Contextual Bandits - "Model the world"

Input $\mathcal{F}$, $f \in \mathcal{F}$ $f: \mathcal{C} \times \mathcal{A} \to \mathbb{R}$, $|\mathcal{A}| = n$

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

Adversary chooses $c_t \in \mathcal{C}$ arbitrarily

Player chooses $a_t \in [n]$

Nature reveals $y_t \in [0,1]$, $\mathbb{E}[y_t \mid c_t, a_t] = f_*(c_t, a_t)$

minimize Regret $= \sum_{t=1}^{T} \max_a f_*(c_t, a) - f_*(c_t, a_t)$

Assume realizability: $f_* \in \mathcal{F}$

UCB: Assume $\exists \phi: \mathcal{C} \times \mathcal{A} \to \mathbb{R}^d$, $\mathcal{F} = \{\langle \theta, \phi(c,a) \rangle, \theta \in \mathbb{R}^d\}$

$$a_t = \arg\max_a \max_{\theta \in \mathcal{E}_t} \langle \theta, \phi(c_t, a) \rangle$$

$$\mathcal{E}_t = \{\theta : \|\theta - \hat{\theta}_t\|_{V_t}^2 \le \beta_t^2\}$$

We showed that for any sequence of $x_1, x_2, \ldots$ and $y_t = \langle \theta, x_t \rangle + \eta_t$ we have $\|\theta_* - \hat{\theta}_t\|_{V_t}^2 \le \beta_t^2$ w.p. $\ge 1 - \delta$. where $V_t = \lambda I + \sum_s x_s x_s^\top$, $\hat{\theta}_t = V_t^{-1} \sum_s x_s y_s$

so apply same result w/ $x_t := \phi(c_t, a_t)$.

$\Rightarrow$ Regret $\leq d\sqrt{T}$.

---

Thompson Sampling

$$V_t = \lambda I + \sum_{s=1}^{t} \phi(c_t, a_t)\, \phi(c_t, a_t)^T$$

$$\hat{\theta}_t = V_t^{-1} \sum_{s=1}^{t} \phi(c_t, a_t)\, y_t$$

At time $t$, draw

$$\tilde{\theta}_t \sim \mathcal{N}(\hat{\theta}_t, \alpha V_t^{-1})$$

Play $a_t = \underset{a \in 1,\dots,n}{\operatorname{argmax}}\ \phi(c_t, a)^T \tilde{\theta}_t$

Greedy

Play $a_t = \underset{a}{\operatorname{argmax}} \langle \theta_t, \phi(c_t, a) \rangle$

Can work! If $\min_a \sum_t \phi(c_t, a)\, \phi(c_t, a)^T \succ 0$

Thompson sampling is great. How do I generalize it to arbitrary $\mathcal{F}$?

Note for linear

$$\bar{\theta}_t = \left(\sum_{s=1}^{t} x_s x_s^T\right)^{-1} \sum_{s=1}^{t} x_s (y_s + z_s'), \quad z_t' \overset{iid}{\sim} \mathcal{N}(0,1)$$

$$= \hat{\theta}_t + V_t^{-1} \underbrace{\sum_{s=1}^{t} x_s z_s'}$$

$$\sim \mathcal{N}\left(0, V_t^{-1} \underbrace{\left(\sum_s x_s x_s^T\right)}_{V_t} V_t^{-1}\right)$$

$$= \mathcal{N}(0, V_t^{-1})$$

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

Generalize Thompson Sampling:

$$\bar{f}_t = \underset{f \in \mathcal{F}}{\operatorname{argmin}} \sum_{s=1}^{t} \left((y_s + z_s') - f(l_s, a_s)\right)^2$$

Play $\underset{a}{\operatorname{argmax}}\ \bar{f}_t(l_t, a)$.

How do we generalize optimism to arbitrary function classes?

Suppose I collected data $\{(c_t, a_t, y_t)\}_t$

~~$\hat{f} = \underset{f \in \mathcal{F}}{\operatorname{argmin}} \sum_t (f(c_t, a_t) - y_t)^2$~~

$\hat{f}_t \leftarrow \text{ORO}(\{(c_s, a_s, y_s)\}_{s=1}^{t-1})$

$$a_t = \underset{a}{\operatorname{argmax}} \max_{f \in H_t} f_t(c_t, a)$$

where

$$H_t = \left\{ f : \sum_{t=1}^{T} \mathbb{E}_{a_t \sim p_t} \left[ \left( \hat{f}_t(c_t, a_t) - f(c_t, a_t) \right)^2 \right] \leq \log |\mathcal{F}| / \delta \right\}$$

Generalized UCB. ↗

Online Regression Oracle At each time $t$ we have $\{(c_s, a_s, y_s)\}_{s=1}^{t-1}$ w/ $\mathbb{E}[y_s | c_s, a_s] = f_*(c_s, a_s)$ and $a_s \sim p_s \in \Delta_n$. An ORO returns $\hat{f}_t$:

$$\sum_{t=1}^{T} \mathbb{E}_{a_t \sim p_t}\left[\left(\hat{f}_t(c_t, a_t) - f_{a_t}(c_t, a_t)\right)^2\right] \le Est(\mathcal{F}, \delta, T)$$

w.p. $\ge 1-\delta$.

**Theorem** Play Exponential weights over $\mathcal{F}$ with loss $\ell_{t,i} = \left(f^i(c_t, a_t) - y_t\right)^2$ and $\xi = 1/2$

and output $\hat{f}_t(c, a) = \sum_i f^i(c, a) p_{t,i}$ then

$$Est(\mathcal{F}, \delta, T) \le \log(|\mathcal{F}|/\delta)$$

**Theorem** $\exists\, \mathcal{F}$ and distribution of contexts such that Generalized UCB suffers linear regret

**Proof** $\mathcal{C} = [m]$ $\mathcal{A} = \{1, 2\}$ Fix $\varepsilon > 0$

$c_t \sim$ uniform $([m])$

$f_*(c, 1) = 1 - \varepsilon$

$f_*(c, 2) = 0$

$y_t = f_{a_t}(c_t, a_t)$

$f_i(c_i, 1) = 0$

$f_i(c_i, 2) = 1$ $\forall i$

$\Rightarrow$ Regret $\ge \min\{|\mathcal{C}|, |\mathcal{F}|\}$