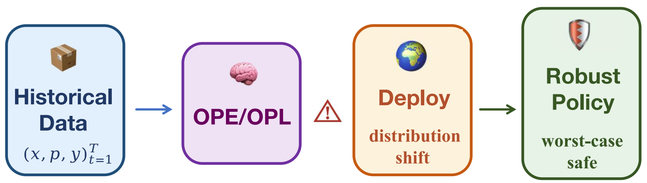
Doubly Robust Distributionally Robust Offline Contextual Pricing

Min Xu[1], Xinyi Yin[1], Yunfan Zhang[2], Yuxuan Han[2], Houcai Shen[1], Caihua Chen[1,*] [1]Nanjing University [2]New York University

ICML International Conference On Machine Learning

1. Introduction



Prior DRO (Discrete)

Inapplicable to continuous prices

DRO-IPW (Leung et al. 2025)

Baseline with theoretical guarantee

Sensitive to propensity estimation; slower regret convergence

Algorithm	Task	Key Guarantee
LDR²O²PE-CP	Evaluation	$\sqrt{Th}$ -consistent & semipar. efficient
CDR²O²PL-CP	Learning	Regret $\widetilde{O}_p(T^{-s/(2s+1)})$, $s = 1,2$

Ours - OPE

Only 3 regressions, semiparameter efficient

Ours - OPL

Sharper regret, requires slower nuisance estimation rates

PRICING MODEL

Valuation: $v(X, Z) = f(X) + Z$

Purchase: $Y = 1[Z \geq P - f(X)]$ Revenue: $R = P \cdot Y$

Robust objective: $R_\delta(\pi) \triangleq \inf_{\mathbf{P} \in \mathcal{U}_{\mathbf{P}_0}(\delta)} \mathbb{E}_{\mathbf{P}}[R(\pi)]$

$R(\pi) \triangleq \mathbb{E}_{\mathbf{P}_0}[r(X, \pi(X), Z)]$

KL uncertainty set: $\mathcal{U}(\delta) = \{P : D_{KL}(P \| \mathbf{P}_0) \leq \delta\}$

DOUBLY ROBUST ESTIMATOR

$$\widehat{W}_T(\pi,\alpha) = \widehat{m}_0(x, \pi(x); \alpha) + \frac{K((p-\pi(x))/h)}{h} \hat{\pi}_0(p|x) \cdot (e^{-py/\alpha} - \widehat{m}_0(x, p; \alpha))$$

2. Method

KEY INSIGHT

Discountinuous reveune $R = P \cdot 1[]$, but outcome regresssion $m_j(x, p; \alpha) = \mathbb{E}[(PY)^j \exp(-PY/\alpha) \mid X = x, P = p]$ is **smooth** via s-differentiability of nose CDF F_Z. Kernel smoothing + localization unlocks near-optimal rates.

DUAL REFORMULATION (STRONG DUALITY)

$R_\delta(\pi) = \sup_{\alpha \geq 0} -\alpha \log W(\pi,\alpha) - \alpha\delta$

$W(\pi,\alpha) = \mathbb{E}[\exp(-r(X,\pi(X), Z)/\alpha)]$.

Reduce infinite-dim problem to scalar α.

LDR2O2PE-CP — Evaluation (3 REGRESSIONS ONLY)

1. **Cross-fit Propensity** $\hat{\pi}_0(p|x)$
2. **Localize at** $\widehat{\alpha}_{\text{init}}$ — Initial IPW estimate ⟶ fit $\widehat{m}_j$ once at this single pont
3. **Solve DR Moment Eq. (Newton-Raphson)** ⟶ output $\widehat{\alpha}$, $\widehat{R}_\delta$

CDR2O2PL-CP — Learning

1. **Continunm Regression via GRF Weights** $\widehat{m}_0 = \sum_t \widehat{\omega}_t(x, p) \exp(-p_t y_t/\alpha)$
2. **Alternating Optimization**

Update $\widehat{\alpha} \leftarrow \operatorname{argmax}_{\alpha \geq 0} -\alpha \log \widehat{W}_T(\hat{\pi},\alpha) - \alpha\delta$

then $\widehat{\pi} \leftarrow \operatorname{argmin}_{\pi \in \Pi} \log \widehat{W}_T(\pi,\widehat{\alpha})$

3. Main Results

OPE CONVERGENCE

$\sqrt{Th}$ - consistent

Neyman Orthogonality

+ semipar. efficient

OPL REGRET

$\widetilde{O}_p(T^{-s/(2s+1)})$

$h = \Theta(T^{-1/(2s+1)})$

DRO regret improvement (s=2) $\widetilde{O}_p(T^{-1/4}) \longrightarrow \widetilde{O}_p(T^{-2/5})$

Method	Action	Robust	Regret
Kitagawa & Athey'18–21	Binary	✗	$O_p(T^{-1/2})$
Dudik, Swaminathan...	Discrete	✗	$O_p(T^{-1/2})$
Kallus, Colangelo, Ai...	Cts.	✗	$O_p(T^{-s/(2s+1)})$
Kallus,Sinian...	Discrete	✓	$O_p(T^{-1/2})$
Leung'25 (DRO-IPW)	Cts.	✓	$O_p(T^{-1/4})$
Ours (DRO-DR)	Cts.	✓	$\widetilde{O}_p(T^{-s/(2s+1)})$

THM 1 — OPE ASYMPTOTIC NORMALITY

Under product rate condition $\varrho_\pi + \varrho_\alpha \wedge \varrho_m \geq (1 - \gamma_h)/2$, and $h = o(T^{-1/(1+2s)})$:

$$\sqrt{Th}(\widehat{\theta} - \theta^*) \xrightarrow{d} N(0, \Sigma^*)$$

Σ^* achieves semiparametric efficiency lower bound. Weaker than the 1/2 threshold for discrete settings.

THM 2 — OPL REGRET (W.P. ≥1–7ε)

$$\mathrm{Reg}(\hat{\pi}) \lesssim \widetilde{O}\left(\log(1/h) \cdot (d_\Pi/Th + \sqrt{d_\Pi/Th}) + T^{-\gamma_\pi - \gamma_m} + h^s\right)$$

Requires only $\gamma_\pi + \gamma_m \geq s/(2s+1)$, also improves the regret of Leung'25 IPW estimator.

4. Experiments

SIMULATION SETUP

CONTEXT: $X \in \mathbb{R}^3 \sim N(0, I_3)$

LOG. POLICY: $P|X \sim N(\mu_0(X), 1)$, $P \in [0.5,3]$

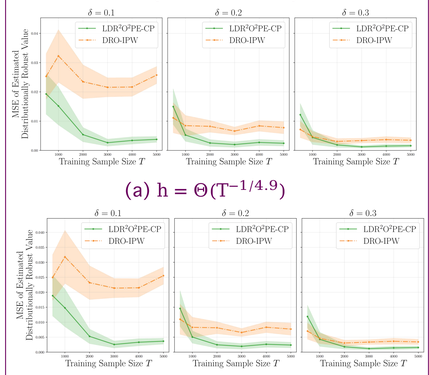
VALUATION: $f(X) = \theta_0 + \theta_1^T X + \theta_2 X_1 X_2$

NOISE: $Z \sim$ Logistic(0,0.2)

$\delta_{train} = 0.2$; Test $\delta \in \{0.1,0.2,0.3\}$

GRF (20 trees), L=5 folds, 20 seeds

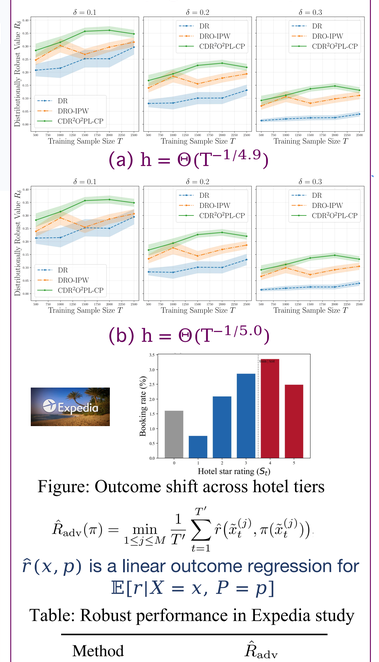
OPE: MSE of R_δ vs. Sample Size T (lower=better)



(a) $h = \Theta(T^{-1/4.9})$

(b) $h = \Theta(T^{-1/5.0})$

OPL: Robust Value $R_\delta(\widehat{\pi})$ at δ=0.2 (higher=better)



(a) $h = \Theta(T^{-1/4.9})$

(b) $h = \Theta(T^{-1/5.0})$

Figure: Outcome shift across hotel tiers

$\hat{R}_{adv}(\pi) = \min_{1 \leq j \leq M} \frac{1}{T'} \sum_{t=1}^{T'} \hat{r}(\tilde{x}_t^{(j)}, \pi(\tilde{x}_t^{(j)}))$

$\hat{r}(x, p)$ is a linear outcome regression for $\mathbb{E}[r|X = x, P = p]$

Table: Robust performance in Expedia study

Method	$\hat{R}_{adv}$
$\hat{\pi}_{Lin}$	6.201 ± 0.683
$\hat{\pi}_{DRO-IPW}$	6.254 ± 0.873
$\hat{\pi}_{CDR^2O^2PL-CP}$	$\mathbf{6.350 \pm 0.598}$

√ KEY FINDINGS

- **OPE:** Faster MSE decay than DRO-IPW for all δ
- **OPL:** Highest robust reward, tighter CI vs. DRO-IPW
- Non-robust DR highly sensitive to distributional shifts
- Both robust methods stable across $\delta \in \{0.1,0.2,0.3\}$

5. Resourses

Paper

Code

DRO · Off-Policy Evaluation / Learning · Contextual Pricing · Doubly Robust · Continuous Treatment

Min Xu, Xinyi Yin, Yunfan Zhang, Yuxuan Han, Houcai Shen, Caihua Chen