Contextual Bandits

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

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

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

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

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

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

$$\leq \sqrt{\left(\sum_{c \in \mathcal{C}} T_c\right) \cdot \sum_{c \in \mathcal{C}} |\mathcal{X}| \log(T_c/\delta)}$$ 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 $\pi : \mathcal{C} \to \mathcal{X}$. Its value is defined as

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

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.

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

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

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

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

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

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

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

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

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

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