CS292F Statistical Foundation of Reinforcement Learning **Spring 2021**

Lecture 8: Linear Bandit (April 21)

Lecturer: Yu-Xiang Wang *Scribes: Xuandong Zhao*

Note: *LaTeX template courtesy of UC Berkeley EECS dept.*

8.1 Problem Setup

In linear bandit, we choose a decision x_t on each round, where the action space is a compact set: $x_t \in D \subset \mathbb{R}^d$. Then we obtain a reward $r_t \in [-1, 1]$. The reward is linear + i.i.d. noise, where $\mathbb{E}\left[r_t \mid x_t = x\right] = \mu^\star \cdot x \in [-1, 1]$ and noise sequence $\eta_t = r_t - \mu^\star \cdot x_t$ is i.i.d. noise.

If $x_0, \ldots x_T$ are our decisions, then our cumulative regret is

$$\mathrm{Reg}_T = T \cdot \langle \mu^\star, x^\star \rangle - \sum_{t=0}^{T} \langle \mu^\star, x_t \rangle$$

where $x^\star \in D$ is an optimal decision for $\mu^\star$, i.e.

$$x^\star \in \mathrm{argmax}_{x \in D}\, \mu^\star \cdot x$$

8.2 LinUCB Algorithm

Algorithm 1: Linear UCB

Input : λ, β_t

for t = *0, 1, 2, ...* **do**

Execute

$$x_t = \mathrm{argmax}_{x \in D} \max_{\mu \in \mathrm{BALL}_t} \langle x, \mu \rangle$$

and observe the reward r_t

Update BALL_{t+1}.

end

LinUCB is based on "optimism in the face of uncertainty," which is described in Algorithm 1. At episode t, we use all previous experience to define an uncertainty region (an ellipse) BALL_t. The center of this region, $\widehat{\mu}_t$, is the solution of the following ridge regression problem:

$$\widehat{\mu}_t = \arg\min_{\theta} \sum_{i=0}^{t-1} \left(x_i^\top \theta - r_i\right)^2 + \lambda \|\theta\|_2^2$$

If we consider the matrix form of x_t that $X_t = [x_0, x_1, \dots, x_{t-1}]^\top \in \mathbb{R}^{t\times d}$ and set $\mathbf{r}_t = [r_0, r_1, \dots, r_{t-1}]^\top \in \mathbb{R}^t$, the solution of the ridge regression is that:

$$\begin{aligned}\widehat{\mu}_t &= \arg\min_\theta \left\|X_t^\top \theta - \mathbf{r}_t\right\|_2^2 + \lambda\|\theta\|_2^2 \\ &= \left(X_t^\top X_t + \lambda I\right)^{-1} X_t^\top \mathbf{r}_t \\ &= \Sigma_t^{-1} \sum_{i=0}^{t-1} r_i x_i\end{aligned}$$

where λ is a parameter and where

$$\Sigma_t = \lambda I + \sum_{i=0}^{t-1} x_i x_i^\top, \text{ with } \Sigma_0 = \lambda I$$

The shape of the region BALL_t is defined through the feature covariance Σ_t. Precisely, the uncertainty region, or confidence ball, is defined as:

$$\mathrm{BALL}_t = \left\{\mu \middle| \left(\mu - \widehat{\mu}_t\right)^\top \Sigma_t \left(\mu - \widehat{\mu}_t\right) \le \beta_t\right\}$$

where β_t is a parameter of the algorithm.

8.3 Regret bound of LinUCB

Our main result here is that we have **sublinear regret**: $R_T \le O^\star(d\sqrt{T})$, **poly dependence** on d and **no dependence** on the cardinality $|D|$.

Theorem 8.1. *Suppose: bounded noise $|\eta_t| \le \sigma$, that $\|\mu^\star\| \le W$, and that $\|x\| \le B$ 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 $t \ge 0$,

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

where c *is an absolute constant.*

To proof the Theorem 8.1, we need two key components. The first is in showing that the confidence region is appropriate.

Proposition 8.2. *(Uniform confidence bound)*
Let $\delta > 0$. We have that

$$\Pr\left(\forall t, \mu^\star \in \mathrm{BALL}_t\right) \ge 1 - \delta.$$

The second main step in analyzing LinUCB is to show that as long as the aforementioned high-probability event holds, we have some control on the growth of the regret. Let us define the instantaneous regret as $\mathrm{regret}_t = \mu^\star \cdot x^\star - \mu^\star \cdot x_t$, the following bounds the sum of the squares of instantaneous regret.

Proposition 8.3. *(Sum of Squares Regret Bound)*
Define:

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

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

$$\sum_{t=0}^{T-1} \text{regret}_t^2 \le 8\beta_T d \log\left(1 + \frac{TB^2}{d\lambda}\right)$$

Using these two results we are able to prove our upper bound as follows:

Proof of Theorem 8.1. By Propositions 8.2 and 8.3 along with the Cauchy-Schwarz inequality, we have, with probability at least $1-\delta$,

$$R_T = \sum_{t=0}^{T-1} \text{regret}_t \le \sqrt{T \sum_{t=0}^{T-1} \text{regret}_t^2} \le \sqrt{8T\beta_T d \log\left(1 + \frac{TB^2}{d\lambda}\right)}.$$

The remainder of the proof follows from using our chosen value of $\beta_T = \sigma^2 \left(2 + 4d \log\left(1 + \frac{TB^2W^2}{d}\right) + 8\log(4/\delta)\right)$ and algebraic manipulations (that $2ab \le a^2 + b^2$). □

8.3.1 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".

Lemma 8.4. *("Width" of Confidence Ball)*
Let $x \in D$. *If* $\mu \in \text{BALL}_t$ *and* $x \in D$. *Then*

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

Proof. By Cauchy-Schwarz, we have:

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

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

Define

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

which is the "normalized width" at time t in the direction of the chosen decision. We now see that the width, $2\sqrt{\beta_t} w_t$, is an upper bound for the instantaneous regret.

Lemma 8.5. *(Instantaneous Regret is bounded by the width of the ellipsoid)*
Fix $t \le T$. If $\mu^\star \in \mathrm{BALL}_t$, then

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

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

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

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

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

where the last step follows from Lemma 8.4 since $\widetilde{\mu}$ and $\mu^\star$ are in BALL_t. Since $r_t \in [-1, 1]$, regret_t is always at most 2 and the first inequality follows. The final inequality is due to that β_t is increasing and larger than 1. □

The following two lemmas prove useful in showing that we can treat the log determinant as a potential function, where can bound the sum of widths independently of the choices made by the algorithm.

Lemma 8.6. *We have:*

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

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

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

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

$$\left(I + v_t v_t^\top\right) v_t = v_t + v_t \left(v_t^\top v_t\right) = \left(1 + w_t^2\right) v_t$$

Hence $\left(1 + w_t^2\right)$ is an eigenvalue of $I + v_t v_t^\top$. Since $v_t v_t^\top$ is a rank one matrix, all other eigenvalues of $I + v_t v_t^\top$ equal 1. Hence, $\det\left(I + v_t v_t^\top\right)$ is $\left(1 + w_t^2\right)$, implying $\det \Sigma_{t+1} = \left(1 + w_t^2\right) \det \Sigma_t$. The result follows by induction. □

Lemma 8.7. *("Potential Function" Bound)*
For any sequence $x_0, \dots x_{T-1}$ such that, for $t < T, \|x_t\|_2 \le B$, we have.

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

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

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

Using the AM-GM inequality,

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

which concludes the proof. □

Finally, we are ready to prove that if $\mu^\star$ always stays within the evolving confidence region, then our regret is under control.

Proof of Proposition 8.3. Assume that $\mu^\star \in \text{BALL}_t$ for all t. We have that:

$$\begin{aligned}\sum_{t=0}^{T-1} \text{regret}_t^2 &\le \sum_{t=0}^{T-1} 4\beta_t \min\left(w_t^2, 1\right) \le 4\beta_T \sum_{t=0}^{T-1} \min\left(w_t^2, 1\right) \\ &\le \max\{8, \frac{4}{\log 2}\}\beta_T \sum_{t=0}^{T-1} \log\left(1 + w_t^2\right) \le 8\beta_T \log\left(\det \Sigma_{T-1}/\det \Sigma_0\right) \\ &= 8\beta_T d\log\left(1 + \frac{TB^2}{d\lambda}\right)\end{aligned}$$

where the first inequality follow from Lemma 8.5, the second from that β_t is an increasing function of t; the third uses that for $0 \le y \le 1, y \ge \log(1+y) \ge \frac{y}{1+y} \ge \frac{y}{2}$,when $w_t^2 \le 1$, $w_t^2 \le 2\log(1+w_t^2)$, and when $w_t^2 > 1$, $4\beta_t = \frac{4}{\log 2}\beta_t \log 2 \le \frac{4}{\log 2}\beta_t \log(1+w_t^2)$; the final two inequalities follow by Lemmas 8.6 and 8.7. □

Then we can do confidence analysis to prove the uniform confidence bound:

Lemma 8.8. *(Self-Normalized Bound for Vector-Valued Martingales; [Abbasi-Yadkori et al., 2011]). Let $\{\varepsilon_i\}_{i=1}^{\infty}$ be a real-valued stochastic process with corresponding filtration $\{\mathcal{F}_i\}_{i=1}^{\infty}$ such that ε_i is $\mathcal{F}_i$ measurable, $\mathbb{E}\left[\varepsilon_i \mid \mathcal{F}_{i-1}\right] = 0$, and ε_i is conditionally σ -sub-Gaussian with $\sigma \in \mathbb{R}^+$. Let $\{X_i\}_{i=1}^{\infty}$ be a stochastic process with $X_i \in \mathcal{H}$ (some Hilbert space) and X_i being $\mathcal{F}_t$ measurable. Assume that a linear operator $\Sigma : \mathcal{H} \to \mathcal{H}$ is positive definite, i.e., $x^\top \Sigma x > 0$ for any $x \in \mathcal{H}$. For any t, define the linear operator $\Sigma_t = \Sigma_0 + \sum_{i=1}^{t} X_i X_i^\top$ (here $xx^\top$ denotes outer-product in $\mathcal{H}$). With probability at least $1-\delta$, we have for all $t \ge 1$:*

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

Proof of Proposition 8.2 . Since $r_\tau = x_\tau \cdot \mu^\star + \eta_\tau$, we have:

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

For any $0 < \delta_t < 1$, using triangle inequality and Lemma 8.8, it holds with probability at least $1 - \delta_t$,

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

where we have also used the triangle inequality and that $\left\|\Sigma_t^{-1}\right\| \le 1/\lambda$. We seek to lower bound $\Pr\left(\forall t, \mu^\star \in \mathrm{BALL}_t\right)$. Note that at $t = 0$, by our choice of λ, we have that BALL_0 contains $W^\star$, so $\Pr\left(\mu^\star \notin \mathrm{BALL}_0\right) = 0$. For $t \ge 1$, let us assign failure probability $\delta_t = \left(3/\pi^2\right)/t^2 \cdot 2\delta$ for the t -th event, which, using the above and union bound, gives us an upper bound on the sum failure probability as

$$1 - \Pr\left(\forall t, \mu^\star \in \mathrm{BALL}_t\right) = \Pr\left(\exists t, \mu^\star \notin \mathrm{BALL}_t\right) \le \sum_{t=1}^{\infty} \Pr\left(\mu^\star \notin \mathrm{BALL}_t\right) < \sum_{t=1}^{\infty} \left(1/t^2\right)\left(3/\pi^2\right) \cdot 2\delta = 1/2 \cdot 2\delta = \delta$$

This along with Lemma 8.7 completes the proof. □

8.4 Remarks

- The regret of LinUCB is optimal up to $\tilde{O}(d\sqrt{T})$
- The analysis of LinUCB is based on strong assumption on realizability.
- For agnostic linear bandits, EXP4 [Auer et al., 2002] can achieve the regret of $O(d\sqrt{T})$, and works in the adversarial settings, but is computationally inefficient.
- In contextual version with a finite list of available actions are given at each t, assuming i.i.d. setting, the "Taming the Monster" algorithm [Agarwal et al., 2014] achieves a regret bound of $O(\sqrt{dkT})$ where k is the number of actions with an oracle-efficient algorithm.

References

[AJKS] Agarwal, Alekh, Nan Jiang and S. Kakade. "Reinforcement Learning: Theory and Algorithms." (2019).

[Abbasi-Yadkori et al., 2011] Yasin Abbasi-Yadkori, David Pal, and Csaba Szepesvari. Improved algorithms for linear stochastic bandits. In Advances in Neural Information Processing Systems, pages 2312–2320, 2011.

[Auer et al., 2002] Peter Auer, Nicolo Cesa-Bianchi, Yoav Freund, and Robert E. Schapire. The nonstochastic multiarmed bandit problem. SIAM Journal of Computing, 32(1):48–77, 2002.

[Agarwal et al., 2014] Alekh Agarwal, Daniel J. Hsu, Satyen Kale, John Langford, Lihong Li, and Robert E. Schapire. Taming the monster: A fast and simple algorithm for contextual bandits. In Proceedings of the 31th International Conference on Machine Learning, pages 1638–1646, 2014