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

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

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

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

Optimism can be a very bad idea.

ε-Greedy

Input: $\varepsilon > 0$

$$P_t(a) = \frac{\varepsilon}{A} + \mathbb{1}\{\hat{a}_t = a\}(1-\varepsilon)$$

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

Obtain $\hat{f}_t$ from $ORO(\{c_s, a_s, y_s\}_{s<t})$

Observe context c_t

w/ prob ε play $a_t \sim \text{uniform}([n])$

w/ prob $1-\varepsilon$ play $a_t = \hat{a}_t = \operatorname{argmax}_a \hat{f}_t(c_t, a)$

Observe reward $y_t \in [0,1]$

Theorem Assume ORO satisfies $Est(\mathcal{F}, T, \delta) \le \log(|\mathcal{F}|/\delta)$.

Then ε-greedy satisfies

$$\sum_{t=1}^{T} f_{t^*}(c_t, a_{t^*}(c_t)) - f_{t^*}(c_t, a_t) \le n^{1/3} T^{2/3} \log(|\mathcal{F}|/\delta)^{1/3}$$

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

proof

$$\sum_{t=1}^{T} f_{t^*}(c_t, a_{t^*}(c_t)) - f_{t^*}(c_t, a_t)$$

$$\le \varepsilon T + \sum_{t=1}^{T} f_{t^*}(c_t, a_{t^*}(c_t)) - f_{t^*}(c_t, \hat{a}_t)$$

$a_* := a_*(c_t)$

$$f_*(c_t, a_*) - f_*(c_t, \hat{a}_t) = f_*(c_t, a_*) - \hat{f}(c_t, a_*) + \hat{f}(c_t, a_*) - \hat{f}(c_t, \hat{a}_t) \le 0 + \hat{f}(c_t, \hat{a}_t) - f_*(c_t, \hat{a}_t)$$

$$P_t(a) = \frac{\varepsilon}{A} + \mathbb{1}\{\hat{a}_t = a\}(1-\varepsilon)$$

$$\le \sum_{a \in \hat{a}_t, a_*} |f_*(c_t, a) - \hat{f}(c_t, a)|$$

$$= \sum_{a \in \hat{a}_t, a_*} \frac{1}{\sqrt{P_t(a)}} \cdot \sqrt{P_t(a)}\, |f_*(c_t, a) - \hat{f}(c_t, a)|$$

$$\le \left(\frac{2n}{\varepsilon}\right)^{1/2} \left(\sum_{a \in \hat{a}_t, a_*} P_t(a)\, |f_*(c_t, a) - \hat{f}(c_t, a)|^2 \right)^{1/2}$$

$$\le \left(\frac{2n}{\varepsilon}\right)^{1/2} \mathbb{E}_{a \sim P_t}\left[\left(f_*(c_t, a) - \hat{f}(c_t, a) \right)^2 \right]^{1/2}$$

$$\text{Regret}_T \le \varepsilon T + \sum_{t=1}^{T} \sqrt{\frac{2n}{\varepsilon} \mathbb{E}_{a \sim P_t}\left[\left(f_*(c_t, a) - \hat{f}(c_t, a) \right)^2 \right]}$$

$$\le \varepsilon T + \sqrt{\frac{2nT}{\varepsilon} \log(|\mathcal{F}|/\delta)}$$

$$\lesssim T^{2/3} \cdot \left(2n \log(|\mathcal{F}|/\delta)\right)^{1/3} \cdot \quad \text{for some } \varepsilon.$$

Inverse Gap Weighting

Fix $\gamma > 0$ and $g : [n] \to \mathbb{R}$, let $\hat{a} = \operatorname{argmax}_a g(a)$.

Define $$IGW_\gamma(a) := p(a) = \frac{1}{\lambda + 2\gamma(g(\hat{a}) - g(a))}$$

where λ is chosen s.t. $\sum_a p(a) = 1$.

$$\frac{1}{\lambda} \leq \sum_a p(a) \leq \frac{n}{\lambda} \Rightarrow \lambda \in [1, n]$$

Lemma Fix some $g : [n] \to \mathbb{R}$ and $p = IGW_\gamma$

Then for any $g_* : [n] \to \mathbb{R}$ we have

$$\mathbb{E}_{a\sim p}[g_*(a_*) - g_*(a)] \leq \frac{n}{\gamma} + \gamma \mathbb{E}_{a\sim p}[(g(a) - g_*(a))^2]$$

where $a_* = \operatorname{argmax}_a g_*(a)$.

Proof

$$\mathbb{E}_{a\sim p}[g_*(a_*) - g_*(a)] = \underbrace{\mathbb{E}_{a\sim p}[g(\hat{a}) - g(a)]}_{(I)} + \underbrace{\mathbb{E}_{a\sim p}[g(a) - g_*(a)]}_{(II)}$$

$$g_*(a_*) - g(\hat{a}) \quad (III)$$

(I) $\mathbb{E}_{a\sim p}[g(\hat{a})-g(a)] = \sum_a p(a)\cdot(g(\hat{a})-g(a))$

$$= \sum_a \frac{(g(\hat{a})-g(a))}{\lambda+2\gamma(g(\hat{a})-g(a))}$$

$$\leq \frac{n}{2\gamma}$$

(II) $u \leq \frac{1}{2\gamma} + \frac{\gamma}{2}u^2$

$$\mathbb{E}_{a\sim p}[g(a)-g_\phi(a)] \leq \frac{1}{2\gamma} + \frac{\gamma}{2}\mathbb{E}_{a\sim p}[(g(a)-g_\phi(a))^2]$$

(III) $u \leq \frac{1}{2\gamma p(a_A)} + \frac{\gamma p(a_A)}{2}u^2$

$$g_\phi(a_A)-g(a_A)$$

$$\leq \frac{1}{2\gamma p(a_A)} + \frac{\gamma}{2}p(a_A)(g_\phi(a_A)-g(a_A))^2$$

$$p(a_n) = \frac{1}{\lambda + 2\gamma\left(g(\hat{a}) - g(a_n)\right)} \geq$$

$$\frac{1}{2\gamma p(a_n)} = \frac{\lambda + 2\gamma\left(g(\hat{a}) - g(a_n)\right)}{2\gamma}$$

$$g(a_n) - g(\hat{a})$$

$$= g_n(a_n) - g(a_n) + g(a_n) - g(\hat{a})$$

$$\leq \frac{1}{2\gamma p(a_n)} + \frac{\gamma}{2} p(a_n)\left(g_n(a_n) - g(a_n)\right)^2 + g(a_n) - g(\hat{a})$$

$$= \frac{\lambda + 2\gamma\left(g(\hat{a}) - g(a_n)\right)}{2\gamma} + \frac{\gamma}{2} p(a_n)\left(g_n(a_n) - g(a_n)\right)^2 + g(a_n) - g(\hat{a})$$

$$\leq \frac{n}{2\gamma} + \frac{\gamma}{2}\mathbb{E}_{a\sim p}\left[\left(g_n(a_n) - g(a_n)\right)^2\right]$$

SquareCB

Input $\gamma > 0$

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

get $\hat{f}_t$ from $\mathrm{ORO}(\{c_s, a_s, y_s\}_{s<t})$

Observe c_t, let $\hat{a}_t = \operatorname{argmax}_a \hat{f}_t(c_t, a)$

Play $a_t \sim p_t(c_t, a) = \dfrac{1}{\lambda + \gamma\left(\hat{f}_t(c_t, \hat{a}_t) - \hat{f}_t(c_t, a)\right)}$

Observe y_t

Theorem SquareCB enjoys regret $\frac{nT}{\gamma} + \gamma \log |\mathcal{F}|/\delta$ w.p. $\geq 1 - \delta$.

$$\sum_t \mathbb{E}_{a \sim p_t}\left[f_{\star}(c_t, a_{\star}(c_t)) - f_{\star}(c_t, a) \right]$$

$$\leq \frac{nT}{\gamma} + \gamma \sum_{t=1}^{T} \mathbb{E}_{a \sim p_t}\left[\left(f_{\star}(c_t, a) - \hat{f}_t(c_t, a) \right)^2 \right]$$

$$\leq \frac{nT}{\gamma} + \gamma \log(|\mathcal{F}|/\delta) \leq \sqrt{nT \log(|\mathcal{F}|/\delta)}.$$