Contextual Bandits

Input: (finite) action set $\mathcal{X}$

for $t = 1, 2, \ldots$

Nature reveals context $c_t \sim \nu$

Player chooses $x_t \in \mathcal{X}$

Nature reveals reward $y_t \in [0,1]$, $\mathbb{E}[y_t \mid c_t, x_t] = r(c_t, x_t)$

Suppose $\text{support}(\nu) =: \mathcal{C}$ and $|\mathcal{C}| < \infty$.

Idea: run MAB on each context $c \in \mathcal{C}$

$$\text{Reward} = \sum_{t=1}^{T} r(c_t, x_t)$$

$$= \sum_{c \in \mathcal{C}} \sum_{t=1}^{T} \mathbb{1}\{c_t = c\} \, r(c, x_t)$$

Regret relative to best action per context

$$\text{Regret} = \sum_{c = \mathcal{C}} \max_{x} \sum_{t=1}^{T} \mathbb{1}\{c_t = c\} \left(r(c, x) - r(c, x_t) \right)$$

Apply MAB to each. $T_c = \sum_{t=1}^{T} \mathbb{1}\{c_t = c\}$

$$\lesssim \sum_{c = \mathcal{C}} \sqrt{|\mathcal{X}| \, T_c \log(T_c / \delta)}$$

$$= \sum_{c \in \mathcal{C}} \sqrt{T_c} \cdot \sqrt{|\mathcal{X}| \log(T_c/\delta)}$$

$$\leq \sqrt{\left(\sum_{c \in \mathcal{C}} T_c \right) \cdot \sum_{c \in \mathcal{C}} |\mathcal{X}| \log(T_c/\delta)} \qquad \text{Cauchy-Schwartz}$$

$$\leq \sqrt{T |\mathcal{X}| |\mathcal{C}| \log(T/\delta)}$$

$$\text{Reward} \geq \sum_{t=1}^{T} \max_{x} r(c_t, x) - \sqrt{T |\mathcal{X}| |\mathcal{C}| \log(T/\delta)}$$

Another simple idea: ignore the context!

$$\text{Reward} = \sum_{t=1}^{T} r(c_t, x_t)$$

$$\text{Regret} = \max_{x} \sum_{t=1}^{T} r(c_t, x) - r(c_t, x_t)$$

$$\leq \sqrt{T \cdot |\mathcal{X}| \cdot \log(T/\delta)}$$

$$\Rightarrow \quad \text{Reward} \geq \max_{x} \sum_{t=1}^{T} r(c_t, x) - \sqrt{T |\mathcal{X}| \cdot \log(T/\delta)}$$

Define a policy $\pi: \mathcal{C} \to \mathcal{X}$. It value is defined as

$$V(\pi) = \mathbb{E}_{c \sim \nu}\left[r(c, \pi(c)) \right].$$

Consider a collection of policies Π (assume finite), and at each time t, we choose $\pi_t \in \Pi$ and play $\pi_t(c_t) = x_t$.

$$\text{Policy-Regret} = \max_{\pi \in \Pi} \sum_{t=1}^{T} r(c_t, \pi(c_t)) - r(c_t, \pi_t(c_t))$$

Ex. For policies that ignore context $\Pi = \{ \pi(c) = x : x \in \mathcal{X} \}$

$|\Pi| = |\mathcal{X}|$

Ex. For MAB at each context $\Pi = \{ \pi(c) = x ; x \in \mathcal{X}, c \in \mathcal{X} \}$

$|\Pi| = |\mathcal{C}| \cdot |\mathcal{X}|$

One idea: Treat policy regret as MAB over policies Π.

$$\Rightarrow \quad \text{Regret} \leq \sqrt{|\Pi| \cdot T \log(T)}$$

Suppose we had full information s.t. we observe $y_{t,x}$ w/ $\mathbb{E}[y_{t,x} \mid x, c_t] = r(c_t, x)$ at every time.

With this knowledge, define $\hat{V}(\pi) = \frac{1}{T} \sum_{t=1}^{T} y_{t,\pi(c_t)}$

$$\hat{V}(\pi) - V(\pi) = \frac{1}{T} \sum_{t=1}^{T} \left(y_{t,\pi(c_t)} - r(c_t, \pi(c_t)) \right)$$

Apply Hoeffding + union bound to have

$$\mathbb{P}\left(\bigcup_{\pi \in \Pi} \{ |\hat{V}(\pi) - V(\pi)| \geq \sqrt{\frac{\log(2|\Pi|/\delta)}{2T}} \right) \leq \delta.$$

$$\mathbb{P}\left(\bigcup_{\pi \in \Pi} \bigcup_{t=1}^{T} |\hat{V}_t(\pi) - V(\pi)| \geq \sqrt{\frac{\log(2T|\Pi|/\delta)}{2t}} \right) \leq \delta$$

Let $\pi_t := \underset{\pi \in \Pi}{\arg\max}\ \hat{V}_{t-1}(\pi)$

$$\text{Regret} = \max_{\pi} \mathbb{E}\left[\sum_{t=1}^{T} r(c_t, \pi(c_t)) - r(c_t, \pi_t(c_t)) \right]$$

$$= \max_{\pi} \sum_{t=1}^{T} V(\pi) - V(\pi_t)$$

$\pi_\emptyset = \arg\max_{\pi} V(\pi)$

$$\leq 1 + \max_{\pi} \sum_{t=2}^{T} V(\pi) - V(\pi_t)$$

$$= 1 + \sum_{t=2}^{T} \underbrace{V(\pi_\emptyset) - \hat{V}_{t-1}(\pi_\emptyset)}_{\leq \frac{1}{\sqrt{t}}} + \underbrace{\hat{V}_{t-1}(\pi_\emptyset) - \hat{V}_{t-1}(\pi_t)}_{\leq 0} + \underbrace{\hat{V}_{t-1}(\pi_t) - V(\pi_t)}_{\lesssim \frac{1}{\sqrt{t}}}$$

$$\leq 1 + \sum_{t=2}^{T} 2 \sqrt{\frac{\log(|\Pi| T/\delta)}{2(t-1)}} \lesssim \sqrt{T \log(|\Pi| T/\delta)}.$$

Off-policy evaluation

Fix a logging policy $\mu(\cdot \mid c) \in \Delta_{\mathcal{X}}$ that prob. dist over actions for each context.

Ex. for any $\lambda \in \Delta_{\Pi}$ I could set $\mu(x \mid c) = \sum_{\pi \in \Pi} \mathbb{1}\{\pi(c) = x\} \lambda_{\pi}$

Ex. Take $\mu(x \mid c) = \frac{1}{|\mathcal{X}|}$

Suppose $\mu(\cdot \mid \cdot)$ is run for T timesteps to produce data $\{(c_t, x_t, P_t, y_t)\}_{t=1}^{T}$ where

$$P_t := \mu(x_t \mid c_t). \qquad x_t \sim \mu(\cdot \mid c_t)$$

$$\hat{V}_{IPS}(\pi) := \frac{1}{T} \sum_{t=1}^{T} \frac{\mathbb{1}\{x_t = \pi(c_t)\}}{P_t} y_t = \frac{1}{T} \sum_{t=1}^{T} \frac{\mathbb{1}\{x_t = \pi(c_t)\}}{\mu(x_t \mid c_t)} y_t$$

$$\mathbb{E}[\hat{V}_{IPS}(\pi)] = \frac{1}{T} \sum_{t=1}^{T} \underbrace{\mathbb{E}\left[\frac{\mathbb{1}\{x_t = \pi(c_t)\}}{\mu(x_t \mid c_t)} y_t\right]}$$

$$= \mathbb{E}\Big[\underbrace{\mathbb{E}\left[\frac{\mathbb{1}\{x_t = \pi(c_t)\}}{\mu(x_t \mid c_t)} y_t \,\Big|\, x_t, c_t\right]}\Big]$$

$$= \frac{\mathbb{1}\{x_t = \pi(c_t)\}}{\mu(x_t \mid c_t)} r(c_t, x_t)$$

$$= \mathbb{E}\Big[\underbrace{\mathbb{E}\left[\frac{\mathbb{1}\{x_t = \pi(c_t)\}}{\mu(x_t \mid c_t)} r(c_t, x_t) \,\Big|\, c_t = c\right]}\Big]$$

Implicitly assumes $\mu(x \mid c) \geq 0 \;\; \forall x, c$

$$= \sum_{x \in \mathcal{X}} \underbrace{\mu(x \mid c_t)}_{= \mathbb{P}(x_t = x \mid c_t)} \frac{\mathbb{1}\{x = \pi(c_t)\}}{\mu(x \mid c_t)} r(c_t, x)$$

$$= r(c_t, \pi(c_t))$$

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

$\mathbb{E}[(X-\mathbb{E}[X])^2]$

$= \mathbb{E}[X^2] - \mathbb{E}[X]^2$

$\le \mathbb{E}[X^2]$

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

$$= V(\pi)$$

Consider $\mathbb{V}ar(\hat{V}_{IPS}(\pi)) = \frac{1}{T}\mathbb{V}ar\left(\frac{\mathbb{1}\{x_t = \pi(c_t)\}}{\mu(x_t|c_t)} y_t\right)$

$$\le \frac{1}{T}\mathbb{E}\left[\left(\frac{\mathbb{1}\{x_t = \pi(c_t)\}}{\mu(x_t|c_t)} y_t\right)^2\right]$$

$y_t \in [0,1]$

$$\le \frac{1}{T}\mathbb{E}\left[\frac{\mathbb{1}\{x_t = \pi(c_t)\}}{\mu(x_t|c_t)^2}\right]$$

$X_t \quad t=1,2,\dots,T$

$$= \frac{1}{T}\mathbb{E}\left[\sum_x \mu(x|c_t)\frac{\mathbb{1}\{x|\pi(c_t)\}}{\mu(x|c_t)^2}\right]$$

$\mathbb{V}(\frac{1}{T}\sum_t X_t) = \frac{1}{T^2}\mathbb{V}(\sum_t X_t)$

$= \frac{1}{T^2}\sum_t \mathbb{V}(X_t)$

$$= \frac{1}{T}\mathbb{E}\left[\sum_x \frac{\mathbb{1}\{x|\pi(c_t)\}}{\mu(x|c_t)}\right]$$

$= \frac{1}{T}\mathbb{V}(X_t)$

for any t.

$$= \frac{1}{T}\mathbb{E}\left[\frac{1}{\mu(\pi(c_t)|c_t)}\right]$$

Lemma Bernstein's inequality. Let $X_1, \dots, X_n$ be independent R.V. s.t. $\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[(X_i - \mathbb{E}[X_i])^2] \le \sigma^2$ and $\max_i |X_i| \le B$. Then w.p. $\ge 1-\delta$:

$$\left|\frac{1}{n}\sum_{i=1}^{n}(X_i - \mathbb{E}[X_i])\right| \le \sqrt{\frac{2\sigma^2\log(2/\delta)}{n}} + \frac{2B\log(2/\delta)}{3n}$$

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

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

Ex. $\mu(x|c) = \frac{1}{|\mathcal{X}|}$ then w.p. $\geq 1-\delta$ $\forall \pi \in \Pi$

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

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

Above is "model-free" approach to estimating $V(\pi)$ in the sense that it makes no attempt to "model the environment" meaning estimating $r(c,x)$ $\forall c,x$.

Example policy class given $\phi: \mathcal{C} \times \mathcal{X} \to \mathbb{R}^d$, for each $\theta \in \mathbb{R}^d$

Define $\pi_\theta(c) = \arg\max_x \langle \phi(c,x), \theta \rangle$

Fact after T rounds $|\Pi| \leq n^d$ policies

Model-Based estimation

If you knew $r(c,x)$ then best possible action (and policy) is to play $x_t = \arg\max_x r(c_t, x)$.

Next best thing: consider a function class $\mathcal{F}$ so that $f \in \mathcal{F}$ has $f: \mathcal{C} \times \mathcal{X} \to \mathbb{R}$, given data $\{(c_t, x_t, p_t, y_t)\}_{t=1}^T$

$$\hat{f} = \arg\min_{f \in \mathcal{F}} \sum_{t=1}^T (f(c_t, x_t) - y_t)^2$$

set
$$\hat{V}_{\mathcal{F}}(\pi) = \frac{1}{T} \sum_{t=1}^T \hat{f}(c_t, \pi(c_t)).$$

Ex. Let $\mathcal{F} = \{\langle \phi(c,x), \theta \rangle : \theta \in \mathbb{R}^d\}$

$$\hat{\theta} = \arg\min_{\theta} \sum_{t=1}^T (\langle \phi(c_t, x_t), \theta \rangle - y_t)^2$$

$$\hat{V}_{\mathcal{F}}(\pi) = \frac{1}{T} \sum_{t=1}^T \langle \phi(c_t, \pi(c_t)), \hat{\theta} \rangle$$