Wasserstein Distributionally Robust Policy Learning with Continuous Context

Wenhao Yang

Management Science and Engineering,
Stanford University.

Joint work with Jose Blanchet, Miao Lu, and Zhengyuan Zhou.

Support from grants AFOSR-FA9550-20-1-0397 and NSF 2118199, 2229012 are gratefully acknowledged

Distributionally robust bandits

- Optimize a worst case objective. (X context, π policy, Y outcome)

$$\max_{\pi} \mathbb{E}_{(X,Y)\sim P}[Y(\pi(X))].$$

Distributionally robust bandits

- Optimize a worst case objective. (X context, π policy, Y outcome)

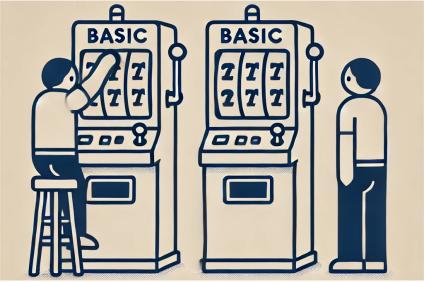
$$\max_{\pi} \mathbb{E}_{(X,Y)\sim P}[Y(\pi(X))]. \qquad \max_{\pi} \inf_{D(Q,P)\leq\rho} \mathbb{E}_{(X,Y)\sim Q}[Y(\pi(X))].$$

Distributionally robust bandits

- Optimize a worst case objective. (X context, π policy, Y outcome)

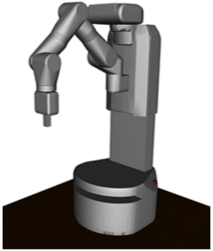
$$\max_{\pi} \mathbb{E}_{(X,Y)\sim P}[Y(\pi(X))]. \qquad \max_{\pi} \inf_{D(Q,P)\leq\rho} \mathbb{E}_{(X,Y)\sim Q}[Y(\pi(X))].$$

- Useful if deploy optimal π on some target domains.



Practical scenarios

- Robotic training.[Peng et al., 2018]



- Medical deployment.[Tang et al., 2024]

Topic of today

- How to solve distributionally robust bandits problems?
  - Wasserstein metric. $W_p(Q,P) = \inf_{\gamma(Q,P)}(\mathbb{E}_\gamma[d(X,Y)^p])^{\frac{1}{p}}$
  - Continuous context. (Discrete is studied in Shen et al. [2023])
  - Separate distribution shifts. (context shifts)

Context shift

- Primal:

$$V_\rho^c(\pi) = \inf_{\substack{P \in \mathcal{P}(\mathcal{X}) \\ W_p(P, P_X^0) \leq \rho}} \mathbb{E}_{X \sim P}\left[R^\pi(X)\right],$$

where $R^\pi(x) = \mathbb{E}[Y | X = x, \pi]$.

Context shift

- Primal:

$$V_\rho^c(\pi) = \inf_{\substack{P \in \mathcal{P}(\mathcal{X}) \\ W_p(P, P_X^0) \le \rho}} \mathbb{E}_{X \sim P}\left[R^\pi(X)\right],$$

  where $R^\pi(x) = \mathbb{E}[Y|X = x, \pi]$.

- Dual:

$$V_\rho^c(\pi) = \sup_{\lambda \in \mathbb{R}_+} \left\{ \mathbb{E}_{X \sim P_X^0} \left[\inf_{x \in \mathcal{X}} \left\{ R^\pi(x) + \lambda d(x, X)^p \right\} \right] - \lambda \rho^p \right\}.$$

Context shift

- Primal:

$$V_\rho^c(\pi) = \inf_{\substack{P \in \mathcal{P}(\mathcal{X}) \\ W_p(P, P_X^0) \leq \rho}} \mathbb{E}_{X \sim P}\left[R^\pi(X)\right],$$

where $R^\pi(x) = \mathbb{E}[Y|X = x, \pi]$.

- Dual:

$$V_\rho^c(\pi) = \sup_{\lambda \in \mathbb{R}_+} \left\{ \mathbb{E}_{X \sim P_X^0}\left[\inf_{x \in \mathcal{X}} \left\{ R^\pi(x) + \lambda d(x, X)^p \right\} \right] - \lambda \rho^p \right\}.$$

- Goal: find an optimal policy within a class Π:

$$\max_{\pi \in \Pi} V_\rho^c(\pi)$$

Estimation

$$V_\rho^c(\pi) = \sup_{\lambda \in \mathbb{R}_+} \left\{ \mathbb{E}_{X \sim P_X^0} \left[\inf_{x \in \mathcal{X}} \left\{ R^\pi(x) + \lambda d(x, X)^p \right\} \right] - \lambda \rho^p \right\}.$$

Estimation

$$V_\rho^c(\pi) = \sup_{\lambda \in \mathbb{R}_+} \left\{ \mathbb{E}_{X \sim P_X^0} \left[\inf_{x \in \mathcal{X}} \left\{ R^\pi(x) + \lambda d(x, X)^p \right\} \right] - \lambda \rho^p \right\}.$$

- Given dataset $\{(X_i, A_i, Y_i)\}_{i=1}^n$, how do we infer the value?

Estimation

$$V_\rho^c(\pi) = \sup_{\lambda\in\mathbb{R}_+}\left\{\mathbb{E}_{X\sim P_X^0}\left[\inf_{x\in\mathcal{X}}\left\{R^\pi(x) + \lambda d(x,X)^p\right\}\right] - \lambda\rho^p\right\}.$$

- Given dataset $\{(X_i, A_i, Y_i)\}_{i=1}^n$, how do we infer the value?
- Non-parametric estimator Nadaraya-Watson:

$$\widehat{R}(x,a) = \frac{\sum_{i=1}^n Y_i K(\frac{x-X_i}{h})1(A_i = a)}{\sum_{i=1}^n K(\frac{x-X_i}{h})1(A_i = a)},$$

Estimation

$$\widehat{V}_{\rho}^{c}(\pi) = \sup_{\lambda \in \mathbb{R}_+} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left[\inf_{x \in \mathcal{X}} \left\{ \widehat{R}^{\pi}(x) + \lambda d(x, X_i)^p \right\} \right] - \lambda \rho^p \right\}.$$

- Is it sample efficient?
- Is it computation efficient?

Statistical error

Theorem 1 [non-asymptotic and asymptotic]

If reward $R^{\pi}(x)$ is Lipschitz:

$$\left| \sup_{\pi \in \Pi} \widehat{V}_{\rho}^{c}(\pi) - \sup_{\pi \in \Pi} V_{\rho}^{c}(\pi) \right| = \mathcal{O}_P(n^{-\frac{1}{2+d}}) + \Delta_n$$

Statistical error

Theorem 1 [non-asymptotic and asymptotic]

If reward $R^\pi(x)$ is Lipschitz:

$$\left|\sup_{\pi\in\Pi}\widehat{V}^c_\rho(\pi)-\sup_{\pi\in\Pi}V^c_\rho(\pi)\right|=\mathcal{O}_P(n^{-\frac{1}{2+d}})+\Delta_n$$

- Where is the dependence on ρ?

$$\Delta_n=\begin{cases}\min\left\{\sqrt{\frac{\log\rho^{-1}}{n}},\rho+\sqrt{\frac{1}{n}}\right\}, & \text{when } \rho \text{ is small.}\\ 0, & \text{when } \rho \text{ is large.}\end{cases}$$

Statistical error

- High-level idea:

$$\rho \text{ is large}$$

Statistical error

- High-level idea:
$$\rho \text{ is large} \Rightarrow V_\rho^c(\pi) = \inf_x R^\pi(x).$$

- Only need to find the minimum.

Statistical error

- High-level idea:

$$\rho \text{ is large} \Rightarrow V_\rho^c(\pi) = \inf_x R^\pi(x).$$

- Only need to find the minimum.
- How large ρ should be?

Statistical error

- High-level idea:
$$\rho \text{ is large} \Rightarrow V_\rho^c(\pi) = \inf_x R^\pi(x).$$

- Only need to find the minimum.
- How large ρ should be?

Theorem 2

$V_\rho^c(\pi) = \inf_x R^\pi(x)$ if and only if
$$\rho^p \geq \mathbb{E}_X\left[d(x_{\inf}, X)^p \mathbf{1}\left(X \neq x_{\inf}\right)\right],$$
where $x_{\inf} \in \arg\inf R^\pi(x)$.

Summary

- Using non-parametric approach can achieve $\mathcal{O}_P(n^{-\frac{1}{2+d}})$ convergence rate.

Summary

- Using non-parametric approach can achieve $\mathcal{O}_P(n^{-\frac{1}{2+d}})$ convergence rate.
- Statistical inference is possible.

Summary

- Using non-parametric approach can achieve $\mathcal{O}_P(n^{-\frac{1}{2+d}})$ convergence rate.
- Statistical inference is possible.
- ρ plays a role in convergence rate.

Summary

- Using non-parametric approach can achieve $\mathcal{O}_P(n^{-\frac{1}{2+d}})$ convergence rate.
- Statistical inference is possible.
- ρ plays a role in convergence rate.
- Reward shift shares the similar results.

Compute the optimal policy

- How do we solve:

$$\begin{aligned}\sup_{\theta} \widehat{V}_{\rho}^{\mathrm{c}}(\pi_\theta) &= \sup_{\theta, \lambda \geq 0} \left\{ \mathbb{E}_{X \sim \widehat{P}_X^0} \left[\inf_{x \in \mathcal{X}} \left\{ \widehat{R}^{\pi_\theta}(x) + \lambda d(x, X)^p \right\} \right] - \lambda \rho^p \right\} \\ &:= \sup_{\theta, \lambda \geq 0} G_\rho(\lambda; \theta)\end{aligned}$$

Compute the optimal policy

- How do we solve:

$$\sup_{\theta} \widehat{V}_{\rho}^{\mathrm{c}}(\pi_\theta) = \sup_{\theta,\lambda\geq 0} \left\{ \mathbb{E}_{X\sim \widehat{P}_X^0} \left[\inf_{x\in\mathcal{X}} \left\{ \widehat{R}^{\pi_\theta}(x) + \lambda d(x,X)^p \right\} \right] - \lambda\rho^p \right\}$$
$$:= \sup_{\theta,\lambda\geq 0} G_\rho(\lambda;\theta)$$

- Step 1: Solve $\inf_{x\in\mathcal{X}}$.

Compute the optimal policy

- How do we solve:

$$\sup_{\theta} \widehat{V}_{\rho}^{\mathrm{c}}(\pi_\theta) = \sup_{\theta,\lambda\geq 0} \left\{ \mathbb{E}_{X\sim \widehat{P}_X^0} \left[\inf_{x\in\mathcal{X}} \left\{ \widehat{R}^{\pi_\theta}(x) + \lambda d(x,X)^p \right\} \right] - \lambda\rho^p \right\}$$
$$:= \sup_{\theta,\lambda\geq 0} G_\rho(\lambda;\theta)$$

- Step 1: Solve $\inf_{x\in\mathcal{X}}$.
- Step 2: Update λ.

Compute the optimal policy

- How do we solve:

$$\sup_{\theta} \widehat{V}^{\mathrm{c}}_{\rho}(\pi_\theta) = \sup_{\theta,\lambda\geq 0} \left\{ \mathbb{E}_{X\sim \widehat{P}^0_X} \left[\inf_{x\in\mathcal{X}} \left\{ \widehat{R}^{\pi_\theta}(x) + \lambda d(x,X)^p \right\} \right] - \lambda \rho^p \right\}$$
$$:= \sup_{\theta,\lambda\geq 0} G_\rho(\lambda;\theta)$$

- Step 1: Solve $\inf_{x\in\mathcal{X}}$.
- Step 2: Update λ.
- Step 3: Update θ.

Overall algorithm

Algorithm 1 Distributionally Robust Policy Gradient under Context Shift (`DRPG-C`)

for step $t = 0, \cdots, T-1$ **do**
  Set $\lambda_{t,0} = M/\rho^p$.
  for step $k = 0, \cdots, T_{\text{dual}} - 1$ **do**
    Set $x_{t,k}(X) = \texttt{MinimizationOracle}(R, \pi_{\theta_t}, \lambda_{t,k}, X, \alpha_x, T_{\text{inner}})$ for all $X \in \text{Supp}(\widehat{P}_X^0)$.
    Set $\lambda_{t,k+1} = \lambda_{t,k} + \alpha_\lambda \cdot (\mathbb{E}_{X \sim P_X^0}[d(x_{t,k}(X), X)^p] - \rho^p)$.
    if $|\mathbb{E}_{X \sim P_X^0}[d(x_{t,k}(X), X)^p] - \rho^p| \leq \varepsilon$ **then**
      break
    end if
  end for
  Update $\theta_{t+1} = \theta_t + \alpha_\theta \cdot \mathbb{E}_{X \sim P_X^0}[\sum_a R(x_{t,k}(X), a) \cdot \nabla_\theta \pi_{\theta_t}(a|x_{t,k}(X))]$.
end for
output: π_{θ_T}.

Theoretical Issues

$$\sup_{\lambda\geq 0} G_\rho(\lambda;\theta) = \sup_{\lambda\geq 0} \mathbb{E}_{X\sim\widehat{P}_X^0}\left[\inf_{x\in\mathcal{X}}\left\{\widehat{R}^{\pi_\theta}(x) + \lambda d(x,X)^p\right\}\right] - \lambda\rho^p\Bigg\}$$

- Optimize x: $\widehat{R}^{\pi_\theta}(x)$ may be non-convex. Let $p = 2$?

Theoretical Issues

$$\sup_{\lambda\geq 0} G_\rho(\lambda;\theta) = \sup_{\lambda\geq 0} \left\{ \mathbb{E}_{X\sim\widehat{P}_X^0}\left[\inf_{x\in\mathcal{X}}\left\{\widehat{R}^{\pi_\theta}(x) + \lambda d(x,X)^p\right\}\right] - \lambda\rho^p \right\}$$

- Optimize x: $\widehat{R}^{\pi_\theta}(x)$ may be non-convex. Let $p=2$?
- Optimize λ: $G_\rho(\lambda;\theta)$ may not be strongly-concave at its optimal point.

Condition [Robust level]

We consider the robust level ρ satisfying that $0 \leq \rho^2 \leq C_R$.

- Under the *robust level condition*,
  - Optimization $\inf_{x \in \mathcal{X}}$ is strongly convex.

Theoretical Issues

Condition [Robust level]

We consider the robust level ρ satisfying that $0 \leq \rho^2 \leq C_R$.

- Under the *robust level condition*,
  - Optimization $\inf_{x\in\mathcal{X}}$ is strongly convex.
  - $G(\lambda, \theta)$ is locally strongly concave.

Theoretical Issues

Condition [Robust level]

We consider the robust level ρ satisfying that $0 \leq \rho^2 \leq C_R$.

- Under the *robust level condition*,
  - Optimization $\inf_{x \in \mathcal{X}}$ is strongly convex.
  - $G(\lambda, \theta)$ is locally strongly concave.
- Convergence rate:

Theoretical Issues

Condition [Robust level]

We consider the robust level ρ satisfying that $0 \leq \rho^2 \leq C_R$.

- Under the *robust level condition*,
  - Optimization $\inf_{x \in \mathcal{X}}$ is strongly convex.
  - $G(\lambda, \theta)$ is locally strongly concave.
- Convergence rate:
  - Iteration complexity $\mathcal{O}(\log \varepsilon^{-1})$ for $\|x_T(X) - x_\lambda^*(X;\theta)\| \leq \varepsilon$.

Theoretical Issues

Condition [Robust level]

We consider the robust level ρ satisfying that $0 \leq \rho^2 \leq C_R$.

- Under the *robust level condition*,
  - Optimization $\inf_{x\in\mathcal{X}}$ is strongly convex.
  - $G(\lambda, \theta)$ is locally strongly concave.
- Convergence rate:
  - Iteration complexity $\mathcal{O}(\log \varepsilon^{-1})$ for $\|x_T(X) - x^*_\lambda(X;\theta)\| \leq \varepsilon$.
  - Iteration complexity $\mathcal{O}(\log \varepsilon^{-1})$ for $|\lambda_T - \lambda^*(\theta)| \leq \varepsilon$.

Theoretical Issues

Condition [Robust level]

We consider the robust level ρ satisfying that $0 \leq \rho^2 \leq C_R$.

- Under the *robust level condition*,
  - Optimization $\inf_{x \in \mathcal{X}}$ is strongly convex.
  - $G(\lambda, \theta)$ is locally strongly concave.
- Convergence rate:
  - Iteration complexity $\mathcal{O}(\log \varepsilon^{-1})$ for $\|x_T(X) - x_\lambda^*(X; \theta)\| \leq \varepsilon$.
  - Iteration complexity $\mathcal{O}(\log \varepsilon^{-1})$ for $|\lambda_T - \lambda^*(\theta)| \leq \varepsilon$.
  - Iteration complexity $\mathcal{O}(\varepsilon^{-2})$ for $\min_{t \leq T} \|\nabla_\theta V_c(\pi_{\theta_t})\| \leq \varepsilon$.

Simulation results

- How robust policy works?

Simulation results

- How robust policy works?
- $\mathcal{X} = [0,1] \times [0,1]$, $\mathcal{A} = \{0,1\}$. Reward function:

$$R(x,a) := 4\mathrm{ReLU}^2\left(x_2 - \left(x_1 - \frac{1}{2}\right)^2 - \frac{1}{2}\right) \cdot \mathbf{1}\{a=0\} + \frac{1}{15} \cdot \mathbf{1}\{a=1\}.$$

Simulation results

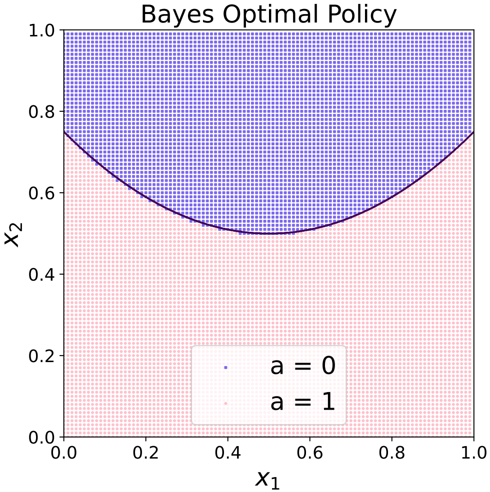
- How robust policy works?
- $\mathcal{X} = [0,1] \times [0,1]$, $\mathcal{A} = \{0,1\}$. Reward function:

$$R(x,a) := 4\mathrm{ReLU}^2\left(x_2 - \left(x_1 - \frac{1}{2}\right)^2 - \frac{1}{2}\right) \cdot \mathbf{1}\{a=0\} + \frac{1}{15} \cdot \mathbf{1}\{a=1\}.$$

- Bayes optimal policy:

$$\pi_\star^{\mathrm{Bayes}}(0|x) = \mathbf{1}\left\{x_2 \geq \left(x_1 - \frac{1}{2}\right)^2 + \frac{1}{2} + \frac{1}{2\sqrt{15}}\right\}.$$

Simulation results



Simulation results

- Linear policy class (Bayes optimal policy is excluded):

$$\Pi_{\Theta,\mathrm{lin}} = \left\{ \pi_\theta(a|x) = \frac{\exp(x^\top \theta_a)}{\sum_{a' \in \{0,1\}} \exp(x^\top \theta'_a)} : \theta_0, \theta_1 \in \mathbb{R}^2 \right\}.$$

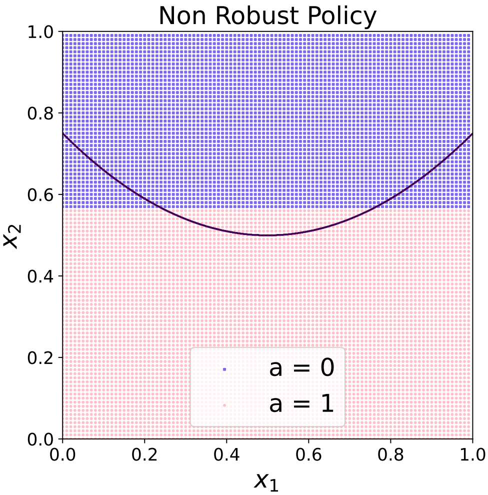
Simulation results

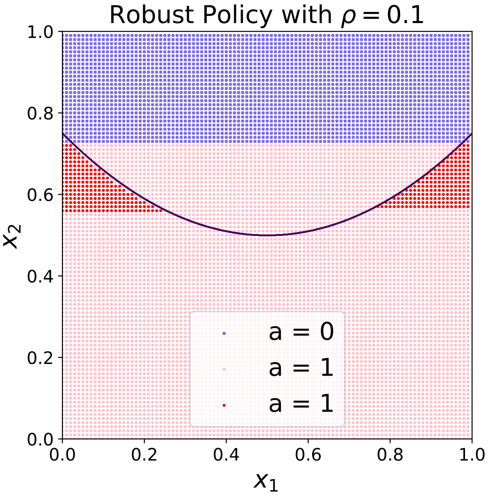
- Linear policy class (Bayes optimal policy is excluded):

$$\Pi_{\Theta,\mathrm{lin}} = \left\{ \pi_\theta(a|x) = \frac{\exp(x^\top \theta_a)}{\sum_{a' \in \{0,1\}} \exp(x^\top \theta'_a)} : \theta_0, \theta_1 \in \mathbb{R}^2 \right\}.$$

- Run standard policy gradient of non-robust case.
- Run `DRPG-C`.

Simulation results





Conclusion

- Leveraging non-parametric estimators to solve Wasserstein distributionally robust bandits.
- Statistical error is controlled by classic non-parametric rate, in both non-asymptotic and asymptotic regime.
- A practical algorithm is developed with theoretical guarantees.

Thank you for listening! Happy to take questions.

Poster at APS Market showcase at Flex C (Oct 22, 2:15-3:30pm)
"Limit Theorems for SGD with Infinite Variance"

Xue Bin Peng, Marcin Andrychowicz, Wojciech Zaremba, and Pieter Abbeel. Sim-to-real transfer of robotic control with dynamics randomization. In *2018 IEEE international conference on robotics and automation (ICRA)*, pages 3803–3810. IEEE, 2018.

Yi Shen, Pan Xu, and Michael M Zavlanos. Wasserstein distributionally robust policy evaluation and learning for contextual bandits. *arXiv preprint arXiv:2309.08748*, 2023.

Yumeng Tang, Ziming Cui, Xishi Wang, Shuyu Xiang, and Yifeng Li. Distributionally robust optimization methods on robust medical diagnosis systems. *Applied and Computational Engineering*, 44:99–107, 2024.