CS861: Theoretical Foundations of Machine Learning

Chapter 4: Stochastic Bandits

Kirthevasan Kandasamy
UW-Madison

Outline

1. Introduction, K-armed bandits
2. Explore-then-commit for K-armed bandits
3. The upper confidence bound algorithm for K-armed bandits
4. Lower bounds for K-armed bandits
5. Linear bandits and martingale concentration

Ch 4.1: Stochastic bandits introduction

So far in class: *passive* learning, i.e a learner learns from already available data.

But often, a learner is working towards a goal, and needs to make decisions about which data to collect, to achieve a goal efficiently.

Sequential/adaptive decision-making:.

- There is a sequence of interactions between a learner and an environment.
- On round t, the learner chooses an action A_t from a a set of possible actions $\mathcal{A}$.
- The environment reveals an observation O_t, and the learner receives a reward $X_t = X_t(O_t, A_t)$.
- The learner wishes to maximize the sum of rewards over T rounds, $\sum_{t=1}^{T} X_t$.

Stochastic bandits

- Let $\mathcal{A}$ denote a set of actions.

- Let $\nu = \{\nu_a; a \in \mathcal{A}\}$, called a *bandit model*, denote a set of distributions indexed by actions in $\mathcal{A}$.

- Let $\mathcal{P}$ denote a family of bandit models.

- On round t, a learner chooses $A_t \in \mathcal{A}$ and observes a *reward* $X_t \sim \nu_{A_t}$.

- Let μ_a be the expected reward of action a, and $\mu^\star$ be the optimal action,

$$\mu_a = \mathbb{E}_{X \sim \nu_a}[X], \qquad \mu^\star = \operatorname*{argmax}_{a \in \mathcal{A}} \mu_a.$$

- **Verbiage:** choose action a = play arm a = pull arm a

Learner's policy

A learner is characterized by a *policy* $\pi = \{\pi_t\}_{t \in \mathbb{N}}$, where, π_t maps the history $\{(A_s, X_s)\}_{s=1}^{t-1}$ to an action A_t:

$$\pi_t : ([K] \times \mathbb{R})^{t-1} \to \mathcal{A}.$$

If π is a randomized policy, π_t maps to a distribution over $\mathcal{A}$, and then an action is sampled from this distribution.

$$\pi_t : ([K] \times \mathbb{R})^{t-1} \to \Delta(\mathcal{A}),$$
$$\Delta(\mathcal{A}) = \begin{cases} \{p \in [0,1]^{|\mathcal{A}|};\ p^\top \mathbf{1}_{|\mathcal{A}|} = 1\} & \text{if } \mathcal{A} \text{ is finite} \\ \{p : \mathcal{A} \to \mathbb{R}_+;\ \int_{\mathcal{A}} p = 1\} & \text{more generally} \end{cases}$$

Regret

We define the regret as,

$$R_T = R_T(\pi, \nu) = T\mu^\star - \mathbb{E}\left[\sum_{t=1}^{T} X_t\right]$$

Here, the expectation $\mathbb{E}$ is with respect to the distribution of actions and rewards $A_1, X_1, A_2, X_2, \ldots,$ induced by the interaction between the policy π and the bandit model ν.

We want $R_T \in o(T)$, i.e $\lim_{T\to\infty} \frac{R_T}{T} = 0$.

K-armed bandits

- Stochastic bandits when actions are a finite set, i.e $\mathcal{A} = [K]$.

- Will assume each ν_i is σ-sub-Gaussian, with σ known.

$$\mathcal{P} = \{\nu = \{\nu_a ; a \in [K]\}, \text{ where } \nu_a \text{ is } \sigma\text{-sub-Gaussian for all } a\}.$$

- Will assume, without loss of generality, that

$$1 \geq \mu_1 \geq \mu_2 \geq \cdots \geq \mu_K \geq 0.$$

(the learner does not know the ordering)

- Denote the *gap*, $\Delta_i = \mu_1 - \mu_i$.

Ch 4.2: The Explore-then-Commit (ETC) algorithm

- **Given:** T (time horizon), m (number of exploration rounds per arm). $(m < T/K)$
- Pull each arm m times in the first mK rounds.
- $A \leftarrow \operatorname{argmax}_{i \in [K]} \widehat{\mu}_i$, where

$$\widehat{\mu}_i = \frac{1}{m} \sum_{s=1}^{mK} \mathbb{1}(A_s = i) X_s.$$

- Pull arm A for the remaining $(T - mK)$ rounds.

ETC Theoretical properties

Let $\mathcal{P}$ be the class of σ-sub-Gaussian K-armed bandit models.

Upper bound: Then for all $\nu \in \mathcal{P}$, π_m^{ETC} satisfies, Recall, $\Delta_i = \mu_1 - \mu_i$.

$$R_T(\pi_m^{\mathrm{ETC}}, \nu) \leq m \sum_{i, \Delta_i > 0} \Delta_i \; + \; (T - mK) \sum_{i, \Delta_i > 0} \Delta_i \exp\left(\frac{-m\Delta_i^2}{4\sigma^2} \right).$$

If we choose $m = K^{-1/3} T^{1/3}$, then

$$\sup_{\nu \in \mathcal{P}} R_T\left(\pi^{\mathrm{ETC}}_{K^{-1/3} T^{1/3}}, \; \nu \right) \; \in \; \tilde{\mathcal{O}}\left(K^{1/3} T^{2/3} \right).$$

Lower bound. Cannot be improved (via a tighter analysis and/or better choice of m),

$$\inf_{m \in [T]} \sup_{\nu \in \mathcal{P}} R_T(\pi_m^{\mathrm{ETC}}, \nu) \in \Omega\left(K^{1/3} T^{2/3} \right).$$

Proof: In homework 3 or 4.

Ch 4.3: The Upper Confidence Bound (UCB) algorithm

Based on the "optimism under uncertainty principle":

Pretend that the environment is as nice as statistically possible, given the data, and then behave myopically.

UCB Algorithm: designing upper confidence bounds

We will construct the following upper confidence (UCB) bound for the mean μ_i of each arm $i \in [N]$ *after* t rounds,

$$N_{i,t} = \sum_{s=1}^{t} \mathbb{1}(A_s = i), \qquad \texttt{samples from arm } i \texttt{ so far.}$$

$$\widehat{\mu}_{i,t} = \frac{1}{N_{i,t}} \sum_{s=1}^{t} \mathbb{1}(A_s = i) X_s, \qquad \texttt{sample mean.}$$

$$e_{i,t} = \sigma \sqrt{\frac{2\log(1/\delta_t)}{N_{i,t}}}, \qquad \texttt{where } \delta_t = 1/(T^2 t).$$

In our proofs, we will show that $\widehat{\mu}_{i,t} + e_{i,t}$ is a UCB for μ_i, *i.e.*, $\mathbb{P}(\mu_i < \widehat{\mu}_{i,t} + e_{i,t}) \geq 1 - \delta_t$.

The UCB Algorithm

- **Given:** time horizon T
- for $t = 1, \dots, K$,
  - Pull arm t and observe $X_t \sim \nu_t$.
- for $t = K + 1, \dots, T$,
  - Pull arm $A_t = \mathrm{argmax}_{i \in [K]} \, \widehat{\mu}_{i,t-1} + e_{i,t-1}$.
  - Observe $X_t \sim \nu_{A_t}$.

Theorem: UCB

Let $\mathcal{P}$ denote the class of σ-sub-Gaussian bandit models and let $\nu \in \mathcal{P}$. Then, the following statements are true:

- Gap-dependent bound: Recall, $\mu_1 < \cdots < \mu_K$, $\Delta_i = \mu_1 - \mu_i$.

$$R_T(\pi^{\mathrm{UCB}}, \nu) \leq 3K + \sum_{i:\Delta_i > 0} \frac{24\sigma^2 \log(T)}{\Delta_i}$$

- Worst-case bound:

$$\sup_{\nu \in \mathcal{P}} R_T(\pi^{\mathrm{UCB}}, \nu) \leq 3K + \sigma\sqrt{96KT\log(T)}.$$

Recall, we want $R_T \in o(T)$. In the worst case, we are guaranteed $\tilde{\mathcal{O}}(\sqrt{KT})$.

But, if gaps are large, we have $R_T \in \mathcal{O}(\log(T))$, as it is easier to distinguish between the arms if gaps are large.

Proof of UCB Theorem: Regret decomposition

Lemma (Regret decomposition). For *any* policy π (not just UCB), we have

$$R_T(\pi, \nu) = \sum_{i;\Delta_i>0} \Delta_i \mathbb{E}[N_{i,T}]$$

where, $N_{i,T} = \sum_{t=1}^{T} \mathbb{1}(A_t = i)$.

Proof. Let us first write the regret as follows,

$$\begin{aligned} R_T &= \sum_{t=1}^{T} (\mu_1 - \mathbb{E}[X_t]) = \sum_{t=1}^{T} \left(\mu_1 - \mathbb{E}\left[\sum_{i=1}^{K} \mathbb{1}(A_t = i) X_t \right] \right) \\ &= \sum_{t=1}^{T} \sum_{i=1}^{K} \mathbb{E}\left[(\mu_1 - X_t) \mathbb{1}(A_t = i) \right] = \sum_{t=1}^{T} \sum_{i=1}^{K} \mathbb{E}\Big[\underbrace{\mathbb{E}\left[(\mu_1 - X_t) \mathbb{1}(A_t = i) \middle| A_t \right]}_{(*)} \Big] \end{aligned}$$

Proof of UCB Theorem: Regret decomposition (cont'd)

Let us simplify $(*)$ as follows:

$$\begin{aligned}
(*) &= \mathbb{1}(A_t = i)\mathbb{E}\left[(\mu_1 - X_t)|A_t\right] \\
&= \mathbb{1}(A_t = i)(\mu_1 - \mu_{A_t}) && \texttt{as } \mathbb{E}[X_t|A_t = a] = \mu_a \\
&= \mathbb{1}(A_t = i)(\mu_1 - \mu_i) && \texttt{as expression is not 0 only when } A_t = i \\
&= \mathbb{1}(A_t = i)\Delta_i.
\end{aligned}$$

Plugging this back into our expression for R_T, we get

$$\begin{aligned}
R_T &= \sum_{t=1}^{T}\sum_{i=1}^{K}\mathbb{E}\big[\mathbb{1}(A_t = i)\Delta_i\big] \\
&= \sum_{i=1}^{K}\Delta_i\mathbb{E}\left[\sum_{t=1}^{T}\mathbb{1}(A_t = i)\right] = \sum_{i=1}^{K}\Delta_i\mathbb{E}\left[N_{i,T}\right]
\end{aligned}$$

Proof of UCB Theorem: Good event

Recall, that the UCB is $\widehat{\mu}_{i,t} + e_{i,t}$, where

$$N_{i,t} = \sum_{s=1}^{t} \mathbb{1}(A_s = i), \qquad \widehat{\mu}_{i,t} = \frac{1}{N_{i,t}} \sum_{s=1}^{t} \mathbb{1}(A_s = i) X_s,$$

$$e_{i,t} = \sigma \sqrt{\frac{2 \log(1/\delta_t)}{N_{i,t}}}, \qquad \text{where } \delta_t = 1/(T^2 t).$$

Define the following good events, G_1 and G_i for all i such that $\Delta_i > 0$,

$$G_1 = \{\forall\, t \geq K,\ \mu_1 < \widehat{\mu}_{1,t} + e_{1,t}\}, \qquad G_i = \{\forall\, t \geq K,\ \mu_i > \widehat{\mu}_{i,t} - e_{i,t}\}.$$

To bound G_1^c, G_i^c, we will assume w.l.o.g that each arm i samples rewards $\{Y_{i,r}\}_{r \geq 1}$ and we observe these samples one by one as we pull each arm. Therefore, we can write,

$$\widehat{\mu}_{i,t} = \frac{1}{N_{i,t}} \sum_{s=1}^{t} \mathbb{1}(A_s = i) X_s = \frac{1}{N_{i,t}} \sum_{r=1}^{N_{i,t}} Y_{i,r}.$$

Proof of UCB Theorem: Good event (cont'd)

Claim. For all i, we have $\mathbb{P}(G_i^c) \leq \frac{1}{T}$.

Proof. We will show this for G_1. The proof for G_i, where $\Delta_i > 0$, follows similarly.

$$
\begin{aligned}
\mathbb{P}(G_1^c) &= \mathbb{P}\left(\exists\, t \geq K \text{ such that } \mu_1 > \widehat{\mu}_{1,t} + e_{1,t}\right) \\
&\leq \sum_{t \geq K} \mathbb{P}\left(\mu_1 > \widehat{\mu}_{1,t} + e_{1,t}\right) \qquad \texttt{union bound} \\
&= \sum_{t \geq K} \mathbb{P}\left(\mu_1 > \frac{1}{N_{1,t}} \sum_{r=1}^{N_{1,t}} Y_{1,r} + \sigma\sqrt{\frac{2\log(1/\delta_t)}{N_{1,t}}}\right) \\
&\leq \sum_{t \geq K} \underbrace{\mathbb{P}\left(\exists\, s \in [1, 2, \ldots, t-K+1],\ \text{s.t } \mu_1 > \frac{1}{s}\sum_{r=1}^{s} Y_{1,r} + \sigma\sqrt{\frac{2\log(1/\delta_t)}{s}}\right)}_{(*)}
\end{aligned}
$$

Proof of UCB Theorem: Good event (cont'd)

Next, we will bound $(*)$ using the union bound and sub-Gaussian concentration,

$$
\begin{aligned}
(*) &\leq \sum_{s=1}^{t-K+1} \mathbb{P}\left(\mu_1 > \frac{1}{s}\sum_{r=1}^{s} Y_{1,r} + \sigma\sqrt{\frac{2\log(1/\delta_t)}{s}}\right) && \texttt{union bound} \\
&\leq \sum_{s=1}^{t-K+1} \underbrace{\exp\left(\frac{-s}{2\sigma^2} \times \sigma^2 \frac{2\log(1/\delta_t)}{s}\right)}_{=\delta_t = \frac{1}{T^2 t}} && \texttt{Sub-Gaussian concentration} \\
&\leq \frac{1}{T^2}. && \texttt{As there are at most } t-K+1 \texttt{ terms in the summation.}
\end{aligned}
$$

Therefore, $\mathbb{P}(G_c^1) \leq \sum_{t \geq K} (*) \leq \frac{1}{T}$.

Proof of UCB Theorem: Gap-dependent bound

Recall, $\quad G_1 = \{\forall\, t \geq K,\ \mu_1 < \widehat{\mu}_{1,t} + e_{1,t}\}\,, \qquad G_i = \{\forall\, t \geq K,\ \mu_i > \widehat{\mu}_{i,t} - e_{i,t}\}\,.$

Next, we will show that, under $G_1 \cap G_i$, for any i such that $\Delta_i > 0$, we have

$$N_{i,T} \leq \frac{24\sigma^2 \log(T)}{\Delta_i^2} + 1.$$

Proof. If an arm is never pulled after round K, then the bound holds trivially. Suppose instead arm i was last pulled on round $t+1 > K$. Then,

$$\mu_i + 2e_{i,t} \underbrace{\geq}_{\text{under } G_i} \widehat{\mu}_{i,t} + e_{i,t} \underbrace{\geq}_{\text{UCB}} \widehat{\mu}_{1,t} + e_{1,t} \underbrace{\geq}_{\text{under } G_1} \mu_1$$

$$\implies \quad \mu_1 - \mu_i \leq 2e_{i,t} = 2\sigma\sqrt{\frac{2\log(T^2 t)}{N_{i,t}}} \leq 2\sigma\sqrt{\frac{2\log(T^3)}{N_{i,t}}}$$

$$\implies \quad N_{i,t} \leq \frac{8\sigma^2\log(T^3)}{\Delta_i^2} \implies \quad N_{i,T} = N_{i,t} + 1 \leq \frac{24\sigma^2\log(T)}{\Delta_i^2} + 1,$$

where the last step uses the fact that i was last pulled in rd $t+1$.

Proof of UCB Theorem: Gap-dependent bound (cont'd)

Let us now decompose $\mathbb{E}[N_{i,T}]$ as follows,

$$\mathbb{E}[N_{i,T}] = \mathbb{E}[N_{i,T} | G_1 \cap G_i] \, \mathbb{P}(G_1 \cap G_i) \; + \; \mathbb{E}[N_{i,T} | G_1^c \cup G_i^c] \, \mathbb{P}(G_1^c \cup G_i^c)$$

Noting that $\mathbb{P}(G_1^c \cup G_i^c) \leq \mathbb{P}(G_1^c) + \mathbb{P}(G_i^c) \leq \frac{2}{T}$, we have

$$\begin{aligned}
\mathbb{E}[N_{i,T}] &\leq \left(\frac{24\sigma^2 \log(T)}{\Delta_i^2} + 1 \right) \cdot 1 \; + \; T \cdot \frac{2}{T} \\
&\leq \frac{24\sigma^2 \log(T)}{\Delta_i^2} + 3
\end{aligned}$$

Therefore,

$$\begin{aligned}
R_T &= \sum_{i:\Delta_i > 0} \Delta_i \mathbb{E}[N_{i,T}] \qquad \texttt{regret decomposition.} \\
&\leq \sum_{i:\Delta_i > 0} \left(\Delta_i \cdot \frac{24\sigma^2 \log(T)}{\Delta_i^2} + 3\Delta_i \right) \leq 3K \; + \sum_{i:\Delta_i > 0} \frac{24\sigma^2 \log(T)}{\Delta_i}
\end{aligned}$$

Proof of UCB Theorem: Worst-case bound

Recall, $R_T = \sum_{i:\Delta_i > 0} \Delta_i \mathbb{E}[N_{i,T}]$ $\qquad \mathbb{E}[N_{i,T}] \leq \frac{24\sigma^2 \log(T)}{\Delta_i^2} + 3$

Letting Δ be a value to be specified shortly, let us write

$$R_T = \sum_{i:\Delta_i \leq \Delta} \Delta_i \mathbb{E}[N_{i,T}] + \sum_{i;\Delta_i > \Delta} \Delta_i \mathbb{E}[N_{i,T}]$$

$$\leq \Delta T + \sum_{i;\Delta_i > \Delta} \Delta_i \left(\frac{24\sigma^2 \log(T)}{\Delta_i^2} + 3 \right) \quad \texttt{as } \sum_i \mathbb{E}[N_{i,T}] \leq T\texttt{.}$$

$$\leq \Delta T + \frac{24\sigma^2 \log(T)}{\Delta} \cdot K + 3K \quad \texttt{at most } K \texttt{ terms in sum.}$$

Now, choosing $\Delta = \sigma\sqrt{\frac{24K\log(T)}{T}}$, we have

$$R_T \leq 2\sigma\sqrt{24KT\log(T)} + 3K.$$

Alternative proof of worst-case bound

We will now look at an alternative proof of the worst-case bound, as this technique will also apply beyond K-armed bandits. For this, let us write the regret as follows,

$$R_T = \mathbb{E}\left[\sum_{t=1}^{T} \mu_1 - X_t\right]$$
$$= \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{E}\left[\mu_1 - X_t \middle| A_t\right]\right]$$
$$= \mathbb{E}\left[\sum_{t=1}^{T} \mu_1 - \mu_{A_t}\right]$$

The quantity $\sum_{t=1}^{T} \mu_1 - \mu_{A_t}$ is referred to as the pseudo-regret.

Alternative proof of worst-case bound (cont'd)

Define the good event $G = \bigcap_{i=1}^{K} G_i$, where

$$G_1 = \left\{\forall\, t \geq K,\ \mu_1 < \widehat{\mu}_{1,t} + e_{1,t}\right\}, \qquad G_i = \left\{\forall\, t \geq K,\ \mu_i > \widehat{\mu}_{i,t} - e_{i,t}\right\}.$$

We already showed $\mathbb{P}(G_i^c) \leq \frac{1}{T}$ so $\mathbb{P}(G^c) \leq \frac{K}{T}$. Now write,

$$\begin{aligned} R_T &= \mathbb{E}\left[\sum_{t=1}^{T} \mu_1 - \mu_{A_t}\right] \\ &= \mathbb{E}\left[\sum_{t=1}^{T} \mu_1 - \mu_{A_t} \middle| G\right] \cdot \underbrace{\mathbb{P}(G)}_{\leq 1} + \underbrace{\mathbb{E}\left[\sum_{t=1}^{T} \mu_1 - \mu_{A_t} \middle| G^c\right]}_{\leq T} \cdot \underbrace{\mathbb{P}(G^c)}_{\leq \frac{K}{T}} \end{aligned}$$

We will now focus on bounding $\mu_1 - \mu_{A_t}$ under G.

Alternative proof of worst-case bound (cont'd)

First note that under G we have,

$$\mu_1 \underbrace{\leq}_{\text{under } G_1} \widehat{\mu}_{1,t-1} + e_{1,t-1} \underbrace{\leq}_{\text{UCB}} \widehat{\mu}_{A_t,t-1} + e_{A_t,t-1} \underbrace{\leq}_{\text{under } G_i} \mu_{A_t} + 2e_{A_t,t-1}$$

Therefore, we have

$$\mu_1 - \mu_{A_t} \leq 2e_{A_t,t-1} = 2\sigma\sqrt{\frac{2\log(T^2(t-1))}{N_{A_t,t-1}}}.$$

This gives us the following bound for $\sum_{t=1}^{T} \mu_1 - \mu_{A_t}$ under G,

$$\sum_{t=1}^{T}(\mu_1 - \mu_{A_t}) = \sum_{t=1}^{K}(\mu_1 - \mu_{A_t}) + \sum_{t=K+1}^{T}(\mu_1 - \mu_{A_t})$$

$$\leq K + \sum_{t=K+1}^{T} 2\sigma\sqrt{\frac{2\log(T^2(t-1))}{N_{A_t,t-1}}} \leq K + \sigma\sqrt{24\log(T)}\underbrace{\sum_{t=K+1}^{T}\frac{1}{\sqrt{N_{A_t,t-1}}}}_{(*)}.$$

Alternative proof of worst-case bound (cont'd)

Let us bound the term $(*)$ as follows,

$$\begin{aligned}
\sum_{t=K+1}^{T} \frac{1}{\sqrt{N_{A_t,t-1}}} &= \sum_{i=1}^{K} \sum_{s=1}^{N_{i,T-1}} \frac{1}{\sqrt{s}} \\
&\leq 2 \sum_{i=1}^{K} \sqrt{N_{i,T-1}} \qquad \texttt{as } \sum_{s=1}^{m} 1/\sqrt{s} \leq 2\sqrt{m}\texttt{, bounding by integral} \\
&= 2K \left(\frac{1}{K} \sum_{i=1}^{K} \sqrt{N_{i,T-1}} \right) \\
&\leq 2K \sqrt{\frac{1}{K} \sum_{i=1}^{K} N_{i,T-1}} \qquad \texttt{Jensen's inequality.} \\
&= 2\sqrt{K(T-1)} \leq 2\sqrt{KT}.
\end{aligned}$$

Alternative proof of worst-case bound (cont'd)

Therefore, under G, we have

$$\sum_{t=1}^{T} \mu_1 - \mu_{A_t} \leq K + \sigma\sqrt{24\log(T)} \underbrace{\sum_{t=K+1}^{T} \frac{1}{\sqrt{N_{A_t,t-1}}}}_{(*)} \leq K + \sigma\sqrt{96KT\log(T)}.$$

Therefore,

$$R_T = \mathbb{E}\left[\sum_{t=1}^{T} \mu_1 - \mu_{A_t} \middle| G\right] \cdot \underbrace{\mathbb{P}(G)}_{\leq 1} + \underbrace{\mathbb{E}\left[\sum_{t=1}^{T} \mu_1 - \mu_{A_t} \middle| G^c\right]}_{\leq T} \cdot \underbrace{\mathbb{P}(G^c)}_{\leq K/T}$$

$$\leq 2K + \sigma\sqrt{96KT\log(T)}$$

$\square$

Ch 4.4: Lower bounds for K-armed bandits

We wish to prove a minimax lower bound of the form,

$$\inf_{\pi} \sup_{\nu \in \mathcal{P}} R_T(\pi, \nu) \in \Omega(?).$$

Reduction to (binary) testing: We will do so by considering *two alternatives* and showing that no policy can simultaneously achieve small regret on both alternatives.

For this, recall the Bretagnolle-Huber inequality: for any event A,

$$P_0(A) + P_1(A^c) \geq \frac{1}{2} e^{-\mathrm{KL}(P_0, P_1)}.$$

Divergence decomposition lemma

The distributions are now over sequences of actions and rewards $A_1, X_1, A_2, X_2, \ldots$ induced by the interaction between a policy and bandit model. The following lemma will be useful in computing the KL divergence between these distributions.

Lemma. Let ν, ν' be two K-armed bandit models. For a fixed (possibly randomized) policy π, let P, P' denote the probability distribution over the sequence of actions and rewards $A_1, X_1, \ldots, A_T, X_T$ under bandit models ν, ν' respectively. Let $\mathbb{E}_\nu$ denote the expectation under bandit model ν. Then, for all $T \geq 1$,

$$\mathrm{KL}(P, P') = \sum_{i=1}^{K} \mathbb{E}_\nu[N_{i,T}]\mathrm{KL}(\nu_i, \nu_i')$$

where $N_{i,T} = \sum_{t=1}^{T} \mathbb{1}(A_t = i)$, and $\nu = \{\nu_i\}_{i\in[K]}$, $\nu' = \{\nu_i'\}_{i\in[K]}$.

Divergence decomposition lemma: intuition

Suppose you had a static (non-adaptive) policy π which pulled the arms $N_1, \ldots, N_K$ times. Then, for two bandit models ν, ν',

$$
\begin{aligned}
\mathrm{KL}(P, P') &= \mathbb{E}_\nu \left[\log \left(\frac{p(A_1, X_1, \ldots, A_T, X_T)}{p'(A_1, X_1, \ldots, A_T, X_T)} \right) \right] \\
&= \mathbb{E}_\nu \left[\log \left(\frac{p(\{\{Y_{i,r}\}_{r=1}^{N_i}\}_{i=1}^{K})}{p'(\{\{Y_{i,r}\}_{r=1}^{N_i}\}_{i=1}^{K})} \right) \right] \qquad \texttt{as only observations differ} \\
&= \mathbb{E}_\nu \left[\log \left(\frac{\prod_{i=1}^{K} \prod_{s=1}^{N_i} p_{\nu_i}(Y_{i,r})}{\prod_{i=1}^{K} \prod_{s=1}^{N_i} p_{\nu_i'}(Y_{i,r})} \right) \right] \\
&= \sum_{i=1}^{K} \mathrm{KL}(\nu_i^{N_i}, {\nu'}_i^{N_i}) = \sum_{i=1}^{K} N_i \mathrm{KL}(\nu_i, \nu_i'). \qquad \texttt{as obs are i.i.d}
\end{aligned}
$$

The divergence decomposition lemma states that a similar result holds in expectation for adaptive policies (even though observations are not i.i.d).

Divergence decomposition lemma: proof

For a policy π, let $\pi(\cdot|a_1, x_1, \ldots, a_{t-1}, x_{t-1})$ be probability distribution over the actions $[K]$ for round t given the history $a_1, x_1, \ldots, a_{t-1}, x_{t-1}$.

Consider any given sequence $a_1, x_1, \ldots, a_T, x_T$ of actions and rewards. Let p, p' denote the densities (*e.g.*, pdfs, pmfs) of P, P' and let $\widetilde{\nu}_i, \widetilde{\nu}'_i$ denote the densities of ν_i, ν'_i respectively. Then,

$$\begin{aligned} p(a_1, x_1, \ldots, a_T, x_T) &= \prod_{t=1}^{T} p(a_t, x_t | a_1, x_1, \ldots, a_{t-1}, x_{t-1}) \\ &= \prod_{t=1}^{T} \pi(a_t | a_1, x_1, \ldots, a_{t-1}, x_{t-1}) \widetilde{\nu}_{a_t}(x_t) \\ \text{Similarly, } p'(a_1, x_1, \ldots, a_T, x_T) &= \prod_{t=1}^{T} \pi(a_t | a_1, x_1, \ldots, a_{t-1}, x_{t-1}) \widetilde{\nu}'_{a_t}(x_t) \end{aligned}$$

Divergence decomposition lemma: proof (cont'd)

Therefore, for any given sequence $a_1, x_1, \dots, a_T, x_T$,

$$\log\left(\frac{p(a_1, x_1, \dots, a_T, x_T)}{p'(a_1, x_1, \dots, a_T, x_T)}\right) = \log\left(\frac{\widetilde{\nu}_{a_1}(x_1) \times \cdots \times \widetilde{\nu}_{a_T}(x_T)}{\widetilde{\nu}'_{a_1}(x_1) \times \cdots \times \widetilde{\nu}'_{a_T}(x_T)}\right) = \sum_{t=1}^{T} \log\left(\frac{\widetilde{\nu}_{a_t}(x_t)}{\widetilde{\nu}'_{a_t}(x_t)}\right).$$

Therefore,

$$\begin{aligned}
\mathrm{KL}(P, P') &= \mathbb{E}_\nu\left[\log\left(\frac{p(A_1, X_1, \dots, A_T, x_T)}{p'(A_1, x_1, \dots, A_T, x_T)}\right)\right] = \sum_{t=1}^{T} \mathbb{E}_\nu\left[\log\left(\frac{\widetilde{\nu}_{A_t}(X_t)}{\widetilde{\nu}'_{A_t}(X_t)}\right)\right] \\
&= \sum_{t=1}^{T} \mathbb{E}_\nu\left[\log\left(\frac{\widetilde{\nu}_{A_t}(X_t)}{\widetilde{\nu}'_{A_t}(X_t)}\right) \sum_{i=1}^{K} \mathbb{1}(A_t = i)\right] \\
&= \sum_{i=1}^{K} \sum_{t=1}^{T} \mathbb{E}_\nu\Bigg[\underbrace{\mathbb{E}_\nu\left[\log\left(\frac{\widetilde{\nu}_{A_t}(X_t)}{\widetilde{\nu}'_{A_t}(X_t)}\right) \mathbb{1}(A_t = i) \,\middle|\, A_t\right]}_{(*)}\Bigg]
\end{aligned}$$

Divergence decomposition lemma: proof (cont'd)

We now observe that $(*)$ can be expressed as follows,

$$(*) = \mathbb{E}_\nu \left[\log\left(\frac{\widetilde{\nu}_{A_t}(X_t)}{\widetilde{\nu}'_{A_t}(X_t)} \right) \mathbb{1}(A_t = i) \middle| A_t \right] = \mathbb{1}(A_t = i)\mathbb{E}_\nu \left[\log\left(\frac{\widetilde{\nu}_{A_t}(X_t)}{\widetilde{\nu}'_{A_t}(X_t)} \right) \middle| A_t \right]$$

$$= \mathbb{1}(A_t = i)\, \underbrace{\mathbb{E}_\nu \left[\log\left(\frac{\widetilde{\nu}_i(X_t)}{\widetilde{\nu}_i(X_t)} \right) \right]}_{=\mathrm{KL}(\nu_i, \nu'_i)} \qquad \texttt{as } = 0 \texttt{ when } A_t \neq i$$

Therefore, noting that there is nothing random in $\mathrm{KL}(\nu_i, \nu'_i)$,

$$\mathrm{KL}(P, P') = \sum_{i=1}^{K} \sum_{t=1}^{T} \mathbb{E}_\nu[\mathbb{1}(A_t = i)]\mathrm{KL}(\nu_i, \nu'_i) = \sum_{i=1}^{K} \mathrm{KL}(\nu_i, \nu'_i)\, \underbrace{\mathbb{E}_\nu \left[\sum_{t=1}^{T} \mathbb{1}(A_t = i) \right]}_{=\mathbb{E}[N_{i,T}]}.$$

□

Minimax lower bounds for K-armed bandits

Theorem. Let $\mathcal{P}$ be the class of σ sub-Gaussian K-armed bandit models, where $K \geq 2$. Then, for some universal constant C,

$$\inf_{\pi} \sup_{\nu \in \mathcal{P}} R_T(\pi, \nu) \geq C\sigma\sqrt{T(K-1)} \in \Omega(\sigma\sqrt{TK}).$$

Proof. Let π be given. Then, for any two bandit models ν, ν', we have

$$\sup_{\nu \in \mathcal{P}} R_T(\pi, \nu) \geq \max\left(R_T(\nu, \pi), R_T(\nu', \pi)\right) \geq \frac{1}{2}\left(R_T(\pi, \nu) + R_T(\pi, \nu')\right).$$

We will lower bound $(R_T(\pi, \nu) + R_T(\pi, \nu'))$ for any policy π.

Minimax lower bound proof: designing alternatives

We will choose ν, ν', *dependent on* π to lower bound $(R_T(\pi,\nu) + R_T(\pi,\nu'))$.

Let $\nu = \{\nu_i = \mathcal{N}(\mu_i, \sigma^2)\}_{i\in[K]}$, where $\mu_1 = \delta$, $\mu_j = 0$ for all $j \neq 1$.
Let $\mathbb{E}_\nu$ denote expectation w.r.t the sequence $A_1, X_1, \ldots, A_T, X_T$ due to π's interaction with ν.

Since $\sum_{i=1}^{K} \mathbb{E}_\nu[N_{i,T}] = T$, there exists some $j \in \{2, \ldots, K\}$ such that $\mathbb{E}_\nu[N_{j,T}] \leq \frac{T}{K-1}$.

Let $\nu' = \{\nu'_i = \mathcal{N}(\mu'_i, \sigma^2)\}_{i\in[K]}$ where

$$\mu'_i = \begin{cases} \mu_i & \text{if } i \neq j, \\ 2\delta & \text{if } i = j \end{cases}$$

Minimax lower bound proof (cont'd)

Recall, Bretagnolle-Huber inequality. for any event A, $P_0(A) + P_1(A^c) \geq \frac{1}{2} e^{-\mathrm{KL}(P_0, P_1)}$.

Let P, P' denote the probability distribution of the sequence $A_1, X_1, \ldots, A_T, X_T$ due to π's interaction with ν and ν' respectively. We have,

$$R_T(\pi, \nu) \geq P\left(N_{1,T} \leq \frac{T}{2}\right) \frac{T\delta}{2}, \qquad R_T(\pi, \nu') \geq P'\left(N_{1,T} > \frac{T}{2}\right) \frac{T\delta}{2}.$$

Therefore,

$$\begin{aligned} R_T(\pi, \nu) + R_T(\pi, \nu') &\geq \frac{T\delta}{2} \left(P\left(N_{1,T} \leq \frac{T}{2}\right) + P'\left(N_{1,T} > \frac{T}{2}\right)\right) \\ &\geq \frac{T\delta}{2} \cdot \frac{1}{2} \exp\left(-\mathrm{KL}(P, P')\right) \qquad \texttt{Bretagnolle-Huber ineq} \end{aligned}$$

Minimax lower bound proof (cont'd)

Recall, we have shown $\quad \sup_{\nu \in \mathcal{P}} R_T(\pi,\nu) \geq \frac{1}{2}\left(R_T(\pi,\nu) + R_T(\pi,\nu')\right) \geq \frac{T\delta}{8} \exp\left(-\mathrm{KL}(P,P')\right)$

$$\mathrm{KL}(P,P') = \sum_{i=1}^{K} \mathbb{E}_\nu[N_{i,T}]\mathrm{KL}(\nu_i,\nu_i') \qquad \text{Divergence decomposition}$$

In our alternatives, ν, ν' only differ in arm j, which, recall, was chosen so that $\mathbb{E}_\nu[N_{j,T}] \leq \frac{T}{K-1}$. Hence,

$$\mathrm{KL}(P,P') = \underbrace{\mathbb{E}_\nu[N_{j,T}]}_{\leq \frac{T}{K-1}} \underbrace{\mathrm{KL}(\nu_j,\nu_j')}_{=\frac{(2\delta)^2}{2\sigma^2}} \leq \frac{2\delta^2 T}{(K-1)\sigma^2}.$$

Therefore,

$$\begin{aligned} \sup_{\nu \in \mathcal{P}} R_T(\pi,\nu) &\geq \frac{T\delta}{8} \exp\left(\frac{-2\delta^2 T}{(K-1)\sigma^2}\right) \\ &\geq \sigma\sqrt{T(K-1)} \cdot \frac{1}{8} e^{-2} \qquad \texttt{Choosing } \delta = \sigma\sqrt{\frac{K-1}{T}} \end{aligned}$$

Gap-dependent lower bound

Theorem (Theorem 16.4 in LS) Let ν be a K-armed bandit model with Gaussian rewards of variance σ^2. Let $\mu = \mu(\nu) \in \mathbb{R}^K$ be the means of the arms in ν. Let

$$\mathcal{P} = \left\{\nu'; \mu_i(\nu') \in [\mu_i, \mu_i + 2\Delta_i],\ \nu'_i = \mathcal{N}(\mu_i(\nu'), 1)\right\}$$

Say that π is a policy such that $R_T(\pi, \nu') \leq cT^p$ for some $c > 0$ and $p \in (0, 1)$ for all $\nu' \in \mathcal{P}$. Then,

$$R_T(\pi, \nu) \geq \frac{1}{2} \sum_{i, \Delta_i > 0} \frac{(1-p)\log(T) + \log(\Delta_i/(8c))}{\Delta_i}$$

Ch 4.5: Stochastic linear bandits

In many practical applications, there could be a very large (potentially infinite) number of arms, but there is additional structure in the problem.

We will next look at a *linear bandit model*, where the arms are in a Euclidean space, and the expected reward is a linear function

Stochastic linear bandits:

- Action space, $\mathcal{A} \subset \mathbb{R}^d$.
- There exists some *unknown true parameter*, $\theta_\star \in \mathbb{R}^d$.
- When we pull action $A_t \in \mathcal{A}$ on round t, we observe

$$X_t = \theta_\star^\top A_t + \epsilon_t,$$

where $\mathbb{E}[\epsilon_t] = 0$ and ϵ_t is σ-sub-Gaussian.

Regret

We can define the regret as,

$$R_T = T \max_{a \in \mathcal{A}} \theta_\star^\top a \; - \; \mathbb{E}\left[\sum_{t=1}^{T} X_t\right]$$
$$= T\theta_\star^\top a_\star - \mathbb{E}\left[\sum_{t=1}^{T} \theta_\star^\top A_t\right]$$

Here, $\mathbb{E}[X_t] = \mathbb{E}[\mathbb{E}[X_t | A_t]] = \mathbb{E}[\theta_\star^\top A_t]$.

We will assume $a_\star \triangleq \mathrm{argmax}_{a \in \mathcal{A}} \theta_\star^\top a$ exists.

The LinUCB Algorithm

Similar idea as UCB for K-armed bandits:
Construct upper confidence bounds (UCB) on the expected reward $\theta_\star^\top a$ of arm a, and choose the arm which maximizes the UCB.

- **Given:** time horizon T
- for $t = 1, \ldots, T$,
  - Pull arm $A_t = \mathrm{argmax}_{a \in \mathcal{A}} \mathrm{UCB}_{t-1}(a)$
  - Observe $X_t = \theta_\star^\top A_t + \epsilon_t$.

Designing a UCB

Recall the regularized ordinary least squares estimator for linear regression,

$$\widehat{\theta}_t \triangleq \operatorname*{argmin}_{\theta \in \mathbb{R}^d} \left(\lambda \|\theta\|_2^2 + \sum_{s=1}^{t} (X_s - \theta^\top A_s)^2 \right)$$

You can verify (after some elementary calculus),

$$\widehat{\theta}_t = V_t^{-1} \sum_{s=1}^{t} A_s X_s, \quad \text{where, } V_t = \lambda I + \sum_{s=1}^{t} A_s A_s^\top .$$

We will use

$$\mathrm{UCB}_{t-1}(a) = \widehat{\theta}_{t-1}^\top a + \beta_{t-1} \|a\|_{V_{t-1}^{-1}} .$$

for an appropriate choice of β_t. Here, $\|x\|_Q = \sqrt{x^\top Q x}$ is the Q–norm.

Intuition. $\widehat{\theta}_{t-1}^\top a$ is an estimate for $\theta_\star^\top a$ and encourages exploitation.
$\|a\|_{V_{t-1}^{-1}}$ is large for under-explored a and encourages exploration.

LinUCB: Upper bound on the regret.

Theorem. We will assume $\|\theta_\star\| \leq B$ (with B known) and that the rewards are bounded $0 \leq \theta_\star^\top a \leq 1$ for all $a \in \mathcal{A}$. Let $L = \max_{a\in\mathcal{A}} \|a\|_2$ (with L known). Then, if we choose

$$\beta_t = \max\left(\frac{1}{2},\ \sqrt{\lambda}B \,+\, \sigma\sqrt{d\log(dT^2)(2+\log(1+tL^2))}\right)$$

the regret satisfies

$$R_T \leq\ 2\ +\ 2\sqrt{2}\beta_T\sqrt{dT\log\left(1+TL^2/(d\lambda)\right)}\ \in\ \tilde{\mathcal{O}}\left(d\sqrt{T}\right).$$

LinUCB: Proof outline.

Our strategy (similar to worst case proof for K-armed bandits) is as follows. First, consider the pseudo-regret,

$$R_T = T \max_{a \in \mathcal{A}} \theta_\star^\top a \; - \; \mathbb{E}\left[\sum_{t=1}^{T} X_t\right] = \mathbb{E}\Bigg[\underbrace{\sum_{t=1}^{T} (\theta_\star^\top a_\star - \theta_\star^\top A_t)}_{=\overline{R}_T \text{ (pseudo-regret)}}\Bigg]$$

Define a good event G, and bound the regret via,

$$R_T = \mathbb{E}[\overline{R}_T | G] \underbrace{\mathbb{P}(G)}_{\leq 1} + \underbrace{\mathbb{E}[\overline{R}_T | G^c]}_{\leq T} \underbrace{\mathbb{P}(G^c)}_{\leq \text{small}}.$$

To bound $\mathbb{E}[\overline{R}_T | G]$,
- Under the good event G: the confidence intervals trap the true means. - Bound the instantaneous pseudo-regret $\theta_\star^\top a_\star - \theta_\star^\top A_t$ under G.

LinUCB Proof: Good event

Let us consider the good event in which the estimate is close to the true expected rewards for all arms on all rounds, *i.e.*,

$$G = \left\{ \left| \theta_\star^\top a - \widehat{\theta}_t^\top a \right| \leq \beta_t \|a\|_{V_t^{-1}}, \quad \text{for all } a \in \mathcal{A},\ t \in [T] \right\}.$$

where, recall

$$V_t = \lambda I + \sum_{s=1}^{t} A_s A_s^\top,$$

$$\widehat{\theta}_t = \operatorname*{argmin}_{\theta \in \mathbb{R}^d} \left(\lambda \|\theta\|_2^2 + \sum_{s=1}^{t} (X_s - \theta^\top A_s)^2 \right) = V_t^{-1} \sum_{s=1}^{t} A_s X_s,$$

Claim 1: $\mathbb{P}(G^c) \leq \frac{1}{T}$.
We will prove this later.

LinUCB Proof: Bounding the instantaneous pseudo-regret under G

Recall, $\quad G = \left\{ \left| \theta_\star^\top a - \widehat{\theta}_t^\top a \right| \le \beta_t \|a\|_{V_t^{-1}}, \quad \text{for all } a \in \mathcal{A},\ t \in [T] \right\}.$

Claim 2: Under G, we have, for all $t \geq 2$,

$$\theta_\star^\top a_\star - \theta_\star^\top A_t \le 2\beta_t \|A_t\|_{V_{t-1}}.$$

Proof. We will upper bound the instantaneous pseudo-regret as follows,

$$\begin{aligned}
&\theta_\star^\top a_\star - \theta_\star^\top A_t \\
&\quad \le \left(\widehat{\theta}_{t-1}^\top a_\star + \beta_{t-1} \|a_\star\|_{V_{t-1}^{-1}} \right) - \left(\widehat{\theta}_{t-1}^\top A_t - \beta_{t-1} \|A_t\|_{V_{t-1}^{-1}} \right) && \texttt{Good event} \\
&\quad \le \left(\widehat{\theta}_{t-1}^\top A_t + \beta_T \|A_t\|_{V_{t-1}^{-1}} \right) - \left(\widehat{\theta}_{t-1}^\top A_t - \beta_{t-1} \|A_t\|_{V_{t-1}^{-1}} \right) && A_t \texttt{ maximizes UCB} \\
&\quad \le 2\beta_{t-1} \|A_t\|_{V_{t-1}^{-1}}
\end{aligned}$$

LinUCB Proof: Bounding $\overline{R}_T$ under G

Recall: *(i)* $0 \leq \theta_\star^\top a \leq 1$ for all $a \in \mathcal{A}$, *(ii)* $\beta_t = \max(1/2, \dots)$

$$\overline{R}_T = \sum_{t=1}^{T} \left(\theta_\star^\top a_\star - \theta_\star^\top A_t\right) \leq 1 + \sum_{t=2}^{T} \left(\theta_\star^\top a_\star - \theta_\star^\top A_t\right)$$

$$\leq 1 + \sum_{t=2}^{T} \min\left(1,\ 2\beta_{t-1}\|A_t\|_{V_{t-1}^{-1}}\right) \quad \texttt{by } \mathit{(i)} \texttt{ and claim 2.}$$

$$\leq 1 + \sum_{t=2}^{T} 2\beta_{t-1} \min\left(1,\ \|A_t\|_{V_{t-1}^{-1}}\right) \quad \texttt{by } \mathit{(ii)}.$$

$$\leq 1 + 2\beta_T \sum_{t=1}^{T} \min\left(1,\ \|A_t\|_{V_{t-1}^{-1}}\right) \quad \texttt{as } \beta_t \texttt{ is increasing in } t.$$

$$\leq 1 + 2\beta_T \sqrt{T \sum_{t=1}^{T} \min\left(1,\ \|A_t\|^2_{V_{t-1}^{-1}}\right)} \quad \texttt{Cauchy-Schwarz ineq.}$$

LinUCB Proof: Bounding $\overline{R}_T$ under G (cont'd)

Claim 3: We will show,

$$\sum_{t=1}^{T} \min\left(1,\ \|A_t\|^2_{V_{t-1}^{-1}}\right) \leq 2d \log\left(1 + \frac{TL^2}{d\lambda}\right)$$

Therefore, under G,

$$\overline{R}_T \leq 1 + 2\beta_T \sqrt{dT} \sqrt{2 \log\left(1 + TL^2/(d\lambda)\right)}$$

Hence,

$$\begin{aligned} R_T &= \mathbb{E}[\overline{R}_T | G] \underbrace{\mathbb{P}(G)}_{\leq 1} + \underbrace{\mathbb{E}[\overline{R}_T | G^c]}_{\leq T} \underbrace{\mathbb{P}(G^c)}_{\leq 1/T}. \\ &\leq 2 + 2\beta_T \sqrt{dT} \sqrt{2 \log\left(1 + TL^2/(d\lambda)\right)} \end{aligned}$$

$\square$

LinUCB Proof: Proof of Claim 3

Claim 3: We will show,

$$\sum_{t=1}^{T} \min\left(1,\ \|A_t\|^2_{V_{t-1}^{-1}}\right) \leq 2d \log\left(1 + \frac{TL^2}{d\lambda}\right)$$

Proof. Recall that $V_t = \lambda I + \sum_{s=1}^{t} A_s A_s^\top$. Let us first consider,

$$\begin{aligned}
\det(V_t) &= \det\left(V_{t-1} + A_t A_t^\top\right) \\
&= \det\left(V_{t-1}^{1/2}\left(I + V_{t-1}^{-1/2} A_t A_t^\top V_{t-1}^{-1/2}\right) V_{t-1}^{1/2}\right) \\
&= \det(V_{t-1})\left(1 + (V_{t-1}^{-1/2} A_t)^\top (V_{t-1}^{-1/2} A_t)\right) \qquad \text{as } \det(I + uv^\top) = 1 + u^\top v\,. \\
&= \det(V_{t-1})(1 + \|A_t\|^2_{V_{t-1}^{-1}}).
\end{aligned}$$

LinUCB Proof: Proof of Claim 3 (cont'd)

$$V_t = \lambda I + \sum_{s=1}^{t} A_s A_s^\top, \qquad \det(V_t) = \det(V_{t-1})(1 + \|A_t\|_{V_{t-1}^{-1}}).$$

Therefore, we obtain

$$\det(V_T) = \underbrace{\det(\lambda I)}_{=\lambda^d} \prod_{t=1}^{T} (1 + \|A_t\|^2_{V_{t-1}^{-1}}).$$

Hence,

$$\log\left(\frac{\det(V_T)}{\lambda^d}\right) = \sum_{t=1}^{T} \log\left(1 + \|A_t\|^2_{V_{t-1}^{-1}}\right)$$

LinUCB Proof: Proof of Claim 3 (cont'd)

$$V_T = \lambda I + \sum_{s=1}^{T} A_s A_s^\top,$$

Now let us consider $\det(V_T)$,

$$\begin{aligned}
\det(V_T) = \prod_{i=1}^{d} \mathrm{eig}_i(V_T) &\le \left(\frac{1}{d}\sum_{i=1}^{d} \mathrm{eig}_i(V_T)\right)^d \qquad \texttt{AM-GM inequality} \\
&= \frac{1}{d^d}\mathrm{trace}(V_T)^d = \frac{1}{d^d}\left(\mathrm{trace}(\lambda I) + \mathrm{trace}\left(\textstyle\sum_{t=1}^{T} A_t A_t^\top\right)\right)^d \\
&= \frac{1}{d^d}\left(d\lambda + \sum_{t=1}^{T} \|A_t\|_2^2\right)^d \\
&\le \frac{1}{d^d}\left(d\lambda + TL^2\right)^d \qquad \text{as } \max_{a\in\mathcal{A}} \|a\|_2 \le L
\end{aligned}$$

LinUCB Proof: Proof of Claim 3 (cont'd)

$$\log\left(\frac{\det(V_T)}{\lambda^d}\right) = \sum_{t=1}^{T} \log\left(1 + \|A_t\|^2_{V_{t-1}}\right), \qquad \det(V_T) \leq \left(\frac{d\lambda + TL^2}{d}\right)^d$$

Putting it altogether,

$$\sum_{t=1}^{T} \min\left(1,\ \|A_t\|^2_{V_{t-1}^{-1}}\right) \leq 2\sum_{t=1}^{T} \log\left(1 + \min\left(1,\ \|A_t\|^2_{V_{t-1}^{-1}}\right)\right)$$

as $x \leq 2\log(1+x)$ for $x \in [0, 2\log(2)]$

$$\leq 2\sum_{t=1}^{T} \log\left(1 + \|A_t\|^2_{V_{t-1}^{-1}}\right)$$

$$\leq 2\log\left(\frac{1}{\lambda^d} \cdot \frac{(d\lambda + TL^2)^d}{d^d}\right) = 2d\log\left(1 + \frac{TL^2}{d\lambda}\right).$$

Proof of Claim 1: From prediction to estimation error

Claim 1. For G defined as follows, we have $\mathbb{P}(G^c) \leq \frac{1}{T}$.

$$G = \left\{ \left| \theta_\star^\top a - \widehat{\theta}_t^\top a \right| \leq \beta_t \|a\|_{V_t^{-1}}, \quad \text{for all } a \in \mathcal{A},\ t \in [T] \right\}$$

Proof. First write

$$\mathbb{P}(G^c) \leq \sum_{t=1}^{T} \mathbb{P}\left(\left| \theta_\star^\top a - \widehat{\theta}_t^\top a \right| > \beta_t \|a\|_{V_t^{-1}}, \quad \text{for some } a \in \mathcal{A} \right)$$

Now, Consider any $a \in \mathcal{A}$,

$$\begin{aligned}
\left| \theta_\star^\top a - \widehat{\theta}_t^\top a \right| &= \left| (\theta_\star - \widehat{\theta}_t)^\top a \right| = \left| ((\theta_\star - \widehat{\theta}_t) V_t^{1/2})^\top (V_t^{-1/2} a) \right| \\
&\leq \|(\theta_\star - \widehat{\theta}_t) V_t^{1/2}\|_2 \cdot \|V_t^{-1/2} a\|_2 \quad \texttt{Cauchy-Schwarz} \\
&= \|\theta_\star - \widehat{\theta}_t\|_{V_t} \cdot \|a\|_{V_t^{-1}} \quad \texttt{as } \|xQ^{1/2}\|_2^2 = x^\top Q x = \|x\|_Q^2 \texttt{ for symmetric } Q
\end{aligned}$$

Therefore, sufficient to show $\|\theta_\star - \widehat{\theta}_t\|_{V_t} \leq \beta_t$ with probability $\geq 1 - 1/T^2$.

Proof of Claim 1 (cont'd)

$$\widehat{\theta}_t = V_t^{-1} \sum_{s=1}^{t} A_s X_s, \quad \text{where,} \qquad V_t = \lambda I + \sum_{s=1}^{t} A_s A_s^\top, \quad X_s = \theta_\star^\top A_s + \epsilon_s.$$

Let $W_t = \sum_{s=1}^t A_s A_s^\top$ so that $V_t = \lambda I + W_t$.

Now write,

$$\widehat{\theta}_t = V_t^{-1} \sum_{s=1}^t A_s (A_s^\top \theta_\star + \epsilon_s) = V_t^{-1} W_t \theta_\star + V_t^{-1} \underbrace{\textstyle\sum_{s=1}^t A_s \epsilon_s}_{\triangleq \xi_t}.$$

Therefore, $\widehat{\theta}_t - \theta_\star = (V_t^{-1} W_t - I)\theta_\star + V_t^{-1}\xi_t$. Hence,

$$\begin{aligned} \|\widehat{\theta}_t - \theta_\star\|_{V_t} &\leq \|(V_t^{-1} W_t - I)\theta_\star\|_{V_t} + \|V_t^{-1}\xi_t\|_{V_t} \quad \texttt{triangle inequality} \\ &= \|(V_t^{-1} W_t - I)\theta_\star\|_{V_t} + \|\xi_t\|_{V_t^{-1}} \quad \texttt{as } \|V_t^{-1}\xi_t\|_{V_t}^2 = \xi_t^\top V_t^{-1} V_t V_t^{-1} \xi_t = \|\xi_t\|_{V_t^{-1}}^2. \end{aligned}$$

Proof of Claim 1 (cont'd)

$$\|\widehat{\theta}_t - \theta_\star\|_{V_t} \le \|(V_t^{-1}W_t - I)\theta_\star\|_{V_t} + \|\xi_t\|_{V_t^{-1}}, \qquad W_t = \sum_{s=1}^{t} A_s A_s^\top, \qquad V_t = \lambda I + W_t.$$

The following calculations show $\|(V_t^{-1}W_t - I)\theta_\star\|_{V_t} \le \sqrt{\lambda}B$, (try at home)

$$\|(V_t^{-1}W_t - I)\theta_\star\|_{V_t}^2 = \theta_\star^\top (V_t^{-1}W_t - I)V_t(V_t^{-1}W_t - I)\theta_\star$$
$$= \theta_\star^\top (V_t^{-1}W_t - I)\underbrace{(W_t - V_t)}_{=-\lambda I}\theta_\star = \lambda\theta_\star^\top \underbrace{(I - V_t^{-1}W_t)}_{\lesssim I}\theta_\star \le \lambda\theta_\star^\top\theta_\star \le \lambda B^2.$$

Recall, we need to show, that the following holds with probability $\ge 1 - 1/T^2$

$$\|(\theta_\star - \widehat{\theta}_t)\|_{V_t} \le \beta_t = \max\left(\frac{1}{2},\ \sqrt{\lambda}B + \sigma\sqrt{d\log(dT^2)(2 + \log(1 + tL^2))}\right)$$

Sufficient to show, the following holds with probability $\ge 1 - 1/T^2$

$$\|\xi_t\|_{V_t^{-1}} \le \sigma\sqrt{d}\sqrt{\log(dT^2)(2 + \log(1 + tL^2))}$$

Proof of Claim 1: Simplifying $\|\xi_t\|_{V_t^{-1}}$ further

Let us write,

$$\begin{aligned}
\|\xi_t\|^2_{V_t^{-1}} = \xi_t^\top V_t^{-1} \xi_t &= \xi_t^\top V_t^{-1/2} \cdot I \cdot V_t^{-1/2} \xi_t \\
&= \xi_t^\top V_t^{-1/2} \left(\sum_{i=1}^{d} e_i e_i^\top \right) V_t^{-1/2} \xi_t \\
&= \sum_{i=1}^{d} \xi_t^\top V_t^{-1/2} e_i e_i^\top V_t^{-1/2} \xi_t \\
&= \sum_{i=1}^{d} (\xi_t^\top V_t^{-1/2} e_i)^2 .
\end{aligned}$$

Proof of Claim 1 (cont'd)

We need to show, the following holds with probability $\geq 1 - 1/T^2$

$$\|\xi_t\|^2_{V_t^{-1}} \leq \sigma^2 d \underbrace{\log(dT^2)(2 + \log(1 + tL^2))}_{(*)}$$

By a union bound,

$$\mathbb{P}\left(\|\xi_t\|^2_{V_t^{-1}} > d\sigma^2(*)\right) = \mathbb{P}\left(\sum_{i=1}^{d}(\xi_t^\top V_t^{-1/2} e_i)^2 > d\sigma^2(*)\right)$$

$$\leq \sum_{i=1}^{d} \mathbb{P}\left((\xi_t^\top V_t^{-1/2} e_i)^2 > \sigma^2(*)\right) = \sum_{i=1}^{d} \mathbb{P}\left(\frac{|\xi_t^\top V_t^{-1/2} e_i|}{\sigma} > \sqrt{(*)}\right)$$

Therefore, sufficient to show, for $a \in \{V_t^{-1/2} e_1, \ldots, V_t^{-1/2} e_d\}$,

$$\mathbb{P}\left(\frac{|\xi_t^\top a|}{\sigma} > \sqrt{(*)}\right) \leq \frac{1}{T^2 d}.$$

Proof of Claim 1: Summary so far

Claim 1. For G defined as follows, we have $\mathbb{P}(G^c) \leq \frac{1}{T}$.

$$G = \left\{ \left| \theta_\star^\top a - \widehat{\theta}_t^\top a \right| \leq \beta_t \|a\|_{V_t^{-1}}, \quad \text{for all } a \in \mathcal{A},\ t \in [T] \right\}$$

1. (Prediction to estimation error) Sufficient to show, the following holds with probability $\geq 1 - 1/T^2$

$$\|(\theta_\star - \widehat{\theta}_t)\|_{V_t} \leq \beta_t = \max\left(\frac{1}{2},\ \sqrt{\lambda} B \ + \ \sigma \sqrt{d \log(dT^2)(2 + \log(1 + tL^2))} \right)$$

2. Sufficient to show, the following holds with probability $\geq 1 - 1/T^2$

$$\|\xi_t\|_{V_t^{-1}} \leq \sigma\sqrt{d}\sqrt{\log(dT^2)(2 + \log(1 + tL^2))}$$

3. Sufficient to show, for $a \in \{V_t^{-1/2} e_1, \ldots, V_t^{-1/2} e_d\}$,

$$\mathbb{P}\left(\frac{|\xi_t^\top a|}{\sigma} > \sqrt{\log(dT^2)(2 + \log(1 + tL^2))} \right) \leq \frac{1}{T^2 d}.$$

Martingale concentration

Recall that $\xi_t = \sum_{s=1}^t A_s \epsilon_s$. We will show, for any $a \in \{V_t^{-1/2} e_1, \ldots, V_t^{-1/2} e_d\}$,

$$\mathbb{P}\left(\frac{|\xi_t^\top a|}{\sigma} > \sqrt{\log(dT^2)(2 + \log(1 + tL^2))} \right) \leq \frac{1}{T^2 d}.$$

Key challenge: The actions and observations are not independent!
Otherwise, we can condition on $A_1, \ldots, A_t$, and use standard sub-Gaussian concentration.

We will use the fact that $\xi_t^\top a$ is a *martingale*.
We can use a variety of martingale concentration results to obtain the above result.

Definition (Martingale). An sequence of random variables $\{Z_t\}_{t\in\mathbb{N}}$ is a martingale w.r.t another sequence $\{Y_t\}_{t\in\mathbb{N}}$ if, $\mathbb{E}[Z_t | Y_{t-1}] = Z_{t-1}$ and $\mathbb{E}[|Z_t|] < \infty$ for all $t \in \mathbb{N}$.

We will present one proof, *without formally introducing martingales*. See LS Chapter 20, for an alternative technique.

Proof of Claim 1: Martingale concentration

We will show, for any $a \in \{V_t^{-1/2} e_1, \dots, V_t^{-1/2} e_d\}$,

$$\mathbb{P}\left(\frac{|\xi_t^\top a|}{\sigma} > \sqrt{\log(dT^2)(2 + \log(1 + tL^2))}\right) \le \frac{1}{T^2 d}.$$

Proof technique from Rusmevichientong and Tsitsiklis, 2008.

Lemma (Corollary 2.2 from de La Pena et al 2004). If A, B are random variables such that $\mathbb{E}\left[e^{\mu A - \frac{\mu^2 B^2}{2}}\right] \le 1$ for all $\mu \in \mathbb{R}$, then for all $\tau \ge \sqrt{2}$, and $y > 0$, we have

$$\mathbb{P}\left(|A| \ge \tau \sqrt{(B^2 + y)\left(1 + \frac{1}{2}\log\left(1 + B^2/y\right)\right)}\right) \le e^{-\tau^2/2}.$$

Cf. If B is a constant and not a R.V, then the condition says that A is B-sub-Gaussian, $\mathbb{E}[e^{\mu A}] \le e^{\frac{\mu^2 B^2}{2}}$. In which case, we know $\mathbb{P}(A > B\tau) \le e^{\frac{-\tau^2}{2}}$.
Note that $\sqrt{(B^2 + y)\left(1 + \log\left(1 + B^2/y\right)\right)} \asymp B$.

Proof of Claim 1: Martingale concentration (cont'd)

Lemma (Corollary 2.2 from de La Pena et al 2004). If A, B are random variables such that $\mathbb{E}\left[e^{\mu A - \frac{\mu^2 B^2}{2}}\right] \leq 1$ for all $\mu \in \mathbb{R}$, then for all $\tau \geq \sqrt{2}$, and $y > 0$, we have

$$\mathbb{P}\left(|A| \geq \tau\sqrt{(B^2+y)\left(1+(1/2)\log\left(1+B^2/y\right)\right)}\right) \leq e^{-\tau^2/2}.$$

We will apply the above result with, $A = \frac{a^\top \xi_t}{\sigma}$ and $B = \|a\|_{W_t}$, $y = \|a\|_2^2$, and $\tau = \sqrt{2\log(T^2 d)}$.

$$A = \frac{a^\top \xi_t}{\sigma} = \frac{1}{\sigma}\sum_{s=1}^{t} a^\top A_s \epsilon_s, \qquad B^2 = a^\top W_t a = \sum_{s=1}^{t} a^\top A_s A_s^\top a.$$

Let us first check that the condition $\mathbb{E}\left[\exp\left(\mu A - \frac{\mu^2 B^2}{2}\right)\right] \leq 1$ holds. Write,

$$\mu A - \frac{\mu^2}{2}B^2 = \sum_{s=1}^{t}\left(\underbrace{\frac{\mu}{\sigma}a^\top A_s \epsilon_s - \frac{\mu^2}{2}(a^\top A_s)^2}_{\triangleq Q_s}\right)$$

Proof of Claim 1: Martingale concentration (cont'd)

We need to show,

$$\mathbb{E}\left[\exp\left(\mu A - \frac{\mu^2}{2}B^2\right)\right] = \mathbb{E}\left[e^{\sum_{s=1}^{t} Q_s}\right] \leq 1.$$

Denote $\mathcal{F}_{s-1} \triangleq (A_1, \epsilon_1, \ldots, A_{s-1}, \epsilon_{s-1})$. We will first bound, $\mathbb{E}[e^{Q_s}|\mathcal{F}_{s-1}]$,

$$\begin{aligned}
\mathbb{E}[e^{Q_s}|\mathcal{F}_{s-1}] &= \exp\left(\frac{-\mu^2}{2}(a^\top A_s)^2\right)\mathbb{E}\left[\exp\left(\frac{\mu}{\sigma}(a^\top A_s)\epsilon_s\right)\Big|\mathcal{F}_{s-1}\right] \\
&\leq \exp\left(\frac{-\mu^2}{2}(a^\top A_s)^2\right)\exp\left(\frac{\sigma^2}{2}\times\left(\frac{\mu}{\sigma}(a^\top A_s)\right)^2\right) = 1
\end{aligned}$$

As, given $\mathcal{F}_{s-1} = \{A_1, \epsilon_1, \ldots, A_{s-1}, \epsilon_{s-1}\}$, A_s is fixed and ϵ_s is σ-sub-Gaussian.

Try at home: Show that $\xi_t^\top a$ is a martingale w.r.t. $\mathcal{F} = \{\mathcal{F}_s\}_s$.

Proof of Claim 1: Martingale concentration (cont'd)

Therefore,

$$
\begin{aligned}
\mathbb{E}\left[\exp\left(\mu A - \frac{\mu^2}{2}B^2\right)\right] &= \mathbb{E}\left[e^{\sum_{s=1}^{t} Q_s}\right] \\
&= \mathbb{E}\left[\left[e^{\sum_{s=1}^{t} Q_s}\middle|\mathcal{F}_{t-1}\right]\right] \\
&= \mathbb{E}\left[e^{\sum_{s=1}^{t-1} Q_s}\mathbb{E}\left[e^{Q_t}\middle|\mathcal{F}_{t-1}\right]\right] \qquad \texttt{as we have fixed } A_1, \epsilon_1, \ldots, A_{t-1}, \epsilon_{t-1} \\
&\le \mathbb{E}\left[e^{\sum_{s=1}^{t-1} Q_s}\right] \le \cdots \le 1.
\end{aligned}
$$

This verifies the condition for the lemma.

Proof of Claim 1: Martingale concentration (cont'd)

Lemma (Corollary 2.2 from de La Pena et al 2004). If A, B are random variables such that $\mathbb{E}\left[e^{\mu A - \frac{\mu^2 B^2}{2}}\right] \leq 1$ for all $\mu \in \mathbb{R}$, then for all $\tau \geq \sqrt{2}$, and $y > 0$, we have

$$\mathbb{P}\left(|A| \geq \tau \sqrt{(B^2 + y)\left(1 + (1/2)\log\left(1 + B^2/y\right)\right)}\right) \leq e^{-\tau^2/2}.$$

- We are applying with, $A = \frac{a^\top \xi_t}{\sigma}$ and $B = \|a\|_{W_t}$, $y = \|a\|_2^2$, and $\tau = \sqrt{2\log(T^2 d)}$.

Therefore, we have the following with probability at least $1 - \frac{1}{T^2 d}$,

$$\begin{aligned}\left|\frac{a^\top \xi_t}{\sigma}\right| &\leq \sqrt{2\log(T^2 d)}\sqrt{(a^\top W_t a + a^\top a)\left(1 + \frac{1}{2}\log\left(1 + \frac{a^\top W_t a}{\|a\|_2^2}\right)\right)} \\ &\leq \sqrt{\log(T^2 d)\,(2 + \log(1 + \mathrm{eig}_1(W_t)))} \cdot \|a\|_{V_t}\end{aligned}$$

As $\mathrm{eig}_1(A) = \max_x \frac{x^\top A x}{x^\top x}$ and $a^\top W_t a + a^\top a = a^\top (W_t + I) a = a^\top V_t a$.

Proof of Claim 1: Martingale concentration (cont'd)

Step 3 of Claim 1 proof: sufficient to show, for $a \in \{V_t^{-1/2} e_1, \ldots, V_t^{-1/2} e_d\}$,

$$\mathbb{P}\left(\frac{|\xi_t^\top a|}{\sigma} > \sqrt{\log(dT^2)(2 + \log(1 + tL^2))} \right) \leq \frac{1}{T^2 d}.$$

We just showed that, for any $a \in \mathbb{R}^d$, with probability at least $1 - 1/T^2 d$,

$$\left| \frac{a^\top \xi_t}{\sigma} \right| \leq \sqrt{\log(T^2 d)\,(2 + \log(1 + \mathrm{eig}_1(W_t)))} \cdot \|a\|_{V_t}$$

The proof is completed by the following observations:

- When $a = V_t^{-1/2} e_i$, $\|a\|_{V_t}^2 = a^\top V_t a = e_i V_t^{-1/2} V_t V_t^{-1/2} e_i = e_i^\top e_i = 1$.
- $\mathrm{eig}_1(W_t) \leq \mathrm{trace}(W_t) = \sum_{s=1}^t (A_s^\top A_s) \leq tL^2$, as $\max_{a \in \mathcal{A}} a^\top a \leq L^2$.

$\square$

LinUCB: Proof summary.

A general recipe for bounding the regret of UCB (optimistic) in structured bandits (e.g linear, generalized linear, kernelized (GP) bandits).

1. First, consider the pseudo-regret, $\overline{R}_T = \sum_{t=1}^{T}(\theta_\star^\top a_\star - \theta_\star^\top A_t)$.
2. Define a good event G, where the confidence intervals trap the true means. Then,

$$R_T = \mathbb{E}[\overline{R}_T | G] \underbrace{\mathbb{P}(G)}_{\leq 1} + \underbrace{\mathbb{E}[\overline{R}_T | G^c]}_{\leq T} \underbrace{\mathbb{P}(G^c)}_{\leq \text{small}}.$$

3. Use martingale concentration to bound $\mathbb{P}(G^c)$.
4. Under G, we can bound the instantaneous pseudo-regret

$$\theta_\star^\top a_\star - \theta_\star^\top A_t \leq 2 \times e_{A_t,t-1}.$$

   where $e_{A_t,t-1}$ is the width of the confidence interval of A_t at round $t-1$.
5. Bound the summation $\sum_{t=1}^{T} e_{A_t,t-1}$ (usually requires setting-specific techniques).