Distributionally Robust Batch Contextual Bandits

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

INFORMS 2021

Stanford University

October 24, 2021

Road map

Stanford University

1. Motivation: distributional shifts in batch contextual bandit
2. Distributionally robust formulation
3. Distributionally robust policy learning
4. Numerical results
5. Extension to f-divergence uncertainty set

Motivation: distributional shifts in batch bandit

Stanford University

Motivation: distributional shifts in batch bandit

Motivation: distributional shifts in batch bandit

Stanford University

Motivation: distributional shifts in batch bandit

Stanford University

A collection of triplets of context, action and rewards in an environment $\mathbf{P}_{\mathrm{a}}$.

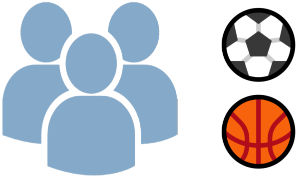
We aim to deploy a robust policy in unknown environments $\mathbf{P}_{\mathrm{b}}$ which are similar but slightly different from the previous environment.

$$\mathbf{P}_{\mathrm{b}} \approx \mathbf{P}_{\mathrm{a}}$$

Main challenges

Stanford University

- Incomplete (bandit-type) data:

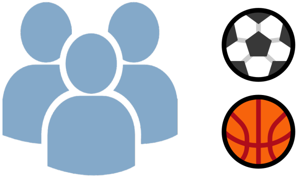


✓	✗	?	✓	?	?	?	✓
?	?	✓	?	✗	✗	✗	?

Main challenges

Stanford University

- Incomplete (bandit-type) data:



✓	✗	?	✓	?	?	?	✓
?	?	✓	?	✗	✗	✗	?

- Distributional shifts: covariate shift and concept drift.

Setting

Stanford University

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

Setting

Stanford University

- 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: $\{(X_i, A_i, Y_i(A_i))\}_{i=1}^{n}$, 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 $A_i \sim \pi_0(\cdot \mid X_i)$ is known.

Setting

Stanford University

- 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: $\{(X_i, A_i, Y_i(A_i))\}_{i=1}^{n}$, 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 $A_i \sim \pi_0(\cdot \mid X_i)$ is known.
- Goal: learn a robust policy that performs well in the presence of unknown distributional shifts.

Assumptions: standard assumptions[1]

Stanford University

Assumption (Standard assumptions)

1. *Unconfoundedness:* $(Y(a^1), Y(a^2), \ldots, Y(a^d))$ *is independent with* A *conditional on* X*, i.e.,*
$$(Y(a^1), Y(a^2), \ldots, Y(a^d)) \perp\!\!\!\perp A \mid X.$$
2. *Overlap: There exists some* $\eta > 0$, $\pi_0(a \mid x) \geq \eta$, $\forall (x, a) \in \mathcal{X} \times \mathcal{A}$.
3. *Bounded reward support:* $0 \leq Y(a^i) \leq M$ *for* $i = 1, 2, \ldots, d$.

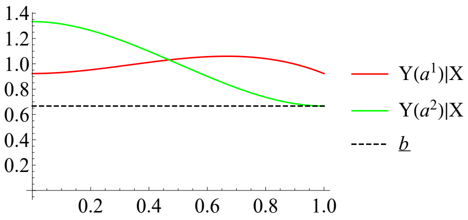
[1] This assumption is standard and commonly adopted in both the causal inference literature (Rosenbaum and Rubin [1983], Imbens [2004], Imbens and Rubin [2015]) and the policy learning literature (Zhang et al. [2012], Zhao et al. [2012], Kitagawa and Tetenov [2018], Swaminathan and Joachims [2015], Zhou et al. [2017]).

Assumptions: positive densities/probabilities

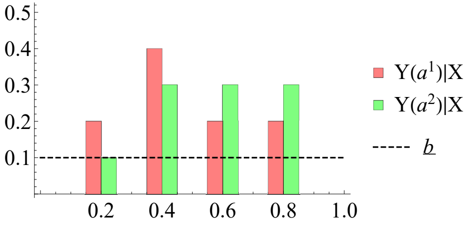
Stanford University

Assumption (Positive densities/probabilities)

1. **Continuous case:** *for any* $i = 1, 2, \ldots, d$, $Y(a^i)|X$ *has a conditional density* $f_i(y_i|x)$, *and* $f_i(y_i|x) \geq \underline{b} > 0$ *over the interval* $[0, M]$ *for any* $x \in \mathcal{X}$.
2. **Discrete case:** *for any* $i = 1, 2, \ldots, d$, $Y(a^i)$ *supported on a finite set* $\mathbb{D}$, *and* $\mathbf{P}_0(Y(a^i) = v|X) \geq \underline{b} > 0$ *for any* $v \in \mathbb{D}$.



(a) Continuous probability distribution



(b) Discrete probability distribution

Distributionally robust formulation

Stanford University

- How to model distributional shifts?
  - Kullback-Leibler divergence: $KL(\mathbf{P}||\mathbf{P}_0) \triangleq \int_{\mathcal{X}\times\prod_{j=1}^{d}\mathcal{Y}_j} \log\left(\frac{\mathrm{d}\mathbf{P}}{\mathrm{d}\mathbf{P}_0}\right)\mathrm{d}\mathbf{P}$.

Distributionally robust formulation

Stanford University

- How to model distributional shifts?
  - Kullback-Leibler divergence: $KL(\mathbf{P}||\mathbf{P}_0) \triangleq \int_{\mathcal{X}\times\prod_{j=1}^{d}\mathcal{Y}_j} \log\left(\frac{\mathrm{d}\mathbf{P}}{\mathrm{d}\mathbf{P}_0}\right)\mathrm{d}\mathbf{P}$.
- Uncertainty set: $\mathcal{U}_{\mathbf{P}_0}(\delta) = \{\mathbf{P} \mid KL(\mathbf{P}||\mathbf{P}_0) \leq \delta\}$.

Distributionally robust formulation

Stanford University

- How to model distributional shifts?
  - Kullback-Leibler divergence: $KL(\mathbf{P}||\mathbf{P}_0) \triangleq \int_{\mathcal{X}\times\prod_{j=1}^{d}\mathcal{Y}_j} \log\left(\frac{\mathrm{d}\mathbf{P}}{\mathrm{d}\mathbf{P}_0}\right)\mathrm{d}\mathbf{P}$.
- Uncertainty set: $\mathcal{U}_{\mathbf{P}_0}(\delta) = \{\mathbf{P} \mid KL(\mathbf{P}||\mathbf{P}_0) \leq \delta\}$.
- Distributionally robust value function (population level):

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

Distributionally robust formulation

Stanford University

- How to model distributional shifts?
  - Kullback-Leibler divergence: $KL(\mathbf{P}||\mathbf{P}_0) \triangleq \int_{\mathcal{X}\times\prod_{j=1}^{d}\mathcal{Y}_j} \log\left(\frac{\mathrm{d}\mathbf{P}}{\mathrm{d}\mathbf{P}_0}\right) \mathrm{d}\mathbf{P}$.
- Uncertainty set: $\mathcal{U}_{\mathbf{P}_0}(\delta) = \{\mathbf{P} \mid KL(\mathbf{P}||\mathbf{P}_0) \leq \delta\}$.
- Distributionally robust value function (population level):

$$\underbrace{Q_{\mathrm{DRO}}(\pi) \triangleq \inf_{\mathbf{P}\in\mathcal{U}_{\mathbf{P}_0}(\delta)} \mathbf{E}_{\mathbf{P}}[Y(\pi(X))].}_{\text{Infinite dimensional optimization. Bandit observations for } \mathbf{P}_0\text{.}}$$

Tractable reformulation and policy evaluation

Stanford University

- Strong duality[2] for the distributionally robust value function:

$$
\begin{aligned}
Q_{\mathrm{DRO}}(\pi) &= \inf_{\mathbf{P}\in\mathcal{U}_{\mathbf{P}_0}(\delta)} \mathbf{E}_{\mathbf{P}}[Y(\pi(X))] \\
&= \sup_{\alpha\geq 0} \left\{-\alpha \log \mathbf{E}_{\mathbf{P}_0}\left[\exp(-Y(\pi(X))/\alpha)\right] - \alpha\delta\right\}
\end{aligned}
$$

[2] Hu and Hong [2013]

Tractable reformulation and policy evaluation

Stanford University

- Strong duality[2] for the distributionally robust value function:

$$
\begin{aligned}
Q_{\mathrm{DRO}}(\pi) &= \inf_{\mathbf{P} \in \mathcal{U}_{\mathbf{P}_0}(\delta)} \mathbf{E}_{\mathbf{P}}[Y(\pi(X))] \\
&= \sup_{\alpha \geq 0} \left\{ -\alpha \log \mathbf{E}_{\mathbf{P}_0} \left[\exp(-Y(\pi(X))/\alpha)\right] - \alpha\delta \right\} \\
&= \sup_{\alpha \geq 0} \left\{ -\alpha \log \mathbf{E}_{\mathbf{P}_0 * \pi_0} \left[\frac{\exp(-Y(A)/\alpha) \mathbf{1}\{\pi(X) = A\}}{\pi_0(A \mid X)} \right] - \alpha\delta \right\}.
\end{aligned}
$$

[2] Hu and Hong [2013]

Tractable reformulation and policy evaluation

Stanford University

- Strong duality[2] for the distributionally robust value function:

$$\begin{aligned} Q_{\mathrm{DRO}}(\pi) &= \inf_{\mathbf{P}\in\mathcal{U}_{\mathbf{P}_0}(\delta)} \mathbf{E}_{\mathbf{P}}[Y(\pi(X))] \\ &= \sup_{\alpha\geq 0}\left\{-\alpha\log\mathbf{E}_{\mathbf{P}_0}\left[\exp(-Y(\pi(X))/\alpha)\right]-\alpha\delta\right\} \\ &= \sup_{\alpha\geq 0}\left\{-\alpha\log\mathbf{E}_{\mathbf{P}_0*\pi_0}\left[\frac{\exp(-Y(A)/\alpha)\mathbf{1}\{\pi(X)=A\}}{\pi_0(A\mid X)}\right]-\alpha\delta\right\}. \end{aligned}$$

where $\mathbf{P}_0 * \pi_0$ denotes the product distribution on the space $\mathcal{X} \times \prod_{j=1}^{d} \mathcal{Y}_j \times \mathcal{A}$.

[2] Hu and Hong [2013]

Tractable reformulation and policy evaluation

Stanford University

- Strong duality[2] for the distributionally robust value function:

$$\begin{aligned} Q_{\mathrm{DRO}}(\pi) &= \inf_{\mathbf{P}\in\mathcal{U}_{\mathbf{P}_0}(\delta)} \mathbf{E}_{\mathbf{P}}[Y(\pi(X))] \\ &= \sup_{\alpha\geq 0}\left\{-\alpha \log \mathbf{E}_{\mathbf{P}_0}\left[\exp(-Y(\pi(X))/\alpha)\right] - \alpha\delta\right\} \\ &= \sup_{\alpha\geq 0}\left\{-\alpha \log \mathbf{E}_{\mathbf{P}_0 * \pi_0}\left[\frac{\exp(-Y(A)/\alpha)\mathbf{1}\{\pi(X)=A\}}{\pi_0(A\mid X)}\right] - \alpha\delta\right\}. \end{aligned}$$

where $\mathbf{P}_0 * \pi_0$ denotes the product distribution on the space $\mathcal{X} \times \prod_{j=1}^{d} \mathcal{Y}_j \times \mathcal{A}$.

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

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

[2] Hu and Hong [2013]

Tractable reformulation and policy evaluation

Stanford University

- Strong duality[2] for the distributionally robust value function:

$$
\begin{aligned}
Q_{\mathrm{DRO}}(\pi) &= \inf_{\mathbf{P} \in \mathcal{U}_{\mathbf{P}_0}(\delta)} \mathbf{E}_{\mathbf{P}}[Y(\pi(X))] \\
&= \sup_{\alpha \geq 0} \{-\alpha \log \mathbf{E}_{\mathbf{P}_0}[\exp(-Y(\pi(X))/\alpha)] - \alpha\delta\} \\
&= \sup_{\alpha \geq 0} \left\{-\alpha \log \mathbf{E}_{\mathbf{P}_0 * \pi_0}\left[\frac{\exp(-Y(A)/\alpha)\mathbf{1}\{\pi(X) = A\}}{\pi_0(A \mid X)}\right] - \alpha\delta\right\}.
\end{aligned}
$$

  where $\mathbf{P}_0 * \pi_0$ denotes the product distribution on the space $\mathcal{X} \times \prod_{j=1}^{d} \mathcal{Y}_j \times \mathcal{A}$.

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

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

[2] Hu and Hong [2013]

Central limit theorem

Stanford University

Theorem

Under assumptions mentioned earlier, 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 $\sigma^2(\alpha) =$

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

A learning algorithm

Stanford University

- How to find a good policy: $\arg\max_{\pi\in\Pi} Q_{\mathrm{DRO}}(\pi)$?

A learning algorithm

Stanford University

- How to find a good policy: $\arg\max_{\pi\in\Pi} Q_{\mathrm{DRO}}(\pi)$?
- 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

Stanford University

- How to find a good policy: $\arg\max_{\pi\in\Pi} Q_{\mathrm{DRO}}(\pi)$?
- 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.

A learning algorithm

Stanford University

- How to find a good policy: $\arg\max_{\pi \in \Pi} Q_{\mathrm{DRO}}(\pi)$?
- 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.
- How does $\hat{\pi}_{\mathrm{DRO}}$ perform?

$$\begin{aligned} R_{\mathrm{DRO}}(\hat{\pi}_{\mathrm{DRO}}) &= \max_{\pi' \in \Pi} Q_{\mathrm{DRO}}(\pi') - Q_{\mathrm{DRO}}(\hat{\pi}_{\mathrm{DRO}}) \\ &= \max_{\pi' \in \Pi} \inf_{\mathbf{P} \in \mathcal{U}_{\mathbf{P}_0}(\delta)} \mathbf{E}_{\mathbf{P}}[Y(\pi'(X))] - \inf_{\mathbf{P} \in \mathcal{U}_{\mathbf{P}_0}(\delta)} \mathbf{E}_{\mathbf{P}}[Y(\hat{\pi}_{\mathrm{DRO}}(X))]. \end{aligned}$$

Statistical performance guarantee

Stanford University

Theorem

Under assumptions mentioned earlier, with probability at least $1-\varepsilon$, we have in the continuous case

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

and in the discrete case

$$R_{\mathrm{DRO}}(\hat{\pi}_{\mathrm{DRO}}) \leq \frac{4M}{\underline{b}\eta\sqrt{n}}\left(24(\sqrt{2}+1)\kappa^{(n)}(\Pi) + 48\sqrt{|\mathbb{D}|\log(2)} + \sqrt{2\log\left(\frac{2}{\varepsilon}\right)}\right),$$

where $\kappa^{(n)}(\Pi)$ represents the complexity of the policy class Π, and $\eta > 0$ is a lower bound for the propensity score (collection policy) $\pi_0(a, x)$ mentioned in the previous assumption.

Statistical performance guarantee

Stanford University

Theorem

Under assumptions mentioned earlier, with probability at least $1-\varepsilon$*, we have in the continuous case*

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

and in the discrete case

$$R_{\mathrm{DRO}}(\hat{\pi}_{\mathrm{DRO}}) \leq \frac{4M}{\underline{b}\eta\sqrt{n}}\left(24(\sqrt{2}+1)\kappa^{(n)}(\Pi)+48\sqrt{|\mathbb{D}|\log(2)}+\sqrt{2\log\left(\frac{2}{\varepsilon}\right)}\right),$$

where $\kappa^{(n)}(\Pi)$ *represents the complexity of the policy class* Π*, and* $\eta > 0$ *is a lower bound for the propensity score (collection policy)* $\pi_0(a,x)$ *mentioned in the previous assumption.*

Statistical performance guarantee

Stanford University

Theorem

Under assumptions mentioned earlier, with probability at least $1-\varepsilon$, we have in the continuous case

$$R_{\mathrm{DRO}}(\hat{\pi}_{\mathrm{DRO}}) \le \frac{4}{\underline{b}\eta\sqrt{n}}\left((\sqrt{2}+1)\kappa^{(n)}(\Pi) + \sqrt{2\log\left(\frac{2}{\varepsilon}\right)} + C\right),$$

and in the discrete case

$$R_{\mathrm{DRO}}(\hat{\pi}_{\mathrm{DRO}}) \le \frac{4M}{\underline{b}\eta\sqrt{n}}\left(24(\sqrt{2}+1)\kappa^{(n)}(\Pi) + 48\sqrt{|\mathbb{D}|\log(2)} + \sqrt{2\log\left(\frac{2}{\varepsilon}\right)}\right),$$

where $\kappa^{(n)}(\Pi)$ represents the complexity of the policy class Π, and $\eta > 0$ is a lower bound for the propensity score (collection policy) $\pi_0(a,x)$ mentioned in the previous assumption.

Remarks on the complexity term $\kappa^{(n)}(\Pi)$

Stanford University

Example

- **Finite class:** For a policy class Π_{Fin} containing a finite number of policies, we have $\kappa^{(n)}(\Pi_{\mathrm{Fin}}) \leq \sqrt{\log(|\Pi_{\mathrm{Fin}}|)}$.
- **Linear class:** For $\mathcal{X} \subset \mathbf{R}^p$, each policy $\pi \in \Pi_{\mathrm{Lin}}$ is parameterized by a set of d vectors $\Theta = \{\theta_a \in \mathbf{R}^p : a \in \mathcal{A}\} \in \mathbf{R}^{p \times d}$, and the mapping $\pi : \mathcal{X} \to \mathcal{A}$ is defined as

$$\pi_\Theta(x) \in \arg\max_{a \in \mathcal{A}} \left\{\theta_a^\top x\right\}.$$

Then, we have $\kappa^{(n)}(\Pi_{\mathrm{Lin}}) \leq C\sqrt{dp \log(d) \log(dp)}$.

- In general, $\kappa^{(n)}(\Pi)$ can be bounded by the VC dimension when $d = 2$, or the graph dimension when $d > 2$.

Simulation, real data experiments, and the selection of δ

Simulation study: benchmark

Stanford University

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

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

$$\overline{\pi}^* \in \arg\max_{\pi \in \overline{\Pi}_0} \mathbf{E}_{\mathbf{P}}[Y(\pi(X))], \text{ and}$$

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

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

Simulation study: benchmark

Stanford University

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

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

$$\overline{\pi}^* \in \arg\max_{\pi \in \overline{\Pi_0}} \mathbf{E}_{\mathbf{P}}[Y(\pi(X))], \text{ and}$$

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

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

- Best policies, but may not in the policy class Π.
- Not learnable, but theoretically easy to compute in the simulation environment, because the policies are the best response for each X.

Simulation study

Stanford University

- 3 actions; 5-dimensional features, but only the first two matter:

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

where the conditional mean $\mu_i(x)$ and conditional variance σ_i are chosen as

$$\begin{aligned} \mu_1(x) &= 0.2x(1), & \sigma_1 &= 0.8, \\ \mu_2(x) &= 1 - \sqrt{(x(1)+0.5)^2 + (x(2)-1)^2}, & \sigma_2 &= 0.2, \\ \mu_3(x) &= 1 - \sqrt{(x(1)+0.5)^2 + (x(2)+1)^2}, & \sigma_3 &= 0.4. \end{aligned}$$

Simulation study

Stanford University

- 3 actions; 5-dimensional features, but only the first two matter:

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

where the conditional mean $\mu_i(x)$ and conditional variance σ_i are chosen as

$$\begin{aligned} \mu_1(x) &= 0.2x(1), & \sigma_1 &= 0.8, \\ \mu_2(x) &= 1 - \sqrt{(x(1) + 0.5)^2 + (x(2) - 1)^2}, & \sigma_2 &= 0.2, \\ \mu_3(x) &= 1 - \sqrt{(x(1) + 0.5)^2 + (x(2) + 1)^2}, & \sigma_3 &= 0.4. \end{aligned}$$

- Bayes policy: $\overline{\pi}^*(x) \in \arg\max_{i=1,2,3}\{\mu_i(x)\}$;
  DRO Bayes policy: $\overline{\pi}^*_{\mathrm{DRO}}(x) \in \arg\max_{i=1,2,3}\left\{\mu_i(x) - \frac{\sigma_i^2}{2\alpha^*(\pi^*_{\mathrm{DRO}})}\right\}$.

Simulation study

Stanford University

- 3 actions; 5-dimensional features, but only the first two matter:

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

where the conditional mean $\mu_i(x)$ and conditional variance σ_i are chosen as

$$\begin{aligned} \mu_1(x) &= 0.2x(1), & \sigma_1 &= 0.8, \\ \mu_2(x) &= 1 - \sqrt{(x(1)+0.5)^2 + (x(2)-1)^2}, & \sigma_2 &= 0.2, \\ \mu_3(x) &= 1 - \sqrt{(x(1)+0.5)^2 + (x(2)+1)^2}, & \sigma_3 &= 0.4. \end{aligned}$$

- Bayes policy: $\overline{\pi}^*(x) \in \arg\max_{i=1,2,3}\{\mu_i(x)\}$;
  DRO Bayes policy: $\overline{\pi}^*_{\mathrm{DRO}}(x) \in \arg\max_{i=1,2,3}\left\{\mu_i(x) - \frac{\sigma_i^2}{2\alpha^*(\pi^*_{\mathrm{DRO}})}\right\}$.
- The linear policy class: $\Pi = \{\pi(x) = \arg\max_{a\in\mathcal{A}} \ \{\theta_a^\top x\} : \theta_a \in \mathbf{R}^p, a \in \mathcal{A}\}$.

Non-linear example with the linear policy class

Stanford University

(a) Bayes policy $\overline{\pi}^*$;

(b) non-DRO linear policy;

(c) Bayes distributionally robust policy $\overline{\pi}^*_{\mathrm{DRO}}$

(d) distributionally robust linear policy $\hat{\pi}_{\mathrm{DRO}}$.

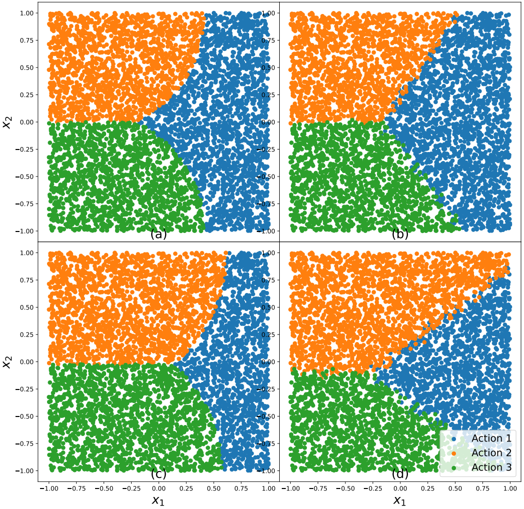


Figure 1: $\sigma_1 = 0.8(\textit{blue}), \sigma_2 = 0.2(\textit{orange}), \sigma_3 = 0.4(\textit{green})$.

Non-linear example with the linear policy class

(a) Bayes policy $\overline{\pi}^*$;

(b) non-DRO linear policy;

(c) Bayes distributionally robust policy $\overline{\pi}^*_{\mathrm{DRO}}$

(d) distributionally robust linear policy $\hat{\pi}_{\mathrm{DRO}}$.

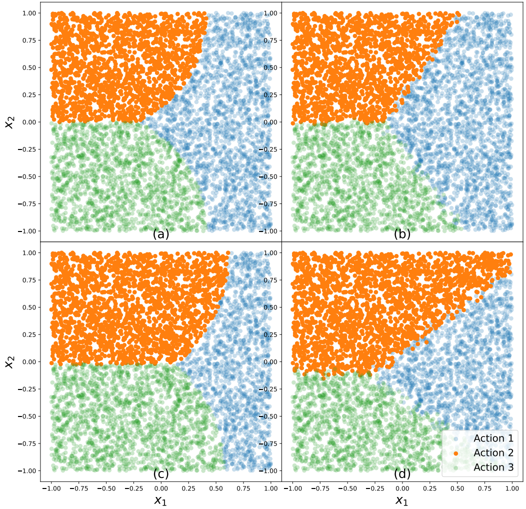


Figure 1: $\sigma_1 = 0.8(\textit{blue}), \sigma_2 = 0.2(\textit{orange}), \sigma_3 = 0.4(\textit{green})$.

Non-linear example with the linear policy class

Stanford University

(a) Bayes policy $\overline{\pi}^*$;

(b) non-DRO linear policy;

(c) Bayes distributionally robust policy $\overline{\pi}^*_{\mathrm{DRO}}$

(d) distributionally robust linear policy $\hat{\pi}_{\mathrm{DRO}}$.

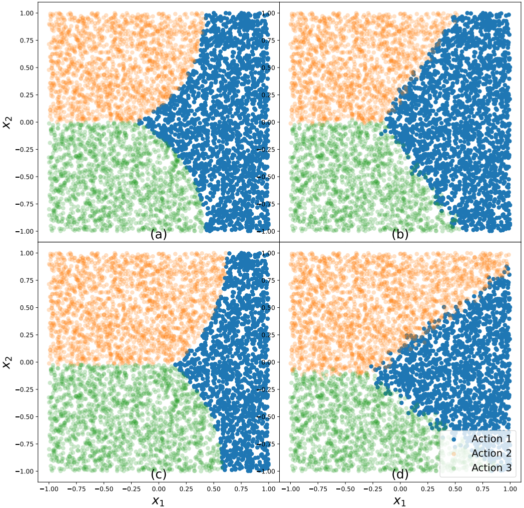


Figure 1: $\sigma_1 = 0.8(\textit{blue}), \sigma_2 = 0.2(\textit{orange}), \sigma_3 = 0.4(\textit{green})$.

[3]

---

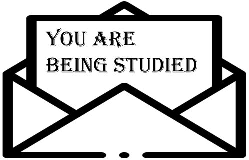
[3]Credit: Getty Images/iStockphoto

Backgrounds

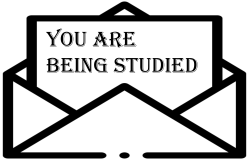
Stanford University

- Dataset Description:[4] 180002 data points on whether individuals voted in the 2006 primary election with their characteristics. There is 1 control and 4 treatments.

(a) Civic



(b) Monitored



(c) Self History

(d) Neighbors

[4] Gerber et al. [2008]

Actions

Stanford University

- There are 5 actions (1 control with probability 5/9 and 4 treatments each with probability 1/9).
    - **Nothing:** No action is performed.
    - **Civic:** A letter with "Do your civic duty" is mailed to the household before the primary election.
    - **Monitored:** A letter with "You are being studied" is mailed to the household before the primary election.
    - **Self History:** A letter with the past voting records of the voter's household is mailed to the household before the primary election.
    - **Neighbors:** A letter with the past voting records of this voter's household and neighbors is mailed to the household.

Actions

Stanford University

- There are 5 actions (1 control with probability 5/9 and 4 treatments each with probability 1/9).
  - **Nothing:** No action is performed.
  - **Civic:** A letter with "Do your civic duty" is mailed to the household before the primary election.
  - **Monitored:** A letter with "You are being studied" is mailed to the household before the primary election.
  - **Self History:** A letter with the past voting records of the voter's household is mailed to the household before the primary election.
  - **Neighbors:** A letter with the past voting records of this voter's household and neighbors is mailed to the household.
- **Neighbors** is dominant for the whole population. To make all actions comparable, we minus an artificial cost of deploying each action: $Y_i(a) = \mathbf{1}\{\text{voter } i \text{ votes in 2006 under action } a\} - c_a$.

Actions

Stanford University

- There are 5 actions (1 control with probability 5/9 and 4 treatments each with probability 1/9).
  - **Nothing:** No action is performed.
  - **Civic:** A letter with "Do your civic duty" is mailed to the household before the primary election.
  - **Monitored:** A letter with "You are being studied" is mailed to the household before the primary election.
  - **Self History:** A letter with the past voting records of the voter's household is mailed to the household before the primary election.
  - **Neighbors:** A letter with the past voting records of this voter's household and neighbors is mailed to the household.
- **Neighbors** is dominant for the whole population. To make all actions comparable, we minus an artificial cost of deploying each action: $Y_i(a) = \mathbf{1}\{\text{voter } i \text{ votes in 2006 under action } a\} - c_a$.
- Goal: learn a distributionally robust policy to maximize voting turnout.

Training and evaluation procedure

Stanford University

- We use decision trees as the policy class.

Training and evaluation procedure

- We use decision trees as the policy class.
- We divide the training and test population based on the *city* (101 cities in the dataset).
  - Natural covariate shifts and concept drifts; e.g., the distribution of *year of birth* is generally different across different cities.
  - Leave-one-out to generate 101 pairs of training set and test set.

Training and evaluation procedure

Stanford University

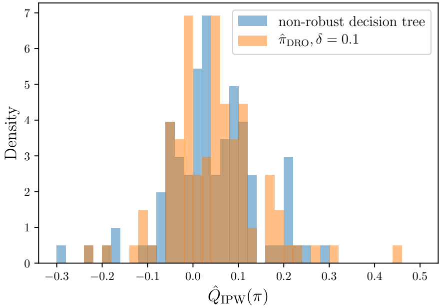
- We use decision trees as the policy class.
- We divide the training and test population based on the *city* (101 cities in the dataset).
  - Natural covariate shifts and concept drifts; e.g., the distribution of *year of birth* is generally different across different cities.
  - Leave-one-out to generate 101 pairs of training set and test set.

		mean	std	min	5% quantile
Non-robust		0.0386	0.0991	-0.2844	-0.1104
Robust	$\delta = 0.1$	0.0458	0.0989	-0.2321	-0.1007
	$\delta = 0.2$	0.0368	0.0895	-0.2314	-0.0785
	$\delta = 0.3$	0.0397	0.0864	-0.2313	-0.0677
	$\delta = 0.4$	0.0383	0.0863	-0.2312	-0.0677

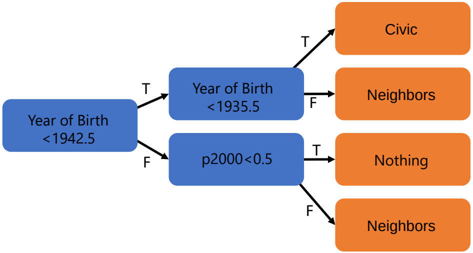
Table 1: Comparison of important statistics for 101 test results.

Results for $\delta = 0.1$

Stanford University



(a) Comparison of test performances between a distributionally robust decision tree and a non-robust decision tree



(b) Example of a distributionally robust tree

How to select the uncertain size δ in practice?

Stanford University

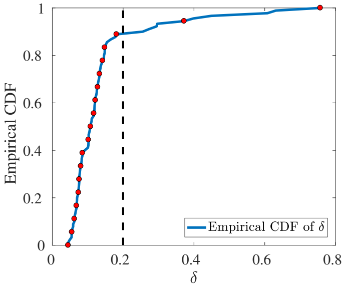
Selecting δ is more like a managerial decision rather than a scientific procedure.

How to select the uncertain size δ in practice?

Stanford University

Selecting δ is more like a managerial decision rather than a scientific procedure.

- Compute δ based on the training data:
  - Estimate distributions of Y using any causal inference/machine learning methods.
  - Randomly split training data into 20 cities ($\mathbf{P}^{20}$) against 80 cities ($\mathbf{P}^{80}$) 100 times.
  - Estimate δ based on $KL(\mathbf{P}^{20}||\mathbf{P}^{80}) = KL(\mathbf{P}_X^{20}||\mathbf{P}_X^{80}) + \mathbf{E}_{\mathbf{P}_X^{20}}[KL(\mathbf{P}_Y^{20}|X||\mathbf{P}_Y^{80}|X)]$.

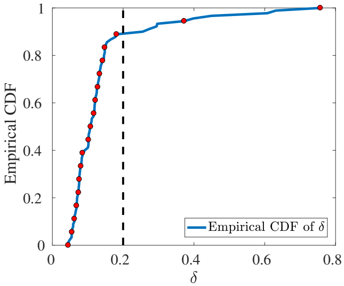


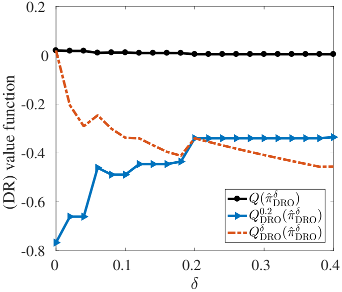
How to select the uncertain size δ in practice?

Stanford University

Selecting δ is more like a managerial decision rather than a scientific procedure.

- Compute δ based on the training data:
  - Estimate distributions of Y using any causal inference/machine learning methods.
  - Randomly split training data into 20 cities ($\mathbf{P}^{20}$) against 80 cities ($\mathbf{P}^{80}$) 100 times.
  - Estimate δ based on $KL(\mathbf{P}^{20}||\mathbf{P}^{80}) = KL(\mathbf{P}^{20}_X||\mathbf{P}^{80}_X) + \mathbf{E}_{\mathbf{P}^{20}_X}[KL(\mathbf{P}^{20}_Y|X||\mathbf{P}^{80}_Y|X)]$.
- Check the performance of $\hat{\pi}^{\delta}_{\mathrm{DRO}}$ using different value functions.
  - Robust policy does not compromise the non-robust value function.
  - The performance is not sensitive to δ, when $\delta \geq 0.2$.





Extension to f-divergence uncertainty set

Stanford University

- Up to now, all of the results are for Kullback-Leibler divergence.
- We can also generalize to f_k-divergence.

Extension to f-divergence uncertainty set

Stanford University

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

Stanford University

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 in the continuous case (similar result for the discrete case)

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

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

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

The paper

Stanford University

Si N, Zhang F, Zhou Z, and Blanchet J. "Distributional Robust Batch Contextual Bandits." arXiv preprint arXiv:2006.05630 (2020). under review.

Thanks!

References I

Stanford University

Alan S Gerber, Donald P Green, and Christopher W Larimer. Social pressure and voter turnout: Evidence from a large-scale field experiment. *American political Science review*, 102(1):33–48, 2008.

Zhaolin Hu and L Jeff Hong. Kullback-leibler divergence constrained distributionally robust optimization. *Available at Optimization Online*, 2013.

Guido W Imbens. Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and statistics*, 86(1):4–29, 2004.

G.W. Imbens and D.B. Rubin. *Causal Inference in Statistics, Social, and Biomedical Sciences*. Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction. Cambridge University Press, 2015. ISBN 9780521885881.

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

Paul R Rosenbaum and Donald B Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

Adith Swaminathan and Thorsten Joachims. Batch learning from logged bandit feedback through counterfactual risk minimization. *Journal of Machine Learning Research*, 16:1731–1755, 2015.

References II

Stanford University

Baqun Zhang, Anastasios A Tsiatis, Marie Davidian, Min Zhang, and Eric Laber. Estimating optimal treatment regimes from a classification perspective. *Stat*, 1(1):103–114, 2012.

Yingqi Zhao, Donglin Zeng, A John Rush, and Michael R Kosorok. Estimating individualized treatment rules using outcome weighted learning. *Journal of the American Statistical Association*, 107(499): 1106–1118, 2012.

Xin Zhou, Nicole Mayer-Hamblett, Umer Khan, and Michael R Kosorok. Residual weighted learning for estimating individualized treatment rules. *Journal of the American Statistical Association*, 112(517): 169–187, 2017.