Contextual Bandits

for $t = 1, 2, \ldots$

Nature reveals context $c_t \in \mathcal{C}$

Player chooses $I_t \in [n]$

Player gets reward $y_t \in [0,1]$, $\mathbb{E}[y_t \mid c_t, I_t] = r(c_t, I_t)$

Model-free

Makes no assumptions about $r : \mathcal{C} \times [n] \to [0,1]$

Given Π collection of policies $\pi \in \Pi$, $\pi : \mathcal{C} \to [n]$

$$R_T^{\Pi} = \max_{\pi \in \Pi} \mathbb{E}\left[\sum_{t=1}^{T} r(c_t, \pi(c_t)) - r(c_t, I_t) \right]$$

where if "proper learning" $I_t = \pi_t(c_t)$ for some $\pi_t \in \Pi$.

We showed $R_T^{\Pi} \lesssim \sqrt{n T \log(|\Pi|)}$

Model-based

Given collection functions $\mathcal{F}$, $f \in \mathcal{F}$, $f : \mathcal{C} \times [n] \to [0,1]$

and assume $\exists f_* \in \mathcal{F} : r(c, i) = f_*(c, i) \ \forall c, i$.

$$R_T^{\mathcal{F}} = \mathbb{E}\left[\sum_{t=1}^{T} \max_i r(c_t, i) - r(c_t, I_t) \right]$$

<u>Claim</u>: Given online learning oracle, $\exists$ alg $R_T^{\mathcal{F}} \lesssim \sqrt{n T \log |\mathcal{F}|}$

Linear models

Assume you know a feature map $\phi(c,i) \to \mathbb{R}^d$

and assume $r(c,i) = \phi(c,i)^T \theta_*$ for some

$\theta_* \in \mathbb{R}^d$. If I collect dataset $(\{c_t, I_t, y_t\})$

$$\hat{\theta} = \underset{\theta}{\operatorname{argmin}} \sum_{t=1}^{T} \left(\phi(c_t, I_t)^T \theta - y_t \right)^2 + \lambda \|\theta\|_2^2$$

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

$$V_t = \sum_{t=1}^{T} \phi(c_t, I_t) \phi(c_t, I_t)^T + \lambda I_d$$

$$UCB_t(i) = \max_{\theta \in \mathcal{E}_t} \langle \theta, \phi(c_t, i) \rangle$$

$$I_t = \underset{i}{\operatorname{argmax}}\ UCB_t(i)$$

$$R_T^{\mathcal{F}} \le d\sqrt{T}\, \text{polylog}(T)$$ for an choice of β_t

Generalizing UCB

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

$$\mathcal{F}^T = \left\{ f \in \mathcal{F} : \sum_{t=1}^{T} \left(f(c_t, I_t) - y_t \right)^2 \leq \sum_{t=1}^{T} \left(\hat{f}_T(c_t, I_t) - y_t \right)^2 + \beta_T \right\}$$

If $\beta_T = 8 \log |\mathcal{F}|/\delta$ then $f_* \in \mathcal{F}^T$ $\forall T$.

$$UCB_t(i) = \max_{f \in \mathcal{F}^T} f(c_t, i)$$

$$I_t = \operatorname*{argmax}_{i \in [n]} UCB_t(i).$$

$$\sum_t \max_i r(c_t, i) - r(c_t, I_t)$$

$$\leq \sum_t UCB_t(I_t) - r(c_t, I_t)$$

$\exists \mathcal{F}$ s.t. UCB suffers regret $\geq \sqrt{T |\mathcal{F}|}$

Cheat-code or index function.

fix m, $m=3$

$000 \to e_1$
$001 \to e_2$
$010 \to e_3$
$011 \to e_4$
100
101 $\vdots$
110
$111 \to e_{2^m}$

$$\begin{bmatrix} 0 \\ 1 \\ 0 \\ e_3 \end{bmatrix} \in \{0,1\}^{m+2^m}$$

$i = 1, \dots, 2^m$

$$f_i = \begin{bmatrix} -\boxed{b(i)} \\ e_i \end{bmatrix}$$

bit string i

UCB explores 2^m bits

resulting regret $\sqrt{2^m T}$

An alg that concentrates on just first m bits, learns $b(i_*)$ and hence i_* and hence f_*.

And suffers regret $\sqrt{mT}$

General Opt Alg. : Square CB

Input $\gamma > 0$

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

Obtain estimate $\hat{f}_t$ given $\{(c_s, I_s, y_s)\}_{s<t}$

Nature reveals c_t

$$P_{t,j} \propto \frac{1}{\lambda + \gamma\left(\max_i \hat{f}_t(c_t, i) - \hat{f}_t(c_t, j)\right)}$$

Play $I_t \sim P_t$ and get y_t

Thm | $R_T^{\mathcal{F}} \le \sqrt{nT\, Est_{s_2}(T)}$

We say $\hat{f}_t$ is an online regression oracle

if $$\sum_{t=1}^{T}\left(\hat{f}_t(c_t, I_t) - f_*(c_t, I_t)\right)^2 \le Est_{s_2}(T)$$

If $\hat{f}_t$ satisfies

$$\sum_{t=1}^{T}\left(\hat{f}_t(c_t, I_t) - y_t\right)^2 \le \sum_{t=1}^{T}\left(f_*(c_t, I_t) - y_t\right) + Est_{s_2}(T)$$

then

$$\sum_t \ell_t(\hat{f}_t) - \sum_t \ell_t(f_*) \le$$

W/ Exp Weights $Est_{s_2}(T) \le \log|\mathcal{F}|$