Logarithmic Regret for Learning Linear Quadratic Regulators Efficiently

Asaf Cassel

Joint work with: Alon Cohen, Tomer Koren

Reinforcement Learning

Reinforcement Learning

	Discrete MDP	Linear Quadratic Regulator (LQR)
Space	$x_t \in S, u_t \in A$	
Transition	Unstructured $x_{t+1} \sim P(\cdot \mid x_t, u_t)$	
Costs	Unstructured $c_t = c(x_t, u_t)$	
Optimal Policy	Dynamic programming	
Problem Size	$\lvert S \rvert, \lvert A \rvert$	

Reinforcement Learning

	Discrete MDP	Linear Quadratic Regulator (LQR)
Space	$x_t \in S, u_t \in A$	$x_t \in \mathbb{R}^d, u_t \in \mathbb{R}^k$
Transition	Unstructured $x_{t+1} \sim P(\cdot \mid x_t, u_t)$	**Linear** $x_{t+1} = A_\star x_t + B_\star u_t + w_t$
Costs	Unstructured $c_t = c(x_t, u_t)$	**Quadratic** $c_t = x_t^\top Q x_t + u_t^\top R u_t$
Optimal Policy	Dynamic programming	$u_t = -K_\star x_t$
Problem Size	$\lvert S \rvert, \lvert A \rvert$	$d, k, \lVert A_\star \rVert, \lVert B_\star \rVert$

“Adaptive Control”

- Transition $x_{t+1} = A_\star x_t + B_\star u_t + w_t$
- Cost $c_t = x_t^\top Q x_t + u_t^\top R u_t$
- Optimal Policy $u_t = -K_\star x_t$
- i.i.d noise $w_t \sim \mathcal{N}(0, \sigma^2 I)$

Minimize regret (costs) when $A_\star, B_\star$ are unknown

Important Milestones:

1. Non-efficient $\sqrt{T}$ regret - Abbasi-Yadkori and Szepesvári (2011)
2. Efficient $T^{2/3}$ regret - Dean et al. (2018)
3. First efficient $\sqrt{T}$ regret - Cohen et al. (2019) , Mania et al. (2019)

"Adaptive Control"

- Transition $x_{t+1} = A_\star x_t + B_\star u_t + w_t$
- Cost $c_t = x_t^\top Q x_t + u_t^\top R u_t$
- Optimal Policy $u_t = -K_\star x_t$
- i.i.d noise $w_t \sim \mathcal{N}(0, \sigma^2 I)$

Minimize regret (costs) when $A_\star, B_\star$ are unknown

Important Milestones:

1. Non-efficient $\sqrt{T}$ regret - Abbasi-Yadkori and Szepesvári (2011)
2. Efficient $T^{2/3}$ regret - Dean et al. (2018)
3. First efficient $\sqrt{T}$ regret - Cohen et al. (2019) , Mania et al. (2019)

Is $\sqrt{T}$ regret optimal? No previous lower bounds

“Adaptive Control”

- Transition $x_{t+1} = A_\star x_t + B_\star u_t + w_t$
- Cost $c_t = x_t^\top Q x_t + u_t^\top R u_t$
- Optimal Policy $u_t = -K_\star x_t$
- i.i.d noise $w_t \sim \mathcal{N}(0,\sigma^2 I)$

Minimize regret (costs) when $A_\star, B_\star$ are unknown

Important Milestones:

1. Non-efficient $\sqrt{T}$ regret - Abbasi-Yadkori and Szepesvári (2011)
2. Efficient $T^{2/3}$ regret - Dean et al. (2018)
3. First efficient $\sqrt{T}$ regret - Cohen et al. (2019) , Mania et al. (2019)

Is $\sqrt{T}$ regret optimal? No previous lower bounds

Noise

Typically $\sqrt{T}$ regret

$\log T$ regret in stochastic bandits

"Adaptive Control"

- Transition $x_{t+1} = A_\star x_t + B_\star u_t + w_t$
- Cost $c_t = x_t^\top Q x_t + u_t^\top R u_t$
- Optimal Policy $u_t = -K_\star x_t$
- i.i.d noise $w_t \sim \mathcal{N}(0, \sigma^2 I)$

Minimize regret (costs) when $A_\star, B_\star$ are unknown

Important Milestones:

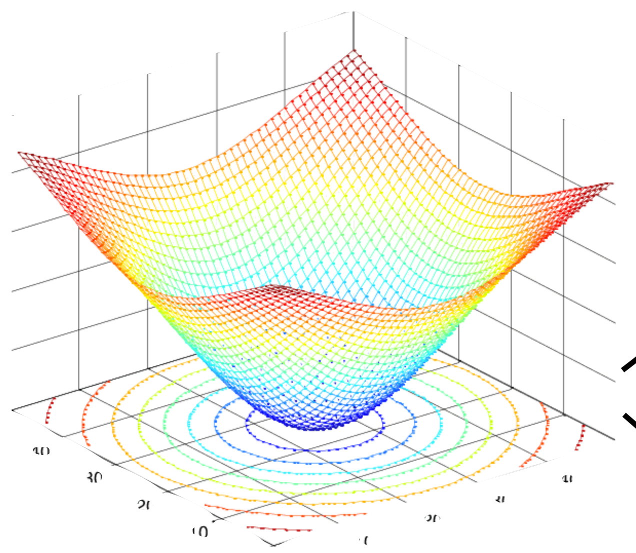
1. Non-efficient $\sqrt{T}$ regret - Abbasi-Yadkori and Szepesvári (2011)
2. Efficient $T^{2/3}$ regret - Dean et al. (2018)
3. First efficient $\sqrt{T}$ regret - Cohen et al. (2019) , Mania et al. (2019)

Is $\sqrt{T}$ regret optimal? No previous lower bounds

Noise

Typically $\sqrt{T}$ regret

$\log T$ regret in stochastic bandits



Objective Structure

Typically $\sqrt{T}$ regret

$\log T$ regret for strongly convex costs

Main Results

$\log T$ regret <u>is</u> possible, sometimes…

- If $A_\star$ unknown ($B_\star$ known) $\Longrightarrow$ efficient algorithm with $\tilde{O}(\log T)$ regret
- If $B_\star$ unknown ($A_\star$ known) $\Longrightarrow$ efficient algorithm with $\tilde{O}\left(\frac{\log T}{\lambda_{\min}(K_\star K_\star^\top)}\right)$ regret

$\tilde{O}$ only hides polynomial dependence on problem parameters

- Transition $x_{t+1} = A_\star x_t + B_\star u_t + w_t$
- Cost $c_t = x_t^\top Q x_t + u_t^\top R u_t$
- Optimal Policy $u_t = -K_\star x_t$
- i.i.d noise $w_t \sim \mathcal{N}(0, \sigma^2 I)$

Main Results

- Transition $x_{t+1} = A_\star x_t + B_\star u_t + w_t$
- Cost $c_t = x_t^\top Q x_t + u_t^\top R u_t$
- Optimal Policy $u_t = -K_\star x_t$
- i.i.d noise $w_t \sim \mathcal{N}(0, \sigma^2 I)$

$\log T$ regret <u>is</u> possible, sometimes…

- If $A_\star$ unknown ($B_\star$ known) $\Longrightarrow$ efficient algorithm with $\tilde{O}(\log T)$ regret
- If $B_\star$ unknown ($A_\star$ known) $\Longrightarrow$ efficient algorithm with $\tilde{O}\left(\frac{\log T}{\lambda_{\min}(K_\star K_\star^\top)}\right)$ regret

$\tilde{O}$ only hides polynomial dependence on problem parameters

… but in general, $\sqrt{T}$ regret is unavoidable

- First* $\Omega(\sqrt{T})$ regret lower bound for the adaptive LQR problem
- Holds even when $A_\star$ is known
- Construction relies on small $\lambda_{\min}(K_\star K_\star^\top)$

* concurrently with Simchowitz and Foster (2020)

Formalities

- Transition $x_{t+1} = A_\star x_t + B_\star u_t + w_t$
- Cost $c_t = x_t^\top Q x_t + u_t^\top R u_t$
- Optimal Policy $u_t = -K_\star x_t$
- i.i.d noise $w_t \sim \mathcal{N}(0,\sigma^2 I)$
- Objective $J = \lim_{T\to\infty} \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T} c_t\right]$

Linear Quadratic Control

Choose $u_1, u_2, \ldots$ that minimize $J = \lim_{T\to\infty} \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T} c_t\right]$

- Optimal policy: $u_t = -K_\star x_t$, Optimal infinite horizon average cost: $J(K_\star)$
- $K_\star := K_\star(A_\star, B_\star, Q, R)$ can be efficiently calculated (Riccati equation)

Formalities

Linear Quadratic Control

Choose $u_1, u_2, \ldots$ that minimize $J = \lim_{T\to\infty} \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T} c_t\right]$

- Optimal policy: $u_t = -K_\star x_t$, Optimal infinite horizon average cost: $J(K_\star)$
- $K_\star := K_\star(A_\star, B_\star, Q, R)$ can be efficiently calculated (Riccati equation)

Learning Objective

Regret minimization under parameter uncertainty.

$$\text{Regret} = \mathbb{E}\left[\sum_{t=1}^{T} (c_t - J(K_\star))\right]$$

- Transition $x_{t+1} = A_\star x_t + B_\star u_t + w_t$
- Cost $c_t = x_t^\top Q x_t + u_t^\top R u_t$
- Optimal Policy $u_t = -K_\star x_t$
- i.i.d noise $w_t \sim \mathcal{N}(0, \sigma^2 I)$
- Objective $J = \lim_{T\to\infty} \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T} c_t\right]$

Formalities

Regret Reparameterization

$$\text{Playing } u_t = -K_t x_t \overset{*}{\Longrightarrow} \text{Regret } \approx \mathbb{E}\left[\sum_{t=1}^{T} (J(K_t) - J(K_\star))\right]$$

*As long as K_t does not change too often

- Transition $x_{t+1} = A_\star x_t + B_\star u_t + w_t$
- Cost $c_t = x_t^\top Q x_t + u_t^\top R u_t$
- Optimal Policy $u_t = -K_\star x_t$
- i.i.d noise $w_t \sim \mathcal{N}(0, \sigma^2 I)$
- Objective $J = \lim_{T\to\infty} \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T} c_t\right]$

Formalities

- Transition $x_{t+1} = A_\star x_t + B_\star u_t + w_t$
- Cost $c_t = x_t^\top Q x_t + u_t^\top R u_t$
- Optimal Policy $u_t = -K_\star x_t$
- i.i.d noise $w_t \sim \mathcal{N}(0, \sigma^2 I)$
- Objective $J = \lim_{T\to\infty} \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T} c_t\right]$

Regret Reparameterization

$$\text{Playing } u_t = -K_t x_t \overset{*}{\Longrightarrow} \text{Regret } \approx \mathbb{E}\left[\sum_{t=1}^{T}(J(K_t) - J(K_\star))\right]$$

*As long as K_t does not change too often

Strong Stability (Cohen et al. 2018)

$$\text{Playing } u_t = -Kx_t \implies \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T} c_t\right] \xrightarrow{\text{exponentially}} J(K)$$

Definition:

$K \in \mathbb{R}^{k\times d}$ is (κ, γ)-strongly stable for $A_\star, B_\star$ if $\exists H, L$ such that:

1. $A_\star + B_\star K = HLH^{-1}$
2. $\|L\| \le 1 - \gamma$, and $\|H\|, \|H^{-1}\|, \|K\| \le \kappa$

A Recipe for $\sqrt{T}$ Regret?

First order estimation

Assuming $J(K)$ is Lipschitz:

$$\text{Regret} \approx \mathbb{E}\left[\sum_{t=1}^{T}(J(K_t) - J(K_\star))\right] \lessapprox \mathbb{E}\left[\sum_{t=1}^{T}\|K_t - K_\star\|\right]$$

- Transition $x_{t+1} = A_\star x_t + B_\star u_t + w_t$
- Cost $c_t = x_t^\top Q x_t + u_t^\top R u_t$
- Optimal Policy $u_t = -K_\star x_t$
- i.i.d noise $w_t \sim \mathcal{N}(0,\sigma^2 I)$
- Objective $J = \lim_{T\to\infty} \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T} c_t\right]$

A Recipe for $\sqrt{T}$ Regret?

First order estimation

Assuming $J(K)$ is Lipschitz:

$$\text{Regret} \approx \mathbb{E}\left[\sum_{t=1}^{T} (J(K_t) - J(K_\star))\right] \lessapprox \mathbb{E}\left[\sum_{t=1}^{T} \|K_t - K_\star\|\right]$$

Perform minimal exploration to get $\|K_t - K_\star\| \leq 1/\sqrt{T}$ and then play K_t:

$$\text{Regret} \approx \sqrt{T} + \text{exploration cost}$$

- Transition $x_{t+1} = A_\star x_t + B_\star u_t + w_t$
- Cost $c_t = x_t^\top Q x_t + u_t^\top R u_t$
- Optimal Policy $u_t = -K_\star x_t$
- i.i.d noise $w_t \sim \mathcal{N}(0, \sigma^2 I)$
- Objective $J = \lim_{T\to\infty} \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T} c_t\right]$

A Recipe for $\sqrt{T}$ Regret?

- Transition $x_{t+1} = A_\star x_t + B_\star u_t + w_t$
- Cost $c_t = x_t^\top Q x_t + u_t^\top R u_t$
- Optimal Policy $u_t = -K_\star x_t$
- i.i.d noise $w_t \sim \mathcal{N}(0, \sigma^2 I)$
- Objective $J = \lim_{T\to\infty} \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T} c_t\right]$

First order estimation

Assuming $J(K)$ is Lipschitz:

$$\text{Regret} \approx \mathbb{E}\left[\sum_{t=1}^{T} (J(K_t) - J(K_\star))\right] \lessapprox \mathbb{E}\left[\sum_{t=1}^{T} \|K_t - K_\star\|\right]$$

Perform minimal exploration to get $\|K_t - K_\star\| \leq 1/\sqrt{T}$ and then play K_t:

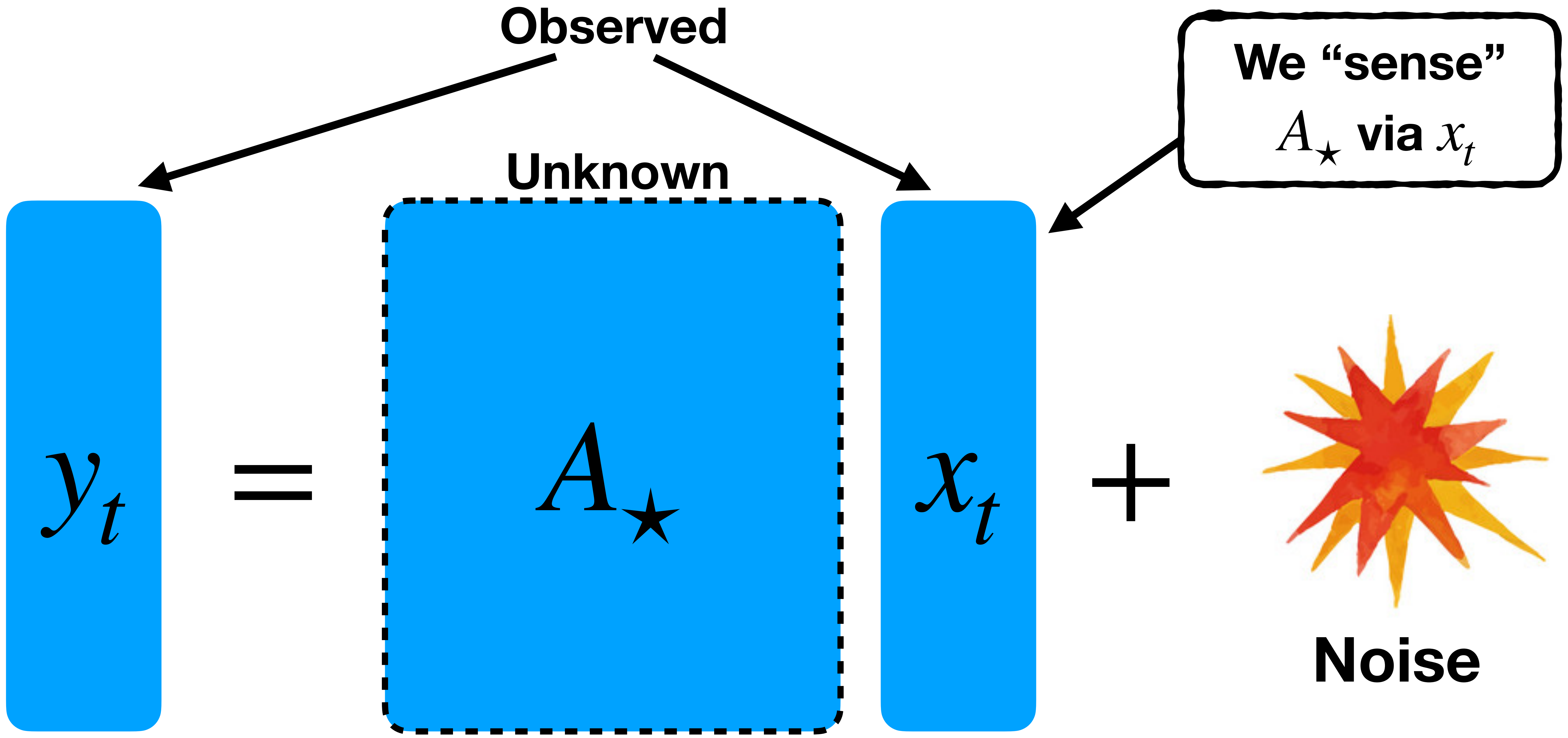
$$\text{Regret} \approx \sqrt{T} + \text{exploration cost}$$

Challenges

- Estimation rate is $\|K_t - K_\star\| \gtrapprox 1/\sqrt{T}$
- Exploration can be expensive! e.g., in previous work $\|K_t - K_\star\| \leq T^{-1/4}$

Case1: Unknown $A_\star$ (Known $B_\star$)

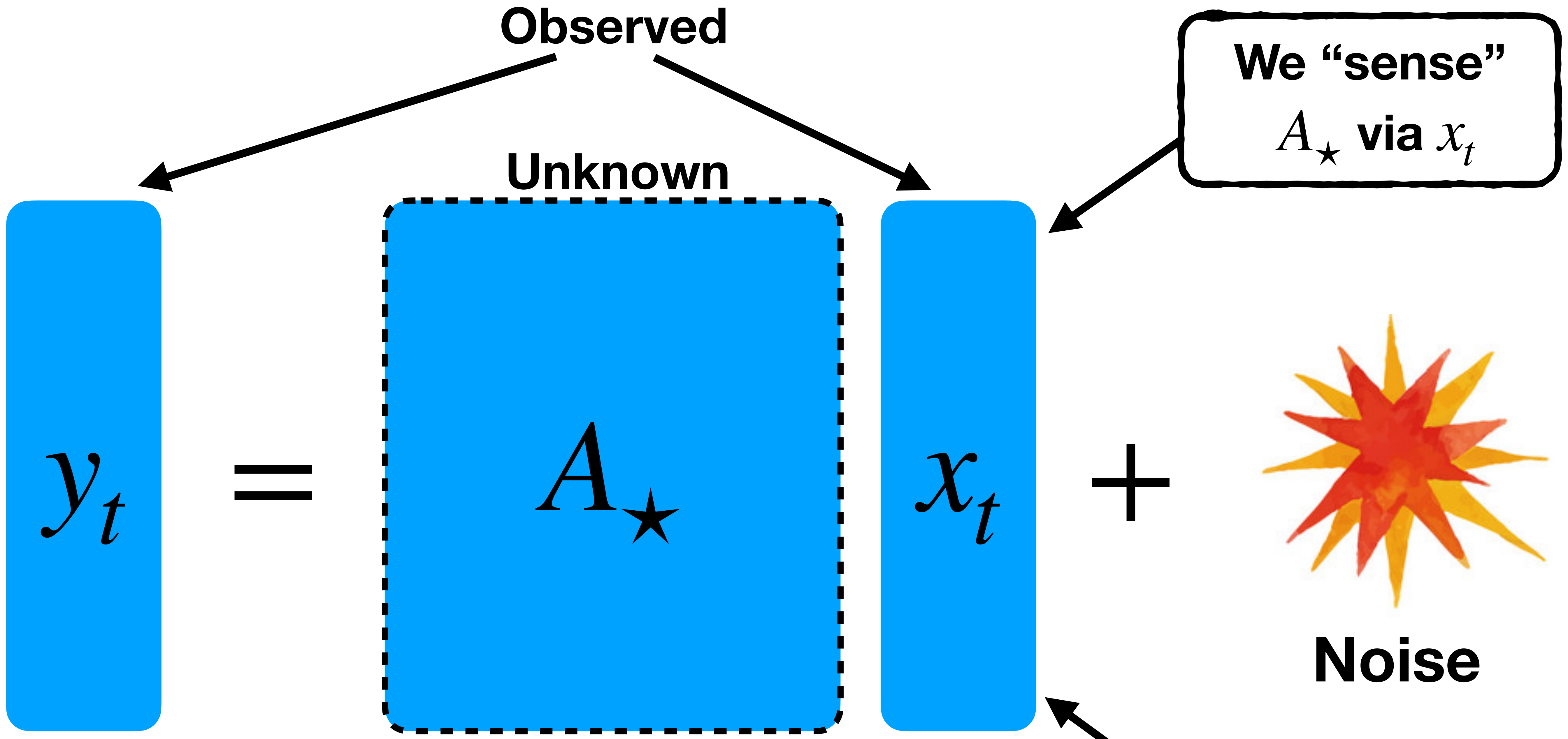
$B_\star$ known $\implies y_t = x_{t+1} - B_\star u_t$



- Transition $x_{t+1} = A_\star x_t + B_\star u_t + w_t$
- Cost $c_t = x_t^\top Q x_t + u_t^\top R u_t$
- Optimal Policy $u_t = -K_\star x_t$
- i.i.d noise $w_t \sim \mathcal{N}(0, \sigma^2 I)$
- Objective $J = \lim_{T\to\infty} \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T} c_t\right]$

Case1: Unknown $A_\star$ (Known $B_\star$)

$B_\star$ known $\implies y_t = x_{t+1} - B_\star u_t$



Free Exploration By w_{t-1} !

<u>Least Squares Estimation ($\hat{A}_t$) Error:</u>

$$\|\hat{A}_t - A_\star\| \propto \frac{\sigma}{\sqrt{\lambda_{\min}\left(\sum_{s=1}^{t} w_s w_s^\top\right)}} \propto T^{-1/2}$$

- Transition $x_{t+1} = A_\star x_t + B_\star u_t + w_t$
- Cost $c_t = x_t^\top Q x_t + u_t^\top R u_t$
- Optimal Policy $u_t = -K_\star x_t$
- i.i.d noise $w_t \sim \mathcal{N}(0, \sigma^2 I)$
- Objective $J = \lim_{T\to\infty} \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T} c_t\right]$

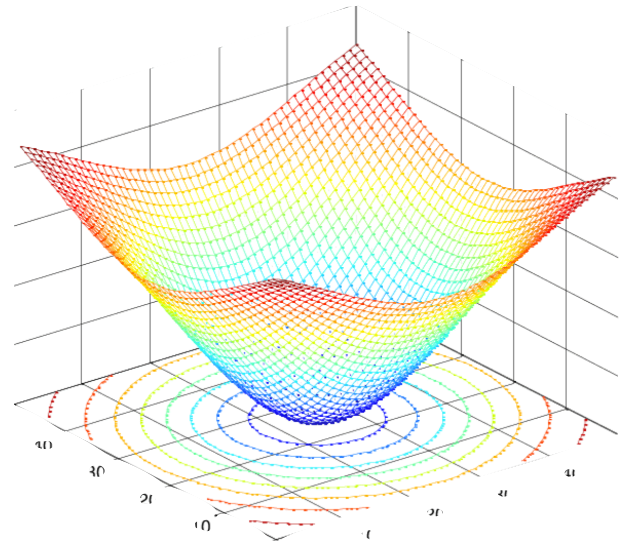
Objective Structure

- Transition $x_{t+1} = A_\star x_t + B_\star u_t + w_t$
- Cost $c_t = x_t^\top Q x_t + u_t^\top R u_t$
- Optimal Policy $u_t = -K_\star x_t$
- i.i.d noise $w_t \sim \mathcal{N}(0, \sigma^2 I)$
- Objective $J = \lim_{T \to \infty} \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T} c_t\right]$

Results by Mania et al. (2019)

- **"Strong Convexity"**:

$$J(K) - J(K_\star) \leq c_1 \|K - K_\star\|^2$$



- **System** estimation $\Longrightarrow$ **Policy** estimation:

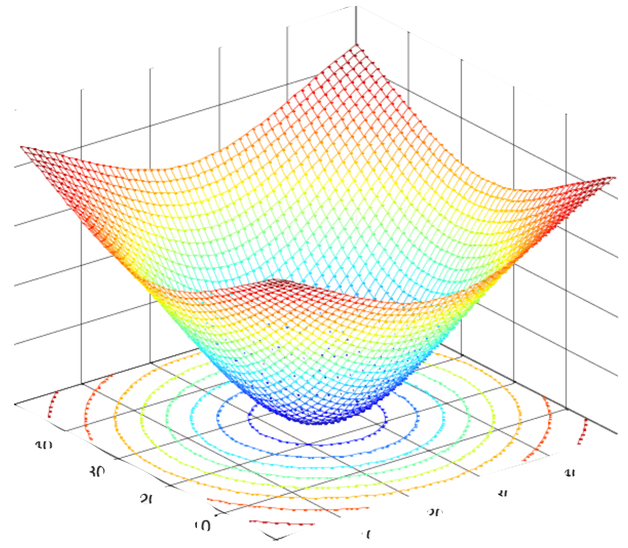
$$\|K_\star(\hat{A}, \hat{B}) - K_\star(A_\star, B_\star)\| \leq c_2 \max\left\{\|\hat{A} - A_\star\|, \|\hat{B} - B_\star\|\right\}$$

Objective Structure

- Transition $x_{t+1} = A_\star x_t + B_\star u_t + w_t$
- Cost $c_t = x_t^\top Q x_t + u_t^\top R u_t$
- Optimal Policy $u_t = -K_\star x_t$
- i.i.d noise $w_t \sim \mathcal{N}(0, \sigma^2 I)$
- Objective $J = \lim_{T\to\infty} \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T} c_t\right]$

Results by Mania et al. (2019)

- "**Strong Convexity**":



$$J(K) - J(K_\star) \le c_1 \|K - K_\star\|^2$$

- **System** estimation $\Longrightarrow$ **Policy** estimation:

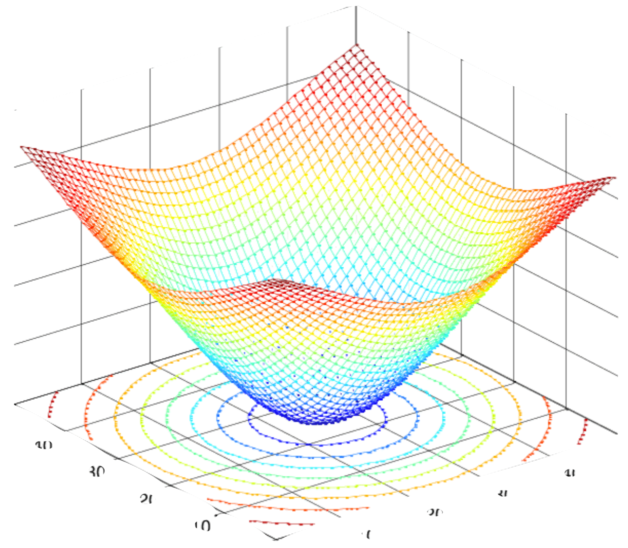
$$\|K_\star(\hat{A}, \hat{B}) - K_\star(A_\star, B_\star)\| \le c_2 \max\left\{\|\hat{A} - A_\star\|, \|\hat{B} - B_\star\|\right\}$$

$\frac{1}{\sqrt{t}}$ estimation $\Longrightarrow$ $\frac{1}{t}$ optimal policy $\overset{?}{\Longrightarrow}$ $\sum_t \frac{1}{t} = \log T$ regret

Objective Structure

Results by Mania et al. (2019)

- "**Strong Convexity**":



$$J(K) - J(K_\star) \le c_1 \|K - K_\star\|^2$$

- **System** estimation $\Longrightarrow$ **Policy** estimation:

$$\|K_\star(\hat{A}, \hat{B}) - K_\star(A_\star, B_\star)\| \le c_2 \max \left\{ \|\hat{A} - A_\star\|, \|\hat{B} - B_\star\| \right\}$$

$\frac{1}{\sqrt{t}}$ estimation $\Longrightarrow$ $\frac{1}{t}$ optimal policy $\overset{?}{\Longrightarrow}$ $\sum_t \frac{1}{t} = \log T$ regret

Not Quite...

- K_t is not stable $\Longrightarrow$ $J(K_t) = \infty$
- Low probability event contributes unbounded regret

- Transition $x_{t+1} = A_\star x_t + B_\star u_t + w_t$
- Cost $c_t = x_t^\top Q x_t + u_t^\top R u_t$
- Optimal Policy $u_t = -K_\star x_t$
- i.i.d noise $w_t \sim \mathcal{N}(0, \sigma^2 I)$
- Objective $J = \lim_{T \to \infty} \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T} c_t\right]$

Algorithm and Abort Mechanism

“Abort”

At every round before playing:

- $\|x_t\|, \|K_t\|$ bounded in high probability bounds? $\Longrightarrow$ Low probability trigger
- Otherwise “abort”: Play K_0 forever $\Longrightarrow$ Constant regret

Assumed Stable

- Transition $x_{t+1} = A_\star x_t + B_\star u_t + w_t$
- Cost $c_t = x_t^\top Q x_t + u_t^\top R u_t$
- Optimal Policy $u_t = -K_\star x_t$
- i.i.d noise $w_t \sim \mathcal{N}(0, \sigma^2 I)$
- Objective $J = \lim_{T\to\infty} \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T} c_t\right]$

Algorithm and Abort Mechanism

"Abort"

At every round before playing:

- $||x_t||, ||K_t||$ bounded in high probability bounds? $\Longrightarrow$ Low probability trigger
- Otherwise "abort": Play K_0 forever $\Longrightarrow$ Constant regret

Assumed Stable

Overall low order regret term!

- Transition $x_{t+1} = A_\star x_t + B_\star u_t + w_t$
- Cost $c_t = x_t^\top Q x_t + u_t^\top R u_t$
- Optimal Policy $u_t = -K_\star x_t$
- i.i.d noise $w_t \sim \mathcal{N}(0, \sigma^2 I)$
- Objective $J = \lim_{T\to\infty} \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T} c_t\right]$

Algorithm and Abort Mechanism

- Transition $x_{t+1} = A_\star x_t + B_\star u_t + w_t$
- Cost $c_t = x_t^\top Q x_t + u_t^\top R u_t$
- Optimal Policy $u_t = -K_\star x_t$
- i.i.d noise $w_t \sim \mathcal{N}(0, \sigma^2 I)$
- Objective $J = \lim_{T\to\infty} \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T} c_t\right]$

"Abort"

<u>At every round before playing:</u>

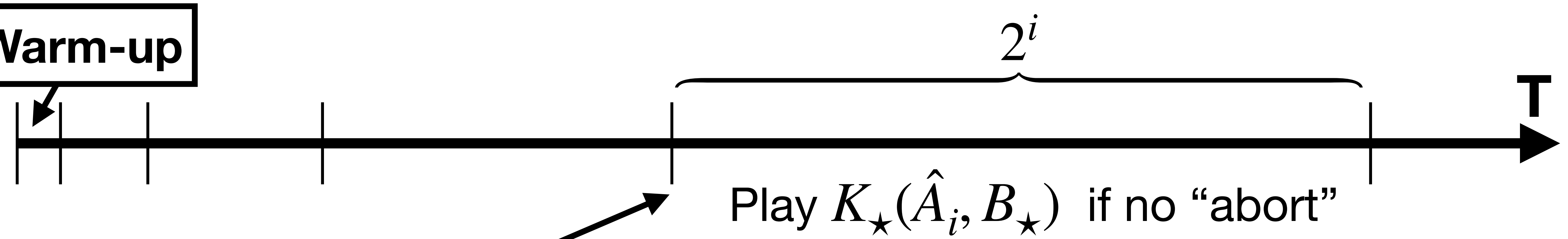
- $\|x_t\|, \|K_t\|$ bounded in high probability bounds? $\Longrightarrow$ <u>Low probability trigger</u>
- Otherwise "abort": Play K_0 forever $\Longrightarrow$ <u>Constant regret</u>

Assumed Stable

Overall low order regret term!

Algorithm for Unknown $A_\star$

Warm-up



<u>Epoch i start:</u>

Estimate (LSE) $\hat{A}_i$

Calculate greedy $K_\star(\hat{A}_i, B_\star)$

Analysis Overview

Regret Decomposition

Regret $\lessapprox \mathbb{E}\left[\sum_{t=1}^{T}(J(K_t) - J(K_\star)) \,\middle|\, \text{no abort}\right]$ + Switching Cost + Abort Cost

$\leq$ **constant** $\cdot$ **#epochs** $\approx \log T$

$\leq$ **constant** $\cdot$ **low probability** $\approx$ **constant**

- Transition $x_{t+1} = A_\star x_t + B_\star u_t + w_t$
- Cost $c_t = x_t^\top Q x_t + u_t^\top R u_t$
- Optimal Policy $u_t = -K_\star x_t$
- i.i.d noise $w_t \sim \mathcal{N}(0, \sigma^2 I)$
- Objective $J = \lim_{T\to\infty} \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T} c_t\right]$

Analysis Overview

Regret Decomposition

Regret $\lessapprox \mathbb{E}\left[\sum_{t=1}^{T}(J(K_t) - J(K_\star)) \,\Big|\, \text{no abort}\right]$ + Switching Cost + Abort Cost

Switching Cost $\leq$ **constant · #epochs** $\approx \log T$

Abort Cost $\leq$ **constant · low probability** $\approx$ **constant**

Putting it all together

$$\mathbb{E}\left[\sum_{t=1}^{T}(J(K_t) - J(K_\star)) \,\Big|\, \text{no abort}\right] \lessapprox \sum_{i=1}^{\#epochs} 2^i \|\hat{A}_i - A_\star\|^2 \lessapprox \#epochs \approx \log T$$

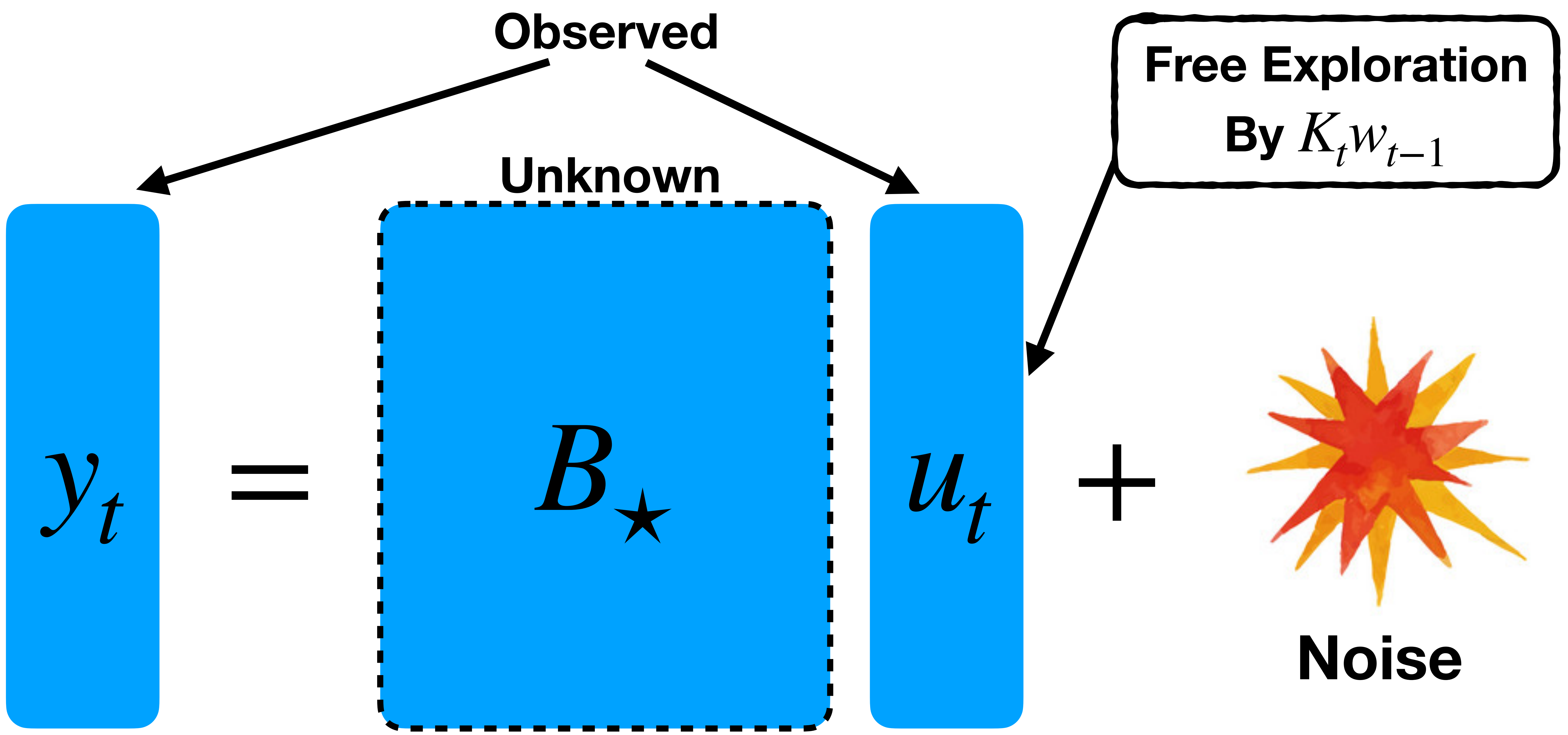
$\|\hat{A}_i - A_\star\|^2 \lessapprox 2^{-(i-1)} =$ epoch length

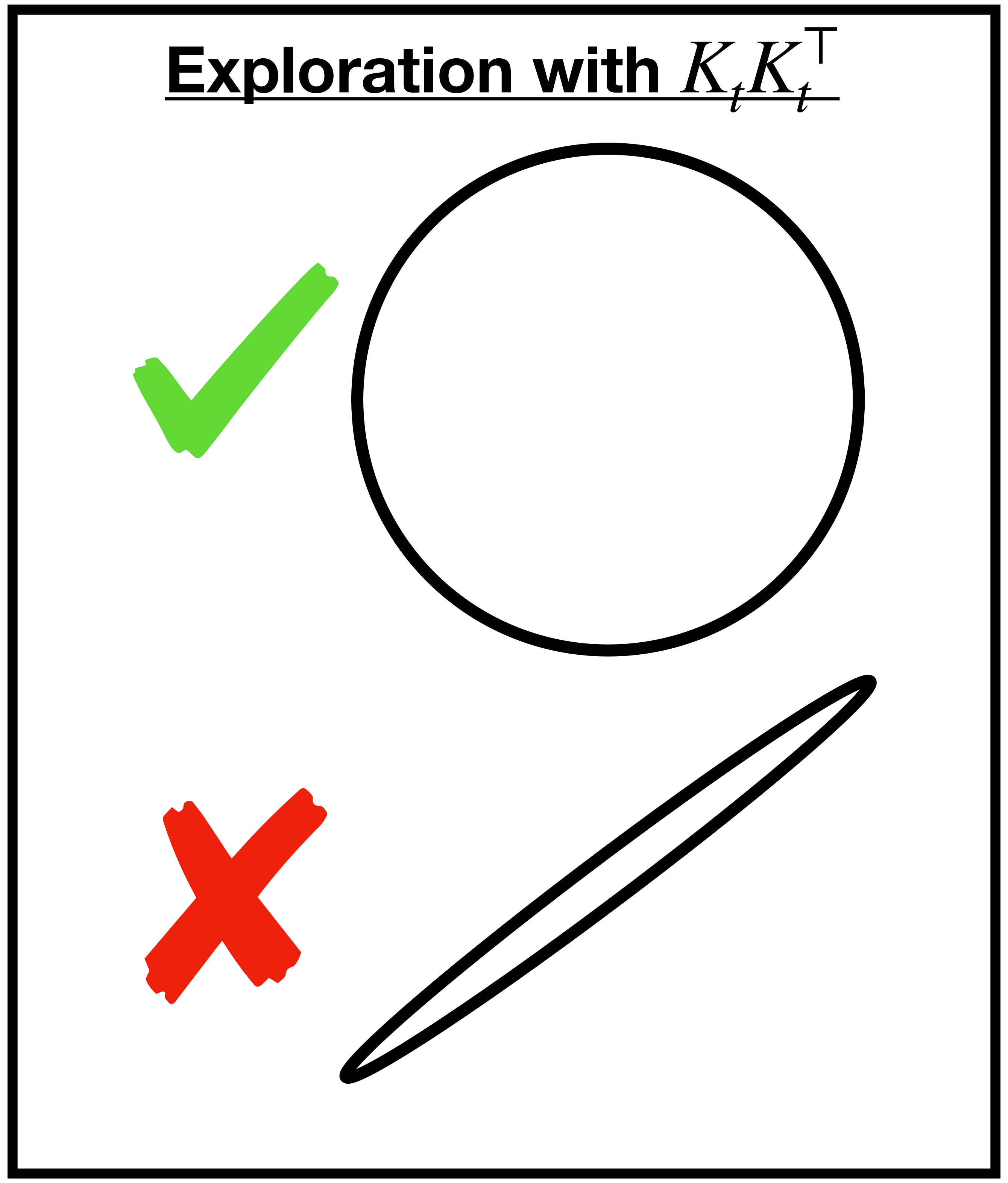
- Transition $x_{t+1} = A_\star x_t + B_\star u_t + w_t$
- Cost $c_t = x_t^\top Q x_t + u_t^\top R u_t$
- Optimal Policy $u_t = -K_\star x_t$
- i.i.d noise $w_t \sim \mathcal{N}(0, \sigma^2 I)$
- Objective $J = \lim_{T\to\infty} \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T} c_t\right]$

Case2: Unknown $B_\star$ (Known $A_\star$)

- Transition $x_{t+1} = A_\star x_t + B_\star u_t + w_t$
- Cost $c_t = x_t^\top Q x_t + u_t^\top R u_t$
- Optimal Policy $u_t = -K_\star x_t$
- i.i.d noise $w_t \sim \mathcal{N}(0, \sigma^2 I)$

Assume $A_\star$ is known $\implies y_t = x_{t+1} - A_\star x_t$





- $K_t K_t^\top \to K_\star K_\star^\top \implies$ Must have $K_\star K_\star^\top \succ \mu_\star I$
- Convergence ensured by Adaptive Warm-up!
- No need to know $\mu_\star$

Lower Bound

Main Ideas

Construction inspired by upper bound $\implies$ $k_\star$ near degenerate

Construction in 1-D

$$\left.\begin{aligned} x_{t+1} &= \frac{1}{2}x_t \pm \varepsilon u_t + w_t \\ c_t &= x_t^2 + u_t^2 \end{aligned}\right. \implies k_\star \approx \mp \varepsilon$$

Lower Bound

Main Ideas

Construction inspired by upper bound $\Longrightarrow$ $k_\star$ near degenerate

Construction in 1-D

$$x_{t+1} = \frac{1}{2}x_t \pm \varepsilon u_t + w_t \quad \Longrightarrow \quad k_\star \approx \mp \varepsilon$$

$$c_t = x_t^2 + u_t^2$$

Learner's Dilemma $\sum_{t=1}^{T} u_t^2$

- Good exploration but Regret $\gtrapprox \sum_{t=1}^{T} u_t^2$
- Bad exploration $\Longrightarrow$ Failed to identify $\operatorname{sign}(k_\star)$

$\varepsilon = T^{-1/4}$ $\Longrightarrow$ **Best Tradeoff gives $\Omega(\sigma^2\sqrt{T})$ regret lower bound**

Summary

- $\log T$ regret is possible sometimes:
  - i) $A_\star$ unknown ($B_\star$ known)
  - ii) $B_\star$ unknown ($A_\star$ known) & $K_\star$ non-degenerate
- In general $\sqrt{T}$ regret is unavoidable

See you at the Q&A session!