Contextual Bandits

Input: Π, $\pi \in \Pi$: $\pi: \mathcal{C} \to \mathcal{X}$ (context to action)

for $t = 1, 2, \ldots$

Nature reveals context $c_t \overset{iid}{\sim} \mathcal{D}$

Player plays $x_t \in \mathcal{X}$ (equiv. $\pi_t \in \Pi$, $x_t = \pi_t(c_t)$)

and recieves reward $r_t \in [0,1]$ w/ $\mathbb{E}[r_t | c_t, x_t] = v(c_t, x_t)$

Value of policy π defined $V(\pi) := \mathbb{E}_{c \sim \mathcal{D}}\left[v(c, \pi(c)) \right]$

Off-policy evaluation

Assume dataset collected by a logging policy where

$x_t \sim \mu(\cdot \mid c_t)$ where $p_t = \mu(x_t | c_t)$

for T steps to construct $\{(c_t, x_t, r_t, p_t)\}_{t=1}^{T}$

IPS-estimator ("Model the bias")

$$\hat{v}(c_t, x) = \frac{\mathbb{1}\{x_t = x\}}{p_t} r_t \quad \text{and} \quad \hat{V}(\pi) = \frac{1}{T} \sum_{t=1}^{T} \hat{v}(c_t, \pi(c_t))$$

We showed $\mathbb{E}[\hat{v}(c_t, x) | c_t] = v(c_t, x) \Rightarrow \mathbb{E}[\hat{V}(\pi)] = V(\pi)$

$$\mathbb{V}\mathrm{ar}(\hat{v}(c_t, x) | c_t) \leq \frac{1}{\mu(x | c_t)} \qquad \mathbb{V}\mathrm{ar}(\hat{V}(\pi)) \leq \mathbb{E}_{c \sim \mathcal{D}}\left[\frac{1}{\mu(\pi(c)|c)}\right] \cdot \frac{1}{T}$$

Bernstein's Inequality Let $X_1, \ldots, X_n$ be independent R.V. w/ $X_i \leq B$, $\mathbb{V}\mathrm{ar}(X_i) \leq \sigma^2$ Then

$$\mathbb{P}\left(\frac{1}{n} \sum X_i - \mathbb{E}[X_i] > \sqrt{\frac{2\sigma^2 \log(1/\delta)}{n}} + \frac{2B \log(1/\delta)}{3n} \right) \leq \delta$$

Recall Hoeffding says if $X_i \in [0, B]$ then

$$\mathbb{P}\left(\frac{1}{n} \sum X_i - \mathbb{E}[X_i] > \sqrt{\frac{B^2 \log(1/\delta)}{2n}} \right) \leq \delta.$$

We always have $r_t \in [0,1]$

$$\hat{V}(c_t, x) = \frac{\mathbb{1}\{x_t = x\}}{p_t} r_t \leq \frac{1}{p_t} \leq \max_{c,x} \frac{1}{\mu(x|c)}$$

$\Rightarrow$ For any fixed $\pi \in \Pi$, w.p. $\geq 1-\delta$.

$$\hat{V}(\pi) - V(\pi) > \sqrt{\mathbb{E}_{c\sim\mathcal{D}}\left[\frac{1}{\mu(\pi(c)|c)}\right] 2\log(1/\delta)\frac{1}{T}} + \frac{2\log(1/\delta)}{3T} \max_{c,x} \frac{1}{\mu(x|c)}$$

$\Rightarrow$ w.p. $\geq 1-\delta$, for all $\pi \in \Pi$ simultaneously

we have

$$|\hat{V}(\pi) - V(\pi)| \leq \sqrt{\mathbb{E}_{c\sim\mathcal{D}}\left[\frac{1}{\mu(\pi(c)|c)}\right] 2\log(2|\Pi|/\delta)\frac{1}{T}} + \max_{x,c} \frac{1}{\mu(x|c)} \cdot \frac{2\log(2|\Pi|/\delta)}{3T}$$

If $\mu(x|c) = \frac{1}{|\mathcal{X}|}$

$$|\hat{V}(\pi) - V(\pi)| \leq \sqrt{\frac{2|\mathcal{X}|\log(2|\Pi|/\delta)}{T}} + \frac{2|\mathcal{X}|\log(2|\Pi|/\delta)}{3T}$$

$$\leq \sqrt{\frac{4|\mathcal{X}|\log(2|\Pi|/\delta)}{T}}$$

$\Longrightarrow$ If $\tau \geq 4|\mathcal{X}|\varepsilon^{-2}\log(2|\Pi|/\delta)$ then under uniform exploration we have

$$\max_{\pi\in\Pi} |\hat{V}_\tau(\pi) - V(\pi)| \leq \varepsilon \quad \text{w.p.} \geq 1-\delta.$$

"Model the World"

Fix some function class $\mathcal{F}$ s.t. $f\in\mathcal{F}$

$$f: \mathcal{C}\times\mathcal{X} \to \mathbb{R}$$

Idea: Ideally $\exists f_* \in \mathcal{F}$: $v(c,x) \approx f_*(c,x)$

Question: Can we learn f_* w/ our dataset

$$\{(c_t, x_t, r_t, p_t)\}_{t=1}^{\tau}$$

$$\hat{f} = \underset{f\in\mathcal{F}}{\arg\min} \frac{1}{\tau}\sum_{t=1}^{\tau}(r_t - f(c_t,x_t))^2$$

fix any $f\in\mathcal{F}$

$$\mathbb{E}\left[(r_t - f(c_t,x_t))^2\right] = \mathbb{E}\left[(r_t - v(c_t,x_t) + v(c_t,x_t) - f(c_t,x_t))^2\right]$$

$$= \mathbb{E}\left[(r_t - v(c_t,x_t))^2\right] + \mathbb{E}\left[(v(c_t,x_t) - f(c_t,x_t))^2\right]$$

$$\leq \frac{1}{4} + \sum_{x\in\mathcal{X}} \mathbb{E}\left[\mathbb{E}\left[\mathbb{1}\{x_t = x\}\left(v(c_t,x) - f(c_t,x)\right)^2 \middle| c_t\right]\right]$$

$$= \frac{1}{4} + \mathbb{E}_c\left[\sum_{x \in \mathcal{X}} \mu(x|c) \; (v(c,x) - f(c,x))^2 \right]$$

$$\mathbb{E}\left[\frac{(r_t - f(c_t, x_t))^2}{p_t} \right] \le \frac{1}{4}\mathbb{E}\left[\sum_x \frac{1}{\mu(x|c)} \right] + \mathbb{E}\left[\sum_x (v(c,x) - f(c,x))^2 \right]$$

Note: If $\mu(x|c) > 0 \quad \forall x, c$ and $v \in \mathcal{F}$

then both above objectives satisfy $\hat{f} \to v$ if $T \to \infty$.

To evaluate a policy π, output

$$\hat{V}(\pi) = \frac{1}{T} \sum_{t=1}^{T} \hat{f}(c_t, \pi(c_t))$$

Concerns / Warning:

- Could be wasteful b/c I care about estimating $V(\pi)$, but learning $\hat{f}$ does not take π into account

  i.e. learning $\hat{f} \in \mathcal{F}$ well enough to estimate $|V(\pi) - \hat{V}(\pi)| \le \varepsilon$ could require way more samples than $|\mathcal{X}| \varepsilon^{-2} \log(|\Pi|)$

  $|\mathcal{F}|$ too big

- Biased. If $\min_{f \in \mathcal{F}} \max_{x,c} |f(x,c) - v(c,x)|$ is not small, then regardless of how much

  $|\mathcal{F}|$ too small

data you collect, estimates are <u>biased</u>.

Doubly Robust Estimators

Learn some $\hat{f} \in \mathcal{F}$ then set

$$\hat{V}_{DR}(c_t, x) = \hat{f}(c_t, x) + \left(r_t - \hat{f}(c_t, x)\right) \frac{\mathbb{1}\{x_t = x\}}{p_t}$$

$$\mathbb{E}[\hat{V}_{DR}(c_t, x) \mid c_t] = V(c_t, x) \quad \text{Unbiased!}$$

and by computing variance, if $\mathbb{E}[(r_t - V(c_t, x_t))^2]$ is small then variance of $\hat{V}_{DR}(c_t, x)$

$$\left(V(c_t, x) - \hat{f}(c_t, x) \right)^2 \frac{1}{p_t}$$

This assume $\hat{f} \perp \{(c_t, x_t, r_t, p_t)\}_{t=1}^{T}$

but in practice $\hat{f}$ is trained on

Linear Contextual Bandits (special case: "Model the world")

Assume $\exists$ known $\phi : \mathcal{C} \times \mathcal{X} \to \mathbb{R}^d$ and unknown $\exists \theta^* \in \mathbb{R}^d$

such that $v(c,x) = \langle \theta^*, \phi(c,x) \rangle$

"Model the world" where $f(c,x) = \langle \theta, \phi(c,x) \rangle$ for $\theta \in \mathbb{R}^d$.

for $t = 1, 2, \ldots$

Nature reveals $c_t \sim \mathcal{D}$

$\mathcal{Z}_t = \{ \phi(c_t, x) : x \in \mathcal{X} \}$

Player chooses $z_t \in \mathcal{Z}_t$

and observes $\langle z_t, \theta^* \rangle + \varepsilon_t$, $\mathbb{E}[\varepsilon_t] = 0$

In practice:

- If $\mathcal{X}$ is unstructured and $c \in \mathbb{R}^p$, $i \in \mathcal{X}$ $\phi(c,i) = \text{vec}(c e_i^T)$
- ϕ is learned by historical data

Natural algorithm is UCB. Construct confidence set C_t : $\theta^* \in C_t$ $\forall t$ w.h.p.

Play $z_t = \arg\max_{z \in \mathcal{Z}_t} \max_{\theta \in C_t} \langle z, \theta \rangle$

$R_T \lesssim d\sqrt{T}$ (ignoring logs)

(Thompson sampling works very well here)