Contextual Bandits w/ policies Π: $\pi \in \Pi$ is a map $\pi : \mathcal{C} \to \mathcal{X}$

Input Π

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

Nature reveals context $c_t \in \mathcal{C}$ is iid from $\mathcal{D}$

Player choose action $x_t \in \mathcal{X}$ (equiv. chooses $\pi_t \in \Pi$ and plays $\pi(c_t)$)

Player receives $r_t \in [0,1]$ where $\mathbb{E}[r_t | c_t, x_t] = V(c_t, x_t)$

Define $V(\pi) = \mathbb{E}_c [V(c, \pi(c))]$

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

$$\max_{\pi \in \Pi} \sum_{t=1}^{T} V(c_t, \pi) - V(c_t, \pi_t) \quad (2)$$

$$\max_{\pi} \sum r_t[\pi(c_t)] - r_t[\pi_t(c_t)] \quad (3)$$

In our stochastic setting these are all within $O(\sqrt{T})$

Idea: play some logging policy μ: in response to context c_t, policy plays $x_t = x$ w.p. $\mu(x|c_t)$

Collect dataset $\{(c_t, x_t, r_t, p_t)\}_{t=1}$ $\quad p_t := \mu(x_t|c_t)$

Estimate $\hat{V}(\pi) = \frac{1}{T} \sum_{t=1}^{T} \frac{\mathbb{1}\{x_t = x\}}{p_t} r_t$

$$\mathbb{E}[\hat{V}(\pi)] = V(\pi)$$

$$\mathbb{E}[(\hat{V}(\pi) - V(\pi))^2] \le \mathbb{E}_{c \sim \mathcal{D}} \left[\frac{1}{\mu(\pi(c)|c)} \right].$$

By Bernstein. w.p. $\geq 1-\delta$, $\forall \pi \in \Pi$ we have

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

where $V_{max} = \max_{c,x} \frac{1}{\mu(x,c)}$.

S-Greedy: Algorithm for "model the bias"/general policies

Input Π

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

Nature reveals c_t

Alg plays x_t uniformly at random

Construct $\hat{V}(\pi)$ as above

$\hat{\pi} = \operatorname{argmax}_{\pi} \hat{V}(\pi)$

for $t = S+1, \ldots, T$

Nature reveals c_t

$x_t = \hat{\pi}(c_t)$

$\mu(x|c_t) = \frac{1}{|\mathcal{X}|}$ for all $x \in \mathcal{X}$.

$\Rightarrow$ w.p. $\geq 1-\delta$ $\forall \pi \in \Pi$ $|\hat{V}(\pi) - V(\pi)| \leq \underbrace{\sqrt{\frac{4|\mathcal{X}|\log(2|\Pi|/\delta)}{S}}}_{:=\varepsilon_S}$

$\pi_* \in \operatorname{argmax}_{\pi\in\Pi} V(\pi)$

$$V(\hat{\pi}) = \underbrace{V(\hat{\pi}) - \hat{V}(\hat{\pi})}_{\geq -\varepsilon_S} + \underbrace{\hat{V}(\hat{\pi}) - \hat{V}(\pi_*)}_{\geq 0} + \underbrace{\hat{V}(\pi_*) - V(\pi_*)}_{\geq -\varepsilon_S} + V(\pi_*)$$

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

$$\text{Regret} = \sum_{t=1}^{T} V(\pi_*) - V(\pi_t)$$

$$= \sum_{t=1}^{S} V(\pi_*) - V(\pi_t) + \sum_{t=S+1}^{T} V(\pi_*) - V(\pi_t)$$

$$\leq S \cdot 1 + (T-S)\underbrace{\left(V(\pi_*) - V(\hat{\pi})\right)}_{\leq 2\varepsilon_S}$$

$$= S + 2(T-S)\sqrt{\frac{4|\mathcal{X}|\log(2|\Pi|/\delta)}{S}}$$

$$= O\left(T^{2/3}\left(|\mathcal{X}|\log(2|\Pi|/\delta)\right)^{1/3}\right)$$

$$S = \left(|\mathcal{X}| T^2 \log(2|\Pi|/\delta)\right)^{1/3}$$

How do we find $\underset{\pi \in \Pi}{\operatorname{argmax}} \hat{V}(\pi)$ efficiently?

$$\hat{V}(\pi) = \frac{1}{T} \sum_{t=1}^{T} \frac{\mathbb{1}\{x_t = \pi(c_t)\}}{p_t} r_t$$

$$= \frac{1}{T} \sum_{t=1}^{T} \frac{(1 - \mathbb{1}\{x_t \neq \pi(c_t)\})}{p_t} r_t$$

$$= \frac{1}{T} \sum_{t=1}^{T} \frac{r_t}{p_t} - \frac{1}{T} \sum_{t=1}^{T} \mathbb{1}\{x_t \neq \pi(c_t)\} \frac{r_t}{p_t}$$

$$\underset{\pi}{\operatorname{argmax}} \hat{V}(\pi) = \underset{\pi}{\operatorname{argmin}} \frac{1}{T} \sum_{t=1}^{T} \mathbb{1}\{x_t \neq \pi(c_t)\} \frac{r_t}{p_t}$$

Think of each $\pi : \mathcal{C} \to \mathcal{X}$ as a classifier over classes $\{1, \ldots, |\mathcal{X}|\}$ and examples in $\mathcal{C}$.

$x_t \in \mathcal{X}$ is a "label", $c_t \in$ example, $\frac{r_t}{p_t}$ is weight

Minimize 0/1 loss on $\{(c_t, x_t, \frac{r_t}{p_t})\}_{t=1}^{T}$

↑ weight of tth point

Ex. Let $f_\theta : \mathcal{C} \to \mathbb{R}^{\mathcal{X}}$ be a neural network parameterized by parameters $\theta \in \mathbb{R}^d$

$$\text{loss}(\theta) = -\frac{1}{T} \sum_{t=1}^{T} \frac{r_t}{p_t} \log \left(\frac{\exp(f_\theta(c_t)[x_t])}{\sum_{x \in \mathcal{X}} \exp(f_\theta(c_t)[x])} \right)$$

Let $\hat{\theta}_s = \arg\min_\theta \text{Loss}(\theta)$

When c_t arrives we play

$$x_t = \underset{x \in \mathcal{X}}{\text{argmax}} \left[f_{\hat{\theta}_s}(c_t) \right]_x$$

Zooming out: $\mathcal{T}$-greedy achieves $R_T \lesssim T^{2/3} \left(|\mathcal{X}| \log(|\Pi|/\delta) \right)^{1/3}$.

But $|\mathcal{C}| = 1$ and $|\Pi| = |\mathcal{X}|$ in the standard MAB then we know $\sqrt{|\mathcal{X}| \cdot T}$ regret is possible.

→ For contextual Bandits, can we achieve $R_T \lesssim \sqrt{|\mathcal{X}| T \cdot \log|\Pi|}$?

Answer: yes.

An elimination alg for stochastic Contextual Bandits

Input Π, δ

$\hat{\Pi}_1 = \Pi$

for $\ell = 1, 2, \ldots$

$\varepsilon_\ell = 2^{-\ell}$, # measurements $\mathcal{T}_\ell$ TBD

regularization γ_ℓ TBD

$$Q_\ell = \underset{Q \in \Delta\hat{\Pi}_\ell}{\text{argmin}} \max_{\pi \in \hat{\Pi}_\ell} \mathbb{E}_c \left[\frac{1}{Q^\gamma(\pi(c)|c)} \right]$$

where $Q^\gamma(x|c) := \gamma + (1 - |\mathcal{X}|\gamma) Q(x|c)$

$Q(x|c) := \sum_{\pi : \pi(c) = x} Q(\pi)$ ← prob dist over actions

$T_\ell = \sum_{i=1}^{\ell} \tau_\ell$

for $t = T_{\ell-1}+1, \ldots, T_\ell$

Nature reveals c_t

Player plays $x_t \sim Q_\ell^\gamma(\cdot \mid c_t)$ to get reward r_t

$$\hat{V}_\ell(\pi) = \frac{1}{\tau_\ell} \sum_{t=1}^{\tau_\ell} \frac{\mathbb{1}\{\pi(c_t) = x_t\}}{p_t} r_t$$

$$\hat{\Pi}_{\ell+1} = \hat{\Pi}_\ell \setminus \{\pi \in \hat{\Pi}_\ell : \max_{\pi'} \hat{V}(\pi') - \hat{V}(\pi) \geq 2\varepsilon_\ell\}.$$

// Q_ℓ is chosen to minimize maximum variance of $\hat{V}(\pi)$:

$$Q_\ell \in \operatorname*{argmin}_{Q \in \Delta_{\hat{\Pi}_\ell}} \max_{\pi \in \hat{\Pi}_\ell} \mathbb{E}\left[\left(\hat{V}_\ell(\pi) - V(\pi)\right)^2\right]$$

Lemma For any policy Π and dist. over contexts

$$\min_{Q \in \Delta_\Pi} \max_{\pi \in \Pi} \mathbb{E}\left[\frac{1}{Q(\pi(c) \mid c)}\right] \leq |X|.$$

Furthermore, if $\gamma \leq \frac{1}{2K}$ then

$$\min_{Q \in \Delta_\Pi} \max_{\pi \in \Pi} \mathbb{E}\left[\frac{1}{Q^\gamma(\pi(c) \mid c)}\right] \leq 2|X|.$$

This ↑ lemma is a corollary of Kiefer-Wolfowitz which we will show next time.

Consequently methane by Bernstein

$$|\hat{V}_\ell(\bar{\pi}) - V(\bar{\pi})| \le \sqrt{\frac{(2|\mathcal{X}|)\cdot 2\log(2|\Pi|/\delta)}{T_\ell}} + \frac{2\log(2|\Pi|/\delta)}{\gamma_\ell 3 T_\ell}$$

$$\le \sqrt{\frac{16|\mathcal{X}|\log(2|\Pi|/\delta)}{T_\ell}} =: \varepsilon_\ell$$

for $\gamma_\ell = \min\left\{\frac{1}{2|\mathcal{X}|}, \sqrt{\frac{2\log(|\Pi|/\delta)}{9|\mathcal{X}|T_\ell}}\right\}$