Contextual Bandits (Stochastic)

Input: Π, $\pi \in \Pi$, $\pi : \mathcal{C} \to [n]$

for $t = 1, 2, \ldots, T$

Nature reveals $c_t \overset{iid}{\sim} \nu$

Player chooses $\pi_t \in \Pi$, $a_t := \pi_t(c_t)$

Receives reward $r_t \in [0,1]$: $\mathbb{E}[r_t \mid c_t] = r(c_t, a_t)$

Minimize regret $\max_{\pi \in \Pi} \sum_{t=1}^{T} r(c_t, \pi(c_t)) - r(c_t, a_t)$

$V(\pi) = \mathbb{E}_{c, a \sim \pi(c)}[r(c,a)] = \mathbb{E}_{c}[r(c, \pi(c))]$

Logging policy $\mu(\cdot \mid c_t) \in \Delta_n$, $p_t = \mu(a_t \mid c_t)$. Assume $\mu(a|c) > 0$ $\forall a, c$

Collect dataset $\{(c_t, a_t, r_t, p_t)\}_t$. Always log sampling probs p_t!

Forget context. Given $\{(a_t, r_t, p_t)\}_t$. Idea: learn some function

f: $f(a_t) \approx r_t$ (e.g. $\hat{f} = \operatorname{argmin}_{f \in \mathcal{F}} \sum_t (f(a_t) - r_t)^2$)

recommend $\operatorname{argmax}_a \hat{f}(a)$

Model the World

Given $\{(c_t, a_t, r_t, p_t)\}$

Learn f: $f(c_t, a_t) \approx r_t$ (e.g. $\hat{f} = \operatorname{argmin}_{f \in \mathcal{F}} \sum_t (f(c_t, a_t) - r_t)^2$)

recommend $\operatorname{argmax}_a \hat{f}(c_t, a)$

Model the Bias

$$\hat{r}(c_t, a) := r_t \frac{\mathbb{1}\{a_t = a\}}{p_t}$$

$$\mathbb{E}[\hat{r}(c_t, a)] = \sum_{a'} \underbrace{P(a_t = a')}_{= \mu(a' | c_t)} r(c_t, a') \frac{\mathbb{1}\{a' = a\}}{\mu(a' | c_t)}$$

$$= r(c_t, a)$$

$$\hat{V}(\pi) = \frac{1}{T} \sum_{t=1}^{T} \hat{r}(c_t, \pi(c_t)), \quad \mathbb{E}[\hat{V}(\pi)] = V(\pi)$$

$$\mathbb{E}[(\hat{V}(\pi) - V(\pi))^2] = \frac{1}{T^2} \mathbb{E} \sum_{t=1}^{T} (\hat{r}(c_t, \pi(c_t)) - r(c_t, \pi(c_t)))^2$$

$$\leq \frac{1}{T^2} \mathbb{E}\left[\sum_t \hat{r}(c_t, \pi(c_t))^2\right]$$

$$= \frac{1}{T^2} \sum_t \mathbb{E}_{c_t}\left[r_t^2 \frac{\mathbb{1}\{a_t = \pi(c_t)\}}{p_t^2}\right]$$

$$\leq \frac{1}{T^2} \sum_t \mathbb{E}_{a, c_t}\left[\frac{\mathbb{1}\{a_t = \pi(c_t)\}}{\mu(a_t | c_t)^2}\right]$$

$$= \frac{1}{T} \mathbb{E}_{c \sim \nu}\left[\frac{1}{\mu(\pi(c) | c)}\right]$$

I want high prob bound on $|\hat{V}(\pi) - V(\pi)|$.

Hoeffding says that if $Z_t \in [a,b]$ and $\mathbb{E}[Z_t] = 0$ then $\mathbb{P}\left(\frac{1}{T}\sum_{t=1}^{T} Z_t \geq |b-a|\sqrt{\frac{\log(1/\delta)}{2T}} \right) \leq \delta$.

$\hat{r}(c_t, \pi(c_t)) \in [0, \frac{1}{\min_{c,a} \mu(a|c)}]$

Bernstein's inequality says that if $Z_t \leq B$ and $\mathbb{E}[Z_t] = 0$, $\mathbb{E}[Z_t^2] \leq \sigma^2$ then

$$\mathbb{P}\left(\frac{1}{T}\sum_{t=1}^{T} Z_t \geq \sqrt{\frac{2\sigma^2 \log(1/\delta)}{T}} + \frac{2B\log(1/\delta)}{3T} \right) \leq \delta.$$

$\Rightarrow$ w.p. $\geq 1-\delta$

$$|\hat{V}(\pi) - V(\pi)| \leq \sqrt{\underbrace{\mathbb{E}_c\left[\frac{1}{\mu(\pi(c)|c)}\right]}_{\leq n \text{ if uniform}} \cdot \frac{2\log(2/\delta)}{T}} + \underbrace{\max_{c,a} \frac{1}{\mu(a|c)}}_{\leq n \text{ if uniform}} \cdot \frac{2\log(2/\delta)}{3T}$$

Side note: There exists an estimator $\hat{\mu}: \mathbb{R}^T \to \mathbb{R}$ s.t. if $\mathbb{E}[Z_t] = 0$, $\mathbb{E}[Z_t^2] \leq \sigma^2$ then

$$\mathbb{P}\left(\hat{\mu}(\{Z_t\}_{t=1}^T) \geq \sqrt{\frac{2\sigma^2 \log(1/\delta)}{T}} \right) \leq \delta.$$

See Catoni's estimator or median of means.

If $\mu(a|c) = \frac{1}{n}$ for c, a then for a $\pi \in \Pi$

$$|\hat{V}(\pi) - V(\pi)| \le \sqrt{\frac{2n \log(2/\delta)}{T}} + \frac{2n \log(2/\delta)}{3T}$$

$$\le \sqrt{\frac{4n \log(2/\delta)}{T}}$$

and $\forall \pi$ w.p. $\ge 1-\delta$

$$|\hat{V}(\pi) - V(\pi)| \le \sqrt{\frac{4n \log(2|\Pi|/\delta)}{T}} = C$$

You collected a data set computed $\hat{V}(\pi)$ for all $\pi \in \Pi$. Which one do you choose?

Natural to choose $\hat{\pi}_{MLE} = \arg\max_{\pi \in \Pi} \hat{V}(\pi)$

Define $C(\pi) = \sqrt{\mathbb{E}\left[\frac{1}{\mu(\pi(s)|s)}\right] \frac{2 \log(2|\Pi|/\delta)}{T}} + 2 \max_{s,a} \frac{1}{\mu(\pi(s)|s)} \frac{\log(\cdot)}{T}$

$$V(\hat{\pi}_{MLE}) \ge \hat{V}(\hat{\pi}_{MLE}) - C(\hat{\pi}_{MLE})$$

$$\ge \hat{V}(\pi_{\star}) - C(\hat{\pi})$$

$$\geq V(\pi_*) - C(\pi_*) - C(\hat{\pi})$$

$$\geq V(\pi_*) - 2 \max_{\pi \in \Pi} C(\pi)$$

Pessimism

Define $\hat{\pi}_{pess} = \operatorname{argmax}_{\pi \in \Pi} \hat{V}(\pi) - C(\pi)$

$$V(\hat{\pi}_{pess}) \geq \hat{V}(\hat{\pi}_{pess}) - C(\hat{\pi}_{pess})$$

$$\geq \hat{V}(\pi_*) - C(\pi_*)$$

$$\geq V(\pi_*) - 2C(\pi_*)$$

Doubly Robust Estimator

$$\hat{r}_{DR}(c_t, a) = \hat{f}(c_t, a) + (r_t - \hat{f}(c_t, a)) \frac{\mathbb{1}\{a_t = a\}}{p_t}$$

$$\mathbb{E}[\hat{r}_{DR}(c_t, a)] = r(c_t, a)$$