Policy Bandit Optimization

for $t = 1, 2, \dots$

Nature reveals c_t

Player choses x_t

Nature reveals y_t s.t. $\mathbb{E}[y_t | c_t, x_t] = r(c_t, x_t)$

Given policy set Π, minimize policy regret

$$\max_{\pi} \sum_{t=1}^{T} r(c_t, \pi(c_t)) - r(c_t, x_t) \qquad \pi_* = \arg\max_{\pi} V(\pi)$$

ξ-greedy

for $t = 1, 2, \dots, \xi$ play $x_t \sim \text{uniform}(\mathcal{X})$ and then set

$$\hat{\pi} = \arg\max_{\pi} \hat{V}_{IPS}(\pi) = \arg\max_{\pi} \frac{1}{\xi} \sum_{t=1}^{\xi} \frac{\mathbb{1}\{x_t = \pi(c_t)\}}{p_t} y_t$$

$p_t = \frac{1}{|\mathcal{X}|}$

By Bernstein's inequality, w.p. $\geq 1 - \delta$

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

$= |\mathcal{X}|$ $\quad = \frac{1}{|\mathcal{X}|}$

$$\leq c \sqrt{\frac{|\mathcal{X}| \log(2|\Pi|/\delta)}{\xi}}$$

$y_t \in [0, 1]$

$$\text{Regret} = \sum_{t=1}^{T} V(\pi_*) - r(c_t, x_t)$$

$$= \left(\sum_{t=1}^{\tau} V(\pi_\star) - r(c_t, x_t) \right) + \underbrace{\sum_{t=\tau+1}^{T} V(\pi_\star) - r(c_t, \hat{\pi}(c_t))}_{(T-\tau)\left(V(\pi_\star) - V(\hat{\pi})\right)}$$

$$\leq \quad \tau \quad + \quad (T-\tau)\left(V(\pi_\star) - V(\hat{\pi})\right)$$

$$V(\pi_\star) - V(\hat{\pi}) = V(\pi_\star) - \hat{V}(\pi_\star) + \underbrace{\hat{V}(\pi_\star) - \hat{V}(\hat{\pi})}_{\leq 0} + \hat{V}(\hat{\pi}) - V(\hat{\pi})$$

$$\leq 2c\sqrt{\frac{|\mathcal{X}| \log(2|\Pi|/\delta)}{\tau}}$$

$$\leq \quad \tau + T \cdot 2c\sqrt{\frac{|\mathcal{X}| \log(2|\Pi|/\delta)}{\tau}}$$

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

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

But how do you find $\hat{\pi}$?

$$\hat{\pi} = \arg\max_{\pi} \hat{V}_{IPS}(\pi) = \arg\max_{\pi} \frac{1}{\tau} \sum_{t=1}^{\tau} \frac{\mathbb{1}\{x_t = \pi(c_t)\}}{p_t} y_t$$

$$= \arg\max_{\pi} \frac{1}{\tau} \sum_{t=1}^{\tau} \frac{\mathbb{1}\{x_t = \pi(c_t)\}}{p_t} y_t$$

$$= \arg\max_{\pi} \frac{1}{\tau} \sum_{t=1}^{\tau} \frac{(1 - \mathbb{1}\{x_t \neq \pi(c_t)\})}{p_t} y_t$$

$$= \arg\max_{\pi} -\frac{1}{\tau} \sum_{t=1}^{\tau} \frac{\mathbb{1}\{x_t \neq \pi(c_t)\}}{p_t} y_t$$

$$= \arg\min_{\pi} \frac{1}{3} \sum_{t=1}^{J} \frac{y_t}{p_t} \mathbb{1}\{x_t \neq \pi(c_t)\}$$

Problem: a minimizer of this is $\pi(c_t) \neq x_t$ for all c_t. Not necessarily converging to π_*!

- Regularization
- Smaller $|\Pi|$
- Doubly robust estimators.

Elimination alg for $\sqrt{T}$-regret bound

Input Π

Init $\Pi_1 = \Pi$, $T_1 = 0$

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

$\varepsilon_\ell = 2^{-\ell}$. Set $\tau_\ell = \ldots$, $T_\ell = T_{\ell-1} + \tau_\ell$

Define $\lambda \in \Delta_{\Pi_\ell}$

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

Nature reveals context c_t

Draw $\pi_t \sim \lambda$, play $x_t = \pi(c_t)$ set $p_t = \sum_{\pi \in \Pi} \lambda_\pi \mathbb{1}\{x_t = \pi(c_t)\}$

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

$$\hat{V}(\pi) = \frac{1}{T_\ell - T_{\ell-1}} \sum_{t=T_{\ell-1}}^{T_\ell} y_t \frac{\mathbb{1}\{\pi(c_t) = x_t\}}{p_t}$$

$$\Pi_{\ell+1} = \Pi_\ell \setminus \{\pi \in \Pi_\ell : \max_{\pi'} \hat{V}(\pi') - \hat{V}(\pi) > 2\varepsilon_\ell\}$$

Define $Q(x \mid c) = \sum_{\pi \in \Pi} \lambda_\pi \mathbb{1}\{\pi(c) = x\}$

$$\min_{\lambda \in \Delta_\Pi} \max_{\pi \in \Pi} \mathbb{E}\left[\left(\hat{V}(\pi) - V(\pi) \right)^2 \right]$$

$$= \min_{\lambda \in \Delta_\Pi} \max_{\pi \in \Pi} \mathbb{E}_c \left[\frac{1}{Q(\pi(c) \mid c)} \right]$$

<u>Lemma</u> For any policy set Π,

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