Linear Bandits

CS 6789: Foundations of Reinforcement Learning

Recap on MAB

Setting:

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

Recap on 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 $\mu_i = \mathbb{E}_{r \sim \nu_i}[r]$

Regret

More formally, we have the following learning objective:

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

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

Regret

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

Regret

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

Regret

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 $T \to \infty$

Outline for Today:

1. Linear Bandit Setting

2. Algorithm: LinUCB

3. Regret analysis of LinUCB

Linear Bandit Setting

We have an action set $D \subset \mathbb{R}^d$

Linear Bandit Setting

We have an action set $D \subset \mathbb{R}^d$

Expected reward of each action $x \in D$ is linear:

$$\mathbb{E}[r \,|\, x] = (\mu^\star)^\top x$$

Linear Bandit Setting

We have an action set $D \subset \mathbb{R}^d$

Expected reward of each action $x \in D$ is linear:

$$\mathbb{E}[r \mid x] = (\mu^\star)^\top x$$

Every time we pick an action $x \in D$, we observe a noisy reward

$$r = \mu^\star \cdot x + \eta$$

Linear Bandit Setting

We have an action set $D \subset \mathbb{R}^d$

Expected reward of each action $x \in D$ is linear:

$$\mathbb{E}[r \,|\, x] = (\mu^\star)^\top x$$

Every time we pick an action $x \in D$, we observe a noisy reward

$$r = \mu^\star \cdot x + \eta$$

Zero mean i.i.d noise

Learning protocol and goal:

For t = 1 to T:

Learning protocol and goal:

```
For t = 1 to T:
    Leaner selects $x_t \in D$ (based on history)
```

Learning protocol and goal:

```
For t = 1 to T:
    Leaner selects $x_t \in D$ (based on history)
    Learner observes a noisy reward, i.e., $r_t = \mu^\star \cdot x_t + \eta_t$
```

Learning protocol and goal:

```
For t = 1 to T:
    Leaner selects $x_t \in D$ (based on history)
    Learner observes a noisy reward, i.e., $r_t = \mu^\star \cdot x_t + \eta_t$
```

Goal: minimize regret

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

Learning protocol and goal:

For t = 1 to T:

Leaner selects $x_t \in D$ (based on history)

Learner observes a noisy reward, i.e., $r_t = \mu^\star \cdot x_t + \eta_t$

Goal: minimize regret

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

$$x^\star = \arg\max_{x \in D} \mu^\star \cdot x$$

Outline for Today:

1. Linear Bandit Setting

2. Algorithm: LinUCB

3. Regret analysis of LinUCB

LinUCB algorithm

Overall idea:

Ridge linear regression for learning $\mu^\star$ + design exploration bonus

LinUCB algorithm

In iteration t:

1. Perform Ridge LR on data $\{x_i, r_i\}_{i=0}^{t-1}$:

   Set $\hat{\mu}_t := \arg\min_{\mu} \sum_{i=0}^{t-1} (\mu^\top x_i - r_i)^2 + \lambda \|\mu\|_2^2$

LinUCB algorithm

In iteration t:

1. Perform Ridge LR on data $\{x_i, r_i\}_{i=0}^{t-1}$:

Set $\hat{\mu}_t := \arg\min_{\mu} \sum_{i=0}^{t-1} (\mu^\top x_i - r_i)^2 + \lambda \|\mu\|_2^2$

2: Set exploration bonus: $b_t(x) = \beta\sqrt{x^\top \Sigma_t^{-1} x}$

LinUCB algorithm

In iteration t:

1. Perform Ridge LR on data $\{x_i, r_i\}_{i=0}^{t-1}$:

Set $\hat{\mu}_t := \arg\min_{\mu} \sum_{i=0}^{t-1} (\mu^\top x_i - r_i)^2 + \lambda \|\mu\|_2^2$

2: Set exploration bonus: $b_t(x) = \beta \sqrt{x^\top \Sigma_t^{-1} x}$

3: Play optimistically, i.e., $x_t = \arg\max_{x \in D} \hat{\mu}_t^\top x_t + b_t(x)$

Outline for Today:

1. Linear Bandit Setting

2. Algorithm: LinUCB

3. Regret analysis of LinUCB

Analysis of Ridge Linear Regression

Recall $\hat{\mu}_t := \arg\min_{\mu} \sum_{i=0}^{t-1} (\mu^\top x_i - r_i)^2 + \lambda \|\mu\|_2^2$

Analysis of Ridge Linear Regression

Recall $\hat{\mu}_t := \arg\min_{\mu} \sum_{i=0}^{t-1} (\mu^\top x_i - r_i)^2 + \lambda \|\mu\|_2^2$

$$\hat{\mu}_t = \Sigma_t^{-1} \sum_{i=0}^{t-1} x_i r_i$$

Analysis of Ridge Linear Regression

Recall $\hat{\mu}_t := \arg\min_{\mu} \sum_{i=0}^{t-1} (\mu^\top x_i - r_i)^2 + \lambda \|\mu\|_2^2$

$$\hat{\mu}_t = \Sigma_t^{-1} \sum_{i=0}^{t-1} x_i r_i$$

$$= \Sigma_t^{-1} \sum_{i=0}^{t-1} x_i (x_i^\top \mu^\star + \eta_i)$$

Analysis of Ridge Linear Regression

Recall $\hat{\mu}_t := \arg\min_{\mu} \sum_{i=0}^{t-1} (\mu^\top x_i - r_i)^2 + \lambda \|\mu\|_2^2$

$$\hat{\mu}_t = \Sigma_t^{-1} \sum_{i=0}^{t-1} x_i r_i$$

$$= \Sigma_t^{-1} \sum_{i=0}^{t-1} x_i (x_i^\top \mu^\star + \eta_i) \; = \Sigma_t^{-1} (\Sigma_t - \lambda I) \mu^\star + \Sigma_t^{-1} \sum_{i=0}^{t-1} x_i \eta_i$$

Analysis of Ridge Linear Regression

Recall $\hat{\mu}_t := \arg\min_{\mu} \sum_{i=0}^{t-1} (\mu^\top x_i - r_i)^2 + \lambda \|\mu\|_2^2$

$$\hat{\mu}_t = \Sigma_t^{-1} \sum_{i=0}^{t-1} x_i r_i$$

$$= \Sigma_t^{-1} \sum_{i=0}^{t-1} x_i (x_i^\top \mu^\star + \eta_i) \; = \Sigma_t^{-1} (\Sigma_t - \lambda I) \mu^\star + \Sigma_t^{-1} \sum_{i=0}^{t-1} x_i \eta_i$$

$$= \mu^\star - \lambda \Sigma_t^{-1} \mu^\star + \Sigma_t^{-1} \sum_{i=0}^{t-1} x_i \eta_i$$

Analysis of Ridge Linear Regression

Recall $\hat{\mu}_t := \arg\min_{\mu} \sum_{i=0}^{t-1} (\mu^\top x_i - r_i)^2 + \lambda \|\mu\|_2^2$

$$\hat{\mu}_t = \Sigma_t^{-1} \sum_{i=0}^{t-1} x_i r_i$$

$$= \Sigma_t^{-1} \sum_{i=0}^{t-1} x_i (x_i^\top \mu^\star + \eta_i) \ = \Sigma_t^{-1} (\Sigma_t - \lambda I) \mu^\star + \Sigma_t^{-1} \sum_{i=0}^{t-1} x_i \eta_i$$

$$= \mu^\star - \lambda \Sigma_t^{-1} \mu^\star + \Sigma_t^{-1} \sum_{i=0}^{t-1} x_i \eta_i$$

$$\hat{\mu}_t - \mu^\star = -\lambda \Sigma_t^{-1} \mu^\star + \Sigma_t^{-1} \sum_{i=0}^{t-1} x_i \eta_i$$

Analysis of Ridge Linear Regression

$$\hat{\mu}_t - \mu^\star = -\lambda \Sigma_t^{-1} \mu^\star + \Sigma_t^{-1} \sum_{i=0}^{t-1} x_i \eta_i$$

Let us look at the training error:

$$\sqrt{(\hat{\mu}_t - \mu^\star)^\top \Sigma_t (\hat{\mu}_t - \mu^\star)}$$

Analysis of Ridge Linear Regression

$$\hat{\mu}_t - \mu^\star = -\lambda \Sigma_t^{-1} \mu^\star + \Sigma_t^{-1} \sum_{i=0}^{t-1} x_i \eta_i$$

Let us look at the training error:

$$\sqrt{(\hat{\mu}_t - \mu^\star)^\top \Sigma_t (\hat{\mu}_t - \mu^\star)} \quad \leq \left\| \lambda \Sigma_t^{-1/2} \mu^\star \right\| + \left\| \Sigma_t^{-1/2} \sum_{i=0}^{t-1} \eta_i x_i \right\|$$

Analysis of Ridge Linear Regression

$$\hat{\mu}_t - \mu^\star = -\lambda \Sigma_t^{-1} \mu^\star + \Sigma_t^{-1} \sum_{i=0}^{t-1} x_i \eta_i$$

Let us look at the training error:

$$\sqrt{(\hat{\mu}_t - \mu^\star)^\top \Sigma_t (\hat{\mu}_t - \mu^\star)} \quad \leq \left\| \lambda \Sigma_t^{-1/2} \mu^\star \right\| + \left\| \Sigma_t^{-1/2} \sum_{i=0}^{t-1} \eta_i x_i \right\|$$

$$\leq \sqrt{\lambda} \|\mu^\star\| + ???$$

Analysis of Ridge Linear Regression

$$\hat{\mu}_t - \mu^\star = -\lambda \Sigma_t^{-1} \mu^\star + \Sigma_t^{-1} \sum_{i=0}^{t-1} x_i \eta_i$$

Let us look at the training error:

$$\sqrt{(\hat{\mu}_t - \mu^\star)^\top \Sigma_t (\hat{\mu}_t - \mu^\star)} \leq \left\| \lambda \Sigma_t^{-1/2} \mu^\star \right\| + \boxed{\left\| \Sigma_t^{-1/2} \sum_{i=0}^{t-1} \eta_i x_i \right\|}$$

$$\leq \sqrt{\lambda} \|\mu^\star\| + ???$$

Self-normalized Martingale bound

Self-normalized Bound for Vector-valued Martingales

Suppose $\{\eta_i\}_{i=0}^{\infty}$ are mean zero random variables, and $|\eta_i| \leq \sigma$;

Let $\{x_i\}_{i=0}^{\infty}$ be any sequence of random vectors with $\|x_i\| \leq 1$, then w/ prob $1 - \delta$, for all $t \geq 1$,

$$\left\| \Sigma_t^{-1/2} \sum_{i=0}^{t-1} x_i \eta_i \right\|^2 \leq \sigma^2 d \cdot \left(\ln\left(\frac{t}{\lambda} + 1\right) + \ln(1/\delta) \right)$$

Analysis of Ridge Linear Regression (Continue)

$$\hat{\mu}_t - \mu^\star = -\lambda\Sigma_t^{-1}\mu^\star + \Sigma_t^{-1}\sum_{i=0}^{t-1} x_i\eta_i$$

Let us look at the training error:

$$\sqrt{(\hat{\mu}_t - \mu^\star)^\top \Sigma_t (\hat{\mu}_t - \mu^\star)} \quad \leq \left\| \lambda\Sigma_t^{-1/2}\mu^\star \right\| + \left\| \Sigma_t^{-1/2}\sum_{i=0}^{t-1} \eta_i x_i \right\|$$

$$\lesssim \sqrt{\lambda} + \sigma\sqrt{d \cdot \ln(T/(\lambda\delta))}$$

Summary for Ridge Linear Regression

$$\hat{\mu}_t - \mu^\star = -\lambda \Sigma_t^{-1} \mu^\star + \Sigma_t^{-1} \sum_{i=0}^{t-1} x_i \eta_i$$

$$\sqrt{(\hat{\mu}_t - \mu^\star)^\top \Sigma_t (\hat{\mu}_t - \mu^\star)} \lesssim \sqrt{\lambda} + \sigma \sqrt{d \ln(T/(\lambda\delta))}$$

Optimism

$$\sqrt{(\hat{\mu}_t - \mu^\star)^\top \Sigma_t (\hat{\mu}_t - \mu^\star)} \lesssim \sqrt{\lambda} + \sqrt{\sigma^2 d \ln(T/(\lambda\delta))}$$

Let's construct uncertainty quantification for each action $x \in D$

$|\hat{\mu}_t \cdot x - \mu^\star \cdot x|$

Optimism

$$\sqrt{(\hat{\mu}_t - \mu^\star)^\top \Sigma_t (\hat{\mu}_t - \mu^\star)} \lesssim \sqrt{\lambda} + \sqrt{\sigma^2 d \ln(T/(\lambda\delta))}$$

Let's construct uncertainty quantification for each action $x \in D$

$$|\hat{\mu}_t \cdot x - \mu^\star \cdot x| \quad \leq \|\hat{\mu}_t - \mu^\star\|_{\Sigma_t} \cdot \|x\|_{\Sigma_t^{-1}}$$

Optimism

$$\sqrt{(\hat{\mu}_t - \mu^\star)^\top \Sigma_t (\hat{\mu}_t - \mu^\star)} \lesssim \sqrt{\lambda} + \sqrt{\sigma^2 d \ln(T/(\lambda\delta))}$$

Let's construct uncertainty quantification for each action $x \in D$

$$|\hat{\mu}_t \cdot x - \mu^\star \cdot x| \leq \|\hat{\mu}_t - \mu^\star\|_{\Sigma_t} \cdot \|x\|_{\Sigma_t^{-1}}$$

$$\lesssim \left(\sqrt{\lambda} + \sigma\sqrt{d \ln(T/(\lambda\delta))}\right) \cdot \| x \|_{\Sigma_t^{-1}}$$

Optimism

$$\sqrt{(\hat{\mu}_t - \mu^\star)^\top \Sigma_t (\hat{\mu}_t - \mu^\star)} \lesssim \sqrt{\lambda} + \sqrt{\sigma^2 d \ln(T/(\lambda\delta))}$$

Let's construct uncertainty quantification for each action $x \in D$

$$|\hat{\mu}_t \cdot x - \mu^\star \cdot x| \leq \|\hat{\mu}_t - \mu^\star\|_{\Sigma_t} \cdot \|x\|_{\Sigma_t^{-1}}$$

$$\lesssim \left(\sqrt{\lambda} + \sigma\sqrt{d \ln(T/(\lambda\delta))}\right) \cdot \| x \|_{\Sigma_t^{-1}}$$

$$b_t(x) := \beta \cdot \|x\|_{\Sigma_t^{-1}}$$

Optimism

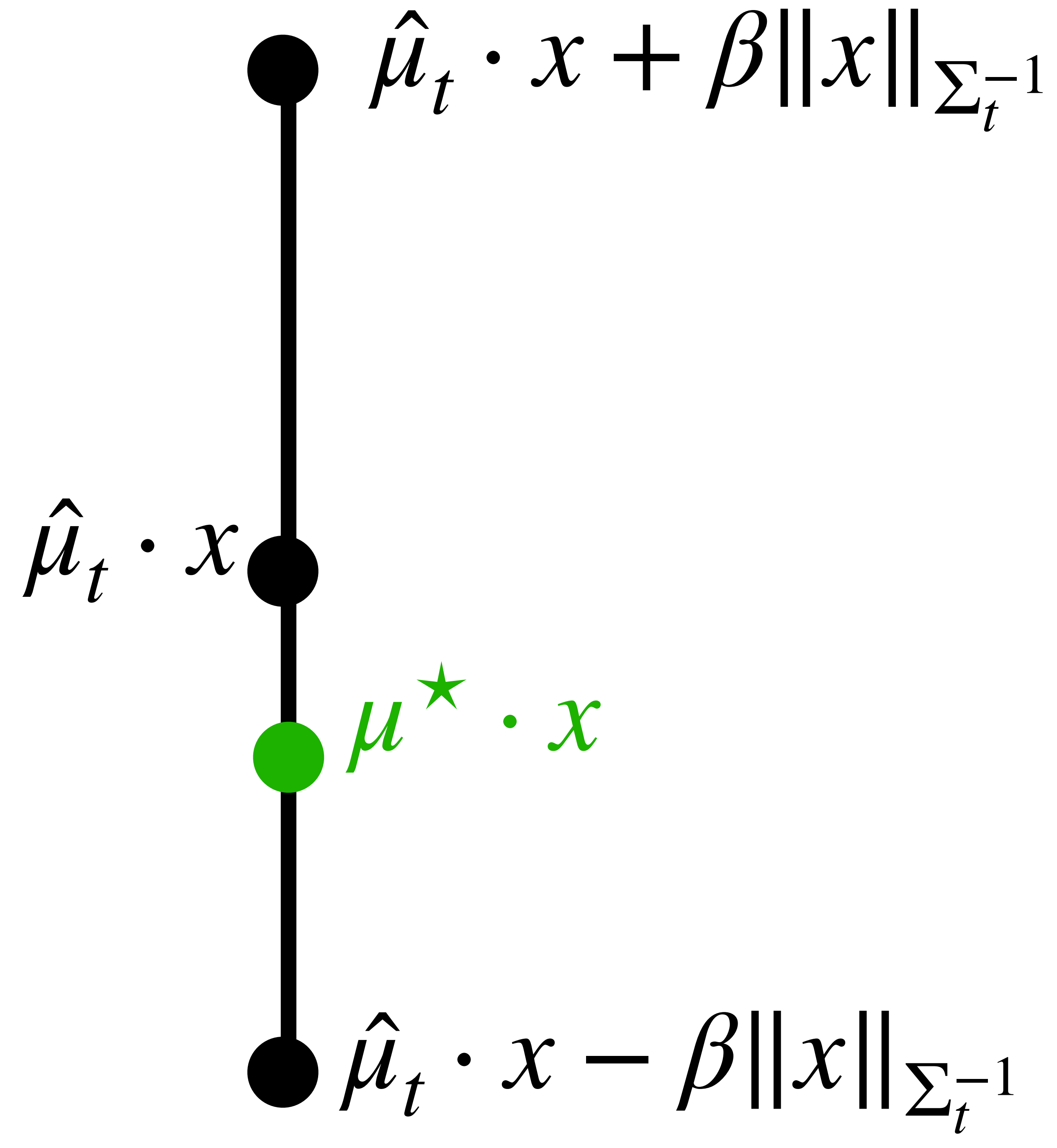
$$\sqrt{(\hat{\mu}_t - \mu^\star)^\top \Sigma_t (\hat{\mu}_t - \mu^\star)} \lesssim \sqrt{\lambda} + \sqrt{\sigma^2 d \ln(T/(\lambda\delta))}$$

Let's construct uncertainty quantification for each action $x \in D$

$$|\hat{\mu}_t \cdot x - \mu^\star \cdot x| \le \|\hat{\mu}_t - \mu^\star\|_{\Sigma_t} \cdot \|x\|_{\Sigma_t^{-1}}$$

$$\boxed{\lesssim \left(\sqrt{\lambda} + \sigma\sqrt{d \ln(T/(\lambda\delta))}\right) \cdot \| x \|_{\Sigma_t^{-1}}}$$

$$b_t(x) := \beta \cdot \|x\|_{\Sigma_t^{-1}}$$

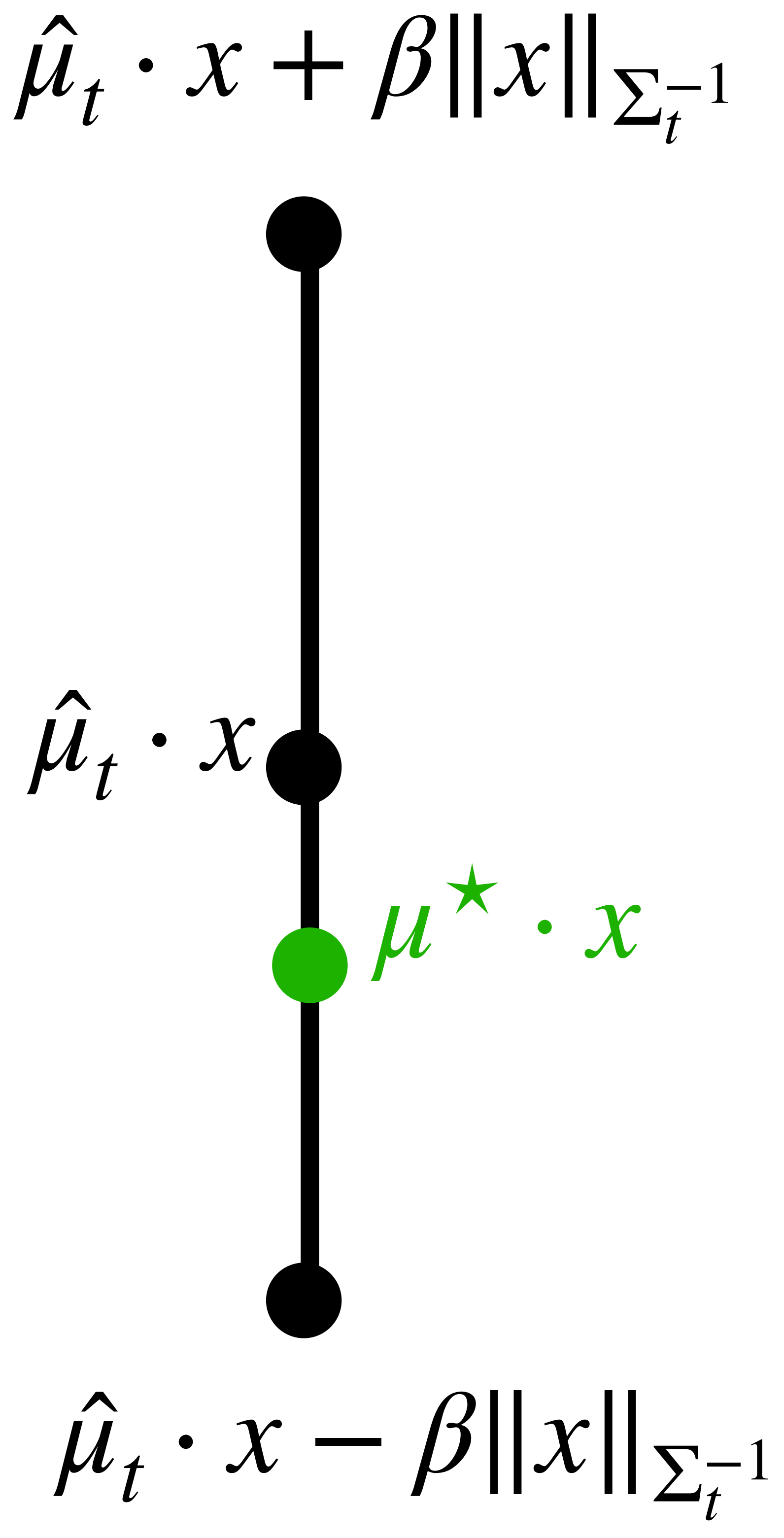


$\hat{\mu}_t \cdot x + \beta\|x\|_{\Sigma_t^{-1}}$
$\hat{\mu}_t \cdot x$
$\mu^\star \cdot x$
$\hat{\mu}_t \cdot x - \beta\|x\|_{\Sigma_t^{-1}}$

Optimism

Optimism: $\mu^\star \cdot x^\star \leq \hat{\mu}_t \cdot x_t + \beta \|x_t\|_{\Sigma_t^{-1}}$

$\forall x \in D$



$\hat{\mu}_t \cdot x + \beta \|x\|_{\Sigma_t^{-1}}$
$\hat{\mu}_t \cdot x$
$\mu^\star \cdot x$
$\hat{\mu}_t \cdot x - \beta \|x\|_{\Sigma_t^{-1}}$

Proof:

Regret

Regret-at-t $= \mu^\star \cdot x^\star - \mu^\star \cdot x_t$

Regret

Regret-at-t $= \mu^\star \cdot x^\star - \mu^\star \cdot x_t$

$$\leq \hat{\mu}_t^\top x_t + \beta \|x_t\|_{\Sigma_t^{-1}} - \mu^\star \cdot x_t$$

Regret

Regret-at-t $= \mu^\star \cdot x^\star - \mu^\star \cdot x_t$

$$\leq \hat{\mu}_t^\top x_t + \beta \|x_t\|_{\Sigma_t^{-1}} - \mu^\star \cdot x_t \quad \leq 2\beta \|x_t\|_{\Sigma_t^{-1}}$$

Regret

Regret-at-t $= \mu^\star \cdot x^\star - \mu^\star \cdot x_t$

$$\leq \hat{\mu}_t^\top x_t + \beta \|x_t\|_{\Sigma_t^{-1}} - \mu^\star \cdot x_t \quad \leq 2\beta \|x_t\|_{\Sigma_t^{-1}}$$

Intuitively this should be convincing already:

Regret

Regret-at-t $= \mu^\star \cdot x^\star - \mu^\star \cdot x_t$

$$\leq \hat{\mu}_t^\top x_t + \beta \|x_t\|_{\Sigma_t^{-1}} - \mu^\star \cdot x_t \quad \leq 2\beta \|x_t\|_{\Sigma_t^{-1}}$$

Intuitively this should be convincing already:

Case 1: x_t is a bad arm, i.e., $2\beta \|x_t\|_{\Sigma_t^{-1}} \geq \mu^\star \cdot (x^\star - x_t) \geq \delta$

Regret

Regret-at-t $= \mu^\star \cdot x^\star - \mu^\star \cdot x_t$

$$\leq \hat{\mu}_t^\top x_t + \beta \|x_t\|_{\Sigma_t^{-1}} - \mu^\star \cdot x_t \quad \leq 2\beta \|x_t\|_{\Sigma_t^{-1}}$$

Intuitively this should be convincing already:

Case 1: x_t is a bad arm, i.e., $2\beta \|x_t\|_{\Sigma_t^{-1}} \geq \mu^\star \cdot (x^\star - x_t) \geq \delta$

x_t falls in the subspace where “data is sparse”, i.e., we explored!

Regret

$$\text{Regret-at-t} = \mu^\star \cdot x^\star - \mu^\star \cdot x_t$$

$$\leq \hat{\mu}_t^\top x_t + \beta \| x_t \|_{\Sigma_t^{-1}} - \mu^\star \cdot x_t \quad \leq 2\beta \| x_t \|_{\Sigma_t^{-1}}$$

Intuitively this should be convincing already:

Case 1: x_t is a bad arm, i.e., $2\beta \| x_t \|_{\Sigma_t^{-1}} \geq \mu^\star \cdot (x^\star - x_t) \geq \delta$

x_t falls in the subspace where “data is sparse”, i.e., we explored!

Case 2: confidence interval $\| x_t \|_{\Sigma_t^{-1}}$ is small

Regret

Regret-at-t $= \mu^\star \cdot x^\star - \mu^\star \cdot x_t$

$$\leq \hat{\mu}_t^\top x_t + \beta \|x_t\|_{\Sigma_t^{-1}} - \mu^\star \cdot x_t \quad \leq 2\beta \|x_t\|_{\Sigma_t^{-1}}$$

Intuitively this should be convincing already:

Case 1: x_t is a bad arm, i.e., $2\beta \|x_t\|_{\Sigma_t^{-1}} \geq \mu^\star \cdot (x^\star - x_t) \geq \delta$

x_t falls in the subspace where “data is sparse”, i.e., we explored!

Case 2: confidence interval $\|x_t\|_{\Sigma_t^{-1}}$ is small

Then regret at this round is small too, i.e., we exploited!

Regret

$$\text{Regret-at-t} = \mu^\star \cdot x^\star - \mu^\star \cdot x_t$$

$$\leq \hat{\mu}_t^\top x_t + \beta \|x_t\|_{\Sigma_t^{-1}} - \mu^\star \cdot x_t \quad \leq 2\beta \|x_t\|_{\Sigma_t^{-1}}$$

More formally, we can show:

$$\text{Regret} \leq \beta \sum_{t=0}^{T-1} \|x_t\|_{\Sigma_t^{-1}}$$

Regret

Regret-at-t $= \mu^\star \cdot x^\star - \mu^\star \cdot x_t$

$$\leq \hat{\mu}_t^\top x_t + \beta \|x_t\|_{\Sigma_t^{-1}} - \mu^\star \cdot x_t \quad \leq 2\beta \|x_t\|_{\Sigma_t^{-1}}$$

More formally, we can show:

Regret $\leq \beta \sum_{t=0}^{T-1} \|x_t\|_{\Sigma_t^{-1}} \leq \beta\sqrt{T} \cdot \sqrt{\sum_{t=0}^{T-1} \|x_t\|^2_{\Sigma_t^{-1}}}$

Regret

Regret-at-t $= \mu^\star \cdot x^\star - \mu^\star \cdot x_t$

$$\leq \hat{\mu}_t^\top x_t + \beta \|x_t\|_{\Sigma_t^{-1}} - \mu^\star \cdot x_t \quad \leq 2\beta \|x_t\|_{\Sigma_t^{-1}}$$

More formally, we can show:

Regret $\leq \beta \sum_{t=0}^{T-1} \|x_t\|_{\Sigma_t^{-1}} \leq \beta\sqrt{T} \cdot \sqrt{\sum_{t=0}^{T-1} \|x_t\|^2_{\Sigma_t^{-1}}}$

$$\lesssim \beta\sqrt{T} \cdot \sqrt{d \ln(T/\lambda + 1)} \qquad \forall \lambda \geq 1$$

Summary

1. To deal w/ infinitely many arms, we introduce linear structure in rewards

Summary

1. To deal w/ infinitely many arms, we introduce linear structure in rewards

2. Analysis of Ridge LR gives us bound on on $|(\mu^\star - \hat{\mu}_t)^\top x|$

Summary

1. To deal w/ infinitely many arms, we introduce linear structure in rewards

2. Analysis of Ridge LR gives us bound on on $|(\mu^\star - \hat{\mu}_t)^\top x|$

3. Optimism in the face of uncertainty: $\mu^\star \cdot x^\star \leq \hat{\mu}_t^\top x_t + \beta \|x_t\|_{\Sigma_t^{-1}}$

Summary

1. To deal w/ infinitely many arms, we introduce linear structure in rewards

2. Analysis of Ridge LR gives us bound on on $|(\mu^\star - \hat{\mu}_t)^\top x|$

3. Optimism in the face of uncertainty: $\mu^\star \cdot x^\star \le \hat{\mu}_t^\top x_t + \beta \|x_t\|_{\Sigma_t^{-1}}$

4. Regret is upper bounded by $\beta \sum_t \|x_t\|_{\Sigma_t} \le \beta \sqrt{T} \sqrt{\sum_t \|x_t\|^2_{\Sigma_t^{-1}}}$