

# Linear Bandits

Kianté Brantley & Sham Kakade

## 1 Linear Bandits

- Setting
- LinUCB
- An Optimal Regret Bound

## 2 Analysis

- Regret Analysis
- Confidence Analysis

# Recap: The “Assumption Ladder” for Linear RL

- **Linear Feature Assumptions:** We explored a hierarchy of conditions for RL:
  - Bellman completeness
  - All-policy realizability
  - Linear  $Q^*$  realizability

# Recap: The “Assumption Ladder” for Linear RL

- **Linear Feature Assumptions:** We explored a hierarchy of conditions for RL:
  - Bellman completeness
  - All-policy realizability
  - Linear  $Q^*$  realizability
- **The Goal:** Can we achieve sample complexities that are  $\text{poly}(d, H)$ ?  
**Crucially**, we sought no dependence on the sizes of the state/action spaces ( $|S|$  or  $|A|$ ).

# Recap: The “Assumption Ladder” for Linear RL

- **Linear Feature Assumptions:** We explored a hierarchy of conditions for RL:

- Bellman completeness
- All-policy realizability
- Linear  $Q^*$  realizability

↳  $\Omega(\min\{2^d, 2^H\})$

- **The Goal:** Can we achieve sample complexities that are  $\text{poly}(d, H)$ ?

**Crucially,** we sought no dependence on the sizes of the state/action spaces ( $|S|$  or  $|A|$ ).

- **Today:** We strip away the horizon complexity and focus on the  $H = 1$  case.

- **Linear Bandits:** How do we explore efficiently in large/infinite action spaces?

## 1 Linear Bandits

- Setting
- LinUCB
- An Optimal Regret Bound

## 2 Analysis

- Regret Analysis
- Confidence Analysis

# Handling Large Action Spaces

- On each round, we must choose a decision  $x_t \in D \subset \mathbb{R}^d$ .

# Handling Large Action Spaces

- On each round, we must **choose a decision**  $x_t \in D \subset R^d$ .
- Obtain a reward  $r_t \in [-1, 1]$ , where

$$r_t = \mu^* \cdot x_t + \eta_t$$

with i.i.d noise  $\eta_t$ .

# Handling Large Action Spaces

- On each round, we must **choose a decision**  $x_t \in D \subset R^d$ .
- Obtain a reward  $r_t \in [-1, 1]$ , where

$$r_t = \mu^* \cdot x_t + \eta_t$$

with i.i.d noise  $\eta_t$ .

- the conditional expectation of  $r_t$  is linear,

$$\mathbb{E}[r_t | x_t = x] = \mu^* \cdot x$$

# Our Objective

If  $x_0, \dots, x_{T-1}$  are our decisions, then our **cumulative regret** is

$$R_T = T\mu^* \cdot x^* - \sum_{t=0}^{T-1} \mu^* \cdot x_t$$

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

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

## 1 Linear Bandits

- Setting
- **LinUCB**
- An Optimal Regret Bound

## 2 Analysis

- Regret Analysis
- Confidence Analysis

# The “Confidence Ball”

After  $t$  rounds, define our uncertainty region  $\text{BALL}_t$ : with center,  $\hat{\mu}_t$ , and shape,  $\Sigma_t$ , using the  $\lambda$ -regularized least squares solution:

$$\hat{\mu}_t = \underset{\mu}{\operatorname{argmin}} \sum_{\tau=0}^{t-1} (\mu \cdot x_{\tau} - r_{\tau})^2 + \lambda \|\mu\|_2^2$$

$$= \Sigma_t^{-1} \sum_{\tau=0}^{t-1} r_{\tau} x_{\tau},$$

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

$$\{ (x_{\tau}, r_{\tau}) \}_{\tau=0}^{t-1}$$

# The “Confidence Ball”

After  $t$  rounds, define our uncertainty region  $\text{BALL}_t$ : with center,  $\hat{\mu}_t$ , and shape,  $\Sigma_t$ , using the  $\lambda$ -regularized least squares solution:

$$\hat{\mu}_t = \underset{\mu}{\operatorname{argmin}} \sum_{\tau=0}^{t-1} (\mu \cdot x_\tau - r_\tau)^2 + \lambda \|\mu\|_2^2$$

$$= \Sigma_t^{-1} \sum_{\tau=0}^{t-1} r_\tau x_\tau,$$

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

with some  
prob  $\geq 1 - \delta$ ,  
 $\mu^* \in \text{BALL}_t$

Define the uncertainty region:

$$\|\hat{\mu}_t - \mu^*\|_{\Sigma_t}^2 \leq \beta_t$$

$$\text{BALL}_t = \left\{ \mu \mid (\hat{\mu}_t - \mu)^\top \Sigma_t (\hat{\mu}_t - \mu) \leq \beta_t \right\},$$

where  $\beta_t$  is a parameter of the algorithm.

# LinUCB (the algo)

- 1 Input:  $\lambda, \beta_t$
- 2 Initialize:  $\Sigma_0 = \lambda I, b_0 = 0$
- 3 For  $t = 0, 1, \dots$ 
  - 1 Compute  $\hat{\mu}_t = \Sigma_t^{-1} b_t$
  - 2 Execute

and observe the reward  $r_t$ .

$$\max_{\mu} \mu \cdot x + \lambda \left( (\mu - \hat{\mu}_t)^\top \Sigma_t (\mu - \hat{\mu}_t) \right)$$

solution  $\lambda (\mu - \hat{\mu}_t) = \Sigma_t^{-1} x$

no optimization

$$x_t = \operatorname{argmax}_{x \in D} x \cdot \hat{\mu}_t$$

fix  $x$

$$x_t = \operatorname{argmax}_{x \in D} \max_{\mu \in \text{BALL}_t} \mu \cdot x$$

$$\|\mu - \hat{\mu}_t\|_{\Sigma_t}^2 \leq \beta_t$$

$$(\mu - \hat{\mu}_t)^\top \Sigma_t (\mu - \hat{\mu}_t) \leq \beta_t$$

# LinUCB (the algo)

- 1 Input:  $\lambda, \beta_t$
- 2 Initialize:  $\Sigma_0 = \lambda I, b_0 = 0$
- 3 For  $t = 0, 1, \dots$ 
  - 1 Compute  $\hat{\mu}_t = \Sigma_t^{-1} b_t$
  - 2 Execute

$$\vec{\mu} = C \vec{\Sigma}_t^{-1} X$$

$$x_t = \operatorname{argmax}_{x \in D} \max_{\mu \in \text{BALL}_t} \mu \cdot x$$

and observe the reward  $r_t$ . Equivalently,

$$x_t \in \operatorname{argmax}_{x \in D} \left( \hat{\mu}_t \cdot x + \sqrt{\beta_t} \sqrt{x^\top \Sigma_t^{-1} x} \right)$$

# LinUCB (the algo)

- 1 Input:  $\lambda, \beta_t$
- 2 Initialize:  $\Sigma_0 = \lambda I, b_0 = 0$
- 3 For  $t = 0, 1, \dots$ 
  - 1 Compute  $\hat{\mu}_t = \Sigma_t^{-1} b_t$
  - 2 Execute

$$x_t = \operatorname{argmax}_{x \in D} \max_{\mu \in \text{BALL}_t} \mu \cdot x$$

and observe the reward  $r_t$ . Equivalently,

$$x_t \in \operatorname{argmax}_{x \in D} \left( \hat{\mu}_t \cdot x + \sqrt{\beta_t} \sqrt{x^\top \Sigma_t^{-1} x} \right)$$

- 3 Update:  $\Sigma_{t+1} = \Sigma_t + x_t x_t^\top$  and  $b_{t+1} = b_t + r_t x_t$ .

# LinUCB Regret Bound

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

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

# LinUCB Regret Bound

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

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

## Theorem

Suppose:  $|\mu^* \cdot x| \leq 1$  and  $\|x\| \leq B$  for all  $x \in D$ ; that the noise is  $\sigma^2$  sub-Gaussian; and that  $\|\mu^*\| \leq W$ . Set  $\lambda = \sigma^2/W^2$  and

$$\beta_t := c \left( \sigma^2 d \log \left( 1 + \frac{tB^2W^2}{d\sigma^2} \right) + \sigma^2 \log(1/\delta) \right)$$

With probability greater than  $1 - \delta$ , 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  $c, c'$  are absolute constants.

$M \propto B$   
 $\sqrt{KT}$   
 $\downarrow$   
 $d \sqrt{T}$

## 1 Linear Bandits

- Setting
- LinUCB
- An Optimal Regret Bound

## 2 Analysis

- Regret Analysis
- Confidence Analysis

In establishing the upper bounds there are two main propositions from which the upper bounds follow. The first is in showing that the confidence region is valid.

## Proposition

*(Confidence) Let  $\delta > 0$ . We have that for our choice of  $\beta_t$ ,*

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

*Equivalently, for all  $t \geq 0$ :*

$$(\hat{\mu}_t - \mu^*)^\top \Sigma_t (\hat{\mu}_t - \mu^*) \leq \beta_t$$

# Sum of Squares Regret Bound

Assuming the confidence event holds, the following controls on the growth of the regret.

## Proposition

*(Sum of Squares Regret Bound) Define:*

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

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

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

# Completing the Proof

**Proof of Theorem 1:** With the two previous Propositions, along with the Cauchy-Schwarz inequality, we have, with probability at least  $1 - \delta$ ,

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

The remainder of the proof follows from using our chosen value of  $\beta_T$  and algebraic manipulations.

## 1 Linear Bandits

- Setting
- LinUCB
- An Optimal Regret Bound

## 2 Analysis

- Regret Analysis
- Confidence Analysis

# “Width” of Confidence Ball

## Lemma

If  $\mu \in \text{BALL}_t$  and  $x \in D$ , then

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

# “Width” of Confidence Ball

## Lemma

If  $\mu \in \text{BALL}_t$  and  $x \in D$ , then

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

**Proof:** By Cauchy-Schwarz, we have:

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

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

# Instantaneous Regret Lemma

**Normalized width** at time  $t$  for decision  $x_t$ :  $w_t := \sqrt{x_t^\top \Sigma_t^{-1} x_t}$

## Lemma

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

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

# Instantaneous Regret Lemma

**Normalized width** at time  $t$  for decision  $x_t$ :  $w_t := \sqrt{x_t^\top \Sigma_t^{-1} x_t}$

## Lemma

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

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

**Proof:** Let  $\tilde{\mu} \in \arg\max_{\mu \in \text{BALL}_t} \mu^\top x_t$ . By the choice of  $x_t$  and the assumption  $\mu^* \in \text{BALL}_t$ , we have **optimism**:

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

# Instantaneous Regret Lemma

**Normalized width** at time  $t$  for decision  $x_t$ :  $w_t := \sqrt{x_t^\top \Sigma_t^{-1} x_t}$

## Lemma

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

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

**Proof:** Let  $\tilde{\mu} \in \arg\max_{\mu \in \text{BALL}_t} \mu^\top x_t$ . By the choice of  $x_t$  and the assumption  $\mu^* \in \text{BALL}_t$ , we have **optimism**:

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

We can decompose the regret by adding and subtracting  $\hat{\mu}_t$ :

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

# Instantaneous Regret Lemma

**Normalized width** at time  $t$  for decision  $x_t$ :  $w_t := \sqrt{x_t^\top \Sigma_t^{-1} x_t}$

## Lemma

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

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

**Proof:** Let  $\tilde{\mu} \in \arg\max_{\mu \in \text{BALL}_t} \mu^\top x_t$ . By the choice of  $x_t$  and the assumption  $\mu^* \in \text{BALL}_t$ , we have **optimism**:

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

We can decompose the regret by adding and subtracting  $\hat{\mu}_t$ :

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

*Note: Since rewards  $r_t \in [-1, 1]$ ,  $\text{regret}_t$  is trivially bounded by 2, yielding the “min”.*

# Geometric Argument: Part 1

The next two lemmas give us a 'geometric' potential function argument, where we can bound the sum of widths independently of the choices made by the algorithm.

# Geometric Argument: Part 1

The next two lemmas give us a 'geometric' potential function argument, where we can bound the sum of widths independently of the choices made by the algorithm.

## Lemma

*We have:*

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

# Geometric Argument: Part 1

The next two lemmas give us a 'geometric' potential function argument, where we can bound the sum of widths independently of the choices made by the algorithm.

## Lemma

*We have:*

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

**Proof:** By the definition of  $\Sigma_{t+1}$ , 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  $v_t^\top v_t = w_t^2$ .

# Geometric Argument: Part 2

## Lemma

For any sequence where  $\|x_t\|_2 \leq B$ , we have:

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

# Geometric Argument: Part 2

## Lemma

For any sequence where  $\|x_t\|_2 \leq B$ , we have:

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

**Proof:** Let  $\sigma_1, \dots, \sigma_d$  be the eigenvalues of  $\sum_{t=0}^{T-1} x_t x_t^\top$ . We have:

$$\sum_{i=1}^d \sigma_i = \text{Tr} \left( \sum_{t=0}^{T-1} x_t x_t^\top \right) = \sum_{t=0}^{T-1} \|x_t\|_2^2 \leq TB^2$$

# Geometric Argument: Part 2

## Lemma

For any sequence where  $\|x_t\|_2 \leq B$ , we have:

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

**Proof:** Let  $\sigma_1, \dots, \sigma_d$  be the eigenvalues of  $\sum_{t=0}^{T-1} x_t x_t^\top$ . We have:

$$\sum_{i=1}^d \sigma_i = \text{Tr} \left( \sum_{t=0}^{T-1} x_t x_t^\top \right) = \sum_{t=0}^{T-1} \|x_t\|_2^2 \leq TB^2$$

Next, apply the AM-GM inequality to the log-determinant:

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

# Proving “Sum of Squares Regret”

## Proof of Proposition 3:

Assume  $\mu^* \in \text{BALL}_t$  for all  $t$ . We have:

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

# Proving “Sum of Squares Regret”

## Proof of Proposition 3:

Assume  $\mu^* \in \text{BALL}_t$  for all  $t$ . We have:

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

# Proving “Sum of Squares Regret”

## Proof of Proposition 3:

Assume  $\mu^* \in \text{BALL}_t$  for all  $t$ . We have:

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

# Proving “Sum of Squares Regret”

## Proof of Proposition 3:

Assume  $\mu^* \in \text{BALL}_t$  for all  $t$ . We have:

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

where the first inequality follows from Lemma 5; the second from that  $\beta_t$  is an increasing function of  $t$ ; the third uses that for  $0 \leq y \leq 1$ ,  $\ln(1 + y) \geq y/2$ ; and the final two follow from Lemmas 6 and 7.

## 1 Linear Bandits

- Setting
- LinUCB
- An Optimal Regret Bound

## 2 Analysis

- Regret Analysis
- Confidence Analysis

## Confidence [Proof of Proposition 2]

Since  $r_\tau = x_\tau \cdot \mu^* + \eta_\tau$ , we decompose the error:

$$\begin{aligned}\hat{\mu}_t - \mu^* &= \Sigma_t^{-1} \sum_{\tau=0}^{t-1} x_\tau (x_\tau^\top \mu^* + \eta_\tau) - \mu^* \\ &= \Sigma_t^{-1} (\Sigma_t - \lambda I) \mu^* - \mu^* + \Sigma_t^{-1} \sum_{\tau=0}^{t-1} \eta_\tau x_\tau \\ &= -\lambda \Sigma_t^{-1} \mu^* + \Sigma_t^{-1} \sum_{\tau=0}^{t-1} \eta_\tau x_\tau\end{aligned}$$

# Confidence [Proof of Proposition 2]

Since  $r_\tau = x_\tau \cdot \mu^* + \eta_\tau$ , we decompose the error:

$$\begin{aligned}\widehat{\mu}_t - \mu^* &= \Sigma_t^{-1} \sum_{\tau=0}^{t-1} x_\tau (x_\tau^\top \mu^* + \eta_\tau) - \mu^* \\ &= \Sigma_t^{-1} (\Sigma_t - \lambda I) \mu^* - \mu^* + \Sigma_t^{-1} \sum_{\tau=0}^{t-1} \eta_\tau x_\tau \\ &= -\lambda \Sigma_t^{-1} \mu^* + \Sigma_t^{-1} \sum_{\tau=0}^{t-1} \eta_\tau x_\tau\end{aligned}$$

Using the norm  $\|z\|_{\Sigma_t} := \sqrt{z^\top \Sigma_t z}$ , the triangle inequality gives:

$$\begin{aligned}\|\widehat{\mu}_t - \mu^*\|_{\Sigma_t} &\leq \|\lambda \Sigma_t^{-1} \mu^*\|_{\Sigma_t} + \left\| \Sigma_t^{-1} \sum_{\tau=0}^{t-1} \eta_\tau x_\tau \right\|_{\Sigma_t} \\ &\leq \sqrt{\lambda} \|\mu^*\|_2 + \underbrace{\left\| \sum_{\tau=0}^{t-1} \eta_\tau x_\tau \right\|_{\Sigma_t^{-1}}}_{??}\end{aligned}$$

# Self-Normalizing Sum

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

Suppose  $\{\varepsilon_i\}_{i=1}^{\infty}$  is a real-valued, mean-zero,  $\sigma$ -sub-Gaussian stochastic process. Let  $\{X_i\}_{i=1}^{\infty}$  be an  $\mathbb{R}^d$ -valued 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  $t \geq 1$ :

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

# Completing the Confidence Bound

For all  $t$ , with probability at least  $1 - \delta$ , we plug the lemma back into our triangle inequality bound:

$$\begin{aligned}\|\hat{\mu}_t - \mu^*\|_{\Sigma_t} &\leq \sqrt{\lambda} \|\mu^*\|_2 + \left\| \sum_{\tau=0}^{t-1} \eta_\tau x_\tau \right\|_{\Sigma_t^{-1}} \\ &\leq \sqrt{\lambda} \|\mu^*\|_2 + \sqrt{\sigma^2 \log \left( \frac{\det(\Sigma_t)}{\det(\Sigma_0) \delta^2} \right)}.\end{aligned}$$

# Completing the Confidence Bound

For all  $t$ , with probability at least  $1 - \delta$ , we plug the lemma back into our triangle inequality bound:

$$\begin{aligned}\|\hat{\mu}_t - \mu^*\|_{\Sigma_t} &\leq \sqrt{\lambda} \|\mu^*\|_2 + \left\| \sum_{\tau=0}^{t-1} \eta_\tau x_\tau \right\|_{\Sigma_t^{-1}} \\ &\leq \sqrt{\lambda} \|\mu^*\|_2 + \sqrt{\sigma^2 \log \left( \frac{\det(\Sigma_t)}{\det(\Sigma_0) \delta^2} \right)}.\end{aligned}$$

Substitute our log-determinant bound (Lemma 7) into the term above:

$$\log \left( \frac{\det(\Sigma_t)}{\det(\Sigma_0)} \right) \leq d \log \left( 1 + \frac{tB^2}{d\lambda} \right)$$

This matches our definition of  $\sqrt{\beta_t}$ , concluding that  $\mu^* \in \text{BALL}_t$ .