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

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

τ-greedy

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

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

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

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

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

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

$y_t \in [0,1]$

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

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

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

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

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

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

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 $\hat{\pi}$?

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

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

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

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

$$= \underset{\pi}{\operatorname{argmin}} \ \frac{1}{S}\sum_{t=1}^{S} \frac{y_t}{p_t} \mathbb{1}\{x_t \neq \pi(c_t)\}$$

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 $S_\ell = 16\,\varepsilon_\ell^{-2} |X| \log\left(\frac{4\ell^2|\Pi|}{\delta}\right)$, $T_\ell = T_{\ell-1} + S_\ell$

Define $\lambda \in \Delta_{\Pi_\ell}$, $Q(x|c) = \sum_{\pi\in\Pi} \lambda_\pi \mathbb{1}\{\pi(c) = x\}$

for $t = T_{\ell-1}+1, \ldots, T_\ell$ where $\gamma = \min\left\{\frac{1}{2|X|}, \sqrt{\frac{\log(2|\Pi|/\delta)}{4|X|S_\ell}}\right\}$

Nature reveals context c_t

Draw $\pi_t \sim \lambda$, Play $x_t = \begin{cases} \pi(c_t) & \text{w.p. } 1-\gamma|X| \\ \text{unif}(X) & \text{w.p. } \gamma|X| \end{cases}$ $p_t = \gamma + (1-\gamma|X|)\,Q(x_t|c_t)$

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

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

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

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,

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

Moreover, if $Q^\gamma(x\mid c) := \gamma + (1-\gamma|\mathcal{X}|)\, Q(x\mid c)$ then $Q^\gamma(x\mid c) \geq \gamma$ but also for $\gamma \leq \frac{1}{2|\mathcal{X}|}$

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

$$\left|\hat{V}^\ell(\pi) - V(\pi)\right| \leq \sqrt{\frac{\mathbb{E}\left[\frac{1}{Q^\gamma(\pi(c)\mid c)}\right] 2\log(2|\Pi|/\delta)}{\tau_\ell}} + \frac{2\log(2|\Pi|/\delta)}{3 \min_{c,x} Q^\gamma(x\mid c)\, \tau_\ell}$$

$$\leq \sqrt{\frac{4|\mathcal{X}|\log(2|\Pi|/\delta)}{\tau_\ell}} + \frac{2\log(2|\Pi|/\delta)}{3\gamma\, \tau_\ell}$$

$$\leq \sqrt{\frac{16|\mathcal{X}|\log(2|\Pi|/\delta)}{\tau_\ell}} \quad \text{if } \gamma = \min\left\{\frac{1}{2|\mathcal{X}|}, \sqrt{\frac{\log(2|\Pi|/\delta)}{4|\mathcal{X}|\tau_\ell}}\right\}$$

$$\le \varepsilon_\ell \quad \text{if} \quad \tau_\ell \ge 16\, \varepsilon_\ell^{-2} |\mathcal{X}| \log(2|\Pi|/\delta)$$

Lemma w.p. $\ge 1-\delta$, $\forall \ell$ $\pi_* \in \Pi_\ell$ and

$$\max_{\pi \in \Pi_\ell} V(\pi_*) - V(\pi) \le 8\varepsilon_\ell.$$

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

$$\le \sum_{\ell=1}^{L} \tau_\ell \Big(\gamma |\mathcal{X}| + \overbrace{\cancel{(1-|\mathcal{X}|\gamma)}}^{\le 1/2} 8\varepsilon_\ell \Big)$$

$$\le \sum_{\ell=1}^{L} \tau_\ell \left(|\mathcal{X}| \cdot \sqrt{\frac{\log(|\Pi|\ell^2/\delta)}{|\mathcal{X}|\tau_\ell}} + 4\varepsilon_\ell \right)$$

Plug in $\tau_\ell = c\, \varepsilon_\ell^{-2} |\mathcal{X}| \log(|\Pi|\ell^2/\delta)$

$$= \sum_{\ell=1}^{L} \tau_\ell \left(c\, \varepsilon_\ell + 4\varepsilon_\ell \right)$$

$$\le \sum_{\ell=1}^{L} \min\left\{ \varepsilon_\ell \tau_\ell,\ \varepsilon_\ell \cdot c\, \varepsilon_\ell^{-2} |\mathcal{X}| \log(|\Pi|\ell^2/\delta) \right\}$$

$$\le \sum_{\ell=1}^{L} \min\left\{ \varepsilon_\ell \tau_\ell,\ c\, \varepsilon_\ell^{-1} |\mathcal{X}| \log(|\Pi|\ell^2/\delta) \right\}$$

$$\le \sum_{\ell=1}^{L} \max_{\varepsilon} \min\left\{ \varepsilon \tau_\ell,\ c\, \varepsilon^{-1} |\mathcal{X}| \log(|\Pi|L^2/\delta) \right\}$$

$\langle x,y\rangle \le \|x\|\cdot\|y\|$

$$= \sum_{\ell=1}^{L} \sqrt{c \quad |\mathcal{X}| \log(|\Pi| L^2/\delta)} \cdot \sqrt{\mathfrak{z}_\ell}$$

$$\le \sqrt{L \cdot c|\mathcal{X}| \log(|\Pi| L^2/\delta)} \cdot \underbrace{\sqrt{\sum_{\ell=1}^{L} \mathfrak{z}_\ell}}_{=\sqrt{T}}$$

$$\le \sqrt{cL|\mathcal{X}| T \log(|\Pi| L/\delta)} \qquad L = \log_2(\Delta_{\min}^{-1}, T)$$

Model-based Contextual Bandits

Given $\mathcal{F}$ and data $\{(c_t, x_t, p_t, y_t)\}$ fit

$$\hat{f} = \underset{f\in\mathcal{F}}{\operatorname{argmin}} \sum_{t=1}^{T} \left(f(c_t, x_t) - y_t\right)^2$$

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

Ex. Linear policy class: $f(c,x) = \phi(x,c)^T\theta$ for $\theta\in\mathbb{R}^d$

and $\phi: \mathcal{C}\times\mathcal{X} \to \mathbb{R}^d$ is a feature map.

Ex. suppose $c_t \in \mathbb{R}^p$ and $\mathcal{X} = \{1,\dots,n\}$ then natural choice $\phi(c,x) = \mathrm{vec}(c\, e_x^\top) \in \mathbb{R}^{pn}$

$$= \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ \boxed{} \\ 0 \\ 0 \\ 0 \end{bmatrix} \begin{matrix} \\ \\ \\ \\ \leftarrow c \\ \\ \\ \\ \end{matrix}$$

Algorithm: UCB

$$\hat{\theta} = \arg\min_{\theta} \sum_{t=1}^{T} \left(y_t - \phi(c_t, x_t)^\top \theta \right)^2 + \gamma \|\theta\|^2$$

Proposition 5. *Fix $\delta \in (0,1)$, $\gamma \geq 0$, and $\theta_* \in \mathbb{R}^d$. Assume for all $t \geq 1$ that $y_t = \langle \theta_*, x_t \rangle + \eta_t$ and $\mathbb{E}[\exp(s\eta_t)|\mathcal{F}_{t-1}] \leq \exp(s^2/2)$ for any $s \in \mathbb{R}$ where $\mathcal{F}_t$ is such that $x_1, y_1, \ldots, x_{t-1}, y_{t-1}, x_t$ are $\mathcal{F}_{t-1}$ measurable. If $S_t = \sum_{i=1}^t x_t \eta_t$, $V_t = \sum_{i=1}^t x_t x_t^\top$, and $\widehat{\theta}_t = (V_t + \gamma I)^{-1} S_t$, then*

$$\|\widehat{\theta}_t - \theta_*\|_{(V_t + \gamma I)} \leq \sqrt{\gamma}\|\theta_*\|_2 + \sqrt{2\log(1/\delta) + \log(\gamma^{-d}|V_t + \gamma I|)}$$

for all $t \geq 1$ simultaneously with probability at least $1-\delta$. Moreover, if $\max_t \|x_t\|_2^2 \leq L$ then $\log(\gamma^{-d}|V_t + \gamma I|) \leq d\log(\frac{tL}{d\gamma} + 1)$.

for $t = 1, 2, \ldots$

Nature reveals context c_t

Player chooses $x_t = \arg\max_{x \in \mathcal{X}} \mathrm{UCB}(c_t, x)$

where $\mathrm{UCB}(c_t, x) = \max_{\theta \in \mathcal{C}_t} \langle \theta, \phi(c_t, x) \rangle$

$$\mathcal{C}_t = \{ \theta : \|\hat{\theta}_t - \theta\|_{V_t + \gamma I} \leq \ldots \}$$

Observe y_t,

Update $\hat{\theta}_t$, V_t

Thompson Sampling

for: $t = 1, 2, 3, \ldots$

Nature reveals c_t

Draw $\theta' \sim N(\hat{\theta}_t, V_t^{-1})$

Play $x_t = \underset{x \in \mathcal{X}}{\arg\max}\ \phi(c_t, x)^T \theta'$