

CS292F StatRL Lecture 8

Exploration in Bandits

Instructor: Yu-Xiang Wang

Spring 2021

UC Santa Barbara

Recap: Lecture 7

- Multi-Armed Bandits
 - Problem setup
 - Regret definition
- No regret learning algorithms
 - Exploration first
 - Eps-greedy
 - Upper Confidence Bound

Recap: Regret bounds

- Exploration first / eps-greedy

$$\tilde{O}(T^{2/3} k^{1/3})$$

- UCB

$$\tilde{O}(\sqrt{Tk})$$

$$\tilde{O}\left(\sum_{a \neq a^*} \frac{1}{\Delta_a}\right)$$

Recap: UCB algorithm for MAB

$$UCB^t(\mathbf{a}) = \tilde{\mu}^t(\mathbf{a}) + \sqrt{\frac{\log(2kT/\delta)}{2n^t(\mathbf{a})}}$$
$$LCB^t(\mathbf{a}) = \tilde{\mu}^t(\mathbf{a}) - \sqrt{\frac{\log(2kT/\delta)}{2n^t(\mathbf{a})}}$$

Actual reward means

By Hoeffding and union bound, with probability $\geq 1 - \delta$, it holds $\forall a \in [k], t \in [T]$:

$$\mu(\mathbf{a}) \in [LCB^t(\mathbf{a}), UCB^t(\mathbf{a})]$$

Claim : In the event that all confidence intervals hold, the regret is at most $\sum_t (UCB^t(a^t) - LCB^t(a^t)) + \delta \cdot T$

Proof: $Reg^t = \mu(a^*) - \mu(a^t)$

$$\leq UCB^t(a^*) - LCB^t(a^t)$$
$$\leq UCB^t(a^t) - LCB^t(a^t)$$

Recap: UCB algorithm for MAB

Gap-independent analysis:

$$\begin{aligned} \sum_{t=0}^{T-1} \mu^* - \mu_{I_t} &\leq 4\sqrt{\ln(TK/\delta)} \sum_{t=0}^{T-1} \sqrt{\frac{1}{N^t(I_t)}} \\ &= 4\sqrt{\ln(TK/\delta)} \sum_a \sum_{i=1}^{N^T(a)} \frac{1}{\sqrt{i}} \leq 8\sqrt{\ln(TK/\delta)} \sum_a \sqrt{N^T(a)} \leq 8\sqrt{\ln(TK/\delta)} \sqrt{K \sum_a N^T(a)} \\ &\leq 8\sqrt{\ln(TK/\delta)} \sqrt{KT}. \end{aligned}$$

Gap-dependent analysis:

$$N^T(a) \leq \frac{4 \ln(TK/\delta)}{\Delta_a^2}.$$

Recap: Linear bandits: problem setup

- Action space is a compact set

$$\text{decision } x_t \in D \subset \mathbb{R}^d.$$

- Reward is linear + i.i.d. noise.

$$\mathbb{E}[r_t | x_t = x] = \mu^* \cdot x \in [-1, 1], \quad \eta_t = r_t - \mu^* \cdot x_t$$

- Agent chooses a sequence of actions

$$x_1, \dots, x_T$$

- The regret is defined similarly

$$\text{Reg}_T = T \cdot \langle \mu^*, x^* \rangle - \sum_{t=1}^T \langle \mu^*, x_t \rangle$$

$$x^* \in \operatorname{argmax}_{x \in D} \mu^* \cdot x$$

Recap: The LinUCB algorithm: Optimism in the Face of Uncertainty.

- Consider the ridge regression at each time t .
- Construct high probability confidence set of the parameter vector
- Choose actions that maximize the UCB.

This lecture

- The regret analysis of LinUCB
- Exploration in Tabular RL
 - Problem setup

Regret bound of LinUCB

Sublinear regret: $R_T \leq O^*(d\sqrt{T})$

poly dependence on d , no dependence on the cardinality $|D|$.

Theorem 5.3 (AJKS)

Suppose: bounded noise $|\eta_t| \leq \sigma$, that $\|\mu^*\| \leq W$, and that $\|x\| \leq B$ for all $x \in D$. Set $\lambda = \sigma^2/W^2$ and

$$\beta_t := \sigma^2 \left(2 + 4d \log \left(1 + \frac{TB^2W^2}{d} \right) + 8 \log(4/\delta) \right).$$

With probability greater than $1 - \delta$, that for all $t \geq 0$,

$$R_T \leq c\sigma\sqrt{T} \left(d \log \left(1 + \frac{TB^2W^2}{d\sigma^2} \right) + \log(4/\delta) \right)$$

where c is an absolute constant.

(Dani, Hayes & Kakde, 2009)

(From this slide onwards mostly taken from Sham Kakade)

Two components of the regret analysis

- Uniform (**over all t**) confidence bound

Proposition 5.5 (AJKS)

(Confidence) Let $\delta > 0$. We have that

$$\Pr(\forall t, \mu^* \in \text{BALL}_t) \geq 1 - \delta.$$

- Sum of Squares Regret bound

Proposition 5.6 (AJKS)

(Sum of Squares Regret Bound) Define:

$$\text{regret}_t = \mu^* \cdot x^* - \mu^* \cdot x_t$$

Suppose $\|x\| \leq B$ for $x \in D$. Suppose β_t is increasing and larger than 1. Suppose $\mu^ \in \text{BALL}_t$ for all t , then*

$$\sum_{t=0}^{T-1} \text{regret}_t^2 \leq 4\beta_T d \log \left(1 + \frac{TB^2}{d\lambda} \right)$$

Proof of the main regret bound

- By Cauchy-Schwarz

$$\sum_{t=0}^{T-1} \text{regret}_t \leq \sqrt{T \sum_{t=0}^{T-1} \text{regret}_t^2} \leq \sqrt{4T\beta_T d \log \left(1 + \frac{TB^2}{d\lambda} \right)}.$$

Plan of the proof

1. First prove the Proposition that bounds the sum of square regret
 - By bounding instantaneous regret
 - And then bounding the sum of squares with “Information Gain”
2. Prove the uniform confidence bound
 - Basically show that the choice of β_t “works”.

“Width” of Confidence Ball

Lemma 5.7 (AJKS)

Let $x \in D$. If $\mu \in \text{BALL}_t$ and $x \in D$. Then

$$|(\mu - \hat{\mu}_t)^\top x| \leq \sqrt{\beta_t x^\top \Sigma_t^{-1} x}$$

Proof: By Cauchy-Schwarz, we have:

$$\begin{aligned} |(\mu - \hat{\mu}_t)^\top x| &= |(\mu - \hat{\mu}_t)^\top \Sigma_t^{1/2} \Sigma_t^{-1/2} x| = |(\Sigma_t^{1/2} (\mu - \hat{\mu}_t))^\top \Sigma_t^{-1/2} x| \\ &\leq \|\Sigma_t^{1/2} (\mu - \hat{\mu}_t)\| \|\Sigma_t^{-1/2} x\| = \|\Sigma_t^{1/2} (\mu - \hat{\mu}_t)\| \sqrt{x^\top \Sigma_t^{-1} x} \leq \sqrt{\beta_t x^\top \Sigma_t^{-1} x} \end{aligned}$$

where the last inequality holds since $\mu \in \text{BALL}_t$. ■

Instantaneous Regret is bounded by the width of the ellipsoid.

Define

$$w_t := \sqrt{x_t^\top \Sigma_t^{-1} x_t}$$

which is the “normalized width” at time t in the direction of our decision.

Lemma 5.8 (AJKS)

Fix $t \leq T$. If $\mu^* \in \text{BALL}_t$, then

$$\text{regret}_t \leq 2 \min(\sqrt{\beta_t} w_t, 1) \leq 2\sqrt{\beta_T} \min(w_t, 1)$$

Proof: Let $\tilde{\mu} \in \text{BALL}_t$ denote the vector which minimizes the dot product $\tilde{\mu}^\top x_t$. By choice of x_t , we have

$$\tilde{\mu}^\top x_t = \max_{\mu \in \text{BALL}_t} \max_{x \in D} \mu^\top x \geq (\mu^*)^\top x^*,$$

where the inequality used the hypothesis $\mu^* \in \text{BALL}_t$. Hence,

$$\begin{aligned} \text{regret}_t &= (\mu^*)^\top x^* - (\mu^*)^\top x_t \leq (\tilde{\mu} - \mu^*)^\top x_t \\ &= (\tilde{\mu} - \hat{\mu}_t)^\top x_t + (\hat{\mu}_t - \mu^*)^\top x_t \leq 2\sqrt{\beta_t} w_t \end{aligned}$$

“Geometric potential” argument: Converting summation to product

Lemma 5.9 (AJKS)

We have:

$$\det \Sigma_T = \det \Sigma_0 \prod_{t=0}^{T-1} (1 + w_t^2).$$

Proof: By the definition of Σ_{t+1} , we have

$$\begin{aligned} \det \Sigma_{t+1} &= \det(\Sigma_t + x_t x_t^\top) = \det(\Sigma_t^{1/2} (I + \Sigma_t^{-1/2} x_t x_t^\top \Sigma_t^{-1/2}) \Sigma_t^{1/2}) \\ &= \det(\Sigma_t) \det(I + \Sigma_t^{-1/2} x_t (\Sigma_t^{-1/2} x_t)^\top) = \det(\Sigma_t) \det(I + v_t v_t^\top), \end{aligned}$$

where $v_t := \Sigma_t^{-1/2} x_t$. Now observe that $v_t^\top v_t = w_t^2$ and ... ■

Taking logarithm (get information gain), then bounding it with data-independent terms.

Lemma

For any sequence x_0, \dots, x_{T-1} such that, for $t < T$, $\|x_t\|_2 \leq B$, we have:

$$\log \left(\det \Sigma_{T-1} / \det \Sigma_0 \right) = \log \det \left(I + \frac{1}{\lambda} \sum_{t=0}^{T-1} x_t x_t^\top \right) \leq d \log \left(1 + \frac{TB^2}{d\lambda} \right).$$

Proof: Denote the eigenvalues of $\sum_{t=0}^{T-1} x_t x_t^\top$ as $\sigma_1, \dots, \sigma_d$, and note:

$$\sum_{i=1}^d \sigma_i = \text{Trace} \left(\sum_{t=0}^{T-1} x_t x_t^\top \right) = \sum_{t=0}^{T-1} \|x_t\|^2 \leq TB^2.$$

Using the AM-GM inequality,

$$\begin{aligned} \log \det \left(I + \frac{1}{\lambda} \sum_{t=0}^{T-1} x_t x_t^\top \right) &= \log \left(\prod_{i=1}^d (1 + \sigma_i / \lambda) \right) \\ &= d \log \left(\prod_{i=1}^d (1 + \sigma_i / \lambda) \right)^{1/d} \leq d \log \left(\frac{1}{d} \sum_{i=1}^d (1 + \sigma_i / \lambda) \right) \leq d \log \left(1 + \frac{TB^2}{d\lambda} \right) \end{aligned}$$

Bounding the Sum of Square Instantaneous Regret

$$\sum_{t=0}^{T-1} \text{regret}_t^2 \leq \sum_{t=0}^{T-1} 4\beta_t \min(w_t^2, 1) \leq 4\beta_T \sum_{t=0}^{T-1} \min(w_t^2, 1)$$

Plan of the proof

1. First prove the Proposition that bounds the sum of square regret
 - By bounding instantaneous regret
 - And then bounding the sum of squares with “Information Gain”
2. Prove the uniform confidence bound
 - Basically show that the choice of β_t “works”.

We need to prove that the true parameter is in the version space w.h.p.

- Recall the version space is:

Proof: Since $r_\tau = \mathbf{x}_\tau \cdot \mu^* + \eta_\tau$, we have:

$$\begin{aligned}\hat{\mu}_t - \mu^* &= \Sigma_t^{-1} \sum_{\tau=0}^{t-1} r_\tau \mathbf{x}_\tau - \mu^* = \Sigma_t^{-1} \sum_{\tau=0}^{t-1} \mathbf{x}_\tau (\mathbf{x}_\tau \cdot \mu^* + \eta_\tau) - \mu^* \\ &= \Sigma_t^{-1} \left(\sum_{\tau=0}^{t-1} \mathbf{x}_\tau (\mathbf{x}_\tau)^\top \right) \mu^* - \mu^* + \Sigma_t^{-1} \sum_{\tau=0}^{t-1} \eta_\tau \mathbf{x}_\tau \\ &= \lambda \Sigma_t^{-1} \mu^* + \Sigma_t^{-1} \sum_{\tau=0}^{t-1} \eta_\tau \mathbf{x}_\tau\end{aligned}$$

By the triangle inequality,

$$\begin{aligned}\sqrt{(\hat{\mu}_t - \mu^*)^\top \Sigma_t (\hat{\mu}_t - \mu^*)} &\leq \left\| \lambda \Sigma_t^{-1/2} \mu^* \right\| + \left\| \Sigma_t^{-1/2} \sum_{\tau=0}^{t-1} \eta_\tau \mathbf{x}_\tau \right\| \\ &\leq \sqrt{\lambda} \|\mu^*\| + ??.\end{aligned}$$

How can we bound “??” To be continued...



Self-normalized Martingale concentration bound.

Lemma (Self-Normalized Bound for Vector-Valued Martingales)

(Abassi et. al '11) Suppose $\{\varepsilon_i\}_{i=1}^{\infty}$ are mean zero random variables (can be generalized to martingales), and ε_i is bounded by σ . Let $\{X_i\}_{i=1}^{\infty}$ be a stochastic process. Define $\Sigma_t = \Sigma_0 + \sum_{i=1}^t X_i X_i^{\top}$. With probability at least $1 - \delta$, we have for all $t \geq 1$:

$$\left\| \sum_{i=1}^t X_i \varepsilon_i \right\|_{\Sigma_t^{-1}}^2 \leq \sigma^2 \log \left(\frac{\det(\Sigma_t) \det(\Sigma_0)^{-1}}{\delta^2} \right).$$

Continue the proof by applying concentration, and the bound for information-gain

$$\begin{aligned}
 \sqrt{(\hat{\mu}_t - \mu^*)^\top \Sigma_t (\hat{\mu}_t - \mu^*)} &= \|(\Sigma_t)^{1/2} (\hat{\mu}_t - \mu^*)\| \\
 &\leq \left\| \lambda \Sigma_t^{-1/2} \mu^* \right\| + \left\| \Sigma_t^{-1/2} \sum_{\tau=0}^{t-1} \eta_\tau x_\tau \right\| \\
 &\leq \sqrt{\lambda} \|\mu^*\| + \sqrt{2\sigma^2 \log(\det(\Sigma_t) \det(\Sigma^0)^{-1} / \delta_t)}.
 \end{aligned}$$

$$\delta_t = (3/\pi^2)/t^2$$

$$1 - \Pr(\forall t, \mu^* \in \text{BALL}_t) = \Pr(\exists t, \mu^* \notin \text{BALL}_t) \leq \sum_{t=1}^{\infty} \Pr(\mu^* \notin \text{BALL}_t) < \sum_{t=1}^{\infty} (1/t^2)(3/\pi^2) = 1/2.$$

Final remarks on Linear Bandits

- The regret of LinUCB is optimal up to
- Strong assumption on realizability.
 - Agnostic linear bandits?
- Contextual version: a finite list of available actions are given at each t .

Exploration in Reinforcement Learning

- Why is it challenging?
 - The reward depends on both s , a
 - Unlike the generative model setting, we cannot just choose any s to explore.
 - The data needs to be actively collected
- We will study
 - Tabular MDP
 - Linear MDPs
 - Both in the finite horizon episodic setting.

Recap: Finite horizon MDPs

- Parameterization / Setup

$$M = (\mathcal{S}, \mathcal{A}, \{P\}_h, \{r\}_h, H, \mu)$$

- Additional notations
 - Q functions
 - V functions
 - Policies

Problem setup: online learning of Finite horizon MDPs

- Agent decides on a policy
- Collect a trajectory
- Agent updates the policy.
- Regret definition

Recap: The need for strategic exploration

UCB-VI: model-based learning by **optimistic** value Iterations

- Construct estimates of the transition kernels
- Design exploration bonuses
 - Idea: based on the uncertainty in the transition kernel estimates
- Update the policy by **optimistic** value iteration

What does value iteration does in finite horizon?

$$\widehat{V}_H^k(s) = 0, \forall s,$$

$$\widehat{Q}_h^k(s, a) = \min \left\{ r_h(s, a) + b_h^k(s, a) + \widehat{P}_h^k(\cdot | s, a) \cdot \widehat{V}_{h+1}^k, H \right\},$$

$$\widehat{V}_h^k(s) = \max_a \widehat{Q}_h^k(s, a), \pi_h^k(s) = \operatorname{argmax}_a \widehat{Q}_h^k(s, a), \forall h, s, a.$$

- Remark:
 - It converges in H steps
 - It produces a non-stationary policy indexed by h

How do we design exploration bonuses?

$$b_h^k(s, a) = H \sqrt{\frac{L}{N_h^k(s, a)}} \quad \text{where } L := \ln(SAHK/\delta)$$

- Intuitively, this encourages exploring new state-action pairs.
- Idea: propagate errors from the estimated transitions over to the rewards.

The regret of UCB-VI

- Theorem (AJKS Thm 6.1):

$$\text{Regret} := \mathbb{E} \left[\sum_{k=0}^{K-1} \left(V^* - V^{\pi^k} \right) \right] \leq 2H^2 S \sqrt{AK \cdot \ln(SAH^2K^2)} = \tilde{O} \left(H^2 S \sqrt{AK} \right)$$

- This is not optimal in H , S , but a simple analysis to start.
- We will talk about the proof next Monday.