# Instance-Dependent Continuous-Time Reinforcement Learning via Maximum Likelihood Estimation

Runze Zhao*[1], Yue Yu*[2], Ruhan Wang[1], Chunfeng Huang[2], Dongruo Zhou[1]
Indiana University Bloomington: [1]Luddy School of Informatics, Computing, and Engineering [2]Department of Statistics

IUB
ICML International Conference On Machine Learning 2026

## 1. PROBLEM SETTING

### How should a CTRL agent schedule measurements?

We learn continuous-time dynamics from finite measurements.

| LOW TOTAL VARIANCE | HIGH TOTAL VARIANCE |
|---|---|
| More frequent measurements per episode | Larger measurement gaps, more episodes |

**CTRL vs. DTRL: measurement scheduling**

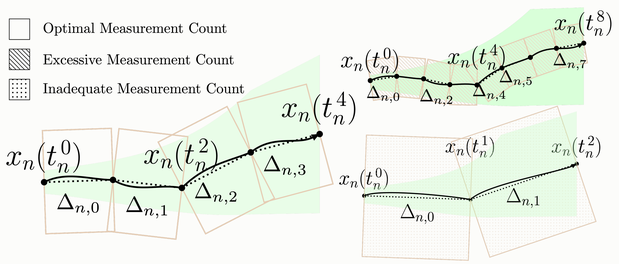
| DTRL | CTRL |
|---|---|
| Fixed uniform grid: $0, \Delta, 2\Delta, \ldots, T$. | Flexible gaps: $\Delta_{n,k} = t_{k+1}^n - t_k^n$. |
| Timing is preset. | Timing is a decision. |

**Takeaway:** DTRL fixes when to measure; CTRL chooses how long to wait before measuring again.

### Why does timing matter?

Continuous dynamics are observed through finite measurements, so the measurement grid itself becomes part of the decision.

### Measurement-grid intuition




Black dots are observations. Brown rectangles mark measurement-grid cost $\Delta_{n,k}^2$. Green shading marks total variance $\mathrm{Var}^{u_n}$.

### Key terms

| | | | |
|---|---|---|---|
| **Episode** | one trajectory over $[0, T]$ | **Measurement** | state observed at $t_n^k$ |
| **Gap** $\Delta$ | time between measurements | **Total variance** | variance of integrated reward |

### Two ways to measure complexity

**Episode complexity**
Number of episodes $N$ needed to learn an $\epsilon$-optimal policy.
More episodes give more independent trajectories.

**Measurement complexity**
Total measurements $\sum_{n=1}^N m_n$, where $m_n$ is the episode-$n$ count.
Finer grids spend more measurements inside episodes.

Scheduling trades off more episodes against more measurements within each episode.

### Mathematical model

Continuous-time controlled dynamics:

$$dx(t) = f^*(x(t), u(x(t)))\, dt + g^*(x(t), u(x(t)))\, dw(t).$$

Instance-dependent variance:

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

**Takeaway:** balance within-episode resolution against more episodes.

## 2. ALGORITHM + THEORY

**CT-MLE in one sentence: CT-MLE** estimates marginal state transition densities by MLE, not path derivatives.

### Four-step CT-MLE workflow

1. **Construct confidence set**
   based on observed data
2. **Compute new models and policy**
   based on optimism
3. **Plan**
   with new models and policy
4. **Collect new measurements**
   and add them to the observed data

New measurements are added to the observed data before the next confidence set is constructed.

### What is used for analysis?

| | |
|---|---|
| **Marginal density MLE** | learn $p_{f,g}(x' \mid u, x, \Delta)$, not $dx/dt$. |
| **Optimistic planning** | plan over statistically plausible models. |
| **Randomized additional measurement** | capture reward-integral behavior. |
| **Variance-aware measurement rule** | match measurement-grid cost to total variance. |

### What CT-MLE estimates

$$p_{f,g}(x' \mid u, x, \Delta)$$

marginal transition density over a finite gap

$x(t_n^k)$ measured state → $p_{f,g}(x' \mid u, x, \Delta)$ MLE target → $x(t_n^{k+1})$ next state

### Theory takeaway

$$\mathrm{Regret}(N) \lesssim \iota\left(d_{8\beta \boldsymbol{m}_N}\beta + \sqrt{d_{\boldsymbol{m}_N}\beta\left(\sum_{n=1}^N \boldsymbol{\Delta}_n^2 + \sum_{n=1}^N \mathrm{Var}^{u_n}\right)}\right).$$

**Note that**

| | |
|---|---|
| $\boldsymbol{m}_N$ | total measurement count up to episode $N$. |
| $d_{\boldsymbol{m}_N}$, $d_{8\beta\boldsymbol{m}_N}$ | complexity terms evaluated at the indicated measurement budgets. |
| $\beta$, $\iota$ | confidence radius and logarithmic factor. |
| $\boldsymbol{\Delta}_n^2 = \sum_k \Delta_{n,k}^2$ | measurement-grid cost for episode $n$. |
| $\mathrm{Var}^{u_n}$ | total variance of the integrated reward under policy $u_n$. |

**Match measurement-grid cost to total variance.**

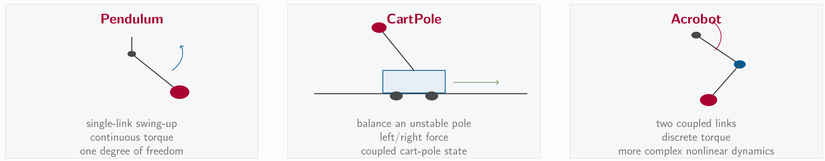
$$\textbf{SET} \quad \boldsymbol{\Delta}_n^2 := \sum_{k=0}^{m_n-1} \Delta_{n,k}^2 \;\approx\; \mathrm{Var}^{u_n}.$$

| | |
|---|---|
| **Low total variance** | smaller gaps, finer grid. |
| **High total variance** | larger gaps, more episodes. |

## 3. EXPERIMENT

### Control environments and setup

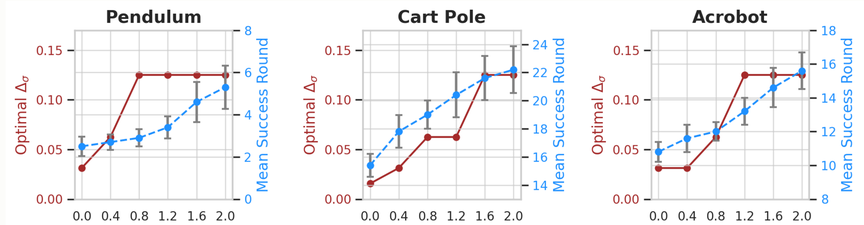
Reward curves use $\sigma = 2.0$. The gap sweep uses fixed uniform gaps $\Delta = 2^{-i}$, $i = 0, \ldots, 7$.




Tasks differ in control structure; reward curves use the same injected noise level.

### Evidence for theory: optimal gap increases with stochasticity

**Higher stochasticity → larger selected gap $\Delta_\sigma$.**

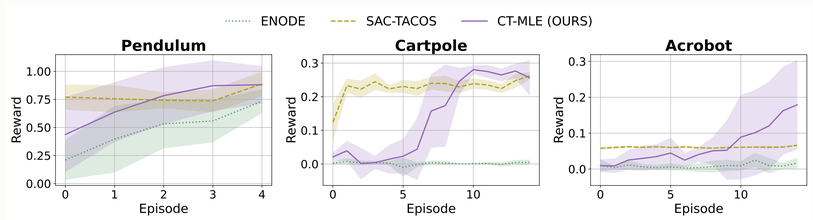



Red series: selected optimal gap $\Delta_\sigma$; the trend is consistent with the variance-aware rule.

**Illustration.**

- Within each task, increase $\sigma$; red gives $\Delta_\sigma$, blue gives the mean success round.
- **Theorem link:** higher $\sum_n \mathrm{Var}^{u_n}$ supports larger $\sum_n \boldsymbol{\Delta}_n^2$, hence larger measurement gaps.
- **Intuition:** more stochastic tasks need more episodes and coarser within-episode observations.

### Reward curves confirm competitive policy quality




At $\sigma = 2.0$, CT-MLE remains competitive; separation is largest on Acrobot.

Reward curves check policy quality; the gap sweep checks the measurement rule.

### Experiment takeaway

Adaptive measurement schedules allocate computation according to instance difficulty: harder stochastic environments benefit from more training episodes rather than denser measurements within each episode.

**Measure finely when variance is low; collect more episodes when variance is high.**

Marginal-density MLE supports a variance-aware measurement schedule.

Paper Code Homepage