1) Setting: Contextual Bandits

Simplified RL setting: simplified version of state: context is memoryless, drawn iid at each timestep

\[ \chi: \text{set of contexts } \chi \]
\[ \mathcal{A} = \{1, \ldots, K\} \text{ a set of discrete actions} \]
\[ \Theta \in \Delta(\chi): \text{context distribution } x_t \sim \Theta \]
\[ r: \chi \times \mathcal{A} \rightarrow \Delta(\mathbb{R}) \text{ noisy reward } r_t \sim r(x_t, a_t) \]
\[ \mathbb{E}\left[r(x, a)\right] = \hat{y}_a(x) \]

T: time horizon

Actions should depend on the context,

\[ \pi: \chi \rightarrow \mathcal{A} \text{ (or } \pi(a|\chi) \text{ stochastic)} \]

Optimal Policy: \[ \pi^*(x) = \arg\max_a y_a(x) \]

Minimize Expected Regret:

\[ R(T) = \sum_{t=1}^{T} \mathbb{E}\left[ y^*(x_t) - y_{a_t}(x_t) \right] \]

2) Tabular Setting

Suppose M contexts

IDEA: run a separate MAB algorithm for each context

Alg: Explore-then-Commit w/ Context

For t = 1, 2, ..., T

Observe \( x_t \)

1) if \( \exists \) arm pulled less than N times - explore

2) otherwise, \( a_t = \arg\max_a \hat{y}_a(x_t) \) - exploit
**Alg: UCB w/ contexts**

For $t = 1, 2, \ldots, T$:

- pull $a_t = \arg\max_a \hat{\mu}_t^a(X_t) + \sqrt{\frac{\log(TM/s)}{N_t^a(X_t)}}$

  \[ \Downarrow \]

  Context-dependent mean & count

**K-M policies:** Regret bounds will be similar to last 2 lectures with K replaced w/ $K$ - $M$

3) Function Approximation

We may never see the same context twice!

- ex: user 1: $\mathcal{E} F$, 22, $\mathcal{C} \mathcal{S} \mathcal{X} = X_1$
- user 2: $\mathcal{E} M$, 21, $\mathcal{E} \mathcal{C} \mathcal{A} \mathcal{N} = X_2$
- user 3: $\mathcal{E} F$, 20, $\mathcal{E} \mathcal{C} \mathcal{A} \mathcal{N} = X_3$

Instead of estimating $\hat{\mu}_a(x)$ with counting, we can use function approximation:

\[ \hat{\gamma}_a(x) = \arg\min_{\gamma \in \mathcal{M}} \sum_{k=1}^T (\gamma(x_k) - \nu_k)^2 \prod \mathcal{E} a_t = a^3 \]

How to get CI on $\hat{\gamma}_a(x)$?

Error bounds for supervised learning
Lemma: for $x_i \sim D$, $\mathbb{E}[y_i] = f_x(x_i)$, $f_x \in \mathcal{F}$

$$\hat{f} = \arg\min_{f \in \mathcal{F}} \sum_{i=1}^{N} (f(x_i) - y_i)^2$$

Then with high probability, $\mathbb{E}_{x \sim D}[|\hat{f}(x) - f_x(x)|] \leq \sqrt{\frac{C_{\mathcal{F}}}{N}}$ complexity of $\mathcal{F}$

Algorithm: Explore-then-Commit w/ SL

1) Pull each arm $N$ times, record $\xi \in (x_i^{a}, r_i^{a})$ for $i = 1, \ldots, N$, $a = 1, \ldots, K$

2) $t = NK + 1, \ldots, T$: pull $a_t = \arg\max_{a} \hat{y}_a(x_t)$

Regret Analysis:

$R(T) = \sum_{t=NK+1}^{T} \mathbb{E}[y^{*}(x_t) - y_{a_t}(x_t)]$

$\mathbb{E}[y^{*}(x_t) - y_{a_t}(x_t)] = \mathbb{E}[(y^{*}(x_t) - \hat{y}^{*}(x_t)) + (\hat{y}^{*}(x_t) - \hat{y}_{a_t}(x_t)) + (\hat{y}_{a_t}(x_t) - y_{a_t}(x_t))]$  \(\leq 0\)

$a_t = \arg\max_{a} \hat{y}_a(x_t)$

$\leq \mathbb{E}_{x \sim D}[|y^{*}(x_t) - \hat{y}^{*}(x_t)|] + \mathbb{E}_{x \sim D}[|\hat{y}_{a_t} - y_{a_t}(x_t)|]$  \(\leq 2\sqrt{\frac{C_{\mathcal{F}}}{N}}\)
\( R(T) \leq N K + 2 T \sqrt{\frac{C M}{N}} \)

\( R(T) \lesssim T^{2/3} (K C M)^{1/3} \)

\( N = \left( \frac{I}{2 K \sqrt{C M}} \right)^{2/3} \)

UCB algorithm?

naive: \( \hat{Y}_a(x) + \sqrt{\frac{C M}{N_a}} \)

we want confidence intervals based on conditional expected error \( \mathbb{E}[|y(x)-\hat{y}(x)| \mid x] \)

next lecture: LinUCB algorithm

\( y_a(x) = \Theta_a^T x \)

general contextual setting

\( y_a(x) = \Theta_a^T \phi(x, a) \)