Policy Learning "without" Overlap: Pessimism and Generalized Empirical Bernstein's Inequality

Ying Jin, Zhimei Ren. Zhuoran Yang, and Zhaoran Wang

April 12, 2024

Presented by Mian Wei

Multi-Arm Bandits

Goal: find the slot machine that tends to give the highest reward, and keep playing it to maximise profit.

Strategies

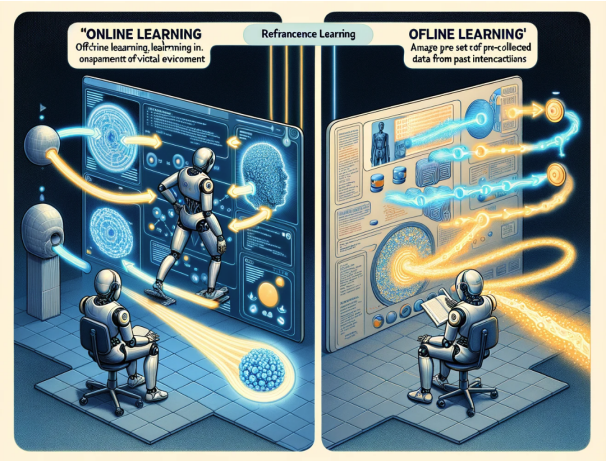
Explore-Then-Commit (ETC): play each machine m times, then pick the machine with the highest average reward and play it exclusively.
Question: how to choose m?

Upper Confidence Bound (UCB): taking into account both how close the estimates are to being maximal and the uncertainties in those estimates. [SB18]

$$\mathrm{UCB}_i(t-1,\delta) = \begin{cases} \infty & \text{if } T_i(t-1) = 0 \\ \hat{\mu}_i(t-1) + \sqrt{\frac{2\log(1/\delta)}{T_i(t-1)}} & \text{otherwise}\,. \end{cases}$$

Based on the optimism in the face of uncertainty principle, overestimate the unknown mean.

Online Learning v.s. Offline Learning

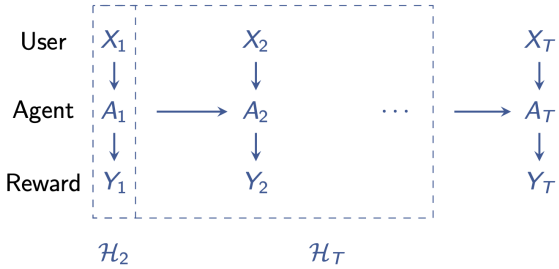


interaction v.s. no interaction

costly and risky v.s. pre-collected data

Offline data collection mechanism

Contextual Bandit Model



Offline data set $\mathcal{D} = \{(X_t, A_t, Y_t)\}_{t=1}^{T}$;
$X_t \overset{\text{iid}}{\sim} P_X$, $\mathbb{P}(A_t = a | X_t = x, \mathcal{H}_t) = e_t(x, a|\mathcal{H}_t)$, $Y_t = \mu(X_t, A_t) + \epsilon_t$.

Offline data collection mechanism

Batched data: behavior policy is fixed

- Logged A/B testing. clinical trials...
- $e_t(x, a|\mathcal{H}_t) \equiv e(x, a)$

Adaptively collected data: behavior policy depends on previous observations

- Observations are mutually dependent due to adaptivity in A_t
- $e_t(x, a|\mathcal{H}_t)$ may diminish to 0 for certain actions

Policy learning with offline data

Goal: learn an optimal policy using the offline data.

Some notations:

- Policy: a deterministic mapping from contexts to actions $\pi : \mathcal{X} \to \mathcal{A}$
- Policy value
$$Q(\pi) = \mathbb{E}[\mu(X, \pi(X))]$$
- Given a policy class Π, the optimal policy is
$$\pi^*(\Pi) = \arg\max_{\pi \in \Pi} Q(\pi)$$
- Learn a policy $\hat{\pi}$ using $\mathcal{D}$ with small suboptimality (regret)
$$\mathcal{L}(\hat{\pi}; \Pi) = Q(\pi^*(\Pi)) - Q(\hat{\pi})$$

Typical two-step greedy procedure

(i) IPW estimator

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

(ii) Pick greedy policy $\hat{\pi}$ that maximized $\hat{Q}(\pi)$

Need the **uniform overlap assumption**:

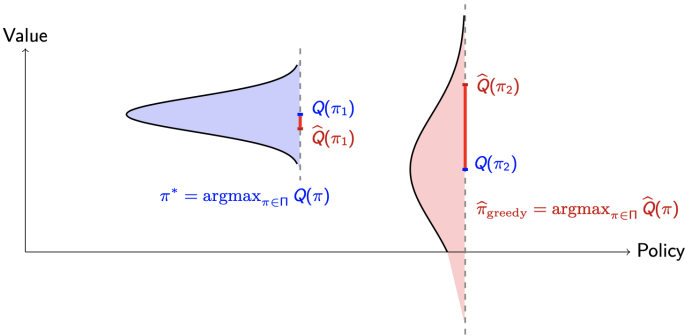
$$\inf_{x,a} e(x,a) \geq \eta > 0$$

or

$$\inf_{x,a} e_t(x,a|\mathcal{H}_t) \geq g_t, \forall t$$

Uniform overlap

Why uniform overlap is important for greedy learner?



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

Uniform overlap

Batched data:

- Mathematically, $\eta > 0$ always exists if $e(x, a) > 0$ for all (x, a)
- Suboptimality bound $\sim \sqrt{\frac{\sigma^2}{\eta \cdot T}}$

Adaptively collected data:

- 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 [ZRAZ23]

Pessimistic policy learning

Decomposing the suboptimality $\mathcal{L}(\hat{\pi})$

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

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

Essentially need $\sup_{\pi \in \Pi} |\hat{Q}(\pi) - Q(\pi)|$ **to be small**

Pessimistic policy learning

Suppose we can quantify (iii) via some $R(\pi)$, so that,

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

Pessimism: optimizing LCBs

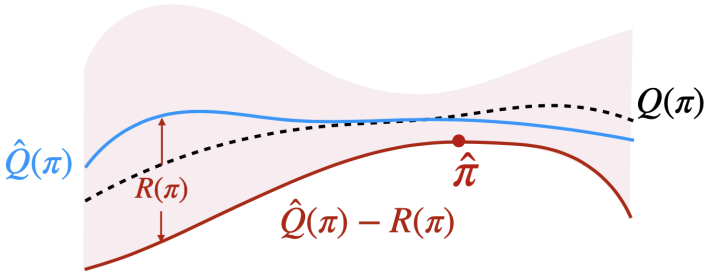
$$\hat{\pi} = \arg\max_{\pi \in \Pi} \{\hat{Q}(\pi) - R(\pi)\}$$

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*

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

Pessimistic policy learning



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

Construction of the LCBs

Point estimate

$$\hat{Q}(\pi) = \frac{1}{T}\sum_{t=1}^{T}\hat{\Gamma}_t(\pi)$$

$$\hat{\Gamma}_t(\pi) = \frac{\mathbb{1}\{A_t = \pi(X_t)\}}{e_t(X_t, \pi(X_t)|\mathcal{H}_t)}Y_t$$

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

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

Construction of the LCBs

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

$$V(\pi) := \max\{V_s(\pi), V_p(\pi), V_h(\pi)\}$$

where

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

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

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

$R(\pi)$ is small if $e(X_t, \pi(X_t)|\mathcal{H}_t)$ is large.

Generalized empirical Bernstein's inequality

Informal idea:

$$|\hat{Q}(\pi) - Q(\pi)| \lesssim \text{N-dim}(\Pi) \cdot V(\pi)$$

where N-dim(Π) is the Natarajan dimension, a generalization of the well-known Vapnik-Chervonenkis (VC) dimension.

VC dimension: cardinality of the largest shattered set.
Examples:

- VC-dim(intervals) $= 2$
- VC-dim(Axis-aligned rectangles) $= 4$

To be simple, Natarajan dimension is a measure of complexity of Π

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

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

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

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

Bounds on N-dim(Π) are available for linear classes, decision trees, neural networks. [JYW21]

Batched data

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

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

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

- Does not require uniform overlap
- Fully data-dependent

Batched data

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

Batched data

Theorem

For any action set $\mathcal{A}$, sample size $T \in \mathbb{N}_+$, any policy class Π and any $C_* > 0$ with $\frac{\text{N-dim}(\Pi)}{C_* T} \leq 1.5$, one has

$$\inf_{\hat{\pi}} \sup_{(\mathcal{C},e)\in\mathcal{R}(C_*,T,\mathcal{A},\Pi)} \mathbb{E}_{\mathcal{C},e}[\mathcal{L}(\hat{\pi};\mathcal{C},\Pi)] \geq 0.12\sqrt{\frac{\text{N-dim}(\Pi)}{C_* T}}$$

where $\mathbb{E}_{\mathcal{C},e}$ means taking expectation w.r.t. the randomness of the offline data generated under $\mathcal{C}$ and $e(\cdot,\cdot)$.

Pessimistic policy leaning is minimax optimal.

Adaptively collected data

Suppose $\log \inf_{x,a} e_t(x,a|\mathcal{H}_t) \geq -(\log T)^\alpha$ for some $\alpha > 0$. If $\alpha = 2$, $\inf_{x,a} e_t(x,a|\mathcal{H}_t) \geq T^{-\log T}$.

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

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

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

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

- Fully data-dependent: small bound if the data estimates optimal policy well
- Non-optimal arms can be explored arbitrarily badly

Adaptively collected data

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* $1 - 2\delta$,

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

Adaptively collected data

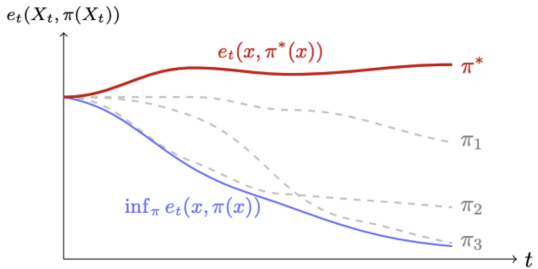
Theorem

For any action set $\mathcal{A}$, sample size $T \in \mathbb{N}_+$, any policy class Π, any $\bar{c} > 0$ and any $\gamma \in (0, 1)$ with $(1-\gamma) \cdot \text{N-dim}(\Pi) \leq 1.5\bar{c} \cdot T^{1-\gamma}$, one has

$$\inf_{\hat{\pi}} \sup_{(\mathcal{C},e)\in\mathcal{R}(\mathcal{C}_*,T,\mathcal{A},\Pi)} \mathbb{E}_{\mathcal{C},e}[\mathcal{L}(\hat{\pi};\mathcal{C},\Pi)] \geq 0.12\sqrt{\frac{\text{N-dim}(\Pi)}{T^{1-\gamma}}} \cdot \sqrt{\frac{1-\gamma}{\bar{c}}}$$

where $\mathbb{E}_{\mathcal{C},e}$ means taking expectation w.r.t. the randomness of the offline data generated under $\mathcal{C}$ and $e(\cdot,\cdot|\mathcal{H}_t)$.

Summary



- A new algorithm that optimized LCBs
- Concentration inequality for policy evaluation without bounded propensities
- Efficient policy learning even when uniform overlap does not hold, it's the overlap for the optimal policy that matters
- In online bandits, people use optimism to guide exploration; for offline data, we should use pessimism to adapt to the uncertainty in the data

Accommodation

Is is worth reading? Maybe.

- Good introductory paper on pessimism in offline learning;
- Clear take-away messages;
- Main contribution is the theoretical derivation;
- Lack of experimental results casts shadows on practical use.

Thank you! Any questions?

Ying Jin, Zhuoran Yang, and Zhaoran Wang.
Is pessimism provably efficient for offline rl?
In *International Conference on Machine Learning*, pages 5084–5096. PMLR, 2021.

Richard S. Sutton and Andrew G. Barto.
Reinforcement Learning: An Introduction.
MIT Press, 2nd edition, 2018.

Ruohan Zhan, Zhimei Ren, Susan Athey, and Zhengyuan Zhou.
Policy learning with adaptively collected data.
Management Science, 2023.