# Reinforcement Learning Policies with local LQR guarantees for Nonlinear Discrete-Time Systems

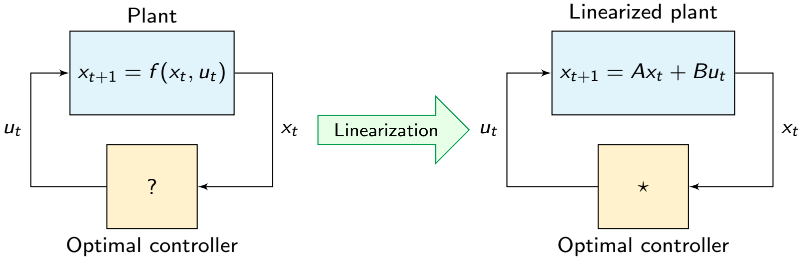
Samuele Zoboli Vincent Andrieu Daniele Astolfi
Giacomo Casadei Jilles S. Dibangoye Madiha Nadri

ANR DeLiCio Project

June 10, 2021

# Framework

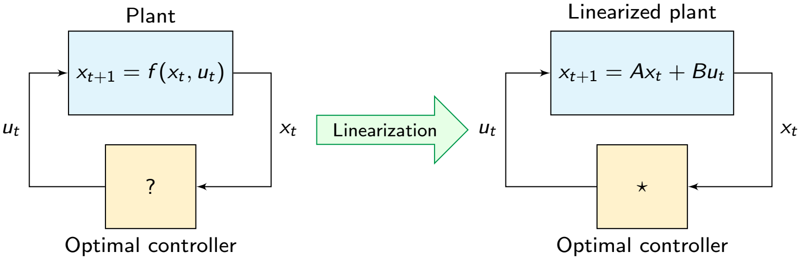
**Control theory (CT):**



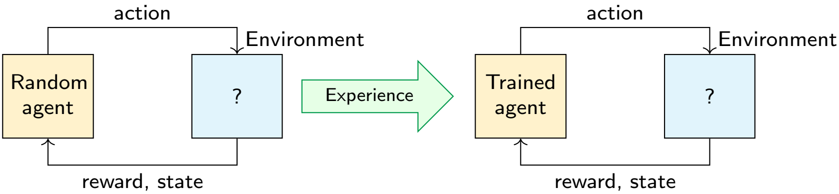

**Reinforcement learning (RL):**

# Framework

**Control theory (CT):**




**Reinforcement learning (RL):**



# Framework

**CT method:**

Pros
- ✓ Stabilizing
- ✓ Guaranteed performances
- ✓ Model based

Cons
- ✗ Local
- ✗ Conservative
- ✗ Model based




**RL method:**

Pros
- ✓ Arbitrary complex systems
- ✓ Arbitrarily large DOA
- ✓ Data driven (model-free)

Cons
- ✗ No guarantees
- ✗ Reward focused
- ✗ Data driven (model-free)

# Framework

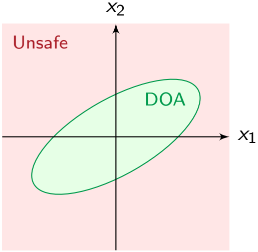
**CT method:**

Pros

- ✓ Stabilizing
- ✓ Guaranteed performances
- ✓ Model based

Cons

- ✗ Local
- ✗ Conservative
- ✗ Model based




**RL method:**

Pros

- ✓ Arbitrary complex systems
- ✓ Arbitrarily large DOA
- ✓ Data driven (model-free)

Cons

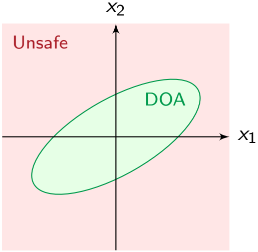
- ✗ No guarantees
- ✗ Reward focused
- ✗ Data driven (model-free)

# Framework

**CT method:**

Pros
- ✓ Stabilizing
- ✓ Guaranteed performances
- ✓ Model based

Cons
- ✗ Local
- ✗ Conservative
- ✗ Model based




**RL method:**

Pros
- ✓ Arbitrary complex systems
- ✓ Arbitrarily large DOA
- ✓ Data driven (model-free)

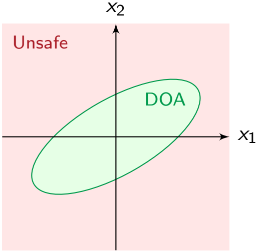
Cons
- ✗ No guarantees
- ✗ Reward focused
- ✗ Data driven (model-free)

# Framework

**CT method:**

Pros
- ✓ Stabilizing
- ✓ Guaranteed performances
- ✓ Model based

Cons
- ✗ Local
- ✗ Conservative
- ✗ Model based

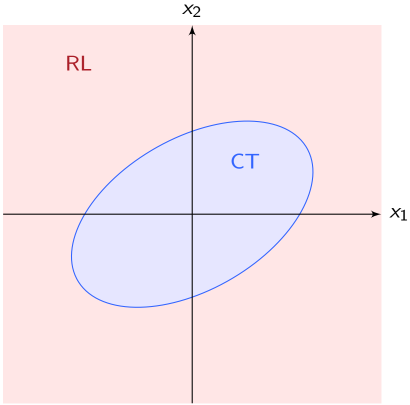



**RL method:**

Pros
- ✓ Arbitrary complex systems
- ✓ Arbitrarily large DOA
- ✓ Data driven (model-free)

Cons
- ✗ No guarantees
- ✗ Reward focused
- ✗ Data driven (model-free)

# Objective




**Merged controller**:

- ✓ "Global" learnt nonlinear controller (**RL**)
- ✓ Local guarantees (**CT**)

# Table of Contents

# Table of Contents

# Undiscounted vs discounted LQR

**Undiscounted**

Cost function:

$$J(x,u) = \sum_{t=0}^{\infty} x_t^\top Q x_t + u_t^\top R u_t,$$

Optimal solution (linear system):

$$u^\star(x_t) = K^\star x_t,$$

$$K^\star = -(R + B^\top P B)^{-1} B^\top P A$$

$P$ solution of DARE[1]

Stability: yes

**Discounted**

Cost function:

$$J_\gamma(x,u) = \sum_{t=0}^{\infty} \gamma^t (x_t^\top Q_\gamma x_t + u_t^\top R_\gamma u_t),$$

$\gamma \in (0, 1]$

Optimal solution (linear system):

$$u_\gamma^\star(x_t) = K_\gamma^\star x_t,$$

$$K_\gamma^\star = -\gamma(R_\gamma + \gamma B^\top P_\gamma B)^{-1} B^\top P_\gamma A$$

$P_\gamma$ solution of discounted DARE[2]

Stability: dependent on $\gamma$ [Postoyan et al.(2016)]

[1] Discrete-time Algebraic Riccati Equation: $P = A^\top P A - A^\top P B (R + B^\top P B)^{-1} + Q$

[2] Discounted DARE: $Q_\gamma + \gamma A^\top (P_\gamma - \gamma P_\gamma B (R_\gamma + \gamma B^\top P_\gamma B)^{-1} B^\top P_\gamma) A$

# Undiscounted vs discounted LQR

## Undiscounted

Cost function:

$$J(x,u) = \sum_{t=0}^{\infty} x_t^\top Q x_t + u_t^\top R u_t,$$

Optimal solution (linear system):

$$u^\star(x_t) = K^\star x_t,$$

$$K^\star = -(R + B^\top P B)^{-1} B^\top P A$$

$P$ solution of DARE[1]

Stability: yes

## Discounted

Cost function:

$$J_\gamma(x,u) = \sum_{t=0}^{\infty} \gamma^t (x_t^\top Q_\gamma x_t + u_t^\top R_\gamma u_t),$$

$\gamma \in (0,1]$

Optimal solution (linear system):

$$u_\gamma^\star(x_t) = K_\gamma^\star x_t,$$

$$K_\gamma^\star = -\gamma(R_\gamma + \gamma B^\top P_\gamma B)^{-1} B^\top P_\gamma A$$

$P_\gamma$ solution of discounted DARE[2]

Stability: dependent on $\gamma$ [Postoyan et al.(2016)]

[1] Discrete-time Algebraic Riccati Equation: $P = A^\top P A - A^\top P B(R + B^\top P B)^{-1} + Q$

[2] Discounted DARE: $Q_\gamma + \gamma A^\top (P_\gamma - \gamma P_\gamma B(R_\gamma + \gamma B^\top P_\gamma B)^{-1} B^\top P_\gamma) A$

# Undiscounted vs discounted LQR

**Undiscounted**

Cost function:

$$\boldsymbol{J}(x,u) = \sum_{t=0}^{\infty} x_t^\top Q x_t + u_t^\top R u_t,$$

Optimal solution (linear system):

$$u^\star(x_t) = \boldsymbol{K}^\star x_t,$$
$$\boldsymbol{K}^\star = -(R + B^\top P B)^{-1} B^\top P A$$

$P$ solution of DARE[1]

Stability: yes

**Discounted**

Cost function:

$$J_\gamma(x,u) = \sum_{t=0}^{\infty} \gamma^t (x_t^\top Q_\gamma x_t + u_t^\top R_\gamma u_t),$$

$\gamma \in (0,1]$

Optimal solution (linear system):

$$u_\gamma^\star(x_t) = K_\gamma^\star x_t,$$
$$K_\gamma^\star = -\gamma(R_\gamma + \gamma B^\top P_\gamma B)^{-1} B^\top P_\gamma A$$

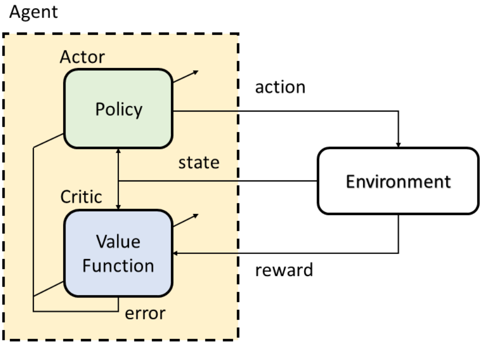
$P_\gamma$ solution of discounted DARE[2]

Stability: dependent on $\gamma$ [Postoyan et al.(2016)]

[1] Discrete-time Algebraic Riccati Equation: $P = A^\top PA - A^\top PB(R + B^\top PB)^{-1} + Q$

[2] Discounted DARE: $Q_\gamma + \gamma A^\top (P_\gamma - \gamma P_\gamma B(R_\gamma + \gamma B^\top P_\gamma B)^{-1} B^\top P_\gamma) A$

# Actor-Critic Reinforcement Learning

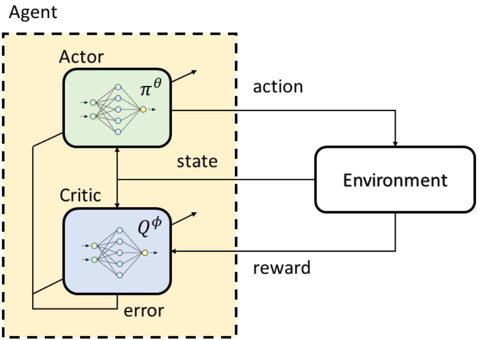



**Goal:** learn optimal policy $\pi$ through experience.

**Typically solved learning:**

- State-value function $J(x) = \sum_{t=0}^{\infty} \gamma^t r(x_t, \pi(x_t))$
- Action-value function $\mathbf{Q}(x, u) = r(x_t, u_t) + \gamma J(x_{t+1})$

# Actor-Critic Reinforcement Learning



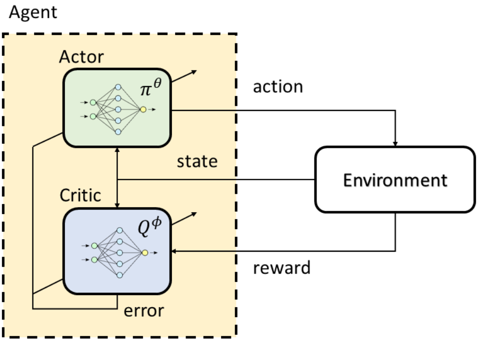

Characteristics:

- Two cost functions (actor and critic)
- Two function approximators (NNs)
- Policy Gradient methods

We focus on deterministic algorithms

- Learn parameters $\theta$ of a deterministic policy
- Learn action-value estimator parameters $\phi$

# Actor-Critic Reinforcement Learning




Characteristics:

- Two cost functions (actor and critic)
- Two function approximators (NNs)
- Policy Gradient methods

We focus on deterministic algorithms

- Learn parameters $\theta$ of a deterministic policy
- Learn action-value estimator parameters $\phi$

# Table of Contents

# What do we start with?

Consider a deterministic discrete-time nonlinear system

$$x_{t+1} = f(x_t, u_t), \qquad x_t \in \mathbb{R}^n, u_t \in \mathcal{U} \subseteq \mathbb{R}$$

Assume we know

- Linearization of the system

$$x_{t+1} = Ax_t + Bu_t$$

- Cost function (infinite horizon)

$$J = \sum_{t=0}^{\infty} x_t^\top Q x_t + u_t^\top R u_t$$

- Optimal LQR gain $K^\star$

$$u_t^\star = K^\star x_t, \quad K^\star = -(R + B^\top P B)^{-1} B^\top P A,$$

# What do we start with?

Consider a deterministic discrete-time nonlinear system

$$x_{t+1} = f(x_t, u_t), \qquad x_t \in \mathbb{R}^n, u_t \in \mathcal{U} \subseteq \mathbb{R}$$

Assume we know

- Linearization of the system

$$x_{t+1} = Ax_t + Bu_t$$

- Cost function (infinite horizon)

$$J = \sum_{t=0}^{\infty} x_t^\top Q x_t + u_t^\top R u_t$$

- Optimal LQR gain $K^\star$

$$u_t^\star = K^\star x_t, \quad K^\star = -(R + B^\top P B)^{-1} B^\top P A,$$

# What do we start with?

Consider a deterministic discrete-time nonlinear system

$$x_{t+1} = f(x_t, u_t), \qquad x_t \in \mathbb{R}^n, u_t \in \mathcal{U} \subseteq \mathbb{R}$$

Assume we know

- Linearization of the system

$$x_{t+1} = Ax_t + Bu_t$$

- Cost function (infinite horizon)

$$\boldsymbol{J} = \sum_{t=0}^{\infty} x_t^\top Q x_t + u_t^\top R u_t$$

- Optimal LQR gain $\boldsymbol{K}^\star$

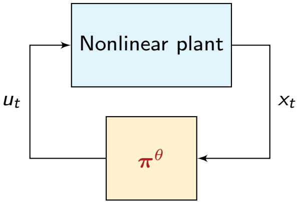
$$u_t^\star = \boldsymbol{K}^\star x_t, \quad \boldsymbol{K}^\star = -(R + B^\top PB)^{-1} B^\top PA,$$

# What do we look for?

## Goal

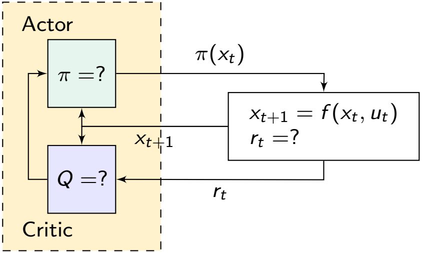
Learn an optimal parametrized control policy $\boldsymbol{\pi}^\theta : \mathbb{R}^n \times \mathbb{R}^p \to \mathcal{U} \subseteq \mathbb{R}$ with parameters $\theta \in \mathbb{R}^p$ such that the origin of the closed-loop nonlinear system is LAS for any $\theta$, namely

$$\frac{\partial \boldsymbol{\pi}^\theta}{\partial x}(0) = \boldsymbol{K}^\star, \forall \theta \in \mathbb{R}^p \tag{1}$$



# Table of Contents

# What do we need?




**Questions**:

- What should the reward be?
- How to structure the policy $\pi$?
- How to estimate the value function $Q$?

# Reward shaping

**RL** requires finite value functions → undiscounted cost functions may not be suitable

$$x_{t+1} = Ax_t + Bu_t$$
$$J = \sum_t^\infty x_t^T Q x_t + u_t^T R u_t$$
$$u = K^\star x_t$$

Undiscounted LQR

**?**

$$x_{t+1} = f(x_t, u_t)$$
$$J = \sum_t^\infty \gamma^t r(x_t, \mu_t) =?$$
$$\pi^\theta =?$$

RL

# From undiscounted to discounted LQR

## Lemma

For any $\gamma \in (0,1]$, the optimal gain $K^\star$ is the optimal solution of the discounted problem $J_\gamma = \sum_t^\infty \gamma^t (x_t^\top Q_\gamma x_t + u_t^\top R_\gamma u_t)$ with $Q_\gamma$, $R_\gamma$ defined as

$$Q_\gamma = \gamma Q + (1-\gamma)P, \qquad R_\gamma = \gamma R, \quad P \text{ solution of DARE.} \tag{2}$$

Moreover, the state-value function $J(x) = \sum_{t=0}^\infty \gamma^t (x_t^\top Q_\gamma x_t + x_t^\top K^{\star\top} R_\gamma K^\star x_t)$ is finite.

Why is it interesting?

- $J_\gamma$ includes the discount factor $\gamma$
- Stability independent from $\gamma$

# From undiscounted to discounted LQR

## Lemma

For any $\gamma \in (0,1]$, the optimal gain $K^\star$ is the optimal solution of the discounted problem $J_\gamma = \sum_t^\infty \gamma^t (x_t^\top Q_\gamma x_t + u_t^\top R_\gamma u_t)$ with $Q_\gamma$, $R_\gamma$ defined as

$$Q_\gamma = \gamma Q + (1-\gamma)P, \qquad R_\gamma = \gamma R, \quad P \text{ solution of DARE.} \tag{2}$$

Moreover, the state-value function $J(x) = \sum_{t=0}^\infty \gamma^t (x_t^\top Q_\gamma x_t + x_t^\top K^{\star\top} R_\gamma K^\star x_t)$ is finite.

Why is it interesting?

- $J_\gamma$ includes the discount factor $\gamma$
- Stability independent from $\gamma$

# From undiscounted to discounted LQR

## Lemma

For any $\gamma \in (0,1]$, the optimal gain $K^\star$ is the optimal solution of the discounted problem $J_\gamma = \sum_t^\infty \gamma^t (x_t^\top Q_\gamma x_t + u_t^\top R_\gamma u_t)$ with $Q_\gamma$, $R_\gamma$ defined as

$$Q_\gamma = \gamma Q + (1-\gamma)P, \qquad R_\gamma = \gamma R, \quad P \text{ solution of DARE.} \tag{2}$$

Moreover, the state-value function $J(x) = \sum_{t=0}^\infty \gamma^t (x_t^\top Q_\gamma x_t + x_t^\top K^{\star\top} R_\gamma K^\star x_t)$ is finite.

Why is it interesting?

- $J_\gamma$ includes the discount factor $\gamma$
- Stability independent from $\gamma$

# Matching objectives

**RL** requires finite value functions → associated discounted problem

$$x_{t+1} = Ax_t + Bu_t$$
$$\boldsymbol{J} = \sum_t x_t^\top Q x_t + u_t^\top R u_t$$
$$u = \boldsymbol{K}^\star x_t$$

Undiscounted LQR

↓

$$x_{t+1} = Ax_t + Bu_t$$
$$J_\gamma = \sum_t \gamma^t (x_t^\top \boldsymbol{Q}_\gamma x_t + u_t^\top \boldsymbol{R}_\gamma u_t)$$
$$u = \boldsymbol{K}^\star x_t$$

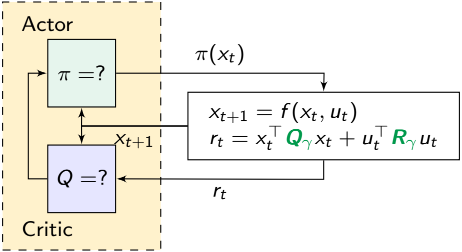
Discounted LQR

↓

$$x_{t+1} = f(x_t, u_t)$$
$$\boldsymbol{J} = \sum_t \gamma^t (x_t^\top \boldsymbol{Q}_\gamma x_t + u_t^\top \boldsymbol{R}_\gamma u_t)$$
$$\pi^\theta = ?$$

RL

# What do we need?




**Questions**:

- How to structure the policy $\pi$?
- How to estimate the value function $Q$?

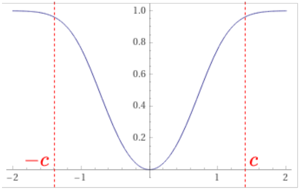
# Control policy structure

Define

$$h(x) = \tanh\left(\alpha \frac{x^\top P x}{c}\right)$$

$$\downarrow$$

saturates at $\{V(x) = c\}$




## Controller

The proposed controller is designed as

$$\pi^\theta(x_t) = u^L(x_t) + u^\theta(x_t) \quad (3)$$

$$= K^\star x_t + h(x_t)\left(\mu^\theta(x_t) - K^\star x_t\right). \quad (4)$$

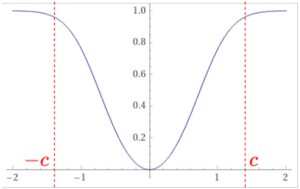
# Control policy structure

Define

$$h(x) = \tanh\left(\alpha \frac{x^\top P x}{c}\right)$$

$$\downarrow$$

saturates at $\{V(x) = c\}$




## Controller

The proposed controller is designed as

$$\pi^\theta(x_t) = u^L(x_t) + u^\theta(x_t) \quad (3)$$

$$= K^\star x_t + h(x_t)\left(\mu^\theta(x_t) - K^\star x_t\right). \quad (4)$$

# Control policy characteristics

$$\pi^\theta(x_t) = u^L(x_t) + u^\theta(x_t)$$
$$= K^\star x_t + h(x_t)\left(\mu^\theta(x_t) - K^\star x_t\right).$$

**Characteristics**:

- $u^L(x_t) \rightarrow$ given by **CT**
- Locally Lipschitz $\mu^\theta$ $\rightarrow$learnt via **RL**
- $u^\theta(x_t) \rightarrow$ higher order term
  - ✓ $\pi^\theta(0) = u^L(0)$
  - ✓ $\frac{\partial \pi^\theta}{\partial x}(0) = K^\star$
- Far from origin $\pi^\theta = \mu^\theta$
- Deterministic Policy Gradient [Silver et al.(2014)]:

$$\Delta\theta \propto \nabla_\theta \pi^\theta(x) = h(x)\nabla_\theta \mu^\theta(x)$$

# Control policy characteristics

$$\begin{aligned}\pi^\theta(x_t) =& u^L(x_t) + u^\theta(x_t)\\ =& K^\star x_t + h(x_t)\left(\mu^\theta(x_t) - K^\star x_t\right).\end{aligned}$$

**Characteristics**:

- $u^L(x_t) \to$ given by **CT**
- Locally Lipschitz $\mu^\theta$ $\to$learnt via **RL**
- $u^\theta(x_t) \to$ higher order term
  - ✓ $\pi^\theta(0) = u^L(0)$
  - ✓ $\frac{\partial \pi^\theta}{\partial x}(0) = K^\star$
- Far from origin $\pi^\theta = \mu^\theta$
- Deterministic Policy Gradient [Silver et al.(2014)]:

$$\Delta\theta \propto \nabla_\theta \pi^\theta(x) = h(x)\nabla_\theta \mu^\theta(x)$$

# Control policy characteristics

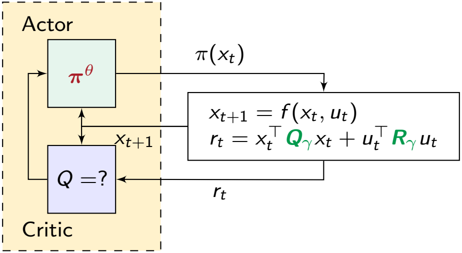
$$\begin{aligned}\boldsymbol{\pi}^{\theta}(x_t) =& \boldsymbol{u}^{L}(x_t) + \boldsymbol{u}^{\theta}(x_t) \\ =& \boldsymbol{K}^{\star} x_t + h(x_t)\left(\boldsymbol{\mu}^{\theta}(x_t) - \boldsymbol{K}^{\star} x_t\right).\end{aligned}$$

**Characteristics**:

- $\boldsymbol{u}^{L}(x_t) \rightarrow$ given by **CT**
- Locally Lipschitz $\boldsymbol{\mu}^{\theta} \rightarrow$learnt via **RL**
- $\boldsymbol{u}^{\theta}(x_t) \rightarrow$ higher order term
  - ✓ $\boldsymbol{\pi}^{\theta}(0) = \boldsymbol{u}^{L}(0)$
  - ✓ $\frac{\partial \boldsymbol{\pi}^{\theta}}{\partial x}(0) = \boldsymbol{K}^{\star}$
- Far from origin $\boldsymbol{\pi}^{\theta} = \boldsymbol{\mu}^{\theta}$
- Deterministic Policy Gradient [Silver et al.(2014)]:

$$\Delta\theta \propto \nabla_{\theta}\boldsymbol{\pi}^{\theta}(x) = h(x)\nabla_{\theta}\boldsymbol{\mu}^{\theta}(x)$$

# What do we need?




**Questions**:

- How to estimate the value function $Q$?

# Discounted LQR value functions

Linear system + optimal discounted LQ controller

- State-value function [Bertsekas et al.(1987)]

$$J^\star_\gamma(x) = x_t^T P_\gamma x_t, \quad P_\gamma \text{ solution of discounted DARE.}$$

- Action-value function [Bradtke et al.(1993)]

$$\mathbf{Q}^\star_\gamma(x,u) = z_t^T \begin{pmatrix} Q_\gamma + \gamma A^\top P_\gamma A & \gamma A^\top P_\gamma B \\ \gamma B^\top P_\gamma A & R_\gamma + \gamma B^\top P_\gamma B \end{pmatrix} z_t, \qquad z_t = \begin{pmatrix} x_t \\ u_t \end{pmatrix}$$

Why is it interesting?

- Good local approximation
- Quadratic

# Discounted LQR value functions

Linear system + optimal discounted LQ controller

- State-value function [Bertsekas et al.(1987)]

$$J^\star_\gamma(x) = x_t^T P_\gamma x_t, \quad P_\gamma \text{ solution of discounted DARE.}$$

- Action-value function [Bradtke et al.(1993)]

$$Q^\star_\gamma(x,u) = z_t^T \begin{pmatrix} Q_\gamma + \gamma A^\top P_\gamma A & \gamma A^\top P_\gamma B \\ \gamma B^\top P_\gamma A & R_\gamma + \gamma B^\top P_\gamma B \end{pmatrix} z_t, \qquad z_t = \begin{pmatrix} x_t \\ u_t \end{pmatrix}$$

Why is it interesting?

- Good local approximation
- Quadratic

# Value function structure

Define $\boldsymbol{Q}^L = \mathbf{Q}^\star_\gamma$ and

$$h(x) = \tanh\left(\alpha\left(\frac{x^\top P x}{c}\right)^{\frac{3}{2}}\right) \rightarrow \text{ saturates at } \{V(x) = c\}$$

## Action-value function

The estimated action-value function is modeled as

$$\boldsymbol{Q}^\pi(x_t, u_t) = \boldsymbol{Q}^L(x_t, u_t) + h(x_t)\left(\boldsymbol{Q}^\phi(x_t, u_t) - \boldsymbol{Q}^L(x_t, u_t)\right). \tag{5}$$

- $\boldsymbol{Q}^L(x_t, u_t) \rightarrow$ computed from **CT** problem
- Locally Lipschitz $\boldsymbol{Q}^\phi \rightarrow$ learnt via **RL**
- Higher order term
  - ✓ Match up to second order
- Exact in the origin

# Value function structure

Define $\mathbf{Q}^L = \mathbf{Q}^\star_\gamma$ and

$$h(x) = \tanh\left(\alpha\left(\frac{x^\top P x}{c}\right)^{\frac{3}{2}}\right) \to \text{ saturates at } \{V(x) = c\}$$

## Action-value function

The estimated action-value function is modeled as

$$\mathbf{Q}^\pi(x_t, u_t) = \mathbf{Q}^L(x_t, u_t) + h(x_t)\left(\mathbf{Q}^\phi(x_t, u_t) - \mathbf{Q}^L(x_t, u_t)\right). \tag{5}$$

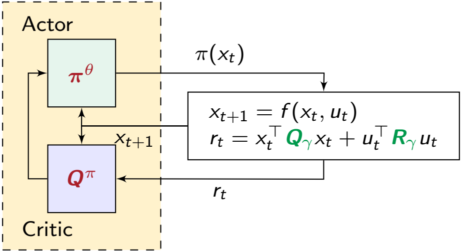
- $\mathbf{Q}^L(x_t, u_t) \to$ computed from **CT** problem
- Locally Lipschitz $\mathbf{Q}^\phi \to$ learnt via **RL**
- Higher order term
  - ✓ Match up to second order
- Exact in the origin

# Final structure

Summing up:

- Close to the equilibrium point → **CT** ensures stability and performances.
- Far from the equilibrium point → **RL** ensure performances.

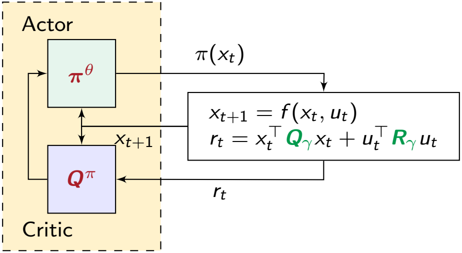
Overall structure:



# Main result

**Theorem**

Let be given an Actor-Critic algorithm. Let the reward function for the **RL** problem be defined as the instantaneous cost for the associated discounted LQR problem. Then by selecting the control policy $\pi^\theta$ and the value function estimator $Q^\pi$, the problem is solved.

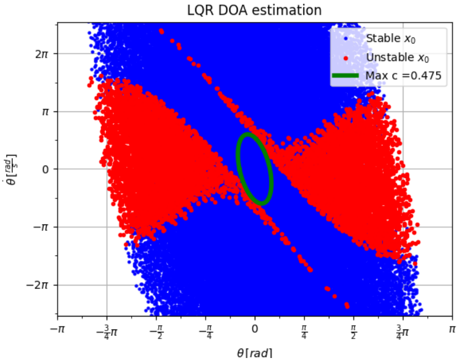


# Table of Contents

# Pendulum example

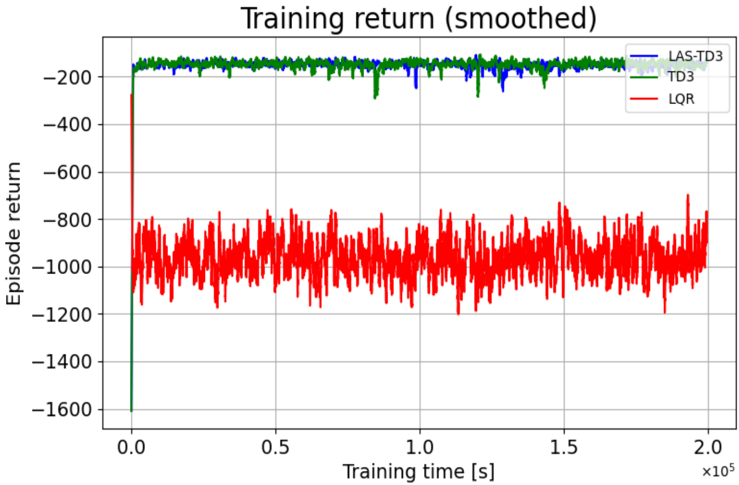
**Goal:** Stabilize in "top" position from any $x_0$
**Difficulty:** Nonlinearities → standard LQR works only locally

$$\alpha_{t+1} = \alpha_t + \omega_t \,\Delta t$$
$$\omega_{t+1} = \omega_t - \frac{3g}{2\ell}\sin(\alpha_t + \pi)\Delta t + \frac{3}{2ml^2}\mathsf{sat}(u(x_t))\Delta t,$$



# Deterministic policy (TD3)

**RL** algorithms learn to swing

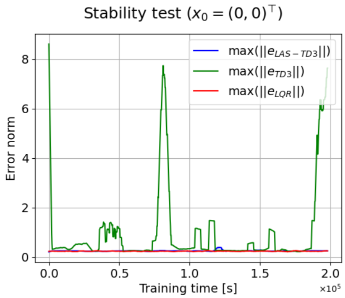


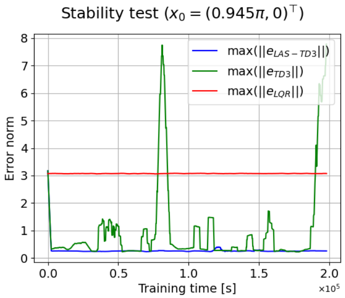
# Deterministic policy (TD3)

Standard **RL** formulations do not account for stability

$$\omega_{t+1} = \omega_t - \frac{3g}{2\ell}\sin(\alpha_t + \pi)\Delta t + \frac{3}{2\tilde{m}\ell^2}[\text{sat}(u(x_t + w_t)) + d_t]\Delta t,$$

$\tilde{m}$: parameter mismatch
$w_t$: random measurement noise
$d_t$: sinusoidal wind

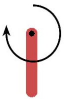


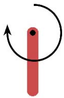
# Deterministic policy (TD3)

Corrupted environment, "bottom" initial condition $x_0 = \begin{pmatrix} 0.945\pi \\ 0 \end{pmatrix}$

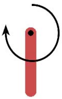
**CT**:LQR



**RL**:TD3



**CT**+**RL**:LAS-TD3



# Table of Contents

# What's next?

**We obtained**:

- ✓ Learnt policy with local guarantees
- ✓ Added linear system identification step

**Next step**:

- Different local controllers

# Thanks

**Thank you!**