Wasserstein Policy Learning for Distributional Outcomes

Yiyan Huang[*†] HUANGYIYAN@GBU.EDU.CN
School of Computing and Information Technology, Great Bay University, Guangdong, China
Cheuk Hang Leung[*] CHLEUNG87@CITYU.EDU.HK
Department of Data Science, City University of Hong Kong, Hong Kong, China
Qi Wu[*] QIWU55@CITYU.EDU.HK
Department of Data Science, City University of Hong Kong, Hong Kong, China
Zhiheng Zhang[*] ZHANGZHIHENG@MAIL.SHUFE.EDU.CN
School of Statistics and Data Science & Institute of Big Data Research, Shanghai University of Finance and Economics, Shanghai, China

Editors: Steve Hanneke and Tor Lattimore

Abstract

Offline policy learning (OPL) usually optimizes the mean of a scalar potential outcome (Kitagawa and Tetenov, 2018; Athey and Wager, 2021). We study OPL when each potential outcome is a probability measure: for each action $a \in \mathcal{A}$, $\mathcal{Y}[a] \in \mathcal{P}_2(\mathbb{R})$. From logged data $\{(X_i, A_i, \mathcal{Y}_i)\}_{i=1}^N$ collected under propensity $f_0(a \mid x)$, and under standard causal assumptions, a policy $\pi : \mathcal{X} \to \mathcal{A}$ induces $\mathcal{Y}[\pi(X)]$. We evaluate it by a Wasserstein-Lipschitz utility of its barycenter,

$$\mu(\pi) \in \arg\min_{\mu \in \mathcal{P}_2(\mathbb{R})} \mathbb{E}\{\mathcal{W}_2^2(\mu, \mathcal{Y}[\pi(X)])\}, \qquad \pi^\star \in \arg\max_{\pi \in \Pi} U(\mu(\pi)).$$

The key simplification is one-dimensional Wasserstein geometry, which gives

$$\mu(\pi)^{-1}(t) = q_\pi(t) := \mathbb{E}[\mathcal{Y}[\pi(X)]^{-1}(t)],\ q_\pi(t) = \mathbb{E}\left[\frac{\mathbf{1}_{\{A=\pi(X)\}}\mathcal{Y}^{-1}(t)}{f_0(A \mid X)}\right] = \mathbb{E}[m_0(\pi(X), X)(t)],$$

where $m_0(a, x)(t) = \mathbb{E}[\mathcal{Y}^{-1}(t) \mid A = a, X = x]$. Hence policy learning reduces to uniform estimation of policy-indexed mean quantile curves over $\Pi \times [0, 1]$. We construct inverse-propensity and cross-fitted doubly robust (Chernozhukov et al., 2018) estimators of q_π, rearrange them into valid quantile functions, and maximize $U(\hat{\mu}(\pi))$ over Π. Let $V = \text{N-dim}(\Pi)$ and let η be the quantile-grid width. Under boundedness, overlap, quantile regularity, and Lipschitz continuity of U, with high probability, the regret of policies satisfy

$$\mathcal{R}(\hat{\pi}^{\mathrm{IPW}}) \le \widetilde{\mathcal{O}}(\sqrt{V/N}) + \mathcal{O}(\eta), \qquad \mathcal{R}(\hat{\pi}^{\mathrm{DR}}) \le C\{\mathcal{V}_N + \eta + r_f r_m + (r_f + r_m)\mathcal{V}_N + r_f \eta\},$$

where $\mathcal{V}_N = \widetilde{\mathcal{O}}(\sqrt{V/N})$. The DR bound exhibits Neyman orthogonality: if both nuisance estimators converge at the required uniform $N^{-1/4}$ rate and $\eta \asymp N^{-1/2}$, then $\mathcal{R}(\hat{\pi}^{\mathrm{DR}}) = \widetilde{\mathcal{O}}(N^{-1/2})$. We complement the upper bounds with a minimax lower bound. For the integrated quantile utility $U_\alpha(\nu) = \int_0^\alpha \nu^{-1}(t)dt$, take two valid base quantile curves q_- and q_+ and define $\Delta_Q := \int_0^\alpha (q_+(t) - q_-(t))dt$. Let $\underline{f}$ denote the overlap lower bound. There exists a universal constant $c_0 > 0$ such that any learner has worst-case expected regret at least $c_0 \Delta_Q \min\{1, \sqrt{\text{N-dim}(\Pi)/(\underline{f}N)}\}$ over this subclass. Thus, up to logarithmic factors and under the one-dimensional Wasserstein setting, distribution-valued outcomes introduce no additional leading-order nonparametric penalty beyond the policy-search complexity of Π.

Keywords: Causal inference, offline policy learning, distribution-valued outcomes, Wasserstein barycenter

∗. Authors are in alphabetical order.
†. Corresponding author.
Extended abstract. Full version appears as [arXiv:2606.19117v1].

Acknowledgments

Yiyan Huang was supported by the Startup Funds of Great Bay University (No. YJKY250111) and the Innovative Team Program for Regular Universities in Guangdong Province (No. 2025KCXTD031). Qi Wu was supported by the CityU-JD Digits Joint Laboratory in Financial Technology and Engineering, the Hong Kong Research Grants Council General Research Fund (Nos. 11219420/9043008 and 11200219/9042900), the HK Institute of Data Science, the InnoHK initiative of the Government of the HKSAR, and the Laboratory for AI-Powered Financial Technologies. Zhiheng Zhang was supported by the Fundamental Research Funds for the Central Universities (No. 2025110602), the Independent Research Project funded by the School of Statistics and Data Science (No. 2026110081), and the Shanghai Engineering Research Center of Finance Intelligence (No. 19DZ2254600).

References

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

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

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