Contextual Bandits

Lucas Janson

CS/Stat 184(0): Introduction to Reinforcement Learning

Fall 2024

Today

- Feedback from last lecture
- Recap
- UCB-VI for linear MDPs
- Recall: Contextual Bandits
- LinUCB

Feedback from feedback forms

1. Thank you to everyone who filled out the forms!

Today

- ✓ Feedback from last lecture
- Recap
- UCB-VI for linear MDPs
- Recall: Contextual Bandits
- LinUCB

Exploration in MDP: make it a bandit and do UCB?

Q: given a discrete MDP, how many unique deterministic policies are there?

$$\left(|A|^{|S|}\right)^{H}$$

So treating each policy as an "arm" and running UCB gives us regret $\tilde{O}(\sqrt{|A|^{|S|H} N})$

This seems bad, so are MDPs just super hard or can we do better?

Tabular UCB-VI

For $n = 1 \to N$:

1. Set $N_h^n(s,a) = \sum_{i=1}^{n-1} \mathbf{1}\{(s_h^i, a_h^i) = (s,a)\}, \forall s, a, h$
2. Set $N_h^n(s,a,s') = \sum_{i=1}^{n-1} \mathbf{1}\{(s_h^i, a_h^i, s_{h+1}^i) = (s,a,s')\}, \forall s, a, a', h$
3. Estimate $\hat{P}^n$: $\hat{P}_h^n(s'|s,a) = \dfrac{N_h^n(s,a,s')}{N_h^n(s,a)}, \forall s, a, s', h$
4. Plan: $\pi^n = \text{VI}\left(\{\hat{P}_h^n, r_h + b_h^n\}_h\right)$, with $b_h^n(s,a) = cH\sqrt{\dfrac{\log(|S||A|HN/\delta)}{N_h^n(s,a)}}$
5. Execute π^n : $\{s_0^n, a_0^n, r_0^n, \ldots, s_{H-1}^n, a_{H-1}^n, r_{H-1}^n, s_H^n\}$

High-level Idea: Exploration Exploitation Tradeoff

Upper bound per-episode regret: $V_0^\star(s_0) - V_0^{\pi^n}(s_0) \leq \hat{V}_0^n(s_0) - V_0^{\pi^n}(s_0)$ by construction of b_h^n

1. What if $\hat{V}_0^n(s_0) - V_0^{\pi^n}(s_0)$ is small?

Then π^n is close to $\pi^\star$, i.e., we are doing exploitation

2. What if $\hat{V}_0^n(s_0) - V_0^{\pi^n}(s_0)$ is large?

Some $b_h^n(s, a)$ must be large (or some $\hat{P}_h^n(\cdot \mid s, a)$ estimation errors must be large, but with high probability any $\hat{P}_h^n(\cdot \mid s, a)$ with high error must have small $N_h^n(s, a)$ and hence high $b_h^n(s, a)$)

Large $b_h^n(s, a)$ means π^n is being encouraged to do (s, a), since it will apparently have very high reward, i.e., exploration

$$\mathbb{E}\left[\text{Regret}_N\right] := \mathbb{E}\left[\sum_{n=1}^{N}\left(V^\star - V^{\pi^n}\right)\right] \leq \widetilde{O}\left(H^2\sqrt{|S||A|N}\right)$$

Today

- ✓ Feedback from last lecture
- ✓ Recap
- UCB-VI for linear MDPs
- Recall: Contextual Bandits
- LinUCB

Linear MDP Definition

Finite horizon time-dependent episodic MDP $\mathcal{M} = \{S, A, H, \{r\}_h, \{P\}_h, s_0\}$

S & A could be large or even continuous, hence poly$(|S|, |A|)$ is not acceptable

$$P_h(s'|s,a) = \mu_h^\star(s') \cdot \phi(s,a), \quad \mu_h^\star : S \mapsto \mathbb{R}^d, \quad \phi : S \times A \mapsto \mathbb{R}^d$$

$$r(s,a) = \theta_h^\star \cdot \phi(s,a), \quad \theta_h^\star \in \mathbb{R}^d$$

Feature map ϕ is known to the learner!

(We assume reward is known, i.e., $\theta^\star$ is known)

Planning in Linear MDP: Value Iteration

$$P_h(\cdot \mid s, a) = \mu_h^\star \phi(s, a), \quad \mu_h^\star \in \mathbb{R}^{|S| \times d}, \ \phi(s, a) \in \mathbb{R}^d$$

$$r_h(s, a) = (\theta_h^\star)^\top \phi(s, a), \quad \theta_h^\star \in \mathbb{R}^d$$

$$V_H^\star(s) = 0, \forall s,$$

$$Q_h^\star(s, a) = r_h(s, a) + \mathbb{E}_{s' \sim P_h(\cdot|s,a)} V_{h+1}^\star(s')$$

$$= \theta_h^\star \cdot \phi(s, a) + \left(\mu_h^\star \phi(s, a)\right)^\top V_{h+1}^\star$$

$$= \phi(s, a)^\top \left(\theta_h^\star + (\mu_h^\star)^\top V_{h+1}^\star\right)$$

$$= \phi(s, a)^\top w_h$$

$$V_h^\star(s) = \max_a \phi(s, a)^\top w_h, \quad \pi_h^\star(s) = \arg\max_a \phi(s, a)^\top w_h$$

Indeed we can show that $Q_h^\pi(\cdot, \cdot)$
Is linear with respect to ϕ as well, for any π, h

UCBVI in Linear MDPs

At the beginning of iteration n:

1. Learn transition model $\{\hat{P}_h^n\}_{h=0}^{H-1}$ from all previous data $\{s_h^i, a_h^i, s_{h+1}^i\}_{i=0}^{n-1}$

2. Design reward bonus $b_h^n(s, a), \forall s, a$

3. Plan: $\pi^{n+1} = \text{VI}\left(\{\hat{P}^n\}_h, \{r_h + b_h^n\}\right)$

How to estimate $\{\hat{P}_h^n\}_{h=0}^{H-1}$?

Denote $\delta(s) \in \mathbb{R}^{|S|}$ with zero everywhere except the entry corresponding to s

Given s, a, note that $\mathbb{E}_{s' \sim P_h(\cdot|s,a)}\left[\delta(s')\right] = P_h(\cdot \mid s, a) = \mu_h^\star \phi(s, a)$

Penalized Linear Regression:

$$\min_{\mu} \sum_{i=1}^{n-1} \|\mu\phi(s_h^i, a_h^i) - \delta(s_{h+1}^i)\|_2^2 + \lambda\|\mu\|_F^2$$

$$A_h^n = \sum_{i=1}^{n-1} \phi(s_h^i, a_h^i)\phi(s_h^i, a_h^i)^\top + \lambda I \qquad \widehat{\mu}_h^n = (A_h^n)^{-1} \sum_{i=1}^{n-1} \delta(s_{h+1}^i)\phi(s_h^i, a_h^i)^\top$$

$$\hat{P}_h^n(\cdot \mid s, a) = \hat{\mu}_h^n \phi(s, a)$$

How to choose $b_h^n(s,a)$?

Chebyshev-like approach, similar to in linUCB (will cover later this lecture):

$$b_h^n(s,a) = \beta\sqrt{\phi(s,a)^\top (A_h^n)^{-1}\phi(s,a)}, \quad \beta = \widetilde{O}(dH)$$

linUCB-VI: Put All Together

For $n = 1 \rightarrow N$:

1. Set $A_h^n = \sum_{i=1}^{n-1} \phi(s_h^i, a_h^i)\phi(s_h^i, a_h^i)^\top + \lambda I$
2. Set $\widehat{\mu}_h^n = (A_h^n)^{-1} \sum_{i=1}^{n-1} \delta(s_{h+1}^i)\phi(s_h^i, a_h^i)^\top$
3. Estimate $\hat{P}^n$: $\hat{P}_h^n(\cdot \mid s, a) = \widehat{\mu}_h^n \phi(s, a)$
4. Plan: $\pi^n = \text{VI}\left(\{\hat{P}_h^n, r_h + b_h^n\}_h\right)$, with $b_h^n(s, a) = cdH\sqrt{\phi(s, a)^\top (A_h^n)^{-1} \phi(s, a)}$
5. Execute π^n : $\{s_0^n, a_0^n, r_0^n, \ldots, s_{H-1}^n, a_{H-1}^n, r_{H-1}^n, s_H^n\}$

$$\mathbb{E}\left[\text{Regret}_N\right] := \mathbb{E}\left[\sum_{n=1}^{N}\left(V^\star - V^{\pi^n}\right)\right] \leq \widetilde{O}\left(H^2 d^{1.5}\sqrt{N}\right)$$

No S, A dependence!

Today

- Feedback from last lecture
- Recap
- UCB-VI for linear MDPs
- Recall: Contextual Bandits
- LinUCB

Recall: (non-contextual) bandit

We have K many arms; label them $1, \ldots, K$

Each arm has an unknown reward distribution, i.e., $\nu_k \in \Delta([0,1])$,

w/ mean $\mu_k = \mathbb{E}_{r \sim \nu_k}[r]$

For $t = 0 \rightarrow T-1$

(based on historical information)

1. Learner pulls arm $a_t \in \{1, \ldots, K\}$

2. Learner observes an i.i.d reward $r_t \sim \nu_{a_t}$ of arm a_t

$$\text{Regret}_T = T\mu^\star - \sum_{t=0}^{T-1} \mu_{a_t} = \sum_{t=0}^{T-1} (\mu^\star - \mu_{a_t})$$

Recall: Beyond simple bandits

In a bandit, we are presented with the same decision at every time

In practice, often decisions are not the same every time

E.g., in online advertising there may not be a single best ad to show all users on all websites:

- maybe some types of users prefer one ad while others prefer another, or
- maybe one type of ad works better on certain websites while another works better on other websites

Which user comes in next is random, but we have some context to tell situations apart and hence learn different optimal actions

Recall: Contextual bandit environment

Context at time t encoded into a variable x_t that we see before choosing our action

x_t is drawn i.i.d. at each time point from a distribution ν_x on sample space $\mathscr{X}$

x_t then affects the reward distributions of each arm, i.e., if we choose arm k, we get a reward that is drawn from a distribution that depends on x_t, namely, $\nu^{(k)}(x_t)$

Accordingly, we should also choose our action a_t in a way that depends on x_t, i.e., our action should be chosen by a function of x_t (a policy), namely, $\pi_t(x_t)$

If we knew everything about the environment, we'd want to use the optimal policy

$$\pi^\star(x_t) := \arg\max_{k \in \{1,\ldots,K\}} \mu^{(k)}(x_t), \qquad \text{where } \mu^{(k)}(x) := \mathbb{E}_{r \sim \nu^{(k)}(x)}[r]$$

$\pi^\star$ is the policy we compare to in computing regret

Recall: Contextual bandit environment

Formally, a contextual bandit is the following interactive learning process:

For $t = 0 \to T - 1$

1. Learner sees context $x_t \sim \nu_x$ — Independent of any previous data
2. Learner pulls arm $a_t = \pi_t(x_t) \in \{1, \ldots, K\}$ — π_t policy learned from all data seen so far
3. Learner observes reward $r_t \sim \nu^{(a_t)}(x_t)$ from arm a_t in context x_t

Note that if the context distribution ν_x always returns the same value (e.g., 0), then the contextual bandit reduces to the original multi-armed bandit

Recall: UCB for contextual bandits

UCB algorithm conceptually identical as long as $|\mathcal{X}|$ finite:

$$\pi_t(x_t) = \arg\max_k \hat{\mu}_t^{(k)}(x_t) + \sqrt{\ln(2TK|\mathcal{X}|/\delta)/2N_t^{(k)}(x_t)}$$

- Added x_t argument to $\hat{\mu}_t^{(k)}$ and $N_t^{(k)}$ since we now keep track of the sample mean and number of arm pulls *separately* for each value of the context
- Added $|\mathcal{X}|$ inside the log because our union bound argument is now over all arm mean estimates $\hat{\mu}_t^{(k)}(x)$, of which there are $K|\mathcal{X}|$ instead of just K

But when $|\mathcal{X}|$ is really big (or even infinite), this will be really bad!

Solution: share information across contexts x_t, i.e., don't treat $\nu^{(k)}(x)$ and $\nu^{(k)}(x')$ as completely different distributions which have nothing to do with one another

Example: showing an ad on a NYT article on politics vs a NYT article on sports: Not *identical* readership, but still both on NYT, so probably still *similar* readership!

Recall: Modeling in contextual bandits

Need a model for $\mu^{(k)}(x)$, e.g., a linear model: $\mu^{(k)}(x) = \theta_k^\top x$

E.g., placing ads on NYT or WSJ (encoded as 0 or 1 in the first entry of x), for articles on politics or sports (encoded as 0 or 1 in the second entry of x) $\Rightarrow x \in \{0,1\}^2$

$|\mathcal{X}| = 4$ $\Rightarrow$ w/o linear model, need to learn 4 different $\mu^{(k)}(x)$ values for each arm k

With linear model there are just 2 parameters: the two entries of $\theta_k \in \mathbb{R}^2$

Lower dimension makes learning easier, but model could be wrong/biased

Today

- ✓ Feedback from last lecture
- ✓ Recap
- ✓ UCB-VI for linear MDPs
- ✓ Recall: Contextual Bandits
- LinUCB

Linear model fitting

Linear model for rewards: $\mu^{(k)}(x) = x^\top \theta^{(k)}$

How to estimate $\theta^{(k)}$? Linear regression

Least squares estimator: $\hat{\theta}_t^{(k)} = \arg\min_{\theta \in \mathbb{R}^d} \sum_{\tau=0}^{t-1} (r_\tau - x_\tau^\top \theta)^2 1_{\{a_\tau = k\}}$

Minimize squared error over time points when arm k selected

Claim: $\hat{\theta}_t^{(k)} = \left(\sum_{\tau=0}^{t-1} x_\tau x_\tau^\top 1_{\{a_\tau = k\}} \right)^{-1} \sum_{\tau=0}^{t-1} x_\tau r_\tau 1_{\{a_\tau = k\}}$

proof: $\nabla_\theta \left[\sum_{\tau=0}^{t-1} (r_\tau - x_\tau^\top \theta)^2 1_{\{a_\tau = k\}} \right] = 2 \sum_{\tau=0}^{t-1} x_\tau (r_\tau - x_\tau^\top \theta) 1_{\{a_\tau = k\}} = 0 \quad \Rightarrow \quad \sum_{\tau=0}^{t-1} x_\tau r_\tau 1_{\{a_\tau = k\}} = \theta \sum_{\tau=0}^{t-1} x_\tau x_\tau^\top 1_{\{a_\tau = k\}}$

Linear model fitting (cont'd)

Recall: $\hat{\theta}_t^{(k)} = \left(\sum_{\tau=0}^{t-1} x_\tau x_\tau^\top 1_{\{a_\tau = k\}} \right)^{-1} \sum_{\tau=0}^{t-1} x_\tau r_\tau 1_{\{a_\tau = k\}}$

Let $A_t^{(k)} = \sum_{\tau=0}^{t-1} x_\tau x_\tau^\top 1_{\{a_\tau = k\}}$ and $b_t^{(k)} = \sum_{\tau=0}^{t-1} x_\tau r_\tau 1_{\{a_\tau = k\}}$

Then $\hat{\theta}_t^{(k)} = \left(A_t^{(k)} \right)^{-1} b_t^{(k)}$

$A_t^{(k)}$ like empirical covariance matrix of the contexts when arm k was chosen

$b_t^{(k)}$ like empirical covariance between contexts and rewards when arm k was chosen

$A_t^{(k)}$ must be invertible, which basically requires $N_t^{(k)} \geq d$

Uncertainty quantification

For UCB, recall that we need confidence bounds on the expected reward of each arm (given context x_t)

Hoeffding was the main tool so far, but it used the fact that our estimate for the expected reward was a sample mean of the rewards we'd seen so far in the same setting (action, context)

With a model, we can use rewards we've seen in other settings → better estimation

But not using sample mean as estimator, so need something other than Hoeffding

Chebyshev's inequality: for a mean-zero random variable Y,

$$|Y| \leq \beta\sqrt{\mathbb{E}[Y^2]} \quad \text{with probability } \geq 1 - 1/\beta^2$$

Apply to $x_t^\top \hat{\theta}_t^{(k)} - x_t^\top \theta^{(k)}$

Uncertainty quantification (cont'd)

Want confidence bounds on our estimated mean rewards for each arm: $x_t^\top \hat{\theta}_t^{(k)}$

Strategy: apply Chebyshev's inequality to $x_t^\top \hat{\theta}_t^{(k)} - x_t^\top \theta^{(k)}$

Need: $\mathbb{E}[x_t^\top \hat{\theta}_t^{(k)} - x_t^\top \theta^{(k)}]$ (make sure it's zero) and $\mathbb{E}\left[(x_t^\top \hat{\theta}_t^{(k)} - x_t^\top \theta^{(k)})^2\right]$

Let $w_t = r_t - \mathbb{E}_{r \sim \nu^{(k)}(x_t)}[r] = r_t - x_t^\top \theta^{(k)}$, and we derive a useful expression for $\hat{\theta}_t^{(k)}$:

$$\hat{\theta}_t^{(k)} = (A_t^{(k)})^{-1} \sum_{\tau=0}^{t-1} x_\tau r_\tau 1_{\{a_\tau = k\}} = (A_t^{(k)})^{-1} \sum_{\tau=0}^{t-1} x_\tau (x_\tau^\top \theta^{(k)} + w_\tau) 1_{\{a_\tau = k\}}$$

$$= (A_t^{(k)})^{-1} \sum_{\tau=0}^{t-1} x_\tau x_\tau^\top 1_{\{a_\tau = k\}} \theta^{(k)} + (A_t^{(k)})^{-1} \sum_{\tau=0}^{t-1} x_\tau w_\tau 1_{\{a_\tau = k\}} = \theta^{(k)} + (A_t^{(k)})^{-1} \sum_{\tau=0}^{t-1} x_\tau 1_{\{a_\tau = k\}} w_\tau$$

Uncertainty quantification (cont'd)

Recall: $\hat{\theta}_t^{(k)} = \theta^{(k)} + (A_t^{(k)})^{-1} \sum_{\tau=0}^{t-1} x_\tau 1_{\{a_\tau = k\}} w_\tau$

Assume for simplicity that we are doing pure exploration, so the actions at each time step are totally independent of everything else.

$$\mathbb{E}_{w_\tau}[x_t^\top \hat{\theta}_t^{(k)} - x_t^\top \theta^{(k)}] = \mathbb{E}_{w_\tau}[x_t^\top (A_t^{(k)})^{-1} \sum_{\tau=0}^{t-1} x_\tau 1_{\{a_\tau=k\}} w_\tau] = x_t^\top (A_t^{(k)})^{-1} \sum_{\tau=0}^{t-1} x_\tau 1_{\{a_\tau=k\}} \mathbb{E}_{w_\tau}[w_\tau] = 0$$

$$\mathbb{E}_{w_\tau}[(x_t^\top \hat{\theta}_t^{(k)} - x_t^\top \theta^{(k)})^2] = \mathbb{E}_{w_\tau}\left[\left(x_t^\top (A_t^{(k)})^{-1} \sum_{\tau=0}^{t-1} x_\tau 1_{\{a_\tau=k\}} w_\tau\right)^2\right]$$

$$= x_t^\top (A_t^{(k)})^{-1} \sum_{\tau=0}^{t-1} \sum_{\tau'=0}^{t-1} x_\tau x_{\tau'}^\top 1_{\{a_\tau=k\}} 1_{\{a_{\tau'}=k\}} \mathbb{E}_{w_\tau}\left[w_\tau w_{\tau'}\right] (A_t^{(k)})^{-1} x_t$$

$$= x_t^\top (A_t^{(k)})^{-1} \sum_{\tau=0}^{t-1} x_\tau x_\tau^\top 1_{\{a_\tau=k\}} \mathbb{E}_{w_\tau}[w_\tau^2] (A_t^{(k)})^{-1} x_t \leq x_t^\top (A_t^{(k)})^{-1} A_t^{(k)} (A_t^{(k)})^{-1} x_t = x_t^\top (A_t^{(k)})^{-1} x_t$$

Chebyshev confidence bounds + intuition

Chebyshev: $x_t^\top \theta^{(k)} \leq x_t^\top \hat{\theta}_t^{(k)} + \beta \sqrt{x_t^\top (A_t^{(k)})^{-1} x_t}$ with probability $\geq 1 - 1/\beta^2$

$$A_t^{(k)} = \sum_{\tau=0}^{t-1} x_\tau x_\tau^\top 1_{\{a_\tau = k\}}$$

Intuition:

UCB term 1: $x_t^\top \hat{\theta}^{(k)}$ large when context and coefficient estimate aligned

UCB term 2: $x_t^\top (A_t^{(k)})^{-1} x_t = \frac{1}{N_t^{(k)}} x_t^\top (\Sigma_t^{(k)})^{-1} x_t$, where

$\Sigma_t^{(k)} = \frac{1}{N_t^{(k)}} A_t^{(k)} = \frac{1}{N_t^{(k)}} \sum_{\tau=0}^{t-1} x_\tau x_\tau^\top 1_{\{a_\tau = k\}}$ is the empirical covariance matrix of contexts when arm k chosen

Large when $N_t^{(k)}$ small or x_t not aligned with historical data

Some issues

Issue 1: All this assumed pure exploration!

Recall from HW 1 that we don't even expect unbiasedness for our arm mean estimates in the simple bandit case, due to adaptivity

So actually, the bounds we got don't really apply…

Issue 2: $A_t^{(k)}$ has to be invertible

Before the dth time that arm k gets pulled, $\hat{\theta}_t^{(k)}$ undefined

Solution (to both issues): regularize

Replace $A_t^{(k)} \leftarrow A_t^{(k)} + \lambda I$ for some $\lambda > 0$

Makes $A_t^{(k)}$ invertible always, and it turns out a bound just like Chebyshev's applies (with more details and a much more complicated proof, which we won't get into)

LinUCB algorithm

For $t = 0 \to T - 1$

Regularization makes $A_t^{(k)}$ invertible

1. $\forall\, k$, define $A_t^{(k)} = \sum_{\tau=0}^{t-1} x_\tau x_\tau^\top 1_{\{a_\tau = k\}} + \lambda I$ and $\hat{\theta}_t^{(k)} = (A_t^{(k)})^{-1} \sum_{\tau=0}^{t-1} x_\tau r_\tau 1_{\{a_\tau = k\}}$
2. Observe context x_t and choose $a_t = \arg\max_k \left\{ x_t^\top \hat{\theta}_t^{(k)} + c_t \sqrt{x_t^\top (A_t^{(k)})^{-1} x_t} \right\}$
3. Observe reward $r_t \sim \nu^{(a_t)}(x_t)$

c_t similar to log term in (non-lin)UCB, in that it depends logarithmically on

i. $1/\delta$ (δ is probability you want the bound to hold with)

ii. t and d implicitly via $\det(A_t^{(k)})$

Can prove $\tilde{O}(\sqrt{T})$ regret bound

Extensions

1. Can always replace contexts x_t with any fixed (vector-valued) function $\phi(x_t)$

   E.g., if believe rewards quadratic in scalar x_t, could make $\phi(x_t) = (x_t, x_t^2)$

2. Instead of fitting different $\theta^{(k)}$ for each arm, we could assume the mean reward is linear in some function of both the context and the action, i.e.,

$$\mathbb{E}_{r \sim \nu^{a_t}(x_t)}[r] = \phi(x_t, a_t)^\top \theta$$

   This is what we did in the linear MDP model! Helpful especially when K is large, since in that case there would be a lot of $\theta^{(k)}$ to fit

Both cases allow a version of linUCB by extension of the same ideas: fit coefficients via least squares and use Chebyshev-like uncertainty quantification to get UCB

More detail on the combined linear model

For $t = 0 \to T-1$

1. $\forall k$, define $A_t = \sum_{\tau=0}^{t-1} \phi(x_\tau, a_\tau)\phi(x_\tau, a_\tau)^\top + \lambda I$ and $\hat{\theta}_t = A_t^{-1} \sum_{\tau=0}^{t-1} \phi(x_\tau, a_\tau) r_\tau$
2. Observe x_t & choose $a_t = \arg\max_k \left\{ \phi(x_t, k)^\top \hat{\theta}_t + c_t \sqrt{\phi(x_t, k)^\top A_t^{-1} \phi(x_t, k)} \right\}$
3. Observe reward $r_t \sim \nu^{(a_t)}(x_t)$

Comments:

i. There is only one A_t and $\hat{\theta}_t$ (not one per arm), so more info shared across k

ii. Good for large K, but step 2's argmax may be hard

Continuous bandit action spaces

In bandits / contextual bandits, we have always treated the action space as discrete

This is because we to some extent treated each arm separately, necessitating trying each arm at least a fixed number of times before real learning could begin

But now with the new combined formulation, there is sufficient sharing across actions that we can learn $\hat{\theta}_t$ and its UCB *without* sampling all arms

This means that in principle, we can now consider continuous action spaces!

This is the power of having a strong model for $\mathbb{E}_{r \sim \nu^{(a_t)}(x_t)}[r]$, and a neural network would serve a similar purpose in place of the combined linear model (UQ less clear)

But in principle, there is no "free lunch", i.e., the hardness of the problem now transfers over to choosing a good model (a bad model will lead to bad performance)

Today

- ✓ Feedback from last lecture
- ✓ Recap
- ✓ UCB-VI for linear MDPs
- ✓ Recall: Contextual Bandits
- ✓ LinUCB

Summary:

- Modeling in MDPs and bandits with large state/action spaces is critical
- When model is linear (in feature space), can still rigorously quantify uncertainty and balance exploration/exploitation

Attendance:
bit.ly/3RcTC9T

Feedback:
bit.ly/3RHtlxy