CS292F StatRL Lecture7 Exploration in Bandits

Instructor: Yu-Xiang Wang

Spring 2021

UC Santa Barbara

Notes / reminders

- Project proposal due today
  - Please submit on Gradescope.
- Start HW1 quickly.
  - It will be more time-consuming than HW0.
  - It will help you with the rest of the class.
- HW2 is to be released this week (hopefully by tomorrow)

Recap: Lecture 6

- Policy gradient methods
  - Policy gradient theorem
  - Unbiased Monte Carlo estimate of the gradient (REINFORCE)
  - Bootstrapping in policy gradient estimates
  - Function approximation and Actor-Critic

- Bandits problem setup
  - Regret definition
  - The need for exploration

Recap: Multi-arm bandits: Problem setup

- No state. k-actions $$a \in \mathcal{A} = \{1, 2, ..., k\}$$
- You decide which arm to pull in every iteration

$$A_1, A_2, ..., A_T$$

- You collect a cumulative payoff of $$\sum_{t=1}^{T} R_t$$
- For MAB, the regret is defined as follow

$$T \max_{a \in [k]} \mathbb{E}[R_t|a] - \sum_{t=1}^{T} \mathbb{E}_{a \sim \pi}\left[\mathbb{E}[R_t|a]\right]$$

“No regret” means sublinear scaling in T. “Linear regret” is very bad.

- “No regret online learning”
- A regret (upper) bound needs to apply to all problem instances
- It suffices to identify one example to get a regret lower bound for a given algorithm.
  - E.g., “Greedy strategy” has linear regret in MAB.
- Minimax lower bounds are information-theoretical
  - They apply to all algorithms.

Recap: “Exploration first” strategy

- Let’s spend the first N step exploring.
  - Play each action for N / k times.

$$Q_t(a) = \frac{\sum_{i=1}^{t-1} R_i \cdot \mathbb{1}_{A_i=a}}{\sum_{i=1}^{t-1} \mathbb{1}_{A_i=a}}$$

- For t = N +1, N+2, …, T:

$$A_t \doteq \underset{a}{\arg\max}\, Q_t(a),$$

This lecture

- Regret analysis for multi-armed bandits
  - Exploration first
  - epsilon-greedy
  - Upper Confidence Bound algorithm (AJKS 5.1)

- Linear bandits. (AJKS 5.2 – 5.3)
  - LinUCB algorithm
  - Regret analysis

Recap: Concentration inequalities --- finite-sample bounds of LLN and CLT

- **Hoeffding's inequality**: Assume $X_1, \ldots, X_n$ are independent and their support bounded:

$$S_n = X_1 + \cdots + X_n$$

$$\mathrm{P}(S_n - \mathrm{E}[S_n] \geq t) \leq \exp\left(-\frac{2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right),$$

- Easy version, if $0<X_i<B$, with probability $1-\delta$:

$$\left|\bar{X} - \mathbb{E}[\bar{X}]\right| \leq \sqrt{\frac{B^2}{2n}\log(2/\delta)}$$

Regret analysis of Exploration First

Regret analysis of Exploration First

ε-Greedy strategy: one way to balance exploration and exploitation

- You choose with probability 1- ε

$$A_t \doteq \underset{a}{\arg\max}\, Q_t(a),$$

- With probability ε, choose an action uniformly at random!
  - Including the argmax.

- Carefully choose ε parameter.

A sketch of the analysis for ε-greedy

- In expectation, each arm is chosen for at least εt times.

- Condition on the number of times, apply Hoeffding's inequality / union bound for all t and a

- Regret bound is

$$\epsilon T \ + \sum_{t=1}^{T} C \sqrt{\frac{k}{\epsilon t}}$$

Optimism-in-the-face of uncertainty: Upper Confidence Bound algorithm

Martingale

- We say that a sequence of r.v. $X_1,\ldots,X_n,\ldots$ is a Martingale if for any n

$$\mathbf{E}(|X_n|) < \infty$$
$$\mathbf{E}(X_{n+1} \mid X_1, \ldots, X_n) = X_n.$$

- Example:
  - Random-walk: Total number of heads minus tails in n coin tosses

Azuma-Hoeffding's inequality

- **Azuma-Hoeffding's inequality**: Assume $X_1, \ldots, X_n$ are Martingale differences

$$S_n = X_1 + \ldots + X_n$$

$$\mathbb{P}\left[S_n \geq \epsilon\right] \leq e^{-\frac{2\epsilon^2}{\sum_{i=1}^{n}(b_i - a_i)^2}}$$

- Apply Azuma-Hoeffding's inequality to our problem

Regret analysis of UCB

Regret analysis of UCB

Summary of Exploration in Multi-Armed Bandits

- Explore-First

- eps-greedy

- UCB

Notes on MAB

- We considered "stochastic setting"
  - Adversarial setting ("a rigged casino")
  - Reward sequence is arbitrary / no expectation in the regret.

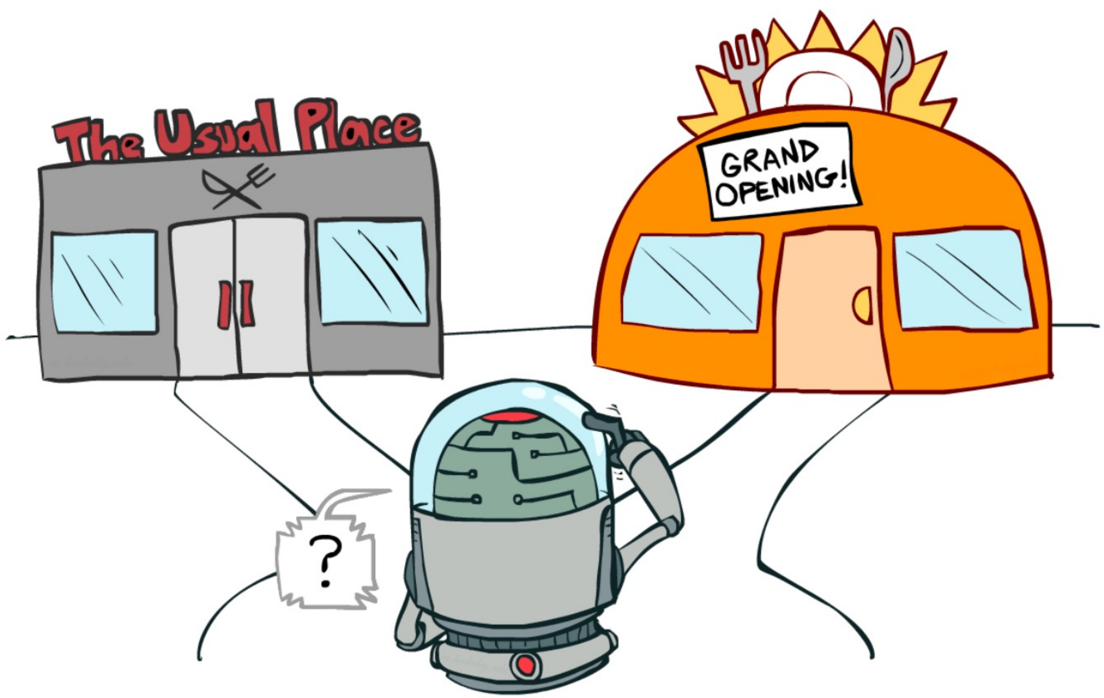
- Exponential weight algorithm for Explore-Exploit. (Exp3) achieves the same regret.
  - Read Auer et al. (2001) The Nonstochastic Multiarmed Bandit Problem

Linear bandits: MAB with an infinite number of actions

- Each action is determined by a "feature vector"

Features of action 1:
[Noodles, Tom Yum Soup, Poor service]

Features of action 2:
[Burger, Fries, Onion Ring, Fried Chicken]



Linear bandits: problem setup

- Action space is a compact set
- Reward is linear + noise.
- Agent chooses a sequence of actions
- The regret is defined similarly

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.

Regret bound of LinUCB

Sublinear regret: $R_T \leq O^{\star}(d\sqrt{T})$
poly dependence on d , no dependence on the cardinality $|D|$.

Theorem 5.3 (AJKS)

Suppose: bounded noise $|\eta_t| \leq \sigma$, that $\|\mu^{\star}\| \leq W$, 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

$$\Pr(\forall t,\ \mu^\star \in \textsc{Ball}_t) \geq 1 - \delta.$$

- Sum of Squares Regret bound

Proposition 5.6 (AJKS)

(Sum of Squares Regret Bound) Define:

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

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

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

Proof of the main regret bound

- By Cauchy-Schwarz

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

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

Let $x \in D$. If $\mu \in \textsc{Ball}_t$ and $x \in D$. Then

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

Proof: By Cauchy-Schwarz, we have:

$$|(\mu - \widehat{\mu}_t)^\top x| = |(\mu - \widehat{\mu}_t)^\top \Sigma_t^{1/2} \Sigma_t^{-1/2} x| = |(\Sigma_t^{1/2}(\mu - \widehat{\mu}_t))^\top \Sigma_t^{-1/2} x|$$
$$\leq \|\Sigma_t^{1/2}(\mu - \widehat{\mu}_t)\| \|\Sigma_t^{-1/2} x\| = \|\Sigma_t^{1/2}(\mu - \widehat{\mu}_t)\| \sqrt{x^\top \Sigma_t^{-1} x} \leq \sqrt{\beta_t x^\top \Sigma_t^{-1} x}$$

where the last inequality holds since $\mu \in \textsc{Ball}_t$. ■

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

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

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

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

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

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

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

"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 $x_0, \ldots x_{T-1}$ 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 $\sigma_1, \ldots \sigma_d$, and note:

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

Using the AM-GM inequality,

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

$$= d\log\left(\prod_{i=1}^{d}\left(1 + \sigma_i/\lambda\right)\right)^{1/d} \leq d\log\left(\frac{1}{d}\sum_{i=1}^{d}(1 + \sigma_i/\lambda)\right) \leq d\log\left(1 + \frac{TB^2}{d\lambda}\right)$$

Bounding the Sum of Square Instantaneous Regret

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

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 $r_\tau = x_\tau \cdot \mu^\star + \eta_\tau$, we have:

$$\begin{aligned}
\widehat{\mu}_t - \mu^\star &= \Sigma_t^{-1} \sum_{\tau=0}^{t-1} r_\tau x_\tau - \mu^\star = \Sigma_t^{-1} \sum_{\tau=0}^{t-1} x_\tau (x_\tau \cdot \mu^\star + \eta_\tau) - \mu^\star \\
&= \Sigma_t^{-1} \left(\sum_{\tau=0}^{t-1} x_\tau (x_\tau)^\top \right) \mu^\star - \mu^\star + \Sigma_t^{-1} \sum_{\tau=0}^{t-1} \eta_\tau x_\tau \\
&= \lambda \Sigma_t^{-1} \mu^\star + \Sigma_t^{-1} \sum_{\tau=0}^{t-1} \eta_\tau x_\tau
\end{aligned}$$

By the triangle inequality,

$$\begin{aligned}
\sqrt{(\widehat{\mu}_t - \mu^\star)^\top \Sigma_t (\widehat{\mu}_t - \mu^\star)} &\leq \left\| \lambda \Sigma_t^{-1/2} \mu^\star \right\| + \left\| \Sigma_t^{-1/2} \sum_{\tau=0}^{t-1} \eta_\tau x_\tau \right\| \\
&\leq \sqrt{\lambda} \|\mu^\star\| \qquad + \qquad ??.
\end{aligned}$$

How can we bound "??" To be continued... $\blacksquare$

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

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

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

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

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.