Double Robustness Is Not a Privacy Certificate: Sensitivity Spillover in Private Policy Selection

Anonymous authors
Paper under double-blind review

Abstract

Doubly robust (DR) scores are a standard method in causal inference and policy evaluation: they make policy-value estimates insensitive to global nuisance estimation error. Differential privacy (DP), however, requires a different form of robustness: the released output must be stable under the replacement of any one individual. This paper shows that these two notions of robustness can differ significantly in private policy selection. We study the problem of selecting a high-value policy from a finite public library using learned DR utilities and the exponential mechanism. Although fixed or frozen-nuisance utilities have constant sensitivity, we identify a *sensitivity spillover* effect: replacing one record in the nuisance-training block can change the learned score map, and that changed score map is then evaluated on all records in the scoring block. We prove a separation showing that double robustness, vanishing nuisance error, and even zero DR population bias can coexist with order-n realized utility sensitivity, invalidating the usual fixed-utility privacy calibration. We then give a sufficient certification route based on deterministic replace-one prediction stability of the nuisance learners, which yields a valid pure-DP exponential mechanism and a regret bound separating library approximation, concentration, DR product remainder, and certified privacy cost. Semi-synthetic experiments confirm that spillover can be large even when DR statistical diagnostics look benign, and that stability-oriented regularization controls the privacy-relevant movement. The investigations in this paper highlight an important future need that the private causal policy selection method requires both orthogonality for statistical robustness and algorithmic stability for individual replacement robustness.

1 Introduction

The growing availability of user-specific observational data has made personalized decision making increasingly central in medicine, education, public programs, and online services (Murphy, 2003; Qian & Murphy, 2011; Zhao et al., 2012; Dudík et al., 2011; Kitagawa & Tetenov, 2018; Athey & Wager, 2021). The basic causal object behind such decisions is a policy value: for a candidate policy π, what mean outcome would be obtained if the population were assigned actions according to π? Estimating this quantity from observational data is statistically delicate because actions are not randomized and counterfactual outcomes are unobserved. A successful line of offline policy evaluation and offline policy learning (OPE/OPL) addresses this difficulty by estimating nuisance functions, such as propensities and outcome regressions, and then evaluating or optimizing policies using inverse probability weighting (IPW) or doubly robust (DR) scores over a constrained class (Kallus, 2018; Zhao et al., 2019; Athey & Wager, 2021). The appeal of doubly robust score is its stability with respect to nuisance error: under standard rate conditions, policy-value estimation can remain valid even when one nuisance component is misspecified, and first-order nuisance errors do not propagate directly into the target value (Chernozhukov et al., 2018). This robustness has made learned DR scores a default interface between flexible machine learning and causal policy evaluation.

In real applications, causal policy evaluation often involves sensitive individual records, and differential privacy (DP) protects such records by requiring the released output to be stable under the replacement of any one individual (Dwork et al., 2006; Dwork & Roth, 2014). Recent work at the intersection of DP and causal inference suggests a natural but unresolved tension. On the one hand, privacy mechanisms introduce

deliberate perturbations, and these perturbations can affect causal estimands in mechanism-specific ways: perturbing outcomes, treatments, covariates, nuisance predictions, estimating equations, or the final released estimand need not have the same statistical effect (Lee et al., 2019; Farzam & Sapiro; Ohnishi & Awan, 2024). On the other hand, several recent private causal methods deliberately build on DR structure, because the DR score is robust to nuisance error and compatible with flexible nuisance learning (Niu et al., 2022; Lebeda et al., 2025; Schröder et al., 2025a;b). These works make DR-based causal estimation appear especially attractive under privacy constraints, but they also reveal that the benefit is not automatic: multi-stage estimator construction and nuisance estimation errors can incur additional variance, sensitivity, or privacy cost under DP (Niu et al., 2022; Lebeda et al., 2025).

A common deployment scenario sits between offline policy evaluation and full offline policy learning. Instead of releasing a value estimate for a single policy (OPE), or optimizing over a rich data-dependent policy class (OPL), the analyst may have a finite public library $\Pi = \{\pi_1, \ldots, \pi_M\}$ of candidate policies and wish to release one policy with high population value. Such libraries arise when policies must be interpretable, pre-specified for regulatory or operational reasons, or generated by a separate public or privacy-accounted procedure. We refer to this problem as *policy selection*. Statistically, policy selection can be viewed as evaluating the candidates in the library and choosing a high-value one. Privately, however, the released identity of the selected policy is itself a data-dependent output and must be stable under the replacement of any one individual.

This selection viewpoint suggests a natural use of the exponential mechanism (McSherry & Talwar, 2007). Given a finite public policy library, one computes a utility for each policy and samples with probability increasing in that utility. For fixed bounded utilities of the form $U_{\mathrm{fix}}(D, \pi) = \sum_i q_\pi(Z_i)$ with $|q_\pi| \leq B$, replacing one record changes only one summand, so the global sensitivity is at most $2B$. The same constant-sensitivity argument also applies to frozen-nuisance DR scores after conditioning on the nuisance functions, as discussed in Section 4. Since learned DR scores are the standard interface between flexible nuisance learning and policy-value estimation, it is tempting to train nuisances form learned DR utilities for all policies in the library, and pass these utilities directly to the exponential mechanism. *The premise of this paper is that this path is NOT seamless.* Double robustness and pure differential privacy control different objects. Double robustness controls population bias or statistical error caused by nuisance estimation. Pure DP controls the worst-case movement of the realized utility under every adjacent pair of datasets. When the DR score map is learned from protected records, replacing one nuisance-training record can change the fitted nuisance functions, and this changed score map is then evaluated across the entire policy-scoring block. Thus a single protected record can affect many summands of the realized utility. We call this effect *sensitivity spillover*. It is invisible to the usual DR product remainder, but it is exactly the object that governs the privacy calibration of the exponential mechanism. This distinction leads to the central question of the paper:

> *When can a learned doubly robust policy-value score be validly used as the utility in a pure-DP policy selection mechanism?*

This terminology is important. The core object of this paper is *private policy selection*: the mechanism releases one policy from a finite public library. OPE and OPL provide the causal estimators and optimization viewpoint that motivate our utility, but our privacy analysis is neither an OPE accuracy theorem, nor an end-to-end private policy learning algorithm for constructing the library. It investigates whether the learned DR utilities used to rank an already fixed library have a valid pure-DP sensitivity calibration.

We answer this question by separating the statistical and privacy roles of learned scores. Section 3 introduces three utility regimes: fixed scores, frozen-nuisance DR scores, and internally learned DR scores. In Section 4.1, we first introduce the constant-sensitivity baseline (Appendix A) for the first two regimes, and then show that internally learned DR utilities introduce a new source of sensitivity: replacing one nuisance-training record can perturb the fitted score map, and that perturbation is evaluated across the full scoring block. Section 4.2 proves that statistical orthogonality can coexist with large global sensitivity, so double robustness alone cannot certify pure-DP calibration. Section 4.3 gives a sufficient certification route: a public deterministic replace-one prediction-stability certificate bounds the spillover and yields a valid pure-DP exponential-mechanism calibration. The resulting regret bound separates approximation, scoring concentration, the DR product remainder, and the certified privacy cost.

Our **contributions** are threefold:

- To our knowledge, this is the first work to study pure-DP policy selection with learned doubly robust utilities. We identify *sensitivity spillover*: one nuisance-training record can perturb the fitted score map evaluated on many scoring records. Our separation result shows that vanishing DR population bias can coexist with order-n realized utility sensitivity, so double robustness alone cannot justify pure-DP calibration.
- We give a sufficient certification route based on public deterministic replace-one prediction stability. The certificate yields $\Delta_{\mathrm{cert}} = \max\{2B_\psi, n\rho_m^\star\}$ and hence a pure-DP exponential mechanism. The accompanying regret bound separates approximation, scoring concentration, the DR product remainder, and the certified privacy cost.
- Experimental results show that standard DR accuracy metrics can fail to reveal large adjacent-dataset movement in learned policy utilities. The experiments further confirm the qualitative prediction in theoretical analysis that stability-oriented regularization reduces nuisance-induced spillover, and a deterministic ERM positive control instantiates the public-certificate route. These results suggest a design principle for private policy selection: orthogonality controls statistical error, while replace-one prediction stability must certify privacy.

2 Related Work

Policy-value estimators based on outcome regression, inverse propensity weighting, and doubly robust scores are standard building blocks for comparing covariate-dependent interventions (Qian & Murphy, 2011; Zhao et al., 2012; Kitagawa & Tetenov, 2018; Kallus, 2017; 2018; Zhao et al., 2019; Athey & Wager, 2021). Several works use these estimators as objectives that are subsequently optimized over policy classes, and recent extensions address many-treatment settings and distribution or concept shifts (Zhu et al., 2025; Si et al., 2023; Kallus et al., 2022; Wang et al., 2025). Orthogonal scores statistically decouple value error from first-order nuisance errors (Chernozhukov et al., 2018). This robustness controls expected error, while privacy calibration requires worst-case utility stability across adjacent datasets. Concurrently, DP causal inference has focused on privatizing specific estimands or uncertainty statements, including private ATE estimation via IPW or covariate balancing (D'Orazio et al., 2015; Lee et al., 2019; Ohnishi & Awan, 2024; Guha & Reiter, 2025; Yuan et al., 2025), locally private randomized inference (Ohnishi & Awan, 2025), private CATE and confidence intervals (Niu et al., 2022; Schröder et al., 2025a;b), and model-agnostic prediction-aggregation mechanisms (Lebeda et al., 2025). Our output is a privately selected policy from a fixed public library. This is a policy-selection problem, rather than only OPE or unrestricted OPL: the learned DR score is used to rank a pre-specified library, and the released object is one selected policy. This shifts the technical challenge because the utility passed to the private selection mechanism is a learned DR policy value whose adjacent-dataset sensitivity is driven by the replaced record and by how that record perturbs the fitted nuisances used to score all other records.

The exponential mechanism and private empirical risk minimization guarantee private selection or optimization for objectives with known bounded sensitivity, convexity, or fixed per-user contributions (McSherry & Talwar, 2007; Dwork & Roth, 2014; Chaudhuri et al., 2011; Bassily et al., 2014). Learned DR policy-value evaluation violates this fixed utility viewpoint. Replacing one nuisance-fitting record alters the fitted propensity and outcome predictions across all scoring records, inducing a sensitivity spillover invisible to usual fixed score calibrations. DP also addresses sequential decision making in bandits and reinforcement learning (Shariff & Sheffet, 2018; Zheng et al., 2020; Garcelon et al., 2022; Huang et al., 2023; Wang & Zhu, 2024; Qiao & Wang, 2023a;b; Liao et al., 2023). These sequential models sidestep the unobserved counterfactuals of static observational settings. Classical algorithmic stability and local sensitivity metrics (Bousquet & Elisseeff, 2002; Shalev-Shwartz et al., 2010; Nissim et al., 2007) do not automatically certify the global worst-case sensitivity required by a pure-DP exponential mechanism. Typical case or local sensitivity bounds would require a modified approximate-DP mechanism, e.g., smooth sensitivity, together with a separate proof controlling bad neighboring datasets. Deterministic replace-one prediction stability provides the auditable certificate used in our pure-DP learned utility calibration.

3 Problem Setup

Policy value and private policy selection. Our paper follows the potential outcome causal inference framework. Consider a dataset of $N = m + n$ records partitioned into a fixed, public, and data-independent split: $D = (D^{\mathrm{tr}}, D^{\mathrm{sc}})$, $D^{\mathrm{tr}} = \{Z_1, \ldots, Z_m\}$, and $D^{\mathrm{sc}} = \{Z_{m+1}, \ldots, Z_{m+n}\}$. We assume each record $Z_i = (X_i, A_i, Y_i)$ is drawn independently from a population distribution P. Throughout, Pf denotes $\mathbb{E}_P[f]$. Here, $X_i \in \mathcal{X}$ denotes the covariate vector, $A_i \in \mathcal{A} = \{1, \ldots, K\}$ the discrete action, and $Y_i \in [\underline{y}, \overline{y}]$ the bounded observed outcome. Let $Y(a)$ denote the potential outcome under action a. We make the standard causal assumptions of consistency, $Y = Y(A)$, and unconfoundedness, $Y(a) \perp A \mid X$ for all $a \in \mathcal{A}$. The population outcome regression and propensity score are respectively defined as:

$$\mu_{0,a}(x) = \mathbb{E}[Y \mid A = a, X = x], \qquad e_{0,a}(x) = \mathbb{P}(A = a \mid X = x).$$

To ensure identification, we assume multi-action overlap: there exists a deterministic constant $\zeta \in (0, 1/K]$ such that $e_{0,a}(x) \geq \zeta$ for all a and x.

The target of the paper is pure-DP policy selection from a public, deterministic, and data-independent library $\Pi_n = \{\pi_1, \ldots, \pi_{M_n}\}$. Let Π be a target policy class and let $\Pi_n \subseteq \Pi$ be the finite release library, where each $\pi_j : \mathcal{X} \to \mathcal{A}$. The expected value of a policy π is defined as $V_P(\pi) = \mathbb{E}[Y(\pi(X))]$. Let $\pi_P^\star \in \arg\max_{\pi \in \Pi} V_P(\pi)$ denote the population oracle over Π, and let $\pi_n^\star \in \arg\max_{\pi \in \Pi_n} V_P(\pi)$ denote the optimal policy within the finite release library. Restricting the output space to the pre-fixed library Π_n decouples the release problem from candidate generation: OPE/OPL tools supply policy-value utilities for comparing the fixed candidates, while the object analyzed here is the privacy and regret of the mechanism that selects and releases one candidate.

The doubly robust score. To evaluate policies efficiently, we utilize the doubly robust (DR) score Chernozhukov et al. (2018). For a given nuisance pair $\eta = (\mu, e)$, the DR score for a record Z under policy π is defined as:

$$\psi_\pi(Z; \eta) = \mu_{\pi(X)}(X) + \frac{\mathbf{1}\{A = \pi(X)\}}{e_{\pi(X)}(X)}\{Y - \mu_{\pi(X)}(X)\}. \tag{1}$$

Assumption 1. *Throughout the privacy analysis, learned nuisance functions are deterministically projected, and the true nuisance functions lie in the same ranges. Thus, for every training sample S, action a, and covariate x,*

$$\underline{\mu} \leq \mu_{0,a}(x), \widehat{\mu}_{S,a}(x) \leq \overline{\mu}, \qquad e_0(x), \widehat{e}_S(x) \in \Delta_K^\zeta := \{p \in \Delta_K : \min_{1 \leq a \leq K} p_a \geq \zeta\}.$$

Under Assumption 1, let $B_\mu = \max\{|\underline{\mu}|, |\overline{\mu}|\}$ and $R_{Y,\mu} = \max\{|\overline{y} - \underline{\mu}|, |\underline{y} - \overline{\mu}|\}$. Then, for any nuisance pair satisfying the deterministic ranges, $|\psi_\pi(Z; \eta)| \leq B_\mu + \zeta^{-1} R_{Y,\mu} =: B_\psi$.

The statistical appeal of DR scores stems from the following identity:

$$P\psi_\pi(\eta) - V_P(\pi) = \sum_{a=1}^{K} \mathbb{E}\left[\mathbf{1}\{\pi(X) = a\}\left(\frac{e_a(X) - e_{0,a}(X)}{e_a(X)}\right)\{\mu_a(X) - \mu_{0,a}(X)\}\right]. \tag{2}$$

This yields the well-known product remainder bound in causal inference (doubly robust property):

$$|P\psi_\pi(\eta) - V_P(\pi)| \leq \zeta^{-1} \sum_{a=1}^{K} \|\mu_a - \mu_{0,a}\|_{L_2(P_X)} \|e_a - e_{0,a}\|_{L_2(P_X)}. \tag{3}$$

The identity (2) and the product-remainder bound (3) are proved in Lemma 3. We state them here because they are the statistical benchmark against which the privacy-sensitivity results below are contrasted.

Differential privacy and utility regimes. We say two split datasets $D = (D^{\mathrm{tr}}, D^{\mathrm{sc}})$ and $D' = (D^{\mathrm{tr}\prime}, D^{\mathrm{sc}\prime})$ are adjacent, denoted $D \sim D'$, if they differ by the replacement of a single record in either the training block or the scoring block. The split is public and index-based: the replaced record remains in the same block

under adjacency. A randomized mechanism $\mathcal{M}$ satisfies ε-differential privacy (ε-DP) if, for all adjacent datasets $D \sim D'$ and all measurable events S,

$$\mathbb{P}\{\mathcal{M}(D) \in S\} \leq e^{\varepsilon}\mathbb{P}\{\mathcal{M}(D') \in S\}.$$

To privately select a policy from Π_n, we use the exponential mechanism with utility $U(D, \pi)$. Given a calibration parameter Δ, the mechanism samples

$$\mathbb{P}(\widehat{\pi} = \pi \mid D) \propto \exp\left\{\frac{\varepsilon U(D,\pi)}{2\Delta}\right\}, \qquad \Delta \geq \sup_{D\sim D'} \sup_{\pi\in\Pi_n} |U(D,\pi) - U(D',\pi)|. \tag{4}$$

For causal-inference readers, (4) is a randomized release rule, not a causal model or an estimating equation. Conditional on the observed dataset, each candidate policy receives a weight that increases with its empirical utility $U(D, \pi)$; the normalizing constant over Π_n turns these weights into probabilities. The scale Δ must be a deterministic upper bound on the largest possible change of any policy utility under a one-record replacement. The factor $\varepsilon/(2\Delta)$ is the standard exponential-mechanism calibration: larger utility gaps make high-utility policies more likely, while a larger sensitivity bound flattens the distribution to maintain pure differential privacy. Thus the privacy question reduces to identifying a valid global sensitivity bound for the realized utility. This is immediate when the score map is fixed before observing the private data, but becomes subtle when the score map itself is learned from protected records. To further distinguish, we consider the following three utility regimes in this paper:

$$U_{\mathrm{fix}}(D^{\mathrm{sc}}, \pi) = \sum_{i\in D^{\mathrm{sc}}} q_\pi(Z_i), \quad U_{\mathrm{fr}}(D^{\mathrm{sc}}, \pi) = \sum_{i\in D^{\mathrm{sc}}} \psi_\pi(Z_i; \eta^\dagger), \quad U_{\mathrm{ls}}(D, \pi) = \sum_{i\in D^{\mathrm{sc}}} \psi_\pi(Z_i; \widehat{\eta}(D^{\mathrm{tr}})).$$

Here U_{fix} is the fixed-score utility, U_{fr} is the frozen-nuisance DR utility, and U_{ls} is the internally learned DR utility. The specific details are as follows:

Regime I: Fixed-score utility U_{fix}. The score map $q_\pi : \mathcal{Z} \to \mathbb{R}$ is bounded and fixed before the private dataset is observed.

Regime II: Frozen-nuisance DR utility U_{fr}. The nuisance object $\eta^\dagger$ is fixed relative to the protected scoring/release block and is conditioned upon in the privacy analysis. For this regime only, the privacy statement concerns replacements in the scoring/release block alone: the two scoring blocks may differ by one record, while $\eta^\dagger$ is held fixed. Auxiliary records used to construct $\eta^\dagger$ lie outside this DP guarantee unless protected by a separate privacy analysis.

Regime III: Internally learned DR utility U_{ls}. The nuisance estimator $\widehat{\eta} = (\widehat{\mu}, \widehat{e})$ is learned from the protected training block. The DP guarantee is with respect to the full split dataset $D = (D^{\mathrm{tr}}, D^{\mathrm{sc}})$, so a neighboring change may occur in either block.

4 Sensitivity Spillover: Separation and Certification

4.1 Sensitivity Spillover Decomposition

We first isolate the sensitivity term that is absent from fixed-score selection. If the score map is fixed before accessing the protected dataset, then replacing one scoring record changes only one bounded summand. The same constant-sensitivity argument applies to frozen-nuisance DR scores after conditioning on the fitted nuisances, and Appendix A records this baseline.

Internally learned scores are different: a replacement in D^{tr} can change the fitted nuisances $\widehat{\eta}(D^{\mathrm{tr}}) = (\widehat{\mu}, \widehat{e})$, hence the score map used on the entire scoring block D^{sc}. This nuisance-induced movement is then accumulated over all n scoring records. We call this effect *sensitivity spillover*. For adjacent nuisance-fitting samples $S \sim S'$ of size m, define

$$\Gamma_{\mu,m} := \sup_{S\sim S'} \max_a \sup_x |\widehat{\mu}_{S,a}(x) - \widehat{\mu}_{S',a}(x)|, \qquad \Gamma_{e,m} := \sup_{S\sim S'} \max_a \sup_x |\widehat{e}_{S,a}(x) - \widehat{e}_{S',a}(x)|. \tag{5}$$

These are deterministic replace-one prediction movements, measuring uniform stability for the fitted prediction functions, so they are different from estimation errors $\|\widehat{\mu} - \mu_0\|_{L_2(P_X)}$ or $\|\widehat{e} - e_0\|_{L_2(P_X)}$.

Lemma 1. *Let $\rho_m := (1+\zeta^{-1})\Gamma_{\mu,m} + \zeta^{-2} R_{Y,\mu}\Gamma_{e,m}$. Under Assumption 1, the internally learned DR utility has global sensitivity*

$$\Delta_{\mathrm{ls}}^{\mathrm{glob}} := \sup_{D \sim D'} \sup_{\pi \in \Pi_n} |U_{\mathrm{ls}}(D, \pi) - U_{\mathrm{ls}}(D', \pi)| \leq \max\{2B_\psi,\, n\rho_m\}. \tag{6}$$

The proof of Lemma 1 is stated in Appendix B.

The two terms correspond to the two possible locations of the neighboring replacement. If it occurs in D^{sc}, only one bounded score changes, which gives the fixed-score term $2B_\psi$. If it occurs in D^{tr}, the scoring records are unchanged but the learned score map can move by ρ_m at each of the n scoring points, which gives the spillover term $n\rho_m$.

Thus bounded realized DR scores control individual summands but not the dataset-dependent movement of the score map. Pure-DP calibration for learned DR utilities therefore requires a deterministic replace-one stability certificate for the nuisance predictions. The next subsection shows that this certificate is not implied by double robustness, nuisance accuracy, or orthogonality alone.

4.2 Double Robustness Does Not Certify Privacy

Lemma 1 shows that learned-score privacy requires a deterministic replace-one stability bound for the fitted nuisances. Such a bound is not implied by standard doubly robust guarantees. The product remainder in (3) controls population-average bias, whereas pure DP requires worst-case control of the realized utility over all adjacent datasets.

The next theorem makes this gap explicit. It constructs a setting in which the usual statistical diagnostics for DR estimation are ideal: one nuisance is exactly correct, the other has vanishing $L_2(P_X)$ error, and the DR population bias is zero. Nevertheless, a single replacement in the nuisance-training block can change the learned score map on a region that is negligible under P_X but can be populated by the scoring block in a worst-case adjacent pair. As a result, the realized learned-score utility can move by n, the full size of the scoring block. Hence the global sensitivity relevant for pure DP can be order-n even when the DR statistical error is zero.

Theorem 1. *There exist a fixed population distribution P, a fixed policy π, and deterministic nuisance learners $\{\widehat{\eta}_N\}_{N\geq 1}$ such that, for every split $N = m + n$, the propensity is exactly correct, the outcome nuisance satisfies*

$$\max_a \|\widehat{\mu}_{N,S,a} - \mu_{0,a}\|_{L_2(P_X)} \leq N^{-2}$$

uniformly over training samples S, and the DR population bias is zero: $P\psi_\pi(\widehat{\eta}_N(S)) = V_P(\pi)$. Nevertheless, there exist adjacent datasets $D \sim D'$ with the same scoring block for which

$$|U_{\mathrm{ls}}(D, \pi) - U_{\mathrm{ls}}(D', \pi)| = n. \tag{7}$$

Consequently, the learned-score utility has worst-case sensitivity at least n for this instance. Moreover, for any fixed calibration $\Delta_0 > 0$ and privacy budget $\varepsilon > 0$, there is a two-policy version with bounded realized scores, correct propensities, vanishing outcome error, and zero DR population bias for which the exponential mechanism calibrated at Δ_0 fails ε-DP whenever $n > 2\Delta_0$.

The proof of Theorem 1 is stated in Appendix C. The construction toggles predictions on a covariate region with vanishing P_X mass. This perturbation is negligible for $L_2(P_X)$ nuisance error and for the DR product remainder, but it can be hit by a worst-case scoring block. Since the learned nuisance is then evaluated on all n scoring records, the single training-record replacement is amplified into an order-n realized utility movement. This is precisely the sensitivity spillover effect: an individually small change in the training block becomes a collective change in the scoring block.

Thus the theorem shows that orthogonality, nuisance accuracy, zero DR population bias, and bounded realized scores do not certify privacy sensitivity. The failure is that pure-DP calibration depends on worst-case

adjacent-dataset movement of the realized utility. Learned-score private selection therefore needs an explicit stability certificate controlling how much the fitted nuisance predictions can change under one replacement. Appendix C gives both the construction and the two-policy likelihood-ratio proof, and Appendix F gives a nearest-neighbor spillover example without a sample-dependent rare-region learner.

4.3 Certified Private Selection via Prediction Stability

The previous subsection gives a negative message, but it also identifies the right positive target. The exponential mechanism cannot be calibrated from a population bias bound. Before release, it needs a deterministic upper bound on how much the realized utility can move under any one-record replacement. For learned DR utilities, Lemma 1 shows that this movement has two channels: a scoring-record replacement contributes the usual bounded-summand term $2B_\psi$, whereas a nuisance-fitting replacement contributes the spillover term accumulated across the scoring block. Therefore the missing ingredient is not another DR accuracy condition, but a certificate that the nuisance learner has small replace-one prediction movement.

We use the following deterministic certificate. Assume the nuisance learner is deterministic, with deterministic tie-breaking, and that any public randomness is fixed before observing the private data. For every adjacent pair of fitting samples $S \sim S'$ of size m, suppose that

$$\max_a \sup_x |\widehat{\mu}_{S,a}(x) - \widehat{\mu}_{S',a}(x)| \le \beta_{\mu,m}, \qquad \max_a \sup_x |\widehat{e}_{S,a}(x) - \widehat{e}_{S',a}(x)| \le \beta_{e,m}. \tag{8}$$

The constants in (8) are public, deterministic upper bounds that hold uniformly over neighboring fitting samples. They are not average-case nuisance errors and are not estimated from the private scoring block. Their role is to certify the algorithmic stability of the learned score map that will be evaluated on the scoring records.

Define

$$\rho_m^\star := (1+\zeta^{-1})\beta_{\mu,m} + \zeta^{-2} R_{Y,\mu}\beta_{e,m}, \qquad \Delta_{\mathrm{cert}} := \max\{2B_\psi,\, n\rho_m^\star\}. \tag{9}$$

By Lemma 1, this Δ_{cert} is a valid global-sensitivity bound for the internally learned DR utility. The certified selection mechanism therefore fits $\widehat{\eta}$ on D^{tr}, evaluates $U_{\mathrm{ls}}(D,\pi)$ on D^{sc}, and releases

$$\widehat{\pi} \sim \Pr(\pi) \propto \exp\left\{\frac{\varepsilon U_{\mathrm{ls}}(D,\pi)}{2\Delta_{\mathrm{cert}}}\right\}, \qquad \pi \in \Pi_n.$$

This construction is the promised bridge between learned causal scores and pure DP: orthogonality will control the statistical error of $P\psi_\pi(\widehat{\eta})$, while the public prediction-stability certificate controls the privacy calibration of the realized utility. We now combine these two ingredients into a regret bound. Let $M_n = |\Pi_n|$ and

$$\mathcal{A}_n(P;\Pi,\Pi_n) := \sup_{\pi\in\Pi} V_P(\pi) - \max_{\pi\in\Pi_n} V_P(\pi),$$

the approximation error of the public finite library. The result is stated in the following theorem.

Theorem 2. *Under the causal identification, bounded-range, public-split, data-independent-library, and stability-certificate conditions above, the certified selection mechanism is pure ε-DP for all users in $D^{\mathrm{tr}} \cup D^{\mathrm{sc}}$. Moreover, suppose that with probability at least $1-\alpha_\eta$ over D^{tr},*

$$\max_a \|\widehat{\mu}_a - \mu_{0,a}\|_{L_2(P_X)} \le r_{\mu,m}, \qquad \max_a \|\widehat{e}_a - e_{0,a}\|_{L_2(P_X)} \le r_{e,m}. \tag{10}$$

Then, with probability at least $1-\alpha_\eta-\alpha-\beta$,

$$\begin{aligned} V_P(\pi_P^\star) - V_P(\widehat{\pi}) \le{}& \mathcal{A}_n(P;\Pi,\Pi_n) + CB_\psi\sqrt{\frac{\log(2M_n/\alpha)}{n}} + C'K\zeta^{-1} r_{\mu,m} r_{e,m} \\ &+ \frac{C''\Delta_{\mathrm{cert}}}{n\varepsilon}\{\log M_n + \log(1/\beta)\}, \end{aligned} \tag{11}$$

where C, C', C'' are universal constants.

The proof of Theorem 2 is stated in Appendix D. Theorem 2 separates the regret into four conceptually distinct terms: finite-library approximation, scoring-sample concentration, the DR product remainder, and the exponential-mechanism price. Only the last term depends on the sensitivity calibration. Since

$$\frac{\Delta_{\mathrm{cert}}}{n\varepsilon} = \max\left\{\frac{2B_\psi}{n\varepsilon}, \frac{\rho_m^\star}{\varepsilon}\right\}, \tag{12}$$

the fixed-score component has the familiar $1/(n\varepsilon)$ scaling, whereas the learned-score component is controlled by the certified spillover radius $\rho_m^\star$. Thus privacy depends on replace-one prediction stability, not on the DR product remainder: orthogonality governs the statistical term $r_{\mu,m}r_{e,m}$, while certification governs the privacy term $\rho_m^\star/\varepsilon$.

This decomposition also makes clear when the learned-score selection mechanism is consistent. The approximation, concentration, DR, and privacy terms must all vanish, and the last requirement is precisely a stability requirement on the nuisance learner.

Corollary 1. *Fix $\varepsilon > 0$. If $\mathcal{A}_n(P;\Pi,\Pi_n) \to 0$, $\log(M_n)/n \to 0$, $Kr_{\mu,m}r_{e,m} \to 0$, and $\frac{\rho_m^\star}{\varepsilon}\{\log M_n + \log(1/\beta)\} \to 0$, then the regret bound in Theorem 2 vanishes for fixed failure probabilities. In particular, if a certifiable strongly convex regularized ERM learner gives $\rho_m^\star = O((m\lambda_m)^{-1})$, then the spillover privacy term vanishes whenever $\frac{\log M_n + \log(1/\beta)}{\varepsilon m\lambda_m} \to 0$, while the DR statistical term requires $r_{\mu,m}r_{e,m} \to 0$.*

The proof of Corollary 1 is stated in Appendix D.

It remains to say when such public stability certificates are available. For some deterministic learning algorithms, they follow directly from the optimization problem defining the nuisance learner.

Remark 1. *For deterministic finite-dimensional strongly convex regularized ERM nuisance learners with bounded features, globally G-Lipschitz losses, L_g-Lipschitz prediction maps, deterministic tie-breaking, and deterministic clipping or simplex projection,*

$$\sup_x |g_{\widehat{\theta}_S}(x) - g_{\widehat{\theta}_{S'}}(x)| \le \frac{2L_gG}{m\lambda_m} \qquad \text{for every } S \sim S'. \tag{13}$$

Hence $\beta_{\mu,m}$ and $\beta_{e,m}$ are explicit for this class, giving $\rho_m^\star = O((m\lambda_m)^{-1})$. Appendix E gives the multi-action statement and proof. More adaptive learners require their own deterministic stability analysis before the pure-DP guarantee applies.

5 Experiments

This section investigates four questions: (1) Can learned nuisance fitting create large adjacent changes in the learned DR utility while DR accuracy diagnostics remain favorable? (2) Does the fixed-score calibration become too small on the resulting likelihood-ratio audit? (3) Do regularization and sample splitting change the empirical movement and regret in the directions predicted by the spillover and regret decompositions? (4) Can the sufficient certification route be instantiated by replacing the flexible learner with a deterministic strongly convex ERM learner whose stability constants are public?

We use flexible boosted-tree and logistic-regression nuisance learners for the stress tests, and a separate deterministic regularized ERM learner as a positive-control certificate. We report nuisance accuracy, DR product diagnostics, empirical replace-one movement, likelihood-ratio audits, regret, and, where applicable, public certificate scales.

5.1 Experimental Setup

Data and semi-synthetic design. A central challenge in empirical causal inference is that counterfactual outcomes are unobserved. Following established semi-synthetic practice (Curth & Van der Schaar, 2021; Curth & Van Der Schaar, 2023; Huang et al., 2024), we use covariates from the ACIC 2016 benchmark (Dorie et al., 2019) and generate treatments and outcomes from known response surfaces. After preprocessing, the design matrix has 4,802 rows and 58 encoded covariates.

Table 1: Separation and privacy audit over 100 repetitions. Appendix G.2 states column definitions.

n	nuis. L_2	DR prod.	unstable move/floor	naive logLR/ε	stable logLR/ε	stable cover
300	0.0950	0.00963	0.37	0.13	0.053	1.00
700	0.0946	0.00958	0.87	0.37	0.148	1.00
1200	0.0952	0.00968	1.49	0.69	0.221	1.00
2000	0.0961	0.00986	2.40	1.17	0.319	1.00
3000	0.0945	0.00955	3.36	1.48	0.336	1.00

The main separation audit uses a rare-region DGP, a finite-sample analogue of Theorem 1. The purpose is to create a low-probability covariate region that is nearly invisible to population-average nuisance error but can still induce large realized movement on stress scoring blocks. Let $R(x)$ indicate that a nonlinear rare-region score exceeds its empirical 95th percentile. We define

$$\begin{aligned} e_0(x) &= \mathrm{clip}_{[\zeta,1-\zeta]}\{\sigma(g_e(x) + 1.7R(x))\}, \qquad \zeta = 0.10, \\ \tau(x) &= 0.20\tanh\{1.4g_\tau(x)\} + 0.11 \cdot \mathbf{1}\{x_1 > 0\} - 0.07 \cdot \mathbf{1}\{x_6 > 0\} + 0.45R(x), \\ \mu_{0,0}(x) &= \mathrm{clip}_{[0.02,0.98]}\{0.25 + 0.45\sigma(g_0(x))\}, \;\; \mu_{0,1}(x) = \mathrm{clip}_{[0.02,0.98]}\{\mu_{0,0}(x) + \tau(x)\}, \end{aligned}$$

where g_e, g_0, g_τ are fixed sparse functions with nonlinear terms. We sample $A \sim \mathrm{Bernoulli}(e_0(X))$ and $Y = \mathrm{clip}_{[0,1]}\{\mu_{0,A}(X) + \xi\}$ with Gaussian noise ξ. The rare region affects both treatment assignment and treatment effects, while clipping enforces the overlap and bounded-outcome conditions used in the theory.

Policy library, learners, and diagnostics. The public policy library contains 160 deterministic policies: structured threshold and rare-region rules, together with random linear threshold rules generated from public randomness. Each repetition uses $m = 1500$ nuisance-fitting records and $n = 1500$ policy-scoring records unless stated otherwise. True policy values are approximated by Monte Carlo evaluation of the known conditional means, and regret for randomized selection mechanisms is reported in expectation under the selection distribution. Propensities are fit by logistic regression and outcome regressions by fixed-seed boosted trees. We compare a more regularized stable variant with a deeper, weakly regularized stress variant. Under the imposed overlap $\zeta = 0.10$, clipped DR scores satisfy $B_\psi = 1 + 1/\zeta = 11$, so the fixed-score sensitivity floor is $2B_\psi = 22$. For adjacent datasets D, D', we audit learned-utility movement by $\widehat{\Delta}_{\mathrm{move}} := \max_{\pi\in\Pi_n} |U_{\mathrm{ls}}(D,\pi) - U_{\mathrm{ls}}(D',\pi)|$. We also report the DR product proxy $\widehat{r}_{\mathrm{DR}} := \frac{1}{2}\{\mathrm{RMSE}(\widehat{\mu}_0) + \mathrm{RMSE}(\widehat{\mu}_1)\}\mathrm{RMSE}(\widehat{e})$, and, in the regularization and split experiments, the prediction-movement proxy $\widehat{\rho}_{\mathrm{pred}}$ from Lemma 1.

Appendix G.1 gives more specific details of preprocessing, DGP, stress-audit construction, policy construction, learner hyperparameters, experiment grid, and remaining diagnostics.

5.2 Results and Analysis

DR accuracy does not imply learned-score stability. We first construct adjacent nuisance-fitting samples that differ in one rare-region anchor record and evaluate the resulting learned utilities on rare-region scoring blocks of increasing size. Table 1 reports averages over 100 repetitions. The first two diagnostic columns behave like ordinary statistical quantities: nuisance L_2 error and the DR product proxy are nearly unchanged as n grows. The utility movement behaves differently, increasing from 0.37 to 3.36 times the fixed-score floor $2B_\psi$. This is the empirical signature of spillover: one changed fitting record perturbs the learned score map, and the perturbation is accumulated over the scoring block. This matches the $n\rho_m$ term in Lemma 1 and the separation in Theorem 1.

The privacy diagnostic shows the consequence of this movement, as visualized in Figure 1. The fixed-score scale $\Delta = 2B_\psi$ would be appropriate if the score map were fixed or frozen, but it becomes too small for the unstable learned-score utility once the observed adjacent movement exceeds that floor. The left panel shows this crossing directly, while the middle panel shows that the audit-calibrated scale for the stable learner stays above the observed movement on the sampled adjacent pairs. The right panel reports the corresponding empirical max log-likelihood ratio divided by ε: at $n = 2000$ and $n = 3000$, naive learned-score exponential-mechanism calibration reaches 1.17 and 1.48, whereas the stability-calibrated audit stays below the target

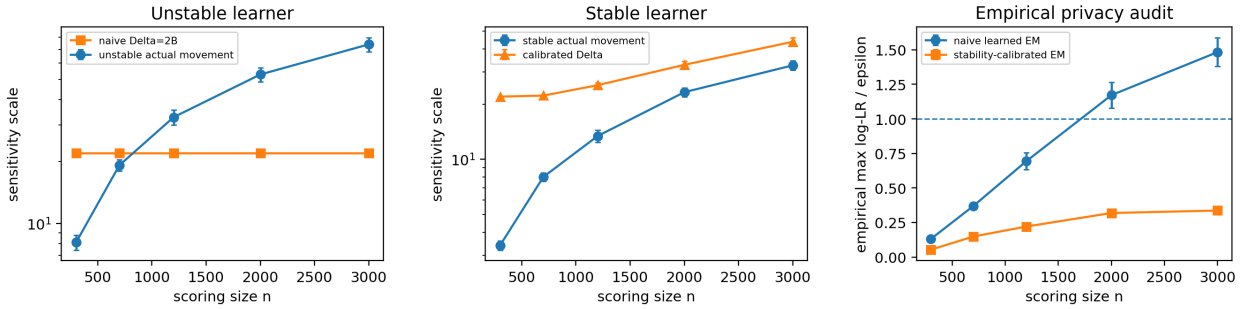

Figure 1: Empirical sensitivity and privacy-loss audits across scoring sizes. Left: learned-score (ls) movement for the unstable learner versus the fixed-score floor $\Delta = 2B_\psi$. Middle: stable learner's movement versus the stability-audit scale. Right: empirical max log-likelihood ratio divided by ε; naive learned-score exponential mechanism uses the fixed-score floor, while stability-calibrated exponential mechanism uses the audit scale. Values above one fail the audited pure-DP likelihood-ratio target.

Table 2: Regularization diagnostic over 100 repetitions. Appendix G.2 states column definitions.

λ	$\widehat{\rho}_{\text{pred}}$	move	audit Δ	audit/floor	floored Δ	floor rate	regret
0.01	8.87	40.18	52.23	2.37	54.70	0.23	0.0211
0.03	8.84	39.72	51.63	2.35	54.13	0.26	0.0210
0.10	8.72	39.27	51.05	2.32	53.55	0.23	0.0209
0.30	8.45	37.39	48.60	2.21	51.11	0.24	0.0207
1.00	7.60	33.94	44.12	2.01	46.55	0.25	0.0201
3.00	4.63	25.41	33.04	1.50	36.15	0.41	0.0185
10.0	2.37	17.54	22.80	1.04	27.76	0.57	0.0167
30.0	1.39	12.94	16.82	0.76	23.90	0.77	0.0157
100	0.48	5.15	6.70	0.30	22.00	1.00	0.0149

line at one. This demonstrates that DR accuracy and privacy sensitivity measure different objects, exactly as motivated by the separation between the product remainder and the spillover term.

Regularization reduces prediction movement and spillover. Theorem 2 identifies deterministic prediction stability as the interface between learned DR utilities and certified private selection. Since XGBoost is outside the strongly convex ERM class certified in Appendix E, this experiment studies the corresponding empirical stability proxy. We hold the adjacent pair, scoring set, and policy library fixed while sweeping the XGBoost regularization parameter λ. Table 2 shows the intended monotone pattern: as λ increases, the prediction-movement proxy falls from 8.87 to 0.48, and the observed learned-utility movement falls from 40.18 to 5.15. The audit scale after applying the fixed-score floor becomes floor-dominated at $\lambda = 100$. Thus regularization targets the empirical prediction movement component of the spillover decomposition. Appendix G.3 reports the corresponding theorem-proxy scale and clarifies the gap between sampled audit diagnostics and theorem-level deterministic certificates. To close this gap constructively, Appendix G.4 replaces the boosted-tree nuisances with a deterministic strongly convex ERM learner, evaluates the resulting public certificate, and verifies that the certified scale covers all audited adjacent movements while retaining small regret.

Sample-splitting decomposition. Holding the public policy library fixed, the split experiment isolates three sample-dependent terms in the regret bound: nuisance quality, scoring-sample concentration, and the learned-score movement component. We fix the total sample size at 3,600 and vary the nuisance-fitting fraction. Table 3 shows the resulting trade-off. Increasing the fitting fraction from 0.20 to 0.80 improves nuisance RMSE from 0.0838 to 0.0608 and reduces observed movement from 2.61 to 0.51, reflecting the stability side of the bound. At the same time, the scoring block shrinks from $2,880$ to 720 records, $1/\sqrt{n}$ increases, and empirical expected regret rises from 0.0232 to 0.0326, reflecting the concentration side of the

Table 3: Sample-split decomposition over 100 repetitions. Appendix G.2 states column definitions.

fit frac.	m	n	nuis. RMSE	$\widehat{\rho}_{\text{pred}}$	move	floored Δ	floor rate	$1/\sqrt{n}$	regret
0.20	720	2880	0.0838	9.04	2.61	22.00	1.00	0.0186	0.0232
0.25	900	2700	0.0785	6.61	2.12	22.00	1.00	0.0192	0.0242
0.35	1260	2340	0.0725	4.61	1.41	22.00	1.00	0.0207	0.0258
0.50	1800	1800	0.0664	2.93	1.01	22.00	1.00	0.0236	0.0284
0.65	2340	1260	0.0631	2.69	0.73	22.00	1.00	0.0282	0.0304
0.75	2700	900	0.0615	1.89	0.56	22.00	1.00	0.0333	0.0320
0.80	2880	720	0.0608	1.80	0.51	22.00	1.00	0.0373	0.0326

bound. The audit scale after applying the fixed-score floor remains 22 for every split because the observed spillover movements are below $2B_\psi = 22$.

In summary, the separation audit shows that statistical DR diagnostics can remain stable while learned-score utility movement grows and fixed-score calibration fails on audited adjacent pairs; the regularization audit shows that regularization targets the empirical prediction-movement component of the spillover decomposition; the sample-split experiment illustrates the trade-off between nuisance accuracy, scoring concentration, and empirical movement. Finally, the certified ERM positive control in Appendix G.4 instantiates the sufficient condition from Section 4.3 with a public deterministic certificate instead of a sampled audit proxy.

6 Conclusion

This paper shows that double robustness, a well-known statistical robustness property in causal inference, is not a privacy certificate for private policy selection: it controls population-level nuisance error, whereas pure DP requires worst-case control of the realized utility under one-record replacement. This distinction leads to sensitivity spillover, where a replacement in the nuisance-training block changes the learned score map and this change is evaluated across all scoring records. Our separation result shows that vanishing nuisance error and even zero DR population bias can coexist with order-n realized utility sensitivity, invalidating the usual fixed-utility calibration of the exponential mechanism. We give a sufficient certification route based on deterministic replace-one prediction stability of the nuisance learners, yielding a valid pure-DP exponential mechanism and a regret decomposition that separates library approximation, concentration, the DR product remainder, and certified privacy cost. The experiments support the same message: statistical DR diagnostics and privacy-relevant adjacent-dataset movement can behave very differently, while stability-oriented regularization can control the spillover term. Our investigations suggest a two-certificate principle for private causal policy selection: orthogonality is needed for robustness to global nuisance estimation error, and algorithmic stability is needed for robustness to individual replacement.

The current study also has several limitations. It applies to selection from a finite, public, data-independent policy library under a public split, and does not by itself cover data-dependent library construction, end-to-end private policy learning, or adaptive policy-class search. In addition, the deterministic uniform stability certificate is also a sufficient condition rather than a universal recipe, and obtaining tight certificates for forests, boosting methods, or neural networks remains important future challenges. Therefore, we hope our results might bring some insights, encouraging future studies to develop auditable stability analyses for richer nuisance learners, combine causal orthogonality with local or smooth sensitivity under approximate-DP guarantees, and extend the framework to large or data-dependent policy classes.

References

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

Raef Bassily, Adam Smith, and Abhradeep Thakurta. Private empirical risk minimization: Efficient algorithms and tight error bounds. In *2014 IEEE 55th annual symposium on foundations of computer science*, pp. 464–473. IEEE, 2014.

Olivier Bousquet and André Elisseeff. Stability and generalization. *Journal of machine learning research*, 2 (Mar):499–526, 2002.

Kamalika Chaudhuri, Claire Monteleoni, and Anand D Sarwate. Differentially private empirical risk minimization. *Journal of Machine Learning Research*, 12(3), 2011.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters, 2018.

Alicia Curth and Mihaela Van der Schaar. On inductive biases for heterogeneous treatment effect estimation. *Advances in Neural Information Processing Systems*, 34:15883–15894, 2021.

Alicia Curth and Mihaela Van Der Schaar. In search of insights, not magic bullets: Towards demystification of the model selection dilemma in heterogeneous treatment effect estimation. In *International conference on machine learning*, pp. 6623–6642. PMLR, 2023.

Vito D'Orazio, James Honaker, and Gary King. Differential privacy for social science inference. *Sloan Foundation Economics Research Paper*, (2676160), 2015.

Vincent Dorie, Jennifer Hill, Uri Shalit, Marc Scott, and Dan Cervone. Automated versus do-it-yourself methods for causal inference: Lessons learned from a data analysis competition. *Statistical Science*, 34(1): 43–68, 2019.

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

Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Foundations and trends® in theoretical computer science*, 9(3-4):211–487, 2014.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference*, pp. 265–284. Springer, 2006.

Amirhossein Farzam and Guillermo Sapiro. Causal inference under differential privacy: Challenges and mitigation strategies. In *NeurIPS 2024 Causal Representation Learning Workshop*.

Evrard Garcelon, Kamalika Chaudhuri, Vianney Perchet, and Matteo Pirotta. Privacy amplification via shuffling for linear contextual bandits. In *International Conference on Algorithmic Learning Theory*, pp. 381–407. PMLR, 2022.

Sharmistha Guha and Jerome P Reiter. Differentially private estimation of weighted average treatment effects for binary outcomes. *Computational Statistics & Data Analysis*, 207:108145, 2025.

Ruiquan Huang, Huanyu Zhang, Luca Melis, Milan Shen, Meisam Hejazinia, and Jing Yang. Federated linear contextual bandits with user-level differential privacy. In *International Conference on Machine Learning*, pp. 14060–14095. PMLR, 2023.

Yiyan Huang, Cheuk H Leung, Siyi Wang, Yijun Li, and Qi Wu. Unveiling the potential of robustness in selecting conditional average treatment effect estimators. *Advances in Neural Information Processing Systems*, 37:135208–135243, 2024.

Nathan Kallus. Recursive partitioning for personalization using observational data. In *International conference on machine learning*, pp. 1789–1798. PMLR, 2017.

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

Nathan Kallus, Xiaojie Mao, Kaiwen Wang, and Zhengyuan Zhou. Doubly robust distributionally robust off-policy evaluation and learning. In *International Conference on Machine Learning*, pp. 10598–10632. PMLR, 2022.

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

Christian Janos Lebeda, Mathieu Even, Aurélien Bellet, and Julie Josse. Model agnostic differentially private causal inference. *arXiv preprint arXiv:2505.19589*, 2025.

Si Kai Lee, Luigi Gresele, Mijung Park, and Krikamol Muandet. Privacy-preserving causal inference via inverse probability weighting. *arXiv preprint arXiv:1905.12592*, 2019.

Chonghua Liao, Jiafan He, and Quanquan Gu. Locally differentially private reinforcement learning for linear mixture markov decision processes. In *Asian Conference on Machine Learning*, pp. 627–642. PMLR, 2023.

Frank McSherry and Kunal Talwar. Mechanism design via differential privacy. In *48th Annual IEEE Symposium on Foundations of Computer Science (FOCS'07)*, pp. 94–103. IEEE, 2007.

Susan A Murphy. Optimal dynamic treatment regimes. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 65(2):331–355, 2003.

Kobbi Nissim, Sofya Raskhodnikova, and Adam Smith. Smooth sensitivity and sampling in private data analysis. In *Proceedings of the thirty-ninth annual ACM symposium on Theory of computing*, pp. 75–84, 2007.

Fengshi Niu, Harsha Nori, Brian Quistorff, Rich Caruana, Donald Ngwe, and Aadharsh Kannan. Differentially private estimation of heterogeneous causal effects. In *Conference on Causal Learning and Reasoning*, pp. 618–633. PMLR, 2022.

Yuki Ohnishi and Jordan Awan. Differentially private covariate balancing causal inference. *arXiv preprint arXiv:2410.14789*, 2024.

Yuki Ohnishi and Jordan Awan. Locally private causal inference for randomized experiments. *Journal of Machine Learning Research*, 26(14):1–40, 2025.

Min Qian and Susan A Murphy. Performance guarantees for individualized treatment rules. *Annals of statistics*, 39(2):1180, 2011.

Dan Qiao and Yu-Xiang Wang. Near-optimal differentially private reinforcement learning. In *International Conference on Artificial Intelligence and Statistics*, pp. 9914–9940. PMLR, 2023a.

Dan Qiao and Yu-Xiang Wang. Offline reinforcement learning with differential privacy. *Advances in neural information processing systems*, 36:61395–61436, 2023b.

Maresa Schröder, Justin Hartenstein, and Stefan Feuerriegel. Private: Differentially private confidence intervals for average treatment effects. *arXiv preprint arXiv:2505.21641*, 2025a.

Maresa Schröder, Valentyn Melnychuk, and Stefan Feuerriegel. Differentially private learners for heterogeneous treatment effects. *arXiv preprint arXiv:2503.03486*, 2025b.

Shai Shalev-Shwartz, Ohad Shamir, Nathan Srebro, and Karthik Sridharan. Learnability, stability and uniform convergence. *The Journal of Machine Learning Research*, 11:2635–2670, 2010.

Roshan Shariff and Or Sheffet. Differentially private contextual linear bandits. *Advances in Neural Information Processing Systems*, 31, 2018.

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

Jingyuan Wang, Zhimei Ren, Ruohan Zhan, and Zhengyuan Zhou. Distributionally robust policy learning under concept drifts. In *International Conference on Machine Learning*, pp. 64244–64281. PMLR, 2025.

Siwei Wang and Jun Zhu. Optimal learning policies for differential privacy in multi-armed bandits. *Journal of Machine Learning Research*, 25(314):1–52, 2024.

Quan Yuan, Xiaochen Li, Linkang Du, Min Chen, Mingyang Sun, Yunjun Gao, Shibo He, Jiming Chen, and Zhikun Zhang. Private: Differentially private average treatment effect estimation for observational data. *arXiv preprint arXiv:2512.14557*, 2025.

Ying-Qi Zhao, Eric B Laber, Yang Ning, Sumona Saha, and Bruce E Sands. Efficient augmentation and relaxation learning for individualized treatment rules using observational data. *Journal of Machine Learning Research*, 20(48):1–23, 2019.

Yingqi Zhao, Donglin Zeng, A John Rush, and Michael R Kosorok. Estimating individualized treatment rules using outcome weighted learning. *Journal of the American Statistical Association*, 107(499):1106–1118, 2012.

Kai Zheng, Tianle Cai, Weiran Huang, Zhenguo Li, and Liwei Wang. Locally differentially private (contextual) bandits learning. *Advances in Neural Information Processing Systems*, 33:12300–12310, 2020.

Ke Zhu, Jianing Chu, Ilya Lipkovich, Wenyu Ye, and Shu Yang. Doubly robust fusion of many treatments for policy learning. In *International Conference on Machine Learning*, pp. 79772–79789. PMLR, 2025.

Appendix

A Fixed and Frozen Utilities

Proposition 1. *If $|q_\pi(z)| \le B$ for all π, z, then $\Delta_{\mathrm{fix}} \le 2B$. If $\eta^\dagger$ is fixed relative to the protected scoring block and $|\psi_\pi(z;\eta^\dagger)| \le B_\psi$ for all π, z, then under scoring-block adjacency with $\eta^\dagger$ held fixed, $\Delta_{\mathrm{fr}} \le 2B_\psi$.*

The proof of Proposition 1 is stated below.

Proof of Proposition 1. Fix an arbitrary policy $\pi \in \Pi_n$. If $D \sim D'$ differ in one scoring record, then all summands except the replaced record coincide. Writing the replaced records as Z and Z', we have

$$|U_{\mathrm{fix}}(D,\pi) - U_{\mathrm{fix}}(D',\pi)| = |q_\pi(Z) - q_\pi(Z')| \le 2B.$$

If the replacement occurs outside the scoring block and the score map is fixed, every summand is identical and the utility change is zero. Taking the supremum over π proves the fixed-score bound.

For frozen DR scores, condition on the fixed value of $\eta^\dagger$. Under the stated adjacency relation, replacing a protected user does not alter $\eta^\dagger$, so the function $z \mapsto \psi_\pi(z;\eta^\dagger)$ is a deterministic bounded score map. If the replacement is in the scoring/release block, the same one-summand calculation gives

$$|U_{\mathrm{fr}}(D,\pi) - U_{\mathrm{fr}}(D',\pi)| \le 2B_\psi.$$

If the replacement is outside the scoring/release block and leaves $\eta^\dagger$ fixed, the frozen utility is unchanged. Taking the supremum over π gives $\Delta_{\mathrm{fr}} \le 2B_\psi$. The proof therefore protects only records whose replacement is covered by this adjacency relation. Auxiliary records used to build $\eta^\dagger$ are not protected by this argument unless their data dependence is separately privatized or included in the sensitivity calculation. □

B DR Score Lipschitz Bound and Sensitivity Decomposition

Lemma 2. *Let $\eta = (\mu, e)$ and $\eta' = (\mu', e')$ satisfy the deterministic ranges. If*

$$\max_a \sup_x |\mu_a(x) - \mu'_a(x)| \le \delta_\mu, \qquad \max_a \sup_x |e_a(x) - e'_a(x)| \le \delta_e,$$

then for every z and π,

$$|\psi_\pi(z;\eta) - \psi_\pi(z;\eta')| \le (1+\zeta^{-1})\delta_\mu + \zeta^{-2} R_{Y,\mu}\delta_e.$$

The proof of Lemma 2 is stated below.

Proof of Lemma 2. Write $z = (x, A, Y)$, set $a = \pi(x)$, and let $I = \mathbf{1}\{A = a\}$. From (1),

$$\psi_\pi(z;\eta) - \psi_\pi(z;\eta') = \{\mu_a(x) - \mu'_a(x)\} + I\left\{\frac{Y - \mu_a(x)}{e_a(x)} - \frac{Y - \mu'_a(x)}{e'_a(x)}\right\}.$$

For the inverse-propensity part, add and subtract $(Y - \mu'_a(x))/e_a(x)$ to obtain

$$\left|\frac{Y - \mu_a(x)}{e_a(x)} - \frac{Y - \mu'_a(x)}{e'_a(x)}\right| \le \frac{|\mu_a(x) - \mu'_a(x)|}{e_a(x)} + |Y - \mu'_a(x)|\left|\frac{1}{e_a(x)} - \frac{1}{e'_a(x)}\right|.$$

The deterministic range assumptions give $e_a(x), e'_a(x) \ge \zeta$ and $|\mu_a(x) - \mu'_a(x)| \le \delta_\mu$. They also give $|Y - \mu'_a(x)| \le R_{Y,\mu}$: since $Y \in [\underline{y}, \overline{y}]$ and $\mu'_a(x) \in [\underline{\mu}, \overline{\mu}]$, the maximum of $|y - u|$ over this rectangle is $\max\{|\overline{y} - \underline{\mu}|, |\underline{y} - \overline{\mu}|\}$. Finally,

$$\left|\frac{1}{e_a(x)} - \frac{1}{e'_a(x)}\right| = \frac{|e_a(x) - e'_a(x)|}{e_a(x)e'_a(x)} \le \zeta^{-2}\delta_e.$$

Combining these bounds and using $I \le 1$ yields

$$\begin{aligned}|\psi_\pi(z;\eta) - \psi_\pi(z;\eta')| &\le \delta_\mu + \left\{\zeta^{-1}\delta_\mu + R_{Y,\mu}\zeta^{-2}\delta_e\right\} \\ &= (1+\zeta^{-1})\delta_\mu + \zeta^{-2}R_{Y,\mu}\delta_e.\end{aligned}$$

□

Proof of Lemma 1. If adjacent datasets differ in one scoring record, D^{tr} is unchanged. Hence $\widehat{\eta}(D^{\mathrm{tr}})$ is fixed and only one bounded score changes, giving

$$|U_{\mathrm{ls}}(D,\pi) - U_{\mathrm{ls}}(D',\pi)| \le 2B_\psi.$$

If they differ in one training record, write the adjacent training samples as $S \sim S'$. The scoring block is unchanged, while the nuisance changes from $\widehat{\eta}_S$ to $\widehat{\eta}_{S'}$. For every fixed π,

$$\begin{aligned}&|U_{\mathrm{ls}}(D,\pi) - U_{\mathrm{ls}}(D',\pi)| \\ &\quad\le \sum_{i\in D^{\mathrm{sc}}} |\psi_\pi\{Z_i;\widehat{\eta}_S\} - \psi_\pi\{Z_i;\widehat{\eta}_{S'}\}| \\ &\quad\le n\left\{(1+\zeta^{-1})\Gamma_{\mu,m} + \zeta^{-2}R_{Y,\mu}\Gamma_{e,m}\right\} = n\rho_m,\end{aligned}$$

where the second inequality is Lemma 2 together with the definitions of $\Gamma_{\mu,m}$ and $\Gamma_{e,m}$. Taking the supremum over π and the maximum over the two possible replacement locations proves the result. □

C Proof of Theorem 1: Separation and Naive Calibration Violation

This appendix proves the separation theorem. It first constructs the order-n learned-score sensitivity example, and then gives the two-policy likelihood-ratio argument showing that a fixed-score calibration can violate pure DP.

C.1 Order-n Learned-Score Sensitivity

Proof of the separation part of Theorem 1. We construct a single population distribution and a triangular sequence of deterministic learners. Let P be the distribution with binary actions $\mathcal{A} = \{1,2\}$, covariates $X \sim \mathrm{Unif}[0,1]$, potential outcomes $Y(1) = Y(2) = 0$, and propensities $e_{0,1}(x) = e_{0,2}(x) = 1/2$ for all x. Let the policy be $\pi_1(x) \equiv 1$. The true outcome nuisances are $\mu_{0,1} = \mu_{0,2} = 0$, and $V_P(\pi_1) = 0$.

For each sample size $N = m + n$, define the rare sets

$$B_N = [0, N^{-4}], \qquad C_N = (N^{-4}, 2N^{-4}],$$

and define the deterministic trigger functional

$$T_N(S) = \mathbf{1}\{\exists j \in S : X_j \in C_N\}.$$

The learner $\widehat{\eta}_N$ maps a nuisance-fitting sample S to

$$\widehat{e}_{N,S,1}(x) = \widehat{e}_{N,S,2}(x) = 1/2,$$

$$\widehat{\mu}_{N,S,1}(x) = \mathbf{1}\{x \in B_N\}T_N(S), \qquad \widehat{\mu}_{N,S,2}(x) = 0.$$

This definition is allowed to depend on N, as stated in the theorem, but the population distribution P is fixed across N.

For every training sample S,

$$\max_a \|\widehat{\mu}_{N,S,a} - \mu_{0,a}\|_{L_2(P_X)} \le \|\mathbf{1}\{\cdot \in B_N\}\|_{L_2(P_X)} = \sqrt{P_X(B_N)} = N^{-2},$$

and $\max_a \|\widehat{e}_{N,S,a} - e_{0,a}\|_{L_2(P_X)} = 0$. Because the propensity nuisance is exactly correct, the product identity (2) gives

$$P\psi_{\pi_1}(\widehat{\eta}_N(S)) - V_P(\pi_1) = 0$$

for every training sample S.

It remains to show large worst-case sensitivity. Construct adjacent training blocks $S \sim S'$ as follows. Let S contain exactly one record with $X \in C_N$, let S' be obtained by replacing that record with a record whose covariate is outside C_N, and let all other training covariates in both samples lie outside C_N. Then $T_N(S) = 1$ and $T_N(S') = 0$. Use a common scoring block for $D = (S, D^{\mathrm{sc}})$ and $D' = (S', D^{\mathrm{sc}})$ with every scoring record satisfying $X_i \in B_N$, $A_i = 2$, and $Y_i = 0$. These records have vanishing probability under P^n, but they are valid points in the sample space. Pure-DP global sensitivity is worst-case over all adjacent datasets in the sample space, not high-probability under P^n.

For each scoring record in this block, $A_i \neq \pi_1(X_i)$, so the inverse-propensity correction in (1) is zero and

$$\psi_{\pi_1}\{Z_i; \widehat{\eta}_N(S)\} = \widehat{\mu}_{N,S,1}(X_i) = 1, \qquad \psi_{\pi_1}\{Z_i; \widehat{\eta}_N(S')\} = \widehat{\mu}_{N,S',1}(X_i) = 0.$$

Therefore each of the n scoring summands changes by one, and

$$|U_{\mathrm{ls}}(D, \pi_1) - U_{\mathrm{ls}}(D', \pi_1)| = n.$$

This proves the order-n sensitivity, correct-propensity, nuisance-accuracy, and zero-bias claims for the single-policy construction. □

C.2 Two-Policy DP Violation Under Naive Calibration

Proof of the two-policy part of Theorem 1. Use the same fixed population distribution as in Appendix C: $X \sim \mathrm{Unif}[0, 1]$, $\mathcal{A} = \{1, 2\}$, $Y(1) = Y(2) = 0$, and $e_{0,1} = e_{0,2} = 1/2$. Let $\Pi_n = \{\pi_0, \pi_1\}$ with $\pi_0(x) \equiv 2$ and $\pi_1(x) \equiv 1$. For $N = m + n$, let

$$B_N = [0, N^{-4}], \qquad C_N = (N^{-4}, 2N^{-4}], \qquad T_N(S) = \mathbf{1}\{\exists j \in S : X_j \in C_N\}.$$

Define the learner, for every training sample S, by

$$\widehat{e}_{N,S,1}(x) = \widehat{e}_{N,S,2}(x) = 1/2, \qquad \widehat{\mu}_{N,S,2}(x) = 0,$$

and

$$\widehat{\mu}_{N,S,1}(x) = \mathbf{1}\{x \in B_N\}\{2T_N(S) - 1\}.$$

This is a fully specified deterministic sample-size-dependent nuisance learner.

We verify the statistical properties first. The true outcome nuisances are zero, so for every S,

$$\max_a \|\widehat{\mu}_{N,S,a} - \mu_{0,a}\|_{L_2(P_X)} = \|\mathbf{1}\{\cdot \in B_N\}\|_{L_2(P_X)} = N^{-2},$$

and the propensity error is zero. Since $\widehat{e}_N = e_0$, Lemma 3 gives

$$P\psi_{\pi_b}(\widehat{\eta}_N(S)) = V_P(\pi_b) = 0, \qquad b \in \{0, 1\},$$

for every training sample S. The realized scores are uniformly bounded. Indeed, for π_0, $\mu_2 \equiv 0$ and $Y = 0$, so $\psi_{\pi_0} \equiv 0$. For π_1, outside B_N the learned mean is zero and the score is zero. On B_N, $\widehat{\mu}_1$ equals either $+1$ or -1. If $A = 2$, the inverse-propensity term is absent and the score is ± 1, while if $A = 1$ the score is

$$\widehat{\mu}_1(X) + 2\{0 - \widehat{\mu}_1(X)\} = -\widehat{\mu}_1(X),$$

also in $\{-1, +1\}$. Hence $|\psi_{\pi_b}(z; \widehat{\eta}_N(S))| \leq 1$ for $b = 0, 1$, all z, and all S.

Now choose adjacent training blocks $S \sim S'$ such that $T_N(S) = 1$ and $T_N(S') = 0$, as in Appendix C. Use the same scoring block in both datasets, with $X_i \in B_N$, $A_i = 2$, and $Y_i = 0$ for all $i = 1, \ldots, n$. Then

$$U(D, \pi_0) = U(D', \pi_0) = 0, \qquad U(D, \pi_1) = n, \qquad U(D', \pi_1) = -n.$$

For the exponential mechanism calibrated at the fixed value Δ_0, put $a = \varepsilon n/(2\Delta_0)$. The probability of outputting π_1 under D is $e^a/(1+e^a)$, while under D' it is $e^{-a}/(1+e^{-a}) = 1/(1+e^a)$. Therefore

$$\frac{\mathbb{P}\{\mathcal{M}(D) = \pi_1\}}{\mathbb{P}\{\mathcal{M}(D') = \pi_1\}} = e^a = \exp\left\{\frac{\varepsilon n}{2\Delta_0}\right\}.$$

If $n > 2\Delta_0$, this ratio is strictly larger than e^ε. Taking the measurable event $\{\pi_1\}$ violates the defining likelihood-ratio inequality for pure ε-DP. □

D Proofs of Theorem 2 and Corollary 1

This appendix proves the certified learned-score selection result in three steps. First, Appendix D.1 proves the pure-DP claim using the sensitivity certificate from Section 4.3. Second, Lemma 4 proves uniform statistical accuracy of the learned DR utilities by combining scoring-sample concentration with the DR product remainder. Third, Lemma 5 gives the finite-library exponential-mechanism oracle inequality, after which we assemble the regret bound in Theorem 2. Corollary 1 is then an immediate consequence of the four vanishing terms in the theorem.

D.1 Differential Privacy Guarantee

Proof of the DP part of Theorem 2. The certificate gives $\Gamma_{\mu,m} \le \beta_{\mu,m}$ and $\Gamma_{e,m} \le \beta_{e,m}$, hence Lemma 1 gives

$$\sup_{D\sim D'} \sup_{\pi\in\Pi_n} |U_{\mathrm{ls}}(D,\pi) - U_{\mathrm{ls}}(D',\pi)| \le \Delta_{\mathrm{cert}}.$$

The standard exponential-mechanism proof then applies. The sensitivity bound implies both $U(D,\pi) \le U(D',\pi) + \Delta_{\mathrm{cert}}$ and $U(D,\rho) \ge U(D',\rho) - \Delta_{\mathrm{cert}}$ for every $\pi, \rho \in \Pi_n$. Hence the numerator ratio is at most $\exp\{\varepsilon/2\}$, and the normalizers satisfy

$$\sum_\rho \exp\{\varepsilon U(D,\rho)/(2\Delta_{\mathrm{cert}})\} \ge e^{-\varepsilon/2} \sum_\rho \exp\{\varepsilon U(D',\rho)/(2\Delta_{\mathrm{cert}})\}.$$

Therefore, for any output π,

$$\frac{\mathbb{P}\{\widehat{\pi} = \pi \mid D\}}{\mathbb{P}\{\widehat{\pi} = \pi \mid D'\}} \le \exp\{\varepsilon/2\} \cdot \exp\{\varepsilon/2\} = e^\varepsilon.$$

Summing over events gives pure ε-DP. □

D.2 Doubly Robust Identity and Product Remainder

Lemma 3. *For any nuisance pair $\eta = (\mu, e)$ satisfying $e_a(x) > 0$ for all a, x and any deterministic policy π,*

$$P\psi_\pi(\eta) - V_P(\pi) = \sum_{a=1}^K \mathbb{E}\left[\mathbf{1}\{\pi(X) = a\} \frac{e_a(X) - e_{0,a}(X)}{e_a(X)} \{\mu_a(X) - \mu_{0,a}(X)\}\right].$$

Consequently, under the overlap range $e_a(x) \ge \zeta$,

$$|P\psi_\pi(\eta) - V_P(\pi)| \le \zeta^{-1} \sum_{a=1}^K \|\mu_a - \mu_{0,a}\|_{L_2(P_X)} \|e_a - e_{0,a}\|_{L_2(P_X)}.$$

The proof of Lemma 3 is stated below.

Proof of Lemma 3. Fix x and let $a = \pi(x)$. By consistency and unconfoundedness,

$$\mathbb{E}[Y \mid A = a, X = x] = \mu_{0,a}(x), \qquad \mathbb{P}(A = a \mid X = x) = e_{0,a}(x).$$

Taking the conditional expectation of the score in (1) given $X = x$ gives

$$\begin{aligned}\mathbb{E}[\psi_\pi(Z;\eta) \mid X = x] &= \mu_a(x) + \frac{e_{0,a}(x)}{e_a(x)}\{\mu_{0,a}(x) - \mu_a(x)\} \\ &= \mu_{0,a}(x) + \left(1 - \frac{e_{0,a}(x)}{e_a(x)}\right)\{\mu_a(x) - \mu_{0,a}(x)\} \\ &= \mu_{0,a}(x) + \frac{e_a(x) - e_{0,a}(x)}{e_a(x)}\{\mu_a(x) - \mu_{0,a}(x)\}.\end{aligned}$$

Since $V_P(\pi) = \mathbb{E}[\mu_{0,\pi(X)}(X)]$, taking expectation over X and decomposing according to the events $\{\pi(X) = a\}$ proves the identity.

We now prove the product-remainder bound explicitly. Starting from the identity,

$$\begin{aligned}&|P\psi_\pi(\eta) - V_P(\pi)| \\ &\le \sum_{a=1}^{K} \mathbb{E}\left[\mathbf{1}\{\pi(X) = a\}\left|\frac{e_a(X) - e_{0,a}(X)}{e_a(X)}\right||\mu_a(X) - \mu_{0,a}(X)|\right] \\ &\le \zeta^{-1}\sum_{a=1}^{K} \mathbb{E}\left[\mathbf{1}\{\pi(X) = a\}|e_a(X) - e_{0,a}(X)||\mu_a(X) - \mu_{0,a}(X)|\right] \\ &\le \zeta^{-1}\sum_{a=1}^{K} \mathbb{E}\left[|e_a(X) - e_{0,a}(X)||\mu_a(X) - \mu_{0,a}(X)|\right] \\ &\le \zeta^{-1}\sum_{a=1}^{K} \left\{\mathbb{E}|e_a(X) - e_{0,a}(X)|^2\right\}^{1/2}\left\{\mathbb{E}|\mu_a(X) - \mu_{0,a}(X)|^2\right\}^{1/2} \\ &= \zeta^{-1}\sum_{a=1}^{K} \|e_a - e_{0,a}\|_{L_2(P_X)}\,\|\mu_a - \mu_{0,a}\|_{L_2(P_X)}\,.\end{aligned}$$

The first inequality is the triangle inequality, the second uses $e_a(X) \ge \zeta$, the third uses $\mathbf{1}\{\pi(X) = a\} \le 1$, and the fourth is Cauchy–Schwarz applied to the ath summand. □

D.3 Regret Bound and Vanishing-Regret Corollary

Lemma 4. *Under the deterministic ranges and nuisance accuracy* (10)*, with probability at least* $1 - \alpha_\eta - \alpha$*,*

$$\sup_{\pi\in\Pi_n} \left|n^{-1}U_{\mathrm{ls}}(D,\pi) - V_P(\pi)\right| \le CB_\psi\sqrt{\frac{\log(2M_n/\alpha)}{n}} + K\zeta^{-1}r_{\mu,m}r_{e,m}.$$

The proof of Lemma 4 is stated below.

Proof of Lemma 4. Define the nuisance event

$$\mathcal{E}_\eta = \left\{\max_a \|\widehat{\mu}_a - \mu_{0,a}\|_{L_2(P_X)} \le r_{\mu,m}, \quad \max_a \|\widehat{e}_a - e_{0,a}\|_{L_2(P_X)} \le r_{e,m}\right\}.$$

By assumption, $\mathbb{P}(\mathcal{E}_\eta) \ge 1 - \alpha_\eta$ over the nuisance-fitting block. Condition on an arbitrary realization of D^{tr}, so that $\widehat{\eta} = \widehat{\eta}(D^{\mathrm{tr}})$ is fixed. Conditional on this training block, the scoring records are independent, and for every fixed $\pi \in \Pi_n$ the variables $\psi_\pi(Z_{m+i};\widehat{\eta})$ satisfy $|\psi_\pi(Z_{m+i};\widehat{\eta})| \le B_\psi$. Hence each summand lies in an interval of length at most $2B_\psi$. Hoeffding's inequality gives, for every $t > 0$,

$$\mathbb{P}\left(\left|n^{-1}U_{\mathrm{ls}}(D,\pi) - P\psi_\pi(\widehat{\eta})\right| > t \,\middle|\, D^{\mathrm{tr}}\right) \le 2\exp\left\{-\frac{nt^2}{2B_\psi^2}\right\}.$$

Choose

$$t = B_\psi \sqrt{\frac{2\log(2M_n/\alpha)}{n}}.$$

Then

$$2\exp\left\{-\frac{nt^2}{2B_\psi^2}\right\} = 2\exp\{-\log(2M_n/\alpha)\} = \frac{\alpha}{M_n}.$$

Thus, for each fixed π,

$$\mathbb{P}\left(\left|n^{-1}U_{\mathrm{ls}}(D,\pi) - P\psi_\pi(\widehat{\eta})\right| > B_\psi\sqrt{\frac{2\log(2M_n/\alpha)}{n}}\,\middle|\, D^{\mathrm{tr}}\right) \le \frac{\alpha}{M_n}.$$

Applying the union bound over the M_n policies,

$$\begin{aligned}&\mathbb{P}\left(\sup_{\pi\in\Pi_n}\left|n^{-1}U_{\mathrm{ls}}(D,\pi) - P\psi_\pi(\widehat{\eta})\right| > B_\psi\sqrt{\frac{2\log(2M_n/\alpha)}{n}}\,\middle|\, D^{\mathrm{tr}}\right)\\ &\quad\le \sum_{\pi\in\Pi_n}\mathbb{P}\left(\left|n^{-1}U_{\mathrm{ls}}(D,\pi) - P\psi_\pi(\widehat{\eta})\right| > B_\psi\sqrt{\frac{2\log(2M_n/\alpha)}{n}}\,\middle|\, D^{\mathrm{tr}}\right) \le \alpha.\end{aligned}$$

Equivalently, there is a conditional scoring event $\mathcal{E}_{\mathrm{sc}}(D^{\mathrm{tr}})$ with conditional probability at least $1-\alpha$ on which

$$\sup_{\pi\in\Pi_n}\left|n^{-1}U_{\mathrm{ls}}(D,\pi) - P\psi_\pi(\widehat{\eta})\right| \le CB_\psi\sqrt{\frac{\log(2M_n/\alpha)}{n}},$$

where C absorbs the displayed constant $\sqrt{2}$. Because the conditional bound holds for every training block, integrating over D^{tr} gives $\mathbb{P}(\mathcal{E}_{\mathrm{sc}}) \ge 1-\alpha$, and hence $\mathbb{P}(\mathcal{E}_\eta \cap \mathcal{E}_{\mathrm{sc}}) \ge 1-\alpha_\eta-\alpha$.

On $\mathcal{E}_\eta$, Lemma 3 gives, uniformly over $\pi\in\Pi_n$,

$$|P\psi_\pi(\widehat{\eta}) - V_P(\pi)| \le \sum_{a=1}^{K}\zeta^{-1}\left\|\widehat{e}_a - e_{0,a}\right\|_{L_2(P_X)}\left\|\widehat{\mu}_a - \mu_{0,a}\right\|_{L_2(P_X)} \le K\zeta^{-1}r_{\mu,m}r_{e,m}.$$

On $\mathcal{E}_\eta\cap\mathcal{E}_{\mathrm{sc}}$, the triangle inequality combines the concentration bound and the product-remainder bound to prove the stated uniform value bound. □

Lemma 5. *Condition on any dataset D. If the exponential mechanism uses utility $U_{\mathrm{ls}} = n\widehat{V}_{\mathrm{ls}}$ and sensitivity Δ_{cert}, then with probability at least $1-\beta$,*

$$\max_{\pi\in\Pi_n}\widehat{V}_{\mathrm{ls}}(\pi) - \widehat{V}_{\mathrm{ls}}(\widehat{\pi}) \le \frac{2\Delta_{\mathrm{cert}}}{n\varepsilon}\{\log M_n + \log(1/\beta)\}.$$

The proof of Lemma 5 is stated below.

Proof of Lemma 5. Let $U^* = \max_{\pi\in\Pi_n}U_{\mathrm{ls}}(D,\pi)$ and

$$t = \frac{2\Delta_{\mathrm{cert}}}{\varepsilon}\{\log M_n + \log(1/\beta)\}.$$

Define the bad set

$$\mathcal{B}_t = \{\pi\in\Pi_n : U_{\mathrm{ls}}(D,\pi) \le U^* - t\}.$$

Since at least one policy attains utility U^*, the exponential-mechanism normalizer is at least $\exp\{\varepsilon U^*/(2\Delta_{\mathrm{cert}})\}$. Therefore,

$$\begin{aligned}\mathbb{P}(\widehat{\pi}\in\mathcal{B}_t \mid D) &\le \frac{|\mathcal{B}_t|\exp\{\varepsilon(U^*-t)/(2\Delta_{\mathrm{cert}})\}}{\exp\{\varepsilon U^*/(2\Delta_{\mathrm{cert}})\}}\\ &\le M_n\exp\{-\varepsilon t/(2\Delta_{\mathrm{cert}})\} = \beta.\end{aligned}$$

Thus, with conditional probability at least $1-\beta$, the selected policy has utility loss at most t. Dividing by n proves the stated inequality for $\widehat{V}_{\mathrm{ls}} = n^{-1}U_{\mathrm{ls}}$. □

Proof of the Regret Bound in Theorem 2. The pure-DP claim was proved in Appendix D.1. We now prove the regret inequality with the advertised probability. Let

$$\widehat{V}_{\mathrm{ls}}(\pi) = n^{-1} U_{\mathrm{ls}}(D, \pi), \qquad \pi_n^\star \in \arg\max_{\pi \in \Pi_n} V_P(\pi).$$

Let $\mathcal{E}_{\mathrm{val}}$ be the event from Lemma 4. It has probability at least $1 - \alpha_\eta - \alpha$ over $D^{\mathrm{tr}}, D^{\mathrm{sc}}$. Conditional on the realized dataset D, Lemma 5 gives an event $\mathcal{E}_{\mathrm{em}}(D)$ over only the exponential-mechanism randomness with conditional probability at least $1 - \beta$ on which

$$\max_{\pi \in \Pi_n} \widehat{V}_{\mathrm{ls}}(\pi) - \widehat{V}_{\mathrm{ls}}(\widehat{\pi}) \le \frac{2\Delta_{\mathrm{cert}}}{n\varepsilon} \{\log M_n + \log(1/\beta)\}.$$

Integrating this conditional probability bound over the data and applying a union bound gives

$$\mathbb{P}(\mathcal{E}_{\mathrm{val}} \cap \mathcal{E}_{\mathrm{em}}(D)) \ge 1 - \alpha_\eta - \alpha - \beta.$$

On this intersection of events,

$$\begin{aligned}
V_P(\pi_P^\star) - V_P(\widehat{\pi}) &= \{V_P(\pi_P^\star) - V_P(\pi_n^\star)\} + \{V_P(\pi_n^\star) - V_P(\widehat{\pi})\} \\
&\le \mathcal{A}_n(P; \Pi, \Pi_n) + \{V_P(\pi_n^\star) - \widehat{V}_{\mathrm{ls}}(\pi_n^\star)\} \\
&\quad + \{\widehat{V}_{\mathrm{ls}}(\pi_n^\star) - \widehat{V}_{\mathrm{ls}}(\widehat{\pi})\} + \{\widehat{V}_{\mathrm{ls}}(\widehat{\pi}) - V_P(\widehat{\pi})\} \\
&\le \mathcal{A}_n(P; \Pi, \Pi_n) + 2 \sup_{\pi \in \Pi_n} \left| \widehat{V}_{\mathrm{ls}}(\pi) - V_P(\pi) \right| \\
&\quad + \frac{2\Delta_{\mathrm{cert}}}{n\varepsilon} \{\log M_n + \log(1/\beta)\} \\
&\le \mathcal{A}_n(P; \Pi, \Pi_n) + 2CB_\psi \sqrt{\frac{\log(2M_n/\alpha)}{n}} + 2K\zeta^{-1} r_{\mu,m} r_{e,m} \\
&\quad + \frac{2\Delta_{\mathrm{cert}}}{n\varepsilon} \{\log M_n + \log(1/\beta)\}.
\end{aligned}$$

Renaming numerical constants as C, C', C'' yields (11). □

Proof of Corollary 1. Apply Theorem 2. Under the stated assumptions, the library approximation term, the scoring-sample concentration term, and the doubly robust product term vanish. The privacy term also vanishes because

$$\frac{\Delta_{\mathrm{cert}}}{n\varepsilon} \{\log M_n + \log(1/\beta)\} = \max\left\{ \frac{2B_\psi}{n\varepsilon}, \frac{\rho_m^\star}{\varepsilon} \right\} \{\log M_n + \log(1/\beta)\} \to 0,$$

where the first component is controlled by $\log(M_n)/n \to 0$ and the second by the assumed stability condition. If $\rho_m^\star = O((m\lambda_m)^{-1})$, the displayed condition follows from $(\log M_n + \log(1/\beta))/(\varepsilon m \lambda_m) \to 0$. □

E Regularized ERM Stability Proof and Extensions

As discussed in Section 4.3, achieving non-vacuous deterministic stability certificates is crucial for applying the certified private selection mechanism. In this appendix, we formally establish the baseline algorithmic stability certificate for scalar regularized Empirical Risk Minimization (ERM) (Proposition 2), explicitly formalize its extensions to multi-component and vector-valued settings (Corollary 2), and provide the complete unified proof.

Proposition 2. *Let a scalar nuisance estimator solve the following regularized objective (with deterministic tie-breaking if necessary):*

$$\widehat{\theta}_S \in \arg\min_\theta \left\{ \frac{1}{m} \sum_{i \in S} \ell(\theta; Z_i) + \frac{\lambda_m}{2} \|\theta\|_2^2 \right\}.$$

Assume the loss function $\ell(\theta; z)$ is convex and G-Lipschitz in θ for all z, which ensures the overall objective is λ_m-strongly convex. Further assume the scalar prediction map $g_\theta(x)$ is L_g-Lipschitz in θ uniformly over x. Then, for any adjacent training samples $S \sim S'$, the maximum pointwise deviation satisfies:

$$\sup_x |g_{\widehat{\theta}_S}(x) - g_{\widehat{\theta}_{S'}}(x)| \leq \frac{2L_g G}{m\lambda_m}.$$

The proof of Proposition 2 is stated below. The Lipschitz loss assumption is used only through a uniform bound on the loss gradient along the parameter region in which the two adjacent minimizers lie. Hence, for a differentiable convex loss, the same proof applies whenever one can exhibit a public set Θ_m containing all possible minimizers and a public constant G such that $\sup_{\theta\in\Theta_m, z} \|\nabla_\theta \ell(\theta; z)\|_2 \leq G$. This localized form is the one used for the ridge outcome learner in the certified ERM experiment below; squared loss is not globally Lipschitz, but the bounded design and ridge penalty give a public norm bound on every fitted parameter vector, and therefore a public gradient bound on the relevant fitted domain.

Corollary 2. *Suppose the conditions of Proposition 2 hold.*

1. ***Coordinate-wise outcome learning and simplex propensity post-processing:*** *If each outcome component μ_a is learned by an independent scalar ERM learner whose objective is averaged over all m records and then deterministically clipped to $[\underline{\mu}, \overline{\mu}]$, then*

$$\beta_{\mu,m} \leq \max_a \frac{2L_{g,\mu,a}G_{\mu,a}}{m\lambda_{\mu,a}}.$$

If, in addition, K coordinate-wise propensity score predictors $g_{e,a}$ are learned by independent scalar ERM objectives averaged over all m records and the preliminary vector $g_e(x) = (g_{e,1}(x), \ldots, g_{e,K}(x))$ is deterministically post-processed as $\widehat{e}(x) = \mathcal{C}_\zeta(g_e(x)) \in \Delta_K^\zeta$, where $\mathcal{C}_\zeta$ is $L_{\mathcal{C},\infty}$-Lipschitz from ℓ_∞ to ℓ_∞, then

$$\beta_{e,m} \leq L_{\mathcal{C},\infty} \max_a \frac{2L_{g,e,a}G_{e,a}}{m\lambda_{e,a}}.$$

2. ***Vector-valued propensity learning:*** *Alternatively, suppose the multi-action propensity is learned jointly via a single vector ERM with a prediction map $g_\theta : \mathcal{X} \to \mathbb{R}^K$ satisfying $\sup_x \|g_\theta(x) - g_{\theta'}(x)\|_2 \leq L_{g,e}\|\theta - \theta'\|_2$. If the deterministic projection map $\mathcal{C}_\zeta : \mathbb{R}^K \to \Delta_K^\zeta$ used to produce $\widehat{e}(x) = \mathcal{C}_\zeta(g_{\widehat{\theta}}(x))$ is $L_{\mathcal{C}}$-Lipschitz from ℓ_2 to ℓ_∞ uniformly in x, then the propensity certificate satisfies:*

$$\beta_{e,m} \leq \frac{2L_{\mathcal{C}}L_{g,e}G_e}{m\lambda_e}.$$

The proof of Corollary 2 is stated below.

Proof of Proposition 2 and Corollary 2. We first prove the scalar stability statement of Proposition 2. Define the empirical objectives on the adjacent samples S and S' as:

$$F_S(\theta) = \frac{1}{m}\sum_{i\in S} \ell(\theta; Z_i) + \frac{\lambda_m}{2}\|\theta\|_2^2, \qquad F_{S'}(\theta) = \frac{1}{m}\sum_{i\in S'} \ell(\theta; Z_i) + \frac{\lambda_m}{2}\|\theta\|_2^2.$$

Let $\theta = \widehat{\theta}_S$ and $\theta' = \widehat{\theta}_{S'}$. By the λ_m-strong convexity of the objectives and the optimality of θ and θ', we obtain the inequalities:

$$F_S(\theta') \geq F_S(\theta) + \frac{\lambda_m}{2}\|\theta' - \theta\|_2^2, \qquad F_{S'}(\theta) \geq F_{S'}(\theta') + \frac{\lambda_m}{2}\|\theta' - \theta\|_2^2.$$

Adding these two inequalities yields:

$$\lambda_m\|\theta' - \theta\|_2^2 \leq \{F_S(\theta') - F_{S'}(\theta')\} + \{F_{S'}(\theta) - F_S(\theta)\}.$$

Because S and S' differ only by the replacement of a single record, say z in S replaced by z' in S', the regularization terms and the losses evaluated on all common records perfectly cancel. The right-hand side simplifies to:

$$\frac{1}{m}\{\ell(\theta';z)-\ell(\theta';z')+\ell(\theta;z')-\ell(\theta;z)\}.$$

Applying the G-Lipschitz continuity of the loss function ℓ with respect to θ, we bound the differences:

$$\ell(\theta';z)-\ell(\theta;z)\le G\|\theta'-\theta\|_2, \qquad \ell(\theta;z')-\ell(\theta';z')\le G\|\theta'-\theta\|_2.$$

Substituting these bounds yields:

$$\lambda_m\|\theta'-\theta\|_2^2\le\frac{2G}{m}\|\theta'-\theta\|_2.$$

If $\theta'=\theta$, the desired bound holds trivially. Otherwise, dividing by $\|\theta'-\theta\|_2$ gives $\|\theta'-\theta\|_2\le 2G/(m\lambda_m)$. Finally, applying the L_g-Lipschitz continuity of the scalar prediction map $g_\theta(x)$ provides the scalar certificate:

$$\sup_x|g_{\widehat{\theta}_S}(x)-g_{\widehat{\theta}_{S'}}(x)|\le L_g\|\widehat{\theta}_S-\widehat{\theta}_{S'}\|_2\le\frac{2L_gG}{m\lambda_m}.$$

We now prove the extensions detailed in Corollary 2. For coordinate-wise outcome learning, deterministic clipping to a fixed interval is a 1-Lipschitz operation. Applying the scalar stability result separately to each action-specific outcome learner immediately yields:

$$\max_a\sup_x|\widehat{\mu}_{S,a}(x)-\widehat{\mu}_{S',a}(x)|\le\max_a\frac{2L_{g,\mu,a}G_{\mu,a}}{m\lambda_{\mu,a}}.$$

For coordinate-wise propensity learning, the scalar argument gives

$$\sup_x\|g_{e,S}(x)-g_{e,S'}(x)\|_\infty\le\max_a\frac{2L_{g,e,a}G_{e,a}}{m\lambda_{e,a}}.$$

The released denominator vector is not obtained by independent coordinate clipping. It is the deterministic post-processing $\widehat{e}_S(x)=\mathcal{C}_\zeta(g_{e,S}(x))\in\Delta_K^\zeta$. By the assumed ℓ_∞-to-ℓ_∞ Lipschitz property of $\mathcal{C}_\zeta$,

$$\max_a\sup_x|\widehat{e}_{S,a}(x)-\widehat{e}_{S',a}(x)|\le L_{\mathcal{C},\infty}\max_a\frac{2L_{g,e,a}G_{e,a}}{m\lambda_{e,a}}.$$

The maximum over actions simply introduces the maximum of deterministic action-specific constants. No probabilistic union bound is required since the certificate is deterministic and holds simultaneously for all $a\in\mathcal{A}$.

For the vector propensity implementation, the identical strong-convexity argument applies to the vector-valued objective, yielding $\|\widehat{\theta}_S-\widehat{\theta}_{S'}\|_2\le 2G_e/(m\lambda_e)$. The vector prediction Lipschitz condition then guarantees:

$$\sup_x\|g_{\widehat{\theta}_S}(x)-g_{\widehat{\theta}_{S'}}(x)\|_2\le\frac{2L_{g,e}G_e}{m\lambda_e}.$$

Because the deterministic post-processing map $\mathcal{C}_\zeta$ is $L_{\mathcal{C}}$-Lipschitz from ℓ_2 to ℓ_∞ uniformly in x, we map this to the maximum element-wise deviation:

$$\max_a\sup_x|\widehat{e}_{S,a}(x)-\widehat{e}_{S',a}(x)|\le\frac{2L_{\mathcal{C}}L_{g,e}G_e}{m\lambda_e}.$$

Thus, both the scalar outcome learners and the multi-action vector propensity learners successfully produce the deterministic certificates required by (8). Substituting these components into (9) confirms the stated $O((m\lambda_m)^{-1})$ order for the spillover modulus $\rho_m^\star$, provided the primitive constants are bounded and the regularization levels are of common order λ_m. □

F Additional Example: Nearest-Neighbor Spillover

Consider binary actions, a constant policy $\pi_1(x) \equiv 1$, fixed propensities $\widehat{e}_1(x) = \widehat{e}_2(x) = 1/2$, and outcomes in $[0, 1]$. Let $\widehat{\mu}_{S,1}$ be clipped one-nearest-neighbor regression using action-1 training records, and let $\widehat{\mu}_{S,2} = 0$. Construct adjacent samples $S \sim S'$ by replacing a single action-1 record at covariate x_0 with outcome 1 in S and outcome 0 in S', with all other action-1 covariates farther from x_0. Then

$$\sup_x |\widehat{\mu}_{S,1}(x) - \widehat{\mu}_{S',1}(x)| = 1.$$

Use a common scoring block with all $X_i = x_0$, $A_i = 2$, and $Y_i = 0$. Since $A_i \neq \pi_1(X_i)$, the inverse-propensity correction is zero and each DR score equals the learned action-1 mean. Therefore

$$|U_{\mathrm{ls}}(D, \pi_1) - U_{\mathrm{ls}}(D', \pi_1)| = n.$$

G Additional Experimental Details

G.1 Experimental Setup Details

ACIC preprocessing. We preprocess the ACIC 2016 covariates as follows. A raw column is treated as numeric if parseable numeric entries account for at least 90% of all rows; otherwise it is treated as categorical. Missing numeric entries are imputed by the column median, and numeric columns are standardized to have zero mean and unit empirical standard deviation. Categorical columns are one-hot encoded, with missing values treated as an additional level. This produces 4,802 rows and 58 encoded covariates.

Rare-region DGP. Let

$$h_R(x) = \sum_{j=1}^{\min\{8,d\}} c_j x_j + \mathbf{1}\{d > 15\}\{0.5 \sin(x_8) + 0.25 x_{11} x_{15}\},$$

where $(c_1, \ldots, c_{\min\{8,d\}})$ is linearly spaced from 1.2 to -0.9. The rare-region indicator is

$$R(x) = \mathbf{1}\{h_R(x) \geq q_{0.95}\},$$

where $q_{0.95}$ is the empirical 95th percentile of the rare-region score on the ACIC covariate pool. For $k = \min\{10, d\}$, define sparse coefficient vectors

$$w_e = \mathrm{linspace}(0.85, -0.65, k), \quad w_0 = \mathrm{linspace}(0.55, -0.35, k), \quad w_\tau = \mathrm{linspace}(-0.75, 0.95, k),$$

padded with zeros outside the first k coordinates. The main rare-region separation audit uses

$$\begin{aligned} g_e(x) &= 1.2 \frac{x^\top w_e}{\sqrt{k}} + 0.4 \sin(x_1) - 0.25 x_2 \mathbf{1}\{x_3 > 0\}, \\ g_0(x) &= \frac{x^\top w_0}{\sqrt{k}} + 0.3 \sin(x_1 x_4) + 0.2 \mathbf{1}\{x_5 > 0\}, \\ g_\tau(x) &= \frac{x^\top w_\tau}{\sqrt{k}}. \end{aligned}$$

The response surfaces used in the main text are those in Section 5.1. Standard samples use Gaussian noise with standard deviation 0.06. In the targeted adjacent-pair construction, non-anchor outcomes use standard deviation 0.04, and the rare anchor outcome is deterministically set to 1 in one fitting sample and 0 in its adjacent counterpart.

Targeted adjacent-pair audit. The rare-region separation audit uses adjacent nuisance-fitting samples designed to expose the spillover channel. Each base fitting sample contains one rare-region anchor and $m-1$ nonrare records. The adjacent sample keeps the same covariates and treatment assignments but flips the anchor outcome from 1 to 0. The stress scoring block is sampled from the rare pool. This construction changes one training record while evaluating the resulting learned score map on a scoring block concentrated in the low-mass region, matching the mechanism highlighted by Theorem 1.

For each audited adjacent pair, we compute

$$\widehat{\Delta}_{\mathrm{move}} = \max_{\pi \in \Pi_n} |U_{\mathrm{ls}}(D, \pi) - U_{\mathrm{ls}}(D', \pi)|.$$

When a reference audit scale is needed in plots, we use

$$\widehat{\Delta}_{\mathrm{audit}} = \max\{2B_\psi,\, 1.3\,\widehat{\Delta}_{\mathrm{move}}\}.$$

Separately, on the same scoring block we compute the empirical prediction movement

$$\widehat{\Gamma}_\mu = \max_{a \in \{0,1\}} \max_{x \in D^{\mathrm{sc}}} |\widehat{\mu}_{S,a}(x) - \widehat{\mu}_{S',a}(x)|,$$
$$\widehat{\Gamma}_e = \max_{x \in D^{\mathrm{sc}}} |\widehat{e}_S(x) - \widehat{e}_{S'}(x)|,$$

and the empirical decomposition proxy

$$\widehat{\rho}_{\mathrm{pred}} = (1 + \zeta^{-1})\widehat{\Gamma}_\mu + \zeta^{-2}\widehat{\Gamma}_e, \qquad \widehat{\Delta}_{\mathrm{proxy}} = \max\{2B_\psi, n\widehat{\rho}_{\mathrm{pred}}\}.$$

We use the term "audit" for empirical replace-one measurements and reserve "certificate" for deterministic public upper bounds of the form (8).

Policy library. For the default library size $M = 160$, the public policy class consists of 18 structured policies and 142 random linear policies. The structured policies are:

$$\begin{array}{cc} \text{always treat}, & \text{never treat}, \\ \mathbf{1}\{x_j \geq 0\}, \quad \mathbf{1}\{x_j < 0\}, & j = 1, \ldots, \min\{6, d\}, \\ \mathbf{1}\{x^\top w_\tau \geq 0\}, & \mathbf{1}\{x^\top w_\tau < 0\}, \\ \mathbf{1}\{h_R(x) \geq q_{0.95}\}, & \mathbf{1}\{h_R(x) < q_{0.95}\}. \end{array}$$

The remaining policies are generated from public randomness. For each random linear rule, we sample a Gaussian direction; with probability 0.70, we retain a uniformly random subset of $\max\{2, \min\{8, d\}\}$ coordinates and zero out the rest. We then normalize the direction to unit Euclidean norm, draw an intercept from $N(0, 0.2^2)$, and use the threshold rule $\mathbf{1}\{x^\top w + b \geq 0\}$. The policy library is generated independently of the private sample.

Nuisance learners. Propensities are fit by logistic regression and clipped to $[\zeta, 1-\zeta]$. In the regularization sweeps, the logistic penalty is $\max\{0.1, \lambda\}$. Outcome regressions are fit separately for the two treatment arms using boosted-tree regressors with squared-error loss, subsampling and column subsampling set to one, one computational thread, and fixed public seeds. The stable default learner uses depth 3, 100 trees, learning rate 0.06, and $\lambda = 6$. The unstable stress learner uses depth 7, 180 trees, learning rate 0.06, and $\lambda = 10^{-8}$. In the regularization audit, the adjacent pair, scoring block, and policy library are held fixed across values of λ so that changes in movement are attributable to learner regularization.

Certified ERM positive-control learner. The certificate experiment differs from the boosted-tree stress tests in two ways. First, every covariate is deterministically clipped to $[-3, 3]$ and rescaled to $[-1, 1]$ before fitting. This public preprocessing gives the feature vector with intercept the deterministic radius $R_x = \sqrt{d+1}$, with $d = 20$ in the reported run. Second, the nuisance learner is a deterministic strongly convex

ERM rather than a flexible boosted tree. For each treatment arm, the outcome nuisance solves the ridge objective for the centered residual $Y - 1/2$,

$$\frac{1}{m}\sum_{i=1}^{m} \mathbf{1}\{A_i = a\}\frac{1}{2}\{\theta_a^\top \widetilde{X}_i - (Y_i - 1/2)\}^2 + \frac{\lambda_\mu}{2}\|\theta_a\|_2^2,$$

and predicts $\mathrm{clip}_{[0,1]}\{1/2 + \theta_a^\top \widetilde{X}\}$. The propensity nuisance solves the ℓ_2-regularized logistic ERM with fixed Newton/backtracking optimization and prediction clipping to $[\zeta, 1-\zeta]$. We use $\lambda_\mu = \lambda_e = 50$, $m = n = 1500$, $\zeta = 0.10$, and the same public policy library construction as in the main experiments.

With the bounded design, the public prediction-stability constants used in the certificate are

$$\beta_{\mu,m} = \frac{2L_{g,\mu}G_\mu}{m\lambda_\mu}, \qquad \beta_{e,m} = \frac{2L_{g,e}G_e}{m\lambda_e}.$$

Here the outcome constant is not obtained from a global Lipschitz property of squared loss. Let $r = Y - 1/2$, so $|r| \le 1/2$. For any possible outcome training sample, comparing the ridge objective at its minimizer $\hat{\theta}_a$ with the objective at 0 gives

$$\frac{\lambda_\mu}{2}\|\hat{\theta}_a\|_2^2 \le \frac{1}{m}\sum_{i=1}^{m} \mathbf{1}\{A_i = a\}\frac{1}{2}r_i^2 \le \frac{1}{8}, \qquad \text{so} \qquad \|\hat{\theta}_a\|_2 \le \frac{1}{2\sqrt{\lambda_\mu}}.$$

For the per-record squared residual loss, $\nabla_\theta \ell_a(\theta; Z) = \mathbf{1}\{A = a\}(\theta^\top \widetilde{X} - r)\widetilde{X}$. Since $\|\widetilde{X}\|_2 \le R_x$, every fitted minimizer satisfies the uniform gradient bound

$$\|\nabla_\theta \ell_a(\hat{\theta}_a; Z)\|_2 \le R_x\left(\frac{1}{2} + \frac{R_x}{2\sqrt{\lambda_\mu}}\right) =: G_\mu.$$

The prediction map $\theta \mapsto 1/2 + \theta^\top \widetilde{X}$, followed by clipping to $[0,1]$, is $L_{g,\mu} = R_x$-Lipschitz in θ. The localized-gradient version of Proposition 2 therefore yields the displayed $\beta_{\mu,m}$. For the logistic propensity learner, the regularized logistic loss has gradient norm at most R_x, and the clipped sigmoid prediction map is $R_x/4$-Lipschitz in θ; hence $G_e = R_x$ and $L_{g,e} = R_x/4$. The resulting learned-score stability radius is

$$\rho_m^\star = (1 + \zeta^{-1})\beta_{\mu,m} + \zeta^{-2}\beta_{e,m}, \qquad \Delta_{\mathrm{cert}} = \max\{2B_\psi, n\rho_m^\star\}.$$

The empirical replace-one movement and likelihood-ratio audits reported below are diagnostics only; the calibration scale itself is the public deterministic quantity Δ_{cert}.

Experiment grid and reported quantities. The main repetitions use $m = 1500$ fitting records, $n = 1500$ scoring records, library size $M = 160$, and 100 repetitions. The separation audit varies $n \in \{300, 700, 1200, 2000, 3000\}$. The regularization audit uses

$$\lambda \in \{0.01, 0.03, 0.1, 0.3, 1, 3, 10, 30, 100\}.$$

The sample-split decomposition fixes total sample size 3600 and varies the training fraction over

$$\{0.20,\ 0.25,\ 0.35,\ 0.50,\ 0.65,\ 0.75,\ 0.80\}.$$

Additional robustness outputs generated by the code vary the DGP regime, overlap level ζ, dimension d, policy-library size, and privacy budget. The certified ERM positive-control experiment uses 100 repetitions with $m = n = 1500$, $d = 20$, $\lambda_\mu = \lambda_e = 50$, and reports both public certificate quantities and empirical adjacent-pair diagnostics.

G.2 Column Definitions for Tables 1–3 and 5

All three tables report averages over 100 repetitions. In Table 1, "nuis. L_2" is nuisance RMSE; "DR prod." is the empirical doubly robust product-remainder proxy; "floor" is the fixed-score sensitivity $2B_\psi$; "unstable move/floor" is $\widehat{\Delta}_{\text{move}}/(2B_\psi)$ for the unstable learner; "naive logLR/ε" is the audited maximum log-likelihood ratio for learned-score exponential mechanism calibrated with the fixed-score floor, divided by ε; "stable logLR/ε" is the same audit after stability-based calibration; and "stable cover" is the fraction of sampled adjacent pairs whose observed movement is bounded by the stability-audit scale.

In Table 2, "$\widehat{\rho}_{\text{pred}}$" is the empirical replace-one prediction-movement proxy; "move" is $\widehat{\Delta}_{\text{move}} = \max_{\pi\in\Pi_n} |U_{\text{ls}}(D,\pi) - U_{\text{ls}}(D',\pi)|$; "audit Δ" is the unfloored audit scale $1.3\widehat{\Delta}_{\text{move}}$; "audit/floor" is that scale divided by $2B_\psi$; "floored Δ" is $\max\{2B_\psi, 1.3\widehat{\Delta}_{\text{move}}\}$ in each repetition; "floor rate" is the fraction of repetitions in which $2B_\psi$ is active; and "regret" is expected exponential-mechanism regret.

In Table 3, "fit frac." is $m/(m+n)$; "nuis. RMSE" is nuisance RMSE; "$\widehat{\rho}_{\text{pred}}$", "move", "floored Δ", "floor rate", and "regret" have the same meanings as in Table 2; and "$1/\sqrt{n}$" is the scoring-sample concentration scale from the regret decomposition.

In Table 5, $\rho^\star_m$ is the public deterministic stability radius computed from the bounded-design ERM constants, "emp. ρ" is the sampled prediction-movement proxy, "cert. Δ" is Δ_{cert}, "move/cert." is the observed utility movement divided by Δ_{cert}, "logLR/ε" is the empirical privacy-loss audit divided by the target privacy level, "cover" is the fraction of audited adjacent pairs covered by Δ_{cert}, "regret" is expected exponential-mechanism regret under the certified scale, and "nuis. RMSE" is the average nuisance RMSE across the two outcome nuisances and propensity nuisance.

For statistical diagnostics, we report nuisance RMSE for $\widehat{\mu}_0, \widehat{\mu}_1, \widehat{e}$, the DR product proxy

$$\widehat{r}_{\text{DR}} = \frac{1}{2}\{\text{RMSE}(\widehat{\mu}_0) + \text{RMSE}(\widehat{\mu}_1)\}\text{RMSE}(\widehat{e}),$$

and empirical regret. For exponential-mechanism policy-selection mechanisms, regret is the expectation under the mechanism's selection distribution rather than a single draw from the mechanism. For sensitivity diagnostics, we report the observed utility movement $\widehat{\Delta}_{\text{move}}$, the audit scale $\widehat{\Delta}_{\text{audit}}$, the prediction-movement proxy $\widehat{\rho}_{\text{pred}}$, and, in the output files, the decomposition scale $\widehat{\Delta}_{\text{proxy}}$.

G.3 Diagnostic Gap Between Sampled Audits and Theorem-Style Proxies

Table 4 compares the sampled adjacent-pair audit scale with the algebraic theorem-proxy scale

$$\widehat{\Delta}_{\text{proxy}} = \max\{2B_\psi, n\widehat{\rho}_{\text{pred}}\}.$$

For boosted-tree nuisances, $\widehat{\rho}_{\text{pred}}$ is an empirical prediction-movement proxy measured on sampled adjacent pairs, not a public deterministic stability certificate. Therefore $\widehat{\Delta}_{\text{proxy}}$ should be read as a theorem-shaped diagnostic scale: it shows the magnitude obtained by plugging the empirical prediction movement into the theorem's spillover formula, but it is not itself a formal pure-DP calibration unless $\widehat{\rho}_{\text{pred}}$ is replaced by a deterministic upper bound of the form (8).

The comparison highlights two points that are useful for interpreting the main experimental claims. First, the empirical audit movement and audit scale decrease with stronger regularization and with larger nuisance-fitting fractions, matching the qualitative direction of the spillover decomposition. Second, the plug-in theorem-proxy scale remains substantially larger than the sampled audit scale for boosted trees. This gap is expected: the audit scale measures observed movement over sampled adjacent pairs, whereas theorem-level pure DP requires a public deterministic upper bound that holds uniformly over adjacent datasets. Thus the boosted-tree experiments should be interpreted as stress-test diagnostics for the spillover mechanism; the formal guarantee in Theorem 2 applies when the empirical proxy is replaced by a certified stability bound.

Table 4: Theorem-proxy and sampled audit scales over 100 repetitions.

setting	move	audit Δ	floored Δ	proxy $\widehat{\Delta}$	proxy/floored	floor rate
0.01	40.18	52.23	54.70	13300	367.7	0.23
0.03	39.72	51.63	54.13	13300	368.3	0.26
0.10	39.27	51.05	53.55	13100	366.3	0.23
0.30	37.39	48.60	51.11	12700	364.2	0.24
1.00	33.94	44.12	46.55	11400	343.5	0.25
3.00	25.41	33.04	36.15	6900	237.0	0.41
10.0	17.54	22.80	27.76	3600	140.1	0.57
30.0	12.94	16.82	23.90	2100	89.45	0.77
100	5.15	6.70	22.00	719.0	32.68	1.00
0.20	2.61	3.39	22.00	26000	1200	1.00
0.25	2.12	2.76	22.00	17900	811.5	1.00
0.35	1.41	1.83	22.00	10800	490.9	1.00
0.50	1.01	1.31	22.00	5300	239.3	1.00
0.65	0.73	0.95	22.00	3400	154.2	1.00
0.75	0.56	0.73	22.00	1700	77.24	1.00
0.80	0.51	0.66	22.00	1300	58.91	1.00

Table 5: Certified ERM positive-control diagnostic over 100 repetitions. Appendix G.2 states column definitions.

$\rho^\star_m$	emp. ρ	$\rho^\star_m$/emp. ρ	cert. Δ	move/cert.	logLR/ε	cover	regret	nuis. RMSE
0.0191	7.49×10^{-4}	26.87	28.61	1.25×10^{-5}	4.88×10^{-6}	1.00	0.0146	0.0423

G.4 Certified ERM Positive Control

Table 5 reports the deterministic ERM certificate described in Appendix G.1. The experiment uses ACIC covariates with public clipping and rescaling, $d = 20$, $m = n = 1500$, $\lambda_\mu = \lambda_e = 50$, $\zeta = 0.10$, and 100 repetitions. Unlike the boosted-tree audits above, the calibration scale is not the sampled adjacent-pair movement and is not obtained by plugging an empirical proxy into the theorem's formula. It is the public deterministic scale $\Delta_{\text{cert}} = \max\{2B_\psi, n\rho^\star_m\}$ computed from the bounded-design ERM constants.

The results verify the intended positive-control behavior. The public stability radius is conservative but non-vacuous for this purpose: $\rho^\star_m = 0.0191$ is about 27 times the sampled prediction-movement proxy, and the resulting certified scale is 28.61. All audited adjacent movements are covered by this scale. The mean observed movement divided by the certified scale is 1.25×10^{-5}, and the maximum over repetitions is 2.36×10^{-5}. The empirical likelihood-ratio audit is also far below the privacy target, with mean logLR/$\varepsilon = 4.88 \times 10^{-6}$ and maximum 9.57×10^{-6}. At the same time, the certified selection mechanism retains useful statistical behavior: the average nuisance RMSE is 0.0423 and the expected regret under the certified scale is 0.0146.

In summary, the main experiments demonstrate why empirical DR accuracy and sampled audit movement are distinct, Section 4.3 gives a sufficient public stability condition, and this positive control instantiates that condition with a deterministic strongly convex ERM learner on the same ACIC covariate design.