CUBO: UNCERTAINTY-AWARE CAUSAL POLICY LEARNING VIA BAYESIAN BOOTSTRAP

Anonymous authors
Paper under double-blind review

ABSTRACT

Classical causal policy learning methods seek treatment rules via empirical welfare maximization. However, this can select a rule with high estimated welfare but substantial downside uncertainty, resulting in poor generalizability of the policy. We propose CUBO, an uncertainty-aware framework for causal policy learning by controlling policy-targeted uncertainty measured by welfare posterior distribution. CUBO combines common Bayesian-bootstrap weights with a nuisance posterior, inducing a joint law of estimated welfare across candidate policies, and the policy is learned by maximizing a reward that combines posterior mean welfare with lower-tail conditional value-at-risk (CVaR). We establish theoretical guarantee on welfare regret bound, and the performance is comprehensively evaluated on IHDP and ACIC datasets. This work provides a novel policy uncertainty quantification from Bayesian perspective, filling the gap of existing methods to learn trustworthy causal policy in real-world applications.

1 INTRODUCTION

Causal policy learning, also known as individualized treatment rule learning, seeks a treatment rule that assigns treatment based on individual baseline characteristics to maximize the expected welfare measured by potential outcomes (Manski, 2004; Kitagawa & Tetenov, 2018; Athey & Wager, 2021). Comparing with classical policy learning tasks, a key difference for causal policy learning is that we can only observe the factual outcome under assigned treatment, and never observe the counterfactual outcome with alternative treatment, which causes extra challenge in welfare identification, estimation and learning procedure.

Existing causal policy learning methods can be categorized as follows. Plug-in methods estimate conditional outcomes or treatment effects and choose the treatment with the largest estimated benefit based on the estimation (Qian & Murphy, 2011). Direct methods instead optimize the rule through reward maximization or weighted classification, including outcome-weighted learning and empirical welfare maximization (Zhao et al., 2012; Kitagawa & Tetenov, 2018). Orthogonal scores extend direct learning to flexible nuisance estimators and observational data (Athey & Wager, 2021). While most of these methods learn the policy by optimizing estimated welfare from empirical mean, such welfare estimation involves uncertainty, and leaving such uncertainty outside the learning objective may result in sub-optimal policy. As illustrated in Figure 1, $\widehat{W}(\pi)$ denotes the estimated welfare under policy π. The density plots indicate the distributions of the estimated welfare for two different policies. Rule π_1 has larger empirical mean welfare than π_2, so a mean-based policy learner prefers π_1. However, if the estimated welfare under π_1 has higher uncertainty and places greater mass below zero, selecting π_1 will instead yield negative welfare and hence sub-optimal treatment assignment in test data with higher probability. This observation implies that uncertainty is crucial in learning trustful policy, and how to quantify policy-targeted uncertainty and design uncertainty-aware causal policy learning method remains an important yet challenging problem.

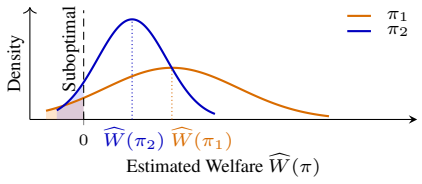


Figure 1: Estimated welfare densities. Policy π_1 has higher mean but greater uncertainty.

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.

However, developing uncertainty-aware causal policy learning method faces three important challenges. First, as effect estimation targets estimation error minimization, whereas policy learning targets welfare maximization, such a difference in target may cause existing methods in causal inference no longer an optimal choice for policy learning, such as causal effect estimation (Fernández-Loría & Provost, 2022; Frauen et al., 2025). Similarly, in terms of uncertainty quantification, CATE estimation uncertainty, measured by confidence intervals, for instance, may not be a good policy-targeted uncertainty measurement.

Second, unlike conditional average treatment effect estimation, the uncertainty in causal policy learning originates from two sources, the estimation error on nuisance parameters, and the finite-sample bias since the welfare is measured on observed samples instead of the true population distribution. This requires decomposition on causal policy uncertainty and new method to quantify and control uncertainty from both sources simultaneously.

Third, existing policy methods that implicitly take account of uncertainty through confidence intervals have limitations (Chernozhukov et al., 2025). Since the validity of confidence interval relies on asymptotic normality assumption, this limits the scope of the method and also involves further asymptotic error. Besides, the method learns the policy by maximizing a lower confidence band. While this controls the policy from producing a low welfare in least-favored cases, this is a local quantile measurement and does not characterize the welfare distribution beneath the quantile. Therefore, it is worthwhile to move beyond interval methods to seek for better uncertainty quantification under Bayesian perspective.

We address these challenges with **C**ausal **U**ncertainty via **B**ayesian b**O**otstrap (CUBO), an uncertainty-aware causal policy learning method under Bayesian framework. We start by decomposing the uncertainty into two sources, the finite-sample uncertainty and parameter-estimation uncertainty. Then, we derive posterior distributions on both sampling and parameter, which enables us to construct the posterior distribution of welfare from the joint posterior distribution. Thereon, we propose a policy-targeted uncertainty quantification from conditional value-at-risk (CVaR) on welfare posterior distribution, which measures the average welfare in the least favorable posterior fraction (Rockafellar & Uryasev, 2000). We further develop the CUBO reward as well as its realization through Bayesian bootstrap (Rubin, 1981), and design an alternating learning procedure to train the policy together with posterior updates. We derive theoretical guarantee on the method, and experiments on IHDP and ACIC datasets compare the performance of our method under five Bayesian backbones with existing causal policy learning methods.

Our main contributions can be summarized as follows:

- This paper addresses an important problem on policy uncertainty, and provides clear decomposition and quantification on uncertainty for causal policy learning.
- This work treats policy-targeted uncertainty through the posterior distribution of welfare from a novel Bayesian perspective, bridging the gap of confidence-based methods, and develops a novel uncertainty measurement through CVaR.
- We provide a thorough uncertainty-aware policy learning framework with both theoretical and experimental guaranties, enabling practitioners to learn more trustful causal policies in real-world scenarios.

Section 2 defines the causal and Bayesian decision framework. Section 3 presents the CUBO objective and algorithm. Sections 4 and 5 give the theoretical and empirical results. Section 6 discusses future works. The appendices review related work and provide additional details and proofs.

2 Causal and Bayesian policy framework

2.1 Causal policy learning

Consider an observational study with the observed data $O = (X, A, Y)$, where $X \in \mathcal{X} = \mathbb{R}^p$ is the vector of covariates, $A \in \{0, 1\}$ is treatment, and Y denotes the outcome. In the study, we observe n units randomly with observation $\mathcal{D}_n = \{O_i\}_{i=1}^n$ randomly sampled from a distribution $\mathcal{P}_0$. Let $\mathbb{E}_0$ and $\mathbb{P}_0$ be the expectations and probability taken over $\mathcal{P}_0$, and let $\mathbb{E}_n$ be the empirical mean. A policy $\pi : \mathcal{X} \to [0, 1]$ is a mapping that assigns treatment based on covariates, and we denote its feasible class as Π. Define a deterministic policy as the special case with $\pi(\mathcal{X}) = \{0, 1\}$. Let $Y(a)$ be the

potential outcomes under treatment assignment $A = a$. Let $\mu_a(x) = \mathbb{E}_0(Y \mid A = a, X = x, w_\mu)$ be the response function, $e(x) = \mathbb{P}_0(A = 1 \mid X = x, w_e)$ be the propensity, $\tau(x) = \mathbb{E}\{Y(1) - Y(0) \mid X = x, w_\tau\}$ be the Conditional Average Treatment Effect (CATE), which follow models with weights w_μ, w_e and w_τ, and denote $\hat{\mu}, \hat{e}, \hat{\tau}$ as their estimators. The true nuisance functions are $e_0(x) = \mathbb{P}_0(A = 1 \mid X = x)$, $\mu_{a,0}(x) = \mathbb{E}_0\{Y(a) \mid X = x\}$, and $\tau_0(x) = \mu_{1,0}(x) - \mu_{0,0}(x)$. Causal policy learning typically seeks to find the optimal rule $\pi^*_\Pi = \arg\max_{\pi \in \Pi} W(\pi; \mathcal{P}_0)$ that maximizes the *welfare* $W(\pi; \mathcal{P}_0)$ in the policy class, which is a deterministic function on population distribution and policy value. We measure welfare as the mean gain relative to never treating with the following welfare

$$W(\pi; \mathcal{P}_0) = \mathbb{E}_0 \left[\pi(X)\{Y(1) - Y(0)\}\right]. \tag{1}$$

Note that $W(\pi; \mathcal{P}_0) + \mathbb{E}_0 Y(0) = \mathbb{E}_0 Y(\pi(X))$, hence the optimal rule also maximizes the mean welfare. To identify the policy value, we impose the following classical causal assumptions (Imbens & Rubin, 2015):

Assumption 1 (Causal identification)**.** *The following conditions hold under $\mathcal{P}_0$.*

1. *(Consistency) If $A = a$, then $Y = Y(a)$ for $a \in \{0, 1\}$.*
2. *(Positivity) There is a constant $\kappa > 0$ such that $\kappa \leq e_0(X) \leq 1 - \kappa$ almost surely.*
3. *(Unconfoundedness) $\{Y(0), Y(1)\} \perp\!\!\!\perp A \mid X$.*

Under Assumption 1, welfare is identified by a score $\phi_{\theta^*}(O; \pi)$ at correctly specified nuisance functions with true parameters θ^*, i.e., $W(\pi; \mathcal{P}_0) = \mathbb{E}_0[\phi_{\theta^*}(O; \pi)]$. For candidate nuisance parameters θ, define $W(\pi; \mathcal{P}, \theta) = \int \phi_\theta(O; \pi)\, \mathcal{P}(dO)$ as the welfare identified over distribution $\mathcal{P}$. There are two main parametrizations on welfare, which lead to different learning procedures. Empirical welfare maximization (EWM) learns the policy directly from maximizing the empirical welfare identified by pseudo-outcomes (Kitagawa & Tetenov, 2018), e.g., for IPW pseudo outcomes,

$$W(\pi; \mathcal{P}_0, \theta) = \mathbb{E}_0 \left[\pi(X) \left\{ \frac{YA}{e(X)} - \frac{Y(1-A)}{1 - e(X)} \right\}\right], \quad \text{with } \theta = w_e,$$

and

$$\hat{\pi} = \arg\max_{\pi \in \Pi} W(\pi; \mathcal{P}_n, \hat{\theta}),$$

where $\mathcal{P}_n = n^{-1} \sum_i \delta_{O_i}$ is the empirical distribution and $\hat{\theta}$ denotes the learned parameters. In this case the welfare parameters contain only nuisance parameters, i.e, $\theta \subset \{w_\mu, w_e\}$. An alternative parametrization, referred as hybrid EWM in Kitagawa & Tetenov (2018) or effect-based offline policy learning (OPL) in Frauen et al. (2025), learns policy through causal effect estimator. The welfare is identified by

$$W(\pi; \mathcal{P}_0, \theta) = \mathbb{E}_0[\pi(X)\{\mu_1(X) - \mu_0(X)\}] = \mathbb{E}_0[\pi(X)\tau(X)], \quad \text{with } \theta = w_\tau, \tag{2}$$

the welfare can therefore be estimated from plug-in CATE estimators, $W(\pi; \mathcal{P}_n, \hat{\theta}) = \frac{1}{n} \sum_{i=1}^n \pi(X_i)\widehat{\tau}(X_i)$ by maximizing empirical welfare gain with a plug-in CATE estimator. This effect-based approach provides an interpretable CATE-based treatment rule, but it is sensitive to effect estimation error and may require a rich policy class to represent the induced rule. Despite their different strengths, direct and hybrid EWM both typically select policies by maximizing estimated mean welfare. However, as policy-value estimates are uncertain due to the existence of finite-sample bias and estimation error, treating estimated mean welfare as the only policy target can favor a policy with a high point estimate but substantial downside risk, leading to suboptimal decisions. Next, we will introduce policy-targeted uncertainty quantification that incorporates this risk into learning.

2.2 Decomposition on welfare uncertainty

Although the welfare $W(\pi; \mathcal{P}_0, \theta)$ is determined by $\mathcal{P}_0$ and parameter θ, in policy learning its empirical estimation $W(\pi; \mathcal{P}_n, \hat{\theta})$ is random. The source of randomness can be explained from the following key decomposition:

$$W(\pi; \mathcal{P}_n, \hat{\theta}) - W(\pi; \mathcal{P}_0, \theta) = \underbrace{\left[W(\pi; \mathcal{P}_n, \hat{\theta}) - W(\pi; \mathcal{P}_0, \hat{\theta})\right]}_{\text{finite-sample uncertainty}} + \underbrace{\left[W(\pi; \mathcal{P}_0, \hat{\theta}) - W(\pi; \mathcal{P}_0, \theta)\right]}_{\text{estimation uncertainty}}. \tag{3}$$

The decomposition in (3) demonstrates that the uncertainty on the estimated welfare comes from two sources, the **finite-sample approximation** and **parameter estimation**. In learning uncertainty-aware CATE, we only need to focus on the second source of uncertainty as the CATE is learned at each individual or subgroup level. However, as in policy learn the welfare is aggregated over the whole population, this further induces finite-sample uncertainty. Besides, the uncertainty on welfare and CATE are different, and policy learned from an uncertainty-aware CATE estimator does not guarantee to be more determined. The above discussion reveals fundamental differences between learning uncertainty-aware policy and CATE estimator as in Jesson et al. (2020), further addressing the importance on developing policy-targeted uncertainty quantification in the next section.

2.3 Posterior Uncertainty via Bayesian Bootstrap

Next, following the welfare uncertainty decomposition in Eq. (3), we seek to control the uncertainty induced from finite-sample bias and parameter estimation from Bayesian perspective as follows.

Finite-sample uncertainty. We represent the finite-sample uncertainty via approximating the unknown population distribution $\mathcal{P}_0$ by a posterior over candidate sampling distribution $\mathcal{P}$. From a Bayesian perspective, we place a Dirichlet-process prior on $\mathcal{P}$, whose conjugate posterior law is $\mathcal{Q}_n$:

$$\mathcal{P} \sim \mathcal{Q}_0 := \mathrm{DP}(\nu), \qquad \mathcal{P} \mid \mathcal{D}_n \sim \mathcal{Q}_n := \mathrm{DP}\left(\nu + \sum_{i=1}^{n} \delta_{O_i}\right), \tag{4}$$

where ν is a finite base measure and δ_{O_i} is a Dirac mass at observation O_i. The Dirichlet process imposes no finite-dimensional parametric form on $\mathcal{P}$, while conjugacy makes posterior simulation tractable (Ye, 2026). Letting the total mass of ν approach zero yields the Bayesian-bootstrap posterior (Rubin, 1981), which assigns random weights to the observed units.

Parameter uncertainty. To control the parameter uncertainty, for fixed policy π, our target is to approximate $\delta_{W(\pi;\mathcal{P}_0,\theta^*)}$, the Dirac distribution of welfare at the true parameter θ^*, with the posterior distribution of $W(\pi;\mathcal{P}_0,\theta)$. Inspired by the idea of Bayesian neural network in causal inference (Jesson et al., 2020), we treat the model weights $\theta \in \mathbb{R}^d$ as random variables, assign a standard Gaussian prior, and update it using the observed-data likelihood:

$$\theta \sim \mathcal{P}_{\theta,0} = \mathcal{N}(0, I_d), \qquad p(\theta \mid \mathcal{D}_n) = \frac{p_0(\theta) L_n(\theta)}{\int p_0(u) L_n(u)\, du}, \qquad \mathcal{P}_{\theta,n}(d\theta) = p(\theta \mid \mathcal{D}_n)\, d\theta, \tag{5}$$

where p_0 is the prior density and L_n is the observed-data likelihood: $\prod_i p_\theta(Y_i \mid X_i, A_i)$ for outcome models or $\prod_i p_\theta(A_i \mid X_i)$ for propensity models. Following Jesson et al. (2020), we fit a variational approximation $q_n(\theta) \approx p(\theta \mid \mathcal{D}_n)$ before policy learning. Appendix B.2 gives the fitting and parameter-sampling details. The parameter posterior induces the welfare posterior by integration. Specifically, for any measurable set $C \subseteq \mathbb{R}$,

$$\mathbb{P}(W(\pi;\mathcal{P}_0,\theta) \in C \mid \mathcal{D}_n) = \int \mathbf{1}\{W(\pi;\mathcal{P}_0,\theta) \in C\}\, \mathcal{P}_{\theta,n}(d\theta). \tag{6}$$

Thus parameter draws $\theta_s \sim \mathcal{P}_{\theta,n}$ give welfare draws $W(\pi;\mathcal{P}_0,\theta_s)$, whose empirical distribution approximates the induced welfare posterior.

Combining the posteriors. To obtain the posterior distribution of $W(\pi;\mathcal{P},\theta)$, we adopt conditional independence of $\mathcal{P}$ and θ given the observed data $\mathcal{D}_n$. The resulting posterior of estimated welfare is

$$\mathbb{P}(W(\pi;\mathcal{P},\theta) \in C \mid \mathcal{D}_n) = \iint \mathbf{1}\{W(\pi;\mathcal{P},\theta) \in C\}\, \mathcal{Q}_n(d\mathcal{P})\, \mathcal{P}_{\theta,n}(d\theta). \tag{7}$$

This posterior can be approximated by pairing Bayesian-bootstrap draws of $\mathcal{P}$ with independent draws of θ from q_n. More generally, a coupled model would derive a joint posterior retaining their dependence; see Lyddon et al. (2018); Stephens et al. (2023) for example. Such dependent posterior constructions are outside the scope of this paper.

3 Method

The framework in Section 2 gives a posterior distribution of estimated welfare for each candidate policy π. In this section, we first discuss how to use this posterior distribution to construct policy-targeted Bayesian uncertainty quantification. Then we incorporate it into the learning objective, and finally describe posterior sampling and optimization procedure.

3.1 POLICY-TARGETED BAYESIAN UNCERTAINTY QUANTIFICATION

To construct uncertainty quantification for policy learning, we start by clarifying its key differences from uncertainty quantification on CATE estimators. First, the CATE uncertainty concerns how precisely the treatment effect is estimated at a given covariate value. Policy learning instead concerns the welfare obtained by assigning treatments across the population, which induces further finite-sample uncertainty as described in Section 2.2. In addition, welfare uncertainty has **uneven importance** across its posterior distribution. For a decision maker concerned about downside risk, posterior distribution on estimated welfare with heavier lower tail are more likely to provide poor policy than equally uncertain distribution. In contrast, uncertainty quantification on CATE via variance or confidence intervals do not directly express this asymmetric preference for downstream policy making. The above observation well motivates us to construct the following policy-targeted uncertainty quantification by controlling the lower tail of the policy's welfare posterior distribution.

For a fixed π, consider the posterior distribution of the welfare $W(\pi; \mathcal{P}, \theta)$ in (7). All posterior probabilities and expectations below condition on $\mathcal{D}_n$ and integrate over both $\mathcal{P} \sim \mathcal{Q}_n$ and $\theta \sim \mathcal{P}_{\theta,n}$. For $\alpha \in (0,1)$, let $q_{\alpha,n}(\pi)$ be the α-quantile of the posterior distribution. For continuous posterior, the lower-tail conditional value-at-risk (CVaR) is defined as

$$C_{\alpha,n}(\pi) = \mathbb{E}[W(\pi; \mathcal{P}, \theta) \mid W(\pi; \mathcal{P}, \theta) \leq q_{\alpha,n}(\pi), \mathcal{D}_n]\,. \tag{8}$$

The CVaR reward averages the lowest α fraction of posterior welfare values. A larger $C_{\alpha,n}(\pi)$ indicates better average welfare in this adverse tail; it does not by itself order the probabilities of falling below every welfare threshold. Its general definition, valid also for discrete posterior draws, is (Rockafellar & Uryasev, 2000):

$$C_{\alpha,n}(\pi) = \sup_{t \in \mathbb{R}} \left[t - \frac{1}{\alpha} \mathbb{E}\Big\{ \big(t - W(\pi; \mathcal{P}, \theta)\big)_+ \mid \mathcal{D}_n \Big\} \right], \tag{9}$$

where $(z)_+ = \max\{z, 0\}$. This representation introduces a scalar threshold t and expresses CVaR through a posterior expectation of the shortfall below it. It permits Monte Carlo estimation (Rockafellar & Uryasev, 2000) and the learning objective below.

Comparison with policy learning with confidence. To further rationalize our policy-targeted uncertainty quantification, we provide the following work which on policy learning with confidence interval which implicitly takes account of policy uncertainty from a frequentist perspective (Chernozhukov et al., 2025). The policy learning with confidence rule maximizes the lower confidence bound:

$$\widehat{\pi}_{\text{conf}} \in \arg\max_{\pi \in \Pi} \left\{ \widehat{W}(\pi) - \widehat{c}_{1-\beta} \widehat{s}(\pi) \right\}. \tag{10}$$

Here $\widehat{W}(\pi)$ is an empirical welfare estimate, $\widehat{s}(\pi)$ its estimated standard error, and $\widehat{c}_{1-\beta}$ a critical value calibrated over the policy class. Both criteria favor protection against low welfare. The confidence method optimizes a simultaneous lower bound with repeated-sample calibration; that bound is not a posterior quantile. CUBO instead averages the entire lower posterior tail, responding to the severity of adverse welfare values beyond a single boundary. This matches our posterior downside criterion, while providing a different uncertainty summary from frequentist confidence guarantees.

3.2 CUBO LEARNING OBJECTIVE

We first estimate the nuisance models and their parameter posterior using the observed data, and then hold this fitted posterior fixed during policy learning. For a policy π, let $\overline{W}_n(\pi) = \mathbb{E}\{W(\pi; \mathcal{P}, \theta) \mid \mathcal{D}_n\}$ denote posterior mean welfare. CUBO balances this mean with the lower-tail CVaR in (9):

$$\begin{aligned} R_{\alpha,\lambda,n}(\pi) &= (1-\lambda)\overline{W}_n(\pi) + \lambda C_{\alpha,n}(\pi) \\ &= \overline{W}_n(\pi) - \lambda\{\overline{W}_n(\pi) - C_{\alpha,n}(\pi)\}, \qquad \lambda \in [0,1]. \end{aligned} \tag{11}$$

The second expression penalizes the gap between posterior mean welfare and its lower-tail average. Larger λ places more weight on unfavorable posterior scenarios; $\lambda = 0$ recovers posterior-mean policy learning and $\lambda = 1$ maximizes lower-tail welfare alone.

Let η parameterize the policy π_η. Given an estimator $\widehat{R}_{\alpha,\lambda,n}$ of this reward, the policy-learning objective is

$$\widehat{\eta} \in \arg\min_{\eta} \widehat{\mathcal{L}}_{\alpha,\lambda,n}(\eta), \qquad \widehat{\mathcal{L}}_{\alpha,\lambda,n}(\eta) := -\widehat{R}_{\alpha,\lambda,n}(\pi_\eta). \tag{12}$$

Only the policy parameters are optimized in this stage. Nuisance-estimation losses are used beforehand to fit the welfare models and their posterior approximation, as detailed in Appendix B.1. The next subsection constructs a Monte Carlo reward estimator and an alternating optimization procedure.

3.3 Sampling and Learning Procedure

Posterior welfare draws. In nuisance posterior distribution, learning q_n from the fitted approximation to $\mathcal{P}_{\theta,n}$ through minimizing the BNN loss function in Appendix B.2. Following (7), draw S pairs of population weights B_s and nuisance parameters θ_s independently from

$$B_s = (B_{1s}, \ldots, B_{ns}) \sim \mathrm{DP}(1), \qquad \mathcal{P}_s = \sum_{i=1}^{n} B_{is}\delta_{O_i}, \qquad \theta_s \sim q_n. \tag{13}$$

Recall that $\phi_\theta(O;\pi)$ is the score that identifies welfare $W(\pi;\mathcal{P},\theta)$ with its expectation on θ. Each pair (B_s,θ_s) gives a posterior draw of estimated welfare:

$$W_s(\pi_\eta) = W(\pi_\eta;\mathcal{P}_s,\theta_s) = \sum_{i=1}^{n} B_{is}\phi_{\theta_s}(O_i;\pi_\eta). \tag{14}$$

We retain the same pairs (B_s,θ_s) throughout policy learning, so candidate policies are compared under common posterior scenarios. To separate the nuisance estimation with policy learning, we apply classical cross-fitting technique such that each observation uses the nuisance draw fitted without its fold. Appendix B.2 describes the details on nuisance fitting and sampling, while Appendix B.3 gives the specific realizations of ϕ_θ.

Loss estimation and learning procedure. Substituting the posterior welfare draws into (11) and using the CVaR representation in (9), we construct the reward estimation as follows

$$\widehat{R}_{\alpha,\lambda,n}(\eta,t) = \frac{1-\lambda}{S}\sum_{s=1}^{S} W_s(\pi_\eta) + \lambda\left[t - \frac{1}{\alpha S}\sum_{s=1}^{S}\big(t - W_s(\pi_\eta)\big)_+\right], \tag{15}$$

and therefore the estimated CUBO reward in (12) is $\widehat{R}_{\alpha,\lambda,n}(\pi_\eta) = \max_t \widehat{R}_{\alpha,\lambda,n}(\eta,t)$, so policy learning minimizes the negative-reward loss $\widehat{\mathcal{L}}_{\alpha,\lambda,n}(\eta,t) = -\widehat{R}_{\alpha,\lambda,n}(\eta,t)$ jointly over (η,t). For a fixed policy and $\lambda > 0$, the loss is convex in threshold t, and the optimal t is given by the α-quantile of the posterior welfare distribution (Rockafellar & Uryasev, 2000).

The policy learning follows from an alternating procedure between threshold t and policy updates. With η fixed, minimize the loss over t by subgradient descent. Then hold t fixed and update η, recomputing welfare after each policy update. Both updates optimize the same negative-reward loss, while q_n and the sampled pairs (B_s,θ_s) remain fixed. The architecture and learning procedure is summarized in Figure 2, with the algorithm given in Algorithm 1 in Appendix B.4.

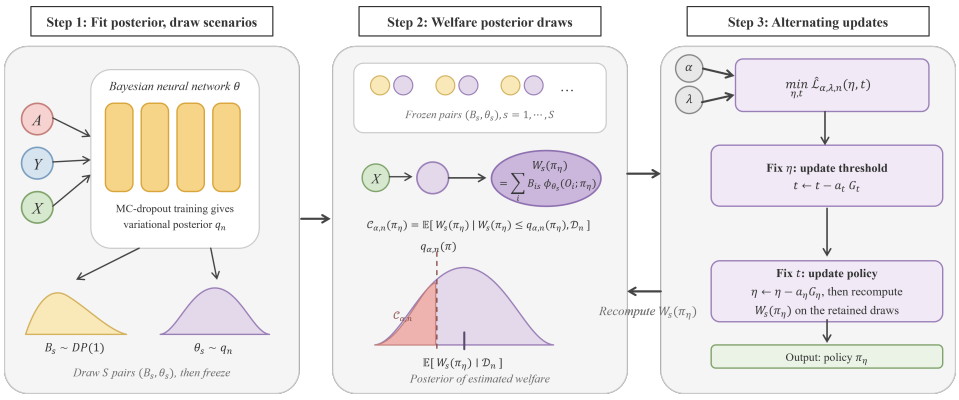


Figure 2: Architecture and learning procedure of CUBO. After nuisance fitting and posterior sampling, each iteration updates the CVaR threshold and the policy alternately.

4 THEORY

Let $I_1, \ldots, I_K$ be the data folds for cross-fitting, with fixed K and sizes $n_k = |I_k| \geq cn$ for fixed $c > 0$. Write q_{-k} for the posterior of nuisance parameters fitted without fold I_k, and q_n for the joint law of $\theta = (\theta_1, \ldots, \theta_K)$ with marginals $\theta_k \sim q_{-k}$. We introduce the following assumption on the sampling and complexity of policy class.

Assumption 2 (Sampling and policy-class regularity)**.** *Suppose that $\mathbb{E}_0|Y(0)| + \mathbb{E}_0|Y(1)| < \infty$. For each fold, the welfare score $\phi_{\theta_k}(O;\pi) = \pi(X)\phi_{\theta_k}(O;1)$, where 1 denotes the treat-all rule, and $|\phi_{\theta_k}(O;1)| \leq M$ for constant $M \in (0,\infty)$. The class Π is nonempty and pointwise measurable, and its subgraphs $\{(x,u) \in \mathcal{X} \times [0,1] : u < \pi(x)\}$ form a VC class of dimension $v < \infty$ (van der Vaart & Wellner, 2023). In other words, for a probability law Q on $\mathcal{X}$, let $N(\epsilon, \Pi, L_2(Q))$ denote the covering number of Π under $\|f\|_{L_2(Q)}$. Then there exists an entropy constant $A_\Pi \geq e$ such that for $0 < \epsilon \leq 1$,*

$$\sup_Q \log N(\epsilon, \Pi, L_2(Q)) \leq v \log(A_\Pi/\epsilon). \tag{16}$$

With $W(\pi; \mathcal{P}_0, \theta_k)$ with parameter θ_k following distribution q_{-k}, define the orthogonal bias score

$$b_n = \sum_{k=1}^{K} \frac{n_k}{n} \int \sup_{\pi \in \Pi} |W(\pi; \mathcal{P}_0, \theta_k) - W(\pi; \mathcal{P}_0)| \, q_{-k}(d\theta_k). \tag{17}$$

Theorem 1 (Posterior welfare-regret bound)**.** *Under Assumptions 1 and 2, let $r_n > 0$ be a deterministic sequence tending to zero such that $b_n = O_{\mathbb{P}_0}(r_n)$. For $\pi^*_\Pi(\mathcal{P}, \theta) \in \arg\max_{\pi \in \Pi} W(\pi; \mathcal{P}, \theta)$ and every deterministic sequence $C_n \to \infty$,*

$$\mathbb{P}_0\Big[W(\pi^*_\Pi; \mathcal{P}_0) - W\{\pi^*_\Pi(\mathcal{P}, \theta); \mathcal{P}_0\} > C_n\{M\sqrt{v/n} + r_n\} \,\Big|\, \mathcal{D}_n\Big] \xrightarrow{\mathbb{P}_0} 0. \tag{18}$$

Theorem 1 bounds the welfare regret of the policy optimized under a posterior draw by the complexity of the class and the nuisance-estimation error. The leading term $M\sqrt{v/n}$ captures the complexity of the policy class, while r_n accounts for posterior uncertainty and nuisance estimation. When $r_n = O(\sqrt{v/n})$, the regret attains the same rate as standard EWM (Ye, 2026). The bias score b_n, and hence the optimal r_n, can be further controlled by a product of nuisance errors, see Proposition 1 in Appendix C for details.

For policy π and threshold t, recall that $\widehat{\mathcal{L}}_{\alpha,\lambda,n}(\pi, t)$ is the loss defined in (15) and set $\mathcal{L}_{\alpha,\lambda,n}(\pi, t) = \mathbb{E}[\widehat{\mathcal{L}}_{\alpha,\lambda,n}(\pi, t) \mid \mathcal{D}_n]$ at fixed (π, t). We have the following concentration bound on the loss, uniform over the stochastic class and conditional on $\mathcal{D}_n$.

Theorem 2 (Uniform loss concentration)**.** *Under Assumptions 1 and 2, generate S iid pairs of (B_s, θ_s) from (13). Then there is a constant $C_\Pi > 0$ depending only on A_Π such that, for fixed $\alpha \in (0,1]$, $\lambda \in [0,1]$, and failure probability $\delta_S \in (0,1)$, with probability at least $1 - \delta_S$ conditional on $\mathcal{D}_n$, integrating only the simulation draws,*

$$\sup_{\pi \in \Pi,\, t \in [-M,M]} |\widehat{\mathcal{L}}_{\alpha,\lambda,n}(\pi, t) - \mathcal{L}_{\alpha,\lambda,n}(\pi, t)| \leq C_\Pi M \left(1 - \lambda + \frac{\lambda}{\alpha}\right) \sqrt{\frac{v + 1 + \log(1/\delta_S)}{S}}, \tag{19}$$

Theorem 2 controls the finite-sample error from approximating the posterior objective using S posterior draws. The factor $1 - \lambda + \lambda/\alpha$ captures the estimation error from the CVaR parameters and increases as α decreases when $\lambda > 0$, since CVaR is then estimated from a smaller tail fraction of the draws. The error is $O(S^{-1/2})$ in the number of simulation draws. The result can be further extended to concentration bound on welfare, see Corollary 2 in Appendix C. Proofs of Theorems 1 and 2 are given in Appendix E.

5 EXPERIMENTS

5.1 EXPERIMENTAL SETUP

Datasets. We evaluate our method in two well-known real-world benchmark datasets for causal inference, IHDP (Hill, 2011; Dehejia, 2005) and ACIC 2016 (Dorie et al., 2019), following the data preprocessing in Jesson et al. (2020). Appendix D.1 reports the preprocessing, split construction, and the scope of the evaluation.

Table 1: Regret and policy accuracy on IHDP and ACIC. Regret is multiplied by 10.

		IHDP				ACIC			
		In-sample		Out-of-sample		In-sample		Out-of-sample	
Family	Method	Regret	PA	Regret	PA	Regret	PA	Regret	PA
Baseline	T-Learner	0.022	97.4%	0.122	94.7%	0.160	93.4%	0.155	94.0%
	CFRNet	0.038	96.4%	0.340	86.7%	0.148	94.3%	0.186	94.0%
	TARNet	0.048	96.8%	0.295	88.0%	0.159	93.4%	0.138	94.6%
	Dragonnet	0.014	98.1%	0.399	84.0%	0.179	93.8%	0.197	93.5%
	CEVAE	0.039	97.0%	0.140	89.3%	0.485	91.0%	0.555	90.2%
	DeRCFR	0.009	98.3%	0.030	97.3%	0.074	95.1%	0.107	94.8%
	ESCFR	0.021	97.7%	0.095	93.3%	0.117	94.6%	0.129	94.2%
	PT-CATE	0.063	96.2%	0.202	93.3%	0.513	89.1%	0.360	91.4%
	DR-EWM	0.067	95.3%	0.174	92.8%	0.157	94.1%	0.239	92.8%
	PoLeCe	0.060	95.1%	0.235	90.7%	0.145	94.7%	0.148	94.0%
CUBO	BCFRNet	0.008	98.7%	0.079	96.1%	0.088	95.2%	0.118	94.8%
	BTARNet	0.016	98.1%	0.226	90.0%	0.083	95.3%	0.096	95.1%
	BT-Learner	0.013	97.9%	**0.013**	97.3%	0.095	95.2%	0.121	94.9%
	BDragonnet	**0.003**	98.9%	0.183	93.3%	0.156	93.5%	0.175	93.3%
	BDeRCFR	**0.003**	**99.4%**	0.014	**98.7%**	**0.069**	**96.1%**	**0.090**	**95.6%**

Baselines. We compare against two families of baselines. (1) *CATE plug-ins* estimate the conditional average treatment effect $\tau(x)$ and deploy $\mathbf{1}\{\widehat{\tau}(x) > 0\}$: T-Learner (Künzel et al., 2019), CFRNet and TARNet (Shalit et al., 2017), Dragonnet (Shi et al., 2019), CEVAE (Louizos et al., 2017), DeRCFR (Wu et al., 2023), and ESCFR (Wang et al., 2023). (2) *Policy learners* optimize a decision rule directly: PT-CATE (Frauen et al., 2025), which retargets CATE estimation toward decision performance, DR-EWM (Athey & Wager, 2021), which maximizes doubly robust scores, and PoLeCe (Chernozhukov et al., 2025), which maximizes a lower confidence bound on welfare. Appendix D.2 gives implementation details.

Evaluation metrics. For an evaluation split of size m, let τ_i be the true CATE conditioning on X_i and $\pi_i = \pi(X_i)$ the treatment probability assigned by a policy, and write $D_m = m^{-1}\sum_i |\tau_i| > 0$. We report the normalized regret $\mathrm{Reg} = m^{-1}\sum_{i=1}^{m} \left[\max(\tau_i, 0) - \pi_i\tau_i\right] / D_m$, which averages the welfare loss of π relative to the oracle rule $\mathbf{1}\{\tau(x) > 0\}$ over the split, and the policy accuracy $\mathrm{PA} = m^{-1}\sum_{i=1}^{m}\left[\pi_i\mathbf{1}\{\tau_i > 0\} + (1-\pi_i)\mathbf{1}\{\tau_i \le 0\}\right]$ over the optimal policy from true effect.

Model setting. CUBO pairs a Bayesian nuisance posterior with policy class Π. We build the nuisance posterior with dropout on five Bayesian CATE backbones: BCFRNet, BTARNet, BT-Learner, BDragonnet, and BDeRCFR. Following Jesson et al. (2020), each Bayesian backbone shares architecture, Gaussian outcome likelihood, and training data with its frequentist counterpart, which disables posterior dropout. Given the posterior-mean effect learner $s(x)$, the policy class is chosen as $\pi_\eta(x) = \sigma\big[\{1 + g_\eta(x)\}s(x)/T + b_\eta(x)\big]$, where (g_η, b_η) are the two five-hidden-layer width-64 ELU networks, T is a fixed temperature, and the initial policy is set as the posterior-mean plug-in. The hyperparameters are chosen as $(\alpha, \lambda) = (0.1, 0.5)$ and we apply $S = 100$ rounds of posterior draws. Appendix D specifies the remaining configurations.

5.2 Benchmark comparison

Table 1 reports in-sample and out-of-sample regret and policy accuracy on IHDP and ACIC. We feature three important observations from the result. (1) Comparing with plug-in methods, CUBO improves policy learning with less regret in every matched backbone on IHDP and on ACIC, and with higher policy accuracy in all but one matched pair. For example, BDragonnet reduces Dragonnet's test regret from 0.399 to 0.183, and BCFRNet reduces CFRNet's from 0.340 to 0.079, while CUBO with the BDeRCFR backbone improves regret from 0.030 to 0.014 over the strongest baseline plug-in. Figure 3 provides a straightforward comparison between the Non-Bayesian plug-in rule and CUBO under three different backbones. (2) CUBO attains the best overall test regret among all methods on both datasets. On IHDP, BT-Learner reaches 0.013, less than half of the test regret 0.030 of the strongest baseline DeRCFR, and BDeRCFR reaches 0.014 with the highest accuracy 98.7%.

(3) Some plug-in baselines have diverging performance on in-sample and out-of-sample data. For example, Dragonnet attains 0.014 in-sample regret yet the highest test regret 0.399, and CEVAE shows low in-sample policy accuracy 97.0% with a comparatively high out-of-sample policy accuracy 89.3%, attributable to the misallocation of a few high-uncertainty test units. This divergence further evidences that targeting the welfare distribution improves generalization.

The results on ACIC support similar conclusion. CUBO with the BDeRCFR backbone reaches the lowest test regret 0.090, a 15.9% decrease from the strongest baseline DeRCFR, and also attains the highest test policy accuracy 95.6%. CUBO with the BDragonnet backbone, as another example, improves over the plug-in counterpart by 11.2% in test regret.

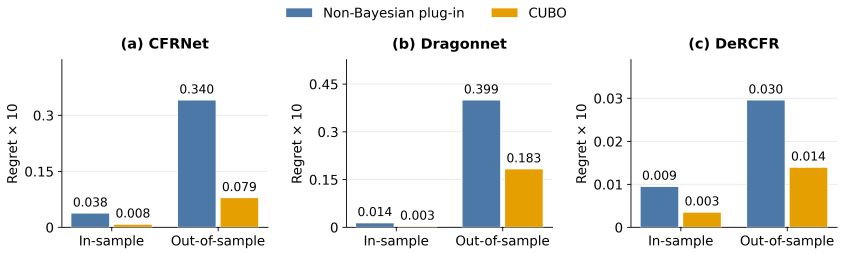


Figure 3: IHDP regret between non-Bayesian plug-in rules and CUBO counterpart.

5.3 Reward and policy-depth ablations

In ablation study, we evaluate the sensitivity of CUBO method with BDeRCFR backbone on IHDP validation data with regard to three hyperparameters, the CVaR tail α, reward weight λ and policy-class complexity featured by the network depth L. Figure 4 varies one setting at a time around the default $(\alpha, \lambda, L) = (0.1, 0.5, 5)$. With mean validation regret minimized at $\lambda = 0.75$ and $\alpha = 0.30$, the ablation on λ implies a mean–tail trade-off, implying that both mean reward maximization and uncertainty control through CVaR are important for policy learning. The ablation on α balances reward estimation error against uncertainty awareness. Larger α improves CVaR estimation with more tail samples, but the measure approaches the posterior mean and becomes less uncertainty-aware. Finally, the depth ablation shows that policy-class complexity affects generalization. While five to ten layers give stable validation regret, shallower policies reverse a few high-benefit units despite slightly higher fitted training reward. The optimal choice is five layers, balancing policy accuracy and computation efficiency.

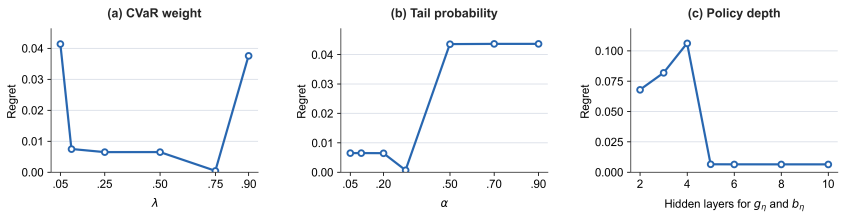


Figure 4: CUBO sensitivity on IHDP validation data. Regret is multiplied by 10.

6 Discussion

In this work, we introduced CUBO, an uncertainty-aware causal policy learning framework that incorporates posterior welfare uncertainty directly into policy optimization. By combining Bayesian-bootstrap sampling with a variational nuisance posterior, CUBO induces a joint welfare distribution across candidate policies and optimizes a reward balancing posterior mean welfare with lower-tail CVaR. CUBO distinguishes aggregate welfare uncertainty from pointwise treatment-effect risk, decouples nuisance fitting from policy updates in a model-agnostic algorithm, and establishes regret guarantees separating policy complexity, nuisance accuracy, and finite posterior simulation. Future directions include extending the lower-tail CVaR objective to continuous or multi-action treatments, incorporating sensitivity parameters for unmeasured confounding, and coupling lower-tail welfare with selective classification to defer high-uncertainty decisions.

Ethics Statement

We have read and adhere to the ICLR Code of Ethics throughout this research. This work does not involve new human-subject recruitment, data collection, or intervention. The real-world datasets are used only for method evaluation. The proposed framework is intended for methodological research, and any real-world deployment would require application-specific consideration of ethical, legal, and fairness requirements.

Reproducibility Statement

The implementation details on experimental configurations have been provided in the appendix. The source code and scripts for reproducing the reported results is provided at the following anonymous repository: `https://anonymous.4open.science/r/CUBO_ICLR-9C58`.

AI use statement

In this work,we used the generative-AI assistant for editing and polishing parts of the manuscript, formatting tables and references, and drafting visualization scripts for experimental figures. We have not used generative-AI tools for problem formulation, theoretical analysis, experimental design or interpretation of results. All AI-assisted text and artifacts were reviewed, verified, and revised by the authors. We take responsibility for the final content of this work, including any text or artifacts produced with the aid of generative AI.

References

Susan Athey and Stefan Wager. Policy learning with observational data. *Econometrica*, 89(1): 133–161, 2021. doi: 10.3982/ECTA15732.

Eli Ben-Michael, D. James Greiner, Kosuke Imai, and Zhichao Jiang. Safe policy learning through extrapolation: Application to pre-trial risk assessment. *Journal of the American Statistical Association*, 120(551):1386–1399, 2025. doi: 10.1080/01621459.2025.2489135.

Gary Chamberlain. Bayesian aspects of treatment choice. In John Geweke, Gary Koop, and Herman Van Dijk (eds.), *The Oxford Handbook of Bayesian Econometrics*, pp. 10–39. Oxford University Press, 2011. doi: 10.1093/oxfordhb/9780199559084.013.0002.

Victor Chernozhukov, Sokbae Lee, Adam M. Rosen, and Liyang Sun. Policy learning with confidence. *arXiv preprint arXiv:2502.10653*, 2025.

Rajeev H. Dehejia. Program evaluation as a decision problem. *Journal of Econometrics*, 125(1–2): 141–173, 2005. doi: 10.1016/j.jeconom.2004.04.006.

Vincent Dorie, Jennifer Hill, Uri Shalit, Marc Scott, and Dan Cervone. Automated versus do-it-yourself methods for causal inference: Lessons learned from a data analysis competition. *Statistical Science*, 34(1):43–68, 2019. doi: 10.1214/18-STS667.

Thomas S. Ferguson. A Bayesian analysis of some nonparametric problems. *The Annals of Statistics*, 1(2):209–230, 1973. doi: 10.1214/aos/1176342360.

Carlos Fernández-Loría and Foster Provost. Causal decision making and causal effect estimation are not the same... and why it matters. *INFORMS Journal on Data Science*, 1(1):4–16, 2022. doi: 10.1287/ijds.2021.0006.

Dylan J. Foster and Vasilis Syrgkanis. Orthogonal statistical learning. *The Annals of Statistics*, 51(3): 879–908, 2023. doi: 10.1214/23-AOS2258.

Dennis Frauen, Valentyn Melnychuk, Jonas Schweisthal, Mihaela van der Schaar, and Stefan Feuerriegel. Treatment effect estimation for optimal decision-making. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-4464.

Yarin Gal and Zoubin Ghahramani. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 1050–1059, 2016.

P. Richard Hahn, Jared S. Murray, and Carlos M. Carvalho. Bayesian regression tree models for causal inference: Regularization, confounding, and heterogeneous effects (with discussion). *Bayesian Analysis*, 15(3):965–1056, 2020. doi: 10.1214/19-BA1195.

Jennifer L. Hill. Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 20(1):217–240, 2011. doi: 10.1198/jcgs.2010.08162.

Guido W. Imbens and Donald B. Rubin. *Causal Inference for Statistics, Social, and Biomedical Sciences*. Cambridge University Press, Cambridge, 2015. doi: 10.1017/CBO9781139025751.

Andrew Jesson, Sören Mindermann, Uri Shalit, and Yarin Gal. Identifying causal-effect inference failure with uncertainty-aware models. In *Advances in Neural Information Processing Systems*, volume 33, pp. 11637–11649, 2020.

Nathan Kallus. Treatment effect risk: Bounds and inference. *Management Science*, 69(8):4579–4590, 2023. doi: 10.1287/mnsc.2023.4819.

Nathan Kallus and Angela Zhou. Confounding-robust policy improvement. In *Advances in Neural Information Processing Systems*, volume 31, pp. 9269–9279, 2018.

Masahiro Kato. General Bayesian policy learning. In *Proceedings of the 42nd Conference on Uncertainty in Artificial Intelligence*, volume 337 of *Proceedings of Machine Learning Research*, pp. 2792–2827, 2026.

Toru Kitagawa and Aleksey Tetenov. Who should be treated? Empirical welfare maximization methods for treatment choice. *Econometrica*, 86(2):591–616, 2018. doi: 10.3982/ECTA13288.

Sören R. Künzel, Jasjeet S. Sekhon, Peter J. Bickel, and Bin Yu. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the National Academy of Sciences*, 116(10):4156–4165, 2019. doi: 10.1073/pnas.1804597116.

Brent R. Logan, Rodney Sparapani, Robert E. McCulloch, and Purushottam W. Laud. Decision making and uncertainty quantification for individualized treatments using Bayesian additive regression trees. *Statistical Methods in Medical Research*, 28(4):1079–1093, 2019. doi: 10.1177/0962280217746191.

Christos Louizos, Uri Shalit, Joris M. Mooij, David Sontag, Richard Zemel, and Max Welling. Causal effect inference with deep latent-variable models. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Alexander R. Luedtke and Mark J. van der Laan. Statistical inference for the mean outcome under a possibly non-unique optimal treatment strategy. *The Annals of Statistics*, 44(2):713–742, 2016. doi: 10.1214/15-AOS1384.

Simon Lyddon, Stephen Walker, and Chris C. Holmes. Nonparametric learning from Bayesian models with randomized objective functions. In *Advances in Neural Information Processing Systems*, volume 31, pp. 2071–2081, 2018.

Charles F. Manski. Statistical treatment rules for heterogeneous populations. *Econometrica*, 72(4):1221–1246, 2004. doi: 10.1111/j.1468-0262.2004.00530.x.

Min Qian and Susan A. Murphy. Performance guarantees for individualized treatment rules. *The Annals of Statistics*, 39(2):1180–1210, 2011. doi: 10.1214/10-AOS864.

R. Tyrrell Rockafellar and Stanislav Uryasev. Optimization of conditional value-at-risk. *The Journal of Risk*, 2(3):21–41, 2000. doi: 10.21314/JOR.2000.038.

Donald B. Rubin. The Bayesian bootstrap. *The Annals of Statistics*, 9(1):130–134, 1981. doi: 10.1214/aos/1176345338.

Uri Shalit, Fredrik D. Johansson, and David Sontag. Estimating individual treatment effect: Generalization bounds and algorithms. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 3076–3085, 2017.

Claudia Shi, David M. Blei, and Victor Veitch. Adapting neural networks for the estimation of treatment effects. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

David A. Stephens, Widemberg S. Nobre, Erica E. M. Moodie, and Alexandra M. Schmidt. Causal inference under mis-specification: Adjustment based on the propensity score (with discussion). *Bayesian Analysis*, 18(2):639–694, 2023. doi: 10.1214/22-BA1322.

Aad W. van der Vaart and Jon A. Wellner. *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer, 2 edition, 2023. doi: 10.1007/978-3-031-29040-4.

Hao Wang, Jiajun Fan, Zhichao Chen, Haoxuan Li, Weiming Liu, Tianqiao Liu, Quanyu Dai, Yichao Wang, Zhenhua Dong, and Ruiming Tang. Optimal transport for treatment effect estimation. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Anpeng Wu, Junkun Yuan, Kun Kuang, Bo Li, Runze Wu, Qiang Zhu, Yueting Zhuang, and Fei Wu. Learning decomposed representations for treatment effect estimation. *IEEE Transactions on Knowledge and Data Engineering*, 35(5):4989–5001, 2023. doi: 10.1109/TKDE.2022.3150807.

Haonan Ye. Nonparametric Bayesian policy learning. *arXiv preprint arXiv:2605.17068*, 2026.

Yingqi Zhao, Donglin Zeng, A. John Rush, and Michael R. Kosorok. Estimating individualized treatment rules using outcome weighted learning. *Journal of the American Statistical Association*, 107(499):1106–1118, 2012. doi: 10.1080/01621459.2012.695674.

A Related work

This appendix distinguishes CUBO from three neighboring lines of work. The distinctions concern the decision target, the source of uncertainty, and whether uncertainty enters reporting or training.

Causal policy and individualized treatment-rule learning. Statistical treatment-rule and ITR methods seek rules that maximize expected outcomes rather than population-average effects (Manski, 2004; Qian & Murphy, 2011). Plug-in approaches estimate conditional outcomes or treatment effects and assign treatment according to the estimated benefit. Direct approaches optimize a welfare surrogate. Outcome-weighted learning formulates treatment assignment as weighted classification (Zhao et al., 2012), while empirical welfare maximization searches for the best rule within a constrained class (Kitagawa & Tetenov, 2018). Doubly robust and orthogonal methods allow flexible nuisance estimators while controlling first-order bias (Athey & Wager, 2021; Foster & Syrgkanis, 2023). Recent decision-targeted work shows why CATE accuracy need not imply accurate treatment decisions (Fernández-Loría & Provost, 2022). The approach of Frauen et al. (2025) therefore retargets two-stage CATE estimators toward decision performance. CUBO shares this policy-centered target and adds distribution-level control of welfare uncertainty.

Uncertainty in causal inference and policy learning. Causal uncertainty quantification usually concerns average or conditional treatment effects. The individual-treatment-effect CVaR studied by Kallus (2023) measures risk across people rather than uncertainty in a policy's aggregate welfare. Inference for an optimal treatment rule is harder because the rule is selected through optimization and the optimum may not be unique (Luedtke & van der Laan, 2016). Bayesian response models such as BART provide posterior distributions for response surfaces and heterogeneous effects (Hill, 2011; Hahn et al., 2020). These distributions can be propagated to individualized recommendations and to the welfare of a prediction-based ITR (Logan et al., 2019). Frequentist policy methods instead use confidence bounds or robust objectives. Policy Learning with Confidence maximizes a simultaneous lower confidence bound built from estimated welfare, standard errors, and a Gaussian-bootstrap critical value (Chernozhukov et al., 2025). Safe and confounding-robust policy learning address extrapolation and identification uncertainty (Kallus & Zhou, 2018; Ben-Michael et al., 2025). CUBO maintains causal identification and places the lower tail of fitted welfare uncertainty inside training.

Bayesian policy learning and posterior risk. Bayesian analyses of treatment choice evaluate actions by integrating policy utility over a posterior distribution (Dehejia, 2005; Chamberlain, 2011). Model-based approaches propagate posterior response surfaces into individualized decisions (Logan et al., 2019). Nonparametric Bayesian policy learning instead places a Dirichlet-process posterior on the reduced-form distribution that determines welfare. This construction induces posterior laws for treatment assignments, welfare, and policy-class comparisons (Ye, 2026). General Bayesian policy learning places a loss-based posterior directly on policy parameters and studies it through PAC-Bayes guarantees (Kato, 2026). CUBO instead combines Bayesian-bootstrap weights with fitted welfare contributions and optimizes posterior mean welfare jointly with lower-tail CVaR.

B Additional methodological details

B.1 Nuisance estimation before policy learning

All nuisance models are learned on the training data before optimizing the policy. The policy weights η are distinct from the nuisance weights θ: even when the two networks have the same architecture or initialization, updating η does not update θ or its fitted posterior. The following losses instantiate the nuisance-fitting stage.

Direct OPL. For IPW welfare, $\theta = w_e$ and the propensity model e_θ is fitted by binary cross-entropy:

$$\widehat{L}(\theta) = -\frac{1}{n}\sum_{i=1}^{n}\left[A_i \log e_\theta(X_i) + (1-A_i)\log\{1-e_\theta(X_i)\}\right]. \tag{20}$$

This is the average negative log-likelihood of $A_i \mid X_i \sim \mathrm{Bernoulli}\{e_\theta(X_i)\}$. Propensities used in welfare scores are clipped to $[\kappa, 1-\kappa]$ to bound fitted inverse-propensity weights.

Effect-based OPL. For CFRNet, $\theta = w_\tau$ collects the weights of a shared representation Φ_θ and two outcome heads $\mu_{a,\theta}$, with $\tau_\theta = \mu_{1,\theta} - \mu_{0,\theta}$. One continuous-outcome nuisance-fitting loss is (Shalit et al., 2017)

$$\begin{aligned}\widehat{L}(\theta) &= \frac{1}{n}\sum_{i=1}^{n}\omega_i\{Y_i - \mu_{A_i,\theta}(X_i)\}^2 + \xi\,\mathrm{IPM}_{\mathcal{G}}(\widehat{p}^{1}_{\Phi_\theta}, \widehat{p}^{0}_{\Phi_\theta}),\\ \omega_i &= \frac{A_i}{2\widehat{p}} + \frac{1-A_i}{2(1-\widehat{p})}, \qquad \widehat{p} = \frac{1}{n}\sum_{i=1}^{n}A_i.\end{aligned} \tag{21}$$

Here $\widehat{p}^{a}_{\Phi_\theta}$ is the empirical distribution of representations in treatment group a, and $\xi \geq 0$ controls balance. The IPM is the supremum over $g \in \mathcal{G}$ of the absolute difference between the two empirical averages of $g(\Phi_\theta)$. For a unit ball in a reproducing kernel Hilbert space, this gives MMD. Representation normalization can fix its scale. This weighted squared-error example specifies a possible nuisance loss; the experiments instead use the matched Gaussian likelihood and unnormalized CFR representation described in Appendix D.3. Weight regularization is included during posterior fitting.

B.2 Learning the parameter posterior and drawing nuisance models

Fitting the nuisance posterior. We fit q_n before policy learning, using the likelihood and Gaussian prior specified in (5). For a variational family $\mathcal{Q}$, the negative evidence lower bound gives the objective

$$q_n \in \arg\min_{q\in\mathcal{Q}}\left\{\mathbb{E}_{\theta\sim q}[-\log \hat{L}(\theta)] + \mathrm{KL}(q \,\|\, p_0)\right\}, \tag{22}$$

where p_0 is the prior density. For direct OPL, the Bernoulli propensity likelihood yields the cross-entropy loss in (20). For effect-based OPL, a Gaussian outcome likelihood yields squared-error fitting; its observation variance determines the loss scale. The weighted factual term in (21) corresponds to variances proportional to $1/\omega_i$. The IPM penalty in that equation can be used to pretrain a balanced representation, which is then fixed during posterior fitting. If the penalty is instead retained in (22), the result is a regularized variational target rather than an approximation to the likelihood-based posterior alone.

Following Jesson et al. (2020); Gal & Ghahramani (2016), we implement approximate variational fitting through dropout training with L_2 weight regularization. Stochastic forward passes estimate the expected likelihood loss, while the weight penalty approximates the prior contribution. The penalty scale accounts for whether the likelihood loss is summed or averaged and, for outcome models, the observation variance. Nuisance settings are fixed before policy learning. In the experiments, dropout probabilities are fixed beforehand, and hyperameters are selected from validation data, which results in a frozen posterior approximation on nuisance parameters.

Drawing nuisance parameters. Let M_ℓ denote the fitted weight matrix of layer ℓ, d_ℓ its input dimension, and $r_\ell \in (0, 1]$ its retention probability. Under the inverted-dropout convention used in training, draw

$$z_{\ell j,s} \sim \text{Bernoulli}(r_\ell), \qquad \Theta_{\ell,s} = M_\ell \,\text{diag}(z_{\ell 1,s}, \ldots, z_{\ell d_\ell,s})/r_\ell, \qquad \theta_s = (\Theta_{\ell,s})_\ell. \tag{23}$$

Masks are sampled independently across layers, coordinates, and posterior draws. Fitted biases and layers without dropout remain fixed. Together, the fitted matrices and mask distribution define q_n.

Each θ_s represents one nuisance function. Its masks are reused across observations and candidate policies, with a common mask for any representation shared by treatment heads. This preserves the dependence among predictions that enter the same welfare draw. Normalization statistics remain fixed, and predictions are conditional means without added observation noise, since the target is expected welfare. The fitted posterior and all sampled functions are retained throughout Algorithm 1.

Cross-fitted construction. Partition the observations into $K \geq 2$ folds I_k, with $n_k = |I_k|$ and $n_k/n \geq c > 0$. The fold assignment and fitting seeds are chosen independently of the observations and held fixed in the analysis. Fit q_{-k} using only $\mathcal{D}_{-k} = \{O_i : i \notin I_k\}$, including nuisance preprocessing, representation learning, and tuning. Thus q_{-k} is a probability kernel measurable with respect to $\mathcal{D}_{-k}$. Let q_n be a joint law for $\theta = (\theta_1, \ldots, \theta_K)$ with these marginals. Independently of B_s, draw $\theta_s \sim q_n$ and set

$$W(\pi; \mathcal{P}_s, \theta_s) = \sum_{k=1}^{K} \sum_{i \in I_k} B_{is} \phi_{\theta_{k,s}}(O_i; \pi), \qquad B_s \sim \text{DP}(1), \tag{24}$$

thus the estimated welfare can be regarded as the mean of score $\phi_{\theta_k}(O_i; \pi)$ from sampling distribution $\mathcal{P}_s = \sum_k \sum_{i \in I_k} B_{is} \delta_{(O_i,k)}$. For a fresh observation, $\mathcal{P}_0$ remains the law of O and $W(\pi; \mathcal{P}_0, \theta_k) = \mathbb{E}_0 \phi_{\theta_k}(O; \pi)$ holds θ_k fixed, and dependence among nuisance draws across folds is allowed. The fitted q_n is frozen during policy optimization. Independent training and evaluation samples give the same construction with $K = 1$, conditional on the training sample.

B.3 Bayesian-bootstrap weights and welfare contributions

Drawing population weights. For a finite base measure ν, the prior $\mathcal{P} \sim \text{DP}(\nu)$ gives the posterior $\text{DP}(\nu + nP_n)$, where $P_n = n^{-1} \sum_i \delta_{O_i}$. Letting the total mass of ν approach zero yields the Bayesian-bootstrap posterior (Ferguson, 1973; Rubin, 1981). Its weights can be sampled by generating independent $E_{is} \sim \text{Exp}(1)$ and setting $B_{is} = E_{is} / \sum_j E_{js}$. Thus $B_s \sim \text{DP}(1)$, with

$$\mathbb{E}(B_{is} \mid \mathcal{D}_n) = \frac{1}{n}, \qquad \text{Var}(B_{is} \mid \mathcal{D}_n) = \frac{n-1}{n^2(n+1)}.$$

For each scenario s, draw B_s independently of the nuisance draw θ_s from Appendix B.2, as required by (7).

Computing welfare draws. For welfare gain relative to no treatment, the direct and effect-based OPL contributions in Section 2 are

$$\phi_\theta(O_i; \pi_\eta) = \begin{cases} \pi_\eta(X_i) \left\{ \dfrac{A_i Y_i}{e_\theta(X_i)} - \dfrac{(1 - A_i) Y_i}{1 - e_\theta(X_i)} \right\}, & \text{direct OPL,} \\ \pi_\eta(X_i) \tau_\theta(X_i), & \text{effect-based OPL.} \end{cases} \tag{25}$$

An orthogonal alternative combines outcome functions $\mu = (\mu_0, \mu_1)$ with a cross-fitted propensity estimate $\widehat{e}$:

$$\Gamma_i(\mu) = \mu_1(X_i) - \mu_0(X_i) + \frac{A_i\{Y_i - \mu_1(X_i)\}}{\widehat{e}(X_i)} - \frac{(1 - A_i)\{Y_i - \mu_0(X_i)\}}{1 - \widehat{e}(X_i)}. \tag{26}$$

Algorithm 1 CUBO with alternating threshold and policy updates

Input: Data $\mathcal{D}_n$, policy π_η, levels $\alpha \in (0,1)$ and $\lambda \in [0,1]$, draws S, constants m_t, m_η, step sizes a_t, a_η.
Fit the nuisance models and q_n using Appendix B.1–B.2.
Draw (B_s, θ_s) independently as in (13), $s = 1, \ldots, S$.
Initialize η and set t to an empirical α-quantile of $\{W_s(\pi_\eta)\}_{s=1}^S$.
while not converged **do**
  Fix η and compute all $W_s(\pi_\eta)$ using (14); set $k = 1$.
  while $\lambda > 0$, condition (27) fails, and $k \le m_t$ **do**
    Compute G_t from (28); update $t \leftarrow t - a_t G_t$ and $k \leftarrow k + 1$.
  end while
  if $\lambda > 0$ and condition (27) still fails **then**
    Set t to an empirical α-quantile of the current welfare draws.
  end if
  Fix the updated threshold t.
  for $j = 1, \ldots, m_\eta$ **do**
    Compute G_η from (28); update $\eta \leftarrow \eta - a_\eta G_\eta$.
    Recompute all $W_s(\pi_\eta)$ with the retained (B_s, θ_s).
  end for
end while
Output: $\pi_{\widehat{\eta}}$ as the final policy.

Then $\phi_\theta(O_i; \pi_\eta) = \pi_\eta(X_i)\Gamma_i(\mu_\theta)$. If a posterior is also fitted for the propensity, its draw replaces $\widehat{e}$ in this score. Under cross-fitting, each observation is evaluated using nuisance functions fitted without its fold.

For any of these contributions, compute $W_s(\pi_\eta) = \sum_i B_{is}\phi_{\theta_s}(O_i; \pi_\eta)$ as in (14). Retaining the same (B_s, θ_s) across policy updates makes changes in W_s reflect changes in the policy under common posterior scenarios. Nuisance predictions can therefore be cached before optimization. As a special case, $q_n = \delta_{\widehat{\theta}}$ retains only population-weight uncertainty with its posterior mean be the empirical fitted welfare, so maximizing that mean recovers empirical welfare maximization. Section 4 covers both this point-mass specialization and the combined nuisance posterior. When each score is evaluated on observations excluded from its nuisance fit and scores are bounded, its regret bound controls the posterior-averaged welfare bias score, and for the orthogonal score, this bias is bounded by a product of nuisance errors.

B.4 CVaR LOSS AND ALTERNATING OPTIMIZATION

The threshold subproblem. We optimize the exact empirical loss in (15). With η fixed and $\lambda > 0$, this loss is convex and piecewise linear in t. Its minimizers are precisely the empirical α-quantiles, characterized by

$$\#\{s : W_s(\pi_\eta) < t\} \;\le\; \alpha S \;\le\; \#\{s : W_s(\pi_\eta) \le t\}. \tag{27}$$

Thus the threshold can be found either by subgradient descent or directly from the ordered welfare draws. This is the CVaR quantile characterization of Rockafellar & Uryasev (2000), applied here to the posterior distribution of estimated welfare.

Subgradients and alternating updates. For each draw, select a hinge subgradient

$$h_s \in \begin{cases} \{1\}, & W_s(\pi_\eta) < t, \\ [0,1], & W_s(\pi_\eta) = t, \\ \{0\}, & W_s(\pi_\eta) > t. \end{cases}$$

For differentiable policy models, the resulting threshold subgradient and policy generalized subgradient are

$$\begin{aligned} G_t &= \lambda \left[\frac{1}{\alpha S} \sum_{s=1}^{S} h_s - 1 \right], \\ G_\eta &= -\frac{1}{S} \sum_{s=1}^{S} \left(1 - \lambda + \frac{\lambda}{\alpha} h_s \right) \nabla_\eta W_s(\pi_\eta). \end{aligned} \tag{28}$$

Away from ties, these are ordinary gradients. The policy update gives additional weight to lower-welfare draws. At ties, any $h_s \in [0,1]$ is valid. When (27) holds and ties are present, distribute the remaining mass $\alpha S - \#\{s : W_s < t\}$ equally among them. If there are no ties, the quantile condition already gives $\#\{s : W_s < t\} = \alpha S$. In either case, $\sum_s h_s = \alpha S$ and $G_t = 0$, accounting for fractional mass at the tail boundary. A fixed automatic-differentiation convention at zero need not return a zero threshold subgradient at a minimizer, so we use (27) to check threshold optimality.

Algorithm 1 first fixes η and updates $t \leftarrow t - a_{t,k} G_t$, warm-starting from the previous threshold. Diminishing step sizes, such as $a_{t,k} = a_{t,1}/k$, avoid persistent oscillation around a kink. If the quantile condition is not reached within m_t steps, set t directly to an empirical α-quantile. With this threshold fixed, take m_η policy steps $\eta \leftarrow \eta - a_\eta G_\eta$, recomputing welfare and the hinge subgradients after each step. All nuisance parameters and posterior draws remain fixed, and no differentiation through the threshold solver is needed.

Posterior mini-batches and validation. The displayed algorithm uses the full fixed posterior draw bank. For less costly policy updates, sample a uniform mini-batch of scenario indices and replace the average over s in (15) and (28) by its mini-batch counterpart. Conditional on (η, t) and the chosen hinge subgradients, these are unbiased estimates of the full-bank loss and update direction. Each selected $W_s(\pi_\eta)$ still aggregates all observations using its original weights B_s. Replacing it by a noisy observation-level estimate inside $(t - W_s)_+$ generally introduces bias because the hinge is nonlinear.

Select (α, λ) and optimization hyperparameters using held-out validation data. For each setting, complete the prescribed policy-update budget and evaluate the final policy; validation does not select an intermediate policy checkpoint. Validation uses fresh population weights and nuisance draws from the frozen fitted posterior. This refreshes Monte Carlo simulation only; independence of validation observations from nuisance fitting is a separate requirement. Compute validation reward from the full welfare draws using (15) at an empirical α-quantile. Equivalently, average the lowest α fraction of the empirical distribution, including fractional boundary mass when αS is not an integer. For deterministic deployment, validate the thresholded rule $\mathbf{1}\{\pi_\eta(x) > 1/2\}$. Choose S large enough that αS adequately resolves the lower tail.

C Additional theoretical results

C.1 Regularity and finite-sample posterior control

Write $\|f\|_{L_\infty(\mathcal{P}_0)}$ for the $\mathcal{P}_0$-essential supremum of $|f|$. The score requirement in Assumption 2 means, for a deterministic $M > 0$, $\phi_{\theta_k}(O;\pi) = \pi(X)\phi_{\theta_k}(O;1)$, and

$$\max\left\{ \|\phi_{\theta_k}(\,\cdot\,;1)\|_{L_\infty(\mathcal{P}_0)}; \max_{i \in I_k} |\phi_{\theta_k}(O_i;1)| \right\} \le M, \tag{29}$$

for every k and q_{-k}-almost every θ_k, almost surely in the data. The treat-all function need not belong to Π. Bounded outcomes and predictions with clipped propensities imply this condition; Gaussian weight priors or dropout fitting alone do not. A VC-subgraph class satisfies (16) by the uniform entropy bound of van der Vaart & Wellner (2023); since the scores are linear in π, the covering arguments below apply verbatim to stochastic rules. Finite policy classes, or compact parameter spaces with continuous policy maps, ensure attainment.

In (17), $W(\pi; \mathcal{P}_0, \theta_k)$ integrates a fresh observation with θ_k fixed. The contraction requirement concerns the fitted nuisance law, not merely its posterior mean. A shrinking regret threshold also requires $C_n(M\sqrt{v_n/n} + r_n) \to 0$.

To state finite-sample bounds, write

$$\pi^*_\Pi(\mathcal{P}, \theta) \in \arg\max_{\pi \in \Pi} W(\pi; \mathcal{P}, \theta), \qquad \mathrm{Reg}(\mathcal{P}_0; \mathcal{P}, \theta) = W(\pi^*_\Pi; \mathcal{P}_0) - W\{\pi^*_\Pi(\mathcal{P}, \theta); \mathcal{P}_0\}. \tag{30}$$

Under Assumption 2 and the cross-fitted construction with fixed K and $n_k \ge cn$, there is a constant $C = C(A_\Pi, K, c) > 0$ such that, for each failure probability $\delta \in (0,1)$, with $\mathbb{P}_0$-probability at least $1 - \delta$,

$$\mathbb{E}[\mathrm{Reg}(\mathcal{P}_0; \mathcal{P}, \theta) \mid \mathcal{D}_n] \le 2\left[CM\sqrt{\frac{v + \log(1/\delta)}{n}} + b_n \right], \tag{31}$$

and, on the same event, simultaneously for all $u > 0$,

$$\mathbb{P}[\mathrm{Reg}(\mathcal{P}_0; \mathcal{P}, \theta) > u \,|\, \mathcal{D}_n] \leq \min\left\{1, \frac{2}{u}\left[CM\sqrt{\frac{v + \log(1/\delta)}{n}} + b_n\right]\right\}. \tag{32}$$

Note that these bounds do not require $b_n \to 0$ as they follow from Lemma 3 and the optimality argument in the proof of Theorem 1.

C.2 Regret guarantees for the CUBO decision

Theorem 1 concerns a rule optimal for each posterior draw. CUBO instead returns one policy by optimizing the mean–CVaR reward. The following corollaries control that decision, first for the ideal posterior reward and then for the finite draw bank used by Algorithm 1.

Corollary 1 (Regret of the CUBO reward maximizer)**.** *Under Assumption 2 and the cross-fitted construction of Appendix B.2, let $\widetilde{\pi} \in \Pi$ have posterior reward gap at most $\varepsilon_{\mathrm{opt}} \geq 0$: $\sup_{\pi\in\Pi} R_{\alpha,\lambda,n}(\pi) - R_{\alpha,\lambda,n}(\widetilde{\pi}) \leq \varepsilon_{\mathrm{opt}}$. For $0 < \alpha \leq 1$, $0 \leq \lambda \leq 1$, and $\delta \in (0,1)$, with $\mathbb{P}_0$-probability at least $1-\delta$,*

$$W(\pi^*_\Pi; \mathcal{P}_0) - W(\widetilde{\pi}; \mathcal{P}_0) \leq 2\left(1 - \lambda + \frac{\lambda}{\alpha}\right)\left[CM\sqrt{\frac{v + \log(1/\delta)}{n}} + b_n\right] + \varepsilon_{\mathrm{opt}}. \tag{33}$$

Here C depends only on A_Π, K, c, and the posterior reward uses the same fitted law as Theorem 1.

The conditional loss in Theorem 2 is explicitly

$$\mathcal{L}_{\alpha,\lambda,n}(\pi, t) = -(1-\lambda)\overline{W}_n(\pi) - \lambda t + \frac{\lambda}{\alpha}\mathbb{E}[\{t - W(\pi; \mathcal{P}, \theta)\}_+ \mid \mathcal{D}_n]. \tag{34}$$

Its empirical and posterior rewards satisfy $\widehat{R}_{\alpha,\lambda,n}(\pi) = -\inf_{t\in[-M,M]} \widehat{\mathcal{L}}_{\alpha,\lambda,n}(\pi, t)$ and $R_{\alpha,\lambda,n}(\pi) = -\inf_{t\in[-M,M]} \mathcal{L}_{\alpha,\lambda,n}(\pi, t)$. The threshold restriction is valid because posterior welfare lies in $[-M, M]$. Denote the bound in Theorem 2 by

$$\epsilon_S(\delta_S) = C_\Pi M\left(1 - \lambda + \frac{\lambda}{\alpha}\right)\sqrt{\frac{v + 1 + \log(1/\delta_S)}{S}}, \tag{35}$$

Taking infima over t in that theorem gives, on the same simulation event,

$$\sup_{\pi\in\Pi} |\widehat{R}_{\alpha,\lambda,n}(\pi) - R_{\alpha,\lambda,n}(\pi)| \leq \epsilon_S(\delta_S). \tag{36}$$

The uniform bound allows selection using the same bank, as required for selection-aware policy control (Chernozhukov et al., 2025). Conditional simulation concentration needs only score linearity, the observed-sample envelope, and policy entropy; it requires neither nuisance consistency nor cross-fitting. It does not establish frequentist coverage of posterior CVaR.

Corollary 2 (Estimated welfare loss and finite-bank regret)**.** *Under Assumption 2 and the cross-fitted construction of Appendix B.2 and the simulation conditions of Theorem 2, let $\widehat{\pi} \in \Pi$ be the returned policy, with empirical reward gap $\xi_{n,S} = \sup_{\pi\in\Pi} \widehat{R}_{\alpha,\lambda,n}(\pi) - \widehat{R}_{\alpha,\lambda,n}(\widehat{\pi})$. For $0 < \delta, \delta_S < 1$ with $\delta + \delta_S < 1$, define*

$$\Delta_{n,S} = \left(1 - \lambda + \frac{\lambda}{\alpha}\right)\left[CM\sqrt{\frac{v + \log(1/\delta)}{n}} + b_n\right] + \epsilon_S(\delta_S). \tag{37}$$

With joint probability at least $1 - \delta - \delta_S$ over the data and simulation, simultaneously for all $\pi \in \Pi$,

$$\left|\inf_{t\in[-M,M]} \widehat{\mathcal{L}}_{\alpha,\lambda,n}(\pi, t) + W(\pi; \mathcal{P}_0)\right| \leq \Delta_{n,S}, \tag{38}$$

and

$$W(\pi^*_\Pi; \mathcal{P}_0) - W(\widehat{\pi}; \mathcal{P}_0) \leq 2\Delta_{n,S} + \xi_{n,S}. \tag{39}$$

This result applies to the returned CUBO policy even when the alternating updates do not reach a global optimum: the actual optimization gap remains explicit. For a returned pair $(\widehat{\pi}, \widehat{t})$, define its joint empirical loss excess by

$$\xi_{\mathrm{pair}} = \widehat{\mathcal{L}}_{\alpha,\lambda,n}(\widehat{\pi}, \widehat{t}) - \inf_{\pi\in\Pi,\, t\in[-M,M]} \widehat{\mathcal{L}}_{\alpha,\lambda,n}(\pi, t).$$

Then $\xi_{n,S} \leq \xi_{\mathrm{pair}}$. On the same event, $W(\widehat{\pi}; \mathcal{P}_0) \geq \widehat{R}_{\alpha,\lambda,n}(\widehat{\pi}) - \Delta_{n,S}$; a numerical guarantee requires justified bounds for the constants and b_n.

C.3 SUFFICIENT NUISANCE-POSTERIOR ACCURACY CONDITIONS

Let $\mathcal{P}_{0,X}$ be the marginal law of X and write $\|f\|_2 = (\int f^2 \, d\mathcal{P}_{0,X})^{1/2}$. Candidate nuisance functions are held fixed when integrating a fresh observation. For a random sequence Z_n and deterministic $a_n > 0$, write $Z_n = O_{\mathbb{P}_0}(a_n)$ when Z_n/a_n is bounded in repeated-sample probability.

Proposition 1 (Score-specific nuisance bias)**.** *Suppose Assumption 1 holds, $\mathbb{E}_0|Y| < \infty$, $\kappa \le e, e_0 \le 1-\kappa$, and the displayed L_2 errors are finite. Let $\Gamma(e,\mu)$ denote the score* (26) *under propensity e and outcome regressions $\mu = (\mu_0, \mu_1)$, and put $\delta_a = \mu_a - \mu_{a,0}$. Then*

$$\mathbb{E}_0\{\Gamma(e,\mu) - \tau_0(X) \mid X\} = (e - e_0)\left\{\frac{\delta_1}{e} + \frac{\delta_0}{1-e}\right\}, \tag{40}$$

so

$$\sup_{\pi\in\Pi} |W(\pi;\mathcal{P}_0,\theta) - W(\pi;\mathcal{P}_0)| \le \kappa^{-1}\|e-e_0\|_2(\|\mu_1-\mu_{1,0}\|_2 + \|\mu_0-\mu_{0,0}\|_2). \tag{41}$$

For the IPW score in (25)*, if $|\mu_{a,0}| \le H$ for a constant $H < \infty$, the same population welfare discrepancy is bounded by $(2H/\kappa)\|e-e_0\|_2$. For the effect-based score with candidate effect function τ_θ, it is bounded by $\|\tau_\theta - \tau_0\|_2$. Integrating these bounds under q_{-k} and weighting by n_k/n gives the corresponding bound on b_n.*

Writing e_{θ_k}, μ_{a,θ_k}, and τ_{θ_k} for the nuisance functions specified by a foldwise draw θ_k, define their posterior root-mean-square errors

$$r_{e,n}^2 = \sum_k \frac{n_k}{n} \int \|e_{\theta_k} - e_0\|_2^2 \, q_{-k}(d\theta_k),$$

$$r_{\mu,n}^2 = \sum_k \frac{n_k}{n} \int (\|\mu_{1,\theta_k} - \mu_{1,0}\|_2 + \|\mu_{0,\theta_k} - \mu_{0,0}\|_2)^2 \, q_{-k}(d\theta_k),$$

$$r_{\tau,n}^2 = \sum_k \frac{n_k}{n} \int \|\tau_{\theta_k} - \tau_0\|_2^2 \, q_{-k}(d\theta_k).$$

Cauchy–Schwarz gives

$$b_n \le \begin{cases} \kappa^{-1} r_{e,n} r_{\mu,n}, & \text{orthogonal score,} \\ (2H/\kappa) r_{e,n}, & \text{IPW score,} \\ r_{\tau,n}, & \text{effect-based score.} \end{cases} \tag{42}$$

Thus deterministic rates $\epsilon_{e,n}, \epsilon_{\mu,n}, \epsilon_{\tau,n}$ with $r_{e,n} = O_{\mathbb{P}_0}(\epsilon_{e,n})$, $r_{\mu,n} = O_{\mathbb{P}_0}(\epsilon_{\mu,n})$, and $r_{\tau,n} = O_{\mathbb{P}_0}(\epsilon_{\tau,n})$ yield the rate r_n in Theorem 1 as, respectively, $\epsilon_{e,n}\epsilon_{\mu,n}$, $\epsilon_{e,n}$, or $\epsilon_{\tau,n}$. Independence between outcome and propensity draws is not needed. Note that these are rates for the fitted nuisance distribution, not merely its mean or a point estimate. With known propensity, the orthogonal bias is zero for every bounded outcome draw, even without outcome-posterior contraction.

If $|Y|, |\mu_{a,\theta}| \le H$ and fitted propensities lie in $[\kappa, 1-\kappa]$, score bounds are $M = 2H(1+\kappa^{-1})$ for the orthogonal score, $M = 2H/\kappa$ for IPW, and $M = 2H$ for an outcome-contrast score. Bounded predictions alone do not bound residual scores when Y is unbounded.

C.4 NUISANCES FITTED ON THE FULL EVALUATION SAMPLE

Proposition 2 (Full-sample nuisance fitting)**.** *Suppose Assumption 2 holds with q_n in place of q_{-k} and the full sample in place of I_k. Instead of cross-fitting, let q_n be any data-dependent law supported on a deterministic nuisance set Θ, with the same function θ evaluated at all observations. Assume the pointwise measurable class*

$$\mathcal{F} = \{\phi_\theta(\,\cdot\,;\pi) : \pi \in \Pi, \theta \in \Theta\}$$

has envelope M and satisfies the following bound for entropy constants $A_F \ge e$ and $w \ge 1$, with Q ranging over probability laws on observations:

$$\sup_Q \log N(M\epsilon, \mathcal{F}, L_2(Q)) \le w \log(A_F/\epsilon), \qquad 0 < \epsilon \le 1,$$

Retain the product posterior and iid scenario simulation. In this setting define

$$b_n = \int \sup_{\pi\in\Pi} |W(\pi;\mathcal{P}_0,\theta) - W(\pi;\mathcal{P}_0)| \, q_n(d\theta).$$

The finite-sample bounds (31)–(32) *and Corollaries 1–2 hold with the sampling term* $CM\sqrt{\{v+\log(1/\delta)\}/n}$ *replaced by* $CM\sqrt{\{v+w+\log(1/\delta)\}/n}$*, where* C *depends only on* A_Π, A_F*. If* $(v+w)/n \to 0$ *and* $b_n = O_{\mathbb{P}_0}(r_n)$ *for deterministic positive* $r_n \to 0$*, the posterior contraction statement holds at rate* $M\sqrt{(v+w)/n} + r_n$*.*

This alternative covers the full-sample fitting permitted by the method, but requires control of the joint policy–nuisance class. Freezing q_n does not remove its dependence on the evaluation observations. The conditional simulation bound still involves only policy complexity, because nuisance parameters are not optimized in the policy-learning stage.

C.5 Consistency, tuning, and validation

Let S_n be the number of posterior draws at sample size n. Write O_P for boundedness in probability and $o_P(1)$ for convergence to zero under the joint law $\mathbb{P}$ of the data, the fitted nuisance law, and the simulation draws, defined in Appendix E. Corollary 2 gives, for balanced fixed folds and possibly growing $\Pi = \Pi_n$,

$$\sup_{\pi\in\Pi_n} W(\pi;\mathcal{P}_0) - W(\widehat{\pi};\mathcal{P}_0) \le O_P\left[\left(1-\lambda+\frac{\lambda}{\alpha}\right)\left\{M\sqrt{\frac{v_n}{n}} + r_n + M\sqrt{\frac{v_n+1}{S_n}}\right\}\right] + \xi_{n,S_n}. \tag{43}$$

Here $b_n = O_{\mathbb{P}_0}(r_n)$. The bound tends to zero if the displayed rate tends to zero and $\xi_{n,S_n} = o_P(1)$. Comparison with a larger policy class $\Pi^{\mathrm{target}} \supseteq \Pi_n$ requires adding its approximation error $\sup_{\pi\in\Pi^{\mathrm{target}}} W(\pi;\mathcal{P}_0) - \sup_{\pi\in\Pi_n} W(\pi;\mathcal{P}_0)$.

For (α,λ) selected from a prespecified grid of J pairs, replace δ_S by δ_S/J in Theorem 2 and take a union bound. The underlying sampling event controls welfare uniformly and is common to all pairs. For L draw banks with the same target, a further union bound uses $\delta_S/(JL)$; each bank must have iid draws internally, although independence across banks is not needed.

When an independent validation sample selects among final policies from prespecified training settings, condition on the training data and randomness. The candidate policies and nuisance law are then fixed, and the $K = 1$ theory applies with validation sample size and a common evaluation score and criterion. An empirical validation maximizer has zero gap within that candidate set. Comparison with the full policy class additionally requires its candidate approximation gap. A fresh draw bank on the same observations only refreshes simulation; it does not make fitted nuisances independent of those observations.

C.6 Deterministic deployment

Proposition 3 (Smooth-to-hard deployment)**.** *For a measurable score* $h : \mathcal{X} \to \mathbb{R}$*, temperature* $T > 0$*, and logistic function* $\sigma(u) = (1+e^{-u})^{-1}$*, define* $\pi_T(x) = \sigma\{h(x)/T\}$ *and* $\overline{\pi}(x) = \mathbf{1}\{h(x) > 0\}$*. Then*

$$|W(\overline{\pi};\mathcal{P}_0) - W(\pi_T;\mathcal{P}_0)| \le \mathbb{E}_0[|\tau_0(X)|e^{-|h(X)|/T}]. \tag{44}$$

Adding this term to a regret bound for π_T *bounds the deployed rule's regret against the same population comparator.*

If the deterministic rule is evaluated and selected directly, the theory instead applies to its deterministic class.

D Experimental details and additional results

This appendix further experimental details, including the dataset construction and evaluation protocol (Appendix D.1), the baseline implementations (Appendix D.2), the Bayesian nuisance models and policy settings (Appendix D.3), and the ablation protocol (Appendix D.4).

D.1 Datasets, splits, and evaluation

Dataset details. The main text introduces IHDP (Hill, 2011; Dehejia, 2005) and ACIC 2016 (Dorie et al., 2019) and their splits following Jesson et al. (2020). We record the remaining construction details here.

IHDP. IHDP originates from the Infant Health and Development Program, a randomized intervention for low-birth-weight infants. Following the standard protocol (Hill, 2011; Shalit et al., 2017), the treated group is subsampled to create covariate imbalance between the arms, and the outcomes are partially simulated so that both potential outcomes of every unit are known. Following Jesson et al. (2020), we use realization 10 of IHDP (Hill, 2011; Dehejia, 2005), with 470 training, 202 validation, and 75 test units on 25 covariates.

ACIC. ACIC 2016 reuses the IHDP covariates within more challenging outcome-generating processes marked by stronger nonlinearity and treatment-effect heterogeneity (Dorie et al., 2019). We use condition 1 of the 2016 release, with the first lexicographic outcome file of trial 0, giving 2,352 training, 1,009 validation, and 1,441 test units on 82 one-hot-encoded covariates. All methods share this setting with the same sample split.

Preprocessing and validation. Nuisance-network gradient updates use training observations. Outcome standardization and the arm counts used by the Gaussian-model prior penalties are computed on the combined training and validation sample; nuisance checkpoints are selected by observed validation loss and then frozen, without refitting. Policy covariate scaling is fitted on training observations only. No test outcomes or test-derived normalization statistics enter these fits. Thus validation participates in nuisance selection and preprocessing: a new Monte Carlo draw bank on that split does not create an independent validation dataset.

Metrics. The main text defines the normalized regret contributions $r_i = \{\max(\tau_i, 0) - \pi_i\tau_i\}/D_m$, the regret $\mathrm{Reg} = m^{-1}\sum_i r_i$, and the policy accuracy PA, with $\tau_i = \mu_{1,0}(X_i) - \mu_{0,0}(X_i)$, $\pi_i = \pi(X_i)$, and $D_m = m^{-1}\sum_i |\tau_i| > 0$ on both benchmarks. These formulas integrate treatment randomization, so PA agrees with action accuracy for a deterministic rule and is expected accuracy for a stochastic policy. Tables display $10\,\mathrm{Reg}$ and $100\,\mathrm{PA}$ with one decimal place. Diagnostic PEHE is $\{m^{-1}\sum_i(\widehat{\tau}(X_i) - \tau_i)^2\}^{1/2}$.

D.2 Baseline implementations

The plug-in baselines are T-Learner (Künzel et al., 2019), CFRNet, TARNet (Shalit et al., 2017), Dragonnet (Shi et al., 2019), CEVAE (Louizos et al., 2017), DeRCFR (Wu et al., 2023), and ESCFR (Wang et al., 2023), each deploying $\mathbf{1}\{\widehat{\tau}(x) > 0\}$. The first four follow the architecture and Gaussian outcome heads in Jesson's released `ucate` implementation, with posterior dropout disabled for the point estimators. CFRNet and BCFRNet use raw representations, unweighted factual likelihood, and the same active MMD coefficient 1. Dragonnet uses the original paper's learned targeted fluctuation. The CEVAE plug-in retains the original Louizos encoder and latent objective, with 32 latent dimensions and 32 antithetic integration nodes. DeRCFR and ESCFR retain their distinct paper-based balancing objectives.

PT-CATE follows the setting in Frauen et al. (2025), with a three-hidden-layer tanh CATE network and ReLU positive-slope network, successive CATE initialization, slope fitting with CATE frozen, and combined learning with the slope frozen. It uses cross-fitted DR pseudo-outcomes from separate tanh outcome and propensity networks. Interior $\gamma \in \{.10, .25, .50, .75\}$ and combined-stage checkpoints are selected by observed validation welfare, and deployment uses the learned CATE sign.

We also include DR-EWM (Athey & Wager, 2021) and PoLeCe (Chernozhukov et al., 2025). DR-EWM learns a stochastic policy in the same signed-multiplier-plus-offset class as CUBO, using cross-fitted DR scores. PoLeCe retains the original simultaneous lower-confidence-bound objective and Appendix H's enumeration of treatment assignments across covariate groups. The groups are application-specific. On each dataset, we fit six partitions using training cross-fitted DR scores. We use regression trees and score-sign classification trees at depths two and three, with classification observations weighted by the absolute DR score, and four- and eight-quantile partitions of a smoothed

DR regression score, adding zero as a cutpoint when it lies inside the training score range. Tree leaves contain at least $\max\{20, \lceil .05n \rceil\}$ training observations. The smoother uses 200 histogram gradient-boosting iterations, at most seven leaves, learning rate .05, and L_2 regularization .1, with cutpoints using training predictions.

All binary assignments on every partition, including treat-none and treat-all, form one finite policy class. Held-out DR scores estimate welfare relative to no treatment and its standard error. A single 95% critical value from 2,000 correlated Gaussian multiplier draws is computed by exact maximization over the entire class, accounting for selection of both the partition and its assignment. No competing learner supplies candidate policies. The class contains 704 assignment entries on IHDP and 608 on ACIC, including duplicates that leave the bootstrap maximum unchanged. The selected IHDP rule uses the four-quantile-plus-zero partition, and ACIC selects the depth-two weighted classification partition. The confidence level, welfare scores, and data splits are unchanged.

D.3 Bayesian nuisance models and policy settings

Matched Gaussian models. BCFRNet, BTARNet, BT-Learner, and BDragonnet use the dropout construction of Jesson et al. (2020). Shared models have three width-200 ELU layers and two width-100 layers per outcome branch, while T-Learner fits a separate five-layer width-200 ELU network for each arm. Each outcome head predicts a Gaussian mean and a positive scale. Fitting uses unweighted factual Gaussian negative log-likelihood, the corresponding Gaussian-prior weight penalty, and each backbone's additional loss. CFRNet and BCFRNet share the unnormalized representation and squared linear MMD coefficient 1. TARNet omits the balancing term by definition. Dragonnet includes propensity cross-entropy and a learned targeted fluctuation with its signed inverse-propensity covariate. The point models retain the same architectures and heads with posterior dropout disabled. All four Bayesian backbones share one dropout schedule per benchmark, with a single exception noted below: IHDP follows the released `ucate` command settings, dropout .5 in hidden layers and .5 before the outcome output, whereas ACIC follows the schedule stated in the paper's Appendix B, dropout .1 in hidden layers and .5 before output. For BDragonnet on ACIC, the released $.1/.5$ schedule leaves the posterior too diffuse for policy optimization, so we instead calibrate the dropout pair on the observed validation split: we prespecify a low-dropout ladder $\{(.1, .1), (.05, .05), (.02, .02)\}$ for (hidden, output) dropout, fit the CUBO method for each candidate, and select the candidate whose final policy attains the highest observed-validation CUBO reward (validation is used for hyperparameter selection only, and test outcomes never enter fitting or selection). The selected pair is $(.1, .1)$.

All these models use Adam with step size .001, batch size 100, at most 2,000 epochs, and validation patience 50. T-Learner fits and stops the two arms separately. Shared models select checkpoints by their validation fitting loss. With an averaged likelihood, shared-layer prior penalties are divided by the non-test sample count, and treatment-head penalties by the corresponding non-test arm counts. The shared weight contribution also includes the dropout retention factor, following the released implementation. Point and Bayesian versions therefore share architecture and likelihood, while their dropout and prior contributions differ as specified by their fitting objectives.

Bayesian DeRCFR. BDeRCFR is our extension of DeRCFR using dropout in its three normalized representations and prediction networks. It retains the two-hidden-layer width-100 architecture, squared-error outcome fitting, and balancing and disentanglement terms, with discrepancy coefficient 1. Hidden/output dropout is $.1/.5$, and an explicit penalty $\|\theta\|_2^2/(2n_{\text{train}})$ replaces optimizer weight decay for the sampled network parameters. Learned sample weights and mutual-information critics remain fixed during posterior sampling. Because balancing and disentanglement penalties remain in fitting, this is a regularized dropout approximation conditional on those fitted quantities, not an exact likelihood posterior. Both DeRCFR versions use the same optimizer, update budget, and factual-MSE validation stopping rule; no separate BDeRCFR hyperparameter search is conducted.

Policy class and initialization. For posterior-mean CATE $s(x) = S^{-1} \sum_s \tau_{\theta_s}(x)$, the deployed policy is

$$\pi_\eta(x) = \sigma\{(1 + g_\eta(x))s(x)/T + b_\eta(x)\}\,.$$

A shared five-hidden-layer width-64 ELU network has two outputs, g_η and b_η. Zero initialization of its final layer gives initial multiplier one and offset zero; the multiplier can subsequently have either

sign. The initial probability $\sigma\{s(x)/T\}$ approximates, but does not equal, the hard Bayesian plug-in $\mathbf{1}\{s(x) > 0\}$. Treatment probabilities are deployed directly. We use $T = 1$ and policy Adam step size .001 for IHDP; ACIC uses the previously validation-selected $T = .05$ and step size .0001. The class and update procedure are otherwise identical.

Posterior draws and optimization. For each backbone, 100 whole-function dropout draws share their masks across observations and candidate policies, and are paired with independent Dirichlet population-weight draws. The effect-based training contribution is $\pi_\eta(X_i)\tau_{\theta_s}(X_i)$. The nuisance model is fitted on these training observations, so this experimental realization does not use cross-fitted orthogonal training scores. The full-sample alternative in Proposition 2 states the additional complexity requirements; freezing a fitted neural model alone does not establish them.

We fix $(\alpha, \lambda) = (.1, .5)$ across all main CUBO backbones. Algorithm 1 uses the full 100-draw bank for 80 cycles, with five Adam policy steps at the fixed threshold per cycle. The threshold uses at most 20 subgradient steps with initial step size .05 and decreasing step size, followed by the exact empirical quantile if needed. A common positive factor, the mean across draws of the sum of absolute weighted training effects floored at .1, scales the entire loss for numerical optimization and does not change its maximizers. Each run returns its final policy after 400 updates.

Validation uses fresh function draws and independent bootstrap weights with observed-outcome orthogonal scores and fixed nuisance functions. These draws are independent Monte Carlo simulations conditional on the fitted nuisance law.

D.4 IHDP VALIDATION ABLATION PROTOCOL

The ablation study is construct on CUBO with BDeRCFR backbone. Each sweep changes one coordinate around $(\alpha, \lambda, L) = (.1, .5, 5)$, where L counts the shared width-64 ELU hidden layers feeding both policy outputs. All settings use the same fitted nuisance posterior, paired draw bank, learning rate .001, and 400-update budget. Across seeds, hidden weights differ but the zero-initialized output layer gives the same initial policy. Validation counterfactual means are read after fitting to evaluate regret, and they do not enter the loss.

For BDeRCFR, the lowest mean validation regret on the displayed scale occurs at $\lambda = .75$ (0.00043), compared with 0.04139 at $\lambda = .05$. The minimum on the tail-probability grid is at $\alpha = .30$ (0.00066), compared with 0.00647 at .05 and 0.04362 at .90. The results suggest that all components are necessary for uncertainty-aware policy learning.

E PROOFS

Conditional expectations below integrate the fitted nuisance law and independent Bayesian-bootstrap weights. Write $\mathbb{P}$ for the joint law of the observed data, the fitted nuisance law, and the simulation draws; $\mathbb{E}$ and unsubscripted probability statements are taken under this law unless stated otherwise. Repeated-sample probabilities integrate only the observed data under $\mathbb{P}_0$. Write $Pf = \int f\, dP$, $P_m = m^{-1}\sum_{i=1}^m \delta_{O_i}$, and $(\pi g)(O) = \pi(X)g(O)$. Unsubscripted $C > 0$ denotes a universal constant unless other dependence is stated; C_Π depends only on A_Π. Constants may increase between displays with their stated dependence. The notation $O_{\mathbb{P}_0}(1)$ and $o_{\mathbb{P}_0}(1)$ denote boundedness and convergence to zero in $\mathbb{P}_0$-probability.

E.1 EMPIRICAL-PROCESS AND BOOTSTRAP BOUNDS

We record the weighted empirical-process bounds needed for the combined posterior. The entropy-integral and symmetrization tools used here are standard; see van der Vaart & Wellner (2023), Sections 2.2, 2.3, and 2.6. The argument explicitly treats the normalization of bootstrap weights and the dependence induced by cross-fitting.

Lemma 1 (Weighted policy complexity)**.** *Under* (16)*, for arbitrary $x_1, \dots, x_m$ and real $c_1, \dots, c_m$, independent Rademacher signs ε_i satisfy*

$$\mathbb{E}_\varepsilon \sup_{\pi\in\Pi} \left| \sum_{i=1}^m \varepsilon_i c_i \pi(x_i) \right| \le C_\Pi \sqrt{v} \left(\sum_i c_i^2 \right)^{1/2}. \tag{45}$$

In particular, for a fixed measurable g with $|g| \le M$ and iid $O_1, \dots, O_m \sim \mathcal{P}_0$,

$$\mathbb{E}_0 \sup_{\pi \in \Pi} |(P_m - \mathcal{P}_0)(\pi g)| \le C_\Pi M \sqrt{v/m}. \tag{46}$$

Proof. If $R = (\sum_i c_i^2)^{1/2} = 0$, the result is immediate. Otherwise let $Q_c = R^{-2} \sum_i c_i^2 \delta_{x_i}$. The process $X_\pi = \sum_i \varepsilon_i c_i \pi(x_i)$ has sub-Gaussian increments with metric $d(\pi, \pi') = R\|\pi - \pi'\|_{L_2(Q_c)}$: indeed, $\cosh u \le e^{u^2/2}$ bounds the moment generating function of every increment by $\exp\{s^2 d(\pi, \pi')^2/2\}$. Add the zero function as an anchor and a reflected copy of the class to control the absolute supremum. This changes covering numbers by at most a fixed factor. The sub-Gaussian entropy integral and (16) bound the expected supremum by

$$CR \int_0^1 \sqrt{1 + v \log(A_\Pi/u)}\, du \le C_\Pi R \sqrt{v}.$$

For completeness, use successive nets of the anchored, reflected class at radii $2^{-j}R$. The expected maximum absolute increment at level j is bounded by $C 2^{-j} R \sqrt{1 + \log N(2^{-j}, \Pi, L_2(Q_c))}$, after changing a universal constant. Summing the telescoping increments gives the displayed integral; the added 1 also covers singleton classes. Pointwise measurability permits passage from finite subcollections to the full supremum. Finally, ordinary iid symmetrization gives

$$\mathbb{E}_0 \sup_\pi |(P_m - \mathcal{P}_0)(\pi g)| \le \frac{2}{m} \mathbb{E}_{0,\varepsilon} \sup_\pi \left| \sum_i \varepsilon_i g(O_i) \pi(X_i) \right| \le C_\Pi M \sqrt{v/m}.$$

□

Lemma 2 (Conditional Bayesian-bootstrap fluctuation)**.** *Under* (16)*, for fixed observations and any fixed score vector $g_1, \dots, g_n$ with $|g_i| \le M$,*

$$\mathbb{E}_B \sup_{\pi \in \Pi} \left| \sum_i (B_i - 1/n) \pi(X_i) g_i \right| \le C_\Pi M \sqrt{v/n}, \qquad B \sim \mathrm{DP}(1). \tag{47}$$

The bound is uniform in the observations and score vector.

Proof. Let $E_i \overset{\text{iid}}{\sim} \mathrm{Exp}(1)$, $T = \sum_i E_i$, and $B_i = E_i/T$. The gamma–Dirichlet factorization gives $T \perp\!\!\!\perp B$, $\mathbb{E}T = n$, and $\mathrm{Var}(T) = n$. With $f_i(\pi) = \pi(X_i) g_i$, the identity $\sum_i (E_i - T/n) f_i(\pi) = T \sum_i (B_i - 1/n) f_i(\pi)$ and the independence of T and B give

$$\begin{aligned}
&\mathbb{E}_E \sup_\pi \left| \sum_i (E_i - T/n) f_i(\pi) \right| \\
&\quad = \mathbb{E}[T]\, \mathbb{E}_B \sup_\pi \left| \sum_i (B_i - 1/n) f_i(\pi) \right| \\
&\quad = n\, \mathbb{E}_B \sup_\pi \left| \sum_i (B_i - 1/n) f_i(\pi) \right|.
\end{aligned}$$

Moreover,

$$\mathbb{E}_E \sup_\pi \left| \sum_i (E_i - T/n) f_i(\pi) \right| \le \mathbb{E}_E \sup_\pi \left| \sum_i (E_i - 1) f_i(\pi) \right| + M \mathbb{E}|T - n|.$$

The last expectation is at most $\sqrt{n}$. Let E_i' be independent copies. Jensen's inequality and the coordinatewise symmetry of $E_i - E_i'$ give

$$\begin{aligned}
\mathbb{E}_E \sup_\pi \left| \sum_i (E_i - 1) f_i(\pi) \right| &\le \mathbb{E}_{E,E'} \sup_\pi \left| \sum_i (E_i - E_i') f_i(\pi) \right| \\
&= \mathbb{E}_{E,E',\varepsilon} \sup_\pi \left| \sum_i \varepsilon_i (E_i - E_i') g_i \pi(X_i) \right|
\end{aligned}$$

$$\leq C_\Pi M\sqrt{v}\,\mathbb{E}\left\{\sum_i (E_i - E_i')^2\right\}^{1/2} \leq C_\Pi M\sqrt{2nv}.$$

The penultimate inequality is Lemma 1; the last uses $\mathbb{E}(E_i - E_i')^2 = 2$. Division by n proves the claim. In particular, no unverified bound on the random denominator T is needed. □

Lemma 3 (Uniform posterior welfare control)**.** *Under Assumption 2 and the cross-fitted construction* (24) *with fixed K and $n_k \geq cn$, there is a constant $C = C(A_\Pi, K, c)$ such that, for every $\delta \in (0,1)$, with $\mathbb{P}_0$-probability at least $1-\delta$,*

$$\mathbb{E}\left[\sup_{\pi\in\Pi} |W(\pi;\mathcal{P},\theta) - W(\pi;\mathcal{P}_0)| \,\middle|\, \mathcal{D}_n\right] \leq CM\sqrt{\frac{v+\log(1/\delta)}{n}} + b_n. \tag{48}$$

This bound does not require the nuisance bias to vanish.

Proof. Let P_{n_k} be the empirical measure on fold I_k. The cross-fitted empirical welfare is $W(\pi;\mathcal{P}_n,\theta) = \sum_k (n_k/n) P_{n_k}\phi_{\theta_k}(\,\cdot\,;\pi)$. For each draw, adding and subtracting this empirical welfare and its foldwise population counterpart gives

$$\begin{aligned}
&\sup_\pi |W(\pi;\mathcal{P},\theta) - W(\pi;\mathcal{P}_0)| \\
&\quad\leq \sup_\pi |W(\pi;\mathcal{P},\theta) - W(\pi;\mathcal{P}_n,\theta)| \\
&\qquad + \sum_k \frac{n_k}{n} \sup_\pi |(P_{n_k} - \mathcal{P}_0)\phi_{\theta_k}(\,\cdot\,;\pi)| \\
&\qquad + \sum_k \frac{n_k}{n} \sup_\pi |W(\pi;\mathcal{P}_0,\theta_k) - W(\pi;\mathcal{P}_0)|.
\end{aligned} \tag{49}$$

The first term is controlled by Lemma 2: conditional on the data and θ, its score vector is fixed and bounded. Independence of the bootstrap weights and θ therefore bounds its integrated expectation by $C_\Pi M\sqrt{v/n}$. The expectation of the last term is exactly b_n.

For the middle term put

$$A_k = \int \sup_\pi |(P_{n_k} - \mathcal{P}_0)\phi_{\theta_k}(\,\cdot\,;\pi)|\, q_{-k}(d\theta_k).$$

Condition separately on $\mathcal{D}_{-k}$, with folds and fitting seeds fixed as specified in the method. The probability kernel q_{-k} is then fixed and the observations in I_k remain iid from $\mathcal{P}_0$. Lemma 1, applied with the fixed function $g(O) = \phi_{\theta_k}(O;1)$, and Fubini's theorem give $\mathbb{E}_0(A_k \mid \mathcal{D}_{-k}) \leq C_\Pi M\sqrt{v/n_k}$. Changing one held-out observation changes A_k by at most $2M/n_k$; integration, absolute values, and suprema preserve this bounded-difference constant. Consequently, McDiarmid's inequality gives

$$\mathbb{P}_0\left[A_k > C_\Pi M\sqrt{v/n_k} + M\sqrt{2\log(K/\delta)/n_k} \,\middle|\, \mathcal{D}_{-k}\right] \leq \delta/K.$$

Integrate out each training sample and union-bound over folds. This does not condition simultaneously on all fitted nuisances and does not assert that cross-fitted scores are iid. On the resulting event, integration of (49) yields

$$C_\Pi M\sqrt{v/n} + \sum_k \frac{n_k}{n}\left\{C_\Pi M\sqrt{v/n_k} + M\sqrt{2\log(K/\delta)/n_k}\right\} + b_n.$$

Use $n_k \geq cn$, fixed K, and $v \geq 1$ to absorb K, c and $\log K$ into C. This proves (48). For a separate training sample, the same reasoning uses $K = 1$ conditional on that sample. □

E.2 Posterior regret and the CUBO decision

Proof of Theorem 1. For this proof write $\rho(\mathcal{P},\theta) = \sup_{\pi\in\Pi} |W(\pi;\mathcal{P},\theta) - W(\pi;\mathcal{P}_0)|$. The optimality of $\pi^*_\Pi(\mathcal{P},\theta)$ gives

$$\mathrm{Reg}(\mathcal{P}_0;\mathcal{P},\theta) = W(\pi^*_\Pi;\mathcal{P}_0) - W(\pi^*_\Pi;\mathcal{P},\theta)$$

$$
\begin{aligned}
&\quad + W(\pi^*_\Pi; \mathcal{P}, \theta) - W\{\pi^*_\Pi(\mathcal{P}, \theta); \mathcal{P}, \theta\} \\
&\quad + W\{\pi^*_\Pi(\mathcal{P}, \theta); \mathcal{P}, \theta\} - W\{\pi^*_\Pi(\mathcal{P}, \theta); \mathcal{P}_0\} \\
&\le 2\rho(\mathcal{P}, \theta).
\end{aligned}
$$

The middle difference is nonpositive, and the two outer differences are each at most ρ. This argument holds for every measurable choice among nonunique maximizers. Conditional expectation and Markov's inequality therefore imply, for all $u > 0$,

$$
\begin{aligned}
\mathbb{E}[\mathrm{Reg}(\mathcal{P}_0; \mathcal{P}, \theta) \mid \mathcal{D}_n] &\le 2\mathbb{E}[\rho(\mathcal{P}, \theta) \mid \mathcal{D}_n], \\
\mathbb{P}[\mathrm{Reg}(\mathcal{P}_0; \mathcal{P}, \theta) > u \mid \mathcal{D}_n] &\le \min\{1, 2\mathbb{E}[\rho(\mathcal{P}, \theta) \mid \mathcal{D}_n]/u\}.
\end{aligned} \tag{50}
$$

Lemma 3 proves (31) and (32) on the same data event, simultaneously over $u > 0$.

For the contraction assertion, the same lemma, for arbitrarily small fixed failure probability, gives $\mathbb{E}[\rho(\mathcal{P}, \theta) \mid \mathcal{D}_n] = O_{\mathbb{P}_0}(M\sqrt{v/n} + r_n)$. In (50), take $u = C_n\{M\sqrt{v/n} + r_n\}$, which is positive. Its right-hand side is $O_{\mathbb{P}_0}(1/C_n) = o_{\mathbb{P}_0}(1)$, proving (18). The conditions $v/n \to 0$ and $r_n \to 0$ make the rate vanish. If exact posterior maximization is replaced by a measurable rule whose posterior welfare gap is at most ε_n, the same argument gives regret at most $2\rho + \varepsilon_n$; its posterior expected regret gains at most $\mathbb{E}(\varepsilon_n \mid \mathcal{D}_n)$. Thus approximation with $\mathbb{E}(\varepsilon_n \mid \mathcal{D}_n) = O_{\mathbb{P}_0}(M\sqrt{v/n} + r_n)$ preserves the contraction rate. An unattained population supremum can likewise be approached by maximizing sequences. □

Proof of Corollary 1. For an integrable random variable U, write $C_\alpha(U) = \sup_{t\in\mathbb{R}}\{t - \alpha^{-1}\mathbb{E}(t - U)_+\}$ for its lower-tail CVaR. For integrable U, U' on a common probability space, this definition and the 1-Lipschitz hinge imply

$$
|C_\alpha(U) - C_\alpha(U')| \le \alpha^{-1} \sup_t |\mathbb{E}(t - U)_+ - \mathbb{E}(t - U')_+| \le \alpha^{-1}\mathbb{E}|U - U'|.
$$

Apply this conditional on the data with $U = W(\pi; \mathcal{P}, \theta)$ and the constant $U' = W(\pi; \mathcal{P}_0)$, whose CVaR equals itself. Combining it with the corresponding bound for the posterior mean gives

$$
\begin{aligned}
&\sup_{\pi\in\Pi} |R_{\alpha,\lambda,n}(\pi) - W(\pi; \mathcal{P}_0)| \\
&\quad \le \left(1 - \lambda + \frac{\lambda}{\alpha}\right) \mathbb{E}\left[\sup_{\pi\in\Pi} |W(\pi; \mathcal{P}, \theta) - W(\pi; \mathcal{P}_0)| \,\middle|\, \mathcal{D}_n\right].
\end{aligned} \tag{51}
$$

Adding and subtracting rewards at π^*_Π and $\widetilde{\pi}$ bounds the welfare regret by twice the supremum in (51) plus $\varepsilon_{\mathrm{opt}}$. Lemma 3 completes the proof. This uses the full fitted product posterior and does not require a posterior density or a unique quantile. □

E.3 Nuisance accuracy and full-sample fitting

Proof of Proposition 1. For fixed candidate functions,

$$
\mathbb{E}_0\{A(Y - \mu_1) \mid X\} = -e_0\delta_1, \qquad \mathbb{E}_0\{(1 - A)(Y - \mu_0) \mid X\} = -(1 - e_0)\delta_0.
$$

Subtracting τ_0 from the expected orthogonal score gives

$$
\delta_1 - \delta_0 - \frac{e_0\delta_1}{e} + \frac{(1 - e_0)\delta_0}{1 - e} = (e - e_0)\left\{\frac{\delta_1}{e} + \frac{\delta_0}{1 - e}\right\}.
$$

Multiply by $\pi(X)$, use $0 \le \pi \le 1$ and $e, 1 - e \ge \kappa$, and apply Cauchy–Schwarz to prove (41). For the IPW score,

$$
\mathbb{E}_0[\phi_\theta(O; 1) \mid X] - \tau_0(X) = (e_0 - e)\left\{\frac{\mu_{1,0}}{e} + \frac{\mu_{0,0}}{1 - e}\right\}.
$$

Its bracket is bounded by $2H/\kappa$, so integration and $\|e - e_0\|_1 \le \|e - e_0\|_2$ give the asserted bound. For the effect-based score, $|\mathbb{E}_0\{\pi(X)(\tau_\theta - \tau_0)\}| \le \|\tau_\theta - \tau_0\|_2$. Finally, integrate over the fold index and nuisance draw, whose joint law assigns mass $(n_k/n)q_{-k}(d\theta_k)$. Cauchy–Schwarz under this probability measure proves (42); it requires no independence between nuisance components. The stated clipping correction follows from the triangle inequality. □

Proof of Proposition 2. For every data realization, posterior integration of the empirical fluctuation is bounded by the deterministic joint class:

$$\int \sup_{\pi\in\Pi} |(\mathcal{P}_n - \mathcal{P}_0)\phi_\theta(\,\cdot\,;\pi)|\, q_n(d\theta) \le \sup_{f\in\mathcal{F}} |(\mathcal{P}_n - \mathcal{P}_0)f|.$$

Let $C_{A_F} > 0$ denote a constant depending only on A_F. Symmetrization and the assumed entropy bound give expected supremum at most $C_{A_F} M\sqrt{w/n}$. Replacing one observation changes that supremum by at most $2M/n$. Thus, with probability at least $1-\delta$, it is at most $C_{A_F} M\sqrt{w/n} + M\sqrt{2\log(1/\delta)/n}$. Lemma 2 still controls bootstrap fluctuation by $C_\Pi M\sqrt{v/n}$, uniformly over the fitted score vector. Adding the population bias and combining the two square-root terms proves

$$\mathbb{E}\left[\sup_{\pi} |W(\pi;\mathcal{P},\theta) - W(\pi;\mathcal{P}_0)| \,\middle|\, \mathcal{D}_n\right] \le CM\sqrt{\frac{v + w + \log(1/\delta)}{n}} + b_n.$$

Replace (48) by this bound in (50) and (51). This proves the posterior and ideal-reward claims. The conditional simulation proof below does not require separation of nuisance-fitting and evaluation observations, so the finite-bank claim follows from the same substitution. □

E.4 Uniform concentration of the estimated loss

Proof of Theorem 2. Throughout this proof the data and fitted nuisance law are fixed. For $t < -M$, every hinge is zero, so increasing t to $-M$ cannot increase the loss. For $t > M$, every hinge equals $t - W$ and the loss has nonnegative slope $\lambda(1/\alpha - 1)$ in t. Decreasing t to M cannot increase it. This justifies restricting both threshold optimizations to $[-M, M]$, including at $\lambda = 0$ or $\alpha = 1$.

For the proof only, write

$$\ell(w,t) = -(1-\lambda)w - \lambda t + \frac{\lambda}{\alpha}(t-w)_+, \qquad a = 1 - \lambda + \lambda/\alpha.$$

On $w, t \in [-M, M]$, this loss is a-Lipschitz in w, and is $\lambda\max\{1, 1/\alpha - 1\} \le a$-Lipschitz in t. Its range lies in $[-M, (2a-1)M]$, an interval of length $2Ma$.

To control the loss class indexed by (π, t), take any law H on posterior pairs supported on the bounded score functions. Write $\omega_i = \mathbb{E}_H B_i$ and $Q_H = \sum_i \omega_i \delta_{X_i}$. Jensen's inequality and the linear score representation yield

$$|W(\pi;\mathcal{P},\theta) - W(\pi';\mathcal{P},\theta)|^2 \le M^2 \sum_i B_i |\pi(X_i) - \pi'(X_i)|^2.$$

Consequently,

$$\begin{aligned}&\|\ell\{W(\pi;\mathcal{P},\theta), t\} - \ell\{W(\pi';\mathcal{P},\theta), t'\}\|_{L_2(H)} \\ &\quad\le aM\|\pi - \pi'\|_{L_2(Q_H)} + a|t - t'|.\end{aligned}$$

Combining an ϵ-net for Π with a threshold grid of mesh $M\epsilon$ gives an $L_2(H)$ cover of the scenario-loss class, at radius $2Ma\epsilon$, with log cardinality at most $C_\Pi(v+1)\log(C_\Pi/\epsilon)$. This holds uniformly in H. Thus the complexity is derived from the policy class and does not assume that posterior integration preserves pseudo-dimension.

Let U_S be the supremum in (19). Shift the scenario loss by M to place it in $[0, 2Ma]$ without changing U_S. Conditional symmetrization for bounded classes (van der Vaart & Wellner, 2023, Thm. 2.14.1) and the entropy-integral argument of Lemma 1 give $\mathbb{E}(U_S \mid \mathcal{D}_n) \le C_\Pi Ma\sqrt{(v+1)/S}$. Replacing one independent draw pair changes U_S by at most $2Ma/S$. McDiarmid's inequality therefore implies

$$\mathbb{P}\left[U_S > C_\Pi Ma\sqrt{(v+1)/S} + Ma\sqrt{2\log(1/\delta_S)/S} \,\middle|\, \mathcal{D}_n\right] \le \delta_S.$$

Increasing C_Π gives (19), with the bound denoted by (35). The inequality $|\inf_t f(t) - \inf_t g(t)| \le \sup_t |f(t) - g(t)|$ also proves (36) on the same event.

□

Proof of Corollary 2. The event in Lemma 3 fails with repeated-sample probability at most δ. The simulation event in Theorem 2 fails with conditional probability at most δ_S for every data realization, hence also unconditionally. A union bound gives their intersection probability at least $1 - \delta - \delta_S$, without an independence assumption. On this event, (51) and (36) imply

$$\sup_{\pi \in \Pi} |\widehat{R}_{\alpha,\lambda,n}(\pi) - W(\pi; \mathcal{P}_0)| \leq \Delta_{n,S}.$$

The identity $\inf_t \widehat{\mathcal{L}}_{\alpha,\lambda,n}(\pi, t) = -\widehat{R}_{\alpha,\lambda,n}(\pi)$ proves (38). Add and subtract estimated rewards at π^*_Π and $\widehat{\pi}$; the two welfare-estimation errors are bounded by $\Delta_{n,S}$ each, and the middle difference is at most $\xi_{n,S}$. This proves (39). If a returned pair has joint empirical loss excess ξ_{pair}, then

$$\xi_{n,S} = \inf_t \widehat{\mathcal{L}}_{\alpha,\lambda,n}(\widehat{\pi}, t) - \inf_{\pi,t} \widehat{\mathcal{L}}_{\alpha,\lambda,n}(\pi, t) \leq \xi_{\text{pair}}.$$

Finally, for arbitrarily small fixed failure probabilities, inserting the rates for b_n, sampling, and simulation proves (43). The fixed constants in the probability bounds are independent of $n, v_n, S_n, \alpha, \lambda$. □

Proof of Proposition 3. The pointwise inequality $|\mathbf{1}\{u > 0\} - \sigma(u/T)| \leq e^{-|u|/T}$ holds for all real u, including zero. Since $W(\pi; \mathcal{P}_0) = \mathbb{E}_0\{\pi(X)\tau_0(X)\}$, integration gives (44). Add and subtract $W(\pi_T; \mathcal{P}_0)$ to obtain the corresponding regret bound. □