Novelty Assessment Report

Paper: Instance-Dependent Continuous-Time Reinforcement Learning via Maximum Likelihood Estimation

PDF URL: https://openreview.net/pdf?id=05NHmcEpNk

Venue: ICLR 2026 Conference Submission

Year: 2026

Report Generated: 2025-12-29

Abstract

Continuous-time reinforcement learning (CTRL) provides a natural framework for sequential decision-making in dynamic environments where interactions evolve continuously over time. While CTRL has shown growing empirical success, its ability to adapt to varying levels of problem difficulty remains poorly understood. In this work, we investigate the instance-dependent behavior of CTRL and introduce a simple, model-based algorithm built on maximum likelihood estimation (MLE) with a general function approximator. Unlike existing approaches that estimate system dynamics directly, our method estimates the state marginal density to guide learning. We establish instance-dependent performance guarantees by deriving a regret bound that scales with the total reward variance and measurement resolution. Notably, the regret becomes independent of the specific measurement strategy when the observation frequency adapts appropriately to the problem's complexity. To further improve performance, our algorithm incorporates a randomized measurement schedule that enhances sample efficiency without increasing measurement cost. These results highlight a new direction for designing CTRL algorithms that automatically adjust their learning behavior based on the underlying difficulty of the environment.

Disclaimer

This report is **AI-GENERATED** using Large Language Models and WisPaper (a scholar search engine). It analyzes academic papers' tasks and contributions against retrieved prior work. While this system identifies **POTENTIAL** overlaps and novel directions, **ITS COVERAGE IS NOT EXHAUSTIVE AND JUDGMENTS ARE APPROXIMATE**. These results are intended to assist human reviewers and **SHOULD NOT** be relied upon as a definitive verdict on novelty.

Note that some papers exist in multiple, slightly different versions (e.g., with different titles or URLs). The system may retrieve several versions of the same underlying work. The current automated pipeline does not reliably align or distinguish these cases, so human reviewers will need to disambiguate them manually.

If you have any questions, please contact: mingzhang23@m.fudan.edu.cn

Core Task Landscape

This paper addresses: **Instance-Dependent Regret Analysis in Continuous-Time Reinforcement Learning**

A total of **11 papers** were analyzed and organized into a taxonomy with **10 categories**.

Taxonomy Overview

The research landscape has been organized into the following main categories:

- **Linear-Quadratic Control Problems**
- **General Continuous-Time MDP Frameworks**
- **Specialized Application Domains**
- **Methodological Foundations and Surveys**

Complete Taxonomy Tree

- Instance-Dependent Regret Analysis in Continuous-Time Reinforcement Learning Survey Taxonomy
- Linear-Quadratic Control Problems
  - Finite-Horizon Episodic LQ Learning ★ (3 papers)
  - [0] Instance-Dependent Continuous-Time Reinforcement Learning via Maximum Likelihood Estimation (Anon et al., 2026) View paper
  - [2] Logarithmic Regret for Episodic Continuous-Time Linear-Quadratic Reinforcement Learning Over a Finite-Time Horizon (Matteo Basei, 2020) View paper
  - [7] Linear Quadratic Reinforcement Learning: Sublinear Regret in the Episodic Continuous-Time Framework (Matteo Basei, 2020) View paper
  - Single-Trajectory LQ Learning (1 papers)
  - [9] Regret Analysis of Certainty Equivalence Policies in Continuous-Time Linear-Quadratic Systems (Mohamad Kazem Shirani Faradonbeh, 2022) View paper
  - Actor-Critic Methods for LQ Systems (1 papers)
  - [1] Sublinear regret for an actor-critic algorithm in continuous-time linear-quadratic reinforcement learning (Yilie Huang, 2024) View paper
- General Continuous-Time MDP Frameworks
  - Average-Reward Continuous-Time MDPs (1 papers)
  - [4] Logarithmic regret bounds for continuous-time average-reward Markov decision processes (Gao Xue-feng, 2022) View paper
  - ODE-Based Model Learning (1 papers)
  - [6] Efficient Exploration in Continuous-time Model-based Reinforcement Learning (Treven, 2023) View paper
  - Local Linearity Exploitation (1 papers)
  - [5] Local Linearity: the Key for No-regret Reinforcement Learning in Continuous MDPs (Davide Maran, 2024) View paper
- Specialized Application Domains
  - Mean-Variance Portfolio Optimization (1 papers)
  - [3] Mean-Variance Portfolio Selection by Continuous-Time Reinforcement Learning: Algorithms, Regret Analysis, and Empirical Study (Jia Yan-Wei, 2024) View paper
  - Jump-Diffusion Linear-Convex Control (1 papers)
  - [8] Reinforcement learning for linear-convex models with jumps via stability analysis of feedback controls (Xin Guo, 2021) View paper
- Methodological Foundations and Surveys
  - Differential and Pointwise Control Theory (1 papers)
  - [11] A Differential and Pointwise Control Approach to Reinforcement Learning (MP Nguyen, n.d.) View paper

- ◦ Mean-Field Games and Stochastic Control (1 papers)
- ◦ [10] Learning in Mean-Field Games and Continuous-Time Stochastic Control Problems (Hu, 2022) View paper

Narrative

Core task: instance-dependent regret analysis in continuous-time reinforcement learning. The field structure reflects a natural division between tractable special cases and more general frameworks. Linear-Quadratic Control Problems form a dense branch where the quadratic cost and linear dynamics enable sharp, often logarithmic regret bounds; works here exploit closed-form solutions and maximum-likelihood estimation to achieve instance-dependent guarantees. General Continuous-Time MDP Frameworks extend beyond LQ settings, addressing broader state and action spaces, jump processes, and nonlinear dynamics, though often at the cost of weaker or sublinear regret rates. Specialized Application Domains apply these techniques to finance, mean-field games, and other areas where continuous-time models arise naturally. Methodological Foundations and Surveys provide overarching perspectives on exploration strategies, certainty equivalence principles, and the interplay between discrete and continuous formulations.

Within the LQ branch, a particularly active line of work focuses on finite-horizon episodic learning. Episodic LQ Logarithmic[2] and Episodic LQ Sublinear[7] illustrate the trade-off between tight instance-dependent bounds and broader applicability: the former achieves logarithmic regret under strong assumptions, while the latter relaxes these at the expense of polynomial rates. Instance-Dependent MLE[0] sits squarely in this episodic LQ cluster, emphasizing maximum-likelihood estimation to refine regret guarantees and exploit problem structure more precisely than sublinear approaches. Its focus on instance-dependent analysis contrasts with works like Actor-Critic Sublinear[1] or Local Linearity[5], which prioritize robustness or local approximations over tight problem-specific bounds. Meanwhile, extensions to average-reward settings (Average-Reward Logarithmic[4]) and jump-diffusion models (Linear-Convex Jumps[8]) show how the core LQ insights scale to richer continuous-time environments, though the original paper remains anchored in the episodic finite-horizon regime where instance-dependent MLE techniques are most directly applicable.

Related Works in Same Category

The following **2 sibling papers** share the same taxonomy leaf node with the original paper:

1. Logarithmic Regret for Episodic Continuous-Time Linear-Quadratic Reinforcement Learning Over a Finite-Time Horizon

Authors: Matteo Basei, Xin Guo, Anran Hu, Yufei Zhang | **Year/Venue**: 2020 • Journal of machine learning research | **URL**: View paper

Abstract

We study finite-time horizon continuous-time linear-quadratic reinforcement learning problems in an episodic setting, where both the state and control coefficients are unknown to the controller. We first propose a least-squares algorithm based on continuous-time observations and controls, and establish a logarithmic regret bound of order $O((\ln M)(\ln\ln M))$, with M being the number of learning episodes. The analysis consists of two parts: perturbation analysis, which exploits the regularity...

Relationship Analysis

Both papers address finite-horizon episodic learning in continuous-time linear-quadratic systems with unknown coefficients, sharing the same taxonomy category of episodic LQ learning. The original paper focuses on instance-dependent regret bounds via maximum likelihood estimation of marginal state densities with general function approximation, achieving variance-aware regret that scales with total reward variance and measurement resolution. The candidate paper establishes logarithmic regret bounds O((ln M)(ln ln M)) for continuous-time LQ systems using least-squares estimation of drift and diffusion coefficients directly, with analysis based on Riccati equation regularity rather than marginal density estimation.

2. Linear Quadratic Reinforcement Learning: Sublinear Regret in the Episodic Continuous-Time Framework

Authors: Matteo Basei, Xin Guo, Anran Hu | **Year/Venue**: 2020 | **URL**: View paper

Abstract

N/A

Relationship Analysis

Both papers belong to the Finite-Horizon Episodic LQ Learning category, addressing episodic learning in continuous-time linear-quadratic systems with unknown coefficients. The original paper focuses on instance-dependent regret analysis using maximum likelihood estimation with general function approximation and adaptive measurement strategies, while the candidate paper specifically targets linear-quadratic systems with sublinear regret guarantees. The key difference is that the original paper employs a broader MLE-based approach with variance-aware bounds applicable to general nonlinear systems, whereas the candidate paper specializes in the linear-quadratic setting with its specific structural properties.

Contributions Analysis

Overall novelty summary. The paper proposes a model-based continuous-time reinforcement learning algorithm using maximum likelihood estimation to achieve instance-dependent regret guarantees. It resides in the Finite-Horizon Episodic LQ Learning leaf, which contains three papers including this work. This leaf sits within the Linear-Quadratic Control Problems branch, representing a moderately populated research direction where tractable dynamics enable sharp theoretical analysis. The focus on instance-dependent bounds through MLE distinguishes this work from sibling papers that pursue either logarithmic regret under strong assumptions or sublinear rates with broader applicability.

The taxonomy reveals that Linear-Quadratic Control Problems form the most developed branch, with three distinct learning paradigms: episodic, single-trajectory, and actor-critic approaches. Neighboring leaves address average-reward continuous-time MDPs and ODE-based model learning in the General Continuous-Time MDP Frameworks branch, which handles nonlinear dynamics and broader state spaces. The paper's episodic LQ setting connects naturally to these general frameworks but exploits quadratic structure for tighter guarantees. The Specialized Application Domains branch shows extensions to finance and jump-diffusion processes, indicating how core LQ insights scale to richer environments beyond the paper's finite-horizon regime.

Among 19 candidates examined across three contributions, no clearly refuting prior work was identified. The CT-MLE algorithm examined 3 candidates with none refutable; the instance-dependent regret bound examined 7 candidates with none refutable; the randomized measurement strategy examined 9 candidates with none refutable. This suggests that within the limited search scope, the specific combination of state marginal density estimation, variance-adaptive measurement, and randomized scheduling appears relatively unexplored. However, the search examined only top-K semantic matches and citations, not an exhaustive literature review, so stronger overlap may exist beyond these 19 papers.

Based on the limited search scope of 19 papers, the work appears to occupy a distinct position within episodic LQ learning by emphasizing measurement-adaptive strategies and state marginal density estimation rather than direct dynamics estimation. The absence of refuting candidates across all contributions suggests novelty in the specific technical approach, though the episodic LQ setting itself is well-established. The analysis cannot rule out substantial prior work outside the examined candidates, particularly in adjacent areas like adaptive sampling or variance-dependent bounds in discrete-time settings.

This paper presents **3 main contributions**, each analyzed against relevant prior work:

Contribution 1: CT-MLE algorithm for continuous-time reinforcement learning

Description: The authors propose CT-MLE, a model-based algorithm that estimates marginal state density using maximum likelihood estimation with general function approximators, rather than directly estimating system dynamics. This approach offers greater modeling flexibility and is compatible with a broad range of policy classes and sampling strategies.

This contribution was assessed against **3 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

1. Comparison of Model-Based and Model-Free Reinforcement Learning Algorithms for Stochastic Linear Quadratic Control

URL: View paper

Brief Assessment

Model-Based vs Model-Free[29] compares model-based and model-free approaches but does not present a specific algorithm using maximum likelihood estimation for marginal state density in continuous-time RL. The candidate focuses on comparing existing paradigms rather than proposing the CT-MLE methodology.

2. Deep Learning-based Approaches for State Space Models: A Selective Review

URL: View paper

Brief Assessment

State Space Review[27] focuses on state space models for dynamical system analysis and time series modeling, not on reinforcement learning algorithms or policy optimization. The candidate does not address continuous-time RL or maximum likelihood estimation for marginal state density in RL contexts.

3. Continuous-Time Reinforcement Learning: Algorithms, Theoretical Analysis, and Financial Applications

URL: View paper

Brief Assessment

Continuous-Time RL Survey[28] is a survey paper covering various continuous-time RL algorithms and methods. It does not present a specific novel algorithm that would refute the novelty of CT-MLE's approach to estimating marginal state density using maximum likelihood estimation with general function approximators.

Contribution 2: Instance-dependent regret bound with variance-adaptive measurement

Description: The authors derive a theoretical regret bound that scales with total reward variance and measurement resolution. When measurement schedules adapt appropriately to problem complexity, the regret becomes nearly independent of the specific measurement strategy, demonstrating instance-dependent adaptivity in continuous-time reinforcement learning.

This contribution was assessed against **7 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

1. Performance Analysis of Least Squares of Continuous-Time Model Based on Sampling Data

URL: View paper

Brief Assessment

Sampled Model Performance[24] focuses on parameter estimation in continuous-time linear stochastic regression models using least squares with sampling data, not on instance-dependent regret bounds with variance-adaptive measurement schedules in reinforcement learning contexts.

2. Dare: The deep adaptive regulator for control of uncertain continuous-time systems

URL: View paper

Brief Assessment

Deep Adaptive Regulator[22] focuses on continuous-time optimal control with unknown environments using physics-informed neural networks, not on instance-dependent regret bounds or variance-adaptive measurement schedules in reinforcement learning.

3. Beyond Worst-case Attacks: Robust RL with Adaptive Defense via Non-dominated Policies

URL: View paper

Brief Assessment

Adaptive Defense[23] addresses adversarial robustness in RL through policy refinement and regret minimization at test time, not variance-adaptive measurement schedules in continuous-time RL.

4. Provably Efficient Model-based Policy Adaptation

URL: View paper

Brief Assessment

Policy Adaptation[25] focuses on model-based policy adaptation across different environments using online learning and adaptive control, not on continuous-time RL with variance-adaptive measurement schedules or instance-dependent regret bounds in the CTRL setting.

5. Adaptive Experience Selection for Policy Gradient

URL: View paper

Brief Assessment

Adaptive Experience Selection[26] focuses on variance reduction in experience replay for policy gradient methods in discrete-time RL, not on continuous-time reinforcement learning with measurement schedules or instance-dependent regret bounds for CTRL systems.

6. Efficient Exploration in Continuous-time Model-based Reinforcement Learning

URL: View paper

Brief Assessment

Efficient Exploration[6] focuses on measurement selection strategies and epistemic uncertainty with Gaussian processes, but does not establish variance-dependent regret bounds that scale with total reward variance as in the original paper.

7. When to Sense and Control? A Time-adaptive Approach for Continuous-Time RL

URL: View paper

Brief Assessment

Time-Adaptive Sensing[21] focuses on optimizing measurement timing to reduce interaction costs in continuous-time systems, not on deriving variance-adaptive regret bounds that scale with total reward variance and measurement resolution as claimed in the original paper.

Contribution 3: Randomized measurement strategy for unbiased reward estimation

Description: The authors introduce a Monte Carlo-type randomized measurement strategy that augments the default measurement grid with additional observation points sampled within each interval. This enables unbiased estimation of reward integrals while maintaining the same order of measurement complexity.

This contribution was assessed against **9 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

1. Improving stochastic policy gradients in continuous control with deep reinforcement learning using the beta distribution

URL: View paper

Brief Assessment

Beta Distribution Gradients[15] focuses on policy gradient methods for bounded action spaces in continuous control, not on measurement strategies for reward estimation in continuous-time systems. The candidate addresses action distribution bias, while the original addresses temporal measurement scheduling.

2. Temporal Difference Learning with Continuous Time and State in the Stochastic Setting

URL: View paper

Brief Assessment

Continuous TD Learning[16] focuses on policy evaluation in continuous-time using TD(0) variants with vanishing time steps, not on randomized measurement strategies for reward integral estimation as proposed in the original paper.

3. Score-based Continuous-time Discrete Diffusion Models

URL: View paper

Brief Assessment

Score-based Diffusion[14] addresses continuous-time diffusion models for categorical discrete data using score matching techniques, not continuous-time reinforcement learning with measurement strategies for reward estimation. The technical domains are fundamentally different.

4. Nonlinear and timeâ□□dependent effects of sparsely measured continuous timeâ□□varying covariates in timeâ□□toâ□□event analysis

URL: View paper

Brief Assessment

Time-Varying Covariates[18] addresses sparse measurements in survival analysis using simulation extrapolation and time-elapsed-since-last-observation effects, not randomized measurement strategies for reward estimation in continuous-time RL systems.

5. Unbiased Simulation of Rare Events in Continuous Time

URL: View paper

Brief Assessment

Rare Events Simulation[19] focuses on unbiased estimation of rare event probabilities in continuous-time Markov processes using epsilon-strong simulation, not on reward estimation in reinforcement learning contexts. The randomized measurement strategy in the original paper augments measurement grids for reward integrals in RL, while the candidate addresses barrier crossing detection in rare event simulation.

6. Accuracy of Discretely Sampled Stochastic Policies in Continuous-time Reinforcement Learning

URL: View paper

Brief Assessment

Discretely Sampled Accuracy[13] focuses on convergence rates of discretely sampled stochastic policies and bias/variance analysis of policy evaluation estimators, not on randomized measurement strategies for unbiased reward integral estimation in continuous-time RL.

7. Unbiased Estimation of the Gradient of the Log-Likelihood for a Class of Continuous-Time State-Space Models

URL: View paper

Brief Assessment

Unbiased Gradient Estimation[20] focuses on unbiased estimation of log-likelihood gradients in continuous-time state-space models using doubly randomized schemes with particle filters. The original paper's randomized measurement strategy addresses unbiased reward integral estimation in reinforcement learning, which is a fundamentally different problem domain and objective.

8. Bias correction for direct spectral estimation from irregularly sampled data including sampling schemes with correlation

URL: View paper

Brief Assessment

Irregular Sampling Bias[17] addresses bias correction in spectral estimation from irregularly sampled data, not reward estimation in reinforcement learning. The randomization serves different purposes in fundamentally different domains (signal processing vs. RL).

9. Stochastic sampling of operator growth dynamics

URL: View paper

Brief Assessment

Operator Growth Sampling[12] focuses on Monte Carlo sampling for quantum operator growth dynamics in spin systems, not reinforcement learning reward estimation in continuous-time control systems.

Appendix: Text Similarity Detection

No high-similarity text segments were detected across any compared papers.

References

- [0] Instance-Dependent Continuous-Time Reinforcement Learning via Maximum Likelihood Estimation View paper
- [1] Sublinear regret for an actor-critic algorithm in continuous-time linear-quadratic reinforcement learning View paper
- [2] Logarithmic Regret for Episodic Continuous-Time Linear-Quadratic Reinforcement Learning Over a Finite-Time Horizon View paper
- [3] Mean-Variance Portfolio Selection by Continuous-Time Reinforcement Learning: Algorithms, Regret Analysis, and Empirical Study View paper
- [4] Logarithmic regret bounds for continuous-time average-reward Markov decision processes View paper
- [5] Local Linearity: the Key for No-regret Reinforcement Learning in Continuous MDPs View paper
- [6] Efficient Exploration in Continuous-time Model-based Reinforcement Learning View paper
- [7] Linear Quadratic Reinforcement Learning: Sublinear Regret in the Episodic Continuous-Time Framework View paper
- [8] Reinforcement learning for linear-convex models with jumps via stability analysis of feedback controls View paper
- [9] Regret Analysis of Certainty Equivalence Policies in Continuous-Time Linear-Quadratic Systems View paper
- [10] Learning in Mean-Field Games and Continuous-Time Stochastic Control Problems View paper
- [11] A Differential and Pointwise Control Approach to Reinforcement Learning View paper
- [12] Stochastic sampling of operator growth dynamics View paper
- [13] Accuracy of Discretely Sampled Stochastic Policies in Continuous-time Reinforcement Learning View paper
- [14] Score-based Continuous-time Discrete Diffusion Models View paper
- [15] Improving stochastic policy gradients in continuous control with deep reinforcement learning using the beta distribution View paper
- [16] Temporal Difference Learning with Continuous Time and State in the Stochastic Setting View paper
- [17] Bias correction for direct spectral estimation from irregularly sampled data including sampling schemes with correlation View paper
- [18] Nonlinear and timeâ□□dependent effects of sparsely measured continuous timeâ□□varying covariates in timeâ□□toâ□□event analysis View paper
- [19] Unbiased Simulation of Rare Events in Continuous Time View paper
- [20] Unbiased Estimation of the Gradient of the Log-Likelihood for a Class of Continuous-Time State-Space Models View paper
- [21] When to Sense and Control? A Time-adaptive Approach for Continuous-Time RL View paper
- [22] Dare: The deep adaptive regulator for control of uncertain continuous-time systems View paper
- [23] Beyond Worst-case Attacks: Robust RL with Adaptive Defense via Non-dominated Policies View paper
- [24] Performance Analysis of Least Squares of Continuous-Time Model Based on Sampling Data View paper
- [25] Provably Efficient Model-based Policy Adaptation View paper
- [26] Adaptive Experience Selection for Policy Gradient View paper
- [27] Deep Learning-based Approaches for State Space Models: A Selective Review View paper
- [28] Continuous-Time Reinforcement Learning: Algorithms, Theoretical Analysis, and Financial Applications View paper
- [29] Comparison of Model-Based and Model-Free Reinforcement Learning Algorithms for Stochastic Linear Quadratic Control View paper