Policy learning 'without' overlap:

Pessimism and generalized empirical Bernstein's inequality

Ying Jin

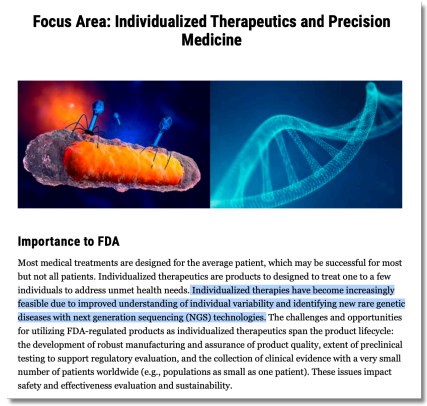
Department of Statistics, Stanford University

Joint work with Zhimei Ren (UChicago stats),
Zhuoran Yang (Yale stats), and Zhaoran Wang (Northwestern IEMS)

Data Driven Decision Making Seminar, January 17, 2022

Precision medicine

figure credit: National Cancer Institute



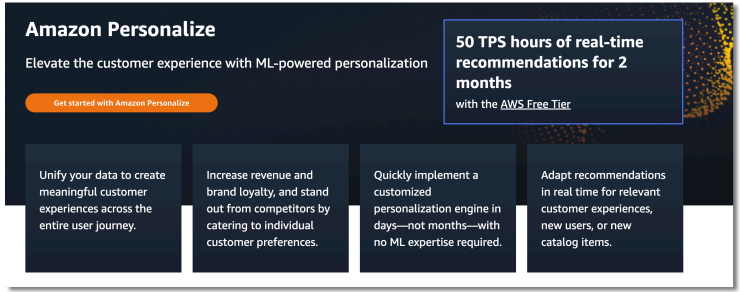
Focus Area: Individualized Therapeutics and Precision Medicine

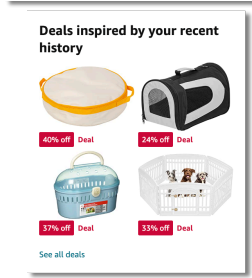
Importance to FDA

Most medical treatments are designed for the average patient, which may be successful for most but not all patients. Individualized therapeutics are products to designed to treat one to a few individuals to address unmet health needs. Individualized therapies have become increasingly feasible due to improved understanding of individual variability and identifying new rare genetic diseases with next generation sequencing (NGS) technologies. The challenges and opportunities for utilizing FDA-regulated products as individualized therapeutics span the product lifecycle: the development of robust manufacturing and assurance of product quality, extent of preclinical testing to support regulatory evaluation, and the collection of clinical evidence with a very small number of patients worldwide (e.g., populations as small as one patient). These issues impact safety and effectiveness evaluation and sustainability.

Given a patient's characteristics, what is the best treatment rule?

Personalized recommendation system





What product should be recommended to a user given their digital information?

Policy learning

characteristics —**policy**→ **action**

- Online experiments?

  Could be costly/unethical ...

- Large amounts of offline data

Adaptive experiment data for a treatment

logged data from a recommendation system

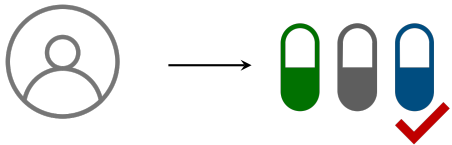
Question: Effectively learn an optimal policy from offline data?

The contextual bandit model

a user's characteristics

$X \in \mathcal{X}$

The contextual bandit model



a user's characteristics
$X \in \mathcal{X}$

an action
$A \in \mathcal{A} = (a_1, \ldots, a_K)$

The contextual bandit model

a user's characteristics
$X \in \mathcal{X}$

an action
$A \in \mathcal{A} = (a_1, \ldots, a_K)$

a reward
$Y \in [0, 1]$

The contextual bandit model

a user's characteristics
$X \in \mathcal{X}$

an action
$A \in \mathcal{A} = (a_1, \ldots, a_K)$

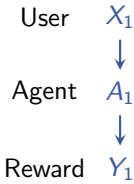
a reward
$Y \in [0, 1]$

$$X \sim P_X, \qquad Y = \mu(X, A) + \epsilon$$

ϵ: independent mean-zero noise

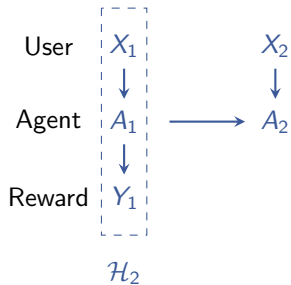
Offline data collection mechanism

Example: bandit algorithms, adaptive experiments, fixed logging policy ...



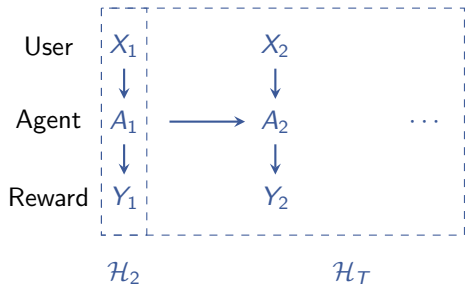
Offline data collection mechanism

Example: bandit algorithms, adaptive experiments, fixed logging policy ...



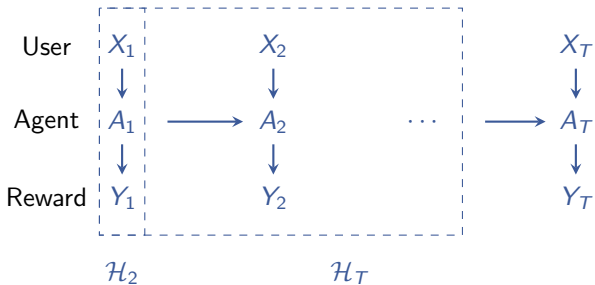
Offline data collection mechanism

Example: bandit algorithms, adaptive experiments, fixed logging policy ...



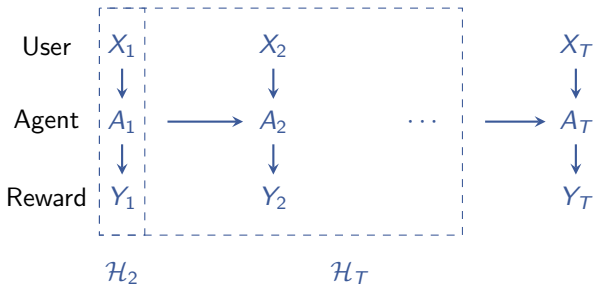
Offline data collection mechanism

Example: bandit algorithms, adaptive experiments, fixed logging policy ...



Offline data collection mechanism

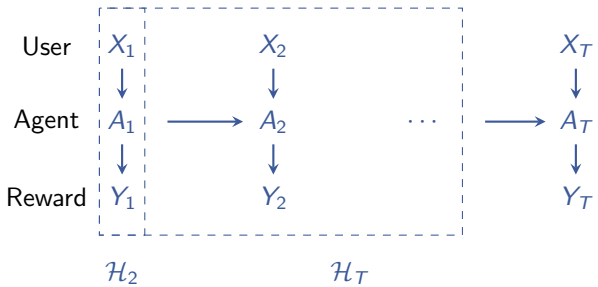
Example: bandit algorithms, adaptive experiments, fixed logging policy ...



Offline data set $\mathcal{D} = \{(X_t, A_t, Y_t)\}_{t=1}^{T}$

Offline data collection mechanism

Example: bandit algorithms, adaptive experiments, fixed logging policy ...



Offline data set $\mathcal{D} = \{(X_t, A_t, Y_t)\}_{t=1}^{T}$

$$X_t \overset{\text{i.i.d.}}{\sim} P_X, \qquad \mathbb{P}(A_t = a \,|\, X_t = x, \mathcal{H}_t) = e_t(x, a \,|\, \mathcal{H}_t), \qquad Y_t = \mu(X_t, A_t) + \epsilon_t$$

known behavior policy

Offline data collection mechanism

$$X_t \overset{\text{i.i.d.}}{\sim} P_X, \qquad \mathbb{P}(A_t = a \,|\, X_t = x, \mathcal{H}_t) = e_t(x, a \,|\, \mathcal{H}_t), \qquad Y_t = \mu(X_t, A_t) + \epsilon_t$$

known behavior policy

- **Batched data**: behavior policy is fixed
  - Logged A/B testing, clinical trials ...
  - $e_t(x, a \,|\, \mathcal{H}_t) \equiv e(x, a) \quad \Rightarrow \quad \{(X_t, A_t, Y_t)\}_{t=1}^{T}$ are i.i.d.

Offline data collection mechanism

$$X_t \overset{\text{i.i.d.}}{\sim} P_X, \qquad \mathbb{P}(A_t = a \,|\, X_t = x, \mathcal{H}_t) = e_t(x, a \,|\, \mathcal{H}_t), \qquad Y_t = \mu(X_t, A_t) + \epsilon_t$$

known behavior policy

- **Batched data**: behavior policy is fixed
  - Logged A/B testing, clinical trials ...
  - $e_t(x, a \,|\, \mathcal{H}_t) \equiv e(x, a) \quad \Rightarrow \quad \{(X_t, A_t, Y_t)\}_{t=1}^T$ are i.i.d.
- **Adaptively collected data**: behavior policy depends on previous observations
  - Adaptive experiments, operation systems using bandit algorithms ...
  - Observations are mutually dependent due to adaptivity in A_t
  - $e_t(x, a \,|\, \mathcal{H}_t)$ may diminish to zero for certain actions

Policy learning with offline data

- Goal: learn an optimal decision rule (policy) using the offline data
  - A policy is a mapping from contexts to actions $\pi : \mathcal{X} \to \mathcal{A}$

Policy learning with offline data

- Goal: learn an optimal decision rule (policy) using the offline data
  - A policy is a mapping from contexts to actions $\pi : \mathcal{X} \to \mathcal{A}$
- Given a policy class Π, the optimal policy (within this class) is

$$\pi^*(\Pi) = \underset{\pi \in \Pi}{\operatorname{argmax}} \; Q(\pi), \quad Q(\pi) = \mathbb{E}\big[\mu(X, \pi(X))\big]$$

Policy learning with offline data

- Goal: learn an optimal decision rule (policy) using the offline data
  - A policy is a mapping from contexts to actions $\pi\colon \mathcal{X} \to \mathcal{A}$
- Given a policy class Π, the optimal policy (within this class) is

$$\pi^*(\Pi) = \underset{\pi\in\Pi}{\operatorname{argmax}}\ Q(\pi), \quad Q(\pi) = \mathbb{E}\big[\mu(X, \pi(X))\big]$$

- Learn a policy $\widehat{\pi}$ using $\mathcal{D}$ with small suboptimality (regret)

$$\mathcal{L}(\widehat{\pi}; \Pi) = Q(\pi^*(\Pi)) - Q(\widehat{\pi})$$

  - Why comparing to a policy class? interpretability (decision trees, linear rules)
  - Model $\mu(x, a)$ within a class? model misspecification

A typical two-step greedy procedure

(i) Evaluation: (A)IPW $\widehat{Q}(\pi)$ (ii) Optimization pick greedy policy $\widehat{\pi}$ that maximizes $\widehat{Q}(\pi)$

A typical two-step greedy procedure

(i) Evaluation: (A)IPW $\widehat{Q}(\pi)$ (ii) Optimization pick greedy policy $\widehat{\pi}$ that maximizes $\widehat{Q}(\pi)$

- Typically impose the (uniform) overlap assumption:

$$\text{(batched)} \quad \inf_{x,a} e(x,a) \geq \eta > 0 \quad \text{or} \quad \text{(adaptive)} \quad \inf_{x,a} e_t(x,a \,|\, \mathcal{H}_t) \geq g_t, \ \forall t$$

A typical two-step greedy procedure

(i) Evaluation: (A)IPW $\widehat{Q}(\pi)$ (ii) Optimization pick greedy policy $\widehat{\pi}$ that maximizes $\widehat{Q}(\pi)$

Prior art [an incomplete list]

batched & overlap & 2 actions	batched & overlap & mult. actions	adaptive & overlap & mult. actions
Zhang, et al. '12	Swaminathan and Joachims '15	Zhan, Ren, Zhou, and Athey '21
Zhao, et al. '15	Zhou, et al. '17	Bibaut, et al. '21'
Kitagawa and Tetenov '18	Kallus '18	
Athey and Wager '21	Zhou, Athey and Wager '22	
...	...	

- **Greedy + uniform overlap is the standard!**

The uniform overlap condition?

$$\text{(batched)} \quad \inf_{x,a} e(x,a) \geq \eta > 0 \quad \text{or} \quad \text{(adaptive)} \quad \inf_{x,a} e_t(x,a \,|\, \mathcal{H}_t) \geq g_t, \ \ \forall t$$

- Batched data: imbalanced budget, costly arms..?
  - Mathematically, $\eta > 0$ always exists if $\mathcal{X}$ is compact and $e(x,a) > 0$ for all (x,a)
  - But this constant usually enters the suboptimality bound $\sim \sqrt{\frac{\sigma^2}{\eta \cdot T}}$ (very small η?)

The uniform overlap condition?

$$\text{(batched)} \quad \inf_{x,a} e(x,a) \geq \eta > 0 \quad \text{or} \quad \text{(adaptive)} \quad \inf_{x,a} e_t(x,a \,|\, \mathcal{H}_t) \geq g_t, \ \forall t$$

- Adaptively collected data: gradually tuned policy, eventually 'discarded' arms..?
  - Existing results impose $e_t(x,a \,|\, \mathcal{H}_t) \geq t^{-\beta}$ for all (x,a) values, a.s.
  - Violated if using a standard Thompson sampling [Zhan, et al., 2021]

The uniform overlap condition?

$$\text{(batched)} \quad \inf_{x,a} e(x,a) \geq \eta > 0 \quad \text{or} \quad \text{(adaptive)} \quad \inf_{x,a} e_t(x,a \,|\, \mathcal{H}_t) \geq g_t, \ \forall t$$

- Batched data: imbalanced budget, costly arms ...
- Adaptively collected data: gradually tuned policy, eventually 'discarded' arms ...

- Policy learning seems 'ruined' by actions that are rarely taken?

Is uniform overlap essential for efficient policy learning?

The uniform overlap condition?

$$\text{(batched)} \quad \inf_{x,a} e(x,a) \geq \eta > 0 \quad \text{or} \quad \text{(adaptive)} \quad \inf_{x,a} e_t(x,a \,|\, \mathcal{H}_t) \geq g_t, \ \forall t$$

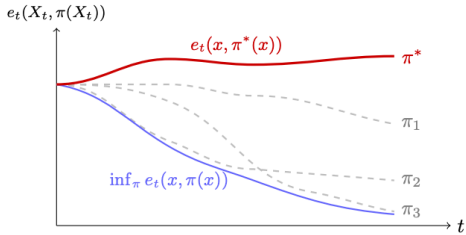
- Batched data: imbalanced budget, costly arms..?
- Adaptively collected data: gradually tuned policy, eventually 'discarded' arms..?

- Policy learning seems 'ruined' by actions that are rarely taken?

Is uniform overlap essential for efficient policy learning? Not always!

This work

- Algorithm: a pessimistic approach to policy learning with known behavior policy
- Guarantee: suboptimality bound only depends on the optimal policy



- Theory: new analytic tools for policy evaluation (generalized empirical Bernstein's ineq)

Part I: why (uniform) overlap is so important (for greedy learner)

Why overlap is important

- In the first step, one estimates $Q(\pi)$ by unbiased IPW estimators

$$\widehat{Q}(\pi) = \frac{1}{T}\sum_{t=1}^{T} \frac{\mathbb{1}\{A_t = \pi(X_t)\}}{e_t(X_t, \pi(X_t)) \,|\, \mathcal{H}_t)} Y_t$$

- If $e_t(X_t, \pi(X_t)) \,|\, \mathcal{H}_t)$ is small, then the variance of $\widehat{Q}(\pi)$ is large, i.e., estimation is inaccurate
- Estimation quality of $\widehat{Q}(\pi)$ can vary with π if $e_t(X_t, \pi(X_t) \,|\, \mathcal{H}_t)$ are not uniformly large

Why overlap is important

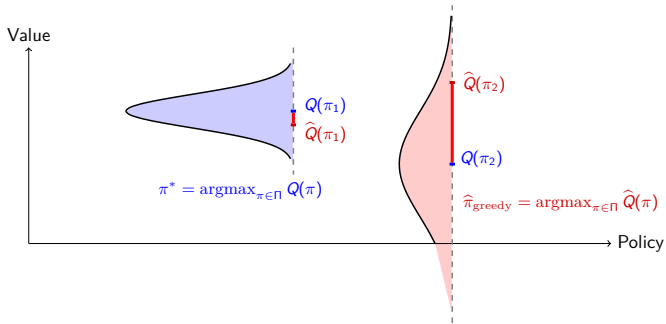
- In the first step, one estimates $Q(\pi)$ by unbiased IPW estimators

$$\widehat{Q}(\pi) = \frac{1}{T}\sum_{t=1}^{T} \frac{\mathbb{1}\{A_t = \pi(X_t)\}}{e_t(X_t, \pi(X_t)) \,|\, \mathcal{H}_t)} Y_t$$

- If $e_t(X_t, \pi(X_t)) \,|\, \mathcal{H}_t)$ is small, then the variance of $\widehat{Q}(\pi)$ is large, i.e., estimation is inaccurate
- Estimation quality of $\widehat{Q}(\pi)$ can vary with π if $e_t(X_t, \pi(X_t) \,|\, \mathcal{H}_t)$ are not uniformly large
- Policy learning/evaluation is a counterfactual learning problem
  - What would a random individual behave had they received actions under π?
  - Learning the distribution induced by π via observed distributions (under behav. policies)
  - The ***overlap*** $e_t(X_t, \pi(X_t) \,|\, \mathcal{H}_t)$ captures discrepancy between two distributions

Uniform overlap is essential for the greedy approach

- An example: no context, 2 actions
  - $\pi_1 = a_1$, $\pi_2 = a_2$ with mean $\mu_1 > \mu_2$, so $\pi^* = \pi_1$
  - Taking a_1 many (T_1) times, but a_2 few (T_2) times
  - Estimators $\widehat{Q}(\pi_1) = \frac{1}{T_1}\sum_{t=1}^{T} Y_t \mathbb{1}\{A_t = 1\}$, $\widehat{Q}(\pi_2) = \frac{1}{T_2}\sum_{t=1}^{T} Y_t \mathbb{1}\{A_t = 2\}$
  - Return $\widehat{\pi}_{\text{greedy}} = \text{argmax}_{\pi\in\Pi}\, \widehat{Q}(\pi)$



Uniform overlap is essential for the greedy approach

- Decomposing the suboptimality $\mathcal{L}(\widehat{\pi}_{\text{greedy}})$

$$Q(\pi^*) - Q(\widehat{\pi}) = \underbrace{Q(\pi^*) - \widehat{Q}(\pi^*)}_{\text{(i) intrinsic uncertainty}} + \underbrace{\widehat{Q}(\pi^*) - \widehat{Q}(\widehat{\pi})}_{\text{(ii) optimization error}} + \underbrace{\widehat{Q}(\widehat{\pi}) - Q(\widehat{\pi})}_{\text{(iii) greedy uncertainty}}$$

- (iii) is the most difficult because bad policies may appear good just by chance

Uniform overlap is essential for the greedy approach

- Decomposing the suboptimality $\mathcal{L}(\widehat{\pi}_{\text{greedy}})$

$$Q(\pi^*) - Q(\widehat{\pi}) = \underbrace{Q(\pi^*) - \widehat{Q}(\pi^*)}_{\text{(i) intrinsic uncertainty}} + \underbrace{\widehat{Q}(\pi^*) - \widehat{Q}(\widehat{\pi})}_{\text{(ii) optimization error}} + \underbrace{\widehat{Q}(\widehat{\pi}) - Q(\widehat{\pi})}_{\text{(iii) greedy uncertainty}}$$

- (iii) is the most difficult because bad policies may appear good just by chance

- Essentially need $\sup_{\pi \in \Pi} |\widehat{Q}(\pi) - Q(\pi)|$ to be small because $\widehat{\pi}$ can be anything
  - Need $e(X_t, \pi(X_t) \,|\, \mathcal{H}_t)$ to be sufficiently large for all $\pi \in \Pi$
  - Uniform overlap $\inf_{x,a} e(x, a \,|\, \mathcal{H}_t)$ when Π is expressive

yes.. uniform overlap is essential for greedy learner

Let's not be greedy!

Part II: An algorithmic change by the pessimism principle

Pessimistic policy learning

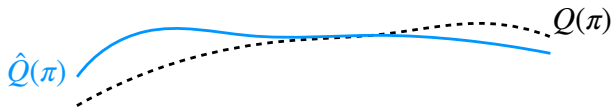
- Suppose we can quantify the uncertainty in $\widehat{Q}(\pi)$ via some $R(\pi)$, so that

$$\mathbb{P}\Big(\big|\widehat{Q}(\pi) - Q(\pi)\big| \leq R(\pi), \quad \forall \pi \in \Pi\Big) \geq 1 - \delta$$

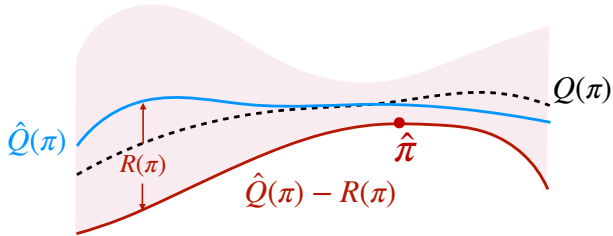
- Pessimism: optimizing **LCBs**, rather than point estimates

$$\widehat{\pi} = \underset{\pi \in \Pi}{\operatorname{argmax}} \ \big\{\widehat{Q}(\pi) - R(\pi)\big\}$$

Pessimistic policy learning



Pessimistic policy learning



- High estimated value, high uncertainty 😟
- Low estimated value, high uncertainty 😟
- High estimated value, low uncertainty 😄
- Good estimated value, low uncertainty 😄

Pessimistic policy learning: upper bounds only depend on π^*

Theorem (Informal, (J., Ren, Yang, and Wang, 2022))

On the event that $|\widehat{Q}(\pi) - Q(\pi)| \leq R(\pi)$ *for all* $\pi \in \Pi$*, we have*

$$\mathcal{L}(\widehat{\pi}) \leq 2R(\pi^*).$$

- Suboptimality is small as long as π^* is accurately estimated!

Pessimistic policy learning: upper bounds only depend on π^*

Theorem (Informal, (J., Ren, Yang, and Wang, 2022))

On the event that $|\widehat{Q}(\pi) - Q(\pi)| \leq R(\pi)$ *for all* $\pi \in \Pi$*, we have*

$$\mathcal{L}(\widehat{\pi}) \leq 2R(\pi^*).$$

- Suboptimality is small as long as π^* is accurately estimated!
- A simple proof
  - The uniform concentration ensures $Q(\pi^*) \leq \widehat{Q}(\pi^*) + R(\pi^*)$ and $Q(\widehat{\pi}) \leq \widehat{Q}(\widehat{\pi}) + R(\widehat{\pi})$
  - Optimality of $\widehat{\pi}$ ensures $\widehat{Q}(\widehat{\pi}) - R(\widehat{\pi}) \geq \widehat{Q}(\pi^*) - R(\pi^*)$
  - Collectively, we have

$$Q(\widehat{\pi}) \geq \widehat{Q}(\widehat{\pi}) - R(\widehat{\pi}) \geq \widehat{Q}(\pi^*) - R(\pi^*) \geq Q(\pi^*) - 2R(\pi^*)$$

Pessimistic policy learning: compared with greedy

- The greedy approach ensures

$$\mathcal{L}(\widehat{\pi}) \leq 2R \quad \text{if} \quad \left|\widehat{Q}(\pi) - Q(\pi)\right| \leq R, \quad \forall \pi \in \Pi$$

- Our pessimistic approach ensures

$$\mathcal{L}(\widehat{\pi}) \leq 2R(\pi^*) \quad \text{if} \quad \left|\widehat{Q}(\pi) - Q(\pi)\right| \leq R(\pi), \quad \forall \pi \in \Pi$$

Pessimistic policy learning: compared with greedy

- The greedy approach ensures

$$\mathcal{L}(\widehat{\pi}) \leq 2R \quad \text{if} \quad \left|\widehat{Q}(\pi) - Q(\pi)\right| \leq R, \quad \forall \pi \in \Pi$$

- Our pessimistic approach ensures

$$\mathcal{L}(\widehat{\pi}) \leq 2R(\pi^*) \quad \text{if} \quad \left|\widehat{Q}(\pi) - Q(\pi)\right| \leq R(\pi), \quad \forall \pi \in \Pi$$

- Greedy optimizes an uninformative LCB:

$$\widehat{\pi}_{\text{greedy}} = \underset{\pi \in \Pi}{\operatorname{argmax}} \left\{\widehat{Q}(\pi) - R\right\} = \underset{\pi \in \Pi}{\operatorname{argmax}} \widehat{Q}(\pi)$$

  and do not need to incorporate such R into optimization

Construction of the LCBs: point estimate

- AIPW point estimator

$$\widehat{Q}(\pi) = \frac{1}{T}\sum_{t=1}^{T} \widehat{\Gamma}_t(\pi), \quad \widehat{\Gamma}_t(\pi) = \widehat{\mu}_t(X_t, \pi(X_t)) + \frac{\mathbb{1}\{A_t = \pi(X_t)\}}{e_t(X_t, \pi(X_t) \,|\, \mathcal{H}_t)} \cdot \big(Y_t - \widehat{\mu}_t(X_t, \pi(X_t))\big),$$

  where $\widehat{\mu}_t(\cdot, \cdot)$ is an estimator for $\mu(\cdot, \cdot)$ constructed using $\mathcal{H}_t$

  - Unbiased and consistent for $Q(\pi)$ for a fixed policy π
  - Standard estimator in causal inference and policy evaluation for adaptively collected data [Athey and Wager (2021); Zhou et al. (2022); Zhan et al. (2021)]

- In this talk, for simplicity we set $\widehat{\mu}_t(\cdot, \cdot) \equiv 0$

Construction of the LCBs: uncertainty bound

- Point estimate $\widehat{Q}(\pi) = \frac{1}{T}\sum_{t=1}^{T} \widehat{\Gamma}_t(\pi)$, where $\widehat{\Gamma}_t(\pi) = \frac{\mathbb{1}\{A_t = \pi(X_t)\}}{e_t(X_t, \pi(X_t)\,|\,\mathcal{H}_t)} \cdot Y_t$

- Idea for $R(\pi)$: approximating the "variance" of $\widehat{Q}(\pi)$

$$R(\pi) \approx \left\{ \frac{1}{T^2} \sum_{t=1}^{T} \widehat{\mathrm{Var}}(\widehat{\Gamma}_t) \right\}^{1/2} ?$$

Construction of the LCBs: uncertainty bound

- Construct $R(\pi) = \beta \cdot V(\pi)$ for a scaling constant $\beta > 0$ to be decided later, and

$$V(\pi) := \max\left\{ V_{\mathrm{s}}(\pi), V_{\mathrm{p}}(\pi), V_{\mathrm{h}}(\pi) \right\}$$

for three data-dependent components

$$V_{\mathrm{s}}(\pi) = \frac{1}{T}\left(\sum_{t=1}^{T} \frac{\mathbb{1}\{A_t = \pi(X_t)\}}{e(X_t, \pi(X_t) \,|\, \mathcal{H}_t)^2} \right)^{1/2} \lesssim \frac{1}{T}\left\{ \sum_{t=1}^{T} \mathrm{Var}\left(\widehat{\Gamma}_t(\pi) \,|\, \mathcal{H}_t, X_t, A_t \right) \right\}^{1/2}$$

$$V_{\mathrm{p}}(\pi) = \frac{1}{T}\left(\sum_{t=1}^{T} \frac{1}{e(X_t, \pi(X_t) \,|\, \mathcal{H}_t)} \right)^{1/2} \lesssim \frac{1}{T}\left\{ \sum_{t=1}^{T} \mathrm{Var}\left(\widehat{\Gamma}(\pi) \,|\, X_t, \mathcal{H}_t \right) \right\}^{1/2}$$

$$V_{\mathrm{h}}(\pi) = \frac{1}{T}\left(\sum_{t=1}^{T} \frac{1}{e(X_t, \pi(X_t) \,|\, \mathcal{H}_t)^3} \right)^{1/4}$$ higher-order error

Also set $V(\pi) = \infty$ if $e(X_t, \pi(X_t) \,|\, \mathcal{H}_t) = 0$ for any t

- $R(\pi)$ is small if $e(X_t, \pi(X_t) \,|\, \mathcal{H}_t)$ is large, i.e., the overlap of the behavior policy for π is large

Pessimism achieves suboptimality that only depends on π^*

The price is more theory

Part III: Generalized empirical Bernstein's inequality

$$\left|\widehat{Q}(\pi) - Q(\pi)\right| \lesssim \text{Complexity}(\Pi) \cdot V(\pi)$$

Batched data

Theorem (J., Ren, Yang, and Wang (2022))

Fix $\delta \in (0,1)$. Let K be the number of actions. For batched data, set

$$R(\pi) = \beta \cdot V(\pi), \qquad \text{for } \beta \geq 10\sqrt{2(\text{N-dim}(\Pi)\log(TK^2) + \log(16/\delta))}.$$

Then with probability at least $1-\delta$, it holds that

$$\left|\widehat{Q}(\pi) - Q(\pi)\right| \leq \beta \cdot V(\pi) \quad \text{for all } \pi \in \Pi.$$

- N-dim(Π) is the Natarajan dimension, a measure of complexity of Π [Zhou et al., 2021; Zhan et al., 2021]
- Bounds on N-dim(Π) are available for linear classes, decision trees, neural networks [Jin, 2021]

Batched data: data-dependent upper bound

Theorem (Continued, J., Ren, Yang, and Wang (2022))

With the previous choice of β, we know with probability at least $1-\delta$,

$$(i)\quad \left|\widehat{Q}(\pi) - Q(\pi)\right| \leq \beta \cdot V(\pi) \quad \textit{for all } \pi \in \Pi, \quad \textit{and} \quad (ii)\quad \mathcal{L}(\widehat{\pi}) \leq 2\beta \cdot V(\pi^*)$$

- **'Without'** overlap: other than that $e(x, \pi^*(x)) > 0$
- The uniform concentration (i) does not need any lower bound on $e(x, \pi(x))$
- Fully **data-dependent**: small bound if the data 'estimates' π^* well, large bound if not (natural as we don't control the data collection process)

Batched data: explicit results

Corollary (J., Ren, Yang, and Wang (2022))

Fix any $\delta \in (0, \exp(-1))$. *Assume there exists some* $C_* > 0$ *such that* $e(x, \pi^*(x)) \geq C_*$ *for* $\mathbb{P}_X$*-almost all* $x \in \mathcal{X}$. *Choose* β *as before. Then with probability at least* $1 - 2\delta$,

$$\mathcal{L}(\widehat{\pi}) \leq \min\left\{2c \cdot \sqrt{\frac{\text{N-dim}(\Pi)\log(TK^2)\{\log(2/\delta)\}^3}{C_* T}}, 1\right\},$$

- The upper bound only depends on how well the optimal policy is explored
- Non-optimal arms can be explored arbitrarily badly and they do not enter the error bound
- $O(1/\sqrt{T})$ learning rate with weaker conditions than the literature [Kitagawa and Tetenov, 2018]
- This is the best effort given the offline data (see paper for matching minimax lower bound)

Adaptively collected data

- More challenging due to (1) dependence and (2) diminishing propensity
- Prior arts [Zhan et al., 2021; Bibaut et al., 2021]: impose $\inf_{x,a} e_t(x, a \mid \mathcal{H}_t) \geq t^{-\beta}$ for all t
  - Need this lower bound to establish concentration of $\widehat{Q}(\pi)$
  - And then the suboptimality bound
- If the uniform overlap does not hold,
  - Can we still **characterize the concentration** of $\widehat{Q}(\pi)$?
  - Can we still **efficiently learn an optimal policy** (when possible)?

Adaptively collected data: empirical Bernstein's inequality

- Suppose $\log \inf_{x,a} e_t(x, a \mid \mathcal{H}_t) \geq -(\log T)^{\alpha}$ for some $\alpha > 0$
  - $\inf_{x,a} e_t(x, a \mid \mathcal{H}_t) \geq T^{-\log T}$ if $\alpha = 2$
  - Essentially no constraint
  - Can be dropped, use a more complex bound $V(\pi) \log V(\pi)$

Adaptively collected data: data-dependent upper bound

Theorem (J., Ren, Yang, and Wang (2022))

Fix any $\delta \in (0, \exp(-1))$. *Suppose* $\log \inf_{x,a} e_t(x, a \,|\, \mathcal{H}_t) \geq -(\log T)^{\alpha}$ *for some* $\alpha > 0$. *Set*

$$R(\pi) = \beta \cdot V(\pi), \quad \text{for } \beta \geq 67 \cdot (\log T)^{\alpha/2} \cdot \sqrt{\text{N-dim}(\Pi) \log(TK^2) + \log(16/\delta)}.$$

Then with probability at least $1 - \delta$, *it holds that*

$$\left|\widehat{Q}(\pi) - Q(\pi)\right| \leq \beta \cdot V(\pi), \quad \text{for all } \pi \in \Pi.$$

Adaptively collected data: data-dependent upper bound

Theorem (Continued, J., Ren, Yang, and Wang (2022))

With the previous choice of β, we know with probability at least $1-\delta$,

$$(i)\quad \left|\widehat{Q}(\pi) - Q(\pi)\right| \leq \beta \cdot V(\pi) \quad \textit{for all } \pi \in \Pi, \quad \textit{and} \quad (ii)\quad \mathcal{L}(\widehat{\pi}) \leq 2\beta \cdot V(\pi^*)$$

- (i) Concentration (essentially) **without** boundedness condition
- Fully **data-dependent**: small bound if the data 'estimates' π^* well, large bound if not
- Non-optimal policies can be explored arbitrarily badly, and they do not enter the upper bound

Adaptively collected data: polynomial-decay case

Corollary (J., Ren, Yang, and Wang, 2022)

Suppose $e_t(X, \pi^*(X) \mid \mathcal{H}_t) \geq \bar{c} \cdot t^{-\gamma}$ *almost surely for some constants* $\bar{c}, \gamma > 0$. *Set* β *as before. Then with probability at least* $1 - 2\delta$,

$$\mathcal{L}(\widehat{\pi}) \leq \min\left\{12c \cdot \sqrt{\frac{\text{N-dim}(\Pi)}{T^{1-\gamma}}} \cdot \frac{(\log(TK^2))^{(1+\alpha)/2} \cdot \log(1/\delta)}{\max\{1, \bar{c}^{3/4}\}}, 1\right\}.$$

- See paper for a matching minimax lower bound

Adaptively collected data: polynomial-decay case

- Result in the literature [Zhan et al., 2021]

$$\mathcal{L}(\widehat{\pi}) \lesssim \sqrt{\frac{\text{N-dim}(\Pi)}{T^{1-\beta}}} \quad \text{if} \quad (A)\ \inf_{x,a} e_t(x,a \,|\, \mathcal{H}_t) \geq t^{-\beta} \quad \text{for all } t$$

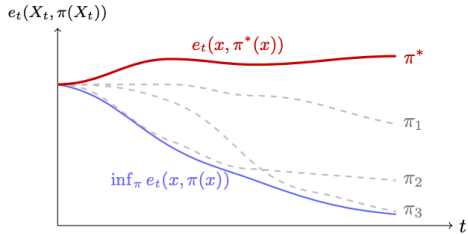
- We show

$$\mathcal{L}(\widehat{\pi}) \lesssim \sqrt{\frac{\text{N-dim}(\Pi)}{T^{1-\gamma}}} \quad \text{if} \quad (B)\ \inf_{x} e_t(x,\pi^*(x) \,|\, \mathcal{H}_t) \geq \overline{c} \cdot t^{-\gamma} \quad \text{for all } t$$

  - If (A) does not hold, then (B) may still hold
  - If (A) holds, then (B) holds with $\gamma \geq \beta$
  - Our upper bound $V(\pi^*)$ holds even when (B) fails

- Policy learning is not necessarily 'ruined' by actions with diminishing propensities (as long as they are not optimal)

Adaptively collected data: the oracle property



- **There are many more scenarios where efficient policy learning is feasible!**

Summary

- A new algorithm that optimizes LCBs
  - Being pessimistic is sometimes good :)
- Concentration inequality for policy evaluation without bounded propensities
  - Data is honest, and our estimation is honest
- Efficient policy learning even when uniform overlap does not hold
  - It's the overlap for the optimal policy that matters (minimax)!

Discussion: the pessimism principle

- In online bandits, people use optimism to guide exploration
- For offline data, we should instead use pessimism to adapt to the uncertainty in the data
  - Exploitation, as we cannot interact with the environment

Discussion: the pessimism principle

- The pessimism principle was initially proposed in offline RL [Jin et al., 2021]
  - **CATE**-based: take $\widehat{\pi}(x) = \operatorname{argmax}_{a \in \mathcal{A}} \{\widehat{\mu}(x, a) - \widehat{\Gamma}(x, a)\}$ assuming a well-specified model
  - Studied in finite-$\mathcal{X}$, linear models, and general function approximations [Zanette et al., 2021; Xie et al., 2021a; Chen and Jiang, 2022; Rashidinejad et al., 2021; Yin and Wang, 2021; Shi et al., 2022; Xie et al., 2021b]
  - Our work is a 'design-based' version that does not require modeling assumptions
- Such ideas also appear in regularized ERM [Maurer and Pontil, 2009]
  - Take $\widehat{f} = \operatorname{argmin}_{f \in \mathcal{F}} \{\widehat{L}(f) + \widehat{V}(f)\}$, where $\widehat{V}(f)$ is an estimate of the uncertainty in $\widehat{L}(f)$
  - Rely on empirical Bernstein's inequality [Bartlett et al., 2005, 2006] that $|\widehat{L}(f) - L(f)| \leq \widehat{V}(f)$
  - Previous works only work for **bounded** loss
  - Our work is a generalized empirical Bernstein's inequality for unbounded weights

Discussion: algorithms & extensions

- This work is more of a theoretical investigation
- Pessimism may be naturally compatible with the tree search algorithm in policy learning
- In general, optimization may be challenging (similar to regularized ERM)
  - Efficient implementation via equivalent formulations? [Duchi and Namkoong, 2016]
- Extensions?
  - From bandits to sequential settings
  - Other reweighting-based methods

Thank you!

Manuscript on arXiv:2212.09900