Contextual bandits (Stochastic)

Input: finite set of arms $\mathcal{X}$

Policy π plays $x_t = \pi(c_t)$

for $t = 1, 2, \ldots$

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

Player chooses action $x_t \in \mathcal{X}$ and receives

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

Goal: choose $x_1, \ldots, x_t, \ldots$ in order to maximize $\sum_{t=1}^{T} v(c_t, x_t)$, total reward.

Finite context set. $\text{support}(\mathcal{D}) = \mathcal{C}$ and $|\mathcal{C}| < \infty$.

Idea: Instantiate a MAB algo (e.g. Elimination, UCB, ...) for each $c \in \mathcal{C}$ and play the cth algo when $c_t = c$. Then w.p. $\geq 1-\delta$

$$\max_{x \in \mathcal{X}} \sum_{t=1}^{T} \mathbb{1}\{c_t = c\}\left(v(c,x) - v(c,x_t)\right) \lesssim \sqrt{|\mathcal{X}|\, T_c \log\left(\frac{|\mathcal{X}|\, T_c}{\delta}\right)}$$

$$\leq \sqrt{|\mathcal{X}|\, T_c \log\left(\frac{|\mathcal{X}|\, T}{\delta}\right)}$$

$$T_c = \sum_{t=1}^{T} \mathbb{1}\{c_t = c\}$$

Union bounding over all $c \in \mathcal{C}$, we have w.p. $\geq 1-\delta$

$$\sum_{c \in \mathcal{C}} \max_{x \in \mathcal{X}} \sum_{t=1}^{T} \mathbb{1}\{c_t = c\}\left(v(c,x) - v(c,x_t)\right) \leq \sum_{c \in \mathcal{C}} \sqrt{|\mathcal{X}|\, T_c \log\left(\frac{|\mathcal{X}|\, T \cdot |\mathcal{C}|}{\delta}\right)}$$

$$\leq \sqrt{|\mathcal{C}| \cdot |\mathcal{X}| \cdot T \cdot \log\left(\frac{|\mathcal{X}| \cdot T \cdot |\mathcal{C}|}{\delta}\right)} \quad (1)$$

This bound is vacuous when $|\mathcal{C}|$ is large.

Idea: Play MAB algo <u>ignoring</u> context altogether.

Note r_t is a R.V. w/ mean $v(c_t, x_t)$

and if $x_t \perp c_t$ then r_t is iid

R.V. w/ mean $\mathbb{E}[r_t \mid x_t] = \mathbb{E}_{c\sim\mathcal{D}}[v(c, x_t) \mid x_t]$

If we play some MAB algo, then w.p. $\geq 1-\delta$

$$\max_{x \in \mathcal{X}} \sum_{t=1}^{T} v(c_t, x) - v(c_t, x_t) \lesssim \sqrt{|\mathcal{X}|\, T \log\left(\frac{|\mathcal{X}| T}{\delta}\right)}. \quad (2)$$

Rearranging

$$(1) \longrightarrow \quad \sum_{t=1}^{T} v(c_t, x_t) \geq \sum_{c \in \mathcal{C}} \max_{x \in \mathcal{X}} \sum_{t=1}^{T} \mathbb{1}\{c_t = c\} v(c, x) - \sqrt{|\mathcal{C}| \cdot |\mathcal{X}| \cdot T \cdots}$$

$$(2) \longrightarrow \quad \sum_{t=1}^{T} v(c_t, x_t) \geq \max_{x \in \mathcal{X}} \sum_{t=1}^{T} v(c_t, x) - \sqrt{|\mathcal{X}| \cdot T \cdots}$$

We always have

$$\sum_{c \in C} \max_{x \in \mathcal{X}} \sum_{t=1}^{T} \mathbb{1}\{c_t = c\} v(c, x) \geq \max_{x \in \mathcal{X}} \sum_{c \in \mathcal{C}} \sum_{t=1}^{T} \mathbb{1}\{c_t = c\} v(c, x)$$

$$= \max_{x \in \mathcal{X}} \sum_{t=1}^{T} v(c_t, x)$$

Ex. suppose $= T$ and $= T/100.$

Then ① $\geq T - \sqrt{|\mathcal{C}||\mathcal{X}|T}$

② $\geq \frac{T}{100} - \sqrt{|\mathcal{X}| \cdot T}$

$$\frac{99}{100} T = \sqrt{T}\left(\sqrt{\mathcal{C}\mathcal{X}} - \sqrt{\mathcal{X}}\right)$$

$T <$

In general: we define a policy $\pi : \mathcal{C} \to \mathcal{X}$. The value of π is defined as

$$V(\pi) := \mathbb{E}_{C \sim \mathcal{D}} \left[V(C, \pi(C)) \right].$$

Consider a collection of policies Π. Then define the policy regret wrt Π as

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

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

$$= \max_{\pi \in \Pi} \mathbb{E} \left[\sum_{t=1}^{T} V(c_t, \pi(c_t)) - V(c_t, x_t) \right]$$

From above, 1 was playing best action per context $\Rightarrow |\Pi| = |\mathcal{X}|^{|\mathcal{C}|}$

2 is best action over all $|\Pi| = |\mathcal{X}|$

We will see later that $R_T \leq \sqrt{T \cdot |\mathcal{X}| \cdot \log |\Pi|}$.

Note if $\Pi_1 \subset \Pi_2$ then $\max_{\pi \in \Pi_2} V(\pi) \geq \max_{\pi \in \Pi_1} V(\pi)$

$\Rightarrow$ the more "complex" your policy class is the higher reward/value is possible. But the regret incurred to learn $\pi_* \in \operatorname{argmax}_{\pi \in \Pi} V(\pi)$

may be larger. Thus you want to pick policy class $|\Pi|$ s.t. $\log|\Pi| \leq T$

Policy evaluation

Aside

Naively, we could estimate $V(\pi)$ for each π by "rolling it out" or playing it: for $\pi \in \Pi$

Play π for τ times to get $\{r_t\}_{t=1}^{\tau}$

Set $\hat{V}(\pi) = \frac{1}{\tau}\sum_{t=1}^{\tau} r_t$.

By Hoeffding $|\hat{V}(\pi) - V(\pi)| \leq \sqrt{\frac{\log(|\Pi|/\delta)}{\tau}} = \varepsilon$ Need $\varepsilon^{-2}|\Pi|\log(|\Pi|/\delta)$ samples.

Smarter way through randomization.

Suppose we have a random exploration policy s.t. at each time t, this policy plays action x where

$$\mathbb{P}(x_t = x \mid c_t) =: \mu(x \mid c_t).$$

Equivalently, I have a distribution $\lambda \in \Delta_\Pi$ and at each time t, sample $\pi_t \sim \lambda$ and play $\pi_t(c_t) = x_t$.

where $\mu(x \mid c_t) = \sum_{\pi\in\Pi} \lambda_\pi \mathbb{1}\{\pi(c_t) = x\}$. $\quad c_t \overset{iid}{\sim} \mathcal{D}$

Suppose we play this policy for T time steps to collect a dataset $\{(c_t, x_t, r_t, p_t)\}_{t=1}^{T}$, $p_t = \mu(x_t \mid c_t)$.

Question: Using collected data, construct estimate $\hat{V}(\pi)$ for $V(\pi)$, $\forall \pi$?

<u>Two strategies</u>: Model the world.

Model the bias.

<u>Model the bias</u>

Fix time t.

For any $x \in \mathcal{X}$: $\quad \hat{v}(c_t, x) = \dfrac{\mathbb{1}\{x_t = x\}}{p_t} r_t$

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

$\frac{1}{p_t}$ = Inverse propensity score

$\hat{v}(c_t, x)$ is IPS estimator

<u>Prop</u> $\mathbb{E}[\hat{v}(c_t, x) \mid c_t] = v(c_t, x)$.

<u>proof</u>

$$\mathbb{E}[\hat{v}(c_t, x) \mid c_t] = \mathbb{E}\left[\frac{\mathbb{1}\{x_t = x\}}{p_t} r_t \,\middle|\, c_t\right]$$

$$= \mathbb{E}\left[\mathbb{E}\left[\frac{\mathbb{1}\{x_t = x\}}{p_t} r_t \,\middle|\, x_t, c_t\right] \,\middle|\, c_t\right]$$

$$= \mathbb{E}\left[\frac{\mathbb{1}\{x_t = x\}}{p_t} v(c_t, x) \,\middle|\, c_t\right]$$

$$= \sum_{x' \in \mathcal{X}} \underbrace{\mathbb{P}(x' = x_t \mid c_t)}_{\substack{= \mu(x_t \mid c_t) \\ = p_t}} \frac{\mathbb{1}\{x' = x\}}{p_t} v(c_t, x)$$

$$= \sum_{x' \in \mathcal{X}} \cancel{p_t} \cdot \frac{\mathbb{1}\{x' = x\}}{\cancel{p_t}} v(c_t, x) = v(c_t, x)$$

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

$$= \mathbb{E}_{c\sim\mathcal{D}}\left[v(c,\pi(c))\right] = V(\pi)$$

What is the variance of $\hat{v}(c_t, x)$?

$$\mathbb{E}\left[\left(\hat{v}(c_t,x) - v(c_t,x)\right)^2 \mid c_t\right]$$

$$\leq \mathbb{E}\left[\hat{v}(c_t,x)^2 \mid c_t\right]$$

$$= \mathbb{E}\left[\frac{\mathbb{1}\{x_t = x\}^2}{p_t^2} r_t^2 \mid c_t\right]$$

$$\leq \mathbb{E}\left[\frac{\mathbb{1}\{x_t = x\}}{p_t^2} \mid c_t\right] \qquad (|r_t| \leq 1)$$

$$= \sum_{x'\in\mathcal{X}} \frac{\mathbb{1}\{x' = x\}}{p_t^2}\mu(x' \mid c_t) = \frac{1}{p_t}$$

$$\Rightarrow \text{Varione}\left(\hat{V}(\pi)\right) \leq \frac{1}{T^2}\sum_{t=1}^{T}\frac{1}{p_t}.$$