1) Setting: Contextual Bandits

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

$\mathcal{X}$: set of contexts x

$\mathcal{A} = \{1, \ldots, K\}$ a set of discrete actions

$\mathcal{D} \in \Delta(\mathcal{X})$: context distribution $x_t \overset{iid}{\sim} \mathcal{D}$

$r: \mathcal{X} \times \mathcal{A} \to \Delta(\mathbb{R})$ noisy reward $r_t \sim r(x_t, a_t)$
depends on contex & action
$\mathbb{E}[r(x,a)] = \mu_a(x)$

T: time Horizon

Actions should depend on the context,
$\pi: \mathcal{X} \to \mathcal{A}$ (or $\pi(a|s)$ stochastic)

Optimal Policy: $\pi^*(x) = \arg\max_a \mu_a(x)$

Minimize Expected Regret:

$$R(T) = \sum_{t=1}^{T} \mathbb{E}_{x_t \sim \mathcal{D}}\left[\mu^*(x_t) - \mu_{a_t}(x_t)\right]$$

$\mu^*(x_t) = \max_a \mu_a(x_t)$

2) Tabular Setting

Suppose M contexts

IDEA: run a separate MAB algorithm for each context

Alg: Explore-then-Commit w/ Context

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

Observe x_t

1) if $\exists$ arm pulled less than N times for context x_t, pull it — explore
2) otherwise, $a_t = \arg\max_a \hat{\mu}_a(x_t)$ — exploit

Alg: UCB w/ contexts

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

$$\text{pull } a_t = \arg\max_a \hat{\mu}_t^a(X_t) + \sqrt{\frac{\log(TKM/\delta)}{N_t^a(X_t)}}$$

context-dependent mean & count

$K \cdot M$ policies: Regret bounds will be similar to last 2 lectures with K replaced w/ $\underbrace{K \cdot M}_{\text{\# policies}}$

K: # arms, M: # contexts

3) Function Approximation

We may never see the same context twice!

ex: user1: {F, 22, CS} = X_1
user 2: {M, 21, econ} = X_2
user 3: {F, 20, econ} = X_3

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

$$\hat{\mu}_a(x) = \underset{\mu \in M}{\arg\min} \sum_{k=1}^{t} (\mu(x_k) - r_k)^2 \, \mathbb{1}\{a_t = a\}$$

M: function class

How to get CI on $\hat{\mu}_a(x)$?
Error bounds for supervised learning

<u>Lemma:</u> for $x_i \sim \mathcal{D}$, $\mathbb{E}[y_i] = f_*(x_i)$ $\quad f_* \in \mathcal{F}$

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

Then with high probability,

$$\mathbb{E}_{x \sim \mathcal{D}}[|\hat{f}(x) - f_*(x)|] \lesssim \sqrt{\frac{C_{\mathcal{F}}}{N}} \leftarrow \text{complexity of } \mathcal{F}$$

<u>Algorithm: Explore-then-Commit w/ SL</u>

1) Pull each arm N times, record $\{\{(x_i^a, r_i^a)\}_{i=1}^N\}_{a=1}^K$
$(t = 1, \dots, NK)$
Estimate $\hat{\mu}_a(x) = \underset{\mu}{\arg\min} \sum_{i=1}^{N} (\mu(x_i^a) - r_i^a)^2$

2) $t = NK+1, \dots, T$: pull $a_t = \underset{a}{\arg\max}\ \hat{\mu}_{a_t}(x_t)$

Regret Analysis:

$$R(T) \overset{\Delta}{=} \underbrace{R_1}_{\leq NK} + \underbrace{R_2}_{= \sum_{NK+1}^{T} \mathbb{E}_{x_t \sim \mathcal{D}}[\mu_{a^*}(x_t) - \mu_{a_t}(x_t)]}$$

$$\mathbb{E}[\mu_{a^*}(x_t) - \mu_{a_t}(x_t)] = \mathbb{E}\Big[(\mu_{a^*}(x_t) - \hat{\mu}_{a^*}(x_t)) + \underbrace{\cancel{\hat{\mu}_{a^*}(x_t) - \hat{\mu}_{a_t}(x_t)}}_{\leq 0} + (\hat{\mu}_{a_t}(x_t) - \mu_{a_t}(x_t))\Big]$$

$a_t = \underset{a}{\arg\max}\ \hat{\mu}_a(x_t)$

$$\leq \mathbb{E}_{x_t \sim \mathcal{D}}[|\mu_{a^*}(x_t) - \hat{\mu}_{a^*}(x_t)|] + \mathbb{E}_{x_t \sim \mathcal{D}}[|\hat{\mu}_{a_t} - \mu_{a_t}(x_t)|]$$

$$\lesssim 2\sqrt{\frac{C_M}{N}}$$

$$R(T) \lesssim \overset{R_1}{NK} + \overset{R_2}{2T\sqrt{\frac{C_M}{N}}} \qquad N = \left(\frac{T}{2K}\sqrt{C_M}\right)^{2/3}$$

$$R(T) \lesssim T^{2/3}(KC_M)^{1/3}$$

UCB algorithm?

Naive: $\underline{\hat{y}_a^t(x)} + \sqrt{\frac{C_M}{N_a^t}}$

We want confidence intervals based on $\underline{\text{conditional}}$ expected error

$$\mathbb{E}[\,|y(x) - \hat{y}(x)|\,|\,x\,]$$

Next lecture: LinUCB algorithm

$$y_a(x) = \theta_a^T x$$

General contextual setting

$$y_a(x) = \theta_*^T \phi(x, a)$$