

Regret Analysis of Continuous-Time Reinforcement Learning

Xin Guo, UC Berkeley

email: xinguo@berkeley.edu

SURE-AI@ Star Midsummer on Reinforcement Learning (Day Two)

Jun 23-25, 2026, Oslo

Outline

- 1 Regret in RL
- 2 Regret Analysis for Continuous-time LQ-RL
- 3 Regret Analysis in Continuous-time LCVX-RL

This lecture is based on the following works

- Basei, G., Hu, Zhang (2020).
Logarithmic regret for episodic continuous-time linear-quadratic reinforcement learning over a finite-time horizon
Journal of Machine Learning Research, arxiv.org/abs/2006.15316
- G., Hu, Zhang (2021).
Reinforcement learning for linear-convex models with jumps via stability analysis of feedback controls
SIAM J. on Control and Optimization, arxiv.org/abs/2104.09311

Outline

- 1 Regret in RL
- 2 Regret Analysis for Continuous-time LQ-RL
- 3 Regret Analysis in Continuous-time LCVX-RL

Reinforcement Learning (RL)

- Learning the (stochastic) environment while trying to find optimal control to achieve certain objectives
- Analytically, RL $\sim$ MDP/control plus learning
- DPP, martingale characterization, verification theorems as covered in earlier lectures

RL is Not Control

For an RL system with parameter θ and learned optimal control α^* , there are

- Identifiability problem:

One learns **how**, but also learns to understand **why**

— Suppose the optimal control α^* is achieved through some learning process, can the parameter θ be identified?

- Lipschitz stability of the value function in terms of the feedback control:

- Learning algorithm naturally of feedback form
- Robustness of control policies

- Regret analysis

When given multiple chances, the best learning strategy to achieve optimality with reasonable cost (e.g. robot driving)

Continuous-time RL problem with linear system

Given $(\Omega, \mathcal{F}, \mathbb{P})$ probability space, $T > 0$, find an $\mathbb{R}^k$ -valued square-integrable process α to minimize

$$J(\alpha) = \mathbb{E} \left[\int_0^T f(t, X_t^\alpha, \alpha_t) dt + g(X_T^\alpha) \right],$$

where for $\forall t \in [0, T]$,

$$dX_t^\alpha = (A^\star X_t^\alpha + B^\star \alpha_t) dt + \sigma dW_t + \int_{\mathbb{R}_0^p} \gamma(u) \tilde{N}(dt, du),$$

- $X_0 = x_0 \in \mathbb{R}^n$, $\sigma \in \mathbb{R}^{n \times d}$, $\gamma: \mathbb{R}_0^p \rightarrow \mathbb{R}^n$ s.t. $\int_{\mathbb{R}_0^p} |\gamma(u)|^2 \nu(du) < \infty$
- f and g appropriate functions
- W being n -dimensional Brownian motion and an independent Poisson random measure $N(dt, du)$ with compensator $\nu(du)dt$ such that $\int_{\mathbb{R}_0^p} \min(1, |u|^2) \nu(du) < \infty$ and $\tilde{N}(dt, du) := N(dt, du) - \nu(du)dt$
- $\mathbb{F} = (\mathcal{F}_t)_{t \in [0, T]}$ completed filtration generated by W and N

Continuous-time RL Problem

- $\theta^* = (A^*, B^*)^\top \in \mathbb{R}^{n \times n} \times \mathbb{R}^{n \times k}$ *unknown*
- **Goal:** simultaneously learn A^* and B^* and take controls to minimize some cost functional

Outline

- 1 Regret in RL
- 2 Regret Analysis for Continuous-time LQ-RL
- 3 Regret Analysis in Continuous-time LCVX-RL

LQ-RL with known $\theta^* = (A^*, B^*)$

- $(A^*, B^*) \in \mathbb{R}^{n \times n} \times \mathbb{R}^{n \times k}$ known

■

$$J(\alpha) = \mathbb{E} \left[\int_0^T ((X_t^\alpha)^\top Q X_t^\alpha + \alpha_t^\top R \alpha_t) dt \right],$$

■

$$dX_t = (A^* X_t + B^* \alpha_t) dt + dW_t, \quad \forall t \in [0, T]; \quad X_0 = x_0,$$

- $Q \in \mathbb{S}_+^n$ positive semi-definite and $R \in \mathbb{S}_+^k$ positive definite given matrices

LQ-RL with $\theta^* = (A^*, B^*)$ is known

With known θ^* , $J(\alpha) = \mathbb{E} \left[\int_0^T ((X_t^\alpha)^\top Q X_t^\alpha + \alpha_t^\top R \alpha_t) dt \right]$,

- the optimal control $\alpha_t^{\theta^*} = \phi^{\theta^*}(t, X_t^{\theta^*}) \quad \mathbb{P} \otimes dt \text{ a.e.}$
- X^{θ^*} the optimal controlled state process
- $\phi^{\theta^*}(t, x) = -R^{-1}(B^*)^\top P_t^{\theta^*} x$ for all $(t, x) \in [0, T] \times \mathbb{R}^n$,
- $(P_t^{\theta^*})_{t \in [0, T]}$ solving the Riccati equation:

$$\begin{cases} \frac{d}{dt} P_t + (A^*)^\top P_t + P_t A^* - P_t (B^* R^{-1} (B^*)^\top) P_t + Q = 0, & t \in [0, T], \\ P_T = 0. \end{cases}$$

LQ-RL

- $\theta^* = (A^*, B^*) \in \mathbb{R}^{n \times n} \times \mathbb{R}^{n \times k}$ unknown
- For any given control α ,

$$J(\alpha) = \mathbb{E} \left[\int_0^T ((X_t^\alpha)^\top Q X_t^\alpha + \alpha_t^\top R \alpha_t) dt \right],$$

- X^α satisfying the dynamics with parameter $\theta^* = (A^*, B^*)$:

$$dX_t = (A^* X_t + B^* \alpha_t) dt + dW_t, \quad \forall t \in [0, T]; \quad X_0 = x_0,$$

- α adapted to the filtration generated by X^α

Episodic LQ-RL formulation

- Given M episodes, for each $i = 1, \dots, M$ episode, exercise a feedback control $\phi^i(t, X_t)$ and observe a trajectory of the controlled process

$$dX_t^{\phi^i} = (A^* X_t^{\phi^i} + B^* \phi^i) dt + dW_t^i \quad t \in [0, T];$$

with $X_0^{\phi^i} = x_0$ and $(W^i)_{i=1}^M$ independent Brownian motions

- Cost of episode i with $\alpha_t^{\phi^i} = \phi^i(t, X_t^{\phi^i})$,

$$J(\alpha^{\phi^i}) = \mathbb{E} \left[\int_0^T f(t, X_t^{\phi^i}, \alpha_t^{\phi^i}) dt + g(X_T^{\phi^i}) \right],$$

- Performance criteria – regret up to episode M :

$$\mathcal{R}(N) = \sum_{i=1}^M \left(J(\alpha^{\phi^i}) - J(\alpha^{\phi^*}) \right),$$

- Goal: learn A^* , B^* and take controls to minimize $\mathcal{R}(M)$

Basic idea

- Starting with an initial estimation θ_0
- In each iteration $\ell = 0, 1, \dots$, execute “best possible” feedback control ϕ^{θ_ℓ} with the estimated parameter θ_ℓ replacing the true parameter θ^* in the Riccati equation, get m_ℓ independent samples
- Updating parameter $\theta_{\ell+1}$ using the LS estimator and based on m_ℓ trajectories
- Choosing m_ℓ appropriately not to exceed the total M episodes

Steps of Analysis

- Parameter estimation bound: $\|\theta_\ell - \theta^*\|$
- Robustness of feedback control: estimate the Lipschitz continuity of the control $\|\phi^{\theta_\ell} - \phi^{\theta^*}\|$, if such feedback control exists uniquely
- Regret bound: estimate the performance gap $J(\phi^{\theta_\ell}) - J(\phi^{\theta^*})$

Main Results

Under appropriate assumptions

(e.g., Q is positive definite and B^* full column rank),

- The regret bound is of a logarithmic scale in M
- θ^* is uniquely identifiable under the optimal control α^{θ^*} .
- Discretization approach will have the same regret order, with an additional discretization error

Logarithmic regret

Theorem 1

Under appropriate technical conditions, for all sufficiently large C and all $\delta \in (0, 3/\pi^2)$, if one set $m_0 = C(-\ln \delta)$ and $m_\ell = 2^\ell m_0$ for all $\ell \in \mathbb{N}$, then with probability at least $1 - \pi^2 \delta / 3$,

$$\mathcal{R}(M) \leq C'(\ln M)(\ln \ln M) + (-\ln \delta)(\ln M), \quad \forall M \in \mathbb{N},$$

where C' is a constant independent of M and δ .

Step I: Parameter estimation and error bound

Consider the state process X^θ controlled by ϕ^θ :

$$dX_t = (A^\star X_t + B^\star \phi^\theta(t, X_t)) dt + dW_t, \quad \forall t \in [0, T]; \quad X_0 = x_0.$$

If we observe (an invertible) $Z_t^\theta := \begin{pmatrix} X_t^\theta \\ \phi^\theta(t, X_t^\theta) \end{pmatrix}$ for all t , then

$$\theta^\star = \left(\mathbb{E} \left[\int_0^T Z_t^\theta (Z_t^\theta)^\top dt \right] \right)^{-1} \mathbb{E} \left[\int_0^T Z_t^\theta (dX_t^\theta)^\top \right],$$

Given m independent trajectories $(X_t^{\theta,i}, \phi^\theta(t, X_t^{\theta,i}))_{t \in [0, T], i=1, \dots, m}$:

$$\hat{\theta} \longleftarrow \left(\frac{1}{m} \sum_{i=1}^m \int_0^T Z_t^{\theta,i} (Z_t^{\theta,i})^\top dt + \frac{1}{m} \mathbb{I} \right)^{-1} \left(\frac{1}{m} \sum_{i=1}^m \int_0^T Z_t^{\theta,i} (dX_t^{\theta,i})^\top \right).$$

Analysis for parameter estimation

- Parameter updating using m i.i.d. trajectories

$$\hat{\theta} \leftarrow \left(\frac{1}{m} \sum_{i=1}^m \int_0^T Z_t^{\theta,i} (Z_t^{\theta,i})^\top dt + \frac{1}{m} \mathbb{I} \right)^{-1} \left(\frac{1}{m} \sum_{i=1}^m \int_0^T Z_t^{\theta,i} (dX_t^{\theta,i})^\top \right).$$

- Estimating the error bound for $\|\theta_\ell - \theta^\star\|$

- $Z_t^\theta = \begin{pmatrix} X_t^\theta \\ \phi^\theta(t, X_t^\theta) \end{pmatrix}$ is Gaussian
- $\int_0^T Z_t^\theta (Z_t^\theta)^\top dt$ and $\int_0^T Z_t^\theta (dX_t^\theta)^\top$ are sub-exponential
- establish corresponding concentration inequalities [here](#)

Parameter estimation error bound

Theorem 2

Under appropriate assumptions, there exist constants $\bar{C}_1, \bar{C}_2 \geq 0$, such that for all $\theta \in \Theta$ and $\delta \in (0, 1/2)$, if $m \geq \bar{C}_1(-\ln \delta)$, then with probability at least $1 - 2\delta$,

$$|\hat{\theta} - \theta^*| \leq \bar{C}_2 \left(\sqrt{\frac{-\ln \delta}{m}} + \frac{-\ln \delta}{m} + \frac{(-\ln \delta)^2}{m^2} \right).$$

- An additional term $\frac{1}{N}$ appears if using discretized algorithm with N being number of discretization.

Step 2: Stability of feedback control

Given an estimate $\theta = (A, B)^\top$ of θ^* , exercise the optimal feedback control ϕ^θ for the estimated LQ control problem:

$$\phi^\theta(t, x) = -R^{-1}B^\top P_t^\theta x, \quad \forall (t, x) \in [0, T] \times \mathbb{R}^n,$$

where P^θ satisfies the Riccati equation with θ :

$$\begin{cases} \frac{d}{dt}P_t + A^\top P_t + P_t A - P_t(BR^{-1}B^\top)P_t + Q = 0, & t \in [0, T]; \\ P_T = 0. \end{cases}$$

- Characterizing the stability of ϕ^θ
- Estimating the performance gap $J(\alpha^\theta) - J(\alpha^{\theta^*})$

Performance gap analysis

Theorem 3

Under appropriate assumptions, for each $\theta \in \Theta$, let ϕ^θ be the optimal feedback control for the LQ system with parameter θ . Then there exists a constant C such that

$$|J(\alpha^\theta) - J(\alpha^{\theta^*})| \leq C|\theta - \theta^*|^2, \quad \forall \theta \in \Theta.$$

- An additional term $\frac{1}{N^2}$ appears if using discretized algorithm with N being number of discretization.

Outline

- 1 Regret in RL
- 2 Regret Analysis for Continuous-time LQ-RL
- 3 Regret Analysis in Continuous-time LCVX-RL

Episodic LCVX-RL formulation

- **Episodic** formulation: given M episodes, for each $i = 1, \dots, M$ episode, exercise a feedback control $\phi^i(t, X_t)$ and observe a **trajectory** of the controlled state process X^{ϕ^i} starting from x_0

$$dX_t^{\phi^i} = (A^* X_t^{\phi^i} + B^* \phi^i) dt + dW_t^i + \int_{\mathbb{R}^p} \gamma(u) \tilde{N}^i(dt, du), \quad t \in [0, T];$$

- **Cost** of episode i :

$$J(\alpha^{\phi^i}) = \mathbb{E} \left[\int_0^T f(t, X_t^{\phi^i}, \alpha_t^{\phi^i}) dt + g(X_T^{\phi^i}) \right],$$

with $\alpha_t^{\phi^i} := \phi^i(t, X_t^{\phi^i})$, $t \in [0, T]$,

- **Performance criteria – regret up to episode M :**

$$\mathcal{R}(M) = \sum_{i=1}^M \left(J(\alpha^{\phi^i}) - J(\alpha^{\phi^*}) \right),$$

- **Goal:** learn A^* , B^* and take controls to minimize $\mathcal{R}(M)$.

Challenges of LC-RL

Key components in the analysis of LQ-RL

- Estimation error based on observed trajectories
Counterpart of sub-exponential in LC-RL??
- Lipschitz stability of value function and control on parameters based on the Riccati equation
Feedback control: existence and stability in LC-RL??

Classical control theory focuses on

- Existence and uniqueness of optimal controls, viscosity solutions, and BSDEs
- Regularity of value functions with respect to state variables

LQ-RL vs LC-RL

	LQ-RL	LC-RL
Stability of feedback control	Stability of Riccati equation	Stability of FBSDE
Parameter estimation	Sub-exponential property	Sub-Weibull property

Assumptions (H)

- $\exists f_0 : [0, T] \times \mathbb{R}^n \times \mathbb{R}^k \rightarrow \mathbb{R}$ and $h : \mathbb{R}^k \rightarrow \mathbb{R} \cup \{\infty\}$ s.t.

$$f(t, x, a) = f_0(t, x, a) + h(a)$$

and for all $(t, x) \in [0, T] \times \mathbb{R}^n$,

- $f(t, \cdot, \cdot)$ is convex in x , strongly convex in a
- $f_0(t, x, \cdot)$ is convex, $f_0(t, \cdot, \cdot)$ is Lipschitz differentiable
- h is proper, lower semicontinuous and convex
- $g : \mathbb{R}^n \rightarrow \mathbb{R}$ is convex and Lipschitz differentiable

h represents control constraints, ℓ_1 -norm or entropy regularizers

Main Result: Sublinear regret

Theorem 4

Given (H) , suppose that θ^* is identifiable under the optimal control α^{θ^*} , and that the pure jump term $\int_0^T \int_{\mathbb{R}_0^p} \gamma(u) \tilde{N}(ds, du)$ is ϑ -sub-exponential with $\vartheta \in [0, 1]$. Now for all sufficiently large C and all $\delta \in (0, 1/4)$, set $m_0 = C(-\ln \delta)^{3+\vartheta}$ and $m_\ell = 2^\ell m_0$ for all $\ell \in \mathbb{N}$. Then with probability at least $1 - 4\delta$,

$$\mathcal{R}(M) \leq C'(\sqrt{M \ln M} + \sqrt{(-\ln \delta)M} + (-\ln \delta)^{3+\vartheta} \ln M), \quad \forall M \in \mathbb{N},$$

where C' is a constant independent of ϑ, M and δ .

- Without jumps, the third term can be reduced to $(-\ln \delta) \ln M$.

Step I: Parameter estimation

- $\int_0^T Z_t^\theta (Z_t^\theta)^\top dt$ and $\int_0^T Z_t^\theta (dX_t^\theta)^\top$ are sub-Weibull [here](#)
- can establish corresponding concentration inequalities [here](#)

Theorem 5

Let $\beta = 3 + \vartheta$. Under appropriate assumptions, there exist constants $\bar{C}_1, \bar{C}_2 \geq 0$, such that for all $\theta \in \Theta$ and $\delta \in (0, 1/2)$, if $m \geq \bar{C}_1 (-\ln \delta)^\beta$, then we have with probability at least $1 - 2\delta$ that

$$|\hat{\theta} - \theta^\star| \leq \bar{C}_2 \left(\sqrt{\frac{-\ln \delta}{m}} + \frac{(-\ln \delta)^\beta}{m} + \frac{(-\ln \delta)^{2\beta}}{m^2} \right),$$

- When there is no jump, the statement holds for $\beta = 1$

Step II: Feedback control for LC

Theorem 6

Under appropriate conditions,

- *the LC control problem with parameter θ admits a unique optimal control of the form $\alpha_t^\theta = \phi^\theta(t, X_t^\theta)$, with $\phi^\theta(t, x) = \psi^\theta(t, x, Y_t^{t,x})$, which is constructed via a coupled forward-backward SDE with parameter θ . [here](#)*
- *The feedback control ϕ^θ is Lipschitz with respect to θ , i.e., there exists some constant C such that*

$$|\phi^\theta(t, x) - \phi^{\theta'}(t, x)| \leq C(1 + |x|)|\theta - \theta'| \text{ for all } (t, x) \in [0, T] \times \mathbb{R}^n.$$

Note: the standard control approach via compactification (El Karoui, Du'huu, Monique (1987), Haussmann, Lepeltier (1990)) passing from open-loop to closed-loop is neither constructive nor sufficient.

Performance gap analysis

Theorem 7

Under appropriate assumptions, for each $\theta \in \Theta$, let ϕ^θ be the optimal feedback control for the LC system with parameter θ . Then there exists a constant C such that

$$|J(\alpha^\theta) - J(\alpha^{\theta^*})| \leq C|\theta - \theta^*|, \quad \forall \theta \in \Theta.$$

- Analysis can be adapted to more general settings.

Summary

This lecture covers

- Regret bound analysis of greedy algorithm for episodic linear-convex with RL with jump
- Lipschitz stability analysis of feedback controls and value function, hopefully inspiring further interests in RL from control and BSDE community.

Open question: generalization to non-linear dynamics?

Thank you

Sub-exponential property

Definition 8

A random variable X with mean $\mu = \mathbb{E}[X]$ is (v, b) -sub-exponential for $v, b \in [0, \infty)$ if $\mathbb{E}[e^{\lambda(X-\mu)}] \leq e^{v^2\lambda^2/2}$ for all $|\lambda| < 1/b$.

Lemma 9

Let $m \in \mathbb{N}$, $v, b \in [0, \infty)$ and $(X_i)_{i=1}^m$ be independent (v, b) -sub-exponential random variables with $\mu = \mathbb{E}[X_i]$ for all $i = 1, \dots, m$. Then for all $\epsilon \geq 0$,

$$\mathbb{P}\left(\left|\frac{1}{m} \sum_{i=1}^m X_i - \mu\right| \geq \epsilon\right) \leq 2 \exp\left(-\min\left\{\frac{m\epsilon^2}{2v^2}, \frac{m\epsilon}{2b}\right\}\right).$$

Adopting the techniques in Cheridito, Soner and Touzi (2005) [back](#)

Maximum principle

Theorem 10

For any given $\theta = (A, B)$, the LC control problem with parameter θ admits a unique optimal control $\alpha^\theta \in \mathcal{H}^2(\mathbb{R}^k)$, which satisfies

$$\alpha_t^\theta = \psi^\theta(t, X_t^\theta, Y_t^\theta), \quad d\mathbb{P} \otimes dt \text{ a.e.},$$

where $\psi^\theta(t, x, y) = \operatorname{argmin}_{a \in \mathbb{R}^k} H(t, x, a, y)$, H is the Hamiltonian, and $(X^\theta, Y^\theta, Z^\theta, M^\theta)$ solves for all $s \in [0, T]$,

$$\begin{cases} dX_s = b(X_s, \psi^\theta(s, X_s, Y_s)) ds + \sigma dW_s + \int_{\mathbb{R}_0^p} \gamma(u) \tilde{N}(ds, du), \\ dY_s = -\partial_x H(s, X_s, \psi^\theta(s, X_s, Y_s), Y_s) ds + Z_s dW_s + \int_{\mathbb{R}_0^p} M_s \tilde{N}(ds, du), \\ X_0 = x_0, \quad Y_T = \nabla g(X_T). \end{cases}$$

back

sub-Weibull property

Definition 11

For every $\alpha > 0$, let $\Psi_\alpha : [0, \infty) \rightarrow \mathbb{R}$ such that $\Psi_\alpha(x) = e^{x^\alpha} - 1$ for all $x \geq 0$, and let $\|\cdot\|_{\Psi_\alpha}$ be the corresponding Ψ_α -Orlicz (quasi-)norm such that for any given random variable X ,

$$\|X\|_{\Psi_\alpha} := \inf \left\{ t > 0 \mid \mathbb{E} \left[\Psi_\alpha \left(\frac{|X|}{t} \right) \right] \leq 1 \right\}.$$

Then a random variable X is said to be sub-Weibull of order $\alpha > 0$, denoted by $X \in \text{subW}(\alpha)$, if $\|X\|_{\Psi_\alpha} < \infty$.

- sub-Gaussian random variable is in $\text{subW}(2)$
- sub-exponential random variable is in $\text{subW}(1)$

back

Concentration inequality for sub-Weibull random variables

Lemma 12

Let $\alpha \in (0, 1]$, $N \in \mathbb{N}$ and $X_1, \dots, X_N \in \text{subW}(\alpha)$ be independent random variables satisfying $\mathbb{E}[X_i] = 0$ for all $i = 1, \dots, N$. Then there exists a constant $C \geq 0$, depending only on α , such that

$$\mathbb{P}\left(\left|\sum_{i=1}^N X_i\right| \geq \varepsilon\right) \leq 2 \exp\left(-C \min\left\{\frac{\varepsilon^2}{\sum_{i=1}^N \|X_i\|_{\Psi_\alpha}^2}, \left(\frac{\varepsilon}{\max_i \|X_i\|_{\Psi_\alpha}}\right)^\alpha\right\}\right), \quad \forall \varepsilon \geq 0.$$

back

Basic idea: greedy least-squares algorithm

- 1: **Input:** Choose an initial estimation θ_0 of $\theta^\star$ and numbers of learning episodes $\{m_\ell\}_{\ell \in \mathbb{N} \cup \{0\}}$.
- 2: **for** $\ell = 0, 1, \dots$ **do**
- 3: Obtain the feedback control ϕ^{θ_ℓ} for LQ control problem with θ_ℓ by Riccati equations.
- 4: Execute ϕ^{θ_ℓ} for m_ℓ independent episodes, and collect the trajectory data $(X_t^{\theta_\ell, i}, \phi^{\theta_\ell}(t, X_t^{\theta_\ell, i}))_{t \in [0, T], i = 1, \dots, m_\ell}$.
- 5: Obtain an updated estimation $\theta_{\ell+1}$ by using the LS estimator and m_ℓ trajectories collected above.
- 6: **end for**