CAUSAL PREDICT-THEN-OPTIMIZE: TREATMENT EFFECT ESTIMATION UNDER DOWNSTREAM OPTIMIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Many operational decisions depend on estimating heterogeneous treatment effects and then using those estimates in constrained downstream optimization. In this setting, accurate causal prediction does not necessarily lead to good decisions, because estimation errors matter most when they occur near decision boundaries and alter the optimizer. We develop causal predict-then-optimize (CPO), which combines doubly robust causal estimation with decision-focused learning, and propose decision-calibrated causal predict-then-optimize (DC-CPO), which explicitly accounts for downstream decision sensitivity. Our framework characterizes causal uncertainty through its orientation relative to optimization boundaries and introduces a boundary-sensitive regularizer that emphasizes prediction errors along decision-relevant directions. The resulting method is designed to improve decisions rather than treatment-effect prediction accuracy alone. Across synthetic and semi-synthetic benchmarks, CPO+ achieves lower downstream regret than prediction-oriented causal learners, while DC-CPO further reduces regret and decision-boundary crossing. Our results show that causal prediction accuracy and downstream decision quality are distinct objectives, and that explicitly calibrating causal estimators to optimization boundaries can lead to better constrained treatment allocation.

1 INTRODUCTION

In many real-world prescriptive analytics problems, a key task is to decide which units should receive treatment when intervention resources are limited, so as to maximize total benefit. Hospitals allocate limited follow-up capacity to patients to maximize treatment benefit; editorial teams allocate limited promotional opportunities to articles to maximize incremental user response. Within the predict-then-optimize paradigm, we formulate this class of problems as causal predict-then-optimize (CPO). Causal inference is used to estimate the unobserved benefit, which then enters a constrained downstream optimization problem to maximize total benefit.

Predict-then-optimize addresses this class of problems. Unknown optimization coefficients are predicted from contextual features and then supplied to a downstream optimizer. Classical approaches train the predictor by minimizing squared error, treating prediction and optimization as independent stages. Elmachtoub & Grigas (2022) show that this separation is suboptimal because not all prediction errors are equally consequential for the downstream decision, and propose SPO+, a convex surrogate that trains predictions by the decision regret they induce. Subsequent work extends this idea to broader settings such as portfolio optimization, energy scheduling, and vehicle routing (Donti et al., 2017; Mandi et al., 2024). The standard SPO formulation assumes that optimization coefficients are observed during training.

In treatment allocation problems, this assumption breaks down. A hospital observes patient outcomes only under the treatment actually administered; an editorial team observes each article's response under its assigned promotional status. For each unit, only one potential outcome is observed, so the benefit entering the optimizer is a conditional average treatment effect (CATE) that must be estimated from observational data subject to confounding and limited overlap. Even when a reasonable CATE estimate is obtained, a second challenge arises from the downstream optimization. For a polyhedral

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.

feasible region, an estimation error along the normal to a decision boundary can alter the optimal allocation, while an error of the same size in a tangential direction leaves that pairwise comparison unchanged. Prediction-oriented losses such as MSE or PEHE weight all directions equally and therefore cannot distinguish these two cases.

To address these challenges, we develop a causal predict-then-optimize (CPO) framework. We define CPO regret to measure the decision loss from optimizing a predicted rather than the true causal coefficient. To train on CPO regret, we use the convex CPO+ surrogate, whose expected pseudo-benefit loss upper-bounds causal regret when either nuisance model is correct. To understand when and why CATE estimation errors lead to poor decisions, we provide a geometric characterization showing that decision risk depends on the orientation and magnitude of estimation uncertainty relative to decision boundaries, and define the Decision Fragility Index (DFI) to quantify this sensitivity. We establish a decision-invariance criterion based on DFI and derive a boundary-crossing probability bound governed by directional uncertainty and decision margins, which in turn yields an expected regret bound. This motivates decision-calibrated causal predict-then-optimize (DC-CPO), which augments CPO+ with a regularizer that penalizes estimation error along directions near active boundaries. For fixed decision geometry, we establish population optimality of the true causal coefficient under conditional residual symmetry, a decomposition connecting the regularizer to directional prediction error, and an oracle inequality. Experiments on synthetic and semi-synthetic benchmarks designs show that DC-CPO achieves lower regret than prediction-focused baselines.

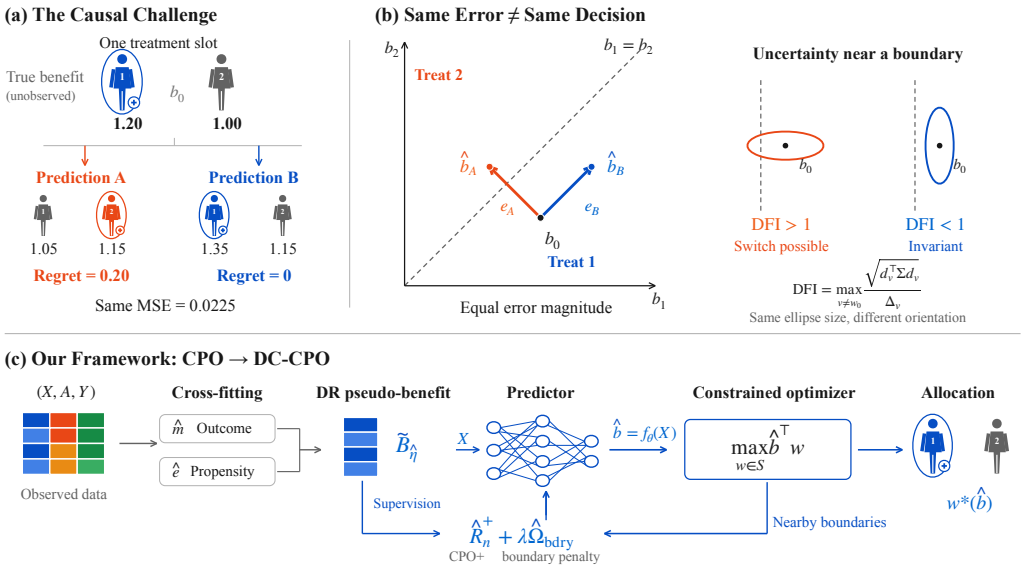


Figure 1: Equal prediction errors can yield different allocations and regret because decision sensitivity depends on error direction relative to decision boundaries. CPO+ learns from doubly robust pseudo-benefits, and DC-CPO adds a penalty along nearby decision-boundary directions.

Our main contributions are summarized as follows:

- We formulate CPO to address the challenge of unobserved benefits. We develop a convex surrogate CPO+ for training causal benefit predictors toward downstream decisions.
- We characterize causal estimation uncertainty relative to downstream decision boundaries. We further define DFI to quantify decision sensitivity and DC-CPO to penalize estimation errors along these critical directions.
- We establish a formal relationship between decision-boundary uncertainty and boundary-crossing probability, and derive an oracle inequality bounding causal regret.

The paper is organized as follows. Section 2 introduces the causal and optimization setup. Section 3 develops CPO and CPO+. Section 4 studies decision uncertainty and introduces DC-CPO. Section 5 establishes the main theoretical properties. Section 6 compares our methods with baselines on synthetic and semi-synthetic benchmarks.

2 PROBLEM SETUP

2.1 PREDICT-THEN-OPTIMIZE WITH CAUSAL COEFFICIENTS

We consider a downstream linear optimization problem. For a given coefficient vector $b \in \mathbb{R}^d$, the decision maker chooses $w \in S$ to maximize

$$V^*(b) = \max_{w \in S} b^\top w, \qquad w^*(b) \in \arg\max_{w \in S} b^\top w, \tag{1}$$

where $S \subseteq \mathbb{R}^d$ is nonempty and compact, and $w \in S$ denotes the downstream decision vector. In the contextual setting, the relevant coefficient is an unknown function $b_0(x) \in \mathbb{R}^d$. A predictor f_θ produces $\widehat{b}(x) = f_\theta(x)$, and the deployed decision is $w^*\{\widehat{b}(x)\}$. The learning objective is therefore to estimate $b_0(x)$ so that the induced decision performs well under the true downstream objective.

In this paper, the benefit entering the optimization problem is determined by a causal treatment effect:

$$b_0(x) = r(x) + B(x)\tau_0(x), \qquad \tau_0(x) = \mathbb{E}[Y(1) - Y(0) \mid X = x], \tag{2}$$

where $r(x) \in \mathbb{R}^d$ and $B(x) \in \mathbb{R}^d$ are known problem-specific mappings, and $\tau_0(x)$ is the conditional average treatment effect (CATE). Because $\tau_0(x)$ is defined through potential outcomes that are not jointly observed for the same unit, neither $\tau_0(x)$ nor $b_0(x)$ is directly observed, and identification from observed data requires causal assumptions.

Write $e_0(x) = \Pr(A = 1 \mid X = x)$. We observe i.i.d. data $O_i = (X_i, A_i, Y_i)$, $i = 1, \ldots, n$, where $A_i \in \{0, 1\}$ is the treatment indicator and $Y_i = Y_i(A_i)$ is the observed outcome. We impose the following standard causal conditions.

Assumption 2.1. (i) **SUTVA:** $Y = AY(1) + (1 - A)Y(0)$, with no interference; (ii) **Ignorability:** $(Y(1), Y(0)) \perp A \mid X$; (iii) **Overlap:** $\epsilon \le e_0(X) \le 1 - \epsilon$ a.s. for some $\epsilon \in (0, 1/2)$; (iv) **Boundedness:** $|Y(a)| \le M_Y$ and $\|B(X)\| \le M_B$ a.s. for finite M_Y, M_B.

Assumption 2.1(i)–(iii) connect the potential outcomes to the observed distribution and identify $\tau_0(x) = \mu_{0,1}(x) - \mu_{0,0}(x)$, $\mu_{0,a}(x) = \mathbb{E}[Y \mid A = a, X = x]$, and hence identify $b_0(x)$ in equation 2. Assumption 2.1(iv) provides the moment and tail control used later for pseudo-benefit concentration and decision-risk analysis.

Our objective is not merely to estimate $b_0(x)$ accurately, but to induce a good downstream decision. For a prediction $\widehat{b}$ and reference coefficient b, define the causal predict-then-optimize (CPO) regret

$$\ell_{\mathrm{CPO}}(\widehat{b}, b) = V^*(b) - b^\top w^*(\widehat{b}). \tag{3}$$

At $b = b_0(x)$, this is the value lost by optimizing the predicted rather than the true causal coefficient.

2.2 DOUBLY ROBUST PSEUDO-BENEFITS

Unlike standard supervised predict-then-optimize problems, $b_0(X_i)$ is not directly observed because only one potential outcome is observed for each unit. We therefore use a doubly robust pseudo-benefit constructed from the observed data.

Let $\eta = (\mu_1, \mu_0, e)$ denote candidate outcome regressions and a propensity score with $0 < e(X) < 1$ almost surely. We assume finite pseudo-benefit moments for the expectations below. Define the AIPW score and the corresponding pseudo-benefit as

$$\Gamma_\eta(O) = \mu_1(X) - \mu_0(X) + \frac{A\{Y - \mu_1(X)\}}{e(X)} - \frac{(1 - A)\{Y - \mu_0(X)\}}{1 - e(X)}, \tag{4}$$

$$\widetilde{B}_\eta(O) = r(X) + B(X)\Gamma_\eta(O) \in \mathbb{R}^d. \tag{5}$$

Its key property is conditional-mean double robustness.

Lemma 2.2 (Conditional-mean double robustness)**.** *Under Assumption 2.1(i)–(iii), if either $\mu_a = \mu_{0,a}$ for both $a \in \{0, 1\}$ or $e = e_0$, then $\mathbb{E}[\widetilde{B}_\eta(O) \mid X = x] = b_0(x)$ for almost every x.*

Thus, although $b_0(X_i)$ is unobserved, the pseudo-benefit has the correct conditional mean whenever either the outcome regression or the propensity score is correctly specified. The proof is given in Appendix B.1.1. In practice, we estimate the nuisance functions by cross-fitting and use $\widetilde{B}_{\widehat{\eta}_i}(O_i)$ as the training signal.

3 CPO AND CPO+

Section 2 provides an observable training signal through the doubly robust pseudo-benefit. In this section, we consider how this signal should be learned when performance is ultimately evaluated by the downstream optimization problem. We first use the CPO regret to measure the decision consequence of prediction error, and then introduce CPO+ as a tractable surrogate for training.

As illustrated in Figure 2, the benefit space is partitioned into regions associated with different optimal decisions. Predictions with similar errors can therefore lead to different decision losses depending on whether they cross a decision boundary. This is the motivation for using the CPO regret in equation 3, rather than a pointwise prediction loss, as the decision-aware learning target.

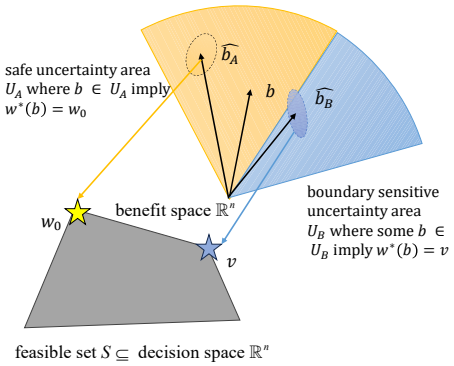


Figure 2: Prediction error and decision loss need not be aligned. The yellow region induces the oracle decision w_0, while the blue region induces the competing vertex v, where w_0 and v are as defined in Section 4.

Directly minimizing CPO is difficult because $w^*(\widehat{b})$ changes discontinuously when $\widehat{b}$ crosses a decision boundary. We therefore adapt the SPO+ construction of Elmachtoub & Grigas (2022) to the benefit-maximization setting and refer to the resulting surrogate as *CPO+*. CPO+ replaces the discontinuous dependence on $w^*(\widehat{b})$ by the support function of the feasible set while retaining the true decision $w^*(b)$ as the reference. Let $\xi_S(u) = \max_{w \in S} u^\top w$ denote the support function of S.

Definition 3.1 (CPO+ loss)**.** For a prediction $\widehat{b}$ and reference coefficient b, define

$$\ell_{\mathrm{CPO}+}(\widehat{b}, b) = \xi_S(2\widehat{b} - b) - 2\widehat{b}^\top w^*(b) + V^*(b). \tag{6}$$

CPO+ preserves the downstream optimization structure of CPO but is convex in the predicted coefficient $\widehat{b}$. As shown below, it also upper-bounds the CPO regret and admits a subgradient that can be computed using the same optimization oracle.

Proposition 3.2 (Basic properties of CPO+)**.** *For every $b, \widehat{b} \in \mathbb{R}^d$:* (i) $\ell_{\mathrm{CPO}}(\widehat{b}, b) \le \ell_{\mathrm{CPO}+}(\widehat{b}, b)$*;* (ii) $\ell_{\mathrm{CPO}+}(\cdot, b)$ *is convex; and* (iii) $2\{w^*(2\widehat{b} - b) - w^*(b)\} \in \partial_{\widehat{b}} \ell_{\mathrm{CPO}+}(\widehat{b}, b)$ *is a computable subgradient.*

Proposition 3.2 explains why CPO+ is useful for decision-focused learning. Part (i) makes it a conservative surrogate for the true CPO regret. Part (ii) provides a convex objective in the predicted coefficient, while Part (iii) shows that its subgradient can be computed using the optimization oracle. A proof is given in Appendix B.1.2.

The pointwise bound in Proposition 3.2 is stated for an observed reference coefficient b. In our setting, $b_0(X)$ is not observed and is replaced during training by the pseudo-benefit $\widetilde{B}_\eta(O)$.

Let $\bar{b}_\eta(x) = \mathbb{E}[\widetilde{B}_\eta(O) \mid X = x]$, and define the support function of the decision-difference set $\xi_{S-S}(u) = \max_{w, w' \in S} u^\top (w - w')$.

Theorem 3.3 (Causal Regret Bound via Pseudo-Benefits)**.** *Under conditional integrability, for almost every x and every $\widehat{b}$,*

$$\ell_{\mathrm{CPO}}\{\widehat{b}, b_0(x)\} \le \mathbb{E}\Big[\ell_{\mathrm{CPO}+}\{\widehat{b}, \widetilde{B}_\eta(O)\} \mid X = x\Big] + \xi_{S-S}\{\bar{b}_\eta(x) - b_0(x)\}. \tag{7}$$

If either nuisance model is correctly specified, the final term vanishes.

Theorem 3.3 connects the observable pseudo-benefit loss to the latent causal decision objective. In particular, if either the outcome regression or the propensity score is correctly specified, every measurable $f : \mathcal{X} \to \mathbb{R}^d$ satisfies

$$\mathbb{E}[\ell_{\mathrm{CPO}}\{f(X), b_0(X)\}] \le \mathbb{E}\Big[\ell_{\mathrm{CPO}+}\{f(X), \widetilde{B}_\eta(O)\}\Big]. \tag{8}$$

Thus, the expected CPO+ loss computed from observable pseudo-benefits provides an upper bound on the expected regret of the decision induced by f under the true causal coefficient $b_0(X)$. The proof is given in Appendix B.1.3.

For estimation, let $\widehat{\eta}_i$ denote nuisance functions fitted without observation i through cross-fitting. The empirical training objective is therefore

$$\widehat{R}_n^+(\theta) = \frac{1}{n}\sum_{i=1}^{n} \ell_{\mathrm{CPO}+}\left\{ f_\theta(X_i), \widetilde{B}_{\widehat{\eta}_i}(O_i) \right\}. \tag{9}$$

Thus, the doubly robust pseudo-benefit supplies an observable training signal, while CPO+ provides a tractable decision-aware surrogate for the downstream CPO regret.

4 Decision-Sensitive Uncertainty Geometry

4.1 Geometric Decision Uncertainty

We characterize decision uncertainty by comparing the directional width of the prediction uncertainty with the margin to an optimizer boundary. Fix x, write $b = b_0(x)$, and let $w_0 = w^*(b)$ be the unique oracle vertex of a polytope S with vertex set $\mathcal{V}$. We fix a measurable, vertex-valued tie-breaking rule. For each $v \in \mathcal{V} \setminus \{w_0\}$, define $d_v = w_0 - v$ and $\Delta_v = b^\top d_v > 0$. Then $\widehat{b}^\top d_v = \Delta_v + d_v^\top(\widehat{b} - b)$. Here d_v is normal to the boundary between w_0 and v, and Δ_v is the corresponding decision margin. A switch toward v therefore requires $d_v^\top(\widehat{b} - b) \le -\Delta_v$.

Let $U = b + E$ be a convex uncertainty set and $h_E(u) = \sup_{e\in E} e^\top u$ its support function. We define

$$\mathrm{DFI}(b, U) = \max_{v \neq w_0} \frac{h_E(v - w_0)}{\Delta_v}, \tag{10}$$

where the numerator measures the largest uncertainty displacement toward a competitor and the denominator is the margin protecting the oracle decision.

Theorem 4.1 (Decision invariance)**.** *The decision fragility index characterizes whether the uncertainty set can cross an optimizer boundary:*

(i) *If* $\mathrm{DFI}(b, U) < 1$*, then* $w^*(b') = w_0$*,* $\forall\, b' \in U$*.*

(ii) *If* $\mathrm{DFI}(b, U) > 1$*, then there exists* $b' \in U$ *and a competitor* $v \neq w_0$ *such that* $b'^\top v > b'^\top w_0$*, so a decision switch is possible.*

Thus DFI converts the geometry of prediction uncertainty into a boundary-crossing certificate: values below one guarantee decision invariance, while exceeding one in some direction makes a decision switch possible. For ellipsoidal uncertainty, DFI depends on $\sqrt{d_v^\top \Sigma d_v}/\Delta_v$; the derivation and geometric examples are given in Appendix B.2.1.

4.2 Boundary-crossing Probability

Starting from the DFI characterization, we now estimate the probability of crossing a decision boundary, which provides a probabilistic measure of downstream decision uncertainty. Let $\widehat{b}_n(x)$ denote an estimator of $b_0(x)$ based on a sample of size n, and let $\mathcal{A}_0$ denote the vertices adjacent to w_0.

Assumption 4.2 (Directional concentration)**.** For fixed x,

$$\widehat{b}_n - b = \frac{1}{n}\sum_{i=1}^{n} \psi_x(O_i) + r_n, \qquad \mathbb{E}[\psi_x(O_i)] = 0,$$

and, for every $v \in \mathcal{A}_0$,

$$d_v^\top \psi_x(O_i) \text{ is } \nu_v^2\text{-sub-Gaussian}, \qquad \Pr\{|d_v^\top r_n| > \rho_{v,n}\} \le \eta_{v,n}, \qquad 0 \le \rho_{v,n} < \Delta_v.$$

Theorem 4.3 (Directional boundary-crossing bound)**.** *Under Assumption 4.2,*

$$\Pr\{w^*(\widehat{b}_n) \neq w_0\} \le \min\left\{ 1, \sum_{v\in\mathcal{A}_0} \left[\exp\left\{ -\frac{n(\Delta_v - \rho_{v,n})^2}{2\nu_v^2} \right\} + \eta_{v,n} \right] \right\}. \tag{11}$$

Theorem 4.3 controls the probability that the estimated benefit induces a decision different from w_0. The bound is governed by $\sqrt{n}(\Delta_v - \rho_{v,n})/\nu_v$ and the remainder probabilities $\eta_{v,n}$; for Gaussian errors with covariance Σ/n, the directional variance is $d_v^\top \Sigma d_v/n$.

Let $\Delta_{\max} = \max_{v \neq w_0} \Delta_v$. We translate this decision-error probability into CPO regret. Under the fixed vertex-valued tie-breaking rule,

$$\mathbb{E}[\ell_{\mathrm{CPO}}(\widehat{b}_n, b)] = \sum_{v \neq w_0} \Delta_v \Pr\{w^*(\widehat{b}_n) = v\} \leq \Delta_{\max} \Pr\{w^*(\widehat{b}_n) \neq w_0\}. \tag{12}$$

Thus, the same directional margins and uncertainties that control boundary-crossing probability also control expected CPO regret. Complete proofs and the Gaussian specialization are given in Appendix B.2.

4.3 DC-CPO Algorithm under DFI Regularization

The DFI and crossing bound show that decision risk is driven by directional uncertainty relative to the decision margin. We therefore use

$$\frac{\left[d_v^\top \{f_\theta(X) - \widetilde{B}_{\widehat{\eta}}(O)\}\right]^2}{\Delta_v^2} \tag{13}$$

as a trainable, margin-weighted proxy based on projected pseudo-benefit error.

DC-CPO estimates the quantities in equation 13 from the current predictor. At outer iteration t, for each observation j, we first compute the current benefit prediction and decision, $\widehat{b}_j^{(t)} = f_{\theta^{(t)}}(X_j)$, $\widehat{w}_j^{(t)} = w^*(\widehat{b}_j^{(t)})$. We then identify the nearby decision boundaries by retaining adjacent competitors whose predicted margins are at most ρ:

$$\widehat{\mathcal{A}}_j^{(t)}(\rho) = \left\{ v \in \mathrm{Adj}(\widehat{w}_j^{(t)}) : 0 \leq (\widehat{b}_j^{(t)})^\top (\widehat{w}_j^{(t)} - v) \leq \rho \right\}. \tag{14}$$

For each retained competitor, the estimated boundary direction, margin, and weight are

$$d_{jv}^{(t)} = \widehat{w}_j^{(t)} - v, \qquad \widehat{\Delta}_{jv}^{(t)} = (\widehat{b}_j^{(t)})^\top d_{jv}^{(t)}, \qquad \omega_{jv}^{(t)} = \min\left\{ \frac{1}{(\widehat{\Delta}_{jv}^{(t)} + \varepsilon)^2}, W \right\}. \tag{15}$$

Thus ρ determines which boundaries are locally relevant, while $\omega_{jv}^{(t)}$ gives greater weight to smaller-margin boundaries; ε and W stabilize the inverse-margin weighting.

Using this estimated geometry, we define

$$\widehat{\Omega}_{\mathrm{bdry}}^{(t)}(\theta) = \frac{1}{n} \sum_{j=1}^{n} \frac{1}{\max\{1, |\widehat{\mathcal{A}}_j^{(t)}(\rho)|\}} \sum_{v \in \widehat{\mathcal{A}}_j^{(t)}(\rho)} \omega_{jv}^{(t)} \left[(d_{jv}^{(t)})^\top \{f_\theta(X_j) - \widetilde{B}_{\widehat{\eta}_j}(O_j)\} \right]^2. \tag{16}$$

The regularizer penalizes projected pseudo-benefit error along the currently relevant decision boundaries.

Holding the estimated geometry fixed, DC-CPO updates

$$\theta^{(t+1)} \in \arg\min_\theta \left\{ \widehat{R}_n^+(\theta) + \lambda \widehat{\Omega}_{\mathrm{bdry}}^{(t)}(\theta) \right\}. \tag{17}$$

Each outer iteration therefore alternates between estimating decision-sensitive geometry and refitting the causal predictor under that geometry.

5 Theoretical Properties of DC-CPO

We analyze one inner DC-CPO update with the decision geometry held fixed. We first establish consistency for the downstream CPO objective, then characterize the population target of the DFI penalty, and finally give finite-sample guarantees.

Algorithm 1 Decision-Calibrated Causal Predict-then-Optimize (DC-CPO)

Require: Observations $\{O_i = (X_i, A_i, Y_i)\}_{i=1}^n$; feasible set S; predictor f_θ; regularization parameter λ; boundary threshold ρ; stabilization parameters ε, W; outer iterations T.

```
Cross-fit nuisance functions η̂_i and construct B̃_{η̂_i}(O_i).
Initialize θ^(0) by minimizing R̂_n^+(θ).
for t = 0, ..., T − 1 do
    for j = 1, ..., n do
        Compute b̂_j^(t) = f_{θ^(t)}(X_j) and ŵ_j^(t) = w*(b̂_j^(t)).
        Construct Â_j^(t)(ρ) using equation 14.
        for v ∈ Â_j^(t)(ρ) do
            Compute (d_jv^(t), Δ̂_jv^(t), ω_jv^(t)) using equation 15.
        end for
    end for
    Form Ω̂_bdry^(t) using equation 16.
    Update θ^(t+1) by solving equation 17 with the current geometry fixed.
end for
return f_{θ^(T)} and x ↦ w*{f_{θ^(T)}(x)}.
```

5.1 Causal Fisher Consistency

Assumption 5.1 (Conditional residual symmetry)**.** The outcome regression is correctly specified and, for $a \in \{0, 1\}$,

$$Y(a) - \mu_{0,a}(X) \mid X \overset{d}{=} -\{Y(a) - \mu_{0,a}(X)\} \mid X.$$

Theorem 5.2 (Causal Fisher consistency)**.** *Under Assumptions 2.1 and 5.1, for almost every x,*

$$b_0(x) \in \arg\min_{a \in \mathbb{R}^d} \mathbb{E}\Big[\ell_{\mathrm{CPO}+}\{a, \widetilde{B}_\eta(O)\} \mid X = x\Big]. \tag{18}$$

Moreover, since the population DFI penalty is minimized at $f(x) = b_0(x)$, the same conclusion holds for the regularized DC-CPO population objective for every $\lambda \geq 0$.

Thus, under conditionally symmetric outcome noise, CPO+ and DC-CPO retain the true causal coefficient as a population minimizer. If the conditional minimizer is unique, it follows that $f^*(x) = b_0(x)$ and hence $w^*\{f^*(x)\} = w^*\{b_0(x)\}$. The proof is given in Appendix B.3.1.

5.2 Decomposition of the DFI Penalty

For fixed decision geometry, let $\mathcal{A}_\rho(X)$ denote the retained competitors, with directions $d_v(X)$, $m_\rho(X) = \max\{1, |\mathcal{A}_\rho(X)|\}$, and let $\omega_v(X)$ be the corresponding nonnegative weights, all fixed as f varies. Define

$$\Omega_\rho(f) = \mathbb{E}\left[\frac{1}{m_\rho(X)} \sum_{v \in \mathcal{A}_\rho(X)} \omega_v(X)\big[d_v(X)^\top\{f(X) - \widetilde{B}_\eta(O)\}\big]^2\right]. \tag{19}$$

Proposition 5.3 (Directional-error decomposition)**.** *If either the outcome regression or the propensity score is correctly specified, then*

$$\Omega_\rho(f) = D_\rho(f) + C_\rho, \tag{20}$$

where $D_\rho(f)$ is the weighted prediction error projected onto the retained decision-boundary directions, and C_ρ is an irreducible pseudo-benefit noise term independent of f.

Proposition 5.3 shows that minimizing the population penalty is equivalent to minimizing weighted directional prediction error. The pseudo-benefit noise only adds an f-independent term and therefore does not change the population minimizer. The definitions of $D_\rho(f)$ and C_ρ, together with the proof, are given in Appendix B.3.2.

5.3 DC-CPO Oracle Inequality

We next quantify how the predictor obtained from the empirical DC-CPO objective compares with the best regularized predictor at the population level. Write $R_\eta^+(f) = \mathbb{E}[\ell_{\mathrm{CPO}+}\{f(X), \widetilde{B}_\eta(O)\}]$ and

$R_{\mathrm{CPO}}(f) = \mathbb{E}[\ell_{\mathrm{CPO}}\{f(X), b_0(X)\}]$. Let $\mathcal{F}$ denote the predictor class and $\lambda \geq 0$. Define

$$\widehat{L}_{n,\lambda}(f) = \widehat{R}^{+}_{n}(f) + \lambda\widehat{\Omega}_{\mathrm{bdry}}(f), \qquad L_{\lambda}(f) = R^{+}_{\eta}(f) + \lambda\Omega_{\rho}(f), \tag{21}$$

where $\widehat{L}_{n,\lambda}$ is the training objective and L_{λ} is its population counterpart. Let $\widehat{f}_{\lambda}$ be an $\varepsilon_{\mathrm{opt}}$-approximate empirical minimizer:

$$\widehat{L}_{n,\lambda}(\widehat{f}_{\lambda}) \leq \inf_{f\in\mathcal{F}} \widehat{L}_{n,\lambda}(f) + \varepsilon_{\mathrm{opt}}.$$

To transfer empirical optimality to population performance, suppose that, with probability at least $1-\delta$,

$$\sup_{f\in\mathcal{F}} \left|\widehat{L}_{n,\lambda}(f) - L_{\lambda}(f)\right| \leq \Gamma_n, \tag{22}$$

For a fixed loss class bounded in $[0, M_{\lambda}]$, a sufficient bound is

$$\Gamma_n = \underbrace{2\mathfrak{R}_n(\mathcal{G}_{\lambda})}_{\text{complexity}} + \underbrace{M_{\lambda}\sqrt{\frac{\log(2/\delta)}{2n}}}_{\text{sampling}} + \underbrace{\Delta_{n,\eta}}_{\text{pseudo-benefit}} + \underbrace{\Delta_{n,\mathrm{geom}}}_{\text{geometry}}. \tag{23}$$

Here $\mathcal{G}_{\lambda}$ is the fixed-geometry DC-CPO loss class and $\mathfrak{R}_n$ its expected absolute Rademacher complexity, while $\Delta_{n,\eta}$ and $\Delta_{n,\mathrm{geom}}$ measure the additional errors from pseudo-benefit and decision-geometry estimation.

Theorem 5.4 (DC-CPO oracle inequality)**.** *Suppose equation 22 holds with probability at least* $1-\delta$ *and either nuisance model is correctly specified. Then, on this event,*

$$R_{\mathrm{CPO}}(\widehat{f}_{\lambda}) \leq \underbrace{\inf_{f\in\mathcal{F}} \left\{R^{+}_{\eta}(f) + \lambda\Omega_{\rho}(f)\right\}}_{\textit{population target}} + \underbrace{2\Gamma_n}_{\textit{estimation error}} + \underbrace{\varepsilon_{\mathrm{opt}}}_{\textit{optimization}}. \tag{24}$$

Theorem 5.4 bounds the true CPO regret by the best regularized population risk, together with finite-sample estimation and optimization errors. The proof is given in Appendix B.3.3.

6 Experiments

Baselines and Metrics. We compare CPO+ and DC-CPO with two prediction-focused baselines, MSE-DR and MAE-DR, which fit the same doubly robust pseudo-outcomes using squared and absolute losses, respectively. We evaluate treatment-effect estimation error using $\mathrm{PEHE} = n_{\mathrm{eval}}^{-1}\sum_{i=1}^{n_{\mathrm{eval}}}\{\widehat{\tau}(X_i) - \tau_0(X_i)\}^2$; downstream decision loss using $R = K^{-1}\sum_{k=1}^{K}[V^*(b_k) - b_k^{\top} w^*(\widehat{b}_k)]$, where $K = K_{\mathrm{eval}}$ is the number of evaluation batches; and decision instability using $P_{\mathrm{cross}} = K^{-1}\sum_{k=1}^{K}\Pr_{\mathrm{boot}}\{w^*(\widehat{b}_k^{\mathrm{boot}}) \neq w^*(\widehat{b}_k) \mid \mathcal{D}\}$, the probability that bootstrap refitting changes the selected allocation, averaged over batches, where $\mathcal{D}$ is the estimation sample. Finally, R-gap $= 100\% \times (\overline{R}_{\mathrm{best}} - \overline{R}_{\text{DC-CPO}})/\overline{R}_{\mathrm{best}}$ measures the relative regret reduction of DC-CPO, where $\overline{R}_{\mathrm{best}} = \min_{m\in\{\text{MSE-DR,MAE-DR}\}}\overline{R}_m$ and bars denote averages across seeds. Lower values are better for PEHE, regret, and P_{cross}, whereas higher values are better for R-gap. Detailed method descriptions and metric definitions are provided in Appendix C.1.

Experimental Settings. For the synthetic experiment, we generate ten-dimensional correlated Gaussian covariates, assign treatment through a logistic propensity model, and generate outcomes with heterogeneous treatment effects and Gaussian noise. We vary model misspecification through $\tau_s(X) = (1-s)\tau_{\mathrm{lin}}(X) + s\tau_{\mathrm{nonlin}}(X)$, where τ_{lin} lies in the predictor's feature span and τ_{nonlin} introduces nonlinear components outside that span; the main setting uses $s = 1$. For downstream allocation, each batch contains 16 individuals with net benefits $b_i = \tau_s(X_i) - c_i$, where $c_i \sim \mathrm{Unif}(0.30, 1.20)$. We maximize total net benefit by selecting at most five individuals, with at most four from either of two equally sized groups formed by ranking X_1. In the main synthetic setting, DC-CPO reduces regret by 10.62%–10.99% relative to the best prediction-focused baseline. We further conduct semi-synthetic experiments using real covariates from Twins, NHANES, and ACTG-175 to evaluate constrained treatment allocation across different health-related covariate distributions. Across the three datasets and both within- and out-of-sample evaluations, DC-CPO reduces regret by 7.46%–16.14% relative to the best prediction-focused baseline. Full DGP, dataset, preprocessing, and implementation details are provided in Appendix C.2.

Table 1: Results on the synthetic DGP and real-covariate semi-synthetic datasets. Parentheses report standard errors over 50 seeds.

		Within				Out			
Database	Method / setting	PEHE	Regret	Crossing	R-Gap	PEHE	Regret	Crossing	R-Gap
Synthetic DGP	MSE-DR	0.659 (0.043)	1.393 (0.063)	0.878		0.652 (0.040)	1.379 (0.056)	0.877	
	MAE-DR	0.589 (0.040)	1.228 (0.039)	0.865		0.594 (0.040)	1.215 (0.039)	0.866	
	CPO+	0.578 (0.035)	1.104 (0.039)	0.857		0.581 (0.036)	1.112 (0.041)	0.854	
	DC-CPO	**0.529** (0.033)	**1.093** (0.035)	**0.845**	10.99%	**0.534** (0.036)	**1.086** (0.036)	**0.843**	10.62%
	DC-CPO$^-$	0.543 (0.034)	1.102 (0.036)	0.850		0.550 (0.037)	1.100 (0.036)	0.849	
Twins	MSE-DR	0.364 (0.025)	0.765 (0.042)	0.756		0.374 (0.026)	0.799 (0.044)	0.760	
	MAE-DR	0.407 (0.014)	0.843 (0.031)	0.757		0.419 (0.014)	0.837 (0.032)	0.757	
	CPO+	0.385 (0.014)	0.704 (0.029)	0.732		0.400 (0.014)	0.721 (0.028)	0.733	
	DC-CPO	**0.347** (0.013)	**0.692** (0.026)	**0.719**	9.54%	**0.363** (0.013)	**0.694** (0.026)	**0.721**	13.14%
	DC-CPO$^-$	0.352 (0.013)	0.694 (0.025)	0.728		0.368 (0.013)	0.703 (0.026)	0.729	
NHANES	MSE-DR	2.043 (0.851)	1.448 (0.085)	0.830		2.438 (1.090)	1.457 (0.082)	0.833	
	MAE-DR	1.308 (0.089)	1.513 (0.043)	0.789		1.313 (0.106)	1.491 (0.049)	0.797	
	CPO+	0.929 (0.040)	1.386 (0.037)	0.776		0.948 (0.050)	1.388 (0.044)	0.787	
	DC-CPO	**0.925** (0.042)	**1.340** (0.038)	**0.757**	7.46%	**0.942** (0.056)	**1.297** (0.042)	**0.766**	10.98%
	DC-CPO$^-$	0.943 (0.043)	1.344 (0.038)	0.764		0.959 (0.057)	1.315 (0.041)	0.772	
ACTG-175	MSE-DR	3.800 (1.104)	1.828 (0.082)	0.839		4.537 (1.403)	1.809 (0.080)	0.837	
	MAE-DR	1.689 (0.114)	1.749 (0.041)	0.814		1.781 (0.123)	1.815 (0.046)	0.817	
	CPO+	**1.022** (0.040)	1.594 (0.042)	0.812		**1.064** (0.042)	1.639 (0.046)	0.817	
	DC-CPO	1.137 (0.058)	**1.499** (0.034)	**0.785**	14.29%	1.196 (0.068)	**1.517** (0.041)	**0.788**	16.14%
	DC-CPO$^-$	1.176 (0.062)	1.516 (0.034)	0.794		1.235 (0.072)	1.546 (0.041)	0.799	

Performance Comparison. Table 1 shows that CPO+ achieves lower regret than MSE-DR and MAE-DR across all reported settings. DC-CPO further reduces regret and boundary-crossing probability relative to CPO+, with all seed-paired 95% confidence intervals for these differences lying below zero. These reductions can coincide with higher PEHE; on ACTG-175, for instance, DC-CPO reduces regret from 1.594 to 1.499 while PEHE rises from 1.022 to 1.137, confirming that decision-aware estimation quality and conventional prediction accuracy are distinct objectives. An oracle-normalized comparison and experiments with alternative feature representations appear in Appendices C.4 and C.6.

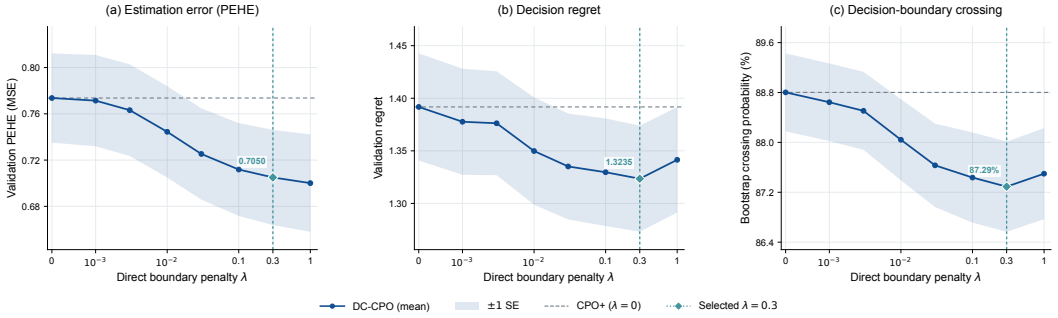


Figure 3: Effect of λ on PEHE, regret, and boundary-crossing probability.

Sensitivity and Ablation. Figure 3 shows that regret and boundary-crossing probability are lowest at $\lambda = 0.3$ among the displayed candidates, while further increases in λ yield only marginal reductions in PEHE, consistent with $\lambda = 0.3$. Beyond this value, both decision metrics increase slightly. Table 1 shows that DC-CPO consistently achieves lower regret and boundary-crossing probability than DC-CPO$^-$. DC-CPO$^-$ retains the projected-residual penalty and active-boundary updates but assigns equal weights to the retained directions, isolating the effect of nonuniform weighting. Additional sensitivity, ablation, and stress-test details are provided in Appendix C.5.

7 Discussion

This paper proposes a causal predict-then-optimize framework that learns treatment effects from observational data for constrained resource allocation, and characterizes the uncertainty of these estimates geometrically through its orientation relative to the decision boundaries. Further extensions include causal predict-then-optimize with multiple treatments and multiple resource constraints. A related formulation seeks the least costly allocation subject to a total causal benefit constraint, adapting to scenarios in which meeting a benefit target takes priority over the budget.

AI USE STATEMENT

In this work, we used generative AI tools to assist with the development and refinement of research methodology and experiments, mathematical derivations and proofs, method implementation, and interpretation of experimental results. Additionally, we used generative AI tools for literature search and summarization, code editing and debugging, and drafting and editing parts of the manuscript. We have reviewed and verified all AI-assisted work, including mathematical claims, experimental implementations, reported results, and cited references. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

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 source datasets are used only to provide empirical covariates, while treatments and potential outcomes are generated semi-synthetically for methodological evaluation. The proposed framework is intended for methodological research on constrained treatment allocation, and any real-world deployment would require application-specific consideration of ethical, legal, and fairness requirements.

REPRODUCIBILITY STATEMENT

The methodological formulation and theoretical results are provided in Sections 2–5, with complete proofs and additional theoretical details in Appendix B. Detailed descriptions of the baselines, synthetic and semi-synthetic data construction, experimental settings, evaluation metrics, sensitivity analyses, and stress tests are provided in Appendix C. Source code, experimental configurations, and scripts for reproducing the reported results will be provided at the following anonymous repository: `https://anonymous.4open.science/r/Causal-predict-then-optimize-B055`.

REFERENCES

Anastasios N. Angelopoulos and Stephen Bates. Conformal prediction: A gentle introduction. *Foundations and Trends in Machine Learning*, 16(4):494–591, 2023. doi: 10.1561/2200000101.

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

Aharon Ben-Tal, Laurent El Ghaoui, and Arkadi Nemirovski. *Robust Optimization*. Princeton University Press, 2009.

Dimitris Bertsimas and Nathan Kallus. From predictive to prescriptive analytics. *Management Science*, 66(3):1025–1044, 2020. doi: 10.1287/mnsc.2018.3253.

Dimitris Bertsimas, David B. Brown, and Constantine Caramanis. Theory and applications of robust optimization. *SIAM Review*, 53(3):464–501, 2011. doi: 10.1137/080734510.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 2018. doi: 10.1111/ectj.12097.

Priya L. Donti, Brandon Amos, and J. Zico Kolter. Task-based end-to-end model learning in stochastic optimization. In *Advances in Neural Information Processing Systems*, volume 30, pp. 5484–5494, 2017.

Othman El Balghiti, Adam N. Elmachtoub, Paul Grigas, and Ambuj Tewari. Generalization bounds in the predict-then-optimize framework. *Mathematics of Operations Research*, 48(4):2043–2065, 2023. doi: 10.1287/moor.2022.1330.

Adam N. Elmachtoub and Paul Grigas. Smart "predict, then optimize". *Management Science*, 68(1): 9–26, 2022. doi: 10.1287/mnsc.2020.3922.

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.

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

Yichun Hu, Nathan Kallus, and Xiaojie Mao. Fast rates for contextual linear optimization. *Management Science*, 68(6):4236–4245, 2022. doi: 10.1287/mnsc.2022.4383.

Yichun Hu, Nathan Kallus, Xiaojie Mao, and Yanchen Wu. Contextual linear optimization with bandit feedback. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Nathan Kallus. Balanced policy evaluation and learning. In *Advances in Neural Information Processing Systems*, volume 31, pp. 8895–8906, 2018.

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.

Jingren Liu, Hanzhang Qin, Junyi Liu, Mabel C. Chou, and Jong-Shi Pang. Offline policy learning with weight clipping and Heaviside composite optimization. *arXiv preprint arXiv:2601.12117*, 2026. Under revision at *Operations Research*.

Jayanta Mandi, James Kotary, Senne Berden, Maxime Mulamba, Víctor Bucarey, Tias Guns, and Ferdinando Fioretto. Decision-focused learning: Foundations, state of the art, benchmark and future opportunities. *Journal of Artificial Intelligence Research*, 80:1623–1701, 2024. doi: 10.1613/jair.1.15320.

James M. Robins, Andrea Rotnitzky, and Lue Ping Zhao. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89(427): 846–866, 1994.

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*, pp. 3076–3085, 2017.

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

Stefan Wager and Susan Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242, 2018. doi: 10.1080/01621459.2017.1319839.

Bryan Wilder, Bistra Dilkina, and Milind Tambe. Melding the data-decisions pipeline: Decision-focused learning for combinatorial optimization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 1658–1665, 2019. doi: 10.1609/aaai.v33i01.33011658.

Zhengyuan Zhou, Susan Athey, and Stefan Wager. Offline multi-action policy learning: Generalization and optimization. *Operations Research*, 71(1):148–183, 2023. doi: 10.1287/opre.2022.2271.

A Related Work

Smart predict-then-optimize and decision-focused learning. Smart predict-then-optimize (SPO) aligns prediction with downstream optimization through a regret-based loss and its convex surrogate, SPO+ (Elmachtoub & Grigas, 2022). Task-based and combinatorial decision-focused learning integrate prediction with optimization (Donti et al., 2017; Wilder et al., 2019; Mandi et al., 2024), while generalization and fast-rate analyses characterize their statistical performance (El Balghiti et al., 2023; Hu et al., 2022). These foundations explain why prediction errors should be evaluated These foundations explain why prediction errors should be evaluated through their decision consequences. In our setting, however, the optimization coefficient is a latent treatment-effect contrast, so fully observed coefficient supervision is unavailable. We construct a causal learning interface through doubly robust pseudo-benefits and establish a bound connecting their expected CPO+ loss to regret under the true causal coefficient.

Treatment effect estimation and policy learning. Augmented weighting and orthogonal estimation provide foundations for learning from incomplete or observational data (Robins et al., 1994; Chernozhukov et al., 2018), while meta-learners, causal forests, and representation learning address heterogeneous treatment-effect estimation (Künzel et al., 2019; Wager & Athey, 2018; Shalit et al., 2017). Policy learning uses causal scores to optimize treatment rules and establish value guarantees (Athey & Wager, 2021; Kallus, 2018; Zhou et al., 2023). Our work builds on these identification and estimation tools while retaining an explicit downstream allocation problem. We study how estimation uncertainty propagates through the optimizer: under coupled constraints, its decision consequences depend on the competing feasible allocations and the directions in which their relative benefits change.

Decision-aware treatment effect estimation. Frauen et al. (2025) show that CATE accuracy need not imply optimal threshold-based treatment decisions and develop retargeted estimation objectives. Liu et al. (2026) address unstable policy-value estimation under small propensities through weight clipping and a Heaviside composite formulation for linear-policy optimization. Both works demonstrate the importance of matching causal estimation to its decision use, focusing on individual treatment rules ($\pi\colon \mathcal{X} \to \{0,1\}$) learned from policy-value objectives. Our setting differs structurally: the decision variable is a joint allocation vector $w \in S$ subject to resource constraints that couple all units, so the optimizer is a linear program over a polytope rather than a per-unit classification. We examine how uncertainty in the causal benefit vector interacts with the boundaries between feasible allocations. Our decision fragility index and directional crossing analysis characterize how projected uncertainty interacts with decision margins. This explains why uncertainty sets with identical volume can have different decision implications and motivates regularizing prediction errors along allocation-boundary directions.

Beyond point predictions in predict-then-optimize. Recent work broadens the information supplied to downstream optimization. Hu et al. (2024) develop induced empirical risk minimization for contextual linear optimization with bandit feedback, where historical decisions reveal only projections of the cost vector. Offline policy learning instead constructs decision signals from estimated causal policy values (Liu et al., 2026). Hu et al. (2026) optimize prediction-set construction under explicit robustness constraints, showing that coverage alone can impose unnecessary conservatism. These approaches establish that useful decision information need not take the form of fully observed coefficient labels. Our framework connects this perspective to causal uncertainty geometry: doubly robust pseudo-benefits supply the training signal, while directional widths and decision margins identify the uncertainty components relevant to allocation changes.

Uncertainty sets, robust optimization, and conformal decision-making. Robust optimization protects decisions against parameter variation within uncertainty sets (Ben-Tal et al., 2009; Bertsimas et al., 2011), and prescriptive analytics connects contextual data to operational decisions (Bertsimas & Kallus, 2020). Conformal prediction provides calibrated sets that can support robust decision procedures (Vovk et al., 2005; Angelopoulos & Bates, 2023). We build on the set-based view of uncertainty to study decision invariance: when does every causal benefit vector in an uncertainty set induce the same optimizer? The decision fragility index answers this question through directional widths relative to allocation margins. It thereby provides a geometric stability certificate and identifies decision-relevant directions for predictor training.

B Proofs and Additional Results

B.1 Causal Identification and CPO+

B.1.1 Proof of Lemma 2.2

Fix x and write $p = e_0(x)$, $e = e(x)$, $m_a = \mu_a(x)$, $m_{0,a} = \mu_{0,a}(x)$. Under Assumption 2.1(i)–(iii), $m_{0,a} = \mathbb{E}[Y(a) \mid X = x] = \mathbb{E}[Y \mid A = a, X = x]$, and hence $\tau_0(x) = m_{0,1} - m_{0,0}$.

Taking conditional expectation of the AIPW score in equation 4 gives

$$\mathbb{E}[\Gamma_\eta(O) \mid X = x] = m_1 - m_0 + \frac{p}{e}(m_{0,1} - m_1) - \frac{1-p}{1-e}(m_{0,0} - m_0).$$

Subtracting $\tau_0(x)$ and rearranging,

$$\begin{aligned}&\mathbb{E}[\Gamma_\eta(O) \mid X = x] - \tau_0(x)\\ &\quad = (e - p)\left\{\frac{m_1 - m_{0,1}}{e} + \frac{m_0 - m_{0,0}}{1 - e}\right\}. \end{aligned} \tag{25}$$

If the outcome regressions are correctly specified, $m_a = m_{0,a}$ for $a \in \{0, 1\}$, the right-hand side is zero. Alternatively, if the propensity score is correctly specified, $e = p = e_0(x)$, it is again zero. Therefore $\mathbb{E}[\Gamma_\eta(O) \mid X = x] = \tau_0(x)$ under either nuisance-correctness branch. Finally, $\mathbb{E}[\widetilde{B}_\eta(O) \mid X = x] = r(x) + B(x)\tau_0(x) = b_0(x)$, which proves the result. □

B.1.2 Proof of Proposition 3.2

Let $\widehat{w} = w^*(\widehat{b})$ and $w_b = w^*(b)$. Since $\xi_S(2\widehat{b} - b) \geq (2\widehat{b} - b)^\top \widehat{w}$,

$$\begin{aligned}\ell_{\mathrm{CPO}+}(\widehat{b}, b) &\geq (2\widehat{b} - b)^\top \widehat{w} - 2\widehat{b}^\top w_b + b^\top w_b\\ &= b^\top(w_b - \widehat{w}) + 2\widehat{b}^\top(\widehat{w} - w_b)\\ &\geq \ell_{\mathrm{CPO}}(\widehat{b}, b),\end{aligned}$$

where the final inequality follows from the optimality of $\widehat{w}$ for $\widehat{b}$. This proves (i).

For fixed b, $\xi_S(2\widehat{b} - b)$ is convex in $\widehat{b}$, while $-2\widehat{b}^\top w_b + V^*(b)$ is affine, proving (ii). Finally, $w^*(2\widehat{b} - b) \in \partial\xi_S(2\widehat{b} - b)$, so the affine chain rule gives $2\{w^*(2\widehat{b} - b) - w^*(b)\} \in \partial_{\widehat{b}}\ell_{\mathrm{CPO}+}(\widehat{b}, b)$, which proves (iii). □

B.1.3 Proof of Theorem 3.3

Fix x and write $Z = \widetilde{B}_\eta(O)$ and $\bar{b} = \mathbb{E}[Z \mid X = x]$. By Proposition 3.2,

$$\mathbb{E}[\ell_{\mathrm{CPO}+}(\widehat{b}, Z) \mid X = x] \geq \mathbb{E}[V^*(Z) - Z^\top w^*(\widehat{b}) \mid X = x].$$

Since $V^* = \xi_S$ is convex, Jensen's inequality gives

$$\mathbb{E}[\ell_{\mathrm{CPO}+}(\widehat{b}, Z) \mid X = x] \geq V^*(\bar{b}) - \bar{b}^\top w^*(\widehat{b}) = \ell_{\mathrm{CPO}}(\widehat{b}, \bar{b}).$$

Let $\delta = \bar{b} - b_0(x)$ and $w_0 = w^*\{b_0(x)\}$. Since $V^*(\bar{b}) \geq \bar{b}^\top w_0$,

$$\begin{aligned}\ell_{\mathrm{CPO}}(\widehat{b}, b_0(x)) - \ell_{\mathrm{CPO}}(\widehat{b}, \bar{b}) &\leq \delta^\top\{w^*(\widehat{b}) - w_0\}\\ &\leq \xi_{S-S}(\delta).\end{aligned}$$

Combining the two inequalities proves equation 7. Under either branch of Lemma 2.2, $\bar{b} = b_0(x)$, so the final term vanishes. □

B.2 Decision-Sensitive Uncertainty Geometry

B.2.1 Decision invariance and ellipsoidal examples

Proof of Theorem 4.1. Let $b' = b + e$ with $e \in E$. For any $v \in \mathcal{V} \setminus \{w_0\}$,

$$\begin{aligned}b'^\top(w_0 - v) &= b^\top(w_0 - v) + e^\top(w_0 - v)\\ &= \Delta_v - e^\top(v - w_0)\\ &\geq \Delta_v - h_E(v - w_0).\end{aligned}$$

If $\mathrm{DFI}(b, U) < 1$, then $h_E(v - w_0) < \Delta_v$ for every $v \neq w_0$, hence $b'^\top(w_0 - v) > 0$ for every competitor. Therefore w_0 strictly dominates every other vertex and is the unique optimizer throughout U.

Conversely, suppose $h_E(v - w_0) > \Delta_v$ for some v. By the definition of the support function, there exists $e \in E$ such that $e^\top(v - w_0) > \Delta_v$. Then, for $b' = b + e$, $b'^\top(v - w_0) = -\Delta_v + e^\top(v - w_0) > 0$, so v is strictly better than w_0. □

Ellipsoidal uncertainty. For $\Sigma \succ 0$ and $r > 0$, let

$$U_n = b + E_n, \qquad E_n = \left\{e : e^\top \Sigma^{-1} e \leq \frac{r^2}{n}\right\},$$

write $e = (r/\sqrt{n})\Sigma^{1/2}z$ with $\|z\|_2 \le 1$. Then

$$h_{E_n}(u) = \frac{r}{\sqrt{n}} \sup_{\|z\|_2 \le 1} z^\top \Sigma^{1/2} u = \frac{r}{\sqrt{n}} \sqrt{u^\top \Sigma u}.$$

Taking $u = v - w_0 = -d_v$ yields

$$\mathrm{DFI}(b, U_n) = \frac{r}{\sqrt{n}} \max_{v \neq w_0} \frac{\sqrt{d_v^\top \Sigma d_v}}{\Delta_v}. \tag{26}$$

Equal global uncertainty, different decision fragility. Consider a two-dimensional problem with decision-boundary normal $d = (1, 0)^\top$ and margin $\Delta > 0$. Let

$$\Sigma_A = \begin{pmatrix} 9 & 0 \\ 0 & 1 \end{pmatrix}, \qquad \Sigma_B = \begin{pmatrix} 1 & 0 \\ 0 & 9 \end{pmatrix}.$$

The matrices have identical eigenvalues $\{9, 1\}$ and hence identical trace, determinant, and ellipsoidal volume, but $d^\top \Sigma_A d = 9$, $d^\top \Sigma_B d = 1$. Their DFI contributions are therefore $\frac{3r}{\sqrt{n}\,\Delta}$ and $\frac{r}{\sqrt{n}\,\Delta}$, respectively. Thus equal global uncertainty can induce different decision fragility because of its orientation relative to the decision boundary.

B.2.2 Proof of Theorem 4.3

Fix $v \in \mathcal{A}_0$ and write $Z_{iv} = d_v^\top \psi_x(O_i)$. By Assumption 4.2, Z_{iv} is centered and ν_v^2-sub-Gaussian, so for every $t > 0$,

$$\Pr\left\{ \frac{1}{n} \sum_{i=1}^n Z_{iv} \le -t \right\} \le \exp\left\{ -\frac{nt^2}{2\nu_v^2} \right\}. \tag{27}$$

For a polytope, the tangent cone at w_0 is generated by directions to its adjacent vertices. Hence $w^*(\widehat{b}_n) \neq w_0$ implies $\widehat{b}_n^\top d_v \le 0$ for some $v \in \mathcal{A}_0$. Since $b^\top d_v = \Delta_v$ and, by Assumption 4.2, $d_v^\top(\widehat{b}_n - b) = n^{-1}\sum_i Z_{iv} + d_v^\top r_n$, on the event $|d_v^\top r_n| \le \rho_{v,n}$,

$$\widehat{b}_n^\top d_v \le 0 \quad \Longrightarrow \quad \frac{1}{n}\sum_{i=1}^n Z_{iv} \le -(\Delta_v - \rho_{v,n}).$$

Therefore, using equation 27 and $\Pr\{|d_v^\top r_n| > \rho_{v,n}\} \le \eta_{v,n}$,

$$\Pr\{\widehat{b}_n^\top d_v \le 0\} \le \exp\left\{ -\frac{n(\Delta_v - \rho_{v,n})^2}{2\nu_v^2} \right\} + \eta_{v,n}. \tag{28}$$

Taking a union bound over $v \in \mathcal{A}_0$ and truncating the resulting bound at one proves Theorem 4.3. □

B.2.3 Proof of the regret decomposition

Under the fixed vertex-valued tie-breaking rule, $w^*(\widehat{b}_n) \in \mathcal{V}$. If $w^*(\widehat{b}_n) = v$, then $\ell_{\mathrm{CPO}}(\widehat{b}_n, b) = b^\top(w_0 - v) = \Delta_v$. Partitioning over the selected vertex therefore gives

$$\mathbb{E}[\ell_{\mathrm{CPO}}(\widehat{b}_n, b)] = \sum_{v \neq w_0} \Delta_v \Pr\{w^*(\widehat{b}_n) = v\} \le \Delta_{\max} \Pr\{w^*(\widehat{b}_n) \neq w_0\}. \tag{29}$$

This is exactly equation 12. □

B.2.4 Gaussian specialization

Suppose $\widehat{b}_n - b \sim \mathcal{N}(0, \Sigma/n)$ with $\Sigma \succ 0$, and let Φ denote the standard normal CDF. For any $v \neq w_0$, the event that v weakly beats w_0 is equivalent to $(\widehat{b}_n - b)^\top(v - w_0) \ge \Delta_v$. Since this scalar is Gaussian with variance $d_v^\top \Sigma d_v / n$,

$$\Pr\{\widehat{b}_n^\top (v - w_0) \ge 0\} = \Phi\left(-\frac{\sqrt{n}\,\Delta_v}{\sqrt{d_v^\top \Sigma d_v}} \right). \tag{30}$$

Because any change from w_0 requires crossing at least one adjacent boundary,

$$\Pr\{w^*(\widehat{b}_n) \neq w_0\} \le \min\left\{ 1, \sum_{v \in \mathcal{A}_0} \Phi\left(-\frac{\sqrt{n}\,\Delta_v}{\sqrt{d_v^\top \Sigma d_v}} \right) \right\}.$$

Moreover, $\Pr\{w^*(\widehat{b}_n) = v\} \le \Pr\{\widehat{b}_n^\top(v - w_0) \ge 0\}$ for every $v \ne w_0$. Combining this observation with equation 12 yields

$$\mathbb{E}[\ell_{\mathrm{CPO}}(\widehat{b}_n, b)] \le \sum_{v \ne w_0} \Delta_v \Phi\left(-\frac{\sqrt{n}\,\Delta_v}{\sqrt{d_v^\top \Sigma d_v}}\right). \tag{31}$$

B.3 Proofs for Section 5

B.3.1 Proof of Theorem 5.2

Fix x and write $b = b_0(x)$ and $Z = \widetilde{B}_\eta(O)$, conditioning throughout on $X = x$. Under Assumption 5.1, the outcome regressions are correctly specified. Let $\varepsilon_a = Y(a) - \mu_{0,a}(X)$ for $a \in \{0, 1\}$. Then $\varepsilon_a \mid X = x \overset{d}{=} -\varepsilon_a \mid X = x$.

Since $\mu_a = \mu_{0,a}$, the centered AIPW score is

$$\Gamma_\eta(O) - \tau_0(X) = \frac{A\varepsilon_1}{e(X)} - \frac{(1 - A)\varepsilon_0}{1 - e(X)}.$$

By conditional ignorability, A is independent of the potential-outcome residuals given X. Hence, conditional on $X = x$, the two mixture components are symmetric about zero, so $\Gamma_\eta(O) - \tau_0(x) \overset{d}{=} -\{\Gamma_\eta(O) - \tau_0(x)\}$. Using equation 5,

$$Z - b \overset{d}{=} -(Z - b), \qquad \text{equivalently} \qquad 2b - Z \overset{d}{=} Z. \tag{32}$$

Fisher consistency of CPO+. Define the conditional CPO+ risk $Q_x(a) = \mathbb{E}[\ell_{\mathrm{CPO}+}(a, Z) \mid X = x]$. By Proposition 3.2, $Q_x(a)$ is convex in a, and a valid subgradient is

$$2\,\mathbb{E}[w^*(2a - Z) - w^*(Z) \mid X = x]\,.$$

Evaluating at $a = b$ and using equation 32, $\mathbb{E}[w^*(2b - Z) \mid X = x] = \mathbb{E}[w^*(Z) \mid X = x]$. Therefore $0 \in \partial Q_x(b)$. Since Q_x is convex,

$$b_0(x) \in \arg\min_{a \in \mathbb{R}^d} \mathbb{E}\Big[\ell_{\mathrm{CPO}+}\{a, \widetilde{B}_\eta(O)\} \mid X = x\Big]. \tag{33}$$

Thus CPO+ is Fisher consistent for the causal benefit coefficient.

Preservation under DFI regularization. By Proposition 5.3, $\Omega_\rho(f) = D_\rho(f) + C_\rho$, where $D_\rho(f) \ge 0$, $D_\rho(b_0) = 0$, and C_ρ does not depend on f. Therefore, for every $\lambda \ge 0$,

$$R_\eta^+(f) + \lambda\Omega_\rho(f) - \{R_\eta^+(b_0) + \lambda\Omega_\rho(b_0)\} = \{R_\eta^+(f) - R_\eta^+(b_0)\} + \lambda D_\rho(f) \ge 0.$$

Hence b_0 also minimizes the regularized DC-CPO population objective. If the conditional minimizer is unique, then $f_\lambda^*(x) = b_0(x)$ almost surely, and therefore $w^*\{f_\lambda^*(x)\} = w^*\{b_0(x)\}$ almost surely. This proves the theorem. □

B.3.2 Directional-Error Decomposition

Recall that $\Omega_\rho(f)$ is the population boundary penalty in equation 19. Define

$$D_\rho(f) = \mathbb{E}\left[\frac{1}{m_\rho(X)} \sum_{v \in \mathcal{A}_\rho(X)} \omega_v(X)\big[d_v(X)^\top\{f(X) - b_0(X)\}\big]^2\right], \tag{34}$$

$$C_\rho = \mathbb{E}\left[\frac{1}{m_\rho(X)} \sum_{v \in \mathcal{A}_\rho(X)} \omega_v(X)\,\mathrm{Var}\{d_v(X)^\top \widetilde{B}_\eta(O) \mid X\}\right]. \tag{35}$$

Fix $X = x$ and $v \in \mathcal{A}_\rho(x)$, and write $Z_v = d_v(x)^\top\{\widetilde{B}_\eta(O) - b_0(x)\}$. If either the outcome regression or the propensity score is correctly specified, Lemma 2.2 gives $\mathbb{E}[Z_v \mid X = x] = 0$. Therefore,

$$\mathbb{E}\Big[\big[d_v(x)^\top\{f(x) - \widetilde{B}_\eta(O)\}\big]^2 \Big| X = x\Big]$$

$$= \left[d_v(x)^\top \{f(x) - b_0(x)\}\right]^2 + \mathrm{Var}\{d_v(x)^\top \widetilde{B}_\eta(O) \mid X = x\}. \tag{36}$$

The cross term vanishes because $\mathbb{E}[Z_v \mid X = x] = 0$.

Multiplying equation 36 by $\omega_v(x)$, averaging over $v \in \mathcal{A}_\rho(x)$, and then taking expectation over X yields $\Omega_\rho(f) = D_\rho(f) + C_\rho$. Since C_ρ is independent of f, minimizing $\Omega_\rho(f)$ is equivalent at the population level to minimizing the directional prediction error $D_\rho(f)$. □

B.3.3 Proof of Theorem 5.4

We separate the proof into two approximation steps. First, the actual training objective differs from an oracle empirical objective because the pseudo-benefits and decision geometry are estimated. Second, the oracle empirical objective differs from its population counterpart through ordinary sampling error. Combining these two deviations with empirical near-optimality yields the population DC-CPO bound, which is then transferred to true causal regret using Theorem 3.3.

Let $\widetilde{L}_{n,\lambda}(f)$ denote the empirical DC-CPO objective computed using $\widetilde{B}_\eta$ and the fixed population decision geometry, so that $L_\lambda(f) = \mathbb{E}[\widetilde{L}_{n,\lambda}(f)] = R^+_\eta(f) + \lambda\Omega_\rho(f)$.

Define $\Delta_{n,\eta}$ as the uniform error from replacing $\widetilde{B}_\eta$ by the estimated pseudo-benefits, and $\Delta_{n,\mathrm{geom}}$ as the uniform error from replacing the population decision geometry by the estimated geometry. By the triangle inequality,

$$\sup_{f\in\mathcal{F}} \left|\widehat{L}_{n,\lambda}(f) - \widetilde{L}_{n,\lambda}(f)\right| \le \Delta_{n,\eta} + \Delta_{n,\mathrm{geom}}. \tag{37}$$

Let $\mathcal{G}_\lambda$ be the fixed-geometry population loss class. If $0 \le g \le M_\lambda$ for every $g \in \mathcal{G}_\lambda$, standard symmetrization and concentration give, with probability at least $1-\delta$,

$$\sup_{f\in\mathcal{F}} \left|\widetilde{L}_{n,\lambda}(f) - L_\lambda(f)\right| \le 2\mathfrak{R}_n(\mathcal{G}_\lambda) + M_\lambda\sqrt{\frac{\log(2/\delta)}{2n}}. \tag{38}$$

Combining the two bounds yields

$$\sup_{f\in\mathcal{F}} \left|\widehat{L}_{n,\lambda}(f) - L_\lambda(f)\right| \le \Gamma_n, \tag{39}$$

where $\Gamma_n = 2\mathfrak{R}_n(\mathcal{G}_\lambda) + M_\lambda\sqrt{\log(2/\delta)/(2n)} + \Delta_{n,\eta} + \Delta_{n,\mathrm{geom}}$.

On the event equation 39, empirical $\varepsilon_{\mathrm{opt}}$-optimality of $\widehat{f}_\lambda$ gives

$$\begin{aligned} L_\lambda(\widehat{f}_\lambda) &\le \widehat{L}_{n,\lambda}(\widehat{f}_\lambda) + \Gamma_n \\ &\le \inf_{f\in\mathcal{F}} \widehat{L}_{n,\lambda}(f) + \Gamma_n + \varepsilon_{\mathrm{opt}} \\ &\le \inf_{f\in\mathcal{F}} L_\lambda(f) + 2\Gamma_n + \varepsilon_{\mathrm{opt}}. \end{aligned} \tag{40}$$

Since $\Omega_\rho(f) \ge 0$, $R^+_\eta(\widehat{f}_\lambda) \le L_\lambda(\widehat{f}_\lambda)$. If either the outcome regression or the propensity score is correctly specified, Theorem 3.3 further gives $R_{\mathrm{CPO}}(\widehat{f}_\lambda) \le R^+_\eta(\widehat{f}_\lambda)$. Therefore,

$$R_{\mathrm{CPO}}(\widehat{f}_\lambda) \le \inf_{f\in\mathcal{F}} \left\{R^+_\eta(f) + \lambda\Omega_\rho(f)\right\} + 2\Gamma_n + \varepsilon_{\mathrm{opt}}, \tag{41}$$

which is equation 24. □

C Experimental Details

C.1 Baselines and Metrics

Shared learning setup. All learned methods use the same observations, cross-fitted doubly robust pseudo-outcomes, feature representation, intervention costs, and downstream feasible set. For observation i, the pseudo-outcome is

$$\widetilde{\Gamma}_i = \widehat{\mu}_1(X_i) - \widehat{\mu}_0(X_i) + \frac{A_i\{Y_i - \widehat{\mu}_1(X_i)\}}{\widehat{e}(X_i)} - \frac{(1-A_i)\{Y_i - \widehat{\mu}_0(X_i)\}}{1-\widehat{e}(X_i)}, \tag{42}$$

where the nuisance estimates are obtained without using observation i. The common treatment-effect predictor is $\widehat{\tau}_\theta(X) = F(X)^\top \theta$. For a decision batch k, predicted and pseudo-benefit vectors are $\widehat{b}_k = F_k\theta - c_k$ and $\widetilde{b}_k = \widetilde{\Gamma}_k - c_k$, respectively. Each batch is treated as one complete allocation problem, with benefit vectors formed by stacking unit-level treatment effects or DR pseudo-outcomes and subtracting intervention costs. The conditional-mean double-robustness property carries over componentwise.

Prediction-focused baselines. MSE-DR and MAE-DR minimize

$$\widehat{\theta}_{\mathrm{MSE}} \in \arg\min_\theta \left[\frac{1}{n} \sum_{i=1}^{n} \{F(X_i)^\top \theta - \widetilde{\Gamma}_i\}^2 + \lambda_0 \|\theta_{-0}\|_2^2 \right] \tag{43}$$

and

$$\widehat{\theta}_{\mathrm{MAE}} \in \arg\min_\theta \left[\frac{1}{n} \sum_{i=1}^{n} |F(X_i)^\top \theta - \widetilde{\Gamma}_i| + \lambda_0 \|\theta_{-0}\|_2^2 \right], \tag{44}$$

respectively, where θ_{-0} excludes the intercept. Each baseline selects its own $\lambda_0 \in \{10^{-4}, 10^{-3}, 10^{-2}\}$ using validation pseudo-regret.

CPO+ and DC-CPO. CPO+ minimizes the average batch loss

$$\begin{aligned} \widehat{R}^+(\theta) = \frac{1}{K_{\mathrm{fit}}} \sum_{k=1}^{K_{\mathrm{fit}}} \Big[& \max_{w \in S} (2\widehat{b}_k - \widetilde{b}_k)^\top w - 2\widehat{b}_k^\top w^*(\widetilde{b}_k) \\ & + \widetilde{b}_k^\top w^*(\widetilde{b}_k) \Big], \end{aligned} \tag{45}$$

where K_{fit} is the number of fitting batches. DC-CPO initializes from CPO+ and alternates between updating the estimated boundary geometry and minimizing $\widehat{R}^+(\theta) + \lambda \widehat{\Omega}^{(t)}_{\mathrm{bdry}}(\theta)$, using the boundary regularizer in equation 16. In this experiment, each indexed observation in the batch-level objective corresponds to one complete allocation problem. Neither CPO+ nor DC-CPO includes an additional ridge penalty.

Oracle. The oracle uses the true treatment effects to form $b_k = \tau_k - c_k$ and solves the same allocation problem. It provides the reference value $V^*(b_k)$ for regret evaluation and is not a learnable comparator.

Evaluation metrics. Let n_{eval} denote the number of evaluation individuals and K_{eval} the number of evaluation batches. We report PEHE in its squared form: $\mathrm{PEHE} = \frac{1}{n_{\mathrm{eval}}} \sum_{i=1}^{n_{\mathrm{eval}}} \{\widehat{\tau}(X_i) - \tau_0(X_i)\}^2$. This metric measures treatment-effect estimation error without taking the square root. Average downstream regret is

$$R = \frac{1}{K_{\mathrm{eval}}} \sum_{k=1}^{K_{\mathrm{eval}}} \left[V^*(b_k) - b_k^\top w^*(\widehat{b}_k) \right], \tag{46}$$

which measures the loss in true total net benefit caused by using the predicted allocation.

To quantify decision instability, we bootstrap the estimation sample and refit the learning pipeline. Let $\widehat{b}_k^{*(r)}$ be the benefit prediction for the fixed evaluation batch k from bootstrap replicate r. We estimate boundary-crossing probability by

$$\widehat{P}_{\mathrm{cross}} = \frac{1}{K_{\mathrm{eval}} M_{\mathrm{boot}}} \sum_{k=1}^{K_{\mathrm{eval}}} \sum_{r=1}^{M_{\mathrm{boot}}} \mathbf{1} \left\{ w^*(\widehat{b}_k^{*(r)}) \neq w^*(\widehat{b}_k) \right\}, \qquad M_{\mathrm{boot}} = 40. \tag{47}$$

A crossing is recorded whenever any component of the allocation vector changes. This metric measures instability around the fitted allocation; Theorem 4.3 uses the oracle allocation as its reference.

Finally, define

$$\overline{R}_{\mathrm{best}} = \min_{m \in \{\text{MSE-DR}, \text{MAE-DR}\}} \overline{R}_m, \qquad R\text{-gap} = 100\% \times \frac{\overline{R}_{\mathrm{best}} - \overline{R}_{\text{DC-CPO}}}{\overline{R}_{\mathrm{best}}}, \tag{48}$$

where bars denote averages across seeds for the same design and evaluation split. The best baseline is selected after averaging each baseline's regret across seeds. A positive R-gap indicates improvement over the best prediction-focused baseline; the percentage is defined when $\overline{R}_{\text{best}} > 0$. Lower values are better for PEHE, regret, and P_{cross}, whereas higher values are better for R-gap.

C.2 Experimental Settings

Synthetic data-generating process. We generate $p = 10$ covariates according to $X \sim \mathcal{N}(0, \Sigma_X)$, $(\Sigma_X)_{jk} = 0.5^{|j-k|}$. Treatment assignment follows $e_0(X) = \sigma(\gamma_e X^\top \alpha)$, $A \mid X \sim \text{Bernoulli}\{e_0(X)\}$, where $\sigma(u) = (1 + \exp(-u))^{-1}$, $\alpha = \frac{(1,1,-1,0.5,0,\ldots,0)^\top}{\|(1,1,-1,0.5,0,\ldots,0)\|_2}$, and $\gamma_e = 1$ in the main setting. The untreated response surface is $m_0(X) = X_1 + 0.5X_2 + X_3^2$. We define two treatment-effect functions:

$$\tau_{\text{lin}}(X) = 1 + 0.5X_1 - 0.5X_2 + 0.3X_3X_4, \tag{49}$$

$$\tau_{\text{nonlin}}(X) = 1 + 0.5X_1 - 0.5X_2 + 0.9\mathbf{1}\{X_3 > 0\} - 0.6\cos(X_4) + 0.45\sin(X_2X_3). \tag{50}$$

The interpolation $\tau_s(X) = (1-s)\tau_{\text{lin}}(X) + s\tau_{\text{nonlin}}(X)$, $s \in \{0, 0.25, 0.50, 0.75, 1\}$, controls misspecification relative to the common predictor class. Specifically, τ_{lin} is linear in the supplied features, including X_3X_4, whereas τ_{nonlin} contains components outside their span. The main comparison uses $s = 1$. Potential and observed outcomes are generated as $Y(0) = m_0(X) + \epsilon_0$, $Y(1) = m_0(X) + \tau_s(X) + \epsilon_1$, $Y = AY(1) + (1-A)Y(0)$, where ϵ_0, ϵ_1 are independent standard Gaussian errors, independent of covariates and treatment assignment.

Downstream allocation. Intervention costs are drawn independently as $c_i \sim \text{Unif}(0.30, 1.20)$, and true net benefits are $b_i = \tau_s(X_i) - c_i$. Individuals are randomly partitioned into batches of 16. Within each batch, sorting by X_1 produces two groups of eight individuals, G_0 and G_1. The allocation problem is

$$\begin{aligned} \max_{w \in \{0,1\}^{16}} \quad & \sum_{i=1}^{16} b_i w_i \\ \text{subject to} \quad & \sum_{i=1}^{16} w_i \le 5, \\ & \sum_{i \in G_g} w_i \le 4, \qquad g \in \{0,1\}. \end{aligned} \tag{51}$$

The total capacity limits the number of interventions, while the group caps prevent allocating all five interventions to one group. Costs enter the objective as deductions from treatment benefits; the resource constraints limit intervention counts. This feasible set contains 6,773 allocations. A deterministic exact optimization oracle is used for all methods.

Real-covariate datasets. We use three empirical covariate designs to examine performance beyond Gaussian inputs. Original treatment assignments and outcomes are not used as experimental labels; treatments and outcomes are generated under the common causal DGP.

Twins. We use the same-sex twin-pair covariate file from the Twins benchmark, treating each twin pair as one decision unit. We retain ten parental and prenatal covariates covering maternal and paternal age and education, live-birth and total-birth order, prenatal-care initiation, the number of prenatal visits, marital status, and maternal smoking. Pair identifiers are used to remove duplicate pairs and are excluded from the predictors, together with birth weights and mortality outcomes. Distinct pairs with identical retained covariates remain separate decision units. The resulting task represents allocating limited family-level prenatal or follow-up support, with simulated treatment benefits rather than effects estimated from the original birth-weight or mortality outcomes.

NHANES. We use the 2017–2018 demographic, body-measurement, and blood-pressure files, merged by participant identifier, and retain adults aged at least 20 years, yielding 5,265 individuals. The ten covariates are age, body mass index, mean systolic and diastolic blood pressure across available readings, waist circumference, pulse, income-to-poverty ratio, household size, sex, and an indicator for some college education or more. Invalid or unavailable sex and education codes are

treated as missing. This task represents allocating a limited preventive or health-support intervention. We evaluate methods on the unweighted empirical covariate distribution.

ACTG-175. The source dataset contains 2,139 participants from an HIV clinical trial. We retain ten pretreatment variables: age, weight, baseline CD4 and CD8 counts, Karnofsky score, days of prior antiretroviral therapy, hemophilia, homosexual activity, intravenous drug use, and prior non-zidovudine antiretroviral therapy. Original treatment assignments and follow-up measurements are excluded. The resulting allocation task represents prioritizing patients for a capacity-limited intervention or intensive follow-up, with simulated benefits rather than estimated effects of the original trial treatments.

Semi-synthetic construction and preprocessing. For each dataset and seed, we sample 1,920 decision units without replacement and partition them into disjoint training, validation, and test sets of sizes 1,536, 192, and 192. Twins uses unique twin-pair identifiers; exact duplicate source records are removed from ACTG-175; NHANES uses unique participant identifiers. Missing covariates are imputed using training-set medians, and covariates are standardized using training-set means and standard deviations. These transformations are then applied to validation and test covariates. Using the resulting ten-dimensional covariates, treatments and potential outcomes are generated by the synthetic DGP with $s = 1$ and $\gamma_e = 1$. Intervention costs, batch construction, and allocation constraints are identical to those in the synthetic experiment.

Learning and implementation. The synthetic experiment uses the same 1,536/192/192 split sizes, with independently generated observations in each split. All methods share the feature representation $F(X) = [1, X_1, \ldots, X_{10}, X_1^2, \ldots, X_{10}^2, X_3X_4]^\top$, with non-intercept features standardized on the fitting sample. Nuisance functions are estimated by five-fold cross-fitting, using ridge outcome regressions within treatment arms and logistic propensity regression on the ten covariates. Both nuisance learners standardize their inputs within the nuisance fitting sample; the ridge parameter and logistic inverse regularization parameter are each set to 1. Estimated propensities are clipped to $[10^{-6}, 1 - 10^{-6}]$.

For baseline hyperparameter selection, validation pseudo-outcomes are constructed from nuisance models fitted on the training set. Each candidate is scored by

$$\widehat{R}_{\mathrm{val}}^{\mathrm{pseudo}}(\theta) = \frac{1}{K_{\mathrm{val}}} \sum_{k=1}^{K_{\mathrm{val}}} \left[V^*(\widetilde{b}_k) - \widetilde{b}_k^\top w^*(\widehat{b}_k) \right]. \tag{52}$$

True treatment effects are used only for evaluation. After selection, training and validation observations are combined into a 1,728-observation estimation sample, and the feature transformation, nuisance estimates, and final predictors are refitted. Within-sample evaluation uses this combined sample; out-of-sample evaluation uses the independent 192-observation test set.

DC-CPO uses $\lambda = 0.3$, $\rho = 1$, $\varepsilon = 0.05$, $W = 400$, and $T = 4$ outer updates. Each update retains adjacent competitors with predicted margins in $[0, \rho]$ and assigns weights $\omega_{kv}^{(t)} = \min\{(\widehat{\Delta}_{kv}^{(t)} + \varepsilon)^{-2}, W\}$. The penalty uses raw DR pseudo-outcomes and is averaged over retained directions within each batch and then over batches. An empty active set contributes zero. MAE-DR is solved through an exact absolute-loss epigraph formulation, and CPO+/DC-CPO through LP/QP cutting-plane separation with the exact allocation oracle.

The evaluation protocol uses 50 seeds, $0, \ldots, 49$, shared across methods. For each seed, 40 bootstrap samples are drawn from the combined estimation sample. Each bootstrap refits nuisance functions, feature scaling, and predictors while keeping the selected hyperparameters fixed. Repeated copies of the same source observation remain in the same cross-fitting fold. Evaluation batches, covariates, and costs remain fixed across bootstrap refits. Metrics are summarized by their means across seeds; standard errors are the sample standard deviations of seed-level metrics divided by $\sqrt{50}$.

C.3 Graphical Comparison of Experimental Results

Figure 4 visualizes the Out and Within results in Table 1, together with stability gains under tighter constraints. On ACTG-175, DC-CPO achieves lower regret than CPO+ despite higher PEHE, illustrating the distinction between prediction accuracy and decision quality.

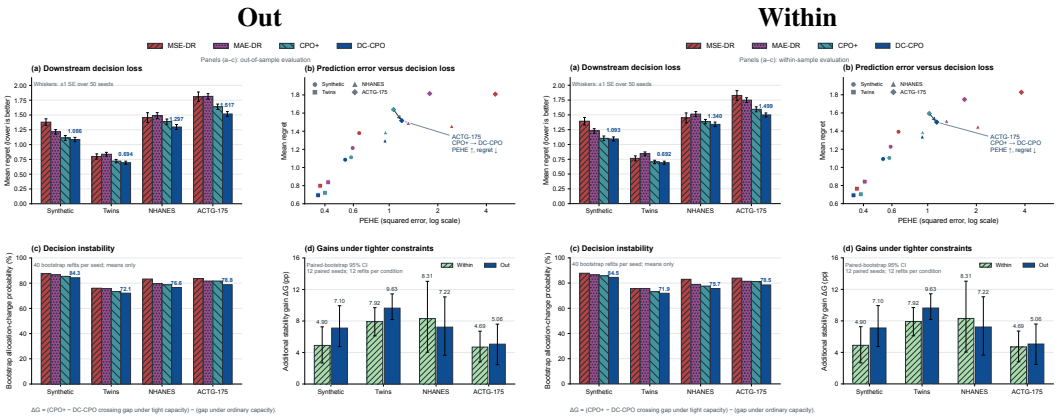


Figure 4: Experimental results for Out (left) and Within (right). In each image, panels (a)–(c) show mean regret, PEHE versus regret (logarithmic PEHE axis), and bootstrap allocation-change probability. Results are averaged over 50 seeds, with ± 1 standard-error bars in (a) and 40 bootstrap refits per seed in (c). Panel (d), repeated in both images, shows the additional stability gain under tighter constraints in percentage points, with paired-bootstrap 95% confidence intervals from 12 paired seeds and 12 refits per condition (Table 3).

C.4 Oracle-Normalized Regret

To express decision losses relative to the oracle net benefit, we report oracle-normalized regret on the Out split: $\mathrm{NR}_m = 100 \times \frac{\overline{R}_m}{\overline{V^\star}}$, $\overline{V^\star} > 0$, where $\overline{R}_m$ and $\overline{V^\star}$ denote the mean regret of method m and the mean oracle net benefit, respectively, using the same evaluation batches and seed aggregation as the main results. We compute the ratio of means rather than the mean of batch-wise ratios.

Table 2: Oracle-normalized regret on the Out split. Lower NR is better. Reduction denotes CPO+ NR minus DC-CPO NR in percentage points (pp). All entries, including reductions, are computed from unrounded results and then rounded for display.

Dataset	CPO+ NR (%)	DC-CPO NR (%)	Reduction (pp)
Synthetic	18.55	18.12	0.43
Twins	11.00	10.58	0.41
NHANES	21.54	20.12	1.41
ACTG-175	27.63	25.58	2.05

Table 2 shows that DC-CPO reduces oracle-normalized regret relative to CPO+ across all four datasets. The reductions range from 0.41 to 2.05 percentage points, corresponding to additional net benefits of 0.41%–2.05% of the mean oracle net benefit, with larger gains on NHANES and ACTG-175.

C.5 Sensitivity and Ablation

Sensitivity to the regularization strength. We examine the effect of λ on the synthetic DGP under the main misspecification setting $s = 1$, using the candidate grid $\Lambda = \{0, 0.001, 0.003, 0.01, 0.03, 0.1, 0.3, 1, 3\}$. This diagnostic experiment uses 800 training observations and 400 validation observations per seed, separately from the sample-size configuration of the main comparison. We repeat the experiment over 50 random seeds and use 40 bootstrap refits per seed and candidate to estimate boundary-crossing probability. All candidates use the same data within each seed. Model fitting uses the training sample, and the reported sensitivity metrics are evaluated on the validation sample.

We select λ by minimizing validation DR pseudo-regret averaged across seeds, which yields $\lambda = 0.3$. The known treatment effects in the synthetic DGP allow us to additionally report PEHE and true regret as diagnostic metrics; these quantities are not used for selecting λ. Boundary-crossing probability is estimated as the frequency with which a bootstrap refit changes the allocation relative to the corresponding model fitted on the original training sample.

Figure 3 displays candidates up to $\lambda = 1$. Within this range, PEHE decreases as regularization strengthens, whereas regret and boundary-crossing probability reach their minima at $\lambda = 0.3$ and

subsequently increase slightly. Thus, stronger regularization can improve treatment-effect estimation without improving either allocation quality or stability.

Ablation of decision-sensitive weighting. To isolate the contribution of boundary weighting, we construct DC-CPO$^-$ by replacing every active-boundary weight with $\omega_{jv}^{(t)} = 1$. The ablated method retains the projected-residual penalty, active-boundary selection, and iterative geometry updates, with the remaining experimental settings unchanged. This comparison therefore assesses the contribution of nonuniform weighting within the same boundary-based regularization framework.

Across the synthetic DGP, Twins, NHANES, and ACTG-175, Table 1 shows that DC-CPO yields lower mean regret and boundary-crossing probability than DC-CPO$^-$ in both Within and Out evaluations. These results support the contribution of decision-sensitive weighting to allocation quality and stability beyond the use of an unweighted boundary penalty.

Table 3: Stability gains of DC-CPO over CPO+ under ordinary and tight resource constraints. All values are expressed in percentage points. Positive ΔG indicates an increased stability advantage under tighter constraints. Intervals are paired-bootstrap 95% confidence intervals. All quantities are calculated before rounding.

Dataset	Evaluation	G_{main}	G_{tight}	ΔG [95% CI]
Synthetic DGP	Within	0.96	5.86	4.90 [2.67, 7.26]
	Out	1.28	8.38	7.10 [4.74, 9.94]
Twins	Within	0.44	8.35	7.92 [6.11, 9.69]
	Out	0.32	9.94	9.63 [8.18, 11.44]
NHANES	Within	2.44	10.75	8.31 [4.04, 13.04]
	Out	2.81	10.03	7.22 [3.65, 11.06]
ACTG-175	Within	3.20	7.90	4.69 [2.81, 6.71]
	Out	4.01	9.07	5.06 [2.46, 7.60]

Tighter resource constraints. We investigate whether tighter resource constraints amplify the stability advantage of DC-CPO over CPO+. We reduce the total allocation capacity from five to two and the capacity of each group from four to one, keeping the batch size at 16. The comparison uses 12 paired random seeds and 12 bootstrap refits per seed under each resource condition. Both resource conditions use the same 12 seeds and bootstrap protocol.

For each resource condition c, we define the stability gain as $G_c = P_{\text{cross},c}^{\text{CPO+}} - P_{\text{cross},c}^{\text{DC-CPO}}$, $\Delta G = G_{\text{tight}} - G_{\text{main}}$. Positive G_c indicates lower crossing probability for DC-CPO, and positive ΔG indicates that this advantage increases under tighter constraints. We report gains in percentage points, together with paired-bootstrap 95% confidence intervals for ΔG.

Table 3 shows that the stability advantage of DC-CPO increases under tighter resource constraints across the synthetic DGP and all three semi-synthetic datasets, in both Within and Out evaluations. For Out evaluation, the increases are 7.10 percentage points on the synthetic DGP, 9.63 on Twins, 7.22 on NHANES, and 5.06 on ACTG-175. All eight paired-bootstrap confidence intervals lie above zero. In these experiments, tighter constraints increase the stability advantage of DC-CPO. The intervals are pointwise.

C.6 Alternative Feature Representations

We examine whether the improvements from decision calibration also hold under alternative feature representations. On Twins, NHANES, and ACTG-175, we compare MSE-DR, CPO+, and DC-CPO using forest-derived and kernel features. The forest representation consists of standardized predictions from 32 trees with a minimum leaf size of 20, augmented with an intercept. The kernel representation appends 64 standardized Nyström-RBF features to the 22-dimensional polynomial representation used in the main experiments. Within each representation, all three methods share the same features, which remain fixed during output-layer optimization.

We retain the 1,536/192/192 training/validation/test split and the cross-fitted DR targets used in the main experiments. The MSE-DR ridge parameter is selected by validation pseudo-regret, after which the nuisance models, feature representation, and predictors are refitted on the combined training

and validation sample. DC-CPO uses $\lambda = 0.3$, $\rho = 1$, $\varepsilon = 0.05$, $W = 400$, and $T = 4$ outer updates. The comparison uses 12 paired seeds and 12 bootstrap refits per seed. Each bootstrap refits the nuisance models, feature representation, and output layers, while keeping the selected hyperparameters fixed. Data splits, evaluation batches, and bootstrap resampling indices are shared across the two representations within each seed.

Table 4: Out-of-sample performance under alternative feature representations, averaged over 12 paired seeds. Crossing denotes bootstrap allocation-change probability, estimated using 12 refits per seed. Lower values are better for all three metrics. Bold values indicate the lowest mean within each dataset–representation combination.

Dataset	Representation	Method	PEHE	Regret	Crossing
Twins	Forest	MSE-DR	0.945	1.273	0.943
		CPO+	0.906	1.321	0.948
		DC-CPO	**0.657**	**1.071**	**0.922**
	Nyström-RBF	MSE-DR	0.942	1.557	0.937
		CPO+	1.075	1.541	0.941
		DC-CPO	**0.671**	**1.017**	**0.872**
NHANES	Forest	MSE-DR	2.835	1.468	0.963
		CPO+	2.371	1.392	0.956
		DC-CPO	**1.827**	**1.254**	**0.944**
	Nyström-RBF	MSE-DR	4.652	2.207	0.968
		CPO+	2.328	1.855	0.960
		DC-CPO	**1.652**	**1.343**	**0.907**
ACTG-175	Forest	MSE-DR	6.469	1.700	0.973
		CPO+	2.905	1.644	0.967
		DC-CPO	**2.547**	**1.492**	**0.958**
	Nyström-RBF	MSE-DR	16.444	2.543	0.976
		CPO+	4.268	2.209	0.969
		DC-CPO	**2.510**	**1.704**	**0.933**

Table 4 shows that DC-CPO achieves lower mean regret and crossing probability than both MSE-DR and CPO+ in all six dataset–representation combinations. For both regret and crossing probability, all six pointwise paired-bootstrap 95% confidence intervals for CPO+ minus DC-CPO lie above zero. These results extend the observed improvements in allocation quality and stability to the forest-derived and kernel representations considered here.