CS292F Statistical Foundation of Reinforcement Learning **Spring 2021**

Lecture 12: OPE in Bandits and in RL (May 3 / May 5)

Lecturer: Yu-Xiang Wang *Scribes: Kaiqi Zhang*

Note: *LaTeX template courtesy of UC Berkeley EECS dept.*

12.1 Contextual bandits model

Contexts: $x_1, \ldots, x_n \sim \lambda$ drawn iid, possibly infinite domain.
Actions: $a_i \sim \mu(a|x_i)$ raken by a randomized "logging" policy.
Reward: $r_i \sim D(r|x_i, a_i)$ revealed only for the action taken.
Value: $v^\mu = \mathbb{E}_{x\sim\lambda}\mathbb{E}_{a\sim\mu(\cdot|)}\mathbb{E}_D[r|x,a]$.
Collect data $(x_i, a_i, r_i)_{i=1}^n$ by above process.

ATE estimation is a special case of off-policy evaluation.

12.2 Direct method

Fit a regression model of the reward $\hat{r}(x,a)$, then for any target policy,

$$\hat{v}^\pi_{DM} = \frac{1}{n}\sum_{i=1}^{n}\sum_{a\in\mathcal{A}}\hat{r}(x_i,a)\pi(a|x_i) \tag{12.1}$$

Low variance, can evaluate on unseen contexts.
Often high bias, the model can be wrong / hard to learn.

12.3 Inverse propensity score / importance sampling

$$\hat{v}^\pi_{IPS} = \frac{1}{n}\sum_{i=1}^{n}\frac{\pi(a_i|x_i)}{\mu(a_i|x_i)}r_i \tag{12.2}$$

Pros: no assumption on rewards, unbiased, computationally efficient.
Cons: high variance when the weight is large.

12.3.1 Performance of importance sampling estimator

Mean:

$$\begin{aligned}\mathbb{E}[\hat{V}_{is}^{\pi}] &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\frac{\pi(A_i|S_i)}{\mu(A_i|S_i)}R_i\right] \\ &= \mathbb{E}_{S_1\sim\lambda}\left[\mathbb{E}_{A_1\sim\mu(\cdot|S_1)}\left[\mathbb{E}[R_1|S_1,A_1\frac{\pi(A_1|S_1)}{\mu(A_1|S_1)}\middle|S_1]\right]\right] \\ &= \mathbb{E}_{s\sim\lambda}\left[\sum_{a\in\mathcal{A}}\mu(a|S_1)\frac{\pi(a|S_1)}{\mu(a|S_1)}r(S_1,a)\right] \\ &= v^{\pi}\end{aligned} \tag{12.3}$$

Variance:

$$\begin{aligned}&\mathrm{Var}\left[\frac{1}{n}\sum_i\frac{\pi(A_i|S_i)}{\mu(A_i|S_i)}R_i\right] \\ &= \frac{1}{n}\mathrm{Var}\left[\frac{\pi(A_i|S_i)}{\mu(A_i|S_i)}R_i\right] \\ &= \frac{1}{n}\left(\mathbb{E}_S\left[\mathrm{Var}_\mu\left[\frac{\pi(A_i|S_i)}{\mu(A_i|S_i)}R_1\middle|S_i\right]\right]+\mathrm{Var}_S\left[\mathbb{E}_\mu\left[\frac{\pi(A_1|S_1)}{\mu(A_1|S_1)}R_1\middle|S_1\right]\right]\right) \\ &= \frac{1}{n}\left(\mathbb{E}_S[\mathbb{E}_{A\sim\mu(\cdot|S)}\rho^2\mathrm{Var}[R_i|S_i,A_i]]+\mathrm{Var}_{A\sum\mu(\cdot|S_1)}[\mathbb{E}[P_1R_1|S_1,A_1]]+\mathrm{Var}_{S_1}[\mathbb{E}_\pi[r(S_i,A_i)|S_i]]\right) \\ &= \frac{1}{n}\left(\mathbb{E}_\mu[\rho^2\sigma^2(S,A)]+\mathbb{E}[\mathrm{Var}_\mu[P_1r(S,A)]|S]+\mathrm{Var}_{S_1}[\mathbb{E}_\pi[r(S_i,A_i)|S_i]]\right) \\ &= \text{reward variance} + \text{logging policy variance} + \text{variance of the context}\end{aligned} \tag{12.4}$$

12.3.2 Importance sampling and direct method are surprisingly similar in some cases

MAB: importance sampling as a regression estimator:

$$\begin{aligned}\hat{V}_{IS} &= \frac{1}{n}\sum_i\frac{\pi(A_i)}{\mu(A_i)}R_i \\ &= \sum_{a\in\mathcal{A}}\pi(a)\frac{1}{n}\sum\frac{R_i\mathbf{1}(A_i=a)}{\mu(A_i)} \\ &= \sum_{a\in\mathcal{A}}\pi(a)\hat{r}_{IS}(a)\end{aligned} \tag{12.5}$$

Regression estimator as an importance sampling:

$$\begin{aligned}
\hat{V}^{\pi}_{DM} &= \sum_a \hat{r}(a)\pi(a) \\
&= \sum_a \frac{1}{N_a} \sum_i R_i \mathbf{1}(A_i = a)\pi(a) \\
&= \sum_a \frac{1}{n}\frac{1}{\hat{\mu}_a} \sum_{i=1}^{n} R_i \mathbf{1}(A_i = a)\pi(a) \\
&= \frac{1}{n} \sum_{i=1}^{n} \frac{\pi(a)}{\hat{\mu}(a)} R_i \mathbf{1}(\hat{\mu}(a) \neq 0) \\
\hat{r}(a) &= \begin{cases} \frac{1}{N_a} \sum_i R_i \mathbf{1}(A_i = a) & \text{if } N_a > 0 \\ 0 & \text{otherwise} \end{cases}
\end{aligned} \tag{12.6}$$

Comparing the MSE of DM and IS.
Bias-variance decomposition:

$$\begin{aligned}
MSE(\hat{V}) &= \mathbb{E}[(\hat{V} - V^{\pi})^2] \\
&= \mathbb{E}[(\hat{V} - \mathbb{E}[\hat{V}] + \mathbb{E}[\hat{V}] - V^{\pi})^2] \\
&= \mathbb{E}[(\hat{V} - \mathbb{E}\hat{V})^2] + \mathbb{E}(\mathbb{E}\hat{[V]} - V^{\pi})^2 + 0 \\
&= \text{var} + \text{bias}
\end{aligned} \tag{12.7}$$

Analyzing DM with plug-in estimator:

$$\begin{aligned}
\hat{V}_{DM} &= \sum_{a'} \pi(a)\hat{r}(a), \\
\mathbb{E}[\hat{V}_{DM}] &= \sum_a \pi(a)\mathbb{E}[\hat{r}(a)] \\
&= \sum_a \pi(a)\mathbb{E}\left[\frac{1}{N_a}\sum R_i \mathbf{1}(A_i = a)\mathbf{1}(N_a > 0)\right] \\
&= \sum_a \pi(a) P(N_a > 0) r(a) \\
bias(\hat{V}_{DM}) &= \mathbb{E}\hat{V}_{DM} - V^{\pi} \\
&= \sum_a \pi(a) r(a)[P(N_a > 0) - 1] \to 0 \\
\text{Var}[\sum_a \pi(a)\hat{r}(a)] &= \mathbb{E}[\text{Var}[\sum_a \pi(a)\hat{r}(a)|N_a, a \in \mathcal{A}]] + \text{Var}[\mathbb{E}[\sum_a \pi(a)\hat{r}(a)|N_a, a \in \mathcal{A}]] \\
&= \mathbb{E}[\sum_a \pi^2(a)\text{Var}(\hat{r}(a)|N_a)] + \text{Var}[\pi(a)r(a)\mathbf{1}(N_a > 0)] \\
&= \mathbb{E}[\sum_a \frac{\pi^2(a)}{N_a}\sigma^2(a)\mathbf{1}(N_a > 0)] \\
&\leq \mathbb{E}[\sum_a \frac{\pi^2(a)}{N_a + 1}\sigma^2(a)] + R^2_{\max} P(\exists a, N_a > 0) \\
&= \sum_a \frac{\pi(a)^2}{(n+1)\mu(a)}\sigma^2(a)(1 - \mu_a))
\end{aligned} \tag{12.8}$$

12.4 Weighted impartance sampling

Self normalization:

$$\begin{aligned}
\hat{V}_{WIS} &= \frac{1}{\sum_{i=1}^n \rho_i} \sum_{j=1}^n \rho_i R_i, \rho_i = \frac{\pi(A_i|S_i)}{\mu(A_i|S_i)}, \\
\mathbb{E}[\sum_{i=1}^n \rho_i] &= \mathbb{E}_\mu \sum_{i=1}^n \frac{\pi(A_i|S_i)}{\mu(A_i|S_i)} \\
&= \sum_{i=1}^n \sum_a \mu(A_i|S_i) \frac{\pi(A_i|S_i)}{\mu(A_i|S_i)} \\
&= n, \\
\hat{V}_{WIS} &\to V^\pi
\end{aligned} \tag{12.9}$$

12.5 Doubly robust estimator for OPE

Using the regression estimator as a baseline. $\hat{r}$ is given, maybe estimated from a different data.

$$\begin{aligned}
\hat{V}_{DR} &= \frac{1}{n} \sum_{i=1}^n \mathbb{E}_\pi(r(S_i, A_i)|S_i) + \rho_i(R_i - \hat{r}(S_i, A_i)) \\
\mathbb{E}[\hat{V}_{DR}] &= \mathbb{E}_{S_i \sim \lambda} \mathbb{E}_\pi(r(S_1, A_1)|S_1) + \mathbb{E}\left[\sum_a \mu(a|S_i) \frac{\pi(a|S_i)}{\mu(a|S_i)} \mathbb{E}[R|A, a] \middle| S_1\right] - \sum_a \pi(a|S_i)\hat{r}(S_i, a) \\
&= \hat{V}^\pi
\end{aligned} \tag{12.10}$$

Double robustness in model-misspecification: Unknown $\mu_i = \mu(A_i|S_i)$.

$$\hat{V}_{DR} = \frac{1}{n} \sum_i \hat{r}^\pi(S_i) + \frac{\pi(A_i|S_i)}{\hat{\mu}(S_i, A_i)}(R_i - \hat{r}(S_i, A_i)) \tag{12.11}$$

If either $\hat{\mu}$ or $\hat{r}$ is consistant (converge to correct value), then $\hat{V}_{DR} \to V^\pi$.
If either $\mathbb{E}[\hat{\mu}] = \mu$ or $\mathbb{E}[\hat{r}] = r$, then $\mathbb{E}[\hat{V}_{DR}] = V^\pi$.

Variance reduction: $\hat{V}_{DR}$ is asymptotically efficient if $(\hat{r} - r); (\frac{1}{\hat{\mu}} - \frac{1}{\mu}) = O(\frac{1}{n^{1/4}})$.

12.6 Lower bounding the minimax risk

Assume λ is a probability density, then under mild moment conditions,

$$\begin{aligned}
&\inf_{\hat{v}} \sup_{D(r|a,x) \in \mathbb{R}(\sigma^2, R_{\max})} \mathbb{E}(\hat{v} - v^\pi)^2 = \Omega[\frac{1}{n}(\mathbb{E}_\mu[\rho^2\sigma^2] + \mathbb{E}_\mu[\rho^2 R_{\max}^2])] \\
&\quad = \text{Randomness in reward + randomness due to context distribution}
\end{aligned} \tag{12.12}$$

Classical optimality theorem: any LAN estimator is greater than

$$\mathbb{E}_{x\sim\mathcal{D}}\left\{\mathbb{E}_\mu[\rho^2\mathrm{Var}[r|x,a]|x]\right\}+\mathrm{Var}_{x\sim\mathcal{D}}\left\{\mathbb{E}_\mu[\rho r|x]\right\} \tag{12.13}$$

Take supremum:

$$\mathbb{E}_\mu[\rho^2\sigma^2]+\mathbb{E}_{x\sim\mathcal{D}}[\mathbb{E}_\mu[\rho R_{\max}|x]^2]. \tag{12.14}$$

The minimax lower bound is bigger:

$$\mathbb{E}_\mu[\rho^2\sigma^2]+\mathbb{E}_\mu[\rho^2 E_{\max}^2] \tag{12.15}$$

Lower than lower bound because of additional assumption about the structure of the model.

12.7 Switch estimator

For each $i=1,\dots n, a\in\mathcal{A}$, if $\pi(a|x_i)/\mu(a|x_i)\le\tau$: use IPS or DR. Else: use regression estimator.

$$MSE(\hat{V}_{SWITCH})\le\frac{2}{n}\mathbb{E}_\mu[(\sigma^2+R_{\max}^2)\rho^2\mathbf{1}(\rho\le\tau)]+\frac{2}{n}\mathbb{E}_\pi[R_{\max}^2\mathbf{1}(\rho\le\tau)]+\mathbb{E}_\pi[\epsilon\mathbf{1}(\rho>\tau)]^2 \tag{12.16}$$

Automatic parameter tuning: $\hat{\tau}=\mathrm{argmin}\,\hat{\mathrm{Var}}_\tau$

12.8 Per-step importance sampling

Importance sampling on the entire trajectory.

$$\begin{aligned}V_{step,IS}^\pi&=\frac{1}{n}\sum_{i=1}^n(\sum_{h=1}^H r_h^{(i)})\frac{d_1\pi_1P_1\dots\pi_{H-1}P_{H-1}\pi_H}{d_1\mu_1P_1\mu_2P_2\dots\mu_{h-1}P_{H-1}\mu_H}\\&=\frac{1}{n}\sum_{i=1}^n\prod_{h=1}^H\rho_h^{(i)}\sum r_h^{(i)},\quad\rho_h^{(i)}=\frac{\pi(a_h^{(i)})}{\mu(a_h^{(i)})}\end{aligned} \tag{12.17}$$

Per-step importance sampling:

$$V_{IS}^\pi=\sum_{h=1}^H\frac{1}{n}\sum_{i=1}^n\left(\prod_{t=1}^H\rho_h^{(i)}\right)r_h^{(i)} \tag{12.18}$$

12.9 Doubly rebust OPE in reinforcement learning

Given a value function approximation:

$$\begin{aligned}V_{DR}^0&:=0\text{ for }t=1,\dots,H\\V_{DR,IS}^{H+1-t}&:=\hat{V}(s_t)+\rho_t(r_t+\gamma V_{DR}^{H-t}-\hat{Q}(s_t,a_t)).\end{aligned} \tag{12.19}$$

Doubly robust estimator is unbiased. Variance:

$$V(\hat{v}_{DR})=\sum_{h=1}^H\sum_{t=1}^h(\prod P_t)^2\mathrm{Var}[(V(s')+r)|s,a] \tag{12.20}$$

12.10 Marginalized importance sampling

Main challenge of OPE in RL: the curse of horizon

$$\hat{v}_{IS}^{\pi} = \frac{1}{n}\sum_{i=1}^{n}\sum_{h=1}^{H}\left[\frac{\pi(a_t^{(i)}|s_t^{(i)})}{\mu(a_t^{(i)}|s_t^{(i)})}\right] \tag{12.21}$$

The variance is exponential in H.

$$\hat{v}_{MIS}^{\pi} = \frac{1}{n}\sum_{i=1}^{n}\sum_{t=1}^{H}\frac{\hat{d}_t^{\pi}(s_t^{(i)})}{\hat{d}_t^{\mu}(s_t^{(i)})}\hat{r}_t^{\pi}(s_t^{(i)}) \tag{12.22}$$

Ideas for estimating the marginalized importance weight:
Idea 1: averaging over multiple visits to the same state.

$$\frac{\hat{d}_t^{\pi}(s)}{\hat{d}_t^{\mu}(s)} = \frac{1}{n}\sum_{i=1}^{n}\prod_{h=1}^{t}\rho_h^{(i)}(s_n^{(i)}, a_n^{(i)})\mathbf{1}(S_t^{(i)} = s) \tag{12.23}$$

Idea 2: recursive estimation:

$$\begin{aligned} d_t^{\pi} &= \sum_{s_{t-1}} P_t^{\pi}(s_t|s_{t-1})d_{t-1}^{\pi}(s_{t-1}), \\ \hat{d}_t^{\pi} &= \hat{P}_t^{\pi}\hat{d}_{t-1}^{\pi}, \\ \hat{P}_t^{\pi}(s_t|s_{t-1}) &= \frac{1}{n_{s_{t-1}}}\sum_{i=1}^{n}\frac{\pi(a_{t-1}^{(i)}|s_{t-1})}{\mu(a_{t-1}^{(i)}|s_{t-1})}\mathbf{1}((s_{t-1}^{(i)}, s_t^{(i)}) = (s_{t-1}, s_t)) \\ \hat{r}_t^{\pi}(s_t) &= \frac{1}{n_{s_t}}\sum_{i=1}^{n}\frac{\pi(a_t^{(i)}|s_t)}{\mu(a_t^{(i)}|s_t)}r_t^{(i)}\mathbf{1}(s_t^{(i)} = s_t) \end{aligned} \tag{12.24}$$

Result: OPE error bound of MIS: The MSE of MIS estimator obeys:

$$\frac{1}{n}\sum_{t=1}^{H}\mathbb{E}_{\mu}\left[\frac{d_t^{\pi}(s_t)^2}{d_t^{\mu}(s_t)^2}\mathrm{Var}\left[\frac{\pi_t(a_t|s_t)}{\mu_t(a_t|s_t)}(V_{t+1}^{\pi}(s_{t+1}) + r_t)\middle| s_t\right]\right] + \tilde{O}(n^{-1.5}) \tag{12.25}$$

Bound $\frac{d_t^{\pi}(s_t)^2}{d_t^{\mu}(s_t)^2} \le \tau_s$, $\frac{\pi_t(a_t|s_t)}{\mu_t(a_t|s_t)} \le \tau_a$, MSE is bounded by $\frac{H^3}{n}\tau_s^2\tau_a^2$.

Challenges of the analysis:

1. Dependent data.
Address: define an approximate martingale.
Consider the data collection in parallel.
Group all data from time h together.
Conditioning on the number of times states are visited.

2. An annoying bias: non-zero probability that some states are not visited at all.
Fictitious estimator:

$$\begin{aligned} \tilde{v}^{\pi} &:= \sum_{t=1}^{H}\sum_{s_t}\tilde{d}_t^{\pi}(s_t)\tilde{t}_t^{\pi}(s_t), \\ \tilde{d}_t^{\pi} &= \tilde{P}_{t,t-1}^{\pi}\tilde{d}_{t-1}^{\pi}, \end{aligned} \tag{12.26}$$

$$\tilde{r}_t^\pi = \begin{cases} \hat{r}_t^\pi & \text{if } n_{s_t} \geq n_d^\mu(s_t)(1-\delta) \\ r_t^\pi & \text{otherwise} \end{cases}$$
$$\tilde{P}_{t,t-1}^\pi = \begin{cases} \hat{P}_{t,t-1}^\pi & \text{if } n_{s_t} \geq n_d^\mu(s_t)(1-\delta) \\ P_{t,t-1}^\pi & \text{otherwise} \end{cases} \tag{12.27}$$

Lemma 12.1. *Multiplicative Chernoff Bound. Let X be a binomial random variable with parameter p, n. For any $\delta > 0$, we have*

$$P[X < (1-\delta)pn] \leq e^{-\frac{\delta^2 pn}{2}} \tag{12.28}$$

$$P(n_{s,t} < nd_t^\mu(s)(1-\delta)) \leq e^{-\frac{\delta^2 n d_t^\mu(s)}{2}} \tag{12.29}$$

Union bound for all t, s:

$$P(Fictitious\ estimator \neq MIS) \leq HS\exp(-\delta^2 n \min_{s,t} d_t\mu(s)),$$
$$\delta = \tilde{O}(\frac{1}{\sqrt{n}}) \tag{12.30}$$

3. Error propagation from recursive estimation.
Empirical / offline version of Bellman equation of variance.

$$\mathrm{Var}[\tilde{v}^\pi] = \sum_{h=0}^{H}\sum_{s_h}\mathbb{E}\left[\frac{\tilde{d}_h^\pi(s_h)^2}{n_{s_h}}\mathbf{1}\left(n_{s_h} \geq \frac{nd_h^\mu(s_h)}{(1-\delta)^{-1}}\right)\right]\mathrm{Var}_\mu\left[\frac{\pi(a_h^{(1)}|s_h)}{\mu(a_h^{(1)}|s_h)}(V_{h+1}^\pi(s_{h+1}^{(1)}) + r_h^{(1)})\middle| s_h^{(1)} = s_h\right] \tag{12.31}$$

Bounding error propagation.

$$\mathbb{E}[\frac{\tilde{d}_h^\pi(s_h)^2}{n_{s_h}}\mathbf{1}\left(n_{s_h} \geq \frac{nd_h^\mu(s_h)}{(1-\delta)^{-1}}\right)] \leq \frac{(1-\delta)^{-1}}{n}\left(\frac{d_h^\pi(s_h)^2}{d_h^\mu(s_h)} + \mathrm{Var}[\tilde{d}_h^\pi(s_h)]\right) \tag{12.32}$$

Bounding the variance requires to bound the covariance:

$$\mathrm{Var}[\tilde{d}_h^\pi(s_h)] \leq \frac{2(1-\delta)^{-1}hd_h^\pi(s_h)}{n} \tag{12.33}$$

MIS is optimal for finite-state / infinite action space. For fully tabular setting, it is not optimal, at least asymptotically.

12.11 Tabular MIS

$$\hat{v}_{MIS}^\pi = \frac{1}{n}\sum_{i=1}^{n}\sum_{t=1}^{H}\frac{\hat{d}_t^\pi(s_t^{(i)})}{\hat{d}_t^\mu(s_t^{(i)})}\hat{r}_t^\pi(s^{(i)}) \tag{12.34}$$

With a minor change to the following recursive estimation:

$$\hat{P}_t^\pi(s_t|s_{t-1}) = \frac{1}{n_{s_{t-1}}}\sum_{i=1}^{n}\frac{\pi(a_{t-1}^{(i)}|s_{t-1})}{\mu(a_{t-1}^{(i)}|s_{t-1})}\mathbf{1}((s_{t-1}^{(i)}, s_t^{(i)}, a_t^{(i)}) = (s_{t-1}, s_t, a_t))$$
$$\hat{r}_t^\pi(s_t) = \frac{1}{n_{s_t}}\sum_{i=1}^{n}\frac{\pi(a_t^{(i)}|s_t)}{\mu(a_t^{(i)}|s_t)}\mathbf{1}(s_t^{(i)} = s_t) \tag{12.35}$$

How to do DM in RL:
1. Estimate MDP.
2. Plug-in the target policy.
TMIS is equivalent to DM.

1. TMIS has an error that is linear in H. ($H^2\tau_a\tau_s/n$)
2. TMIS is better than MC even when we are doing on-policy evaluation.

TMIS is equivalent to DM - a model-based approcach.

$$\hat{V}_{MIS} = \frac{1}{n}\sum_{i=1}^{n}\sum_{t=1}^{H}\frac{\hat{d}^{\pi}(s_\tau^{(i)})}{d^{\mu}(s_\tau^{(i)})}\hat{r}_\tau^{\pi}(s_\tau^{(i)}) = \sum_{s\in\mathcal{S}}\sum_{t=1}^{H}\hat{d}_\tau^{\pi}(s)\hat{r}_t(s) = \hat{V}_{DM} \tag{12.36}$$