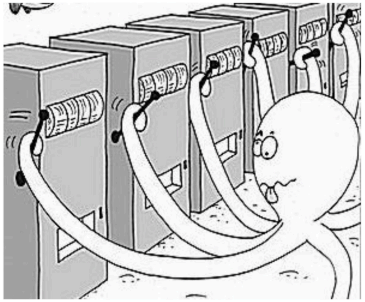
# Linear Bandits

Sham M. Kakade and Wen Sun

# Outline

# Intro to MAB



**Setting:**

We have K many arms: $a_1, \ldots, a_K$

Each arm has a unknown reward distribution, i.e., $\nu_i \in \Delta([0,1])$,
w/ mean $\mu_i = \mathbb{E}_{r \sim \nu_i}[r]$

**Example:** $a_i$ has a Bernoulli distribution $\nu_i$ w/ mean $\mu_i := p$:

Every time we pull arm $a_i$, we observe an i.i.d reward $r = \begin{cases} 1 & \text{w/ prob } p \\ 0 & \text{w/ prob } 1-p \end{cases}$

# Intro to MAB

**More formally, we have the following learning objective:**

$$\mu^\star = \max_{i \in [K]} \mu_i$$

$$\text{Regret}_T = T\mu^\star - \sum_{t=0}^{T-1} \mu_{I_t}$$

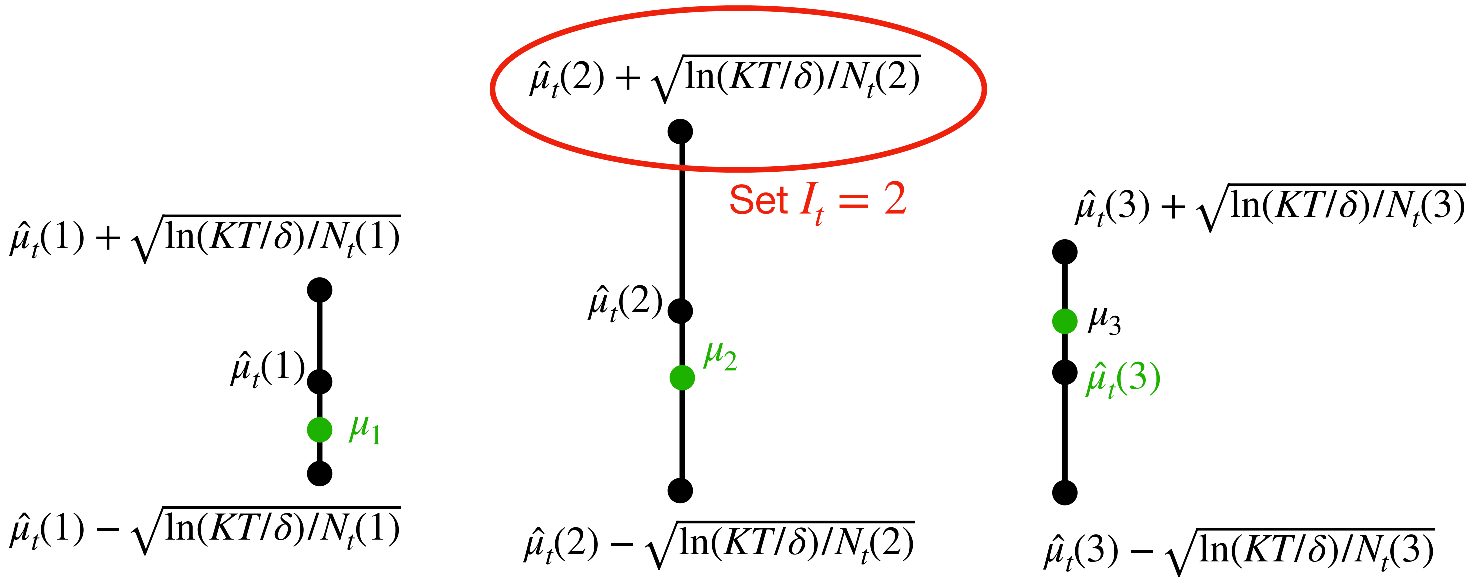
Total expected reward if we pulled best arm over T rounds

Total expected reward of the arms we pulled over T rounds

Goal: no-regret, i.e., $\text{Regret}_T / T \to 0$, as $T \to \infty$

# UCB: Optimism in the face of Uncertainty

Given the confidence interval, we pick arm that has the **highest Upper-Conf-Bound:**



# UCB Regret:

[Theorem (informal)] With high probability, UCB has the following regret:

$$\text{Regret}_T = \widetilde{O}\left(\sqrt{KT}\right)$$

$$\frac{\text{Regret}_T}{T} = \frac{\widetilde{O}\left(\sqrt{KT}\right)}{T} \approx \frac{\sqrt{K}}{\sqrt{T}}$$

# Generalization in RL

- (distribution free) Agnostic learning is not possible in RL:
  we showed that to get $O(\log|\Pi|)$ sample complexity we need either:
  - $\mathrm{poly}(|\mathcal{S}|)$ samples OR
  - $\exp(H)$ samples.

  in order to learn the best policy in some policy class.
- upshot: we need stronger assumptions for RL analysis.

# Outline

# Handling Large Actions Spaces

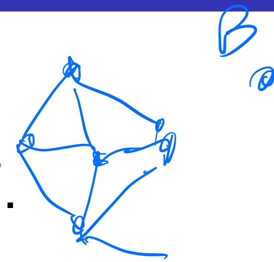
- On each round, we must choose a decision $x_t \in D \subset R^d$.

# Handling Large Actions 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

$$\mathbb{E}[r_t | x_t = x] = \mu^\star \cdot x \in [-1, 1],$$

# Handling Large Actions 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

$$\mathbb{E}[r_t | x_t = x] = \mu^\star \cdot x \in [-1, 1],$$

"paths on graph"

  - so the the conditional expectation of $r_t$ is linear)
  - Also, we have the *noise sequence*,

$\mu^* \leftarrow$ edge cost vector.

$$\eta_t = r_t - \mu^\star \cdot x_t$$

is i.i.d noise.

see cumulative path length.

model due to Abe & Long '99

$[x]_j = 1\{$ edge j taken.

# Our Objective

"Data set" at time $t$ $\{(\vec{x}_0, r_0), \ldots (\vec{x}_{t-1}, r_{t-1})\}$

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

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

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

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

# Outline

1. Recap
2. Linear Bandits
   - Setting
   - LinUCB
   - An Optimal Regret Bound
3. Analysis
   - Regret Analysis
   - Confidence Analysis

# The "Confidence Ball"

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

$$\widehat{\mu}_t = \arg\min_{\mu} \sum_{\tau=0}^{t-1} \|\mu \cdot x_\tau - r_\tau\|_2^2 + \lambda\|\mu\|_2^2$$

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

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

# The "Confidence Ball"

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

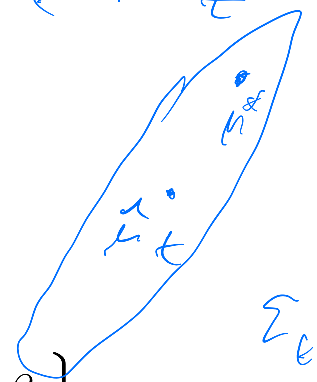
$$\widehat{\mu}_t = \arg\min_{\mu} \sum_{\tau=0}^{t-1} \|\mu \cdot x_\tau - r_\tau\|_2^2 + \lambda\|\mu\|_2^2$$

hope

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

$\mu^* \in \text{Ball}_t$

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




Define the uncertainty region:

$\Sigma_t^{-1}$

$$\textsc{Ball}_t = \left\{\mu \mid (\widehat{\mu}_t - \mu)^\top \Sigma_t^{-1} (\widehat{\mu}_t - \mu) \le \beta_t\right\},$$

shape of the ellipse

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

# LinUCB (the algo)

1. Input: $\lambda$, $\beta_t$
2. For $t = 0, 1, \ldots$
   1. Execute
   $$x_t = \text{argmax}_{x \in D} \max_{\mu \in \textsc{Ball}_t} \mu \cdot x$$
   and observe the reward $r_t$.
   2. Update $\textsc{Ball}_{t+1}$.

# LinUCB Regret Bound

Sublinear regret: $R_T \leq O^\star(d\sqrt{T})$ MAB $\longleftrightarrow$ $O\sqrt{|D|T}$

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

# LinUCB Regret Bound

Sublinear regret: $R_T \leq O^\star(d\sqrt{T})$ does $\hat{\mu}_t \to \mu$ ??

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

## Theorem

*Suppose: bounded noise $|\eta_t| \leq \sigma$, that $\|\mu^\star\| \leq W$, and that $\|x\| \leq B$ for all $x \in D$. Set $\lambda = \sigma^2/W^2$ and*

$$\beta_t := \sigma^2\left(2 + 4d\log\left(1 + \frac{tB^2W^2}{d}\right) + 8\log(4/\delta)\right).$$

*With probability greater than $1-\delta$, that for all ~~$t \geq 0$~~,* $T \geq 0$

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

*where $c$ is an absolute constant.*

due to Dani, Hayes, K. '09

# Outline

1. Recap
2. Linear Bandits
   - Setting
   - LinUCB
   - An Optimal Regret Bound
3. Analysis
   - Regret Analysis
   - Confidence Analysis

# Confidence

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*

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

$\forall T$

$$E[R_T] \leq d\sqrt{T}$$

# 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:*

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

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

$O(\beta_T d \lg T)$

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

# Completing the Proof

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

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

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

# Outline

# "Width" of Confidence Ball

## Lemma

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

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

# “Width” of Confidence Ball

**Lemma**

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

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

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

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

$$\le \|\Sigma_t^{1/2}(\mu - \widehat{\mu}_t)\|_2 \|\Sigma_t^{-1/2} x\|_2 = \|\Sigma_t^{1/2}(\mu - \widehat{\mu}_t)\| \sqrt{x^\top \Sigma_t^{-1} x} \le \sqrt{\beta_t x^\top \Sigma_t^{-1} x}$$

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

# Instantaneous Regret Lemma

Define


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

which is the “normalized width” at time $t$ in the direction of our decision.

# Instantaneous Regret Lemma

Define

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

which is the "normalized width" at time $t$ in the direction of our decision.

**Lemma**

*Fix $t \le T$. If $\mu^\star \in \textsc{Ball}_t$, then*

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

Assume all the rewards are bounded by 1

$$reg_t \le 2\sqrt{\beta_T}\, w_t$$

# Instantaneous Regret Lemma

Define

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

which is the "normalized width" at time $t$ in the direction of our decision.

## Lemma

*Fix $t \leq T$. If $\mu^\star \in \text{BALL}_t$, then* True for LinUCB algo

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

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

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

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

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

# Geometric Argument: Part 1

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

# Geometric Argument: Part 1

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

## Lemma

*We have:*

$\Sigma_t = \lambda I + \sum_{\tau < t} x_\tau x_\tau^\top$

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

# Geometric Argument: Part 1

The next two lemmas give us 'geometric' potential function argument, where 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

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

$\det(I + aa^\top) = 1 + \|a\|^2$

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

# Geometric Argument: Part 2

## Lemma

*For any sequence* $x_0, \ldots x_{T-1}$ *such that, for* $t < T$, $\|x_t\|_2 \le B$, *we have:*

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

# Geometric Argument: Part 2

**Lemma**

*For any sequence $x_0, \ldots x_{T-1}$ such that, for $t < T$, $\|x_t\|_2 \leq B$, we have:*

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

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

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

Using the AM-GM inequality,

$U \mathrm{Diag}(\sigma_1 \ldots \sigma_d) V^T$

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

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

# Proving "sum of squares regret" Proposition

$\mathrm{reg}_t \leq 2\sqrt{\beta_t}\, \min(w_t, 1)$

**Proof:**[Proof of Proposition 3] Assume $\mu^\star \in \textsc{Ball}_t$ for all $t$. We have:

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

$8 \qquad\qquad 8$

$$\leq 4\beta_T \sum_{t=0}^{T-1} \ln(1 + w_t^2) \leq 4\beta_T \log\left( \det \Sigma_{T-1} / \det \Sigma_0 \right)$$

$8$

$$= 4\beta_T d \log\left(1 + \frac{TB^2}{d\lambda}\right)$$

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

# Outline

# Confidence [Proof of Proposition 2]

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

$$\widehat{\mu}_t - \mu^\star = \Sigma_t^{-1} \sum_{\tau=0}^{t-1} r_\tau x_\tau - \mu^\star = \Sigma_t^{-1} \sum_{\tau=0}^{t-1} x_\tau (x_\tau \cdot \mu^\star + \eta_\tau) - \mu^\star$$

$$= \Sigma_t^{-1} \left( \sum_{\tau=0}^{t-1} x_\tau (x_\tau)^\top \right) \mu^\star - \mu^\star + \Sigma_t^{-1} \sum_{\tau=0}^{t-1} \eta_\tau x_\tau$$

$$= \lambda \Sigma_t^{-1} \mu^\star + \Sigma_t^{-1} \sum_{\tau=0}^{t-1} \eta_\tau x_\tau$$

# Confidence [Proof of Proposition 2]

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

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

By the triangle inequality,

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

How can we bound "??" To be continued... ■

# Self-Normalizing Sum

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

*(Abassi et. al '11) Suppose $\{\varepsilon_i\}_{i=1}^{\infty}$ are mean zero random variables (can be generalized to martingales), and $\varepsilon_i$ is bounded by $\sigma$. Let $\{X_i\}_{i=1}^{\infty}$ be a stochastic process. Define $\Sigma_t = \Sigma_0 + \sum_{i=1}^{t} X_i X_i^\top$. With probability at least $1-\delta$, we have for all $t \geq 1$:*

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

# Continued... [Proof of Proposition 2]

**Proof:**

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

We seek to lower bound $\Pr(\forall t,\ \mu^\star \in \textsc{Ball}_t)$. Assign failure probability $\delta_t = (3/\pi^2)/t^2$ for the $t$-th event, which gives us:

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

This along with Lemma 7 completes the proof. ■