# CMPT 419/983: Theoretical Foundations of Reinforcement Learning

Lecture 2

Sharan Vaswani

September 15, 2023

# Recap

- **Input**: $K$ arms (possible actions), $T$ rounds. $\mu_a := \mathbb{E}_{r\sim\nu_a}[r]$ is the (unknown) expected reward obtained by choosing action $a$.
- **Protocol**: In each round $t \in [T]$, the bandit algorithm chooses action $a_t \in [K]$ and observes reward $R_t \sim \nu_{a_t}$.
- **Objective**: Minimize $\text{Regret}(T) := \sum_{t=1}^{T} [\mu^* - \mathbb{E}[R_t]] = \sum_{a=1}^{K} \Delta_a \, \mathbb{E}[N_a(T)]$.
- **Assumption**: $\eta_t := R_t - \mu_{a_t}$ is 1 sub-Gaussian i.e. for all $\lambda \in \mathbb{R}$, $\mathbb{E}[\exp(\lambda \eta_t)] \le \exp\left(\frac{\lambda^2}{2}\right)$.
- **Concentration for sub-Gaussian r.v.**: If $X$ is centered and $\sigma$ sub-Gaussian, then for any $\epsilon \ge 0$, $\Pr[X \ge \epsilon] \le \exp\left(-\frac{\epsilon^2}{2\sigma^2}\right)$. For $n$ i.i.d r.v's $X_i$ s.t. $\mathbb{E}[X_i] = \mu$, if $\hat{\mu} := \frac{1}{n}\sum_{i=1}^{n} X_i$ and $X_i - \mu$ is $\sigma$ sub-Gaussian, then $\Pr[|\hat{\mu} - \mu| \ge \epsilon] \le \exp\left(-\frac{n\epsilon^2}{2\sigma^2}\right)$
- **Explore-then-Commit (ETC)**: Under a sub-Gaussian assumption, ETC results in $O(\sqrt{KT})$ regret when exploring for $m = O\left(\frac{1}{\Delta^2}\right)$ rounds, while it can only result in $O(T^{2/3})$ regret when $m$ is set indepndent of $\Delta$.

# $\epsilon$-greedy Algorithm

**Algorithm** $\epsilon$-greedy (EG)

**Input**: $\{\epsilon_t\}_{t=1}^{T}$
**for** $t = 1 \to K$ **do**
    Select arm $a_t = t$ and observe $R_t$
**end for**
Calculate empirical mean reward for arm $a \in [K]$ as $\hat{\mu}_a(K) := \frac{\sum_{t=1}^{K} R_t \, \mathcal{I}\{a_t = a\}}{N_a(K)}$
**for** $t = K + 1 \to T$ **do**
    Select arm $\begin{cases} a_t = \arg\max_{a \in [K]} \hat{\mu}_a(t-1) \ w.p \ 1 - \epsilon_t \\ a_t \sim \mathcal{U}\{1, 2, \dots, K\} \ w.p \ \epsilon_t \end{cases}$
    Observe reward $R_t$ and update for $a \in [K]$:

$$N_a(t) = N_a(t-1) + \mathcal{I}\{a_t = a\} \quad ; \quad \hat{\mu}_a(t) = \frac{N_a(t-1)\,\hat{\mu}_a(t-1) + R_t \, \mathcal{I}\{a_t = a\}}{N_a(t)}$$

**end for**

- EG with $\epsilon_t = \epsilon$ can result in linear regret.
- For $K = 2$, EG with $\epsilon_t = O\left(\frac{1}{\Delta^2 t}\right)$ incurs $O\left(\frac{\log(T)}{\Delta^2}\right)$ regret.

Prove in Assignment 1!

# Upper Confidence Bound (UCB) Algorithm

- Based on the principle of *optimism in the face of uncertainty*.

**Algorithm** Upper Confidence Bound

**Input**: $\delta$
For each arm $a \in [K]$, initialize $U_a(0, \delta) := \infty$.
**for** $t = 1 \to T$ **do**
  Select arm $a_t = \arg\max_{a \in [K]} U_a(t-1, \delta)$ *(Choose the lower-indexed arm in case of a tie)*
  Observe reward $R_t$ and update for $a \in [K]$:

$$N_a(t) = N_a(t-1) + \mathcal{I}\{a_t = a\} \quad ; \quad \hat{\mu}_a(t) = \frac{N_a(t-1)\,\hat{\mu}_a(t-1) + R_t\,\mathcal{I}\{a_t = a\}}{N_a(t)}$$

$$U_a(t, \delta) = \hat{\mu}_a(t) + \sqrt{\frac{2\,\log(1/\delta)}{N_a(t)}}$$

**end for**

- Intuitively, UCB pulls a "promising" arm (with higher empirical mean $\hat{\mu}_a$) or one that has not been explored enough (with lower $N_a(t)$).

# UCB – Regret Analysis

**Claim**: UCB with $\delta = \frac{1}{T^2}$ achieves the following problem-dependent bound on the regret,

$$\text{Regret}(\text{UCB}, T) \leq 2 \sum_{a=1}^{K} \Delta_a + \sum_{a \in [K] | \Delta_a > 0} \frac{16 \log(T)}{\Delta_a}$$

*Proof*: Without loss of generality, assume that arm 1 is the best arm. Using the regret decomposition, we know that $\text{Regret}(\text{UCB}, T) = \sum_a \Delta_a \, \mathbb{E}[N_a(T)]$. Define a threshold $\tau_a$ and $\hat{\mu}_{a,\tau_a}$ as the mean for arm $a$ after pulling it for the first $\tau_a$ times. Define a "good" event $G_a$ for each $a \neq 1$.

$$G_a = \left\{ \mu_1 < \min_{t \in [T]} U_1(t, \delta) \right\} \cap \left\{ \hat{\mu}_{a,\tau_a} + \sqrt{\frac{2 \log(1/\delta)}{\tau_a}} < \mu_1 \right\}$$

Consider two cases when bounding $\mathbb{E}[N_a(T)]$. Using the law of total expectation,

$$\begin{aligned} \mathbb{E}[N_a(T)] &= \mathbb{E}[N_a(T) | G_a] \Pr[G_a] + \mathbb{E}[N_a(T) | G_a^c] \Pr[G_a^c] \\ &\leq \underbrace{\mathbb{E}[N_a(T) | G_a]}_{\text{Term (i)}} + T \underbrace{\Pr[G_a^c]}_{\text{Term (ii)}} \qquad (N_a(T) \leq T \text{ for all } a, \Pr[G_a] \leq 1) \end{aligned}$$

# UCB – Regret Analysis

Recall that $G_a = \{\mu_1 < \min_{t\in[T]} U_1(t,\delta)\} \cap \left\{\hat{\mu}_{a,\tau_a} + \sqrt{\frac{2\log(1/\delta)}{\tau_a}} < \mu_1\right\}$. We will show (by contradiction) that Term (i) $= \mathbb{E}[N_a(T)|G_a] \leq \tau_a$.

Suppose $\mathbb{E}[N_a(T)|G_a] > \tau_a$, then there is a round $t$ s.t. $N_a(t-1) = \tau_a$, $a_t = a$. Since $a_t = \arg\max_a U_a(t-1,\delta)$, it follows that $U_a(t-1,\delta) > U_1(t-1,\delta)$. However, we know that,

$$U_a(t-1,\delta) = \hat{\mu}_a(t-1) + \sqrt{\frac{2\log(1/\delta)}{N_a(t-1)}} = \hat{\mu}_a(t-1) + \sqrt{\frac{2\log(1/\delta)}{\tau_a}}$$

(By assumption, $N_a(t-1) = \tau_a$)

$$= \hat{\mu}_{a,\tau_a} + \sqrt{\frac{2\log(1/\delta)}{\tau_a}}$$

(Since arm $a$ has been pulled $\tau_a$ times)

$$\leq \mu_1 < U_1(t-1,\delta)\,,$$

(Since we are conditioning on $G_a$)

which is a contradiction. Hence, $\mathbb{E}[N_a(T)|G_a] \leq \tau_a$.

# UCB – Regret Analysis

Bounding Term (ii) $= \Pr[G_a^c] \leq \Pr\left[\mu_1 \geq \min_{t\in[T]} U_1(t,\delta)\right] + \Pr\left[\hat{\mu}_{a,\tau_a} + \sqrt{\frac{2\log(1/\delta)}{\tau_a}} \geq \mu_1\right]$.

$$\begin{aligned}
\left\{\mu_1 \geq \min_{t\in[T]} U_1(t,\delta)\right\} &= \left\{\mu_1 \geq \min_{t\in[T]}\left\{\hat{\mu}_1(t) + \sqrt{\frac{2\log(1/\delta)}{N_1(t)}}\right\}\right\} \\
&= \left\{\mu_1 \geq \min_{s\in[T]}\left\{\hat{\mu}_{1,s} + \sqrt{\frac{2\log(1/\delta)}{s}}\right\}\right\} \\
&= \bigcup_{s=1}^{T}\left\{\mu_1 \geq \hat{\mu}_{1,s} + \sqrt{\frac{2\log(1/\delta)}{s}}\right\} \\
\implies \Pr\left[\mu_1 \geq \min_{t\in[T]} U_1(t,\delta)\right] &\leq \sum_{s=1}^{T}\Pr\left[\mu_1 \geq \hat{\mu}_{1,s} + \sqrt{\frac{2\log(1/\delta)}{s}}\right] \qquad \text{(Union Bound)} \\
&\leq \sum_{s=1}^{T}\delta = \delta T \qquad \text{(Using concentration for sub-Gaussian r.v's)}
\end{aligned}$$

# UCB – Regret Analysis

Recall that Term (ii) $= \Pr[G_a^c] \leq \delta T + \Pr\left[\hat{\mu}_{a,\tau_a} + \sqrt{\frac{2\log(1/\delta)}{\tau_a}} \geq \mu_1\right]$. Assume that $\tau_a$ is chosen such that $\Delta_a - \sqrt{\frac{2\log(1/\delta)}{\tau_a}} \geq \frac{\Delta_a}{2}$.

$$\Pr\left[\hat{\mu}_{a,\tau_a} + \sqrt{\frac{2\log(1/\delta)}{\tau_a}} \geq \mu_1\right] = \Pr\left[\hat{\mu}_{a,\tau_a} - \mu_a + \sqrt{\frac{2\log(1/\delta)}{\tau_a}} \geq \Delta_a\right] \leq \Pr\left[\hat{\mu}_{a,\tau_a} - \mu_a \geq \frac{\Delta_a}{2}\right]$$

$$\leq \exp\left(-\frac{\tau_a \Delta_a^2}{8}\right)$$

(Using concentration for sub-Gaussian r.v's)

Putting everything together,

$$\implies \Pr[G_a^c] \leq \delta T + \exp\left(-\frac{\tau_a \Delta_a^2}{8}\right)$$

$$\implies \mathbb{E}[N_a(T)] \leq \tau_a + T\left[\delta T + \exp\left(-\frac{\tau_a \Delta_a^2}{8}\right)\right]$$

# UCB – Regret Analysis

Recall that $\mathbb{E}[N_a(T)] \leq \tau_a + T\left[\delta T + \exp\left(-\frac{\tau_a \Delta_a^2}{8}\right)\right]$.

$$
\begin{aligned}
\mathbb{E}[N_a(T)] &\leq \frac{8\log(1/\delta)}{\Delta_a^2} + T\,[\delta T + \delta] && \left(\text{Setting } \tau_a = \tfrac{8\log(1/\delta)}{\Delta_a^2}\right) \\
&\leq \frac{8\log(1/\delta)}{\Delta_a^2} + 2\delta T^2 \\
&= \frac{16\log(T)}{\Delta_a^2} + 2 && \left(\text{Setting } \delta = 1/T^2\right)
\end{aligned}
$$

$$
\implies \text{Regret}(\text{UCB}, T) = \sum_a \Delta_a \mathbb{E}[N_a(T)] = 2\sum_{a=1}^{K} \Delta_a + \sum_{a=2}^{K} \frac{16\log(T)}{\Delta_a} \quad \square
$$

# UCB – Regret Analysis

**Claim**: For $\Delta \leq 1$, UCB with $\delta = \frac{1}{T^2}$ achieves the following worst-case regret,

$$\text{Regret}(\text{UCB}, T) \leq 2K + 8\sqrt{K\;T\;\log(T)}$$

*Proof*: Define $C > 0$ to be a constant to be tuned later. From the regret decomposition result,

$$\begin{aligned}
\text{Regret}(\text{UCB}, T) = \sum_{a=1}^{K} \Delta_a\, \mathbb{E}[N_a(T)] &= \sum_{a \mid \Delta_a < C} \Delta_a\, \mathbb{E}[N_a(T)] + \sum_{a \mid \Delta_a \geq C} \Delta_a\, \mathbb{E}[N_a(T)] \\
&\leq CT + \sum_{a \mid \Delta_a \geq C} \Delta_a\, \mathbb{E}[N_a(T)] && \text{(Since } \textstyle\sum_{a=1}^{K} N_a(T) = T\text{)} \\
&\leq CT + \sum_{a \mid \Delta_a \geq C} \left[\frac{16\log(T)}{\Delta_a} + 2\Delta_a\right] && \text{(From the previous slide)} \\
&\leq CT + \left[\frac{16K\;\log(T)}{C} + \sum_{a \mid \Delta_a \geq C} 2\Delta_a\right] && \text{(Setting } C = \sqrt{\tfrac{16K\log(T)}{T}}\text{)}
\end{aligned}$$

$$\implies \text{Regret}(\text{UCB}, T) \leq 8\sqrt{K\;T\;\log(T)} + 2K\Delta_a \leq 2K + 8\sqrt{K\;T\;\log(T)}$$

# UCB vs ETC

- Similar to best-tuned ETC, UCB results in an $\tilde{O}(\sqrt{KT})$ problem-independent regret.
- Unlike best-tuned ETC, UCB does not need to know the gaps $\Delta$ to set algorithm parameters, but does require knowledge of the horizon $T$.

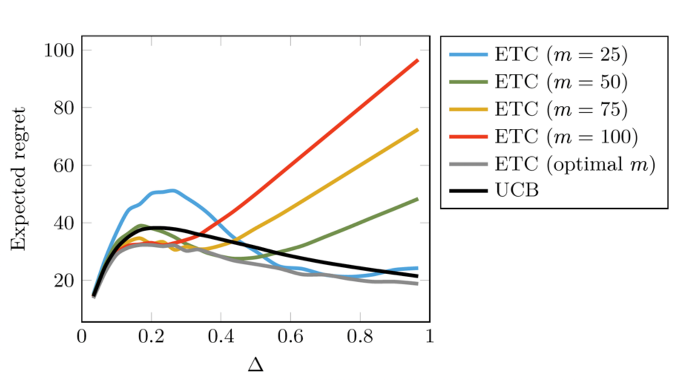


**Figure 1:** For $K = 2$, $T = 1000$, Gaussian rewards, comparing UCB and ETC(m) as a function of the gap $\Delta$.

# Improvements to UCB

- **Problem**: UCB requires knowledge of $T$ and hence, the number of rounds needs to be fixed.
- *Sol*: Define UCB as $\hat{\mu}_a(t) + \sqrt{\frac{2\log(f(t))}{N_a(t)}}$ where $f(t) := 1 + t\log^2(t)$. No dependence on $T$, but results in the same $O(\sqrt{KT\log(T)})$ worst-case regret. (see [LS20, Chapter 8])
- **Lower-Bound**: For a fixed $T$ and for every bandit algorithm, there exists a stochastic bandit problem with rewards in $[0, 1]$ such that $\text{Regret}(T) = \Omega(\sqrt{KT})$. (see [LS20, Chapter 15]).
- **Problem**: UCB is sub-optimal by a $\sqrt{\log(T)}$ factor compared to the lower-bound. Is it possible to develop an algorithm that does not incur this log factor?
- *Sol*: [Lat18, MG17] propose modifications of UCB that achieve $O(\sqrt{KT})$ regret.

# Stochastic Linear Bandits

# Stochastic Linear Bandits

- MAB treat each arm (e.g. drug choice) independently. But the arms (and their rewards) can be dependent. E.g., drugs with similar chemical composition can have similar side-effects.
- Stochastic Linear Bandits can model linear dependence between different arms. For this, we require *feature vectors* $X_a \in \mathbb{R}^d$ for each arm $a \in [K]$.

- **Reward Model**: For an unknown vector $\theta^* \in \mathbb{R}^d$, the mean reward for arm $a$ is given as: $\mu_a = \langle X_a, \theta^* \rangle$. Hence, arms with similar feature vectors will have similar mean rewards.
- Similar to the MAB setting, on pulling arm $a_t$ at round $t$, we observe the reward $R_t = \mu_{a_t} + \eta_t = \langle X_t, \theta^* \rangle + \eta_t$. We will assume that $\eta_t$ is conditionally 1 sub-Gaussian, i.e. if $\mathcal{H}_{t-1} := \{X_1, R_1, \ldots, X_t\}$ is the *history* of interactions until round $t$, then for all $\lambda \in \mathbb{R}$, $\mathbb{E}[\exp(\lambda \eta_t) | \mathcal{H}_{t-1}] \leq \exp(\lambda^2/2)$.

- $\text{Regret}(T) := \sum_{t=1}^{T} \left[\max_{a \in [K]} \langle X_a, \theta^* \rangle - \mathbb{E}[R_t]\right] = T \max_{a \in [K]} \langle X_a, \theta^* \rangle - \sum_{t=1}^{T} \mathbb{E}[R_t]$.
- In the special case, when all the arms are independent, i.e. $d = K$ and $\forall a \in [K]$, $X_a = e_a$ where $\forall i \in [d], i \neq a$, $e_a[i] = 0$ and $e_a[a] = 1$. Hence, $\mu_a = \theta^*_a$ and the linear bandit setup strictly generalizes MAB.

# Stochastic Linear Bandits – Estimating $\hat{\mu}_a(t)$

At round $t$, we have collected the following data: $\{X_s, R_s\}_{s=1}^t$. **Q**: How do we estimate $\hat{\mu}_a(t)$?

By solving regularized ridge regression, i.e. for a regularization parameter $\lambda \geq 0$,

$$\hat{\theta}_t := \arg\min_{\theta} \left\{ \frac{1}{2} \sum_{s=1}^{t} \left[\langle X_s, \theta \rangle - R_s\right]^2 + \frac{\lambda}{2} \|\theta\|^2 \right\}$$

Setting the derivative to zero to solve the above minimization problem,

$$\sum_{s=1}^{t} \left[ X_s \left[ \langle X_s, \hat{\theta}_t \rangle - R_s \right] \right] + \lambda \hat{\theta}_t = 0$$

$$\implies \underbrace{\left[ \sum_{s=1}^{t} X_s X_s^T + \lambda I_d \right]}_{:= V_t \in \mathbb{R}^{d \times d}} \hat{\theta}_t = \underbrace{\sum_{s=1}^{t} X_s R_s}_{:= b_t \in \mathbb{R}^{d \times 1}} \implies V_t \hat{\theta}_t = b_t \implies \hat{\theta}_t = V_t^{-1} b_t$$

Hence, the empirical mean for each arm after $t$ rounds: $\hat{\mu}_a = \langle X_a, \hat{\theta}_t \rangle = X_a^T V_t^{-1} b_t$

# Linear UCB

**Algorithm** Linear Upper Confidence Bound

1: **Input**: $\{\beta_t\}_{t=1}^T$, $V_0 = \lambda I_d \in \mathbb{R}^{d\times d}$

2: For each arm $a \in [K]$, initialize $U_a(0,\delta) := \infty$.

3: **for** $t = 1 \to T$ **do**

4: Select arm $a_t = \arg\max_{a\in[K]} U_a(t-1,\delta)$ *(Choose the lower-indexed arm in case of a tie)*

5: Observe reward $R_t$ and update:

$$V_t = V_{t-1} + X_t X_t^T \quad ; \quad b_t = b_{t-1} + R_t X_t \quad ; \quad \hat{\theta}_t = V_t^{-1} b_t$$

$$U_a(t) = \langle X_a, \hat{\theta}_t \rangle + \sqrt{\beta_t}\, \|X_a\|_{V_t^{-1}} \qquad \text{(where } \|x\|_A := \sqrt{x^\intercal A x}\text{)}$$

6: **end for**

In the special case, when all the arms are independent, Linear UCB with $\beta_t = \beta = 2\log(1/\delta)$ is equivalent to UCB, and hence, Linear UCB strictly generalizes UCB.

Prove this in Assignment 1!

# Linear UCB – Regret Analysis

**Claim**: $U_a(t) := \langle X_a, \hat{\theta}_t \rangle + \sqrt{\beta_t}\ \|X_a\|_{V_t^{-1}} = \max_{\theta \in \mathcal{C}_t} \langle \theta, X_a \rangle$ where $\mathcal{C}_t = \left\{ \theta \mid \left\| \theta - \hat{\theta}_t \right\|_{V_t}^2 \leq \beta_t \right\}$.

$\mathcal{C}_t$ is an ellipsoid centered at $\hat{\theta}_t$ with the principle axes being the eigenvectors of $V_t$ and the corresponding lengths being the reciprocal of the eigenvalues. As $t$ increases, the eigenvalues of matrix $V_t$ increases and the volume of the ellipsoid decreases.

Prove this in Assignment 1! For the subsequent proof, we will use this equivalence.

**Claim**: Assuming (i) $\|\theta^*\| \leq 1$, (ii) $\|X_a\| \leq 1$ for all $a$ and (iii) $R_t \in [0, 1]$, UCB with $\sqrt{\beta_t} = \sqrt{d \log\left(\frac{\lambda d + t}{\lambda\, d}\right) + 2\log(1/\delta)} + \sqrt{\lambda}$ achieves the following worst-case bound on the regret,

$$\text{Regret}(\text{LinUCB}, T) \leq O\left(d\sqrt{T}\log(T)\right)$$

# Linear UCB – Regret Analysis

*Proof*: Define a "good" event $G := \{\forall t \in [T] | \theta^* \in \mathcal{C}_t := \left\{\theta \mid \left\|\theta - \hat{\theta}_t\right\|_{V_t}^2 \leq \beta_t\right\}$, and denote the instantaneous expected regret at round $t$ as $r_t = \max_a \langle X_a, \theta^* \rangle - \langle X_t, \theta^* \rangle$. Using the law of total expectation,

$$\begin{aligned}
\text{Regret}(\text{LinUCB}, T) &= \mathbb{E}[\text{Regret}(\text{LinUCB}, T) | G] \Pr[G] + \mathbb{E}[\text{Regret}(T) | G^c] \Pr[G^c] \\
&\leq \mathbb{E}[\text{Regret}(\text{LinUCB}, T) | G] + T \Pr[G^c] \\
&\qquad\qquad (\text{Regret}(\text{LinUCB}, T) \leq T \text{ and } \Pr[G] \leq 1) \\
&= \sum_{t=1}^{T} \mathbb{E}[r_t | G] + T \Pr[G^c] \leq \sqrt{T \sum_{t=1}^{T} [\mathbb{E}[r_t | G]]^2} + T \Pr[G^c] \\
&\qquad (\text{Cauchy Schwarz inequality: } \langle x, y \rangle \leq \|x\| \, \|y\| \text{ with } x, y \in \mathbb{R}^T \text{ and } x[t] = 1, y[t] = r_t)
\end{aligned}$$

# Linear UCB – Regret Analysis

Recall that $\text{Regret}(\text{LinUCB}, T) \le \sqrt{T \sum_{t=1}^{T} [\mathbb{E}[r_t|G]]^2} + T \Pr[G^c]$. Let us first bound $\mathbb{E}[r_t|G]$. If event $G$ happens, then $\theta^* \in \mathcal{C}_t$. Hence, for all $a \in [K]$,

$$\langle \theta^*, X_a \rangle \le \max_{\theta \in \mathcal{C}_t} \langle \theta, X_a \rangle = U_a(t) \le U_{a_t}(t)$$

(Using the equivalence on Slide 15 and the algorithm)

$$\implies \max_{a \in [K]} \langle \theta^*, X_a \rangle \le U_{a_t}(t) = \max_{\theta \in \mathcal{C}_t} \langle \theta, X_t \rangle = \langle \tilde{\theta}_t, X_t \rangle \qquad (\tilde{\theta}_t := \arg\max_{\theta \in \mathcal{C}_t} \langle \theta, X_t \rangle)$$

$$\implies \mathbb{E}[r_t|G] = \mathbb{E}[\max_a \langle X_a, \theta^* \rangle - \langle X_t, \theta^* \rangle | G] \le \langle \tilde{\theta}_t - \theta^*, X_t \rangle$$

$$\le \left\| \tilde{\theta}_t - \theta^* \right\|_{V_t} \|X_t\|_{V_t^{-1}}$$

(Cauchy Schwarz inequality with $x, y \in \mathbb{R}^d$ and $x = V_t^{1/2}(\tilde{\theta}_t - \theta^*)$, $y = V_t^{-1/2} X_t$)

$$\le \left[ \left\| \tilde{\theta}_t - \hat{\theta}_t \right\|_{V_t} + \left\| \theta^* - \hat{\theta}_t \right\|_{V_t} \right] \|X_t\|_{V_t^{-1}} \qquad \text{(Triangle inequality)}$$

$$\implies \mathbb{E}[r_t|G] \le 2\sqrt{\beta_t} \, \|X_t\|_{V_t^{-1}} \qquad (\text{Since } \theta^*, \tilde{\theta}_t \in \mathcal{C}_t)$$

# Linear UCB – Regret Analysis

Putting everything together,

$$\text{Regret}(\text{LinUCB}, T) \leq \sqrt{T \sum_{t=1}^{T} [\mathbb{E}[r_t | G]]^2 + T \Pr[G^c]} \leq 2\sqrt{T \sum_{t=1}^{T} \beta_t \, \|X_t\|^2_{V_t^{-1}} + T \Pr[G^c]}$$

$$\leq 2\sqrt{T \, \beta_T \sum_{t=1}^{T} \|X_t\|^2_{V_t^{-1}} + T \Pr[G^c]} \qquad \text{(Since } \beta_t \leq \beta_T \text{ for all } t \in [T]\text{)}$$

We will prove the following results: (i) $\sum_{t=1}^{T} \|X_t\|^2_{V_t^{-1}} \leq 2d \log\left(\frac{\lambda d + T}{\lambda d}\right)$ and (ii) $\sqrt{\beta_t} = \sqrt{d \log\left(\frac{\lambda d + t}{\lambda d}\right) + 2\log(T)} + \sqrt{\lambda}$, $\Pr[G^c] \leq \frac{1}{T}$.

Given these results,

$$\text{Regret}(\text{LinUCB}, T) \leq 2\sqrt{2d \, T \, \beta_T \log\left(\frac{\lambda d + T}{\lambda d}\right)} + 1 = O\left(d\sqrt{T} \log(T)\right) \quad \square$$

# Linear UCB – Regret Analysis

**Claim**: If $\|X_a\| \leq 1$ for all $a$, $\sum_{t=1}^{T} \|X_t\|^2_{V_t^{-1}} \leq 2d \log\left(\frac{\lambda d + T}{\lambda d}\right)$.

*Proof*:

$$V_t = V_{t-1} + X_t X_t^{\intercal} = V_{t-1}^{1/2}\left[I_d + V_{t-1}^{-1/2} X_t X_t^{\intercal} V_{t-1}^{-1/2}\right] V_{t-1}^{1/2}$$

$$\implies \det[V_t] = \det[V_{t-1}^{1/2}] \det\left[I_d + V_{t-1}^{-1/2} X_t X_t^{\intercal} V_{t-1}^{-1/2}\right] \det[V_{t-1}^{1/2}]$$

$$(\det[XY] = \det[X]\det[Y])$$

$$= \det[V_{t-1}] \det\left[I_d + V_{t-1}^{-1/2} X_t \, [V_{t-1}^{-1/2} X_t]^{\intercal}\right] \quad (\det[X^{1/2}] = \sqrt{\det[X]})$$

$$= \det[V_{t-1}] \left(1 + \left\|V_{t-1}^{-1/2} X_t\right\|^2\right) = \det[V_{t-1}] \left(1 + \|X_t\|^2_{V_t^{-1}}\right)$$

$$(\text{Matrix Determinant Lemma: } \det[I_d + x x^{\intercal}] = 1 + x^{\intercal} x = 1 + \|x\|^2)$$

$$\implies \ln\left(1 + \|X_t\|^2_{V_t^{-1}}\right) = \ln\left(\frac{\det[V_t]}{\det[V_{t-1}]}\right)$$

# Linear UCB – Regret Analysis

Recall that $\ln\left(1+\|X_t\|^2_{V_t^{-1}}\right)=\ln\left(\frac{\det[V_t]}{\det[V_{t-1}]}\right)$.

Hence, $\sum_{t=1}^T \ln\left(1+\|X_t\|^2_{V_t^{-1}}\right)=\ln\left(\frac{\det[V_T]}{\det[V_0]}\right)$. For any $x\geq 0$, $x\leq 2\ln(1+x)$. Hence, $\sum_{t=1}^T \|X_t\|^2_{V_t^{-1}}\leq 2\sum_{t=1}^T \ln(1+\|X_t\|^2_{V_t^{-1}})$, implying,

$$\sum_{t=1}^T \|X_t\|^2_{V_t^{-1}}\leq 2\sum_{t=1}^T \ln(1+\|X_t\|^2_{V_t^{-1}})=2\ln\left(\frac{\det[V_T]}{\det[V_0]}\right)$$

$$\det[V_T]\leq\left(\frac{\mathrm{Tr}[V_T]}{d}\right)^d \quad \left(\det[A]=\prod\lambda_i=\left(\left(\prod\lambda_i\right)^{1/d}\right)^d\leq\left(\frac{\sum\lambda_i}{d}\right)^d=\left(\frac{\mathrm{Tr}[A]}{d}\right)^{1/d}\right)$$

$$=\left(\frac{\mathrm{Tr}[V_0+\sum_{t=1}^T X_t X_t^\intercal]}{d}\right)^d\leq\left(\frac{\mathrm{Tr}[V_0]+T}{d}\right)^d=\left(\frac{d\,\lambda+T}{d}\right)^d$$

(Since $\|X_t\|\leq 1$)

$$\implies\sum_{t=1}^T \|X_t\|^2_{V_t^{-1}}\leq 2\ln\left(\left(\frac{(d\,\lambda+T)/d}{(\det[V_0])^{1/d}}\right)^d\right)=2d\,\log\left(\frac{\lambda d+T}{\lambda d}\right)\quad\square$$

# Digression – (Super)-Martingales

**Martingale**: Sequence of random variables for which, at a particular time, the conditional expectation of the next value in the sequence is equal to the present value, regardless of all prior values.

A sequence of random variables – $M_1, M_2, \dots$ is a discrete-time martingale if for all $t$,

$$\mathbb{E}[|M_t|] \leq \infty \quad ; \quad \mathbb{E}[M_t | M_1, M_2, \dots M_{t-1}] = M_{t-1}$$

*Example 1*: An unbiased random walk
*Example 2*: Gambler's fortune: Suppose $M_t$ is a gambler's fortune after $t$ tosses of a fair coin, where the gambler wins $1 if the coin comes up heads and loses $1 if it comes up tails.

**Super-Martingale**: A sequence of random variables – $M_1, M_2, \dots$ is a discrete-time super-martingale if for all $t$,

$$\mathbb{E}[|M_t|] \leq \infty \quad ; \quad \mathbb{E}[M_t | M_1, M_2, \dots M_{t-1}] \leq M_{t-1}$$

# Linear UCB – Regret Analysis

**Claim**: If (i) $\|\theta^*\| \leq 1$ and (ii) $\|X_a\| \leq 1$ for all $a$, for $\sqrt{\beta_t} = \sqrt{d \log\left(\frac{\lambda d + t}{\lambda d}\right) + 2\log(T)} + \sqrt{\lambda}$ and $G := \{\forall t \in [T] | \theta^* \in \mathcal{C}_t := \left\{\theta \,|\, \left\|\theta - \hat{\theta}_t\right\|_{V_t}^2 \leq \beta_t\right\}$, $\Pr[G^c] \leq \frac{1}{T}$.

*Proof*: Define $S_t := \sum_{s=1}^t \eta_s X_s$ and $K_t := \sum_{s=1}^t X_s X_s^\intercal$. We will prove the claim in 4 steps:

(i) $\left\|\theta - \hat{\theta}_t\right\|_{V_t} \leq \|S_t\|_{V_t^{-1}} + \sqrt{\lambda}$.

(ii) $M_t(z) = \exp\left(\langle z, S_t\rangle - \frac{1}{2}\|z\|_{K_t}^2\right)$ is a non-negative super-martingale with $M_0(z) = 1$.

(iii) Use the fact that a mixture of super-martingales given by $\bar{M}_t = \int_z M_t(z) h(z)\, dz$ is also a non-negative super-martingale for any probability density function $h(z)$.

(iv) Use the maximal inequality for super-martingales to bound $\Pr\left[\sup_{t \in [T]} \log(\bar{M}_t(z)) \geq \log(1/\delta)\right]$ and hence bound $\left\|\theta - \hat{\theta}_t\right\|_{V_t}$.

# Linear UCB – Regret Analysis

**Part (i)**: If $S_t := \sum_{s=1}^{t} \eta_s X_s$ and $K_t := \sum_{s=1}^{t} X_s X_s^{\intercal}$, then $\left\| \theta^* - \hat{\theta}_t \right\|_{V_t} \leq \| S_t \|_{V_t^{-1}} + \sqrt{\lambda}$.

*Proof*:

$$
\begin{aligned}
b_t &= \sum_{s=1}^{t} X_s R_s = \sum_{s=1}^{t} X_s \left[ \langle X_s, \theta^* \rangle + \eta_s \right] \\
&= \sum_{s=1}^{t} X_s^{\intercal} X_s \theta^* + \sum_{s=1}^{t} X_s \eta_s = S_t + \sum_{s=1}^{t} X_s^{\intercal} X_s \theta^*.
\end{aligned}
$$

$$
\implies \hat{\theta}_t = V_t^{-1} b_t = V_t^{-1} S_t + V_t^{-1} \left[ \sum_{s=1}^{t} X_s^{\intercal} X_s \right] \theta^*
$$

$$
\begin{aligned}
\left\| \theta^* - \hat{\theta}_t \right\|_{V_t} &= \left\| V_t^{-1} S_t + \left( V_t^{-1} K_t - I_d \right) \theta^* \right\|_{V_t} = \| S_t \|_{V_t^{-1}} + \sqrt{\theta^{*\intercal} \left( V_t^{-1} K_t - I_d \right) \left( K_t - V_t \right) \theta^*} \\
&= \| S_t \|_{V_t^{-1}} + \sqrt{\lambda} \sqrt{\theta^{*\intercal} \left( I_d - V_t^{-1} K_t \right) \theta^*}
\end{aligned}
$$

$$
\implies \left\| \theta^* - \hat{\theta}_t \right\|_{V_t} \leq \| S_t \|_{V_t^{-1}} + \sqrt{\lambda} \, \| \theta^* \| \leq \| S_t \|_{V_t^{-1}} + \sqrt{\lambda} \quad \square
$$

# Linear UCB – Regret Analysis

**Part (ii)**: If $S_t := \sum_{s=1}^{t} \eta_s X_s$ and $K_t := \sum_{s=1}^{t} X_s X_s^{\mathsf{T}}$, $M_t(z) = \exp\left(\langle z, S_t\rangle - \frac{1}{2}\, \|z\|_{K_t}^2\right)$ is a non-negative super-martingale with $M_0(z) = 1$.

*Proof*: It is clear that $M_t(z) = \exp\left(\langle z, S_t\rangle - \frac{1}{2}\, \|z\|_{K_t}^2\right)$ is non-negative and $M_0(z) = 1$. By our assumption on the noise, $\mathbb{E}[\exp(\lambda \eta_t) | \mathcal{H}_{t-1}] \leq \exp\left(\frac{\lambda^2}{2}\right)$. Setting $\lambda = \langle z, X_t\rangle$, implies that $\mathbb{E}[\exp(\langle z, X_t\rangle \eta_t) | \mathcal{H}_{t-1}] \leq \exp\left(\frac{\|z\|_{X_t X_t^{\mathsf{T}}}^2}{2}\right) \implies \mathbb{E}\left[\exp(\langle z, X_t\rangle \eta_t) - \frac{\|z\|_{X_t X_t^{\mathsf{T}}}^2}{2} | \mathcal{H}_{t-1}\right] \leq 1$ (*).

$$
\begin{aligned}
\mathbb{E}[M_t(z)|\mathcal{H}_{t-1}] &= \mathbb{E}\left[\exp\left(\langle z, S_{t-1} + \eta_t X_t\rangle - \frac{1}{2}\, \|z\|_{K_{t-1}+X_t X_t^{\mathsf{T}}}^2\right) |\mathcal{H}_{t-1}\right] \\
&= \mathbb{E}\left[\exp\left(\langle z, \eta_t X_t\rangle - \frac{1}{2}\, \|z\|_{X_t X_t^{\mathsf{T}}}^2\right) |\mathcal{H}_{t-1}\right] \mathbb{E}\left[\exp\left(\langle z, S_{t-1}\rangle - \frac{1}{2}\, \|z\|_{K_{t-1}}^2\right) |\mathcal{H}_{t-1}\right] \\
&= M_{t-1}(z)\, \mathbb{E}\left[\exp\left(\langle z, \eta_t X_t\rangle - \frac{1}{2}\, \|z\|_{X_t X_t^{\mathsf{T}}}^2\right) |\mathcal{H}_{t-1}\right] \\
\implies \mathbb{E}[M_t(z)|\mathcal{H}_{t-1}] &\leq M_{t-1}(z) && \text{(Using (*))}
\end{aligned}
$$

# Linear UCB – Regret Analysis

**Fact 1**: For a probability density $h$, if $M_t(z)$ is a non-negative super-martingale with $M_0(z) = 1$, the "mixture" $\bar{M}_t := \int_z M_t(z)\, h(z)\, dz$ is also a non-negative super-martingale with $\bar{M}_0 = 1$.

**Fact 2**: For a non-negative super-martingale $\bar{M}_t$ s.t. $\bar{M}_0 = 1$, for any $\epsilon > 0$, $\Pr[\sup_{t \in [T]} \bar{M}_t \geq \epsilon] \leq \frac{1}{\epsilon}$.

In order to construct $\bar{M}_t$, we will choose $h = \mathcal{N}(0, H^{-1})$ and $H = \lambda I_d$.

$$\bar{M}_t = \int_z M_t(z)\, h(z)\, dz = \frac{1}{\sqrt{(2\pi)^d \det[H^{-1}]}} \int_z \exp\left(\langle z, S_t \rangle - \frac{1}{2} \|z\|_{K_t}^2 - \frac{1}{2} \|z\|_H^2\right) dz$$

From **Fact 1**, $\bar{M}_t$ is a non-negative super-martingale, and hence using **Fact 2** with $\epsilon = {}^1\!/\!{}_\delta$

$$\Pr\left[\sup_{t \in [T]} \bar{M}_t \geq \epsilon\right] = \Pr\left[\sup_{t \in [T]} \log(\bar{M}_t) \geq \log(\epsilon)\right] = \Pr\left[\sup_{t \in [T]} \log(\bar{M}_t) \geq \log(1/\delta)\right] \leq \delta$$

In the last part of the proof, we will relate $\bar{M}_t$ to $\|S_t\|_{V_t^{-1}}$.

# Linear UCB – Regret Analysis

Recall that $\bar{M}_t = \int_z M_t(z)\, h(z)\, dz = \frac{1}{\sqrt{(2\pi)^d \ \det[H^{-1}]}} \int_z \exp\left(\langle z, S_t\rangle - \frac{1}{2}\, \|z\|_{K_t}^2 - \frac{1}{2}\, \|z\|_H^2\right) dz.$

$$\langle z, S_t\rangle - \frac{1}{2}\, \|z\|_{K_t}^2 - \frac{1}{2}\, \|z\|_H^2 = \frac{1}{2}\, \|S_t\|_{(K_t+H)^{-1}}^2 - \frac{1}{2}\, \left\|z - (K_t + H)^{-1} S_t\right\|_{(K_t+H)}^2$$

$$\implies \int_z M_t(z)\, h(z)\, dz = \frac{\exp\left(\frac{1}{2}\, \|S_t\|_{V_t^{-1}}^2\right)}{\sqrt{(2\pi)^d \ \det[H^{-1}]}} \int_z \exp\left(-\frac{1}{2}\, \left\|z - V_t^{-1} S_t\right\|_{V_t}^2\right) dz$$

$$\implies \bar{M}_t = \frac{\exp\left(\frac{1}{2}\, \|S_t\|_{V_t^{-1}}^2\right)}{\sqrt{(2\pi)^d \ \det[H^{-1}]}} \sqrt{(2\pi)^d \ \det[V_t^{-1}]} = \sqrt{\frac{\det[H]}{\det[V_t]}} \exp\left(\frac{1}{2}\, \|S_t\|_{V_t^{-1}}^2\right)$$

(Integral of a Gaussian density)

# Linear UCB – Regret Analysis

Putting everything together, we know that for all $t \in [T]$, w.p $1-\delta$, $\log(\bar{M}_t) \le \log(1/\delta)$. Using the result from the previous slide, w.p $1-\delta$, for all $t \in [T]$

$$\frac{1}{2}\|S_t\|^2_{V_t^{-1}} + \frac{1}{2}\log\left(\frac{\det[H]}{\det[V_t]}\right) \le \log(1/\delta) \implies \|S_t\|_{V_t^{-1}} \le \sqrt{\log\left(\frac{\det[V_t]}{\lambda^d}\right) + 2\log(1/\delta)}$$

$$\implies \|S_t\|_{V_t^{-1}} \le \sqrt{d\log\left(\frac{\lambda d + t}{\lambda d}\right) + 2\log(1/\delta)}$$

From **Part (i)**, we know that,

$$\left\|\theta^* - \hat{\theta}_t\right\|_{V_t} \le \|S_t\|_{V_t^{-1}} + \sqrt{\lambda} \le \underbrace{\sqrt{d\log\left(\frac{\lambda d + t}{\lambda d}\right) + 2\log(1/\delta)} + \sqrt{\lambda}}_{:=\sqrt{\beta_t}}$$

Hence, we have shown that w.p. $1-\frac{1}{T}$, $\left\|\theta^* - \hat{\theta}_t\right\|^2_{V_t} \le \beta_t$, and hence $\Pr[G^c] \le \frac{1}{T}$ $\square$

# References i


- Tor Lattimore, *Refining the confidence level for optimistic bandit strategies*, The Journal of Machine Learning Research **19** (2018), no. 1, 765–796.
- Tor Lattimore and Csaba Szepesvári, *Bandit algorithms*, Cambridge University Press, 2020.
- Pierre Ménard and Aurélien Garivier, *A minimax and asymptotically optimal algorithm for stochastic bandits*, International Conference on Algorithmic Learning Theory, PMLR, 2017, pp. 223–237.