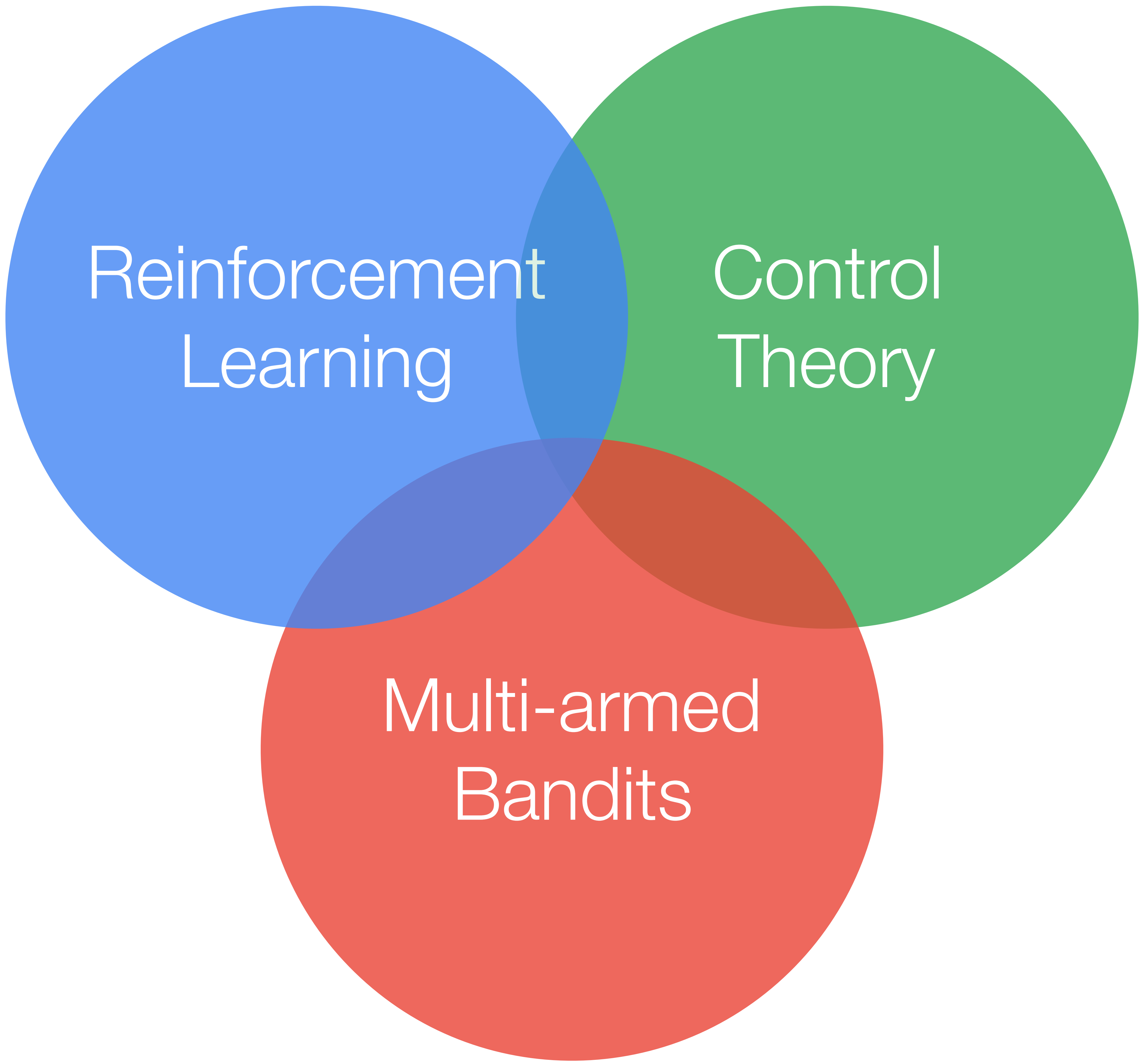
Learning Linear Quadratic Regulators Efficiently with Only $\sqrt{T}$ Regret

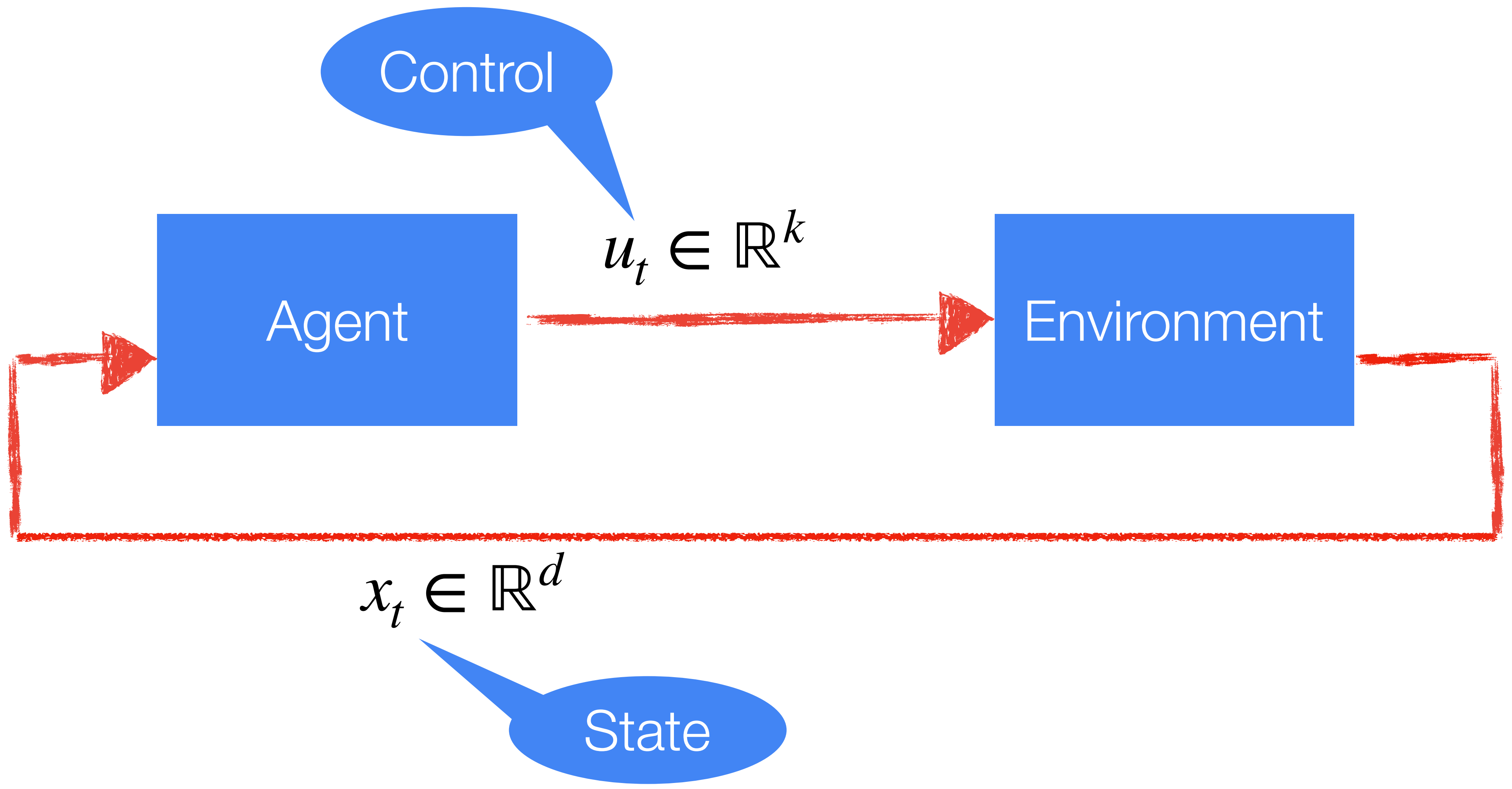
Alon Cohen

Joint work with: Tomer Koren and Yishay Mansour

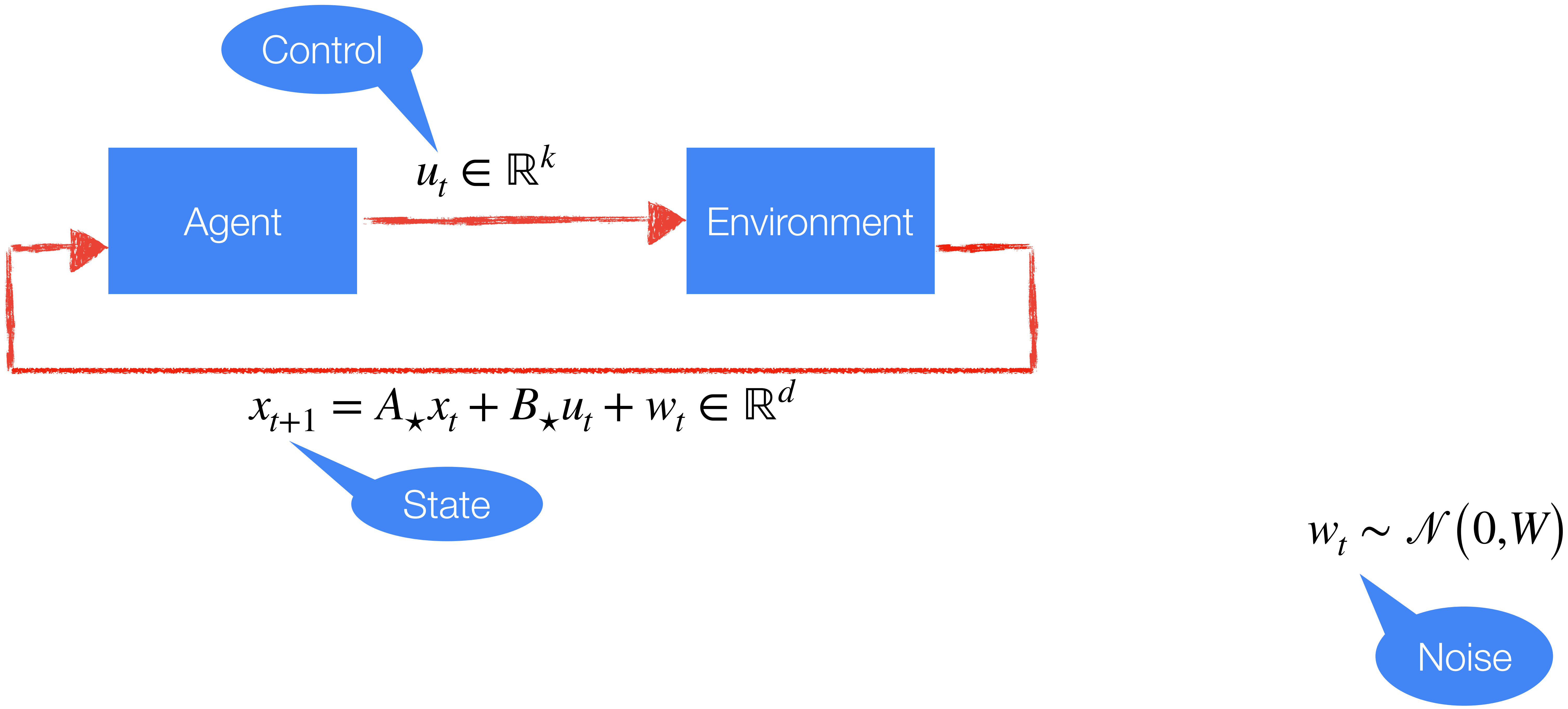

Reinforcement
Learning
Control
Theory
Multi-armed
Bandits

Linear Quadratic Control

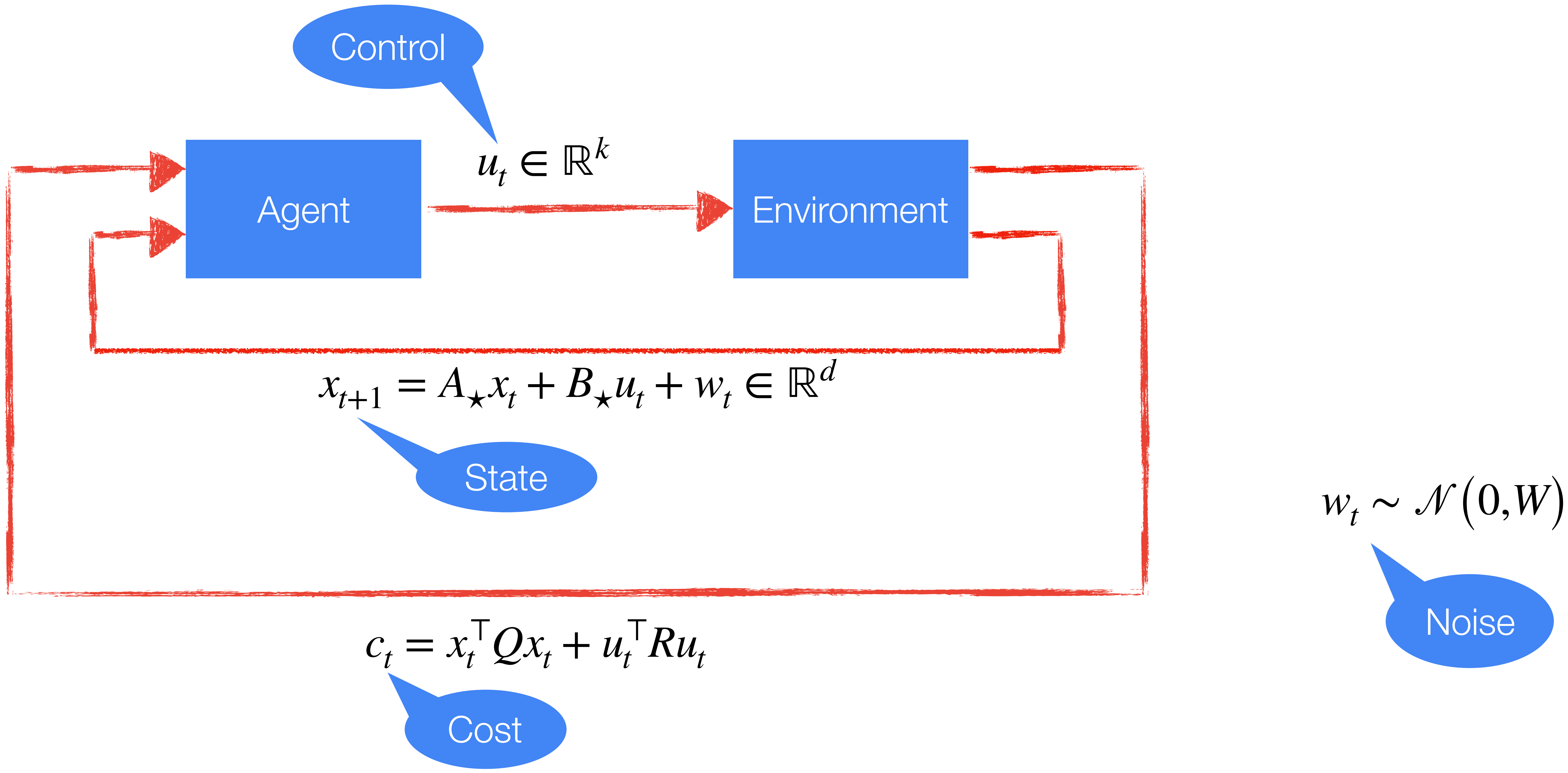
Linear Quadratic Control



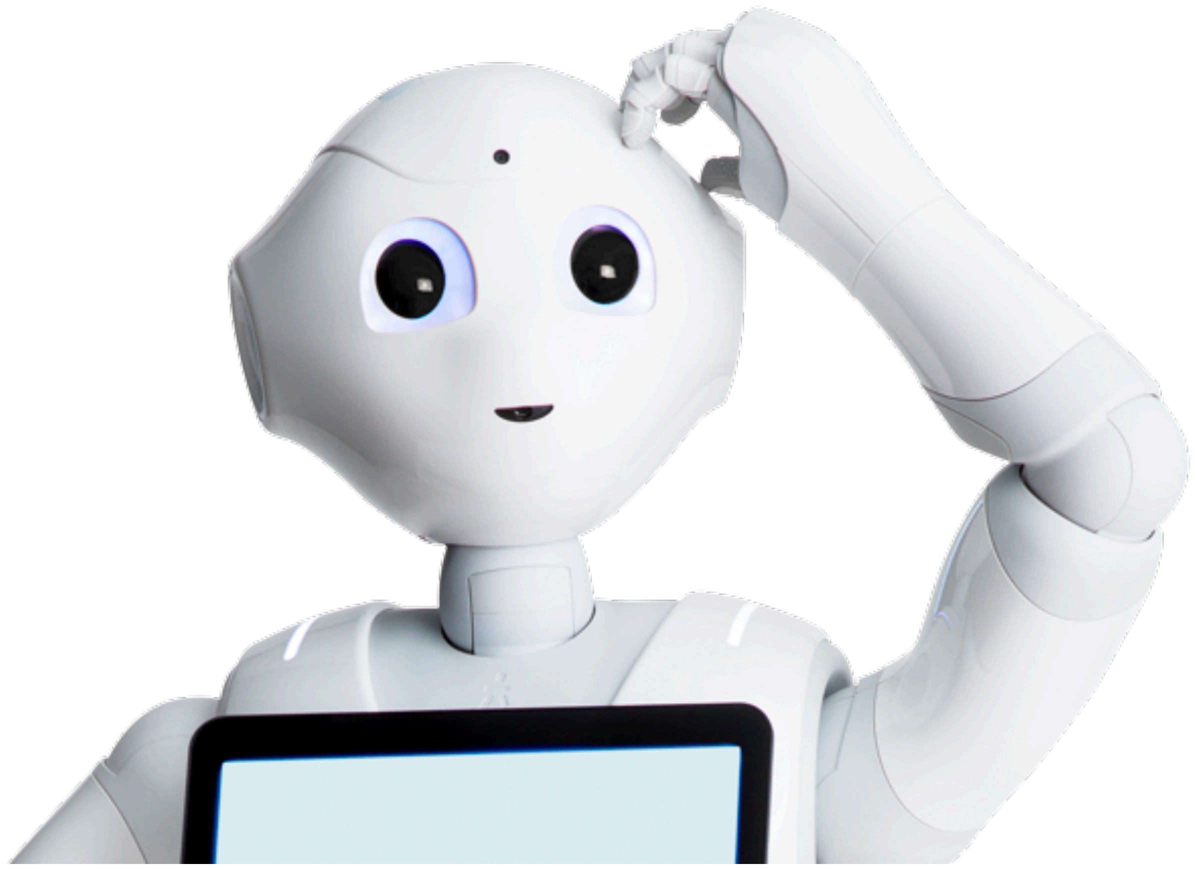
Linear Quadratic Control

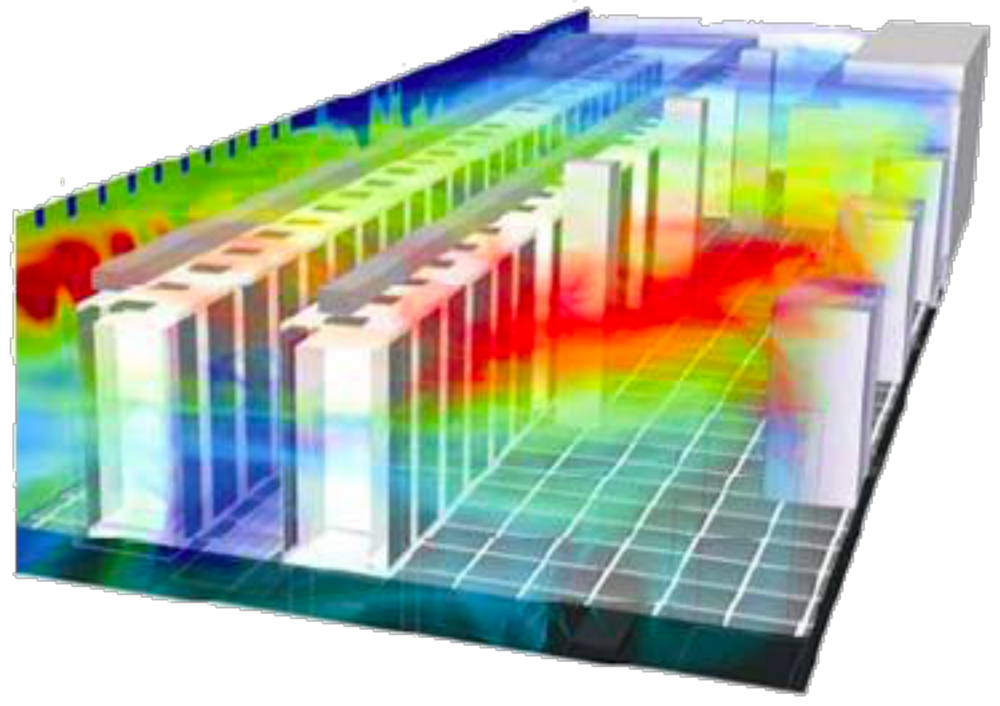


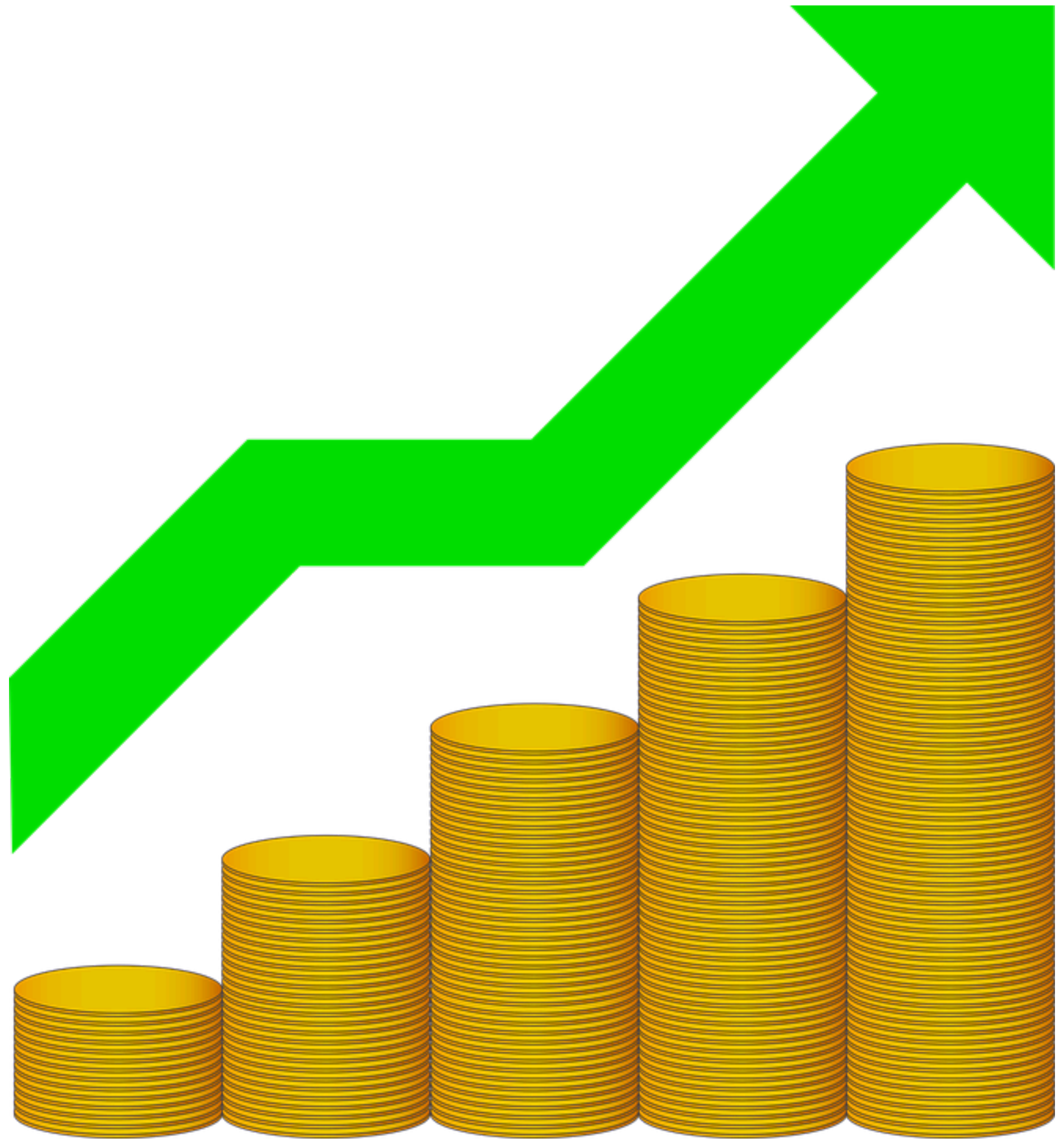
Linear Quadratic Control

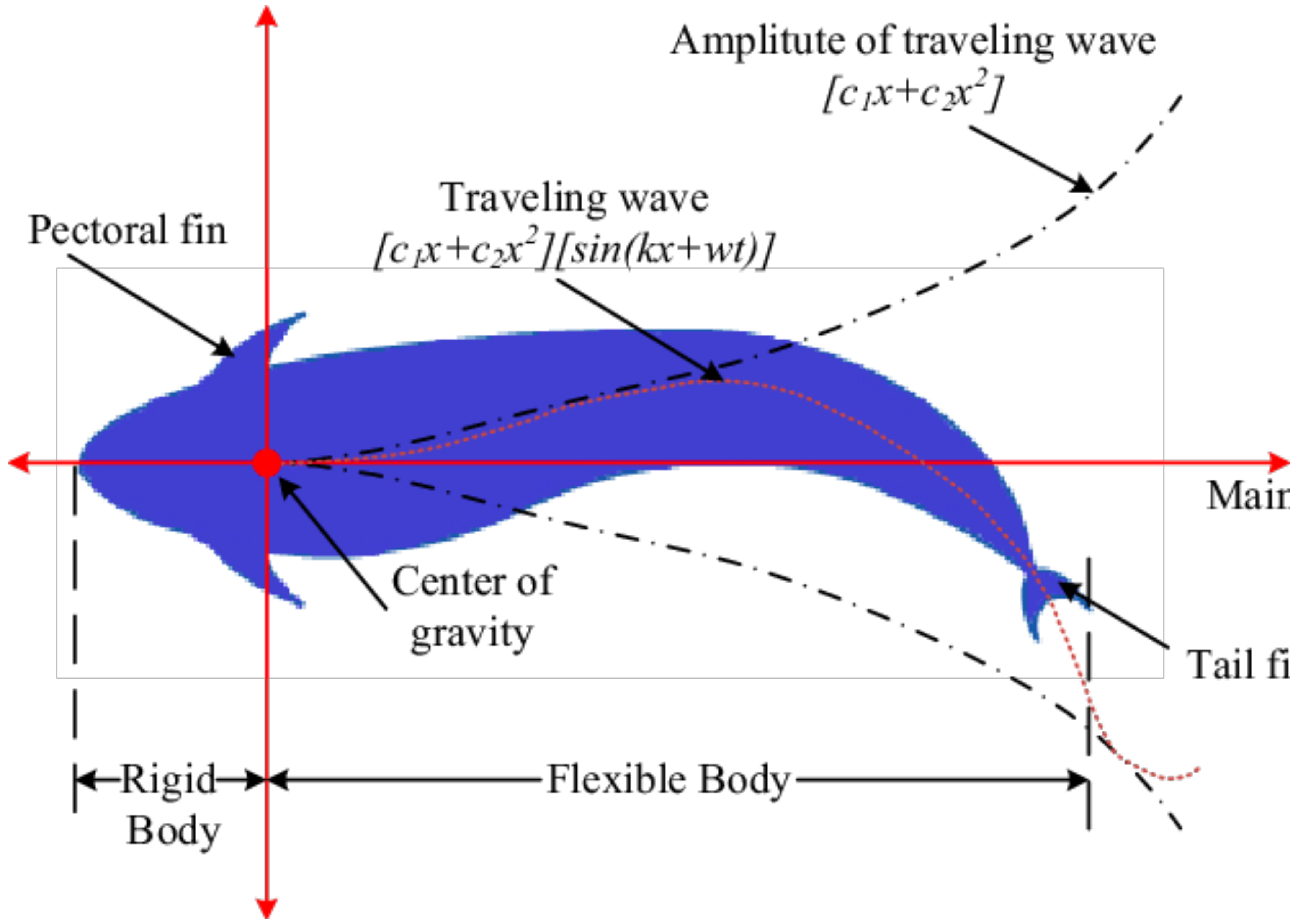


Applications

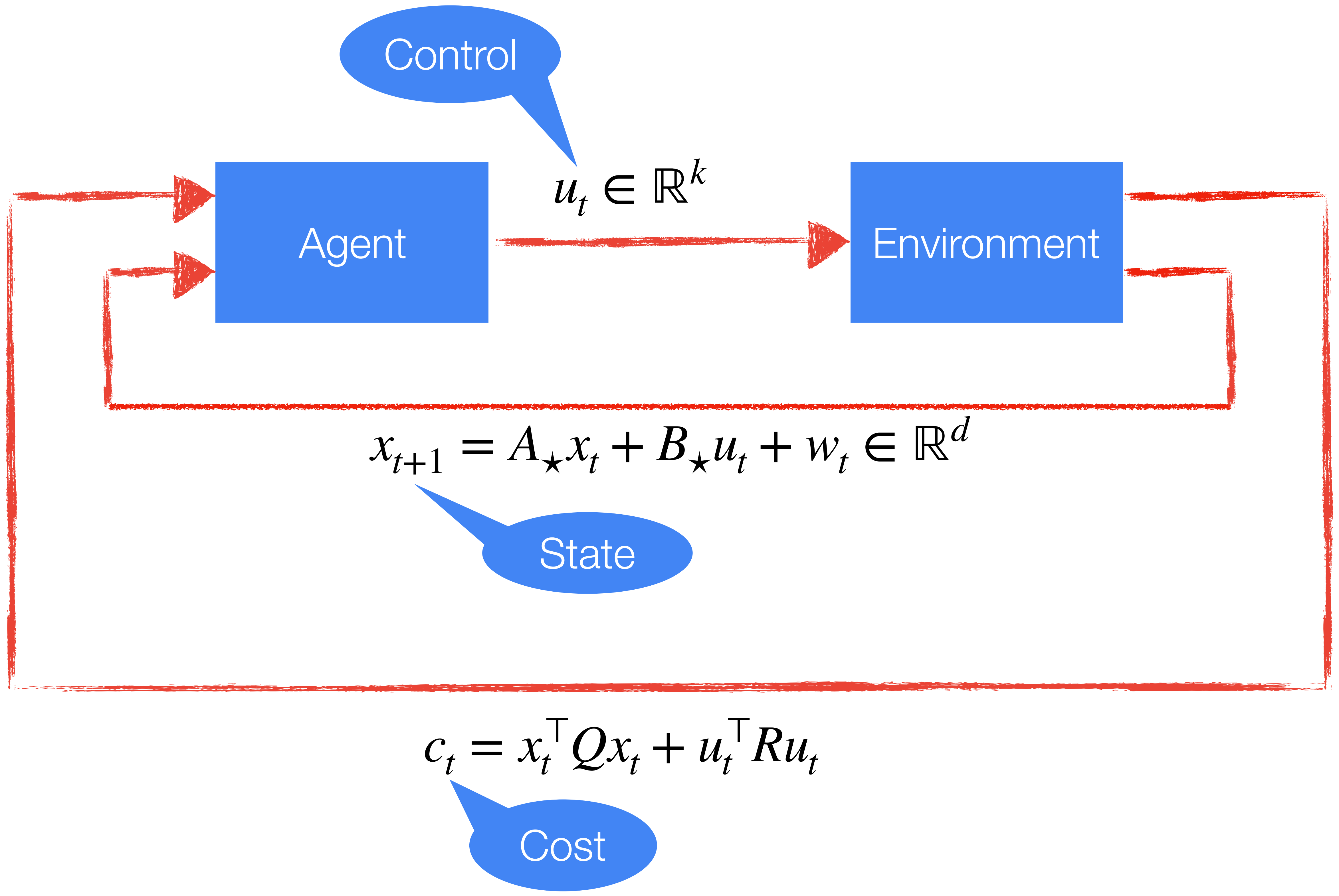








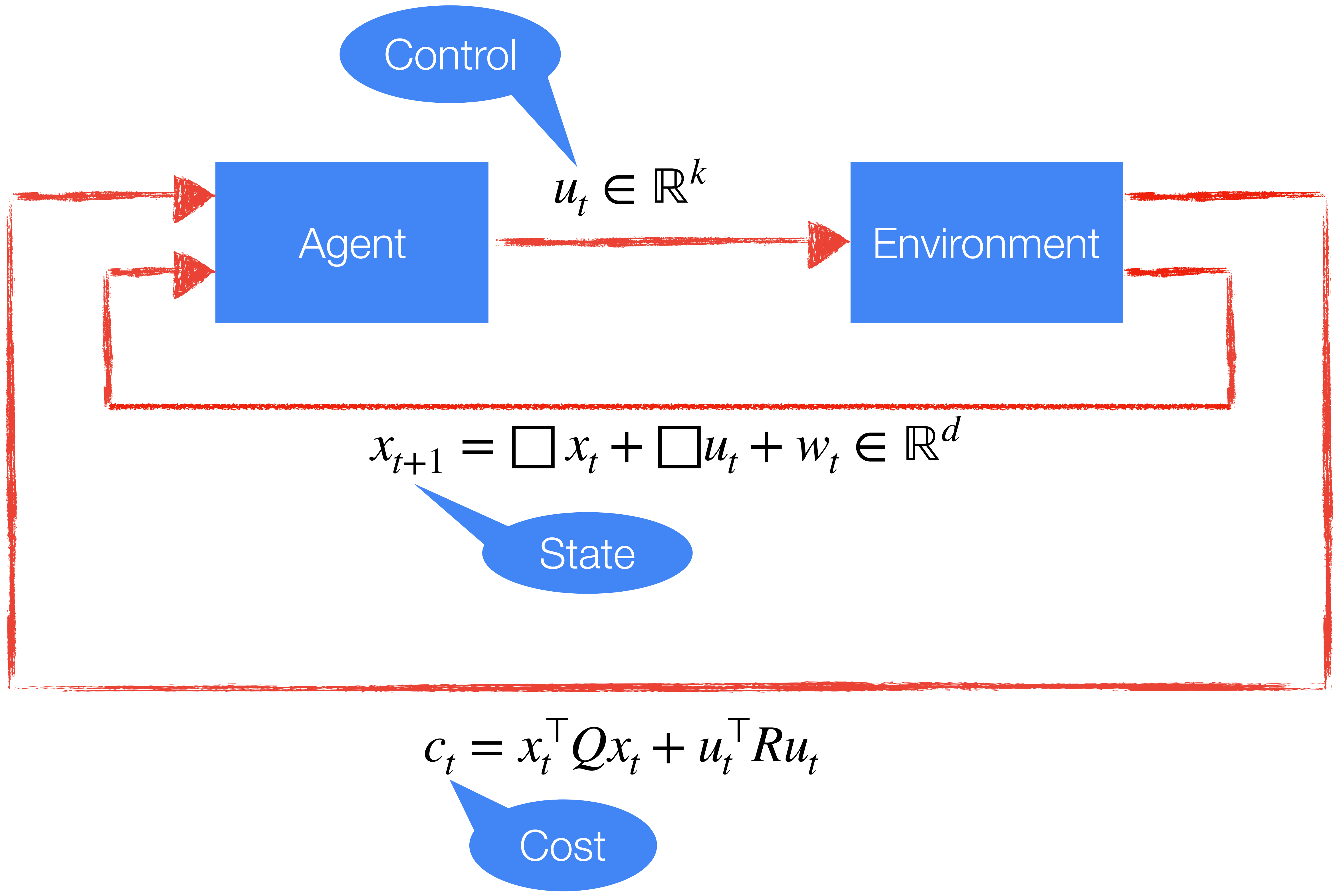
Planning in LQRs



- Policy: $\pi : x_t \longmapsto u_t$
- Optimal policy stabilizes the system in minimum cost.
- For infinite horizon: $\pi^\star(x) = Kx$

Dimitri P. Bertsekas, Dynamic Programming and Optimal Control, 2005.

Learning in LQRs



Goal: minimize the regret

$$\mathbf{R}_T = \sum_{t=1}^{T} \mathbf{cost}_t(Alg) - \min_K \sum_{t=1}^{T} \mathbf{cost}_t(K)$$

Abbasi-Yadkori and Szepesvári, 2011
Ibrahimi et al., 2012
Faradonbeh et al., 2017
Ouyang et al., 2017
Abeille and Lazaric, 2017, 2018
Dean et al. 2018, 2019

Our Result

- **First poly-time** algorithm for online learning of linear-quadratic control systems with $\widetilde{O}(\sqrt{T})$ regret.

- Resolve an open question of Abbasi-Yadkori and Szepesvári (2011) and Dean, Mania, Matni, Recht, and Tu (2018).

	Regret	Efficient

Our Result

- **First poly-time** algorithm for online learning of linear-quadratic control systems with $\widetilde{O}(\sqrt{T})$ regret.
- Resolve an open question of Abbasi-Yadkori and Szepesvári (2011) and Dean, Mania, Matni, Recht, and Tu (2018).

	Regret	Efficient
Abbasi-Yadkori and Szepesvári, 2011	$\mathbf{exp}(d)\sqrt{T}$	✘

Our Result

- **First poly-time** algorithm for online learning of linear-quadratic control systems with $\widetilde{O}(\sqrt{T})$ regret.

- Resolve an open question of Abbasi-Yadkori and Szepesvári (2011) and Dean, Mania, Matni, Recht, and Tu (2018).

	Regret	**Efficient**
Abbasi-Yadkori and Szepesvári, 2011	$\mathbf{exp}(d)\sqrt{T}$	✗
Ibrahimi et al., 2012	$\mathbf{poly}(d)\sqrt{T}$	✗

Our Result

- **First poly-time** algorithm for online learning of linear-quadratic control systems with $\widetilde{O}(\sqrt{T})$ regret.

- Resolve an open question of Abbasi-Yadkori and Szepesvári (2011) and Dean, Mania, Matni, Recht, and Tu (2018).

	Regret	Efficient
Abbasi-Yadkori and Szepesvári, 2011	**exp**$(d)\sqrt{T}$	✗
Ibrahimi et al., 2012	**poly**$(d)\sqrt{T}$	✗
Dean et al., 2018	**poly**$(d)T^{2/3}$	✓

Our Result

- **First poly-time** algorithm for online learning of linear-quadratic control systems with $\widetilde{O}(\sqrt{T})$ regret.

- Resolve an open question of Abbasi-Yadkori and Szepesvári (2011) and Dean, Mania, Matni, Recht, and Tu (2018).

	Regret	Efficient
Abbasi-Yadkori and Szepesvári, 2011	**exp**$(d)\sqrt{T}$	✗
Ibrahimi et al., 2012	**poly**$(d)\sqrt{T}$	✗
Dean et al., 2018	**poly**$(d)T^{2/3}$	✓
Ours	**poly**$(d)\sqrt{T}$	✓

Our Result

- **First poly-time** algorithm for online learning of linear-quadratic control systems with $\widetilde{O}(\sqrt{T})$ regret.

- Resolve an open question of Abbasi-Yadkori and Szepesvári (2011) and Dean, Mania, Matni, Recht, and Tu (2018).

	Regret	Efficient
Abbasi-Yadkori and Szepesvári, 2011	$\textbf{exp}(d)\sqrt{T}$	✗
Ibrahimi et al., 2012	$\textbf{poly}(d)\sqrt{T}$	✗
Dean et al., 2018	$\textbf{poly}(d)T^{2/3}$	✓
Ours	$\textbf{poly}(d)\sqrt{T}$	✓

* Recent paper by Mania et al., 2019 can be used to derive a result similar to ours.

Solution Techniques

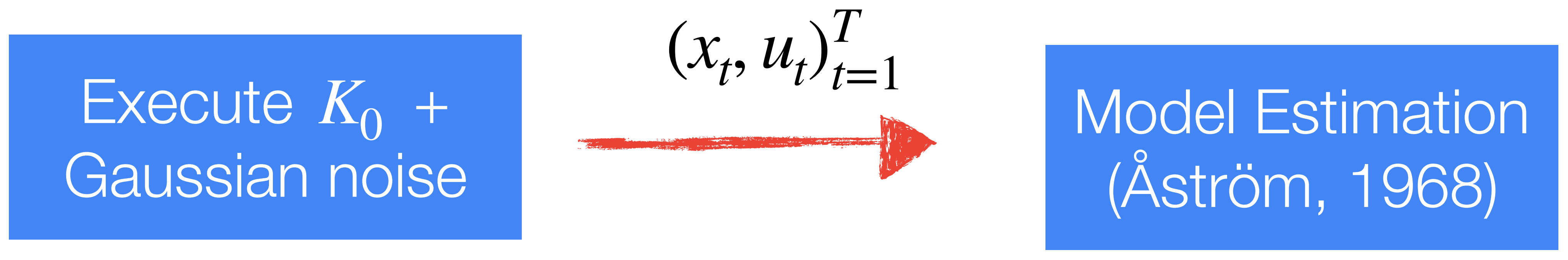
Explore-then-Exploit (Dean et al., 2018)

Execute K_0 +
Gaussian noise

$$u_t = K_0 x_t + \mathcal{N}\left(0, \varepsilon^2 I\right)$$

Solution Techniques

Explore-then-Exploit (Dean et al., 2018)

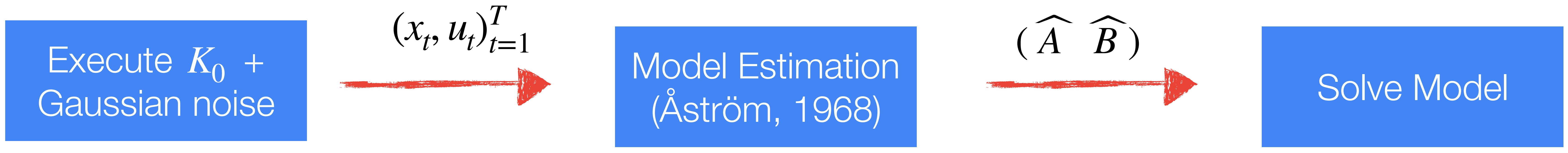


$$u_t = K_0 x_t + \mathcal{N}\left(0, \varepsilon^2 I\right)$$

$$(\hat{A}\ \hat{B}) = \arg\min_{(A\ B)} \sum_{t=1}^{T} \|A x_t + B u_t - x_{t+1}\|^2$$

Solution Techniques

Explore-then-Exploit (Dean et al., 2018)



$$u_t = K_0 x_t + \mathcal{N}\left(0, \varepsilon^2 I\right)$$

$$(\hat{A}\ \hat{B}) = \arg\min_{(A\ B)} \sum_{t=1}^{T} \|Ax_t + Bu_t - x_{t+1}\|^2$$

Solution Techniques

Explore-then-Exploit (Dean et al., 2018)

Execute K_0 + Gaussian noise → $(x_t, u_t)_{t=1}^{T}$ → Model Estimation (Åström, 1968) → $(\widehat{A}\ \ \widehat{B})$ → Solve Model → $\widehat{K}$ → Execute

$$u_t = K_0 x_t + \mathcal{N}\left(0, \varepsilon^2 I\right)$$

$$(\hat{A}\ \hat{B}) = \arg\min_{(A\ B)} \sum_{t=1}^{T} \|Ax_t + Bu_t - x_{t+1}\|^2$$

$$\mathbf{R}_T = O\left(T^{2/3}\right)$$

Solution Techniques

Based on UCRL

Optimism in the Face of Uncertainty (Abbasi-Yadkori and Szepesvári, 2011)

$\Theta_t \ni (A_\star \; B_\star)$

Solution Techniques

Based on UCRL

Optimism in the Face of Uncertainty (Abbasi-Yadkori and Szepesvári, 2011)

$\Theta_t \ni (A_\star \; B_\star)$

Find Optimistic Policy

$$\pi_t = \arg \min_{\pi,\ (A\ B) \in \Theta_t} J_{(A\ B)}(\pi)$$

Solution Techniques

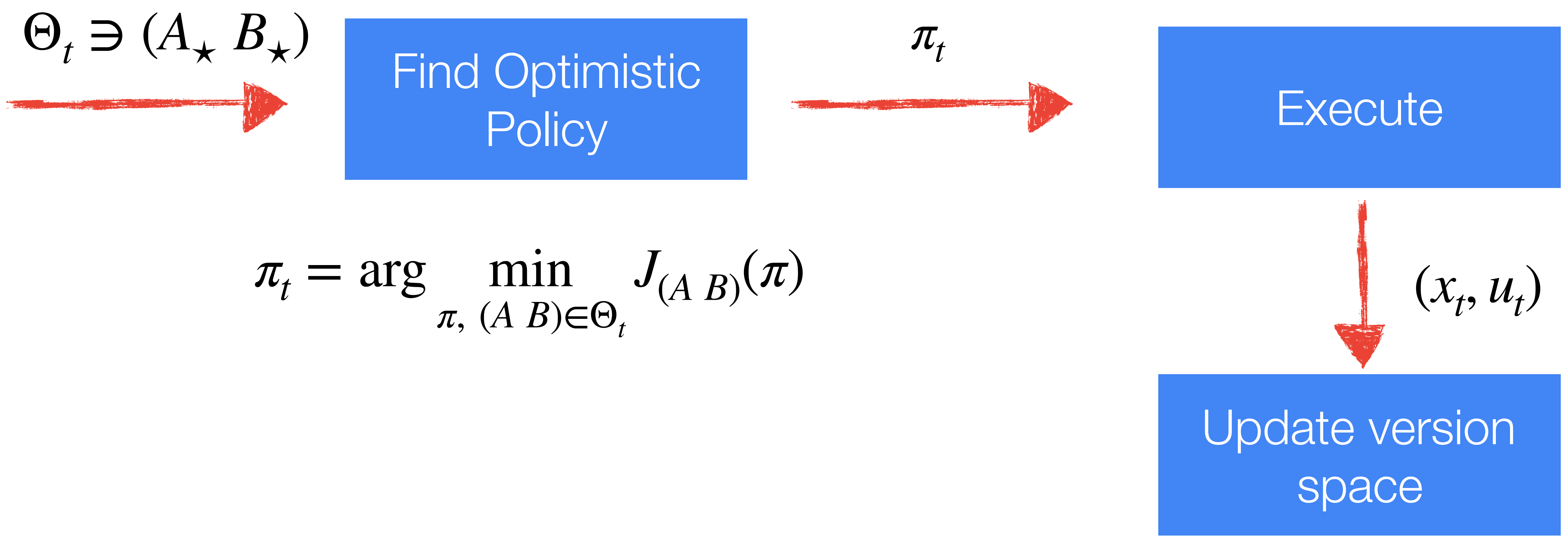
Optimism in the Face of Uncertainty (Abbasi-Yadkori and Szepesvári, 2011)

$$\pi_t = \arg \min_{\pi,\ (A\ B) \in \Theta_t} J_{(A\ B)}(\pi)$$

Solution Techniques

Based on UCRL

Optimism in the Face of Uncertainty (Abbasi-Yadkori and Szepesvári, 2011)

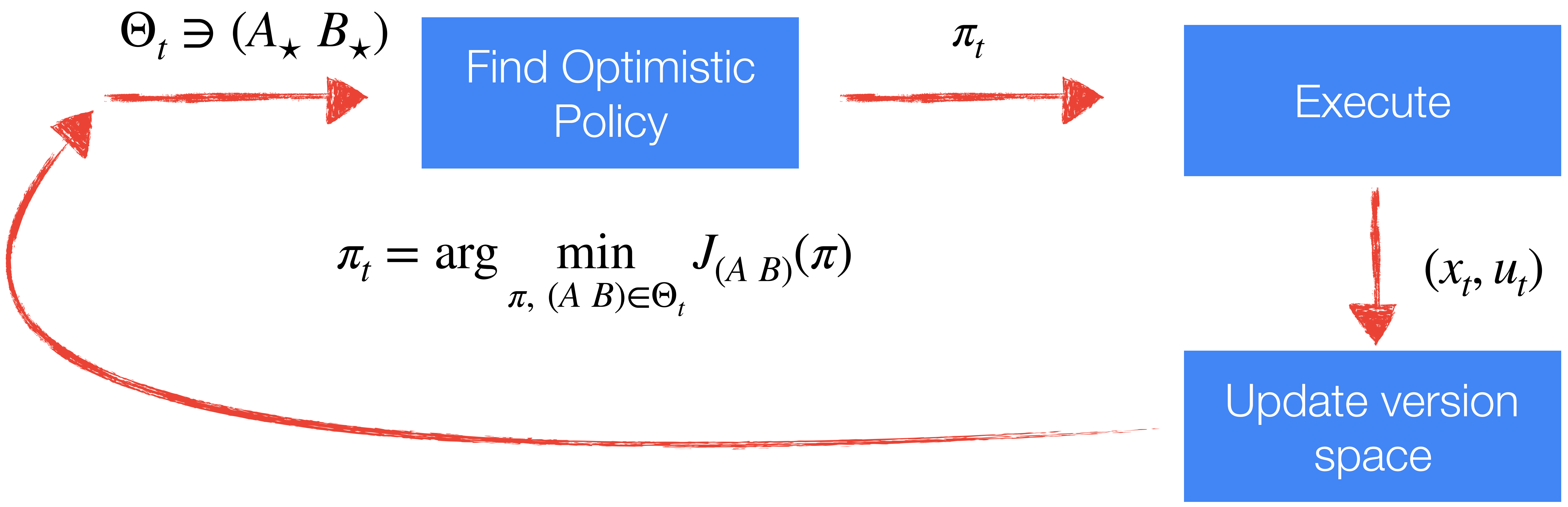


$$\pi_t = \arg \min_{\pi,\ (A\ B) \in \Theta_t} J_{(A\ B)}(\pi)$$

Solution Techniques

Based on UCRL

Optimism in the Face of Uncertainty (Abbasi-Yadkori and Szepesvári, 2011)



$$\pi_t = \arg \min_{\pi,\ (A\ B) \in \Theta_t} J_{(A\ B)}(\pi)$$

Optimistic in the sense that:

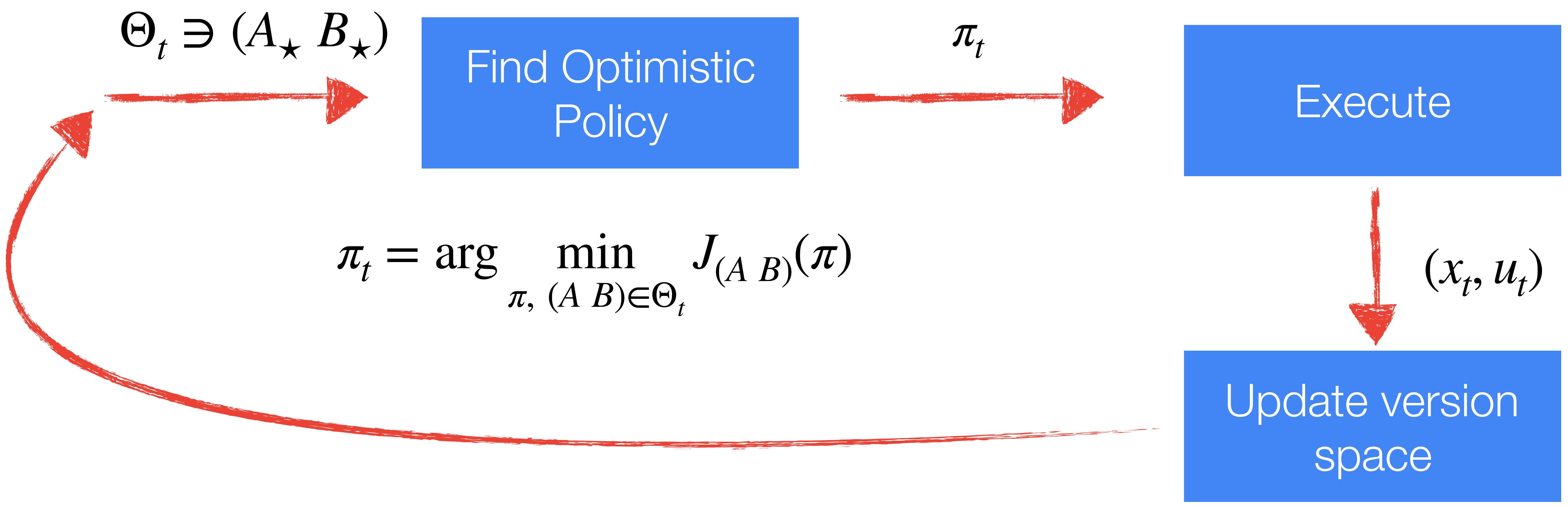
$$\min_{\pi,\ (A\ B) \in \Theta_t} J_{(A\ B)}(\pi) \leq J(\pi^\star)\,.$$

$$\mathbf{R}_T = O\left(\sqrt{T}\right)$$

Solution Techniques

Based on UCRL

Optimism in the Face of Uncertainty (Abbasi-Yadkori and Szepesvári, 2011)



$$\pi_t = \arg \min_{\pi,\ (A\ B) \in \Theta_t} J_{(A\ B)}(\pi)$$

$$\mathbf{R}_T = O(\sqrt{T})$$

Optimistic in the sense that:

$$\min_{\pi,\ (A\ B) \in \Theta_t} J_{(A\ B)}(\pi) \leq J(\pi^\star)\,.$$

Caveat: $J_{(A\ B)}(\pi)$ not convex in policy parameters.

Convex (SDP) Formulation

Cohen et al., 2018

Convex re-parameterization: $$\Sigma = \mathbb{E}\left[\begin{pmatrix} x \\ u \end{pmatrix}\begin{pmatrix} x \\ u \end{pmatrix}^\top\right].$$

Steady-state covariance matrix

$$\Sigma = \begin{pmatrix} \Sigma_{xx} & \Sigma_{xu} \\ \Sigma_{ux} & \Sigma_{uu} \end{pmatrix}$$

LQ Control:

$$x_{t+1} = A_\star x_t + B_\star u_t + w_t$$
$$c_t = x_t^\top Q x_t + u_t^\top R u_t$$

Convex (SDP) Formulation

Cohen et al., 2018

Convex re-parameterization: $\Sigma = \mathbb{E}\left[\begin{pmatrix} x \\ u \end{pmatrix}\begin{pmatrix} x \\ u \end{pmatrix}^\top\right].$

Steady-state covariance matrix

$$\min_{\Sigma \succeq 0} \quad \Sigma \bullet \begin{pmatrix} Q & 0 \\ 0 & R \end{pmatrix}$$

s.t. $\Sigma_{xx} = (A_\star \; B_\star)\,\Sigma\,(A_\star \; B_\star)^\top + W.$

Lemma: $K = \Sigma_{ux}\Sigma_{xx}^{-1}$ is optimal for LQR.

$\Sigma = \begin{pmatrix} \Sigma_{xx} & \Sigma_{xu} \\ \Sigma_{ux} & \Sigma_{uu} \end{pmatrix}$

LQ Control:

$x_{t+1} = A_\star x_t + B_\star u_t + w_t$

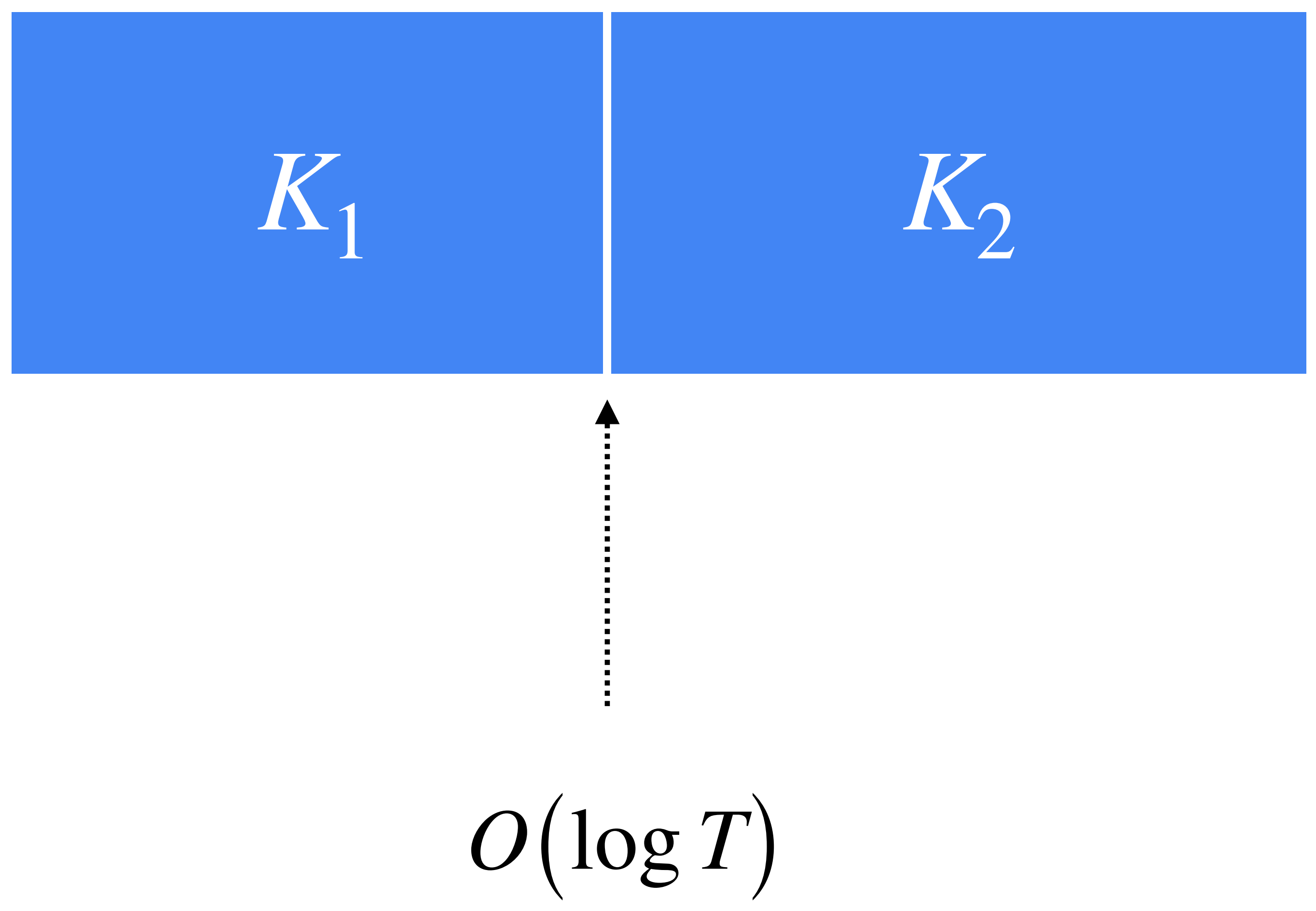
$c_t = x_t^\top Q x_t + u_t^\top R u_t$

Intuition for Our Algorithm

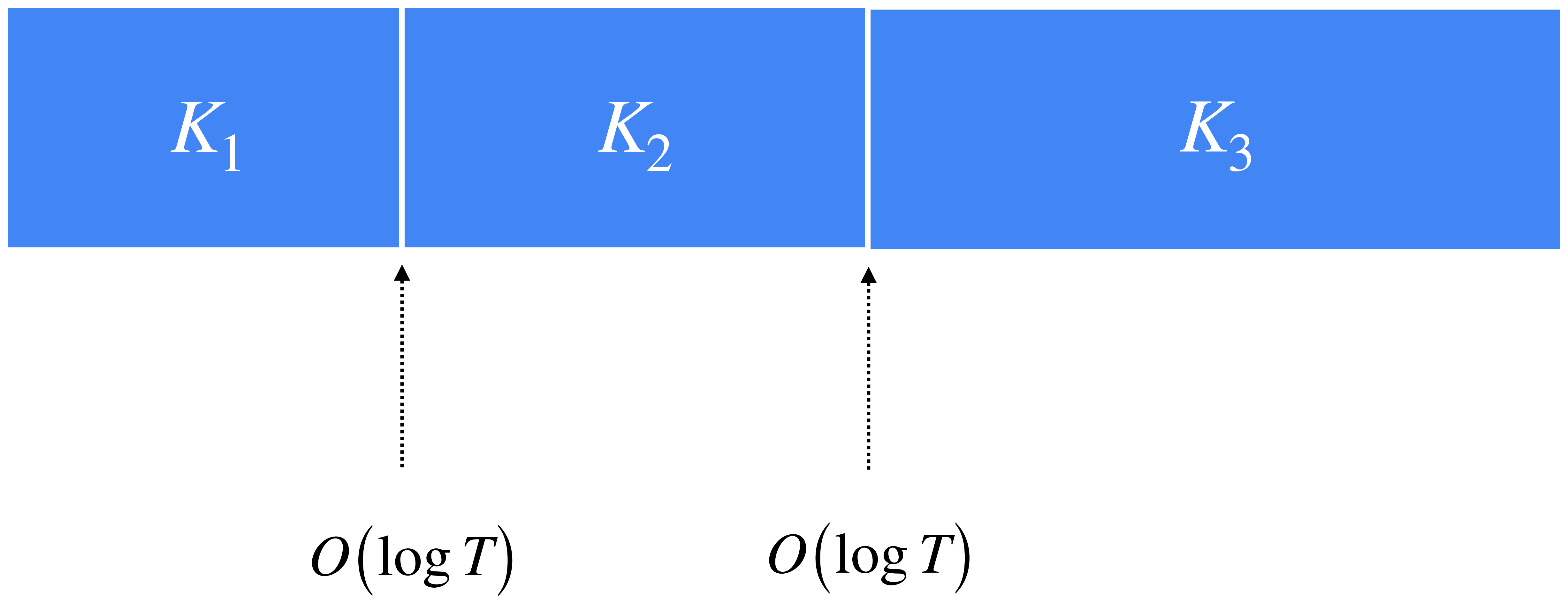
Intuition for Our Algorithm

K_1

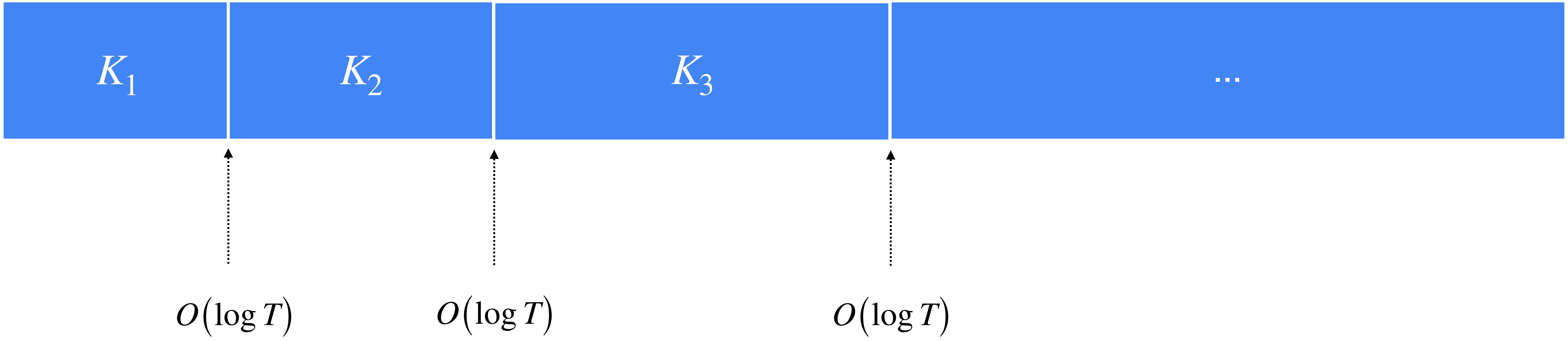
Intuition for Our Algorithm



Intuition for Our Algorithm



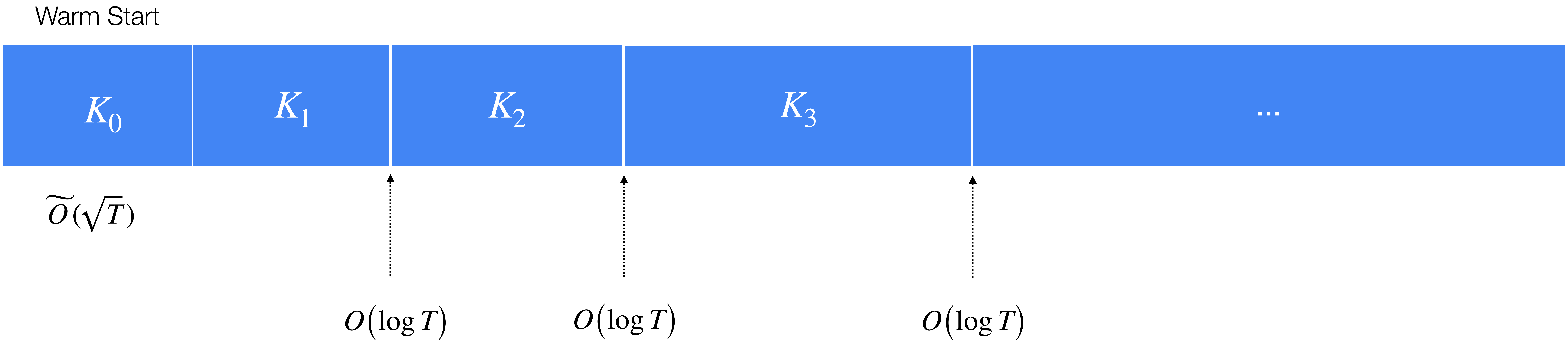
Intuition for Our Algorithm



$O(\log T)$ epochs with high probability.

$\widetilde{O}(\sqrt{T})$ regret in total.

Intuition for Our Algorithm



$O(\log T)$ epochs with high probability.

$\widetilde{O}(\sqrt{T})$ regret in total.

Our Algorithm: OSLO (i)

- After warm start: $\|(A_0\ B_0) - (A_\star\ B_\star)\|_F^2 \le O\left(1/\sqrt{T}\right)$.
- Maintain: $V_t = \lambda I + \frac{1}{\beta}\sum_{s=1}^{t-1} z_s z_s^\top$, where $z_s = \begin{pmatrix} x_s \\ u_s \end{pmatrix}$.
- Run in epochs:
  - Compute K_t using a semidefinite program.
  - Execute fixed K_t during epoch.
  - Epoch ends when $\det(V_t)$ is doubled.

Optimistic

Our Algorithm: OSLO (ii)

At epoch start:

- Estimate $A_\star, B_\star$ from past observations

$$(A_t \ B_t) = \arg\min_{(A\ B)} \frac{1}{\beta} \sum_{s=1}^{t-1} \|(A\ B)z_s - x_{s+1}\|^2 + \lambda \|(A\ B) - (A_0\ B_0)\|_F^2$$

- Compute optimistic policy by solving

$$\Sigma_t = \arg\min_{\Sigma \succeq 0} \quad \Sigma \bullet \begin{pmatrix} Q & 0 \\ 0 & R \end{pmatrix}$$

$$\textbf{s.t.} \quad \Sigma_{xx} \succeq (A_t\ B_t)\Sigma(A_t\ B_t)^\top + W - \mu(\Sigma \bullet V_t^{-1})I$$

Replaces hard problem in Abbasi-Yadkori & Szepesvári

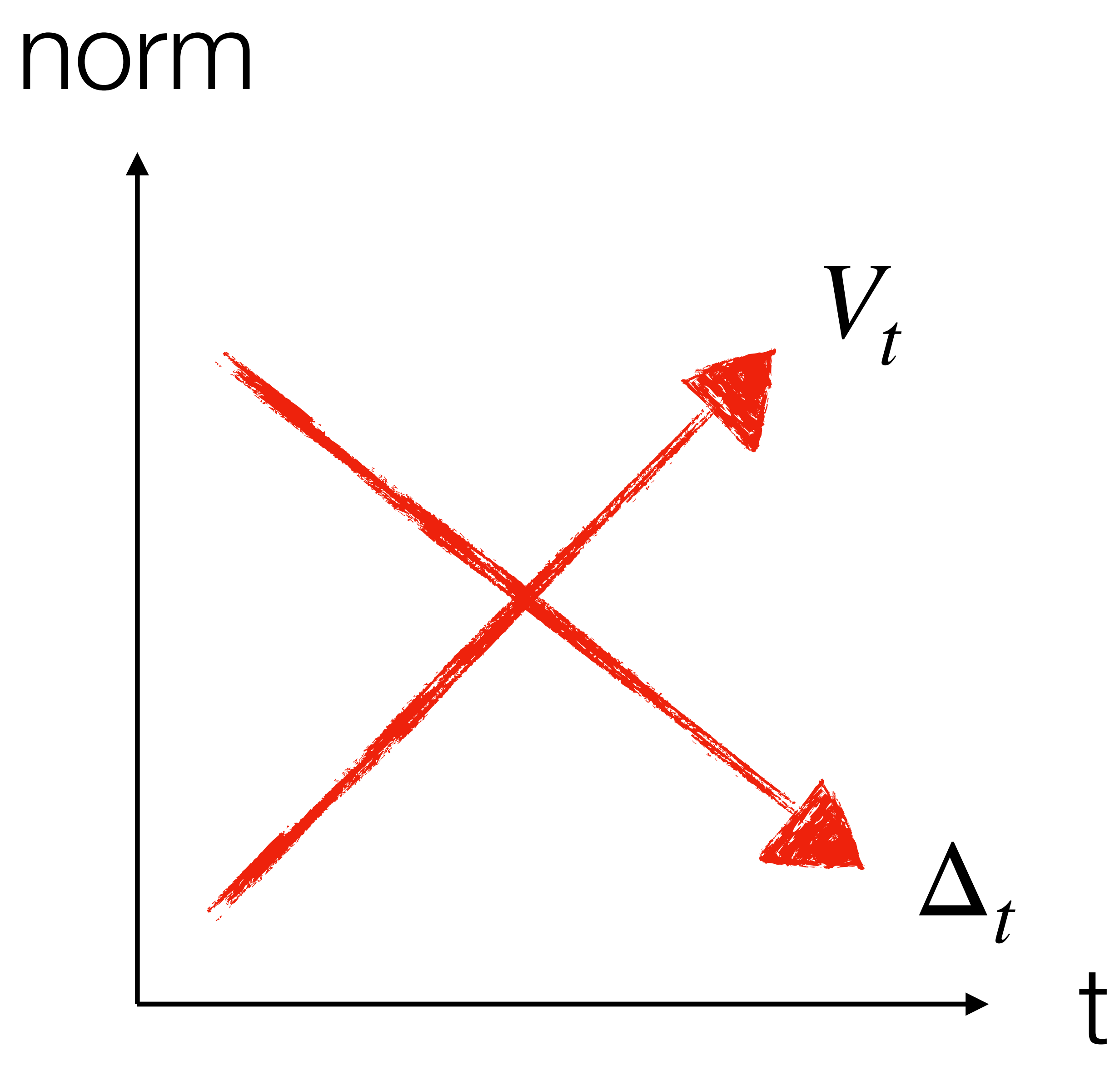
- Output: $K_t = (\Sigma_t)_{ux}(\Sigma_t)_{xx}^{-1}$

$$\Sigma = \begin{pmatrix} \Sigma_{xx} & \Sigma_{xu} \\ \Sigma_{ux} & \Sigma_{uu} \end{pmatrix}$$

Parameter Estimation

Lemma (Abbasi-Yadkori and Szepesvari, 2011)

Let $\Delta_t = (A_t\ B_t) - (A_\star\ B_\star)$. With high probability $\mathbf{tr}\left(\Delta_t V_t \Delta_t^\top\right) \leq 1$.



$$\|V_t\| = \Theta(t)$$

$$\|\Delta_t\| = \Theta(1/\sqrt{t})$$

"Almost" the regret $= \sum_{t=1}^{T} \|\Delta_t\| = O\left(\sqrt{T}\right)$ (disregarding switches and warm start)

MDP vs. LQR: Boundedness of States

- Unlike in MDPs states may be unbounded.
  - Low probability if K is stable, but may have unpredictable effect on expectation.
  - System may destabilize when switching between policies too often.
- Main technique:
  - Generate “sequentially stable” policies.
  - Keep states bounded with high probability: $||x_t|| \lessapprox \frac{\kappa}{\gamma}\sqrt{d \log T}$ **w.h.p**

Summary

- First efficient algorithm for learning LQRs with $\widetilde{O}(\sqrt{T})$ regret.
- Solved open problem.
- Shown connection between MAB, RL, control and convex optimization.
- Open Problems:
  - No lower bound!
  - Evidence that the correct rate is $O\left(\log T\right)$ (Mania et al., 2019) .

Thank You!

Poster #159