PROXIMAL POLICY LEARNING WITH NONUNIQUE OUTCOME CONFOUNDING BRIDGES

Anonymous authors
Paper under double-blind review

ABSTRACT

Learning treatment policies from observational data with unmeasured confounding requires valid comparisons of candidate policies. Proximal causal inference uses proxy variables and outcome confounding bridges for this purpose, but the observed bridge equation can admit multiple solutions. Existing results on nonunique bridges show that some causal functionals remain identified, yet they do not resolve policy choice when admissible bridges imply different rankings. In this paper, we characterize when policy value contrasts are invariant over the observed bridge solution set and, when they are not, formulate minimax regret over the joint bridge solution set. We derive an observed-moment dual and a convex empirical program for randomized mixtures of a fixed finite collection of policies. We establish a finite-sample excess-regret bound and give conditions under which a finite dual radius exactly represents the population objective. Under finite-feature rank and interior conditions, this exact representation yields root-n excess worst-case regret at zero bridge residual tolerance for a fixed finite policy hull when the global optimization error is root-n. We also relate worst-case regret over observed bridge solutions to worst-case regret over compatible causal models. Controlled experiments show the decision effects of selecting one bridge and relaxing joint restrictions, and quantify the tradeoff between minimizing worst-case regret and maximizing worst-case policy value.

1 INTRODUCTION

Learning whom to treat is a central decision problem in causal inference. Policy learning addresses this problem by choosing a treatment policy that maps observed characteristics available at deployment to treatment decisions, with the goal of maximizing welfare over a prespecified policy class (Kitagawa & Tetenov, 2018; Athey & Wager, 2021). In observational studies, valid policy comparisons require a causal identification strategy because unmeasured factors that affect both treatment and outcomes can invalidate comparisons based only on recorded deployment covariates. Proximal causal inference addresses unmeasured confounding using proxy variables and confounding bridge functions (Miao et al., 2018). Under suitable causal and proxy assumptions, an outcome confounding bridge satisfies an observed conditional moment equation and represents policy values. This representation underlies kernel proximal estimators, conditional treatment effect estimators, and proximal treatment regime learners (Mastouri et al., 2021; Sverdrup & Cui, 2023; Qi et al., 2024). The observed bridge equation, however, may admit multiple solutions.

Existing work has shown that bridge uniqueness is not always required for causal identification. Some causal functionals can remain identified under nonunique bridges (Zhang et al., 2023), and a linear functional can be strongly identified even when the function solving a conditional moment restriction is weakly or nonuniquely identified (Bennett et al., 2022). A separate literature studies treatment choice under partial identification using minimax regret and related decision criteria (Manski, 2004; Stoye, 2009; Cui, 2021; Pu & Zhang, 2021; d'Adamo, 2021). These results leave an unresolved problem for proximal policy learning. Policy decisions depend on value contrasts among candidate policies, and admissible bridges can agree on some contrasts while disagreeing on others. When such disagreement changes policy rankings, identifying other functionals does not determine which policy to choose. In addition, restrictions on the bridge class can link several policy comparisons, so the feasible comparisons cannot generally be recovered from separate marginal intervals.

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

Policy learning with nonunique bridges therefore has two requirements. First, one must determine whether a policy value contrast is invariant over the observed bridge solution set. Different bridges may give the same contrast, change its magnitude while preserving its sign, or reverse the ranking of two policies. Second, when relevant rankings remain unresolved, the decision criterion must compare policies over admissible bridges while retaining the restrictions that a common bridge imposes on all comparisons. Shared coefficients or a global norm bound, for example, can link policy contrasts across covariate values. Replacing the resulting joint set by the Cartesian product of marginal contrast intervals can introduce combinations that no single bridge satisfies and can change the preferred policy. This leads to the central question of this paper: *When is a policy value contrast invariant over the observed bridge solution set, and how should a policy be learned when relevant contrasts are not invariant?*

In this paper, we address this question by first characterizing exactly when a policy value contrast is invariant over the bridge solution set. When relevant contrasts are not invariant, we define minimax regret over the joint bridge solution set, with the candidate and comparator policies evaluated under the same bridge. We derive an observed-moment dual and a convex empirical program for randomized mixtures of a fixed finite collection of policies. We establish a finite-sample excess-regret bound and give conditions under which a finite dual radius exactly represents the population objective. Under finite-feature rank and interior conditions, the resulting learner achieves $O_P(n^{-1/2})$ excess worst-case regret at zero bridge residual tolerance when the global optimization error is $O_P(n^{-1/2})$, even when admissible bridges reverse policy rankings. We also show that policy contrasts generated by observed bridge solutions can form a wider range than those attainable by compatible latent causal models and bound the resulting gap between bridge-based and causal worst-case regret.

Our main **contributions** are threefold. **(i)** We characterize when a policy value contrast is invariant over the observed bridge solution set and formulate minimax regret over the joint bridge solution set when relevant contrasts are not invariant. Evaluating candidate and comparator policies under the same bridge preserves restrictions across covariate values. **(ii)** We derive an observed-moment dual and a convex empirical learner for randomized mixtures of a fixed finite collection of policies. We establish a finite-sample excess-regret bound and an explicit sufficient dual radius for exact population representation. Under the conditions of Theorem 3, these results yield $O_P(n^{-1/2})$ excess worst-case regret at zero residual tolerance when the global empirical optimization error is $O_P(n^{-1/2})$. **(iii)** We characterize the relation between regret over observed bridge solutions and worst-case causal regret over compatible latent models, including the gap that can arise when the observed bridge contrast range is wider. Controlled experiments examine the consequences of selecting one bridge, replacing joint restrictions by marginal intervals, and choosing between minimax regret and maximin policy value.

2 Related Work

Proximal learning and nonunique bridges. Proxy-variable identification connects observed data to causal effects under restrictions on latent confounding (Miao et al., 2018). Kernel proximal estimation (Mastouri et al., 2021), conditional treatment effect learning (Sverdrup & Cui, 2023), and proximal treatment regime learning (Qi et al., 2024) develop this approach for estimation and decisions. Qi et al. (2024) allow policy classes using different covariate and proxy information. Our policies use only deployment covariates X. Zhang et al. (2023) establish proximal inference without bridge uniqueness, and Bennett et al. (2022) analyze strongly identified functionals of weakly identified functions. Ghassami et al. (2023) derive proxy-based bounds on causal effects without point identification of the bridge function. These results on nonunique bridges concern functionals or bounds when a bridge is not point identified. We study the policy decision when admissible bridges can reverse a policy comparison. Candidate and comparator are evaluated under the same bridge, with the joint bridge restrictions retained across covariate groups.

Treatment choice under uncertainty. Empirical welfare maximization and observational policy learning connect estimation error to policy regret (Kitagawa & Tetenov, 2018; Athey & Wager, 2021). Minimax regret also has a long treatment-choice history (Manski, 2004; Stoye, 2009). Under partial identification, it appears in decision analysis (Cui, 2021), instrumental-variable policy learning (Pu & Zhang, 2021), orthogonal policy learning (d'Adamo, 2021), choice among new treatments (Higbee, 2025), and conditional linear programming (Ben-Michael, 2025). Remark 2

of d'Adamo (2021) already defines regret against the best policy in the same class under a common effect function. Its learning analysis uses rectangular identified sets across covariates. We retain the cross-covariate restrictions imposed by a prespecified bridge class.

Conditional moments and robust policy objectives. Minimax conditional-moment estimation provides empirical norm penalties and function-class tools (Dikkala et al., 2020). Kernel conditional moments also support sensitivity-based policy learning (Ishikawa & He, 2023). Related robust policy methods use hidden-confounding sensitivity models (Kallus & Zhou, 2018) or distributional uncertainty (Si et al., 2020; Kallus et al., 2022). Conditional linear programming in Ben-Michael (2025) keeps the constraint matrix fixed while estimating its objective and right-hand side. We also estimate the bridge operator, allowing it to be rank deficient. Our analysis connects a bounded empirical dual to the population regret objective and gives sufficient conditions for exact representation of every relevant policy contrast (Theorem 3).

3 Problem Setup

3.1 Observed data and outcome confounding bridges

We observe n independent copies of $O = (Y, A, Z, W, X)$, where Y is the outcome, $A \in \{0, 1\}$ is treatment, $X \in \mathcal{X}$ contains deployment covariates, and Z, W are treatment-inducing and outcome-inducing proxies. A nonempty prespecified class Π contains measurable policies $\pi : \mathcal{X} \to [0, 1]$. Deployment treats with probability $\pi(X)$. For a latent causal model M, define

$$V_M(\pi) = \mathbb{E}_M[\pi(X)Y(1) + (1 - \pi(X))Y(0)], \qquad \mathrm{Reg}_M(\pi) = \sup_{\pi' \in \Pi} V_M(\pi') - V_M(\pi).$$

The target population is the observed population, and the comparator belongs to the same policy class.

Assumption 1 (Causal and proxy validity)**.** *There is a latent variable U such that $Y = Y(A)$, $Y(a) \perp\!\!\!\perp A \mid (U, X)$, and $0 < P(A = a \mid U, X) < 1$ for $a = 0, 1$. The proxies satisfy $Y \perp\!\!\!\perp Z \mid (A, U, X)$ and $W \perp\!\!\!\perp (A, Z) \mid (U, X)$. An integrable outcome confounding bridge $h^\dagger = (h_0^\dagger, h_1^\dagger)$ satisfies $\mathbb{E}[Y \mid A = a, U, X] = \mathbb{E}[h_a^\dagger(W, X) \mid A = a, U, X]$. All policy values below are finite.*

Let $\mathsf{S} = (A, Z, X)$, $\mathbb{G} = L_2(P_\mathsf{S})$, and let $\mathbb{H}$ be a real separable Hilbert space of bridge pairs. The bounded operator $Th = \mathbb{E}[h_A(W, X) \mid \mathsf{S}]$ and $m = \mathbb{E}[Y \mid \mathsf{S}] \in \mathbb{G}$ define the observed bridge equation $Th = m$. Write T^* for its adjoint. The prespecified bridge class $\mathcal{H}_B \subset \mathbb{H}$ is nonempty, closed, convex, contained in the radius-B ball, and contains $h^\dagger$. Write

$$L_\pi(h) = \mathbb{E}[\pi(X)h_1(W, X) + (1 - \pi(X))h_0(W, X)], \qquad L_{\pi'}(h) - L_\pi(h) = \langle r_{\pi' - \pi}, h \rangle.$$

Here $\Pi - \Pi = \{\pi' - \pi : \pi, \pi' \in \Pi\}$, and, for $\delta \in \Pi - \Pi$, $r_\delta \in \mathbb{H}$ is the Riesz representative of the continuous linear functional $h \mapsto \mathbb{E}[\delta(X)\{h_1(W, X) - h_0(W, X)\}]$. For the model M under consideration, write $h_M^\dagger$ for its causal bridge. Assumption 1 gives $Th_M^\dagger = m$ and $V_M(\pi) = L_\pi(h_M^\dagger)$ for every $\pi \in \Pi$. We suppress the model subscript when working within one fixed model. Appendix C.1 gives the derivation.

3.2 Policy comparisons and minimax regret

For residual tolerance $\rho \geq 0$, define

$$\begin{aligned} \mathcal{C}_\rho &= \{h \in \mathcal{H}_B : \|Th - m\|_\mathbb{G} \leq \rho\}, \quad U_\rho(r) = \sup_{h \in \mathcal{C}_\rho} \langle r, h \rangle, \\ R_\rho(\pi) &= \sup_{\pi' \in \Pi} U_\rho(r_{\pi' - \pi}), \qquad\qquad R_\rho^* = \inf_{\pi \in \Pi} R_\rho(\pi). \end{aligned} \tag{1}$$

The exact bridge solution set is $\mathcal{C}_0$. Its support function U_0 gives the largest feasible value of a policy contrast. The objective R_ρ is worst-case regret over $\mathcal{C}_\rho$, since

$$R_\rho(\pi) = \sup_{h \in \mathcal{C}_\rho} \left\{ \sup_{\pi' \in \Pi} L_{\pi'}(h) - L_\pi(h) \right\}.$$

A minimax-regret policy minimizes this objective. Its minimum R_ρ^* can be positive.

When does nonuniqueness affect a comparison? A unique bridge reduces the problem to maximizing its policy value. For nonunique bridges, only differences between feasible solutions can change a comparison. The following result applies the null-space criterion for invariant linear functionals (Zhang et al., 2023, Proposition 2.3) to policy contrasts over $\mathcal{C}_0$. If all contrasts are invariant, Assumption 1 makes $R_0(\pi)$ the usual causal regret within Π.

Proposition 1 (Invariant policy contrasts with nonunique bridges)**.** *Let $\mathcal{D}_0 = \operatorname{cl}\operatorname{span}\{h - h' : h, h' \in \mathcal{C}_0\}$. A policy value contrast is constant over $\mathcal{C}_0$ if and only if $r_\delta \perp \mathcal{D}_0$. If $\mathcal{H}_B$ is a Hilbert ball containing an exact bridge strictly inside it, then $\mathcal{D}_0 = \ker T$ and invariance is equivalent to $r_\delta \in \operatorname{cl}\operatorname{range}(T^*)$. If the stronger equation $T^* f = r_\delta$ has a solution $f \in \mathbb{G}$, the invariant contrast equals $\langle f, m\rangle$.*

The relevant contrast interval is

$$[-U_0(-r_{\pi'-\pi}),\ U_0(r_{\pi'-\pi})]\,. \tag{2}$$

Constancy makes this interval a single point. Preservation of its sign is weaker: a nonconstant interval may still rank two policies in the same order. Rankings reverse across bridges exactly when the lower endpoint is negative and the upper endpoint is positive. A common optimal policy may still exist if reversals involve only other policies. Thus contrast identification, pairwise ordering, and a common optimum are distinct properties. Appendix C.1 gives the proof of the proposition.

Joint restrictions and alternative criteria. The conditional moment equation acts at each covariate value, while $\mathcal{H}_B$ can link policy contrasts across covariate values through shared coefficients or a global norm bound. Replacing the resulting joint contrast set by the Cartesian product of its marginal intervals removes these restrictions and can change the minimax-regret policy. A policy based on a selected bridge maximizes $L_\pi(\widetilde{h})$ for one solution $\widetilde{h}$. A maximin value rule maximizes $\inf_{h\in\mathcal{C}_\rho} L_\pi(h)$. Maximin value protects the worst-case policy value, while minimax regret measures loss relative to the best policy under each bridge. A bridge-dependent baseline added to every policy value cancels from regret but can change maximin value.

Relation to causal regret. Let $\mathfrak{M}(P, \mathcal{H}_B)$ contain the latent models inducing the observed law P and satisfying Assumption 1 with a causal bridge in $\mathcal{H}_B$. Define

$$R_{\mathrm{ca}}(\pi) = \sup_{M\in\mathfrak{M}(P,\mathcal{H}_B)} \mathrm{Reg}_M(\pi), \qquad R^*_{\mathrm{ca}} = \inf_{\pi\in\Pi} R_{\mathrm{ca}}(\pi). \tag{3}$$

Every compatible model supplies an exact bridge, so $R_{\mathrm{ca}} \le R_0 \le R_\rho$. The bridge objective therefore bounds causal regret from above. Its contrast range can be wider than the causally attainable range: Appendix A.3 gives an observed law with causal effect one and a bridge contrast of $5/4$. Proposition 4 bounds the resulting regret gap. The bridge class, tolerance, and policy class specify the target (Appendix A.5).

4 Method

The worst-case comparison in Equation (1) is an optimization over solutions of a conditional moment equation. We express this comparison through observed moments, then learn a policy by optimizing the resulting empirical criterion.

4.1 A dual representation of policy comparisons

For $\mathcal{C}_0 \neq \varnothing$, every $r \in \mathbb{H}$ and $\rho \ge 0$ satisfy

$$U_\rho(r) = \inf_{f\in\mathbb{G}} \{\langle f, m\rangle + \rho\|f\|_{\mathbb{G}} + \sigma_{\mathcal{H}_B}(r - T^* f)\}\,, \tag{4}$$

where $\sigma_{\mathcal{H}_B}(u) = \sup_{h\in\mathcal{H}_B}\langle u, h\rangle$ is the support function of the bridge class. For a radius-B ball, the last term is $B\|r - T^* f\|_{\mathbb{H}}$. The first term equals $\mathbb{E}[f(\mathsf{S})Y]$ when $Y \in L_2$. The support penalty evaluates the unresolved residual $r - T^* f$, allowing policy comparisons to be represented even without an exact adjoint solution. Appendix C.3 proves Equation (4) at both zero and positive tolerance, and shows $U_\rho(r) \downarrow U_0(r)$ as $\rho \downarrow 0$ for each fixed r.

4.2 FINITE FEATURES AND EMPIRICAL POLICY LEARNING

Prespecify arm-specific bridge features $\phi_a(W, X) \in \mathbb{R}^{d_a}$ for $a \in \{0, 1\}$ and test features $\psi(\mathsf{S}) \in \mathbb{R}^q$, with $\|\phi_a\| \le \kappa_H$ and $\|\psi\| \le 1$. Set $h_{\theta,a} = \theta_a^\top \phi_a$, $\theta = (\theta_0^\top, \theta_1^\top)^\top \in \mathbb{R}^{d_0+d_1}$, and $f_\beta = \beta^\top \psi$. In the finite-feature analysis, $\mathcal{H}_B = \{h_\theta : \|\theta\| \le B\}$, with the direct-sum coefficient norm as Hilbert norm after removal of redundant bridge coordinates. Restrict dual coefficients to $\|\beta\| \le \Lambda$. Define $\xi = ((1-A)\phi_0^\top, A\phi_1^\top)^\top$ and

$$M_T = \mathbb{E}[\xi\psi^\top], \quad v = \mathbb{E}[\psi Y], \quad \Gamma = \mathbb{E}[\psi\psi^\top], \quad r_\delta = \mathbb{E}[(-\delta\phi_0^\top, \delta\phi_1^\top)^\top], \qquad \delta \in \Pi - \Pi.$$

Thus $M_T \in \mathbb{R}^{(d_0+d_1)\times q}$ contains the bridge and test moments, v contains outcome moments, and Γ measures test-function norms. Suppressing ρ, Λ in the notation, the restricted population criterion is

$$S(r) = \inf_{\|\beta\| \le \Lambda} \{\beta^\top v + \rho\sqrt{\beta^\top \Gamma \beta} + B\|r - M_T\beta\|\}, \tag{5}$$
$$D(\pi) = \sup_{\pi' \in \Pi} S(r_{\pi'-\pi}).$$

Training-sample averages give $\widehat{M}_T, \widehat{v}, \widehat{\Gamma}, \widehat{r}_\delta$ and the empirical objective $\widehat{D}$. The radius Λ controls the approximation of the dual while leaving $\mathcal{C}_\rho$ unchanged.

For a finite hull $\Pi = \mathrm{conv}\{p_1, \ldots, p_J\}$, let $\pi_w = \sum_k w_k p_k$ for $w \in \Delta_J = \{w \ge 0 : \sum_k w_k = 1\}$. Convexity of the support criterion places a worst comparator at a vertex. Define

$$\widehat{g}_j(w, \beta) = \widehat{v}^\top \beta + \rho\|\widehat{\Gamma}^{1/2}\beta\| + B\|\widehat{r}_{p_j - \pi_w} - \widehat{M}_T\beta\|.$$

The empirical learner solves

$$\min_{w \in \Delta_J,\, t,\, \beta_1, \ldots, \beta_J} t \quad \text{subject to} \quad \|\beta_j\| \le \Lambda, \quad \widehat{g}_j(w, \beta_j) \le t \quad (j = 1, \ldots, J). \tag{6}$$

This is a convex program. Its decision variables are the policy weights w and one dual vector β_j for each comparator. Each vector evaluates a different policy comparison under the common bridge restrictions. After computing the sample moments, the program optimizes these variables jointly and returns $\widehat{\pi}(x) = \sum_j \widehat{w}_j p_j(x)$. Deployment requires only the prespecified policies and their learned weights. Appendix B derives the program and gives its implementation in Algorithm 1.

A moment-residual variant. A coefficient-norm variant uses $\rho\|\beta\|_2$ in place of $\rho\sqrt{\beta^\top\Gamma\beta}$. This corresponds to

$$\mathcal{C}_\rho^{\mathrm{mom}} = \{h_\theta : \|\theta\| \le B,\ \|M_T^\top\theta - v\|_2 \le \rho\}.$$

When the test moments recover the full bridge equation, $\mathcal{C}_0^{\mathrm{mom}} = \mathcal{C}_0$. At positive tolerance the sets can differ. All fitted comparisons below use $\rho = 0$. Appendix C.4 provides the corresponding dual and error bound.

5 THEORETICAL ANALYSIS

The empirical procedure minimizes $\widehat{D}$, while the target is R_ρ. We first control the difference between sample and population dual objectives. We then determine when the finite dual represents minimax regret over the joint bridge solution set.

5.1 FINITE-SAMPLE EXCESS REGRET

Assume $|Y| \le M_Y$ and the fixed-class measurability conditions in Appendix C.5. For failure probability $\alpha \in (0, 1)$, choose positive α_m, α_Γ with $\alpha_m + \alpha_\Gamma = \alpha$. Define

$$E_n = e_n(\alpha_m) + \rho\Lambda\left\{\frac{1 + \sqrt{2\log(1/\alpha_\Gamma)}}{\sqrt{n}}\right\}^{1/2}.$$

Here e_n controls the outcome moments, bridge and test moments, and policy contrast moments using their empirical Rademacher complexities, as defined in Equation (22). The second term controls estimation of the test-function norm. For fixed finite features, policy hull, radii, and confidence level, $e_n = O_P(n^{-1/2})$.

Theorem 2 (Uniform estimation and excess regret)**.** *With probability at least* $1-\alpha$*, simultaneously for* $\pi \in \Pi$*,* $|\widehat{D}(\pi) - D(\pi)| \le E_n$*. If* $\widehat{D}(\widehat{\pi}) \le \inf_{\pi\in\Pi} \widehat{D}(\pi) + e_{\mathrm{opt}}$*, then*

$$R_\rho(\widehat{\pi}) - R_\rho^* \le c + 2E_n + e_{\mathrm{opt}}, \qquad c = \inf_{\pi\in\Pi} D(\pi) - R_\rho^* \ge 0. \tag{7}$$

The result allows singular Γ *and uses the same sample to select the policy and dual coefficients.*

Appendix C.6 proves the theorem. Equation (7) separates population representation error c, sampling error E_n, and global empirical optimization error e_{opt}. Because the bound is uniform over Π, the same sample may be used to select $\widehat{\pi}$. At $\rho = 0$, the norm-estimation term disappears. At fixed positive ρ, its square-root dependence gives an $n^{-1/4}$ contribution without a lower eigenvalue condition on Γ. The next result controls c and states when the finite dual represents the exact bridge target.

5.2 Exact representation with a finite dual radius

Theorem 3 (Finite-radius representation with nonunique bridges)**.** *In the finite-feature setting, assume* $M_T^\top \theta = v$ *if and only if* $Th_\theta = m$*. Let* M_T *have positive rank and smallest nonzero singular value* $s_* > 0$*. Let* $\theta_{\min}$ *be the minimum-norm solution of* $M_T^\top \theta = v$*, with* $\|\theta_{\min}\| < B$*, and set*

$$a = \sqrt{B^2 - \|\theta_{\min}\|^2}, \quad N = \ker(M_T^\top), \quad K_\Pi = \sup_{\delta\in\Pi-\Pi} \|r_\delta\| < \infty, \quad \Lambda_0 = \frac{BK_\Pi}{s_* a}.$$

If $\Lambda \ge \Lambda_0$*, then at* $\rho = 0$*, every policy contrast satisfies*

$$S(r_\delta) = U_0(r_\delta) = r_\delta^\top \theta_{\min} + a\|P_N r_\delta\|, \qquad D = R_0, \quad c = 0, \tag{8}$$

where P_N *projects onto* N*. For* $b_\rho = \sup_{\delta\in\Pi-\Pi}\{U_\rho(r_\delta) - U_0(r_\delta)\}$*, the same radius at* $\rho \ge 0$ *gives*

$$R_\rho(\pi) \le D(\pi) \le R_0(\pi) + \rho\Lambda_0, \qquad c \le \rho\Lambda_0, \quad b_\rho \le \rho\Lambda_0. \tag{9}$$

Appendix C.7 proves the theorem. Every exact bridge coefficient is $\theta_{\min} + z$, with $z \in N$ and $\|z\| \le a$. The moments determine the component orthogonal to N. The term $a\|P_N r_\delta\|$ measures the remaining variation in a policy contrast. In particular, the contrast interval has both signs when $|r_\delta^\top \theta_{\min}| < a\|P_N r_\delta\|$.

The finite representation follows from an explicit dual coefficient vector. Write $r_R = P_{N^\perp} r_\delta$ and $r_N = P_N r_\delta$, and let $M_T^\dagger$ denote the Moore–Penrose inverse. The vector

$$\beta_* = M_T^\dagger \left(r_R - \frac{\|r_N\|}{a}\theta_{\min} \right), \qquad \|\beta_*\| \le \frac{B\|r_\delta\|}{s_* a} \le \Lambda_0 \tag{10}$$

has residual $r_\delta - M_T\beta_* = r_N + (\|r_N\|/a)\theta_{\min}$. Since $\theta_{\min} \perp r_N$, substituting this vector into the dual objective gives exactly $r_\delta^\top \theta_{\min} + a\|r_N\|$. When $r_N = 0$, the residual vanishes. When $r_N \ne 0$, the support penalty represents the variation left by the bridge equation. Both cases therefore admit exact finite-dual evaluation. The same appendix gives a boundary example showing why an interior solution matters.

The sufficient radius grows as the population interior margin a or moment strength s_* decreases, while a larger Λ increases the sampling bound E_n. Choosing Λ therefore requires balancing population representation and sampling error.

Combining the two theorems yields

$$R_\rho(\widehat{\pi}) - R_\rho^* \le \rho\Lambda_0 + 2E_n + e_{\mathrm{opt}}. \tag{11}$$

For a fixed distribution, finite features, a finite policy hull, and fixed $\Lambda \ge \Lambda_0$, taking $\rho = 0$ and $e_{\mathrm{opt},n} = O_P(n^{-1/2})$ gives $R_0(\widehat{\pi}) - R_0^* = O_P(n^{-1/2})$ (Corollary 8, Appendix C.8). The null-space contrast may remain nonzero as n grows. Thus the rate concerns learning the optimal worst-case decision under persistent nonuniqueness. At $\rho = 0$ under the same exact-representation conditions, a finite policy class with regret gap $\Delta > 0$ has an optimal empirical choice whenever $2E_n + e_{\mathrm{opt}} < \Delta$. Appendix C.9 gives the corresponding bound for regret over compatible causal models.

The corresponding causal guarantee. To compare the learned objective with compatible causal models, let $\Delta_M(\delta) = \mathbb{E}_M[\delta(X)\{Y(1) - Y(0)\}]$ and define

$$g_{\mathrm{bc}} = \sup_{\delta \in \Pi - \Pi} \left\{ U_0(r_\delta) - \sup_{M \in \mathfrak{M}(P, \mathcal{H}_B)} \Delta_M(\delta) \right\}.$$

Proposition 4 gives $R_{\mathrm{ca}} \leq R_0 \leq R_{\mathrm{ca}} + g_{\mathrm{bc}}$. Under the exact-dual conditions at $\rho = 0$, the same learning event therefore implies

$$R_{\mathrm{ca}}(\widehat{\pi}) - R^*_{\mathrm{ca}} \leq g_{\mathrm{bc}} + 2E_n + e_{\mathrm{opt}}. \tag{12}$$

The additional term measures a property of the model classes. It vanishes when their worst-case policy contrasts agree, for example when every admissible bridge represents all policy values in a compatible causal model. Larger samples reduce the estimation terms but leave this population discrepancy unchanged. This motivates evaluating both excess worst-case regret and causal loss under the generating model in the experiments.

6 Experiments

We test the consequences of selecting one bridge, discarding joint restrictions, and changing the decision criterion. Each comparison changes one choice while keeping the observed sample, bridge class, and policy class fixed. A separate learning curve examines estimation of the population minimax-regret target. Appendix D gives the generating laws and fitting details.

6.1 Experimental Setup

Data and common information. Table 1 reports synthetic and semi-synthetic experiments with $n = 4{,}000$ and 200 attempted paired datasets per comparison. Each comparison has its own declared generating law. The semi-synthetic experiments resample 4,802 covariate rows from the 2016 Atlantic Causal Inference Conference (ACIC) and simulate treatment, proxies, and outcomes with hidden confounding. A prespecified nonlinear score defines two covariate groups. Within each comparison, both methods use the same observations and policy class. Training and evaluation use $\rho = 0$. The synthetic bridge-selection and synthetic criterion designs use constant treatment probabilities. The synthetic joint-restriction design and all ACIC comparisons use one probability per group. Dual radii are fixed from population representation bounds before sampling (Appendix D.4). Sample-dependent fitting then uses observed data only.

Selecting a bridge. We compare minimax regret over the full empirical bridge set with selecting its minimum-norm bridge and maximizing that bridge's policy value. Both rules use the same equality constraints and coefficient ball.

Keeping joint restrictions. The two population group contrasts have the form $(c_- + \zeta, c_+ + \zeta)$, where c_j is one tenth of the group's mean covariate score and ζ is a common unknown shift. A bridge evaluates both covariate groups simultaneously. The rectangular comparator uses the Cartesian product of the two marginal contrast intervals. We change only this restriction and evaluate both learned policies on the original joint set.

Choosing the criterion. We compare minimax regret and maximin value on the same empirical bridge set. Every population bridge agrees on the value of standard care ($A = 0$), while the bridge-implied treatment contrast can have either sign. Minimax regret measures loss relative to the best policy under each bridge. Maximin value maximizes the worst-case policy value.

Evaluation. The first two comparisons report excess bridge regret $\mathcal{E}_0(\pi) = R_0(\pi) - \min_{\pi' \in \Pi} R_0(\pi')$. The criterion comparison reports $R_0(\pi)$ and worst-case policy value $W_0(\pi) = \inf_{h \in \mathcal{C}_0} L_\pi(h)$. Lower $\mathcal{E}_0$ and R_0, and higher W_0, are better. We also report generating-world causal loss $C(\pi) = \sup_{\pi' \in \Pi} V_{\mathrm{gen}}(\pi') - V_{\mathrm{gen}}(\pi)$, where V_{gen} is causal policy value under the simulated model. Population evaluation uses exact moments under the declared law. Appendix D.1 defines paired uncertainty, and Appendix E.2 gives supplementary statistics. The two-world learning experiment uses 30 datasets per sample size.

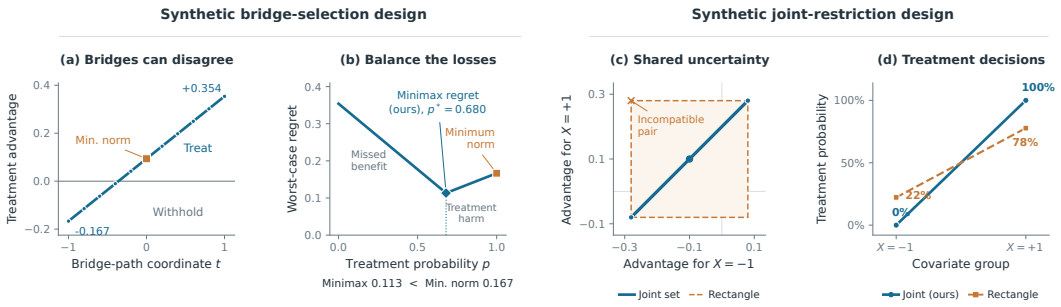


Figure 1: **Population comparisons in the two synthetic designs.** Panels (a,b) show bridge selection. Panels (c,d) show joint restrictions, where a common shift couples the group contrasts and the rectangle admits incompatible pairs.

6.2 Main Results

Table 1: **Policy performance in three controlled comparisons at $n = 4{,}000$.** Means are in 10^{-3} units over paired fits with nonempty empirical bridge sets. The last two column groups report two objectives for the same criterion comparison. MR is minimax regret, Min. norm selects the minimum-norm bridge, Rectangle uses marginal contrast intervals, and MV is maximin value. Arrows give the preferred direction, and bold marks the better mean within each pair.

	Bridge selection $\mathcal{E}_0 \downarrow$		Joint restrictions $\mathcal{E}_0 \downarrow$		Criterion $R_0 \downarrow$		Criterion $W_0 \uparrow$	
Data	MR	Min. norm	MR	Rectangle	MR	MV	MR	MV
Synthetic	**7.07**	53.35	**13.71**	28.85	**63.74**	253.94	442.52	**500.00**
ACIC*	**14.26**	56.07	**20.53**	22.59	**66.30**	208.00	436.39	**483.90**

*Of 200 attempted datasets per comparison, ACIC gives 198 valid bridge-selection pairs, 183 joint-restriction pairs and 194 criterion pairs. The remaining 2, 17 and 6 attempted datasets yield an empty empirical bridge set. Each synthetic comparison has 200 valid pairs.

Minimax regret lowers excess regret relative to minimum-norm bridge selection. Table 1 compares the fitted policies. In the synthetic bridge-selection design, admissible bridges give treatment contrasts from $-.167$ to $.354$, as shown in Figure 1(a). The minimum-norm bridge always treats. Population minimax regret treats with probability $.680$ and reduces worst-case regret from $.167$ to $.113$ in panel (b). Fitted excess regret is 7.07×10^{-3}, compared with 53.35×10^{-3} for minimum-norm selection. With ACIC covariates, the paired difference in excess regret, MR minus minimum norm, has 95% interval $[-43.99, -39.63] \times 10^{-3}$. Appendix E.1 reports other proxy strengths.

Retaining joint restrictions reduces worst-case policy regret. In the synthetic joint-restriction design, the two covariate groups have contrasts $-.1+\zeta$ and $.1+\zeta$. Their difference remains $.2$ even when either sign is uncertain. Figure 1(c) displays this restriction. Joint minimax regret treats only the positive-score group at the population optimum. The rectangle permits incompatible shifts and chooses probabilities $(.222, .778)$ for the negative- and positive-score groups (panel d). For the fitted policies in Table 1, the paired excess-regret intervals (MR minus rectangle) are $[-16.22, -14.06] \times 10^{-3}$ for synthetic data and $[-2.85, -1.27] \times 10^{-3}$ for ACIC covariates. Appendix E.4 examines an additional population bridge class with group interactions.

Minimax regret and maximin value optimize different criteria. In the synthetic criterion design, standard care has value $1/2$ under every bridge. The bridge-implied treatment contrast ranges from $-.07377$ to $.25394$. Population maximin value withholds treatment and keeps the $1/2$ floor. Minimax regret treats with probability $.77489$, balancing treatment harm against forgone benefit (Appendix D.3). Table 1 shows both consequences after fitting: MR has lower R_0 and lower W_0. The ACIC criterion comparison has the same tradeoff, with paired intervals (MR minus MV)

Table 2: **Generating-world causal loss $C \downarrow$.** The fitted policies and valid pairs are the same as in Table 1. Means are in 10^{-3} units, and bold marks the lower mean within each pair.

	Bridge selection		Joint restrictions		Decision criterion	
Data	MR	Min. norm	MR	Rectangle	MR	MV
Synthetic	**22.73**	33.33	**4.89**	20.66	**11.04**	50.00
ACIC	**23.41**	32.73	**3.65**	6.33	**6.56**	39.08

$[-152.42, -130.97] \times 10^{-3}$ for R_0 and $[-50.69, -44.33] \times 10^{-3}$ for W_0. Regret thus limits opportunity loss at the cost of lower worst-case policy value.

Causal performance in the generating worlds. Table 2 evaluates the fitted policies in their generating causal worlds. MR has lower mean causal loss in all six comparisons. In the criterion comparison, C falls from 50.00 to 11.04 in the synthetic design and from 39.08 to 6.56 with ACIC covariates, in 10^{-3} units. Table 1 simultaneously shows a lower W_0 for MR. Thus, in these generating models, lower causal loss is accompanied by a lower worst-case bridge value. This is consistent with the objectives: minimax regret controls opportunity loss under each bridge, while maximin value maximizes the worst-case policy value.

Larger samples approach the population target. In the two-world design, the population dual with $(B, \rho, \Lambda) = (3, 0, 20)$ equals R_0 for every constant policy. Figure 2 shows how population excess regret $\mathcal{E}_0$ changes with training size. Its mean decreases from .09610 at $n = 500$ to .02035 at 20,000 and .00758 at 100,000. The last point has a 95% interval of [.00531, .00985]. Appendix E.3 reports optimization accuracy and empirical feasibility. The population bridge set and its optimal minimax-regret policy remain fixed as n grows. Only the estimated moments change, so the curve measures finite-sample learning of the fixed population target in Corollary 8.

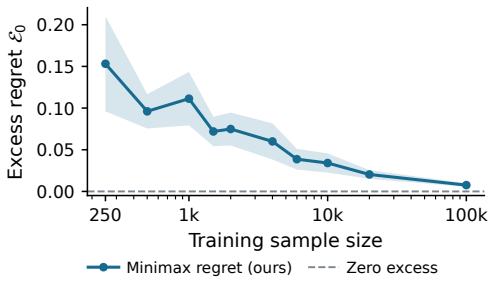


Figure 2: **Population excess regret $\mathcal{E}_0$ versus training size.** Mean over 30 fitted policies with pointwise 95% intervals.

7 Conclusion

We study policy learning when the observed proximal outcome bridge equation admits multiple solutions and policy rankings need not be identified. We characterize when policy value contrasts are invariant and, when they are not, optimize minimax regret over the joint bridge solution set. An observed-moment dual yields a convex learner for finite policy hulls while preserving restrictions shared across policy comparisons. The finite-radius representation and finite-sample bounds show that this target admits root-n excess worst-case regret under the stated rank, interior, and optimization conditions. Experiments isolate the decision consequences of selecting one bridge, relaxing joint restrictions, and replacing minimax regret with maximin value.

Limitations and future work. The theory of this paper prespecifies the bridge features, test features, bridge radius, dual radius, and finite policy hull. An interesting next step is to learn these classes from data while retaining valid regret guarantees. Beyond this setting, a sharper characterization of which observed bridges are attainable by compatible causal models could reduce the gap between bridge and causal regret. Finally, extending the joint-bridge optimization to longitudinal treatment decisions remains an open problem.

AI USE STATEMENT

Generative AI tools assisted with manuscript polishing and paper revision. The authors are responsible for the scientific claims, proofs, and experimental results.

REFERENCES

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

Peter L Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of machine learning research*, 3(Nov):463–482, 2002.

Eli Ben-Michael. Partial identification via conditional linear programs: estimation and policy learning. *arXiv preprint arXiv:2506.12215*, 2025.

Andrew Bennett, Nathan Kallus, Xiaojie Mao, Whitney Newey, Vasilis Syrgkanis, and Masatoshi Uehara. Inference on strongly identified functionals of weakly identified functions. *arXiv preprint arXiv:2208.08291*, 2022.

Yifan Cui. Individualized decision-making under partial identification: three perspectives, two optimality results, and one paradox. *Harvard Data Science Review*, 3(3):1–19, 2021.

Riccardo d'Adamo. Orthogonal policy learning under ambiguity. *arXiv preprint arXiv:2111.10904*, 2021.

Nishanth Dikkala, Greg Lewis, Lester Mackey, and Vasilis Syrgkanis. Minimax estimation of conditional moment models. *Advances in Neural Information Processing Systems*, 33:12248–12262, 2020.

AmirEmad Ghassami, Ilya Shpitser, and Eric Tchetgen Tchetgen. Partial identification of causal effects using proxy variables. *arXiv preprint arXiv:2304.04374*, 2023.

Samuel D Higbee. Policy learning with new treatments. *Quantitative Economics*, 16(4):1409–1456, 2025.

Kei Ishikawa and Niao He. Kernel conditional moment constraints for confounding robust inference. In *International Conference on Artificial Intelligence and Statistics*, pp. 650–674. PMLR, 2023.

Nathan Kallus and Angela Zhou. Confounding-robust policy improvement. *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.

Charles F Manski. Statistical treatment rules for heterogeneous populations. *Econometrica*, 72(4): 1221–1246, 2004.

Afsaneh Mastouri, Yuchen Zhu, Limor Gultchin, Anna Korba, Ricardo Silva, Matt Kusner, Arthur Gretton, and Krikamol Muandet. Proximal causal learning with kernels: Two-stage estimation and moment restriction. In *International conference on machine learning*, pp. 7512–7523. PMLR, 2021.

Wang Miao, Zhi Geng, and Eric J Tchetgen Tchetgen. Identifying causal effects with proxy variables of an unmeasured confounder. *Biometrika*, 105(4):987–993, 2018.

Hongming Pu and Bo Zhang. Estimating optimal treatment rules with an instrumental variable: A partial identification learning approach. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 83(2):318–345, 2021.

Zhengling Qi, Rui Miao, and Xiaoke Zhang. Proximal learning for individualized treatment regimes under unmeasured confounding. *Journal of the American Statistical Association*, 119(546):915–928, 2024.

Nian Si, Fan Zhang, Zhengyuan Zhou, and Jose Blanchet. Distributionally robust policy evaluation and learning in offline contextual bandits. In *International Conference on Machine Learning*, pp. 8884–8894. PMLR, 2020.

Maurice Sion. On general minimax theorems. 1958.

Jörg Stoye. Minimax regret treatment choice with finite samples. *Journal of Econometrics*, 151(1): 70–81, 2009.

Erik Sverdrup and Yifan Cui. Proximal causal learning of conditional average treatment effects. In *International Conference on Machine Learning*, pp. 33285–33298. PMLR, 2023.

Jeffrey Zhang, Wei Li, Wang Miao, and Eric Tchetgen Tchetgen. Proximal causal inference without uniqueness assumptions. *Statistics & probability letters*, 198:109836, 2023.

A POLICY COMPARISONS AND CAUSAL INTERPRETATION

This appendix gives the precise invariance criterion, the relation between regret over bridge solutions and compatible causal models, and two constructions used in the main text.

A.1 EXACT BRIDGES AND INVARIANT CONTRASTS

$$\mathcal{C}_0 = \{h \in \mathcal{H}_B : Th = m\}. \tag{13}$$

Proposition 1 in the main text characterizes invariant contrasts. Appendix C.1 proves the result. Constancy over the bridge solution set is sufficient, but not necessary in general, for point identification across compatible models. Every such model supplies a feasible bridge. The observed law in Appendix A.3 shows why the converse can fail.

The proposition distinguishes constancy of a contrast from preservation of its sign. Only a change of sign reverses the ranking of two policies. In infinite-dimensional spaces, membership in $\mathrm{cl}\,\mathrm{range}(T^*)$ can hold without a finite-norm adjoint solution. This distinguishes an invariant comparison from its representation by an exact adjoint solution with finite $L_2(P_{\mathsf{S}})$ norm.

A.2 THE GAP BETWEEN BRIDGE AND CAUSAL REGRET

For $\delta = \pi' - \pi$, define $\Delta_M(\delta) = V_M(\pi') - V_M(\pi) = \mathbb{E}_M[\delta(X)\{Y(1) - Y(0)\}]$.

Proposition 4 (Relation to compatible causal models)**.** *Assume $\mathfrak{M}(P, \mathcal{H}_B)$ is nonempty and $K_\Pi := \sup_{\delta \in \Pi - \Pi} \|r_\delta\|_{\mathbb{H}} < \infty$. For every $\delta \in \Pi - \Pi$, every compatible causal contrast belongs to the observed bridge contrast range. Define*

$$g_{\mathrm{bc}} = \sup_{\delta \in \Pi - \Pi} \left\{ U_0(r_\delta) - \sup_{M \in \mathfrak{M}(P, \mathcal{H}_B)} \Delta_M(\delta) \right\} \geq 0. \tag{14}$$

Then, for every $\pi \in \Pi$,

$$R_{\mathrm{ca}}(\pi) \leq R_0(\pi) \leq R_{\mathrm{ca}}(\pi) + g_{\mathrm{bc}}, \qquad R^*_{\mathrm{ca}} \leq R^*_0 \leq R^*_{\mathrm{ca}} + g_{\mathrm{bc}}. \tag{15}$$

If $R_0(\widetilde{\pi}) \leq R^*_0 + \varepsilon$, then $R_{\mathrm{ca}}(\widetilde{\pi}) - R^*_{\mathrm{ca}} \leq \varepsilon + g_{\mathrm{bc}}$. Thus $g_{\mathrm{bc}} = 0$ makes the exact R_0 criterion equal to minimax regret over the prespecified causal model class. A sufficient condition is every $h \in \mathcal{C}_0$ is a causal bridge in some compatible model and represents all policies there.

Appendix C.1 proves the proposition. Equality of the relevant support values is enough for $g_{\mathrm{bc}} = 0$. The bridge realizability condition above supplies one way to obtain it. The next example has a strictly wider bridge contrast range.

A.3 AN OBSERVED BRIDGE OUTSIDE THE CAUSAL RANGE

Take constant X, Z, $U \sim \mathrm{Bernoulli}(1/2)$, $W = U$, treatment probabilities $P(A = 1 \mid U = 1) = 3/4$, $P(A = 1 \mid U = 0) = 1/4$, and deterministic potential outcomes $Y(0) = 0, Y(1) = 1$. The bridge $h^\dagger_0 = 0, h^\dagger_1 = 1$ satisfies the causal assumptions.

Every compatible causal model with the observed law $Y = A$ and latent positivity has $Y(0) = 0, Y(1) = 1$ almost surely. Exchangeability transfers each treatment-conditional degenerate distribution to the population conditional on U, and positivity prevents hiding an unobserved outcome stratum. The causal contrast between constant policies is consequently one in every such model.

Yet $\widetilde{h}_0(w) = 1/4 - w$, $\widetilde{h}_1(w) = 1$ also satisfies the observed bridge equation because $P(W = 1 \mid A = 0) = 1/4$. Its population policy contrast is $5/4$. Both bridges lie in a Euclidean radius-2 ball on their four function values. Thus the exact bridge range strictly contains the causally attainable contrasts.

A.4 COMPATIBLE WORLDS WITH OPPOSITE POLICY PREFERENCES

Let $U \sim \mathrm{Bernoulli}(1/2)$ and, conditionally on $U = u$, draw $A, W, Y(0), Y(1)$ independently with probabilities $e_u, w_u, \mu_{0u}, \mu_{1u}$. Independently draw a fair binary Z, set $Y = Y(A)$, and take constant

X. Use

	M_+	M_-
(e_0, e_1)	$(1/5, 4/5)$	$(29/34, 5/34)$
(w_0, w_1)	$(1/5, 4/5)$	$(151/200, 49/200)$
(μ_{00}, μ_{01})	$(1/10, 7/10)$	$(247/250, 127/1450)$
(μ_{10}, μ_{11})	$(3/10, 9/10)$	$(1323/1450, 3/250)$.

The explicit bridge values $(h_a(0), h_a(1))$ for $a = 0, 1$ are $(-1/10, 9/10), (1/10, 11/10)$ under M_+ and $(-2501/7250, 10299/7250), (-3049/7250, 9751/7250)$ under M_-. Substitution in $(1 - w_u)h_a(0) + w_u h_a(1) = \mu_{au}$ verifies bridge validity. Both bridges are in a common Euclidean radius-3 ball. Direct marginalization gives, in both models,

(A, W)	$Y = 0$	$Y = 1$
$(0, 0)$	$147/500$	$23/500$
$(0, 1)$	$12/125$	$8/125$
$(1, 0)$	$8/125$	$12/125$
$(1, 1)$	$23/500$	$147/500$.

The Z law is also identical. All positivity and proxy restrictions hold, but causal effects are τ_+ and $-\tau_-$, where $\tau_+ = 1/5$ and $\tau_- = 274/3625$ are the two positive effect magnitudes. Every learner based on any number of observed samples therefore has the same expected probability q of selecting treatment under both models. Its maximum expected causal regret across the two models, relative to their optimal constant policies, is at least $\max\{\tau_+(1 - q), \tau_- q\}$, whose minimum is $\tau_+\tau_-/(\tau_+ + \tau_-) = 274/4995 > 0$.

The positive lower bound persists with unlimited observed data because the two models have the same observed law and prefer opposite treatments.

A.5 Prespecifying the target and reporting sensitivity

The bridge class $\mathcal{H}_B$ (including its radius B), population residual tolerance ρ, and policy class Π determine which bridges and policy comparisons enter R_ρ. Bridge features also define the decision target when they define $\mathcal{H}_B$. Dual features, the dual radius Λ, and solver tolerance affect how that target is computed.

The containment $h_M^\dagger \in \mathcal{H}_B$ is a modeling assumption, including a bound on components in $\ker T$ that the observed moments leave unresolved. We fix $\mathcal{H}_B$, B, ρ, and Π before drawing outcomes. Appendix E.4 varies the allowed bridge interactions to study how this choice affects the decision target.

B Convex optimization over a finite policy hull

For a finite policy hull $\Pi = \mathrm{conv}\{p_1, \ldots, p_J\}$, write $\pi_w = \sum_j w_j p_j$ with $w_j \geq 0$ and $\sum_j w_j = 1$. The support criterion is convex in each policy contrast. A comparator mixture therefore has support at most the largest candidate support, and

$$\widehat{D}(\pi_w) = \max_{1 \leq j \leq J} \widehat{S}(\widehat{r}_{p_j - \pi_w}).$$

The comparator maximum is attained at a supplied candidate. The learner optimizes over the entire mixture simplex. The finite-feature implementation solves the convex program

$$\begin{aligned} \min_{w, t, \beta_1, \ldots, \beta_J} \quad & t \\ \text{subject to} \quad & \widehat{v}^\top \beta_j + \rho\sqrt{\beta_j^\top \widehat{\Gamma} \beta_j} + B\|\widehat{r}_{p_j - \pi_w} - \widehat{M}_T \beta_j\| \leq t, \quad \|\beta_j\| \leq \Lambda \quad (j = 1, \ldots, J), \\ & w_j \geq 0, \quad \sum_j w_j = 1. \end{aligned}$$

Because each coefficient ball is compact and the objective continuous, the inner infima are attained, so this epigraph formulation represents the empirical criterion exactly. Singular Gram matrices are allowed: a positive semidefinite square root implements the penalty without taking an inverse.

For the coefficient-penalty variant, replace each $\rho\|\widehat{\Gamma}^{1/2}\beta_j\|_2$ term in the convex program by $\rho\|\beta_j\|_2$. The coefficient ball, policy hull and comparator construction are unchanged. The target and applicable statistical bound are those specified in Appendix C.4.

The resulting empirical learner can be implemented as follows.

Algorithm 1 Minimax-regret policy learning from observed proximal moments

Require: Observations $O_{1:n}$ and fixed $\phi_0, \phi_1, \psi, p_{1:J}, B, \rho, \Lambda$

1: Set $\phi_{a,i} = \phi_a(W_i, X_i)$, $\psi_i = \psi(\mathsf{S}_i)$, and $\xi_i = ((1 - A_i)\phi_{0,i}^\top, A_i\phi_{1,i}^\top)^\top$ for each i.
2: Compute $\widehat{M}_T = n^{-1}\sum_i \xi_i\psi_i^\top$, $\widehat{v} = n^{-1}\sum_i \psi_i Y_i$, and $\widehat{\Gamma} = n^{-1}\sum_i \psi_i\psi_i^\top$.
3: Form $\widehat{r}_{p_j - \pi_w} = n^{-1}\sum_i (-(p_j - \pi_w)(X_i)\phi_{0,i}^\top, (p_j - \pi_w)(X_i)\phi_{1,i}^\top)^\top$.
4: Solve Equation (6) to global objective accuracy e_{opt}, obtaining a feasible weight vector $\widehat{w}$.
5: **return** $\widehat{\pi}(x) = \sum_j \widehat{w}_j p_j(x)$.

The finite hull makes both the comparator maximum and policy optimization explicit. The resulting global objective error e_{opt} enters the learning bound in Theorem 2.

C Learning guarantees and proofs

We first establish the causal representation and observed-moment dual, then compare the two penalty norms. The sampling analysis is followed by finite-radius exactness, the root-n consequence, and the causal-regret transfer.

C.1 Causal representation and comparison statements

For any treatment a, proxy independence and the latent bridge equation give

$$\mathbb{E}[Y \mid A = a, Z, X] = \mathbb{E}[\mathbb{E}[Y \mid A = a, U, X] \mid A = a, Z, X] = \mathbb{E}[h_a^\dagger(W, X) \mid A = a, Z, X].$$

The second equality uses $W \perp\!\!\!\perp (A, Z) \mid (U, X)$. Thus $Th^\dagger = m$. Consistency, exchangeability and positivity give $\mathbb{E}[Y(a) \mid U, X] = \mathbb{E}[Y \mid A = a, U, X] = \mathbb{E}[h_a^\dagger(W, X) \mid U, X]$. Taking expectations after multiplying by the bounded X-measurable policy weights proves $V_M(\pi) = L_\pi(h^\dagger)$ for every policy, using the same bridge. Consequently, every causal contrast attainable by a compatible model belongs to the range $\{\langle r_\delta, h\rangle : h \in \mathcal{C}_0\}$ over the exact bridge solution set. Taking suprema over the same policy class proves the causal regret upper bound.

Comparison invariance (Proposition 1). Constancy is exactly $\langle r_\delta, h - h'\rangle = 0$ for every feasible pair, hence orthogonality to its closed linear span. Every exact bridge difference lies in $\ker T$. If $h_\circ$ is strictly inside the radius-B ball, every $v \in \ker T$ admits $h_\circ + tv$ in that ball for sufficiently small $|t|$. Therefore the span of feasible differences contains $\ker T$, and equality follows. The identity $(\ker T)^\perp = \mathrm{cl}\,\mathrm{range}(T^*)$ follows because the orthogonal complement of $\mathrm{range}(T^*)$ is $\ker T$. For an exact adjoint solution, $\langle r_\delta, h\rangle = \langle f, Th\rangle = \langle f, m\rangle$ for every $h \in \mathcal{C}_0$. The equality $\mathbb{E}[fY] = \langle f, m\rangle$, when used, follows by truncating f, applying the conditional expectation identity to the bounded truncations, and using dominated convergence with $|fY|$ and $|fm|$ as dominators. The latter is integrable by Cauchy–Schwarz.

Relation to compatible causal models (Proposition 4). The representation above places every compatible-model contrast in the exact bridge range. Hence $U_0(r_\delta) - \sup_M \Delta_M(\delta) \geq 0$. Cauchy–Schwarz gives $|\langle r_\delta, h\rangle| \leq BK_\Pi$, so $g_{\mathrm{bc}} \leq 2BK_\Pi$ and all required infima are finite. Commuting the model and comparator suprema gives

$$R_{\mathrm{ca}}(\pi) = \sup_{\pi' \in \Pi} \sup_M \Delta_M(\pi' - \pi), \qquad R_0(\pi) = \sup_{\pi' \in \Pi} U_0(r_{\pi' - \pi}).$$

The inner quantities differ by at most g_{bc}. Taking suprema, then policy infima, proves both inequalities in Equation (15). If $R_0(\widetilde{\pi}) \leq R_0^* + \varepsilon$, the same inequalities yield $R_{\mathrm{ca}}(\widetilde{\pi}) \leq R_{\mathrm{ca}}^* + g_{\mathrm{bc}} + \varepsilon$. Under the sufficient condition in Proposition 4, every bridge contrast is also attained by a compatible causal model, giving equality of contrast ranges and $g_{\mathrm{bc}} = 0$.

C.2 Restricted computation and approximation error

For a computational Hilbert space $\mathbb{F}$ satisfying $\|f\|_{\mathbb{G}} \le C_F\|f\|_{\mathbb{F}}$ and fixed radius Λ, define

$$S_{\mathbb{F}}(r) = \inf_{\|f\|_{\mathbb{F}}\le\Lambda}\{\langle f, m\rangle + \rho C_F\|f\|_{\mathbb{F}} + B\|r - T^*f\|_{\mathbb{H}}\}, \quad D_{\mathbb{F}}(\pi) = \sup_{\pi'} S_{\mathbb{F}}(r_{\pi'-\pi}). \tag{16}$$

The restricted criterion upper-bounds R_ρ. To separate computational approximation from residual tolerance, define

$$\begin{aligned} a_\rho &= \sup_{\delta\in\Pi-\Pi}\{S_{\mathbb{F}}(r_\delta) - U_\rho(r_\delta)\}, \\ b_\rho &= \sup_{\delta\in\Pi-\Pi}\{U_\rho(r_\delta) - U_0(r_\delta)\}, \\ a_{\mathrm{ex}} &= \sup_{\delta\in\Pi-\Pi}\{S_{\mathbb{F}}(r_\delta) - U_0(r_\delta)\} \le a_\rho + b_\rho. \end{aligned} \tag{17}$$

The subscript "ex" denotes comparison with the exact-bridge target while retaining the fitted criterion's tolerance ρ. a_{ex} is not a_ρ evaluated at zero. These quantities measure approximation at population moments. If every contrast has an $f \in \mathbb{F}$ with $\|f\|_{\mathbb{F}} \le K \le \Lambda$ and $\|r_\delta - T^*f\|_{\mathbb{H}} \le e$, then $a_{\mathrm{ex}} \le 2Be + \rho C_F K$.

C.3 Observed-moment dual of the bridge support

Theorem 5 (Duality for the bridge set)**.** *Under the Hilbert assumptions of Section 3, with $\mathcal{C}_0 \neq \varnothing$, Equation* (4) *holds for every $r \in \mathbb{H}$ and $\rho \ge 0$. For each fixed r, $U_\rho(r) \downarrow U_0(r)$ as $\rho \downarrow 0$.*

Proof. For fixed r, define $F(h, f) = \langle r, h\rangle + \langle f, m - Th\rangle + \rho\|f\|_{\mathbb{G}}$. For fixed h, its infimum over $f \in \mathbb{G}$ equals $\langle r, h\rangle$ when $\|Th - m\| \le \rho$ and $-\infty$ otherwise. In the latter case choose $f = t(Th - m)/\|Th-m\|$ and let $t \uparrow \infty$. The bridge class is weakly compact, since a closed bounded convex set in a Hilbert space is weakly compact. For fixed f, F is affine weakly continuous in h. For fixed h, it is convex norm-continuous in f. Sion's theorem (Sion, 1958) therefore exchanges the supremum in h and infimum in f, proving Equation (4). The support of a Hilbert ball is $B\|r - T^*f\|$. Weak compactness permits the argument at zero tolerance even when the dual infimum is not attained.

The sets $\mathcal{C}_\rho$ are nested weakly compact sets. For a sequence $\rho_j \downarrow 0$, take maximizing bridges. A weakly converging subsequence exists, its limit lies in $\mathcal{C}_0$ by weak lower semicontinuity of the residual norm, and the objective converges by weak continuity. This proves the pointwise limit. Monotonicity extends it from sequences to $\rho \downarrow 0$. Uniform comparisons across policies are controlled separately by the approximation bounds below. □

For every feasible h and admissible computational f,

$$\langle r, h\rangle = \langle f, m\rangle + \langle f, Th - m\rangle + \langle r - T^*f, h\rangle \le \langle f, m\rangle + \rho C_F\|f\|_{\mathbb{F}} + B\|r - T^*f\|.$$

Taking suprema in h and then infima in f proves conservatism. The inequalities involving $a_\rho, b_\rho, a_{\mathrm{ex}}$ follow by adding the two support differences pointwise and taking suprema. If $\|r - T^*f\| \le e$ and $\|f\|_{\mathbb{F}} \le K$, evaluating on any exact feasible bridge gives $U_0(r) \ge \langle f, m\rangle - Be$, while the same f gives $S_{\mathbb{F}}(r) \le \langle f, m\rangle + \rho C_F K + Be$. Thus $a_{\mathrm{ex}} \le 2Be + \rho C_F K$ uniformly.

The approximation requirement can be summarized by the curve

$$\kappa(e) = \sup_{\delta\in\Pi-\Pi}\inf\{\|f\|_{\mathbb{F}} : \|r_\delta - T^*f\|_{\mathbb{H}} \le e\}.$$

The empty infimum is infinite. When $\Lambda > \kappa(e)$ is finite, apply the preceding argument with any $K \in (\kappa(e), \Lambda]$ and let $K \downarrow \kappa(e)$. At $\Lambda = \kappa(e)$, dual functions attaining the necessary infima must exist. The strict and boundary cases distinguish approximate feasibility from attainment within the fixed dual ball.

C.4 Conditional-residual and moment-residual versions

Use the fixed finite-feature bridge ball from Section 4. Put $d_\theta(\mathsf{S}) = Th_\theta(\mathsf{S}) - m(\mathsf{S})$ and $g(\theta) = M_T^\top\theta - v = \mathbb{E}[\psi(\mathsf{S})d_\theta(\mathsf{S})]$. The coefficient-penalty criterion is

$$S_{\mathrm{E}}(r) = \inf_{\|\beta\|\le\Lambda}\{\beta^\top v + \rho\|\beta\|_2 + B\|r - M_T\beta\|\}, \qquad D_{\mathrm{E}}(\pi) = \sup_{\pi'} S_{\mathrm{E}}(r_{\pi'-\pi}).$$

Cauchy–Schwarz and $\|\psi(\mathsf{S})\| \le 1$ imply $\|g(\theta)\|_2 \le (\mathbb{E}\|\psi\|^2)^{1/2}\|d_\theta\|_{L^2} \le \|d_\theta\|_{L^2}$. Thus $\mathcal{C}_\rho \subseteq \mathcal{C}_\rho^{\mathrm{mom}}$. Define $U_\rho^{\mathrm{mom}}(r) = \sup_{h_\theta \in \mathcal{C}_\rho^{\mathrm{mom}}} r^\top\theta$ and $R_\rho^{\mathrm{mom}}(\pi) = \sup_{\pi'} U_\rho^{\mathrm{mom}}(r_{\pi'-\pi})$. Under the existing nonempty exact-bridge assumption,

$$U_\rho^{\mathrm{mom}}(r) = \inf_{\beta\in\mathbb{R}^q}\{\beta^\top v + \rho\|\beta\|_2 + B\|r - M_T\beta\|\}. \tag{18}$$

To verify this equality, introduce a residual vector z with $\|z\| \le \rho$ and equality $M_T^\top\theta - v = z$. Maximizing over the compact product of the coefficient and residual balls and minimizing its bilinear Lagrangian over β gives the displayed infimum by the same minimax argument as Theorem 5. Infeasible pairs have inner infimum $-\infty$. Consequently

$$U_\rho \le U_\rho^{\mathrm{mom}} \le S_{\mathrm{E}}, \qquad S \le S_{\mathrm{E}},$$

where the second inequality uses $\Gamma \preceq I$.

A finite dual radius as a residual penalty. For any finite matrix M_T and vectors v, r, including empirical moments,

$$S_{\mathrm{E}}(r) = \sup_{\|\theta\|\le B}\left\{r^\top\theta - \Lambda\big(\|M_T^\top\theta - v\|_2 - \rho\big)_+\right\}. \tag{19}$$

Indeed, express $B\|r - M_T\beta\|$ as the supremum of $(r - M_T\beta)^\top\theta$ over the ball. Sion's theorem interchanges this supremum and the infimum over the compact dual ball. For fixed θ, minimizing $\beta^\top(v - M_T^\top\theta) + \rho\|\beta\|$ gives $-\Lambda(\|M_T^\top\theta - v\| - \rho)_+$. Thus a finite radius penalizes violations of the moment constraint. It need not impose the hard empirical set exactly.

For indicators of every positive-probability atom of S, write $p_j = P(\mathsf{S} = s_j) > 0$ and $d_j = d_\theta(s_j)$. Then $g_j = p_j d_j$, $\Gamma = \mathrm{diag}(p_j)$, and

$$\|d_\theta\|^2_{L^2(P_{\mathsf{S}})} = g^\top\Gamma^{-1}g = \sum_j p_j d_j^2, \qquad \|g\|_2^2 = \sum_j p_j^2 d_j^2. \tag{20}$$

For unequal atom probabilities, the two norms are not scalar multiples on the full cell-residual space, so rescaling ρ does not in general equate the constraints. In general, zero tolerance gives $\mathcal{C}_0^{\mathrm{mom}} = \mathcal{C}_0$ only when the test moments express the full conditional restriction, as they do for these complete atom indicators.

The coefficient variant is Equation (16) with $C_F = 1$ and the finite-feature computational norm. If coefficients are redundant as functions in $L^2(P_{\mathsf{S}})$, use the minimum Euclidean coefficient norm: projecting away redundant directions leaves both population and sample moments unchanged almost surely and decreases the penalty and radius. For i.i.d. observations, bounded fixed classes and the stated measurability conditions, Theorem 7 applies under its global empirical optimization inequality. It bounds excess in the original R_ρ by $a_\rho + 2e_n + e_{\mathrm{opt}}$. Its approximation term includes both

$$S_{\mathrm{E}}(r) - U_\rho(r) = \underbrace{S_{\mathrm{E}}(r) - U_\rho^{\mathrm{mom}}(r)}_{\text{finite dual restriction}} + \underbrace{U_\rho^{\mathrm{mom}}(r) - U_\rho(r)}_{\text{moment-set relaxation}}. \tag{21}$$

Both differences are nonnegative. Removing the dual coefficient bound recovers U_ρ^{mom} and does not in general remove the second difference. The coefficient penalty itself is fixed across population and sample, so this bound uses e_n. The empirical-Gram-norm term in E_n applies only when that norm is estimated. The two contributions therefore have different statistical roles. For a positive fitting tolerance and zero-tolerance evaluation, the same event gives

$$R_0(\widehat{\pi}) - R_0^* \le a_{\mathrm{ex}} + 2e_n + e_{\mathrm{opt}}.$$

Apply Equation (26) with $Q = D_{\mathrm{E}}$ and use $R_0 \le D_{\mathrm{E}} \le R_0 + a_{\mathrm{ex}}$. Here a_{ex} retains the training tolerance and dual restriction, so the zero-tolerance evaluation does not remove their learning cost.

C.5 Uniform sampling-error control

Let $\mathbb{H}_0$ and $\mathbb{F}$ be separable reproducing kernel Hilbert spaces (RKHSs), with measurable kernels k and ℓ. The bridge-pair space is the direct sum $\mathbb{H} = \mathbb{H}_0 \oplus \mathbb{H}_0$. Assume $|Y| \le M_Y$, $\sup_v k(v,v) \le$

κ_H^2, and $\sup_s \ell(s,s) \le \kappa_F^2$ almost surely. The policy, bridge, and computational classes and their radii are fixed before observing the sample, and the function classes below are measurable and separable. These conditions give finite envelopes for the three moment classes. The display below uses a common arm RKHS. For arm-specific spaces, use $\mathbb{H} = \mathbb{H}_0 \oplus \mathbb{H}_1$ with kernels k_0, k_1 satisfying $\sup_{a,v} k_a(v,v) \le \kappa_H^2$, and replace k_V in component a by $k_{a,V}$. In the diagonal bound for $\mathcal{G}_T$ below, K_{ii} may be replaced by $\max_a k_a(V_i, V_i)$. Treatment gating keeps its envelope at $\Lambda\kappa_F\kappa_H$, and the contrast envelope remains $\sqrt{2}\kappa_H$. The deviation and finite-hull bounds therefore apply with the same constants. Put $V = (W, X)$, $Pg = \mathbb{E}[g(O)]$, and $P_n g = n^{-1}\sum_{i=1}^n g(O_i)$. Write $k_i = k(V_i, \cdot)$ and $\ell_i = \ell(\mathsf{S}_i, \cdot)$. The Hilbert-valued classes needed for the three moments are

$$\begin{aligned}
\mathcal{G}_Y &= \{O \mapsto f(\mathsf{S})Y : \|f\|_{\mathbb{F}} \le \Lambda\},\\
\mathcal{G}_T &= \{O \mapsto (1\{A=0\}f(\mathsf{S})k_V, 1\{A=1\}f(\mathsf{S})k_V) : \|f\|_{\mathbb{F}} \le \Lambda\},\\
\mathcal{G}_\Delta &= \{O \mapsto (-\delta(X)k_V, \delta(X)k_V) : \delta \in \Pi - \Pi\}.
\end{aligned}$$

Their envelopes are $M_1 = M_Y\Lambda\kappa_F$, $M_2 = \Lambda\kappa_F\kappa_H$, and $M_3 = \sqrt{2}\kappa_H$. For any bounded, measurable separable class $\mathcal{G}$ with values in a Hilbert space $\mathbb{E}_0$, define

$$\widehat{\mathfrak{R}}_{\mathbb{E}_0}(\mathcal{G}) = \mathbb{E}_\sigma\left[\sup_{g\in\mathcal{G}}\left\|n^{-1}\sum_i \sigma_i g(O_i)\right\|_{\mathbb{E}_0}\,\middle|\, O_{1:n}\right],$$

where the σ_i are independent uniform signs, independent of the sample. For a scalar class $\mathcal{A}$, write $\widehat{\mathfrak{R}}_{\mathrm{abs}}(\mathcal{A}) = \mathbb{E}_\sigma[\sup_{a\in\mathcal{A}}|n^{-1}\sum_i \sigma_i a(O_i)| \mid O_{1:n}]$. Scalar complexities here use this absolute-value convention. The class $\mathcal{G}_Y$ is symmetric, so the ordinary scalar convention agrees.

Lemma 6 (Bounded Hilbert-valued deviation)**.** *If $\|g(O)\| \le M$ for every $g \in \mathcal{G}$ and $\eta \in (0,1)$, then with probability at least $1-\eta$,*

$$\sup_g \|(P_n - P)g\| \le 2\widehat{\mathfrak{R}}_{\mathbb{E}_0}(\mathcal{G}) + 6M\sqrt{\frac{\log(2/\eta)}{2n}}.$$

Proof. Let Z be the supremum on the left. An independent ghost sample, conditional Jensen, exchange of each paired observation using an independent sign, and the triangle inequality give $\mathbb{E}Z \le 2\mathbb{E}\widehat{\mathfrak{R}}_{\mathbb{E}_0}(\mathcal{G})$. Boundedness and separability justify these expectations. Replacing one observation changes either Z or $\widehat{\mathfrak{R}}_{\mathbb{E}_0}$ by at most $2M/n$. Applying the bounded-differences inequality to each quantity, with failure probability $\eta/2$ each and $q = \sqrt{\log(2/\eta)/(2n)}$, gives $Z \le \mathbb{E}Z + 2Mq$ and $\mathbb{E}\widehat{\mathfrak{R}}_{\mathbb{E}_0} \le \widehat{\mathfrak{R}}_{\mathbb{E}_0} + 2Mq$. Combining proves the bound. For $M = 0$ it is immediate. This is the standard symmetrization and bounded-differences argument. See also Bartlett & Mendelson (2002). □

Applying this lemma to the three classes and taking a union bound yields

$$\begin{aligned}
e_n(\alpha) = 2\widehat{\mathfrak{R}}(\mathcal{G}_Y) + 2B\widehat{\mathfrak{R}}_{\mathbb{H}}(\mathcal{G}_T) + 2B\widehat{\mathfrak{R}}_{\mathbb{H}}(\mathcal{G}_\Delta)\\
+ 6(M_Y\Lambda\kappa_F + B\Lambda\kappa_F\kappa_H + B\sqrt{2}\kappa_H)\sqrt{\frac{\log(6/\alpha)}{2n}}.
\end{aligned} \tag{22}$$

Let $\widehat{D}_{\mathbb{F}}$ replace the population moments in (16) by empirical means. All classes and envelopes are fixed before observing the sample.

Theorem 7 (Uniform estimation and policy learning)**.** *With probability at least $1-\alpha$, simultaneously for every $\pi \in \Pi$ and every $M \in \mathfrak{M}(P, \mathcal{H}_B)$,*

$$|\widehat{D}_{\mathbb{F}}(\pi) - D_{\mathbb{F}}(\pi)| \le e_n(\alpha), \qquad \mathrm{Reg}_M(\pi) \le \widehat{D}_{\mathbb{F}}(\pi) + e_n(\alpha). \tag{23}$$

If $e_{\mathrm{opt}} \ge 0$ verifies the global inequality $\widehat{D}_{\mathbb{F}}(\widehat{\pi}) \le \inf_{\pi\in\Pi}\widehat{D}_{\mathbb{F}}(\pi) + e_{\mathrm{opt}}$, then on the same event

$$R_\rho(\widehat{\pi}) - R_\rho^* \le a_\rho + 2e_n(\alpha) + e_{\mathrm{opt}}, \tag{24}$$

$$\mathrm{Reg}_M(\widehat{\pi}) \le R_0^* + a_{\mathrm{ex}} + 2e_n(\alpha) + e_{\mathrm{opt}}. \tag{25}$$

Theorem 7 applies to a fixed computational norm. Theorem 2 additionally estimates the $L_2(P_{\mathsf{S}})$ norm used by the finite-feature learner.

Proof of Theorem 7. The RKHS identities $r_\delta = (-\mathbb{E}[\delta k_V], \mathbb{E}[\delta k_V])$ and $T^* f = (\mathbb{E}[1\{A = 0\} f k_V], \mathbb{E}[1\{A = 1\} f k_V])$ follow by testing the defining inner products. On the joint deviation event, the reverse triangle inequality bounds the residual-norm change by the sum of the two Hilbert moment errors. The outcome moment contributes the scalar error. The penalty is identical in population and sample. The pointwise error is at most e_n, uniformly in f, δ. Taking infima and then comparator suprema proves the first inequality of Theorem 7. Applying the population causal bound proves its causal-regret upper bound. For later use, uniform deviation $\sup_\pi |\widehat{Q}(\pi) - Q(\pi)| \le e$ and empirical e_{opt}-near-optimality imply, for every π,

$$Q(\widehat{\pi}) \le \widehat{Q}(\widehat{\pi}) + e \le \widehat{Q}(\pi) + e + e_{\mathrm{opt}} \le Q(\pi) + 2e + e_{\mathrm{opt}}. \tag{26}$$

With $Q = D_{\mathbb{F}}$, take infima and use $R_\rho \le D_{\mathbb{F}} \le R_\rho + a_\rho$ or $R_0 \le D_{\mathbb{F}} \le R_0 + a_{\mathrm{ex}}$. This proves both learning inequalities without assuming that a policy infimum is attained. □

Data-computable radius components. Let K, L be the bridge and computational kernel Gram matrices. Jensen and sign cancellation give

$$\widehat{\mathfrak{R}}(\mathcal{G}_Y) \le \frac{\Lambda}{n}\sqrt{\sum_i Y_i^2 L_{ii}}, \qquad \widehat{\mathfrak{R}}_{\mathbb{H}}(\mathcal{G}_T) \le \frac{\Lambda}{n}\sqrt{\sum_i K_{ii} L_{ii}}.$$

For the second inequality, bound the random operator norm by its Hilbert–Schmidt norm before taking the second moment. For a convex hull of J policy vertices p_j, define $z_{jk,i} = (-[p_k(X_i) - p_j(X_i)]k_i, [p_k(X_i) - p_j(X_i)]k_i)$. Every policy difference is a convex combination of these J^2 differences. Convexity of the norm and inclusion of the vertices make the supremum equal the largest vertex-difference norm. Hence

$$\widehat{\mathfrak{R}}_{\mathbb{H}}(\mathcal{G}_\Delta) \le \frac{1}{n}\sqrt{\sum_{j,k=1}^{J}\sum_i \|z_{jk,i}\|^2}.$$

This bound is conservative, is computed from the observed sample, and covers all mixtures. An exact sign average or a sharper analytical bound can reduce this radius. Monte Carlo approximations require an additional error allowance.

C.6 Sampling radius and proof for the empirical dual norm

With e_n defined by Equation (22), split $\alpha_m + \alpha_\Gamma = \alpha$ into positive terms and define

$$\epsilon_n = \frac{1 + \sqrt{2\log(1/\alpha_\Gamma)}}{\sqrt{n}}, \qquad E_n = e_n(\alpha_m) + \rho\Lambda\sqrt{\epsilon_n}. \tag{27}$$

The exact norm identity is $\|f_\beta\|_{\mathbb{G}}^2 = \beta^\top \Gamma \beta$. Cauchy–Schwarz in the dual argument proves $R_\rho \le D$. Since $\|\psi\| \le 1$, $\Gamma \preceq I$ and $D \le D_{\mathbb{F}}$. To control the empirical $L_2(P_n)$ dual norm, put $Q_i = \psi_i \psi_i^\top$. Its Frobenius norm is at most one. Independence gives

$$\mathbb{E}\left\| n^{-1}\sum_i (Q_i - \mathbb{E}Q_i)\right\|_F^2 = n^{-1}\mathbb{E}\|Q_1 - \mathbb{E}Q_1\|_F^2 \le n^{-1}.$$

The expected Frobenius error is at most $n^{-1/2}$ by Jensen. Replacing one observation changes that error by at most $2/n$. The bounded-differences inequality therefore implies, with probability $1-\alpha_\Gamma$, $\|\widehat{\Gamma} - \Gamma\|_{\mathrm{op}} \le \|\widehat{\Gamma} - \Gamma\|_F \le \epsilon_n$. For all β in the same coefficient ball, the scalar inequality $|\sqrt{a} - \sqrt{b}| \le \sqrt{|a-b|}$ for $a, b \ge 0$ gives

$$\left|\sqrt{\beta^\top \widehat{\Gamma}\beta} - \sqrt{\beta^\top \Gamma \beta}\right| \le \|\beta\|\sqrt{\epsilon_n} \le \Lambda\sqrt{\epsilon_n}.$$

The inequality holds for singular matrices. The joint moment event at failure level α_m and this matrix event have probability at least $1-\alpha$ by a union bound, without requiring independence. The uniform objective error is E_n. For every compatible M, $\mathrm{Reg}_M(\pi) \le R_\rho(\pi) \le D(\pi) \le \widehat{D}(\pi) + E_n$ on this event. Passing the uniform moment bounds through the common infimum and supremum gives the objective-deviation inequality in Theorem 2. Equation (26), with $Q = D$ and $e = E_n$, gives $D(\widehat{\pi}) \le \inf_\pi D(\pi) + 2E_n + e_{\mathrm{opt}}$. Together with $R_\rho \le D$, this proves Equation (7). All objectives are finite for fixed bounded features and radii, and $0 \le c \le \sup_{\delta \in \Pi - \Pi}\{S(r_\delta) - U_\rho(r_\delta)\}$. A small optimal-value gap c need not bound the largest support-function error.

For varying confidence allocations, the empirical-norm term is $O\{n^{-1/4}(1 + \log(1/\alpha_{\Gamma,n}))^{1/4}\}$ at fixed positive tolerance and radii.

C.7 Proof of the sufficient finite dual radius

Theorem 3 uses the same bridge coefficient norm and observed moments as the learner. All quantities below are population quantities, and r ranges over $\{r_\delta : \delta \in \Pi - \Pi\}$.

Proof. Use S_ρ to display the otherwise implicit tolerance argument. The observed-moment definitions give $\beta^\top(M_T^\top \theta - v) = \langle f_\beta, Th_\theta - m\rangle_{\mathbb{G}}$ and $\|f_\beta\|_{\mathbb{G}}^2 = \beta^\top \Gamma \beta$. Cauchy–Schwarz therefore gives weak duality $U_\rho \le S_\rho$ in the original residual norm for every $\rho \ge 0$.

1. Exact bridge geometry. Minimum norm implies $\theta_{\min} \in N^\perp = \mathrm{range}(M_T)$. The exact set is $\theta_{\min} + \{z \in N : \|z\| \le a\}$, so its support is $r^\top \theta_{\min} + a\|P_N r\|$.

2. A bounded dual witness. Write $r_R = P_{N^\perp} r$ and $r_N = P_N r$. If $r_N \ne 0$, put $k = \|r_N\|/a$ and $\theta_* = \theta_{\min} + a r_N/\|r_N\|$, and take β_* from Equation (10). The input to $M_T^\dagger$ lies in $\mathrm{range}(M_T)$, hence $r - M_T\beta_* = k\theta_*$. Also

$$\|\beta_*\| \le \frac{\|r_R\| + (\|\theta_{\min}\|/a)\|r_N\|}{s_*} \le \frac{B\|r\|}{s_* a} \le \Lambda_0.$$

The second inequality is Cauchy–Schwarz on two scalar coordinates and uses $1 + \|\theta_{\min}\|^2/a^2 = B^2/a^2$. Since $v = M_T^\top \theta_{\min}$, $\|\theta_*\| = B$, and $\theta_{\min} \perp r_N$,

$$v^\top \beta_* + B\|r - M_T\beta_*\| = r^\top \theta_{\min} - k\|\theta_{\min}\|^2 + kB^2 = r^\top \theta_{\min} + a\|r_N\|.$$

Weak duality for the original conditional restriction gives the reverse inequality. If $r_N = 0$, choose $\beta_* = M_T^\dagger r$. The residual vanishes, its norm is at most $\|r\|/s_* \le \Lambda_0$, and the same equality follows. This also covers $r = 0$. Therefore every contrast has a feasible exact support witness, uniformly bounded by Λ_0.

3. Positive tolerance. The same witness adds at most $\rho\|\beta_*\| \le \rho\Lambda_0$ because $\Gamma \preceq I$. Thus $U_0 \le U_\rho \le S_\rho \le U_0 + \rho\Lambda_0$. Taking the same comparator supremum, and then policy infima, proves all claims. The argument uses only the common support bounds. □

If $M_T = 0$, exact consistency gives $v = 0$ and $\theta_{\min} = 0$. $\beta = 0$ attains $B\|r\|$ directly, so no nonzero-singular-value condition is needed. The strict interior condition is substantive. For $M_T = (1,0)^\top$, $v = b \in (0, B)$, and $r = (0,1)^\top$, bridges span (b,t) with $|t| \le \sqrt{B^2 - b^2}$ and reverse the sign of the contrast. An exact dual witness is $\beta = -b/\sqrt{B^2 - b^2}$. Its magnitude diverges as $b \uparrow B$. At $b = B$ the infimum for this direction is approached only as $\beta \to -\infty$.

For the finite-policy consequence in the text, define $\Delta = \min_{\pi \in \Pi : R_0(\pi) > R_0^*}\{R_0(\pi) - R_0^*\}$ when this set is nonempty. A returned policy with empirical error at most e_{opt} that fails to minimize R_0 would have excess at least Δ, contradicting Equation (11) at $\rho = 0$. If all policies minimize R_0, selection is automatic. The finite-class gap conclusion concerns selection among isolated policies. The convex-hull result below controls excess regret when randomized mixtures are also admissible.

C.8 Root-n excess worst-case regret with unresolved comparisons

Corollary 8 (Root-n learning with unresolved bridge comparisons)**.** *Fix a distribution P and the finite-feature setting of Theorems 2 and 3, including their boundedness and measurability assumptions. Let $\Pi = \operatorname{conv}\{p_1, \dots, p_J\}$ for a fixed finite collection of prespecified measurable policies taking values in $[0,1]$. Keep the feature maps, B, and a finite radius $\Lambda \geq \Lambda_0$ fixed as n grows, and set $\rho = 0$. For n i.i.d. observations from P, let the sample-measurable policy $\widehat{\pi}_n \in \Pi$ satisfy, almost surely,*

$$\widehat{D}(\widehat{\pi}_n) \leq \inf_{\pi \in \Pi} \widehat{D}(\pi) + e_{\mathrm{opt},n}, \qquad e_{\mathrm{opt},n} \geq 0.$$

If $e_{\mathrm{opt},n} = O_P(n^{-1/2})$, then

$$R_0(\widehat{\pi}_n) - R_0^* = O_P(n^{-1/2}). \tag{28}$$

The condition $P_N r_\delta = 0$ is unnecessary. The constants depend on the fixed moment strength, interior margin, feature bounds, dual radius, and policy hull.

Proof of Corollary 8. The finite-feature class has $\kappa_F = 1$. The moment-complexity bounds in Appendix C.5 give, on every sample,

$$\widehat{\mathfrak{R}}(\mathcal{G}_Y) \leq \frac{\Lambda M_Y}{\sqrt{n}}, \qquad \widehat{\mathfrak{R}}_{\mathbb{H}}(\mathcal{G}_T) \leq \frac{\Lambda \kappa_H}{\sqrt{n}}.$$

For the comparison class, use the J^2-vertex bound already proved in Appendix C.5. Since $\|z_{jk,i}\| \leq \sqrt{2}\kappa_H$,

$$\widehat{\mathfrak{R}}_{\mathbb{H}}(\mathcal{G}_\Delta) \leq \sqrt{2} J \kappa_H / \sqrt{n}.$$

All three bounds hold over the entire hull, including the selected mixture. For any fixed $\alpha \in (0,1)$, take $\alpha_m = \alpha_\Gamma = \alpha/2$. At $\rho = 0$, Equation (22) gives $E_n = e_n(\alpha/2) \leq C_\alpha/\sqrt{n}$, where the finite, sample-independent constant is

$$C_\alpha = 2\Lambda M_Y + 2B\Lambda\kappa_H + 2\sqrt{2}BJ\kappa_H + 6(\Lambda M_Y + B\Lambda\kappa_H + \sqrt{2}B\kappa_H)\sqrt{\frac{\log(12/\alpha)}{2}}.$$

Theorem 3 gives $c = 0$, so Theorem 2 yields, with probability at least $1 - \alpha$,

$$R_0(\widehat{\pi}_n) - R_0^* \leq 2C_\alpha/\sqrt{n} + e_{\mathrm{opt},n}.$$

To conclude $O_P(n^{-1/2})$, fix any $\varepsilon > 0$ and choose $\alpha < \varepsilon/2$. Tightness of $\sqrt{n}e_{\mathrm{opt},n}$ supplies $K_\varepsilon < \infty$ such that its probability of exceeding K_ε is below $\varepsilon/2$ for all sufficiently large n. A union bound then controls the probability that the scaled excess exceeds $2C_\alpha + K_\varepsilon$ by ε. The population geometry controls the target, while the empirical dual remains defined even when an empirical exact-bridge ball is empty. □

C.9 From empirical learning to causal regret

Under Theorem 2 and Proposition 4, use the same prespecified finite-feature bridge space, policy class and nonempty causal model class throughout. On the event of probability at least $1 - \alpha$ of that theorem, every sample-measurable $\widehat{\pi}$ satisfying its empirical optimization inequality obeys

$$R_{\mathrm{ca}}(\widehat{\pi}) - R_{\mathrm{ca}}^* \leq g_{\mathrm{bc}} + b_\rho + c + 2E_n + e_{\mathrm{opt}}. \tag{29}$$

Here g_{bc} measures the largest difference between an upper support value $U_0(r_\delta)$ and the corresponding supremum over compatible causal models, while b_ρ measures the population tolerance cost. The observed bridge equation alone does not determine g_{bc}.

Proof. Equation (17) gives $R_\rho(\pi) \leq R_0(\pi) + b_\rho$ for every π, hence $R_\rho^* \leq R_0^* + b_\rho$ by taking infima, without assuming attainment. Proposition 4 and $R_{\mathrm{ca}} \leq R_0 \leq R_\rho$ then imply

$$\begin{aligned} R_{\mathrm{ca}}(\widehat{\pi}) - R_{\mathrm{ca}}^* &\leq R_\rho(\widehat{\pi}) - R_\rho^* + R_\rho^* - R_{\mathrm{ca}}^* \\ &\leq R_\rho(\widehat{\pi}) - R_\rho^* + b_\rho + g_{\mathrm{bc}}. \end{aligned}$$

Apply Equation (7) on the same event. □

D EXPERIMENTAL DESIGN AND EVALUATION

The methods share observed training data and a policy class within each comparison. This appendix gives the data-generating processes, fitted objectives, evaluation rules, and population calculations behind the figures.

D.1 METHODS AND EVALUATION

Observed information. Every fitted method receives the same observed sample within its comparison and returns a policy in the same declared class. The minimax-regret learner estimates bridge and test moments and minimizes the observed-moment dual. Minimum norm selects the smallest coefficient vector in the same empirical bridge set and maximizes its plug-in value. Rectangular relaxation computes each group's treatment-contrast interval from that set, then permits all combinations of these intervals. Maximin value keeps the set and maximizes its lowest policy value. None of these fitted rules uses latent variables or simulated counterfactual outcomes. Population quantities define evaluation references and the fixed simulation settings.

Numerical computation. All main comparisons fit and evaluate zero-tolerance targets. For the synthetic bridge-selection and criterion designs, $B = 2$ and $\Lambda = 125$. The four matched comparisons in Appendix D.4 use a common $\Lambda = 2415$, fixed from their population representation bounds before sampling. CLARABEL uses tolerance 10^{-9} with at most 500 iterations. Exact affine-ball support calculations verify the empirical objectives on the declared policy classes. Appendix E.3 gives global optimization bounds for the learning curve.

Paired evaluation and uncertainty. For a fitted policy $\widehat{\pi}$, evaluate R_0, $\mathcal{E}_0 = R_0 - R_0^*$ and W_0 on its design's exact population bridge set. The secondary causal loss is $C(\widehat{\pi}) = \sup_{\pi\in\Pi} V_{\mathrm{gen}}(\pi) - V_{\mathrm{gen}}(\widehat{\pi})$ under that generating model. Lower C is better. Each main comparison uses 200 paired datasets. Means and paired intervals use the nonempty empirical bridge sets, with all counts reported in Table 1. For differences $d_1, \ldots, d_m$ between policies fitted on the same data, report $\bar{d} \pm t_{m-1,.975} s_d/\sqrt{m}$. Holm corrections use three prespecified families. The first contains five tests: $\mathcal{E}_0$ for the synthetic joint-restriction comparison and the ACIC joint-restriction and bridge-selection comparisons, plus R_0 and W_0 for the ACIC criterion comparison. The second contains the two synthetic criterion tests. The third contains the three synthetic proxy-strength comparisons against minimum norm. Reported 95% intervals are ordinary paired intervals. The learning curve uses pointwise intervals over 30 repetitions per size.

D.2 DECISION-RELEVANT AMBIGUITY WITH PARTIAL PROXIES

Let U, V be independent uniform signs in $\{-1, 1\}$. Given (U, V), draw treatment, three proxies, and the potential-outcome noises independently, with

$$P(A = 1 \mid U, V) = \tfrac{1}{2} + \tfrac{1}{5}U + \tfrac{1}{10}V, \qquad P(Y(a) = 1 \mid U, V) = \tfrac{1}{2} - \tfrac{1}{30}a + \tfrac{3}{20}U + \tfrac{1}{10}V,$$
$$P(Q = q \mid U) = \tfrac{1}{2}(1 + qU/2), \qquad P(R = r \mid V) = \tfrac{1}{2}(1 + rV/2),$$
$$P(Z = z \mid V) = \tfrac{1}{2}(1 + szV), \qquad q, r, z \in \{-1, 1\}.$$

Only $(A, Q, R, Z, Y(A))$ is observed. The outcome proxy is $W = (Q, R)$, and there are no deployment covariates. The policy class contains all constant treatment probabilities $p \in [0, 1]$. Both latent factors confound treatment and outcome. The proxy Z measures V, leaving the contribution of U to policy contrasts unresolved.

The bridge features are $(1, Q, R)/\sqrt{3}$ separately for each treatment arm, with joint coefficient radius $B = 2$. The four indicators of (A, Z) are the test features. The bridge $h_a^\dagger = 1/2 - a/30 + 3Q/10 + R/5$ satisfies the latent outcome equation, and its normalized coefficient squared norm is $131/60 < 4$. For every $s \in (0, 1]$, the observed moment matrix has rank four and nullity two. Adding and subtracting the two rows of $M_T^\top$ within each treatment arm, then dividing the difference by s, gives constraints independent of s. Hence all three positive strengths below have the same exact bridge set. Their always-treat minus never-treat contrast interval is

$$[\ell, u] = \frac{34}{363} \pm \sqrt{\frac{223249}{3294225}} = [-.16666244, .35399026].$$

The interval contains both signs. The population minimum-norm bridge has contrast $34/363 > 0$ and chooses $p = 1$. For any constant policy,

$$R_0(p) = \max\{(1-p)u, -p\ell\}, \qquad p^* = \frac{u}{u-\ell} = .67989710, \qquad R_0^* = .11331331.$$

Thus minimum-norm selection incurs population excess regret .05334913. The generating treatment effect is $-1/30$, so its causal loss is $C(p) = p/30$.

For Figure 1(a), take the support witnesses h_-, h_+ attaining ℓ, u symmetrically about the minimum-norm bridge. The path $h_t = (1-t)h_-/2 + (1+t)h_+/2$, $t \in [-1,1]$, lies in $\mathcal{C}_0$ by convexity. Its midpoint is the minimum-norm bridge, and its bridge-implied treatment contrast is $34/363 + t\sqrt{223249/3294225}$. The path spans the full contrast interval used to compute the regret curve in panel (b).

Before sampling, we fixed $n = 4{,}000$, 200 independent paired datasets per $s \in \{.25, .5, .75\}$, training and evaluation tolerance $\rho = 0$, and $\Lambda = 125$. The middle strength is the main reporting setting. The sufficient population radii from Theorem 3 are 42.574, 21.407, and 14.420. All sample moments, including policy-value directions, are estimated from the observed data. Minimax regret uses the observed-moment dual. The minimum-norm rule selects a bridge from exactly the same empirical equality-and-ball set. The finite support permits an independent affine-ball calculation of the empirical optimum for every fitted dataset.

D.3 Criterion choice with identified standard care

This design uses the treatment and proxy laws of Appendix D.2 with $s = .5$, and changes the potential-outcome means to

$$P\{Y(a) = 1 \mid U, V\} = \frac{1}{2} + \frac{a}{20} + \frac{3aU}{20} + \frac{V}{10}.$$

Standard care has no U term. Before sampling, we specify arm-specific bridge features $(1, R)/\sqrt{3}$ for $a = 0$ and $(1, Q, R)/\sqrt{3}$ for $a = 1$, with joint coefficient radius $B = 2$. Thus the bridge class fixes the Q coefficient of standard care to zero. Both methods use this same class and all constant treatment probabilities $p \in [0,1]$. The causal bridges are $h_0^\dagger = 1/2 + R/5$ and $h_1^\dagger = 11/20 + 3Q/10 + R/5$. Their normalized coefficient squared norm is $867/400 < 4$, and their conditional means given (U, V) equal the displayed outcome means.

With coefficients ordered as $(\theta_{00}, \theta_{0R}, \theta_{10}, \theta_{1Q}, \theta_{1R})$ and the four tests ordered by (A, Z), the population moments are

$$M_T = \frac{1}{80\sqrt{3}} \begin{pmatrix} 22 & 18 & 0 & 0 \\ -7 & 3 & 0 & 0 \\ 0 & 0 & 18 & 22 \\ 0 & 0 & 4 & 4 \\ 0 & 0 & -3 & 7 \end{pmatrix}, \qquad v = \frac{1}{800}(96, 96, 105, 147)^\top.$$

The matrix has rank four. Its unit null direction is $\nu = (0, 0, -5, 24, 2)^\top/\sqrt{605}$, so the exact bridge set is $\theta_{\min} + t\nu$ with $|t| \le a$, where

$$\theta_{\min} = \sqrt{3}\left(\frac{1}{2}, \frac{1}{5}, \frac{357}{605}, \frac{651}{6050}, \frac{1113}{6050}\right)^\top, \qquad a^2 = \frac{117923}{60500}.$$

Every bridge gives standard-care value $1/2$. Its bridge-implied treatment contrast lies in

$$[\ell, u] = \frac{109}{1210} \pm \sqrt{\frac{117923}{4392300}} = [-.07376993, .25393522].$$

Consequently the two population objectives have the simple forms

$$R_0(p) = \max\{u(1-p), -\ell p\}, \qquad W_0(p) = \inf_{h \in \mathcal{C}_0} L_p(h) = \frac{1}{2} + \ell p.$$

Maximin value uniquely selects $p = 0$. Minimax regret uniquely selects $p = u/(u-\ell) = .77488932$, equalizing the two possible opportunity losses at $R_0^* = .05716353$. Its worst-case

policy value is .44283647. For no treatment, the corresponding quantities are $R_0 = .25393522$ and $W_0 = .5$. The generating treatment effect is $1/20$, giving $C(p) = (1-p)/20$. These formulas distinguish the protection supplied by each criterion.

We fixed $n = 4{,}000$, 200 paired datasets, $\rho = 0$, and $\Lambda = 125$ before sampling. The sufficient population radius is 24.189 for policy contrasts and is no larger for policy-value directions. Each method estimates every moment from the same observed rows. All 200 empirical sets were nonempty and both methods completed every fit. The two primary paired comparisons are R_0 and W_0, with a two-test Holm correction. For minimax regret minus maximin value, the regret difference has ordinary 95% interval $[-.19118, -.18921]$. The worst-case policy-value difference has interval $[-0.05803, -0.05692]$ and adjusted $p < .001$. These paired comparisons quantify the reduction in worst-case policy value associated with lower worst-case regret in Table 1.

D.4 Matched zero-tolerance comparisons

Observed information and policy class. Let U, V be independent fair signs, independent of the observed covariate X. Conditional on (U, V, X), treatment A, proxy variables Q, R, Z, and the two binary potential outcomes are independent. Their probabilities are

$$\begin{aligned} \Pr(A = 1 \mid U, V) &= \tfrac{1}{2} + \tfrac{1}{5}U + \tfrac{1}{10}V, \\ \Pr(Q = q \mid U) &= \tfrac{1}{2}(1 + qU/2), \\ \Pr(R = r \mid V) &= \tfrac{1}{2}(1 + rV/2), \\ \Pr(Z = z \mid V) &= \tfrac{1}{2}(1 + zV/2), \end{aligned} \qquad q, r, z \in \{-1, 1\}.$$

The potential-outcome means $\mu_a(U, V, X)$ are specified below, and $Y(a) \mid U, V, X \sim \mathrm{Bernoulli}(\mu_a)$. Fitting observes (X, A, Q, R, Z, Y) with $Y = Y(A)$. The propensity lies in $[.2, .8]$. Both latent signs affect treatment and outcomes.

Write $G = -$ when $X \le 0$ and $G = +$ when $X > 0$, with group masses $\omega_j = \Pr(G = j)$. The policy class is the square $[0, 1]^2$, represented by the four vertices $(0, 0), (0, 1), (1, 0), (1, 1)$. For arm-specific bridge features $\phi_a(X, Q, R)$, define

$$\xi = \big((1 - A)\phi_0^\top, A\phi_1^\top\big)^\top, \qquad h_\theta(A, X, Q, R) = \xi^\top \theta.$$

For the test vector $\psi(A, Z, X)$, let $M_T = \mathbb{E}[\xi\psi^\top]$ and $v = \mathbb{E}[\psi Y]$. The common bridge set is

$$\mathcal{C}_0 = \{\theta : M_T^\top \theta = v,\ \|\theta\|_2 \le B\}.$$

Its sample version replaces every expectation by the average of the same n observed rows. This includes the policy direction

$$q(p) = \mathbb{E}\Big[\big((1 - p_G)\phi_0^\top, p_G\phi_1^\top\big)^\top\Big]$$

and the group masses. Write $\widehat{q}(p)$ for the sample policy direction.

Fixed ACIC covariate population. We use the public 2016 Atlantic Causal Inference Conference ACIC covariates. The ACIC cells sample covariate indices uniformly with replacement from the fixed 4,802-row pool. We retain the numeric columns `x_1`, `x_10`, `x_11`, `x_12`, `x_13`, `x_14` prespecified before generating treatment or outcomes. Missing numeric entries are filled by their column median. If $r_{i\ell}$ is the average rank of row i in column ℓ, tied entries receive the same rank, and

$$t_{i\ell} = 2(r_{i\ell} - .5)/4802 - 1.$$

Define

$$f_i = \frac{t_{i1}t_{i2} + (2t_{i3}^2 - 1) + \tanh(t_{i4} + t_{i5})}{3}, \qquad X_i = \frac{f_i - \overline{f}}{\max_k |f_k - \overline{f}|}, \qquad \overline{f} = \frac{1}{4802}\sum_i f_i.$$

The scalar score uses the first five selected columns. The resulting fixed population has $\mathbb{E}[X] = 0$, $|X| \le 1$, positive variance and positive mass in both groups. Population evaluation averages over this pool and enumerates the 64 combinations of (U, V, A, Q, R, Z). Outcomes enter that calculation through their Bernoulli means. The transformation and population distribution are fixed before generating treatment or outcomes.

D.4.1 WHOLE BRIDGE SET AND MINIMUM-NORM SELECTION

The bridge-selection cell uses

$$\mu_a = \tfrac{1}{2} - \tfrac{a}{30} + \tfrac{3}{20}U + \tfrac{1}{10}V + \tfrac{1}{20}X.$$

Let $b(X) = (1, X)^\top/\sqrt{2}$. Both arms use $\phi_a = b(X) \otimes (1, Q, R)^\top/\sqrt{3}$, giving twelve coefficients in total, and $B = 2\sqrt{2}$. The eight test coordinates are $1\{A = a, Z = z\}b(X)$ for $a \in \{0, 1\}$ and $z \in \{-1, 1\}$. A true bridge is

$$h_a^\dagger = \tfrac{1}{2} - \tfrac{a}{30} + \tfrac{3}{10}Q + \tfrac{1}{5}R + \tfrac{1}{20}X.$$

The equalities $\mathbb{E}[Q \mid U] = U/2$ and $\mathbb{E}[R \mid V] = V/2$ give $\mathbb{E}[h_a^\dagger \mid U, V, X] = \mu_a$. Its normalized coefficient vector lies strictly inside the declared ball.

The minimum-norm comparator selects

$$\widehat{\theta}_{\min} = \arg\min_{\theta \in \widehat{\mathcal{C}}_0} \|\theta\|_2$$

and maximizes this bridge's value over the same policy square. It treats a group when its fitted contrast is positive and withholds treatment when that contrast is negative. Contrasts within 10^{-12} of zero give probability $1/2$. Minimax regret uses the complete set $\widehat{\mathcal{C}}_0$. The generating-model treatment effect averages to $-1/30$ in each group, so its causal loss is $C(p) = (\omega_- p_- + \omega_+ p_+)/30$.

D.4.2 JOINT RESTRICTIONS AND A MARGINAL RECTANGLE

The synthetic and ACIC joint-restriction cells share

$$\mu_0 = \tfrac{1}{2} - \tfrac{1}{20}X - \tfrac{29}{480}U + \tfrac{1}{10}V,$$
$$\mu_1 = \tfrac{1}{2} + \tfrac{1}{20}X + \tfrac{7}{160}U + \tfrac{1}{10}V.$$

Each arm uses $\phi_a = (1, X, Q, R, XR)^\top/\sqrt{5}$, and the ten concatenated coefficients satisfy $\|\theta\|_2 \le \sqrt{5}$. Equivalently, the raw coefficients in $h_a = c_a + d_a X + e_a Q + f_a R + k_a XR$ lie in a unit Euclidean ball. The eight tests are $1\{A = a\}(1, Z, X, XZ)^\top/2$ for $a = 0, 1$. The true bridges are

$$h_0^\dagger = \tfrac{1}{2} - \tfrac{1}{20}X - \tfrac{29}{240}Q + \tfrac{1}{5}R,$$
$$h_1^\dagger = \tfrac{1}{2} + \tfrac{1}{20}X + \tfrac{7}{80}Q + \tfrac{1}{5}R.$$

Their latent conditional expectations equal μ_0, μ_1. All Bernoulli means lie strictly between zero and one for $|X| \le 1$.

The geometry can be computed directly in raw coefficients. Put $t = 2a - 1$. The null direction for arm a is

$$n_a = (1, 0, -24t/5, -2t/5, 0)^\top.$$

The true coefficient vector is orthogonal to both arm null directions and has squared norm $17489/28800$. It is therefore the minimum-norm center. Writing $h_a = h_a^\dagger + z_a n_a^\top (1, X, Q, R, XR)^\top$, the unit ball gives

$$z_0^2 + z_1^2 \le \frac{5}{121}\left(1 - \frac{17489}{28800}\right).$$

Since Q, R have zero marginal means and are independent of X, every conditional bridge contrast satisfies

$$\mathbb{E}[h_1 - h_0 \mid X] = X/10 + \zeta, \qquad \zeta = z_1 - z_0, \qquad |\zeta| \le d_{\text{shift}}, \qquad d_{\text{shift}}^2 = \frac{11311}{348480}.$$

Thus both groups share one shift. For synthetic $X \sim \text{Unif}\{-1, 1\}$, the joint minimax policy is $(0, 1)$, and its worst regret is $(d_{\text{shift}} - .1)/2$. The rectangle gives

$$p_-^{\text{rect}} = \frac{d_{\text{shift}} - .1}{2d_{\text{shift}}}, \qquad p_+^{\text{rect}} = \frac{d_{\text{shift}} + .1}{2d_{\text{shift}}}.$$

The sample comparison constructs the rectangle from the empirical joint set itself. For group j, let d_j be its conditional bridge-contrast direction, so $d_j^\top \theta = \mathbb{E}[h_1 - h_0 \mid G = j]$. Estimate d_j and ω_j from the common observed sample, and calculate

$$\widehat{\ell}_j = \inf_{\theta \in \widehat{\mathcal{C}}_0} \widehat{d}_j^\top \theta, \qquad \widehat{u}_j = \sup_{\theta \in \widehat{\mathcal{C}}_0} \widehat{d}_j^\top \theta.$$

The rectangular rule minimizes

$$\sum_{j \in \{-,+\}} \widehat{\omega}_j \max\{(1-p_j)\widehat{u}_j, -p_j \widehat{\ell}_j, 0\}.$$

When an interval straddles zero, this gives $p_j = \widehat{u}_j/(\widehat{u}_j - \widehat{\ell}_j)$. A nonnegative interval gives probability one, a nonpositive interval gives zero, and an interval containing only zero gives $1/2$. The joint learner uses the full empirical coefficient ball and equations. Both rules are evaluated on the same population joint regret. The common-shift formula describes the population geometry. Sample comparisons use the constraints estimated from the observed data.

D.4.3 Minimax regret and maximin value

The criterion cell uses

$$\mu_a = \tfrac{1}{2} + \tfrac{a}{20} + \tfrac{3a}{20}U + \tfrac{1}{10}V + \tfrac{1}{20}X.$$

Arm zero uses $b(X) \otimes (1, R)^\top/\sqrt{3}$ and arm one uses $b(X) \otimes (1, Q, R)^\top/\sqrt{3}$, giving ten coefficients. The eight tests are the same as in the bridge-selection cell, and $B = 2\sqrt{2}$. The true bridges are

$$h_0^\dagger = \tfrac{1}{2} + \tfrac{1}{5}R + \tfrac{1}{20}X, \qquad h_1^\dagger = \tfrac{11}{20} + \tfrac{3}{10}Q + \tfrac{1}{5}R + \tfrac{1}{20}X.$$

The arm-zero moment block is invertible, so its bridge is fixed throughout the exact population solution set. Arm-one ambiguity leaves treatment rankings unresolved. The two rules use the same empirical set and policy square, optimizing respectively

$$\inf_p \sup_{\theta \in \widehat{\mathcal{C}}_0} \max_{p'} \{\widehat{q}(p') - \widehat{q}(p)\}^\top \theta \quad \text{and} \quad \sup_p \inf_{\theta \in \widehat{\mathcal{C}}_0} \widehat{q}(p)^\top \theta.$$

We report R_0 and W_0 together to show the tradeoff between missed benefit and worst-case policy value. The generating-model effect averages to $1/20$, giving $C(p) = (1 - \omega_- p_- - \omega_+ p_+)/20$.

D.4.4 Connection to the conditional bridge equation

For any bridge in a declared class, define its conditional residual

$$e_\theta(A, Z, X) = \mathbb{E}[Y - h_\theta(A, X, Q, R) \mid A, Z, X].$$

Independence of X from (U, V, A, Q, R, Z) implies that this residual is affine in X within each (A, Z) cell. It therefore belongs to the span of the eight tests. If $M_T^\top \theta = v$, then $\mathbb{E}[\psi e_\theta] = 0$. Writing $e_\theta = \alpha^\top \psi$ yields $\mathbb{E}[e_\theta^2] = \alpha^\top \mathbb{E}[\psi e_\theta] = 0$. Conversely, a zero conditional residual gives every test moment. Hence the population moment equations are equivalent to the complete conditional bridge equation within each declared feature class.

The test Gram matrix is nonsingular because $\mathrm{Var}(X) > 0$, every (A, Z) cell has positive probability, and X is independent of (A, Z). For the joint-restriction design, reorder the raw bridge coordinates as $(1, Q, R, X, XR)$. Up to the positive feature-normalization factors, centering X separates each arm's matrix into

$$K_a = \begin{pmatrix} 1/2 & t/20 \\ t/10 & 0 \\ t/20 & 1/8 \end{pmatrix} \quad \text{and} \quad \mathbb{E}[X^2] L_a, \qquad L_a = \begin{pmatrix} 1/2 & t/20 \\ t/20 & 1/8 \end{pmatrix}.$$

Here K_a has rank two and $\det L_a = 3/50$. Each arm thus has rank four. The joint-restriction design has rank eight and nullity two. In the bridge-selection design, the nonsingular covariate Gram matrix multiplies a rank-two proxy block for each arm, giving rank eight and nullity four. Removing Q from the criterion criterion design's arm-zero bridge makes that four-dimensional block invertible. Its total rank is eight and nullity is two.

For the numerical checks, let $\theta_{\min}$ be the minimum-norm solution, P_N the orthogonal projector onto $\ker M_T^\top$, and $a = (B^2 - \|\theta_{\min}\|^2)^{1/2}$, as in Theorem 3. For any direction d, the affine-ball support is

$$\sup_{\theta \in \mathcal{C}_0} d^\top \theta = d^\top \theta_{\min} + a\|P_N d\|.$$

We use this expression to compute population references and to check the finite-dual fits against an independent optimization over the policy square.

Let s_* be the smallest positive singular value of the normalized matrix M_T. The sufficient witness-norm bound for direction d is $B\|d\|/(s_* a)$. Before drawing outcomes, we maximize this bound over the four population cells and the required vertex-value, vertex-contrast directions used by the finite dual. The synthetic joint-cell maximum is about 38.1, and the largest ACIC bound is about 1207. The fixed rule

$$\Lambda = \max\left\{125, \left\lceil 2 \max_{\text{cells},d} \frac{B\|d\|}{s_* a} \right\rceil\right\} = 2415$$

sets one radius for all four comparisons. This is a design constant computed from the known simulation laws. Fitted bridge moments, value directions and group frequencies use only the observed sample.

Each cell uses 200 paired datasets of size 4,000, with uncertainty reported as specified in Appendix D.1.

D.5 Two-world design

The two-world design implements the analytical pair in Appendix A.4. The two causal worlds share the observed distribution but prefer opposite constant policies, with effect magnitudes τ_+, τ_- defined and evaluated in Appendix A.4. Let $\theta = (h_0(0), h_0(1), h_1(0), h_1(1))$. The exact population bridge constraint and bridge ball are

$$\begin{pmatrix} 17/25 & 8/25 & 0 & 0 \\ 0 & 0 & 8/25 & 17/25 \end{pmatrix} \theta = \begin{pmatrix} 11/50 \\ 39/50 \end{pmatrix}, \qquad \|\theta\| \le 3.$$

For the contrast between always treating and never treating, the support interval has endpoints

$$\ell, u = \frac{175}{353} \mp \sqrt{\frac{448173}{498436}}.$$

For a constant treatment probability ω,

$$R_0(\pi_\omega) = \max\{u(1-\omega), -\ell\omega\}, \qquad \omega^*_{R_0} = \frac{u}{u-\ell}.$$

If the decision problem is restricted to the two displayed causal worlds, the minimax treatment probability is instead $725/999$. The two reference policies differ because the exact bridge solution set contains variation beyond these two latent models.

The observed cells give $P(A=1) = 1/2$, $\mathbb{E}[AY] = 0.39$, and $\mathbb{E}[(1-A)Y] = 0.11$. Consistency therefore writes any compatible average treatment effect as

$$0.28 + \mathbb{E}[(1-A)Y(1)] - \mathbb{E}[AY(0)].$$

The binary observed law, latent exchangeability, and strict positivity force both potential outcomes to be binary almost surely, so the two unobserved terms lie in $[0, 1/2]$. Hence every compatible causal effect is in $[-0.22, 0.78]$, an outer bound that additional proxy restrictions may tighten. The upper bridge endpoint is $u = 1.4439904289\ldots$, so the always-treat versus never-treat direction gives $g_{\text{bc}} \ge 0.6639904289\ldots$ for the support gap defined in Proposition 4.

Corollary 9 (Exact restricted dual in the two-world design)**.** *With the unnormalized four-cell test basis, $B = 3$, $\rho = 0$, and $\Lambda = 20$, every constant randomized policy π_ω, $\omega \in [0,1]$, satisfies $D(\pi_\omega) = R_0(\pi_\omega)$. Consequently $c = 0$ on this policy class and Theorem 2 bounds excess R_0 by $2E_n + e_{\text{opt}}$. The discrepancy from compatible-model causal regret is quantified separately above.*

Proof. With the four unnormalized cell indicators ordered by $(A, Z) = (0,0), (0,1), (1,0), (1,1)$, the finite-feature moments and the always-treat minus never-treat contrast are

$$M_T = \frac{1}{100}\begin{pmatrix} 17 & 17 & 0 & 0 \\ 8 & 8 & 0 & 0 \\ 0 & 0 & 8 & 8 \\ 0 & 0 & 17 & 17 \end{pmatrix}, \quad v = \frac{1}{200}(11, 11, 39, 39)^\top, \quad r = \frac{1}{2}(-1, -1, 1, 1)^\top.$$

Because the four indicators exhaust the positive-probability support of (A, Z), $M_T^\top \theta = v$ is the complete finite conditional bridge equation in this design. Its minimum-norm solution and the unit null direction relevant to r are

$$\theta_{\min} = \frac{1}{706}(187, 88, 312, 663)^\top, \qquad d = \frac{1}{\sqrt{706}}(8, -17, 17, -8)^\top.$$

Here $\|\theta_{\min}\|^2 = 821/706$, $\theta_{\min} \perp d$, and $P_{\ker(M_T^\top)} r = (9/\sqrt{706})d$. The two support witnesses in the $B = 3$ ball are therefore

$$\theta_\pm = \frac{1}{706}\big(187 \pm 8\sqrt{5533}, 88 \mp 17\sqrt{5533}, 312 \pm 17\sqrt{5533}, 663 \mp 8\sqrt{5533}\big)^\top.$$

They satisfy $M_T^\top \theta_\pm = v$ and $\|\theta_\pm\| = 3$, and give

$$\sup_{\theta: M_T^\top \theta = v,\ \|\theta\| \le 3} r^\top \theta = \frac{175}{353} + \frac{9\sqrt{5533}}{706} = u, \qquad \sup_{\theta: M_T^\top \theta = v,\ \|\theta\| \le 3} (-r)^\top \theta = -\ell.$$

It remains to show that the radius-20 restricted dual attains both values. Writing $q = \sqrt{5533}$, define

$$\beta_+ = (b_{0+}, b_{0+}, b_{1+}, b_{1+})^\top, \quad b_{0+} = -\frac{25(425 + 1683/q)}{6001}, \quad b_{1+} = \frac{25(50 - 702/q)}{706},$$
$$\beta_- = (b_{0-}, b_{0-}, b_{1-}, b_{1-})^\top, \quad b_{0-} = \frac{25(425 - 1683/q)}{6001}, \qquad b_{1-} = -\frac{25(50 + 702/q)}{706}.$$

Direct substitution gives

$$r - M_T\beta_+ = \frac{9}{q}\theta_+, \qquad -r - M_T\beta_- = \frac{9}{q}\theta_-.$$

Since $74 < q < 75$, the displayed expressions imply $\|\beta_+\| < 4$ and $\|\beta_-\| < 4$, hence both are strictly feasible for $\Lambda = 20$. For any feasible bridge θ and any β,

$$\beta^\top v + 3\|r - M_T\beta\| = \theta^\top M_T\beta + 3\|r - M_T\beta\| \ge \theta^\top r.$$

The residual identities make this weak-duality inequality an equality for (θ_+, β_+) in direction r and for (θ_-, β_-) in direction $-r$. Thus the restricted and exact supports agree in both directions. Every difference of constant randomized policies is ar for some $a \in [-1, 1]$. For $a \in [0, 1]$, the feasible witness $a\beta_+$ gives $S(ar) \le (a\beta_+)^\top v + 3\|ar - M_T(a\beta_+)\| = aU_0(r) = U_0(ar)$. For $a \in [-1, 0)$, the feasible witness $|a|\beta_-$ gives $S(ar) \le (|a|\beta_-)^\top v + 3\|ar - M_T(|a|\beta_-)\| = |a|U_0(-r) = U_0(ar)$. Both witnesses stay within the radius-20 ball since $|a| \le 1$. Weak duality gives the reverse inequality, including at $a = 0$ with witness $\beta = 0$. Thus $S(ar) = U_0(ar)$. Taking the same comparator supremum on both sides proves $D(\pi_\omega) = R_0(\pi_\omega)$ for all $\omega \in [0, 1]$, and hence $c = 0$. The residual norm equals $27/\sqrt{5533}$ because $\|\theta_\pm\| = 3$. □

Learning-curve setup. We use sample sizes 250, 500, 1,000, 1,500, 2,000, 4,000, 6,000, 10,000, 20,000, and 100,000, with 30 independent datasets at each size. Every fit uses the same observed law, constant-policy class, four (A, Z) cell indicators, and $(B, \rho, \Lambda) = (3, 0, 20)$. The evaluation target is population excess regret $\mathcal{E}_0$.

E Additional experimental results

This appendix reports additional proxy strengths, details of the matched comparisons and learning curve, and sensitivity to bridge interactions.

E.1 Partial-proxy strength and finite-sample learning

For the main setting $s = .5$, the paired excess-regret interval, minimax regret minus minimum norm, is $[-.04711, -.04544]$. The remaining strengths $s = .25$ and .75 give mean minimax excess regret .00816 and .00677, compared with .05335 for minimum-norm selection. Their paired differences have ordinary 95% intervals $[-.04629, -.04408]$ and $[-.04739, -.04577]$. All three prespecified Holm-adjusted tests have $p < .001$. Minimum norm chooses $p = 1$ on all 600 datasets. Minimax regret has mean generating-world causal losses .02293 and .02262 at these two strengths, compared with $1/30$ for minimum-norm selection.

All 600 datasets yielded nonempty rank-four empirical bridge sets and valid fits. Their sufficient dual radii were at most $121.230 < 125$. Because the exact population bridge set is identical across the three positive proxy strengths, these comparisons change the sampling information without changing the population target.

E.2 Matched comparisons and generating-world loss

The synthetic joint and three ACIC comparisons use 800 datasets in total. The empirical equality constraints intersect the coefficient ball in 775 of them. The other 25 sets are empty, so the exact-set comparisons use the valid pairs counted in Table 1. Every fixed seed was evaluated, with no replacement of infeasible datasets. All five prespecified paired tests have Holm-adjusted $p < .001$. Table 2 evaluates the same fitted policies in their generating worlds using the formulas in Appendix D.

E.3 Learning when the population dual is exact

Figure 2 uses the 300 datasets described in Appendix D.5. For each fit, feasible dual coefficients bound the returned policy's empirical objective from above, and a feasible conic-dual witness bounds its global optimum from below. Rational arithmetic with outward rounding for square roots gives a gap below 2.14×10^{-7} for every fit. Optimization error is therefore small relative to the plotted excess regret.

The exact empirical bridge set is empty in 91 of the 300 samples. The bounded dual remains defined on these samples as the residual-penalized criterion in Equation (19), whose value can be negative. Its population counterpart equals R_0 by Corollary 9. Theorem 2 connects the empirical criterion to this target through moment estimation error.

E.4 Allowing group interactions in the bridge class

This separate population calculation varies the allowed bridge interactions across two covariate groups. Let $U \sim \mathrm{Unif}\{-1, 0, 1\}$ and $X \sim \mathrm{Unif}\{-1, 1\}$ independently. Conditional on (U, X), draw A, Z, W, $Y(0)$ and $Y(1)$ independently. Write $e_u = P(A = 1 \mid U = u)$, $z_u = P(Z = 1 \mid U = u)$ and $r_W = P(W = U \mid U)$. Each other proxy value has probability $(1 - r_W)/2$. For $u = (-1, 0, 1)$, we set $(e_u) = (.2, .5, .8)$, $(z_u) = (.3, .5, .7)$ and $r_W = .8$. With $\phi(X, W) = (1, X, W, W^2)^\top$, the binary potential outcomes have conditional means

$$\mu_a(U, X) = E\{\phi(X, W) \mid U, X\}^\top \theta_a^0,$$

where

$$\theta_0^0 = (b_0, -\kappa/2, b_W, b_2)^\top, \qquad \theta_1^0 = (b_0, \kappa/2, -b_W, b_2)^\top.$$

We fix $\kappa = .1$ and

$$b_0 = \frac{78165}{230822}, \qquad b_W = \frac{22815}{461644}, \qquad b_2 = \frac{86697}{461644}.$$

These fixed coefficients determine the intercept, linear proxy term, and quadratic proxy term. The opposite covariate coefficients give group effects $\pm\kappa$. Thus $Y(a) \mid U, X \sim \mathrm{Bernoulli}(\mu_a(U, X))$, and the observed outcome is $Y = Y(A)$.

The bridge class is $h_a(X, W) = \phi(X, W)^\top \theta_a$ with $\|\theta\|_2 \le 2.6$. The test features are the eight indicators for (A, Z, X), and the policy class is the convex hull of the four deterministic policies on

binary X. The population moment operator has rank six and nullity two. The exact contrast set is the coupled line segment

$$\{(-\kappa + t,\ \kappa + t) : |t| \le d_\kappa\}, \qquad d_{.1} = \sqrt{\frac{65658466251}{1331969892100}} \approx 0.222023100274.$$

Replacing this segment by the Cartesian product of its two marginal intervals is the rectangular relaxation. The population objectives give optimal treatment probabilities $(0, 1)$ for the coupled set and approximately $(0.274798, 0.725202)$ for the relaxation. Evaluating the latter under the joint bridge solution set gives excess $R_0 = 0.027480$.

Enlarge each arm's basis from $(1, X, W, W^2)$ to $(1, X, W, W^2, XW, XW^2)$. The original eight coefficients retain norm at most 2.6, while the four added coefficients have norm at most η. Let $\mathcal{C}_{0,\eta}$ impose these two balls and the same exact bridge equation. Each row of Table 3 evaluates both policies using its own target $R_{0,\eta}$. Excess regret subtracts that target's optimum. Zero therefore means optimal for the stated bridge class.

Table 3: **Joint restrictions with additional bridge interactions.** These are population calculations. Excess regret and generating-world causal loss C are in 10^{-3} units, and lower is better for both. The joint policy is $(0, 1)$ throughout and has $C = 0$. The rectangular policy has $p_+ = 1 - p_-$. Numerical excess below 10^{-6} is displayed as zero.

	Excess bridge regret ↓		Rectangular policy	
η	Joint	Rectangle	p_-	$C \downarrow$
0	0.00	27.48	0.275	27.48
0.025	0.00	27.03	0.277	27.72
0.05	0.00	26.55	0.280	27.96
0.1	0.00	25.56	0.284	28.42
0.25	0.00	22.20	0.297	29.68
0.5	0.00	15.59	0.314	31.42

The joint policy attains each declared target optimum. The rectangle incurs positive excess throughout this grid, but that gap decreases as interactions expand. Its generating-world causal loss increases. The columns therefore distinguish loss against the declared bridge class from loss in the generating causal world.