV}(Q_-, Q_+) < \frac{1}{4}. \tag{33}$$

Under $\mathbb{P}_-$, the protected-subgroup value difference of a deployment with candidate weight $\widehat{\lambda}$ is

$$D_{1,-}(\widehat{\lambda}) = \widehat{\lambda} D_{1,-}(d) = -\widehat{\lambda} ph.$$

With zero subgroup tolerance, safety holds only when $\widehat{\lambda} = 0$. Uniform $(1-\alpha)$ safety under $\mathbb{P}_-$ therefore implies

$$Q_-(\widehat{\lambda} = 0) \ge 1 - \alpha. \tag{34}$$

Combining equation 33 and equation 34,

$$Q_+(\widehat{\lambda} = 0) \ge 1 - \alpha - \frac{1}{4} \ge \frac{5}{8}, \tag{35}$$

where the last inequality uses $\alpha \le 1/8$.

Under $\mathbb{P}_+$, both the overall and protected-subgroup value differences of the candidate are positive. The oracle therefore chooses $\lambda = 1$ and attains $W_+^* = \Gamma$. Since $D_{0,+}(\widehat{\lambda}) = \widehat{\lambda}\Gamma$,

$$\begin{aligned}\mathbb{E}_+\Big[W_+^* - D_{0,+}(\widehat{\lambda})\Big] &= \Gamma\,\mathbb{E}_+[1-\widehat{\lambda}] \\ &\geq \Gamma Q_+(\widehat{\lambda}=0) \\ &\geq \frac{5}{8}\Gamma. \end{aligned} \tag{36}$$

Together with equation 25 and equation 26, this proves the theorem. □

Corollary 4 (Local minimax regret at the boundary scale)**.** *Under the setting of Theorem 2, set* $\Gamma = \delta_n$ *and let* $\mathcal{P}_{p,n} := \{\mathbb{P}_-, \mathbb{P}_+\}$*. Let* $\mathcal{A}_\alpha(\mathcal{P}_{p,n})$ *be the class of possibly randomized deployment rules that are* $(1-\alpha)$*-safe under both distributions in* $\mathcal{P}_{p,n}$*. Then*

$$\frac{5}{256}\sqrt{\frac{p}{n}} \;\leq\; \inf_{\widehat{\lambda}\in\mathcal{A}_\alpha(\mathcal{P}_{p,n})} \sup_{\mathbb{P}\in\mathcal{P}_{p,n}} \mathbb{E}_{\mathbb{P}}\Big[W_{\mathbb{P}}^* - D_{0,\mathbb{P}}(\widehat{\lambda})\Big] \;\leq\; \frac{1}{32}\sqrt{\frac{p}{n}}. \tag{37}$$

Proof. Set $\Gamma = \delta_n = (1/32)\sqrt{p/n}$. Since $np \geq 1$,

$$\frac{\Gamma}{p} = \frac{1}{32\sqrt{np}} \leq \frac{1}{32} < \frac{1}{16},$$

so this choice satisfies the range required by Theorem 2. The theorem implies that every rule in $\mathcal{A}_\alpha(\mathcal{P}_{p,n})$ has expected regret at least

$$\frac{5}{8}\Gamma = \frac{5}{256}\sqrt{\frac{p}{n}}$$

on $\mathbb{P}_+$. Taking the supremum over $\mathcal{P}_{p,n}$ and then the infimum over uniformly safe rules gives the lower bound.

For the upper bound, consider the rule that always deploys the incumbent, $\widehat{\lambda} = 0$. It is safe with probability one under both distributions. Its regret is zero under $\mathbb{P}_-$, where the candidate violates the protected-subgroup constraint and the oracle also chooses the incumbent, and equals Γ under $\mathbb{P}_+$. Hence

$$\inf_{\widehat{\lambda}\in\mathcal{A}_\alpha(\mathcal{P}_{p,n})} \sup_{\mathbb{P}\in\mathcal{P}_{p,n}} \mathbb{E}_{\mathbb{P}}\Big[W_{\mathbb{P}}^* - D_{0,\mathbb{P}}(\widehat{\lambda})\Big] \leq \Gamma = \frac{1}{32}\sqrt{\frac{p}{n}}.$$

Combining the two inequalities proves the stated local minimax rate. □

Corollary 4 gives a $\Theta(\sqrt{p/n})$ local minimax regret rate when the oracle gain is on the boundary scale. Theorem 2 also permits larger Γ: even when the overall gain is fixed and positive, a protected-subgroup value difference of order $\sqrt{p/n}$ cannot be uniformly distinguished from an unsafe alternative without nonvanishing conservatism. The upper confidence bounds contain additional logarithmic factors because they control multiple candidates and evaluation populations simultaneously.

A.5 Proof of Theorem 3

Proof. Let

$$\mathcal{E}_D := \bigcap_{s\in\mathcal{S}} \bigcap_{j=0}^{M} \Big\{|\widehat{D}_{sj} - D_{sj}| \leq c_{sj}\Big\}$$

be the simultaneous confidence event from Theorem 1, and fix an arbitrary realization in $\mathcal{E}_D$. The proof is then deterministic.

For every $s \in \mathcal{S}$ and $j \in \{0, \ldots, M\}$,

$$D_{sj} - c_{sj} \leq \widehat{D}_{sj} \leq D_{sj} + c_{sj}.$$

Since $\underline{D}_{sj} = \widehat{D}_{sj} - c_{sj}$, it follows that

$$D_{sj} - 2c_{sj} \le \underline{D}_{sj} \le D_{sj}. \tag{38}$$

For any $\theta \in \Delta_M$, define

$$\underline{D}_s(\theta) := \sum_{j=0}^{M} \theta_j \underline{D}_{sj}, \qquad C_s(\theta) := \sum_{j=0}^{M} \theta_j c_{sj}.$$

Because all mixture weights are nonnegative, summing equation 38 gives, simultaneously for every $s \in \mathcal{S}$ and every $\theta \in \Delta_M$,

$$D_s(\theta) - 2C_s(\theta) \le \underline{D}_s(\theta) \le D_s(\theta). \tag{39}$$

The fact that equation 39 holds for every θ is important, because the selected mixture $\widehat{\theta}$ is itself audit-data dependent.

Let

$$\widehat{\mathcal{C}} := \{\theta \in \Delta_M : \underline{D}_s(\theta) \ge -\Delta_s \text{ for all } s \in \mathcal{H}\}$$

be the feasible set of the empirical linear program. The baseline vector $e_0 = (1, 0, \ldots, 0)$ satisfies $D_{s0} = \widehat{D}_{s0} = c_{s0} = \underline{D}_{s0} = 0$ for every s, and therefore $e_0 \in \widehat{\mathcal{C}}$ because $\Delta_s \ge 0$. Hence the empirical feasible set is nonempty.

For a protected subgroup $s \in \mathcal{H}$, feasibility of $\widehat{\theta}$ and the upper inequality in equation 39 imply

$$D_s(\widehat{\theta}) \ge \underline{D}_s(\widehat{\theta}) \ge -\Delta_s. \tag{40}$$

For the overall population, optimality of $\widehat{\theta}$ together with feasibility of e_0 gives

$$D_0(\widehat{\theta}) \ge \underline{D}_0(\widehat{\theta}) \ge \underline{D}_0(e_0) = 0. \tag{41}$$

Thus the desired safety statements hold on $\mathcal{E}_D$.

It remains to establish the achievable-value comparison. Define the confidence-adjusted population feasible set

$$\mathcal{C}^- := \{\theta \in \Delta_M : D_s(\theta) - 2C_s(\theta) \ge -\Delta_s \text{ for all } s \in \mathcal{H}\}.$$

If $\theta \in \mathcal{C}^-$, then by the lower inequality in equation 39,

$$\underline{D}_s(\theta) \ge D_s(\theta) - 2C_s(\theta) \ge -\Delta_s \qquad (s \in \mathcal{H}).$$

Hence

$$\mathcal{C}^- \subseteq \widehat{\mathcal{C}}. \tag{42}$$

Let θ^- be an optimizer of

$$\max_{\theta \in \mathcal{C}^-} \{D_0(\theta) - 2C_0(\theta)\}.$$

The set $\mathcal{C}^-$ is nonempty because it contains e_0. By equation 42, θ^- is feasible for the empirical problem. Therefore,

$$\begin{aligned} D_0(\widehat{\theta}) &\ge \underline{D}_0(\widehat{\theta}) \\ &\ge \underline{D}_0(\theta^-) \\ &\ge D_0(\theta^-) - 2C_0(\theta^-) \\ &= \max_{\theta \in \mathcal{C}^-} \{D_0(\theta) - 2C_0(\theta)\}. \end{aligned} \tag{43}$$

Expanding the definitions of $D_s(\theta)$ and $C_s(\theta)$ in equation 43 gives exactly the constrained maximum stated in the theorem.

Finally, Theorem 1 gives

$$\mathbb{P}(\mathcal{E}_D) \ge 1 - \frac{\alpha}{2} \ge 1 - \alpha.$$

Hence the safety event in Definition 1 holds with probability at least $1 - \alpha$, while the achievable-value comparison holds on the same event. □

A.6 Proof of Corollary 2

Proof. Fix a realization of the simultaneous confidence event $\mathcal{E}_D$ from Theorem 1. Recall

$$C_s(\theta) = \sum_{j=0}^{M} \theta_j c_{sj}, \qquad D_s(\theta) = \sum_{j=0}^{M} \theta_j D_{sj}.$$

Let θ^* be an optimal solution of the oracle problem. Then

$$D_s(\theta^*) \geq -\Delta_s \quad (s \in \mathcal{H}), \qquad D_0(\theta^*) = W^*. \tag{44}$$

Theorem 3 implies, on $\mathcal{E}_D$,

$$D_0(\widehat{\theta}) \geq \max_{\theta \in \mathcal{C}^-} \{D_0(\theta) - 2C_0(\theta)\}, \tag{45}$$

where

$$\mathcal{C}^- = \{\theta \in \Delta_M : D_s(\theta) - 2C_s(\theta) \geq -\Delta_s \text{ for all } s \in \mathcal{H}\}.$$

For the first claim, suppose

$$D_s(\theta^*) + \Delta_s \geq 2C_s(\theta^*) \qquad (s \in \mathcal{H}).$$

Then

$$D_s(\theta^*) - 2C_s(\theta^*) \geq -\Delta_s \qquad (s \in \mathcal{H}),$$

so $\theta^* \in \mathcal{C}^-$. Evaluating equation 45 at θ^* and using equation 44,

$$D_0(\widehat{\theta}) \geq D_0(\theta^*) - 2C_0(\theta^*) = W^* - 2C_0(\theta^*).$$

Rearranging,

$$W^* - D_0(\widehat{\theta}) \leq 2C_0(\theta^*).$$

For the second claim, assume $\Delta_s > 0$ for all $s \in \mathcal{H}$ and define

$$\rho^* := \min\left\{1, \min_{s \in \mathcal{H}} \frac{\Delta_s}{\Delta_s + 2C_s(\theta^*)}\right\}.$$

Since $C_s(\theta^*) \geq 0$, we have $0 < \rho^* \leq 1$, and for every protected subgroup $s \in \mathcal{H}$,

$$\rho^* \{\Delta_s + 2C_s(\theta^*)\} \leq \Delta_s. \tag{46}$$

Let $e_0 = (1, 0, \ldots, 0)$ denote the baseline vector and set

$$\widetilde{\theta} := \rho^* \theta^* + (1 - \rho^*) e_0.$$

Because both D_s and C_s are linear in θ, while $D_s(e_0) = C_s(e_0) = 0$,

$$D_s(\widetilde{\theta}) = \rho^* D_s(\theta^*), \qquad C_s(\widetilde{\theta}) = \rho^* C_s(\theta^*). \tag{47}$$

Using oracle feasibility in equation 44, equation 46, and equation 47,

$$\begin{aligned} D_s(\widetilde{\theta}) - 2C_s(\widetilde{\theta}) &= \rho^* \{D_s(\theta^*) - 2C_s(\theta^*)\} \\ &\geq -\rho^* \{\Delta_s + 2C_s(\theta^*)\} \\ &\geq -\Delta_s. \end{aligned} \tag{48}$$

Thus $\widetilde{\theta} \in \mathcal{C}^-$. Substituting this feasible comparison point into equation 45,

$$\begin{aligned} D_0(\widehat{\theta}) &\geq D_0(\widetilde{\theta}) - 2C_0(\widetilde{\theta}) \\ &= \rho^* \{W^* - 2C_0(\theta^*)\}. \end{aligned} \tag{49}$$

Therefore

$$W^* - D_0(\widehat{\theta}) \leq (1 - \rho^*) W^* + 2\rho^* C_0(\theta^*),$$

which proves the general regret bound. □

A.7 Proof of Corollary 3

Proof. Condition on the training σ-field $\mathcal{F}_{\mathrm{tr}}$ and fix an oracle optimizer θ^* for the realized candidate library. Under this conditioning, D_{sj}, p_{sj}, r_{sj}, and θ^* are fixed. Define the event

$$\mathcal{E}_r := \bigcap_{s\in\mathcal{S}} \bigcap_{j=0}^{M} \left\{ |\widehat{D}_{sj} - D_{sj}| \le c_{sj} \le r_{sj} \right\}.$$

By Corollary 1,

$$\mathbb{P}(\mathcal{E}_r \mid \mathcal{F}_{\mathrm{tr}}) \ge 1-\alpha \qquad \text{almost surely.} \tag{50}$$

Fix a realization in $\mathcal{E}_r$. Because this event is contained in the simultaneous confidence event, Theorem 3 and Corollary 2 both apply.

For $s \in \mathcal{S}$, define

$$R_s^* := \sum_{j=0}^{M} \theta_j^* r_{sj}, \qquad C_s(\theta^*) := \sum_{j=0}^{M} \theta_j^* c_{sj}.$$

Since $\theta_j^* \ge 0$ and $c_{sj} \le r_{sj}$ on $\mathcal{E}_r$,

$$0 \le C_s(\theta^*) \le R_s^* \qquad (s \in \mathcal{S}). \tag{51}$$

Suppose first that

$$D_s(\theta^*) + \Delta_s \ge 2R_s^* \qquad (s \in \mathcal{H}).$$

By equation 51,

$$D_s(\theta^*) + \Delta_s \ge 2C_s(\theta^*) \qquad (s \in \mathcal{H}).$$

Hence the first part of Corollary 2 yields

$$W^* - D_0(\widehat{\theta}) \le 2C_0(\theta^*) \le 2R_0^*.$$

Because $r_{00} = 0$,

$$\begin{aligned} 2R_0^* &= 2\sum_{j=1}^{M} \theta_j^* r_{0j} \\ &= 2L \sum_{j=1}^{M} \theta_j^* \left(\sqrt{\frac{3p_{0j}\ell_\alpha}{n_0-1}} + \frac{7\ell_\alpha}{n_0-1} \right), \end{aligned} \tag{52}$$

which is the first stated bound.

For the general case, assume $\Delta_s > 0$ for all $s \in \mathcal{H}$ and define

$$\bar{\rho} := \min\left\{ 1,\ \min_{s\in\mathcal{H}} \frac{\Delta_s}{\Delta_s + 2R_s^*} \right\}.$$

The coefficient ρ^* in Corollary 2 is

$$\rho^* = \min\left\{ 1,\ \min_{s\in\mathcal{H}} \frac{\Delta_s}{\Delta_s + 2C_s(\theta^*)} \right\}.$$

By equation 51,

$$0 < \bar{\rho} \le \rho^* \le 1. \tag{53}$$

Corollary 2 gives

$$W^* - D_0(\widehat{\theta}) \le (1-\rho^*)W^* + 2\rho^* C_0(\theta^*). \tag{54}$$

To compare the right-hand side with the desired expression, consider

$$f(\rho) := (1-\rho)W^* + 2\rho C_0(\theta^*).$$

If $2C_0(\theta^*) < W^*$, then f is decreasing. By equation 53,

$$f(\rho^*) \le f(\bar{\rho}) = (1-\bar{\rho})W^* + 2\bar{\rho}C_0(\theta^*) \le (1-\bar{\rho})W^* + 2\bar{\rho}R_0^*.$$

Together with equation 54, this yields the desired inequality.

If instead $2C_0(\theta^*) \geq W^*$, then equation 51 implies $2R_0^* \geq W^*$. On $\mathcal{E}_r$, Theorem 3 guarantees $D_0(\widehat{\theta}) \geq 0$, and hence

$$W^* - D_0(\widehat{\theta}) \leq W^* \leq (1-\bar{\rho})W^* + 2\bar{\rho}R_0^*.$$

Thus in both cases

$$W^* - D_0(\widehat{\theta}) \leq (1-\bar{\rho})W^* + 2\bar{\rho}\sum_{j=0}^{M}\theta_j^* r_{0j}.$$

The two conclusions hold on the single event $\mathcal{E}_r$. By equation 50, their conditional probability is at least $1-\alpha$ whenever the corresponding premises hold for the realized candidate library. Taking expectations over the training stage gives the unconditional statements. □

B Extension to Observational Data

The main finite-sample analysis assumes that the logging propensities $e_s(a \mid x)$ are recorded and therefore known. We now consider the observational setting in which the propensity scores must instead be estimated from data. The candidate policies and all nuisance estimators are constructed using the training sample only and are frozen before the audit data are observed. Thus, throughout this subsection we again condition on the training σ-field $\mathcal{F}_{\mathrm{tr}}$.

For each $s \in \mathcal{S}$, let

$$m_s(x,a) := \mathbb{E}_s[Y \mid X = x, A = a]$$

denote the true outcome regression. Let $\widetilde{e}_s(a \mid x)$ and $\widetilde{m}_s(x,a)$ denote propensity and outcome-regression estimators obtained from the training data. We assume that the estimated propensity is bounded away from zero,

$$\widetilde{e}_s(a \mid x) \geq \frac{\kappa}{2}, \qquad \forall\, s \in \mathcal{S},\ a \in \mathcal{A}, \tag{55}$$

and that $\widetilde{m}_s$ is clipped so that $|\widetilde{m}_s(x,a)| \leq B$. The lower bound in equation 55 may be enforced by propensity clipping.

For notational convenience, define

$$q_{j,a}(x) := \mathbf{1}\{d_j(x) = a\} - \mathbf{1}\{b(x) = a\}.$$

Replacing the true propensity in equation 1 by its estimate gives the observational paired score

$$\widetilde{Z}_{sj}(O) = \sum_{a \in \mathcal{A}} q_{j,a}(X)\left[\widetilde{m}_s(X,a) + \frac{\mathbf{1}\{A=a\}}{\widetilde{e}_s(a \mid X)}\{Y - \widetilde{m}_s(X,a)\}\right]. \tag{56}$$

As in the known-propensity setting, $\widetilde{Z}_{sj}(O) = 0$ whenever $d_j(X) = b(X)$. Moreover, under Assumption 2 and equation 55,

$$|\widetilde{Z}_{sj}(O)| \leq \widetilde{L} := 2B(1 + 2\kappa^{-1}) \qquad \text{almost surely.} \tag{57}$$

The difference from the main analysis is that $\widetilde{Z}_{sj}$ need not be exactly unbiased for D_{sj}. The resulting bias has the usual doubly robust product structure, localized to the region on which the candidate and incumbent disagree.

Proposition 2 (Doubly robust remainder under estimated propensities)**.** *Suppose Assumptions 1 and 2 hold and equation 55 is satisfied. Conditional on $\mathcal{F}_{\mathrm{tr}}$,*

$$\mathbb{E}_s[\widetilde{Z}_{sj}(O)] - D_{sj} = \mathbb{E}_s\left[\sum_{a \in \mathcal{A}} q_{j,a}(X)\frac{\widetilde{e}_s(a \mid X) - e_s(a \mid X)}{\widetilde{e}_s(a \mid X)}\big\{\widetilde{m}_s(X,a) - m_s(X,a)\big\}\right]. \tag{58}$$

Consequently,

$$\Big|\mathbb{E}_s[\widetilde{Z}_{sj}(O)] - D_{sj}\Big| \leq \beta_{sj}, \tag{59}$$

where

$$\beta_{sj} := \frac{2}{\kappa}\sum_{a \in \mathcal{A}} \big\|q_{j,a}\big(\widetilde{e}_{s,a} - e_{s,a}\big)\big\|_{2,P_s}\,\big\|q_{j,a}\big(\widetilde{m}_{s,a} - m_{s,a}\big)\big\|_{2,P_s}, \tag{60}$$

with $\widetilde{e}_{s,a}(x) := \widetilde{e}_s(a \mid x)$, $e_{s,a}(x) := e_s(a \mid x)$, and analogously for $\widetilde{m}_{s,a}$ and $m_{s,a}$.

Proof. Fix $s \in \mathcal{S}$ and condition on $\mathcal{F}_{\mathrm{tr}}$. For every $a \in \mathcal{A}$,

$$\begin{aligned}
&\mathbb{E}_s\left[\widetilde{m}_s(X,a) + \frac{\mathbf{1}\{A=a\}}{\widetilde{e}_s(a \mid X)}\{Y - \widetilde{m}_s(X,a)\} \,\middle|\, X, \mathcal{F}_{\mathrm{tr}}\right] \\
&\quad = \widetilde{m}_s(X,a) + \frac{e_s(a \mid X)}{\widetilde{e}_s(a \mid X)}\{m_s(X,a) - \widetilde{m}_s(X,a)\} \\
&\quad = m_s(X,a) + \frac{\widetilde{e}_s(a \mid X) - e_s(a \mid X)}{\widetilde{e}_s(a \mid X)}\{\widetilde{m}_s(X,a) - m_s(X,a)\}. \qquad (61)
\end{aligned}$$

Multiplying equation 61 by $q_{j,a}(X)$, summing over a, and taking expectations gives

$$\begin{aligned}
\mathbb{E}_s[\widetilde{Z}_{sj}(O)] &= \mathbb{E}_s\left[\sum_{a\in\mathcal{A}} q_{j,a}(X) m_s(X,a)\right] \\
&\quad + \mathbb{E}_s\left[\sum_{a\in\mathcal{A}} q_{j,a}(X)\frac{\widetilde{e}_s(a \mid X) - e_s(a \mid X)}{\widetilde{e}_s(a \mid X)}\{\widetilde{m}_s(X,a) - m_s(X,a)\}\right]. \qquad (62)
\end{aligned}$$

By Assumption 1, the first term equals D_{sj}, proving equation 58.

For the second claim, using $\widetilde{e}_s(a \mid X) \geq \kappa/2$ gives $1/\widetilde{e}_s(a \mid X) \leq 2/\kappa$. Hence,

$$\begin{aligned}
\left|\mathbb{E}_s[\widetilde{Z}_{sj}] - D_{sj}\right| &\leq \frac{2}{\kappa}\sum_{a\in\mathcal{A}} \mathbb{E}_s\left[|q_{j,a}(X)||\widetilde{e}_{s,a}(X) - e_{s,a}(X)||\widetilde{m}_{s,a}(X) - m_{s,a}(X)|\right] \\
&= \frac{2}{\kappa}\sum_{a\in\mathcal{A}} \mathbb{E}_s\left[|q_{j,a}(X)\{\widetilde{e}_{s,a}(X) - e_{s,a}(X)\}|\,|q_{j,a}(X)\{\widetilde{m}_{s,a}(X) - m_{s,a}(X)\}|\right] \\
&\leq \frac{2}{\kappa}\sum_{a\in\mathcal{A}} \left\|q_{j,a}(\widetilde{e}_{s,a} - e_{s,a})\right\|_{2,P_s} \left\|q_{j,a}(\widetilde{m}_{s,a} - m_{s,a})\right\|_{2,P_s}, \qquad (63)
\end{aligned}$$

where the equality uses $q_{j,a} \in \{-1, 0, 1\}$ and hence $|q_{j,a}| = q_{j,a}^2$, while the final inequality follows from Cauchy–Schwarz. This is exactly equation 60. □

Proposition 2 makes two properties explicit. First, the remainder vanishes if either the propensity model or the outcome regression is correctly specified. Second, both nuisance errors are multiplied by $q_{j,a}$, so only their errors on the candidate–incumbent disagreement region enter the bias. Thus the local structure exploited by DiSPU is retained even when the logging propensity must be estimated.

We next state how this remainder modifies the finite-sample certification step. Define

$$\widetilde{D}_{sj} := \frac{1}{n_s}\sum_{i=1}^{n_s} \widetilde{Z}_{sj}(O_{is}), \qquad \widetilde{\sigma}_{sj}^2 := \frac{1}{n_s - 1}\sum_{i=1}^{n_s}\left(\widetilde{Z}_{sj}(O_{is}) - \widetilde{D}_{sj}\right)^2. \qquad (64)$$

Let $\alpha_{\mathrm{aud}} > 0$ and $\alpha_\eta \geq 0$ denote failure probability budgets satisfying

$$\alpha_{\mathrm{aud}} + \alpha_\eta \leq \alpha,$$

and define

$$\ell_{\mathrm{aud}} := \log\left(\frac{4N}{\alpha_{\mathrm{aud}}}\right), \qquad N = M|\mathcal{S}|.$$

The corresponding empirical Bernstein radius is

$$\widetilde{c}_{sj} := \sqrt{\frac{2\widetilde{\sigma}_{sj}^2 \ell_{\mathrm{aud}}}{n_s}} + \frac{14\widetilde{L}\ell_{\mathrm{aud}}}{3(n_s - 1)}, \qquad j \geq 1, \qquad (65)$$

with $\widetilde{c}_{s0} = 0$.

The disagreement-adaptive sampling behavior is unchanged. In particular, since $\widetilde{Z}_{sj}(O) = 0$ whenever $d_j(X) = b(X)$ and $|\widetilde{Z}_{sj}(O)| \leq \widetilde{L}$, the deterministic argument of Proposition 1 gives

$$\widetilde{\sigma}_{sj}^2 \leq \frac{n_s}{n_s - 1}\widetilde{L}^2\widehat{p}_{sj}, \qquad (66)$$

and therefore

$$\widetilde{c}_{sj} \le \widetilde{L}\sqrt{\frac{2\widehat{p}_{sj}\ell_{\mathrm{aud}}}{n_s-1}} + \frac{14\widetilde{L}\ell_{\mathrm{aud}}}{3(n_s-1)}. \tag{67}$$

Thus propensity estimation introduces an additional nuisance remainder, but does not remove the $\sqrt{p_{sj}/n_s}$ localization of the sampling uncertainty.

To obtain a finite-sample safety certificate, however, the nuisance remainder itself must also be controlled. Suppose there exist nonnegative, training-measurable quantities $\{\overline{\beta}_{sj}\}$ such that, with $\overline{\beta}_{s0} := 0$,

$$\mathbb{P}\left(\beta_{sj} \le \overline{\beta}_{sj} \text{ for all } s \in \mathcal{S},\ j = 1, \dots, M\right) \ge 1 - \alpha_\eta. \tag{68}$$

We then define the observational-data lower confidence bound

$$\underline{D}_{sj}^{\mathrm{obs}} := \widetilde{D}_{sj} - \widetilde{c}_{sj} - \overline{\beta}_{sj}, \qquad j \ge 1, \tag{69}$$

and set $\underline{D}_{s0}^{\mathrm{obs}} = 0$.

Corollary 5 (Safe deployment with estimated propensities)**.** *Under the conditions of Proposition 2, suppose the nuisance certificate equation 68 holds. Then, with probability at least* $1 - \alpha_{\mathrm{aud}} - \alpha_\eta$*, simultaneously for all* $s \in \mathcal{S}$ *and* $j \in \{0, \dots, M\}$*,*

$$D_{sj} \ge \underline{D}_{sj}^{\mathrm{obs}}. \tag{70}$$

Consequently, replacing $\underline{D}_{sj}$ *in equation 4 by* $\underline{D}_{sj}^{\mathrm{obs}}$ *yields deployment weights* $\widehat{\theta}_{\mathrm{obs}}$ *satisfying*

$$D_0(\widehat{\theta}_{\mathrm{obs}}) \ge 0, \qquad D_s(\widehat{\theta}_{\mathrm{obs}}) \ge -\Delta_s, \quad \forall s \in \mathcal{H}, \tag{71}$$

with probability at least $1 - \alpha_{\mathrm{aud}} - \alpha_\eta \ge 1 - \alpha$*.*

Proof. Conditional on $\mathcal{F}_{\mathrm{tr}}$, the scores $\widetilde{Z}_{sj}(O_{is})$ are i.i.d. across i for each fixed (s, j) and are bounded by $\widetilde{L}$ according to equation 57. Applying the same empirical Bernstein argument as in the proof of Theorem 1, with per-comparison failure probability α_{aud}/N, gives

$$\left|\widetilde{D}_{sj} - \mathbb{E}_s[\widetilde{Z}_{sj}]\right| \le \widetilde{c}_{sj} \tag{72}$$

simultaneously over all s and $j \ge 1$ with probability at least $1 - \alpha_{\mathrm{aud}}$.

On the nuisance-certification event in equation 68, Proposition 2 gives

$$\left|\mathbb{E}_s[\widetilde{Z}_{sj}] - D_{sj}\right| \le \overline{\beta}_{sj}.$$

Therefore, on the intersection of the two events,

$$\left|\widetilde{D}_{sj} - D_{sj}\right| \le \widetilde{c}_{sj} + \overline{\beta}_{sj} \tag{73}$$

for every (s, j). A union bound over the two sources of failure gives probability at least $1-\alpha_{\mathrm{aud}}-\alpha_\eta$. equation 70 follows immediately.

The proof of Theorem 3 then applies without change: feasibility of the confidence-adjusted linear program and the coordinatewise lower bounds imply all protected-subgroup constraints, while the incumbent remains feasible with objective value zero and therefore guarantees non-degradation in the overall population. □

Corollary 5 separates the two sources of uncertainty introduced by observational logging. The sampling component $\widetilde{c}_{sj}$ remains disagreement-adaptive and retains the leading $\sqrt{p_{sj}/n_s}$ behavior established in the main analysis, whereas $\overline{\beta}_{sj}$ accounts for uncertainty from nuisance estimation. The latter also inherits a form of disagreement localization through the factors $q_{j,a}$ in equation 60.

Importantly, standard asymptotic product-rate conditions on $\widetilde{e}_s - e_s$ and $\widetilde{m}_s - m_s$ are sufficient for making the remainder asymptotically negligible, but they are not by themselves computable finite-sample safety certificates. A finite-sample guarantee in the observational setting therefore additionally requires a valid bound such as $\overline{\beta}_{sj}$, obtained from model-specific validation, an external propensity guarantee, a randomized audit mechanism, or an explicit sensitivity bound. In the absence of such a certificate, the observational paired estimator may still be useful for estimation, but the finite-sample safety guarantee of Definition 1 does not automatically follow.

C Experimental Implementation

C.1 Datasets

Synthetic dataset. We construct a binary-action contextual-bandit environment with a finite covariate space that permits exact evaluation of all policy values. Each context is represented as $X = (G, U) \in \{0,1\}^{12}$, where $G = (G_1, G_2, G_3)$ contains the protected-attribute indicators and $U = (U_1, \ldots, U_9)$ contains the covariates available to the policy learner. These indicators follow

$$P(G = (0,0,0)) = 0.50, \qquad P(G \in \{(1,0,0),(0,1,0),(0,0,1)\}) = 0.10,$$

and

$$P(G \in \{(1,1,0),(1,0,1),(0,1,1),(1,1,1)\}) = 0.05,$$

while $U \mid G$ is uniform over $\{0,1\}^9$. Consequently, each marginal protected subgroup $G_k = 1$ has probability 0.25, and each pairwise intersection has probability 0.10. We use the six protected subgroups G_1, G_2, G_3, $G_1 \cap G_2$, $G_1 \cap G_3$, and $G_2 \cap G_3$. The incumbent policy is $b(U) = \mathbf{1}\{U_1 = U_2 = U_3 = 1\}$. To induce localized and subgroup-dependent treatment effects, define $Z_k = 2U_k - 1$ and the four disjoint regions

$$\begin{aligned} R_1 &= \{U_1 = U_2 = U_3 = U_4 = 1\}, \\ R_2 &= \{U_1 = U_2 = U_3 = 1,\ U_4 = 0\}, \\ R_3 &= \{U_1 = 0,\ U_5 = U_6 = U_7 = 1\}, \\ R_4 &= \{U_1 = 0,\ U_5 = U_6 = 1,\ U_7 = 0\}. \end{aligned}$$

The conditional mean under action 0 is

$$\mu_0(X) = \begin{cases} 0.34, & X \in R_1 \cup R_2, \\ 0.60, & X \in R_3 \cup R_4, \\ 0.50, & \text{otherwise}, \end{cases} \quad + \ 0.015 Z_8 + 0.01 Z_9,$$

and $\mu_1(X) = \mu_0(X) + \eta(X)$, where

$$\eta(X) = \begin{cases} 0.28 - 0.08G_1 + 0.02(2G_2 - 1) + 0.01Z_8, & X \in R_1, \\ -0.20 + 0.75G_1 - 0.08G_3 + 0.01Z_8, & X \in R_2, \\ 0.32 - 0.85G_2 + 0.04G_1 + 0.01Z_9, & X \in R_3, \\ 0.32 - 0.85G_3 + 0.04G_2 + 0.01Z_8, & X \in R_4, \\ -0.05 + 0.01Z_8 Z_9, & \text{otherwise}. \end{cases}$$

Logging actions are randomized as $A \sim \text{Bernoulli}(1/2)$, and we draw an uncentered binary outcome $\widetilde{Y} \mid X, A \sim \text{Bernoulli}(\mu_A(X))$ and use the centered outcome $Y = \widetilde{Y} - 1/2 \in \{-1/2, 1/2\}$. Thus the logging propensity is known and equal to $1/2$, and the outcome bound in Assumption 2 holds with $B = 1/2$. Candidate policies are restricted to functions of U, so protected-group indicators are not directly available to the policy learner, whereas the nuisance outcome regression may use the full covariate vector X. In each run, we generate $50{,}000$ training observations, using $40{,}000$ for model fitting and $10{,}000$ for candidate construction and selection. A separate audit stream is generated independently and evaluated at nested prefixes $n_{\text{aud}} \in \{10{,}000, 25{,}000, 50{,}000, 100{,}000, 200{,}000\}$. Protected-subgroup audit samples are obtained by filtering this common audit stream.

Semi-synthetic dataset. We construct a semi-synthetic benchmark from the 2018 American Community Survey (ACS) 1-Year Public Use Microdata Sample, using real individual-level covariates but generating all logging actions and policy outcomes synthetically. Our implementation considers the ACSIncome task. We retain individuals with $\text{AGEP} > 16$, $\text{PINCP} > 100$, $\text{WKHP} > 0$, and $\text{PWGTP} \geq 1$, and define the original prediction label as $\mathbf{1}\{\text{PINCP} > 50{,}000\}$. Twenty percent of the eligible records are reserved exclusively for constructing the semi-synthetic environment, and the remaining 80% form the empirical covariate population from which all training and audit logs are sampled. On the construction split, we fit a histogram gradient-boosting classifier to the original task label using five-fold out-of-fold predictions, calibrate these predictions using isotonic regression, and refit the classifier on the full construction split. Its calibrated prediction on the empirical

population, clipped to $[0.05, 0.95]$, is denoted by $q_{\text{env}}(X)$ and is used only to determine the baseline outcome level. Protected groups are defined by $G_1 = \mathbf{1}\{\text{SEX} = 2\}$, $G_2 = \mathbf{1}\{\text{RAC1P} \neq 1\}$, and $G_3 = \mathbf{1}\{\text{AGEP} \leq 30\}$, together with the three pairwise intersections. To define the incumbent and heterogeneous update effects without directly using these protected attributes, let

$$z(X) = \text{SCHL} + \frac{1}{2} F_{\text{mid}}(R),$$

where $R = \text{WKHP}$, and F_{mid} is the empirical mid-rank distribution function estimated on the construction split. Let $q_{.25}, q_{.375}, q_{.50}, q_{.75}, q_{.875}$ denote the corresponding construction-sample quantiles of z. We define the incumbent as $b(X) = \mathbf{1}\{z(X) \geq q_{.75}\}$ and four disjoint regions

$$R_1 = \{z \geq q_{.875}\}, \quad R_2 = \{q_{.75} \leq z < q_{.875}\}, \quad R_3 = \{q_{.25} \leq z < q_{.375}\}, \quad R_4 = \{q_{.375} \leq z < q_{.50}\}.$$

Writing $V = \mathbf{1}\{G_2 = 1, G_3 = 1\}$, the gain from switching away from the incumbent is set to

$$\gamma(X) = \begin{cases} 0.06, & X \in R_1, \\ 0.12 - 0.60V, & X \in R_2, \\ 0.20 - 0.80V, & X \in R_3, \\ 0.28 - 1.00V, & X \in R_4, \\ -0.04, & \text{otherwise.} \end{cases}$$

We then set $\eta(X) = (1 - 2b(X))\gamma(X)$, $c(X) = 0.5 + 0.1\{2q_{\text{env}}(X) - 1\}$, and

$$\mu_0(X) = c(X) - \frac{1}{2}\eta(X), \qquad \mu_1(X) = c(X) + \frac{1}{2}\eta(X).$$

This construction gives $(1 - 2b(X))\{\mu_1(X) - \mu_0(X)\} = \gamma(X)$, creating regions in which changing the incumbent action is beneficial on average while potentially harmful for the prespecified protected subgroup $G_2 \cap G_3$. All probabilities remain in $[0.05, 0.95]$ without outcome clipping. For every experimental replicate, contexts are sampled uniformly with replacement from the empirical population, logging actions satisfy $A \sim \text{Bernoulli}(1/2)$, and outcomes are generated as $\widetilde{Y} \mid X, A \sim \text{Bernoulli}(\mu_A(X))$ followed by the same centering $Y = \widetilde{Y} - 1/2$. We use $40{,}000$ observations for candidate and nuisance-model fitting, $10{,}000$ independent observations for candidate selection, and $50{,}000$ observations for independent auditing; subgroup audit samples are again formed by filtering the common audit sample according to their prespecified memberships.

C.2 Candidate Policy Construction

For all experiments, candidate policies are constructed using training data only and are frozen before the audit sample is accessed. We use four T-learner families: logistic regression, decision tree, histogram gradient boosting, and random forest. For each learner, separate outcome models are fitted for the two actions, yielding estimated conditional means $\widehat{m}_0(x)$ and $\widehat{m}_1(x)$. Given the incumbent policy b, we define the estimated gain from switching the incumbent action as $\widehat{g}(x) = \{1 - 2b(x)\}\{\widehat{m}_1(x) - \widehat{m}_0(x)\}$. Thus, $\widehat{g}(x) > 0$ indicates that the learned outcome models predict an improvement from replacing the incumbent action by the alternative action. To generate policy updates of different locality, each learner is combined with a disagreement budget $\rho \in \{0.02, 0.05, 0.10, 0.20\}$. For a given ρ, we retain the incumbent action except on observations with the largest positive values of $\widehat{g}(x)$, subject to an empirical disagreement rate of at most ρ on the independent candidate-selection split. Formally, the resulting candidate takes the form

$$d_\rho(x) = \begin{cases} 1 - b(x), & x \in \widehat{\mathcal{R}}_\rho, \\ b(x), & \text{otherwise,} \end{cases}$$

where $\widehat{\mathcal{R}}_\rho$ is determined entirely from the training and selection data. Ties at the disagreement threshold are resolved deterministically so that identical policy covariates are assigned the same action. Combining the four learner families with the four disagreement budgets yields $M = 16$ candidate policies in each experimental replicate. In the synthetic experiment, candidate policies use only the public covariates U and do not observe the protected-group indicators. In the semi-synthetic experiment, candidate policies are likewise constructed from the prespecified policy covariates, while the nuisance outcome regression used for auditing may use the full covariate vector. The complete candidate library and nuisance models are fixed before any audit observations are generated or inspected.