Distributionally Robust Batch Contextual Bandits

Nian Si
Joint work with Fan Zhang, Zhengyuan Zhou, and Jose Blanchet

INFORMS 2020

October 3, 2020

Road map

Motivation: Distributional Shift in Batch Bandit

A collection of bandit observational data: $\{(X_i, A_i, Y_i)\}_{i=1}^n \overset{i.i.d.}{\sim} \mathbf{P}_{\mathrm{a}} * \pi_0$, given the known collection policy $A_i \sim \pi_0(\cdot \mid X_i)$.

Motivation: Distributional Shift in Batch Bandit

A collection of bandit observational data: $\{(X_i, A_i, Y_i)\}_{i=1}^{n} \overset{i.i.d.}{\sim} \mathbf{P}_{\mathrm{a}} * \pi_0$, given the known collection policy $A_i \sim \pi_0(\cdot \mid X_i)$.

How to design a robust policy for the environment $\mathbf{P}_{\mathrm{b}} \approx \mathbf{P}_{\mathrm{a}}$?

Setting

- Context: $X \in \mathcal{X}$; Actions: $A \in \mathcal{A} = \{a^1, a^2, \ldots, a^d\}$; Rewards: $(Y(a^1), Y(a^2), \ldots, Y(a^d)) \in \prod_{j=1}^{d} \mathcal{Y}_j$.

Setting

- Context: $X \in \mathcal{X}$; Actions: $A \in \mathcal{A} = \{a^1, a^2, \dots, a^d\}$; Rewards: $(Y(a^1), Y(a^2), \dots, Y(a^d)) \in \prod_{j=1}^{d} \mathcal{Y}_j$.
- Batch bandit data: $\{(X_i, A_i, Y_i)\}_{i=1}^{n}$, where $(X_i, Y_i(a^1), Y_i(a^2), \dots, Y_i(a^d)) \overset{i.i.d.}{\sim} \mathbf{P}_0$, and $A_i \sim \pi_0(\cdot \mid X_i)$.

Setting

- Context: $X \in \mathcal{X}$; Actions: $A \in \mathcal{A} = \{a^1, a^2, \ldots, a^d\}$; Rewards: $(Y(a^1), Y(a^2), \ldots, Y(a^d)) \in \prod_{j=1}^d \mathcal{Y}_j$.
- Batch bandit data: $\{(X_i, A_i, Y_i)\}_{i=1}^n$, where $(X_i, Y_i(a^1), Y_i(a^2), \ldots, Y_i(a^d)) \overset{i.i.d.}{\sim} \mathbf{P}_0$, and $A_i \sim \pi_0(\cdot \mid X_i)$.
- Assumptions: unconfoundedness, overlapping, bounded reward support, and positive densities.

Setting

- Context: $X \in \mathcal{X}$; Actions: $A \in \mathcal{A} = \{a^1, a^2, \dots, a^d\}$; Rewards: $(Y(a^1), Y(a^2), \dots, Y(a^d)) \in \prod_{j=1}^{d} \mathcal{Y}_j$.
- Batch bandit data: $\{(X_i, A_i, Y_i)\}_{i=1}^{n}$, where $(X_i, Y_i(a^1), Y_i(a^2), \dots, Y_i(a^d)) \overset{i.i.d.}{\sim} \mathbf{P}_0$, and $A_i \sim \pi_0(\cdot \mid X_i)$.
- Assumptions: unconfoundedness, overlapping, bounded reward support, and positive densities.
  - For any $i = 1, 2, \dots, d$, $Y(a^i)|X$ has a non-zero conditional density $f_i(y_i|x) \geq \underline{b} > 0$ over the interval $[0, M]$.

Setting

- Context: $X \in \mathcal{X}$; Actions: $A \in \mathcal{A} = \{a^1, a^2, \ldots, a^d\}$; Rewards: $(Y(a^1), Y(a^2), \ldots, Y(a^d)) \in \prod_{j=1}^{d} \mathcal{Y}_j$.
- Batch bandit data: $\{(X_i, A_i, Y_i)\}_{i=1}^{n}$, where $(X_i, Y_i(a^1), Y_i(a^2), \ldots, Y_i(a^d)) \overset{i.i.d.}{\sim} \mathbf{P}_0$, and $A_i \sim \pi_0(\cdot \mid X_i)$.
- Assumptions: unconfoundedness, overlapping, bounded reward support, and positive densities.
  - For any $i = 1, 2, \ldots, d$, $Y(a^i)|X$ has a non-zero conditional density $f_i(y_i|x) \geq \underline{b} > 0$ over the interval $[0, M]$.
- Goal: learn a robust policy that performs well in the presence of the distributional shifts.

Distributionally Robust Formulation and Policy Evaluation

- Uncertainty set: $\mathcal{U}_{\mathbf{P}_0}(\delta) = \{\mathbf{P} \ll \mathbf{P}_0 \mid KL(\mathbf{P}||\mathbf{P}_0) \leq \delta\}$.

Distributionally Robust Formulation and Policy Evaluation

- Uncertainty set: $\mathcal{U}_{\mathbf{P}_0}(\delta) = \{\mathbf{P} \ll \mathbf{P}_0 \mid KL(\mathbf{P}||\mathbf{P}_0) \leq \delta\}$.
- Strong duality for the distributionally robust value function:

$$Q_{\mathrm{DRO}}(\pi) := \inf_{\mathbf{P} \in \mathcal{U}_{\mathbf{P}_0(\delta)}} \mathbf{E}_{\mathbf{P}}[Y(\pi(X))]$$

Distributionally Robust Formulation and Policy Evaluation

- Uncertainty set: $\mathcal{U}_{\mathbf{P}_0}(\delta) = \{\mathbf{P} \ll \mathbf{P}_0 \mid KL(\mathbf{P}||\mathbf{P}_0) \leq \delta\}$.
- Strong duality for the distributionally robust value function:

$$Q_{\mathrm{DRO}}(\pi) := \inf_{\mathbf{P} \in \mathcal{U}_{\mathbf{P}_0(\delta)}} \mathbf{E}_{\mathbf{P}}[Y(\pi(X))]$$

$$= \sup_{\alpha \geq 0} \{-\alpha \log \mathbf{E}_{\mathbf{P}_0} [\exp(-Y(\pi(X))/\alpha)] - \alpha\delta\}$$
$$= \sup_{\alpha \geq 0} \left\{-\alpha \log \mathbf{E}_{\mathbf{P}_0 * \pi_0} \left[\frac{\exp(-Y(A)/\alpha)\mathbf{1}\{\pi(X) = A\}}{\pi_0(A \mid X)}\right] - \alpha\delta\right\}.$$

Distributionally Robust Formulation and Policy Evaluation

- Uncertainty set: $\mathcal{U}_{\mathbf{P}_0}(\delta) = \{\mathbf{P} \ll \mathbf{P}_0 \mid KL(\mathbf{P}||\mathbf{P}_0) \leq \delta\}$.
- Strong duality for the distributionally robust value function:

$$Q_{\mathrm{DRO}}(\pi) := \inf_{\mathbf{P} \in \mathcal{U}_{\mathbf{P}_0(\delta)}} \mathbf{E}_{\mathbf{P}}[Y(\pi(X))]$$

$$= \sup_{\alpha \geq 0} \{-\alpha \log \mathbf{E}_{\mathbf{P}_0}[\exp(-Y(\pi(X))/\alpha)] - \alpha\delta\}$$

$$= \sup_{\alpha \geq 0} \left\{-\alpha \log \mathbf{E}_{\mathbf{P}_0 * \pi_0}\left[\frac{\exp(-Y(A)/\alpha)\mathbf{1}\{\pi(X) = A\}}{\pi_0(A \mid X)}\right] - \alpha\delta\right\}.$$

- Finite-sample estimate: $\hat{Q}_{\mathrm{DRO}}(\pi) = \sup_{\alpha \geq 0}\{-\alpha \log \hat{W}_n(\pi, \alpha) - \alpha\delta\}$, where

$$\hat{W}_n(\pi, \alpha) = \frac{1}{\sum_{i=1}^n \frac{\mathbf{1}\{\pi(X_i)=A_i\}}{\pi_0(A_i|X_i)}} \sum_{i=1}^n \frac{\mathbf{1}\{\pi(X_i) = A_i\}}{\pi_0(A_i \mid X_i)} \exp(-Y_i(A_i)/\alpha).$$

Distributionally Robust Formulation and Policy Evaluation

- Uncertainty set: $\mathcal{U}_{\mathbf{P}_0}(\delta) = \{\mathbf{P} \ll \mathbf{P}_0 \mid KL(\mathbf{P}||\mathbf{P}_0) \leq \delta\}$.
- Strong duality for the distributionally robust value function:

$$Q_{\mathrm{DRO}}(\pi) := \inf_{\mathbf{P} \in \mathcal{U}_{\mathbf{P}_0(\delta)}} \mathbf{E}_{\mathbf{P}}[Y(\pi(X))]$$

$$= \sup_{\alpha \geq 0} \{-\alpha \log \mathbf{E}_{\mathbf{P}_0} [\exp(-Y(\pi(X))/\alpha)] - \alpha\delta\}$$

$$= \sup_{\alpha \geq 0} \left\{-\alpha \log \mathbf{E}_{\mathbf{P}_0 * \pi_0} \left[\frac{\exp(-Y(A)/\alpha)\mathbf{1}\{\pi(X) = A\}}{\pi_0(A \mid X)}\right] - \alpha\delta\right\}.$$

- Finite-sample estimate: $\hat{Q}_{\mathrm{DRO}}(\pi) = \sup_{\alpha \geq 0}\{-\alpha \log \hat{W}_n(\pi, \alpha) - \alpha\delta\}$, where

$$\hat{W}_n(\pi, \alpha) = \frac{1}{\sum_{i=1}^n \frac{\mathbf{1}\{\pi(X_i)=A_i\}}{\pi_0(A_i|X_i)}} \sum_{i=1}^n \frac{\mathbf{1}\{\pi(X_i) = A_i\}}{\pi_0(A_i \mid X_i)} \exp(-Y_i(A_i)/\alpha).$$

Central Limit Theorem

Theorem

Under standard assumptions, for any policy $\pi \in \Pi$, we have

$$\sqrt{n}\left(\hat{Q}_{\mathrm{DRO}}(\pi) - Q_{\mathrm{DRO}}(\pi)\right) \Rightarrow \mathcal{N}\left(0, \sigma^2(\alpha^*)\right),$$

where α^* is the optimal dual variable, defined by

$$\alpha^* = \arg\max_{\alpha \geq 0} \left\{-\alpha \log \mathbf{E}_{\mathbf{P}_0}\left[\exp(-Y(\pi(X))/\alpha)\right] - \alpha\delta\right\},$$

and

$$\sigma^2(\alpha) = \frac{\alpha^2}{\mathbf{E}\left[W_i(\pi,\alpha)\right]^2}\mathbf{E}\left[\frac{1}{\pi_0\left(\pi(X)|X\right)}\left(\exp\left(-Y(\pi(X))/\alpha\right)\right.\right.$$
$$\left.\left. - \mathbf{E}\left[\exp\left(-Y(\pi(X))/\alpha\right)\right]\right)^2\right].$$

A Learning Algorithm

- How to find a good policy:

$$\arg\max_{\pi \in \Pi} \inf_{\mathbf{P} \in \mathcal{U}_{\mathbf{P}_0}(\delta)} \mathbf{E}_{\mathbf{P}}[Y(\pi(X))]?$$

A Learning Algorithm

- How to find a good policy:

$$\arg\max_{\pi \in \Pi} \inf_{\mathbf{P} \in \mathcal{U}_{\mathbf{P}_0}(\delta)} \mathbf{E}_{\mathbf{P}}[Y(\pi(X))]?$$

- Given a policy class Π, learn a distributionally robust policy:

$$\begin{aligned} \hat{\pi}_{\mathrm{DRO}} &= \arg\max_{\pi \in \Pi} \hat{Q}_{\mathrm{DRO}}(\pi) \\ &= \arg\max_{\pi \in \Pi} \sup_{\alpha \geq 0} \{-\alpha \log \hat{W}_n(\pi, \alpha) - \alpha\delta\} \end{aligned}$$

A Learning Algorithm

- How to find a good policy:

$$\arg\max_{\pi\in\Pi} \inf_{\mathbf{P}\in\mathcal{U}_{\mathbf{P}_0}(\delta)} \mathbf{E}_{\mathbf{P}}[Y(\pi(X))]?$$

- Given a policy class Π, learn a distributionally robust policy:

$$\begin{aligned}\hat{\pi}_{\mathrm{DRO}} &= \arg\max_{\pi\in\Pi} \hat{Q}_{\mathrm{DRO}}(\pi) \\ &= \arg\max_{\pi\in\Pi} \sup_{\alpha\geq 0}\{-\alpha \log \hat{W}_n(\pi,\alpha) - \alpha\delta\}\end{aligned}$$

- Alternatively update π and α;

A Learning Algorithm

- How to find a good policy:

$$\arg\max_{\pi \in \Pi} \inf_{\mathbf{P} \in \mathcal{U}_{\mathbf{P}_0}(\delta)} \mathbf{E}_{\mathbf{P}}[Y(\pi(X))]?$$

- Given a policy class Π, learn a distributionally robust policy:

$$\begin{aligned} \hat{\pi}_{\mathrm{DRO}} &= \arg\max_{\pi \in \Pi} \hat{Q}_{\mathrm{DRO}}(\pi) \\ &= \arg\max_{\pi \in \Pi} \sup_{\alpha \geq 0} \{-\alpha \log \hat{W}_n(\pi, \alpha) - \alpha\delta\} \end{aligned}$$

- Alternatively update π and α;
  - Using Newton-Raphson method to update α; converge fast empirically.

Statistical Performance Guarantee

Theorem

Under assumptions mentioned above, with probability at least $1-\varepsilon$, we have

$$\max_{\pi'\in\Pi}\inf_{\mathbf{P}\in\mathcal{U}_{\mathbf{P}_0}(\delta)}\mathbf{E}_{\mathbf{P}}[Y(\pi'(X))]-\inf_{\mathbf{P}\in\mathcal{U}_{\mathbf{P}_0}(\delta)}\mathbf{E}_{\mathbf{P}}[Y(\pi(X))]$$

$$\leq \frac{4}{\underline{b}\eta\sqrt{n}}\left((\sqrt{2}+1)\kappa^{(n)}(\Pi)+\sqrt{2\log\left(\frac{2}{\varepsilon}\right)}+C\right),$$

where $\kappa^{(n)}(\Pi)$ is the entropy integral defined via the Hammer distance in Π, $\eta>0$ is a lower bound for the propensity score (collection policy) $\pi_0(a,x)$, and C is a universal constant.

Statistical Performance Guarantee

Theorem

Under assumptions mentioned above, with probability at least $1 - \varepsilon$, we have

$$\max_{\pi' \in \Pi} \inf_{\mathbf{P} \in \mathcal{U}_{\mathbf{P}_0}(\delta)} \mathbf{E}_{\mathbf{P}}[Y(\pi'(X))] - \inf_{\mathbf{P} \in \mathcal{U}_{\mathbf{P}_0}(\delta)} \mathbf{E}_{\mathbf{P}}[Y(\pi(X))]$$
$$\leq \frac{4}{\underline{b}\eta\sqrt{n}} \left((\sqrt{2}+1)\kappa^{(n)}(\Pi) + \sqrt{2\log\left(\frac{2}{\varepsilon}\right)} + C \right),$$

where $\kappa^{(n)}(\Pi)$ is the entropy integral defined via the Hammer distance in Π, $\eta > 0$ is a lower bound for the propensity score (collection policy) $\pi_0(a, x)$, and C is a universal constant.

Simulation Study: Benchmark

Benchmark: let $\overline{\Pi}$ denotes the class of all measurable mappings from contexts $\mathcal{X}$ to the action set $\mathcal{A}$.

- Bayes policy $\bar{\pi}^*$:

$$\bar{\pi}^* \in \arg\max_{\pi\in\overline{\Pi}} \mathbf{E}_{\mathbf{P}}[Y(\pi(X))], \text{ and}$$

- Bayes DRO policy $\bar{\pi}^*_{\mathrm{DRO}}$:

$$\overline{\pi}^*_{\mathrm{DRO}} \in \arg\max_{\pi\in\overline{\Pi}} \inf_{\mathbf{P}\in\mathcal{U}_{\mathbf{P}_0(\delta)}} \mathbf{E}_{\mathbf{P}}[Y(\pi(X))].$$

Simulation Study: Benchmark

Benchmark: let $\overline{\Pi}$ denotes the class of all measurable mappings from contexts $\mathcal{X}$ to the action set $\mathcal{A}$.

- Bayes policy $\bar{\pi}^*$:

$$\bar{\pi}^* \in \arg\max_{\pi \in \overline{\Pi}} \mathbf{E}_{\mathbf{P}}[Y(\pi(X))], \text{ and}$$

- Bayes DRO policy $\bar{\pi}^*_{\mathrm{DRO}}$:

$$\overline{\pi}^*_{\mathrm{DRO}} \in \arg\max_{\pi \in \overline{\Pi}} \inf_{\mathbf{P} \in \mathcal{U}_{\mathbf{P}_0(\delta)}} \mathbf{E}_{\mathbf{P}}[Y(\pi(X))].$$

- Easy to compute, because the policies are the best response for each X.

Simulation Study: A Linear Example

- A linear example: 5-dimensional features, but only the first two matters:

$$Y(i)|X \sim \mathcal{N}(\beta_i^\top X, \sigma_i^2), \text{ for } i = 1, 2, 3.$$

for $\beta_1 = (1, 0, 0, 0, 0), \beta_2 = (-1/2, \sqrt{3}/2, 0, 0, 0), \beta_3 = (-1/2, -\sqrt{3}/2, 0, 0, 0)$. and $\sigma_1 = 0.2, \sigma_2 = 0.5, \sigma_3 = 0.8$.

Simulation Study: A Linear Example

- A linear example: 5-dimensional features, but only the first two matters:

  $$Y(i)|X \sim \mathcal{N}(\beta_i^\top X, \sigma_i^2), \text{ for } i = 1, 2, 3.$$

  for $\beta_1 = (1, 0, 0, 0, 0), \beta_2 = (-1/2, \sqrt{3}/2, 0, 0, 0), \beta_3 = (-1/2, -\sqrt{3}/2, 0, 0, 0)$. and $\sigma_1 = 0.2, \sigma_2 = 0.5, \sigma_3 = 0.8$.
- The linear policy class:
  $\Pi = \{\pi(x) = \arg\max_{a \in \mathcal{A}} \ \{\theta_a^\top x\} : \theta_a \in \mathbf{R}^p, a \in \mathcal{A}\}$.

Simulation Study: A Linear Example

- A linear example: 5-dimensional features, but only the first two matters:

$$Y(i)|X \sim \mathcal{N}(\beta_i^\top X, \sigma_i^2), \text{ for } i = 1, 2, 3.$$

for $\beta_1 = (1, 0, 0, 0, 0), \beta_2 = (-1/2, \sqrt{3}/2, 0, 0, 0), \beta_3 = (-1/2, -\sqrt{3}/2, 0, 0, 0)$. and $\sigma_1 = 0.2, \sigma_2 = 0.5, \sigma_3 = 0.8$.

- The linear policy class:
$\Pi = \{\pi(x) = \arg\max_{a\in\mathcal{A}} \left\{\theta_a^\top x\right\} : \theta_a \in \mathbf{R}^p, a \in \mathcal{A}\}$.
- Collection policy π_0:

	Region 1	Region 2	Region 3
Action 1	0.50	0.25	0.25
Action 2	0.25	0.50	0.25
Action 3	0.25	0.25	0.50

Table 1: The probabilities of selecting an action based on π_0 in the linear example.

Linear Example

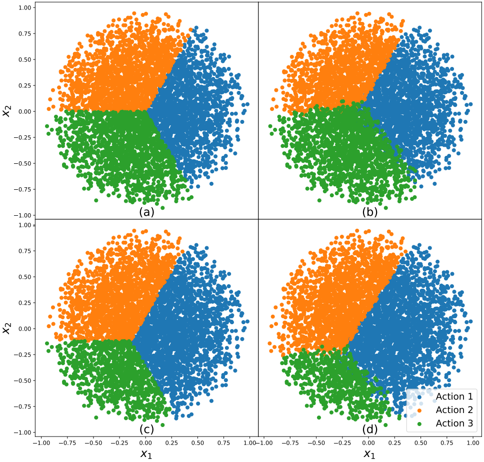

Figure 1: (a) Bayes policy $\bar{\pi}^*$; (b) non-DRO linear policy; (c) Bayes distributionally robust policy $\overline{\pi}^*_{\mathrm{DRO}}$; (d) distributionally robust linear policy $\hat{\pi}_{\mathrm{DRO}}$.

Non-linear Example with the Linear Policy Class

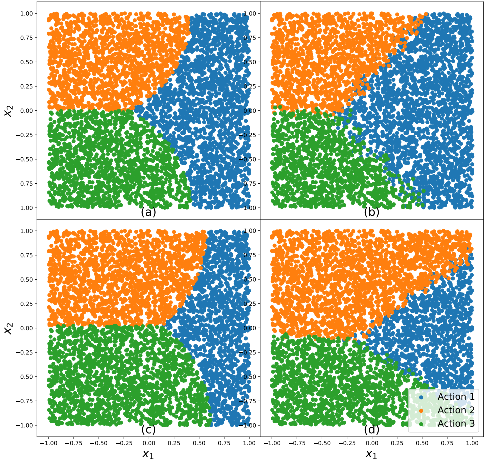

Figure 2: (a) Bayes policy $\bar{\pi}^*$; (b) non-DRO linear policy; (c) Bayes distributionally robust policy $\overline{\pi}^*_{\mathrm{DRO}}$; (d) distributionally robust linear policy $\hat{\pi}_{\mathrm{DRO}}$. $\sigma_1 = 0.8, \sigma_2 = 0.2, \sigma_3 = 0.4$.

Extension to f-divergence Uncertainty Set

For $f_k(t) \triangleq \frac{t^k - kt + k - 1}{k(k-1)}$, define f-divergence as

$$D_k(\mathbf{P}||\mathbf{P}_0) \triangleq \int f_k\left(\frac{d\mathbf{P}}{d\mathbf{P}_0}\right) d\mathbf{P}_0.$$

Extension to f-divergence Uncertainty Set

For $f_k(t) \triangleq \frac{t^k - kt + k - 1}{k(k-1)}$, define f-divergence as

$$D_k(\mathbf{P}||\mathbf{P}_0) \triangleq \int f_k\left(\frac{d\mathbf{P}}{d\mathbf{P}_0}\right) d\mathbf{P}_0.$$

Theorem

Under assumptions mentioned above, with probability at least $1-\varepsilon$, we have

$$\max_{\pi' \in \Pi} \inf_{\mathbf{P} \in \mathcal{U}^k_{\mathbf{P}_0}(\delta)} \mathbf{E}_{\mathbf{P}}[Y(\pi'(X))] - \inf_{\mathbf{P} \in \mathcal{U}^k_{\mathbf{P}_0}(\delta)} \mathbf{E}_{\mathbf{P}}[Y(\pi(X))]$$

$$\leq \frac{4c_k(\delta)}{\underline{b}\eta\sqrt{n}} \left((\sqrt{2}+1)\kappa^{(n)}(\Pi) + \sqrt{2\log\left(\frac{2}{\varepsilon}\right)} + C \right),$$

where $c_k(\delta) \triangleq (1 + k(k-1)\delta)^{1/k}$.

Reference

Si, Nian, Fan Zhang, Zhengyuan Zhou, and Jose Blanchet. "Distributional Robust Batch Contextual Bandits." arXiv preprint arXiv:2006.05630 (2020).

The short version has been accepted in ICML 2020.

Thanks!