Indiana University Bloomington
Luddy School of Informatics, Computing, and Engineering

Instance-Dependent Continuous-Time RL via Maximum Likelihood Estimation

Variance-adaptive learning via maximum likelihood estimation

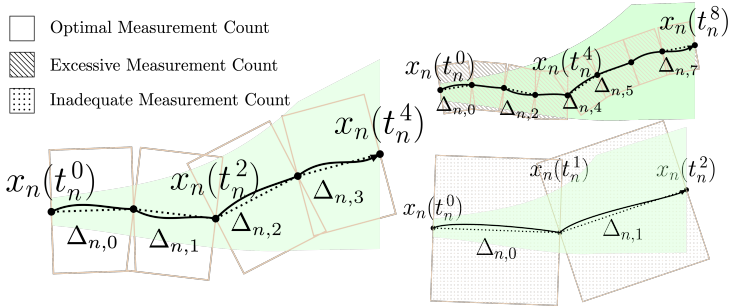
ICML 2026

Runze Zhao, Yue Yu, Ruhan Wang, Chunfeng Huang, Dongruo Zhou

Idea: learn where the system may go next, and look at the system at a resolution matched to how noisy the task is.

Problem Formulation: Measurement in Continuous Time

Question: Can a continuous-time RL algorithm adapt its learning effort to easy versus hard instances?



Paper Figure 1: the same trajectory can be observed with different measurement gaps $\Delta_{n,k}$.

Too few measurements

Misses hidden changes.

Too many measurements

Wastes measurement budget.

Right resolution

Match sampling to variability.

Instance Difficulty: Total Reward Variance

Policy value

A policy earns total reward over the whole time interval:

$$\text{total reward} = \int_0^T b(x(t), u(x(t)))\, dt.$$

Instance difficulty

The paper measures hardness by the variance of this total reward:

$$\mathrm{Var}^u = \mathrm{Var}\left[\int_0^T b(x(t), u(x(t)))\, dt\right].$$

Low-variance instance

Dynamics are stable. Repeated trials produce similar total rewards.

High-variance instance

The same policy can lead to very different rewards, so the algorithm needs more learning effort.

The goal is not just low worst-case regret; it is regret that reflects the instance actually faced by the learner.

CT-MLE with Marginal Density Estimation

Modeling step

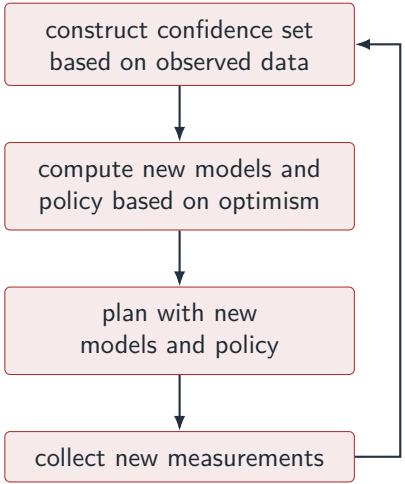
Estimate the transition density

$$p_{f,g}(x_{\text{next}} \mid x, u, \Delta)$$

from measured states.

Benefit

Avoid directly estimating derivatives $\dot{x}(t)$ from sparse, noisy trajectory snapshots.



Regret Guarantee

$$\mathrm{Regret}(N) \lesssim \widetilde{O}\left(d^2 + d\sqrt{\underbrace{\sum_{n,k} \Delta_{n,k}^2}_{\text{how coarsely we look}} + \underbrace{\sum_{n} \mathrm{Var}^{u_n}}_{\text{how noisy the instance is}}} \right)$$

Measurement term

Large gaps make the continuous trajectory harder to reconstruct.

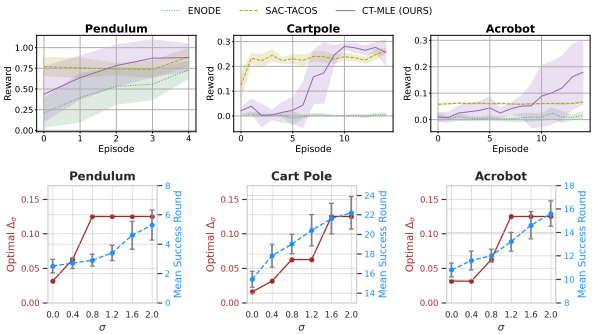
Variance term

Noisy instances require more episodes, even with good measurements.

Consequence

Choose measurement gaps according to the reward variance, and the algorithm becomes adaptive to the problem instance rather than tied to one fixed grid.

Empirical Validation



Paper Figures 2 and 3: performance across tasks, and measurement gap versus stochasticity.

- Purple curve: CT-MLE often reaches stronger final performance.
- Harder stochastic tasks show a larger advantage.
- Best measurement gap increases with volatility σ, matching the theory.

Takeaway

CT-MLE learns what can happen next, then spends measurement effort at a resolution matched to the task's variability.