

A control-theoretical perspective of continuous time reinforcement learning

Xin Guo *UC Berkeley*

email: xinguo@berkeley.edu

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

June 23-25, 2026, Oslo



arxiv

Continuous-time-space RL: recap of Day One

- ▶ Consider continuous-time dynamics driven by SDE over a finite time horizon
- ▶ dynamic:

$$dX_t^a = b(t, X_t^a, a_t)dt + \sigma(t, X_t^a, a_t)dW_t, \quad t \in [0, T], \quad X_0^a = x_0 \sim \nu$$

- ▶ reward:

$$\mathbb{E} \left[\int_0^T e^{-\beta t} r(t, X_t^a, a_t) dt + e^{-\beta T} g(X_T^a) \right]$$

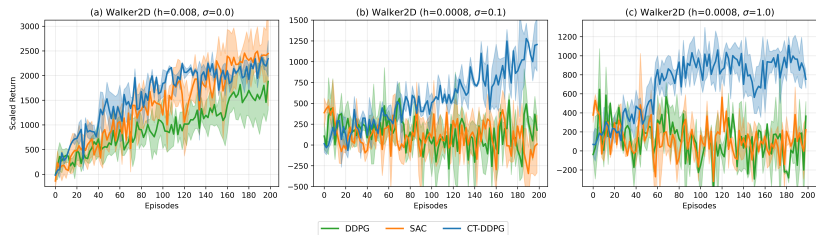
- ▶ The coefficients b, σ, r, g are **unknown**.

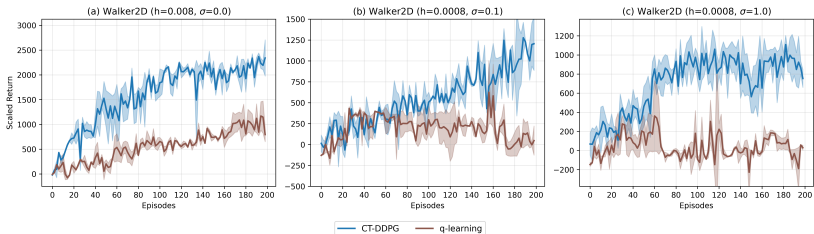
Reinforcement learning: recap

- ▶ Significant development in discrete-time-space, while limited progress in continuous-time-space
- ▶ **Q:** What about discretization for continuous-time-space RL?

Bottleneck: inefficiency of discrete-time RL in continuous framework

Standard discrete-time RL algorithms typically degrade in **continuous-time, stochastic** environments





Deterministic policy: recap

- ▶ In classical control, strict control is conceptually more natural than relaxed control
- ▶ For sufficiently diffusive controlled process, exploration is **unnecessary**, for instance in regret analysis for LQR and LCV cases (Basei, G. Hu, Zhang (2022), G., Hu, Zhang (2023)).

Deterministic policy: recap

Approximate the (optimal) deterministic policy in a parametric form $\{\mu_\phi\}_{\phi \in \mathbb{R}^k}$, and optimize the policy parameterization by gradient methods.

- ▶ Consider a parameterized Markov policy $\mu_\phi : [0, T] \times \mathbb{R}^n \rightarrow \mathcal{A}$

$$dX_t^\phi = b(t, X_t^\phi, \mu_\phi(t, X_t^\phi))dt + \sigma(t, X_t^\phi, \mu_\phi(t, X_t^\phi))dW_t$$

- ▶ Maximize the value function over $\phi \in \mathbb{R}^k$:

$$V^\phi(t, x) := \mathbb{E} \left[\int_t^T e^{-\beta(s-t)} r(s, X_s^\phi, \mu_\phi(s, X_s^\phi)) ds + e^{-\beta(T-t)} g(X_T^\phi) \mid X_t^\phi = x \right]$$

Theorem (Martingale characterization)

Let $\hat{V} \in C^{1,2}([0, T] \times \mathbb{R}^n)$ and $\hat{q} \in C([0, T] \times \mathbb{R}^n \times \mathcal{A})$ satisfy

$$\hat{V}(T, x) = g(x), \quad \hat{q}(t, x, \mu_\phi(t, x)) = 0,$$

and for a neighborhood $\mathcal{O}_{\mu_\phi(t, x)}$ of $\mu_\phi(t, x)$, it holds that $\forall t, x, a \in \mathcal{O}_{\mu_\phi(t, x)}$, there exists a control process $(\alpha_s)_{s \geq t}$ with $\lim_{s \searrow t} \alpha_s = a$, a.s.

$$\left(e^{-\beta s} \hat{V}(s, X_s^{t, x, a}) + \int_t^s e^{-\beta u} (r - \hat{q})(u, X_u^{t, x, a}, \alpha_u) du \right)_{s \in [t, T]} \text{ is a martingale}$$

where

$$dX_s^{t, x, a} = b(s, X_s^{t, x, a}, \alpha_s) ds + \sigma(s, X_s^{t, x, a}, \alpha_s) dW_s, \quad X_t^{t, x, a} = x,$$

Then $\forall (t, x, a) \in [0, T] \times \mathbb{R}^n \times \mathcal{O}_{\mu_\phi(t, x)}$,

$$\hat{V}(t, x) = V^\phi(t, x), \quad \hat{q}(t, x, a) = A^\phi(t, x, a).$$

Remarks on martingale characterization

(1) Martingale condition + local continuity:

$$\left\{ \begin{array}{l} \left(e^{-\beta(s-t)} \hat{V}(s, X_s^{t,x,a}) + \int_t^s e^{-\beta(u-t)} (r - \hat{q})(u, X_u^{t,x,a}, \alpha_u) du \right)_{s \in [t, T]} \\ \lim_{s \searrow t} \alpha_s = a \end{array} \right.$$

$$\Rightarrow \quad \hat{q}(t, x, a) = \mathcal{L}[\hat{V}](t, x, a) + r(t, x, a)$$

(2) Bellman condition: $\hat{q}(t, x, \mu_\phi(t, x)) = 0$

$$\Rightarrow \quad \mathcal{L}[\hat{V}](t, x, \mu_\phi(t, x)) + r(t, x, \mu_\phi(t, x)) = 0$$

(3) Terminal value constraint: $\hat{V}(T, x) = g(x)$

Policy evaluation

Let V_θ, q_ψ be the parameterized value and advantage function.
For martingale condition, one can utilize semi-gradient Temporal Difference (TD) update:

$$\Delta := V_\theta(t, x_t) - \int_t^{t+\delta} e^{-\beta(s-t)} [r - q_\psi](s, x_s, a_s) ds - e^{-\beta\delta} V_\theta(t+\delta, x_{t+\delta})$$

$$\theta \leftarrow \theta - \eta \partial_\theta V_\theta(t, x_t) \cdot \Delta$$

$$\psi \leftarrow \psi - \eta \partial_\psi q_\psi(t, x_t, a_t) \cdot \Delta$$

Issues of one-step TD in continuous RL: variance blow-up

One-step (stochastic) semi-gradient update is typically used in RL (???)

$$G_{\theta,h} := \frac{1}{h} \mathbb{E} \left[\partial_{\theta} V_{\theta}(\tilde{x}_t) \left(V_{\theta}(\tilde{x}_t) - (r_t - A_{\psi}(\tilde{x}_t, a_t)) \cdot h - e^{-\beta h} V_{\theta}(\tilde{x}_{t+h}) \right) \right]$$

$$g_{\theta,h} := \frac{1}{h} \left[\partial_{\theta} V_{\theta}(\tilde{x}_t) \left(V_{\theta}(\tilde{x}_t) - (r_t - A_{\psi}(\tilde{x}_t, a_t)) \cdot h - e^{-\beta h} V_{\theta}(\tilde{x}_{t+h}) \right) \right]$$

Proposition (one-step TD suffers from variance blow up)

Suppose $\underline{C} \cdot I \preceq \sigma \sigma^{\top} \preceq \overline{C} \cdot I$ for some $0 < \underline{C} \leq \overline{C}$.

$$\lim_{h \rightarrow 0} \mathbb{E}[g_{\theta,h}] = \lim_{h \rightarrow 0} G_{\theta,h} = \Theta(1),$$

$$\lim_{h \rightarrow 0} h \cdot \text{Var}(g_{\theta,h}) = \Theta(1).$$

Implementation with discretization (CT-DDPG)

- ▶ h being the discretization stepsize.
- ▶ $\tilde{x}_t$ the concatenation (t, x_t) for notation compactness
- ▶ Multi-step TD:

$$\Delta = V_{\theta}(\tilde{x}_{kh}) - \sum_{\ell=0}^{L-1} e^{-\beta \ell h} [r_{(k+\ell)h} - q_{\psi}(\tilde{x}_{(k+\ell)h}, a_{(k+\ell)h})] h - e^{-\beta Lh} V_{\theta}(\tilde{x}_{(k+L)h})$$

- ▶ $L > 1$ is the essential reason for the efficiency and stability in practice

Stability of multi-step TD in CT-DDPG

$$G_{\theta,h,L} :=$$

$$\mathbb{E}\left[\partial_{\theta} V_{\theta}(\tilde{x}_t) \left(V_{\theta}(\tilde{x}_t) - \sum_{\ell=0}^{L-1} e^{-\beta \ell h} [r_{t+\ell h} - q_{\psi}(\tilde{x}_{t+\ell h}, a_{t+\ell h})] h - e^{-\beta L h} V_{\theta}(\tilde{x}_{t+L h}) \right)\right]$$

$$g_{\theta,h,L} :=$$

$$\partial_{\theta} V_{\theta}(\tilde{x}_t) \left(V_{\theta}(\tilde{x}_t) - \sum_{\ell=0}^{L-1} e^{-\beta \ell h} [r_{t+\ell h} - q_{\psi}(\tilde{x}_{t+\ell h}, a_{t+\ell h})] h - e^{-\beta L h} V_{\theta}(\tilde{x}_{t+L h}) \right)$$

Proposition (multi-step TD is more stable)

If $Lh \equiv \delta > 0$, then

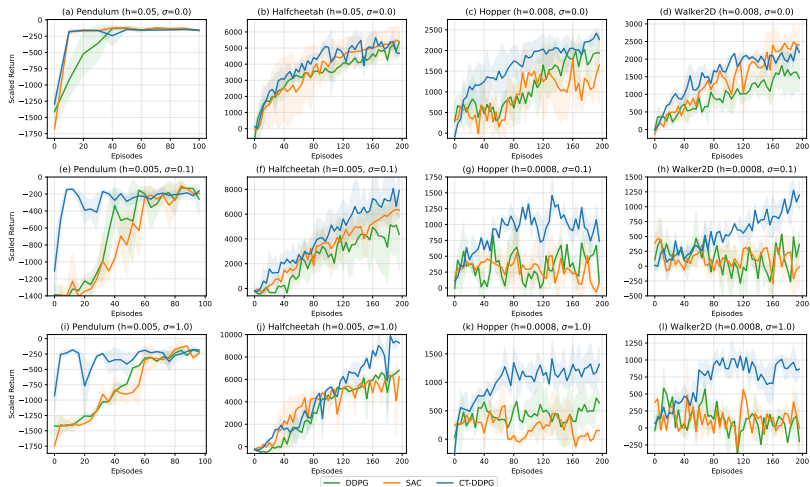
$$\lim_{h \rightarrow 0} \mathbb{E}[g_{\theta,h,\frac{\delta}{h}}] = \lim_{h \rightarrow 0} G_{\theta,h,\frac{\delta}{h}} = \Theta(1).$$

$$\overline{\lim}_{h \rightarrow 0} \text{Var}(g_{\theta,h,\frac{\delta}{h}}) = \mathcal{O}(1).$$

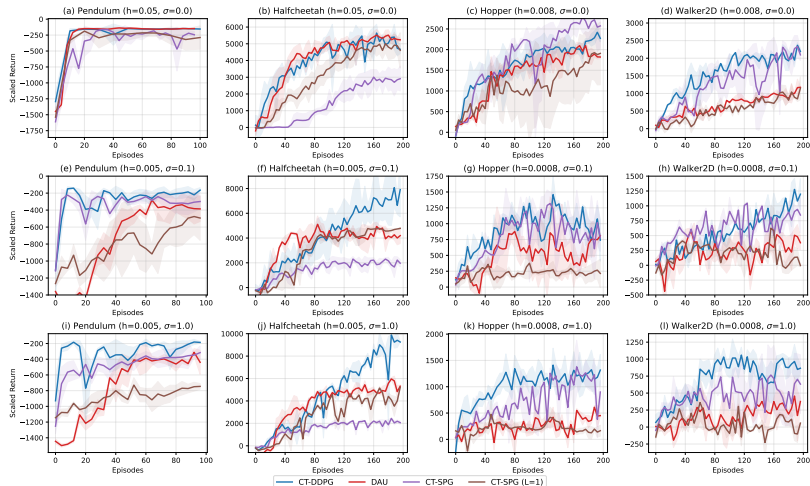
Experiments

- ▶ Evaluate CT-DDPG on a suite of continuous robotic control benchmarks from OpenAI Gym, with varying discretization stepsize and dynamic noise levels
- ▶ Compare CT-DDPG against discrete-time algorithms including DDPG (?), SAC (?), and continuous-time algorithm with stochastic Gaussian policy and q-learning (?)
 - ▶ The original q-learning uses one-step TD ($L = 1$), we evaluate q-learning with multi-step TD ($L > 1$) for fair comparison
 - ▶ We also test CT-DDPG with $L = 1$ for ablations

Results: advantages over discrete-time algorithms



Results: advantages over continuous-time stochastic policy algorithms



Additional References I

- Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International conference on machine learning*, pages 1861–1870. Pmlr, 2018.
- Yanwei Jia and Xun Yu Zhou. q-learning in continuous time. *Journal of Machine Learning Research*, 24(161):1–61, 2023.
- David Silver, Guy Lever, Nicolas Heess, Thomas Degris, Daan Wierstra, and Martin Riedmiller. Deterministic policy gradient algorithms. In *International conference on machine learning*, pages 387–395. Pmlr, 2014.
- Corentin Tallec, Léonard Blier, and Yann Ollivier. Making deep q-learning methods robust to time discretization. In *International Conference on Machine Learning*, pages 6096–6104. PMLR, 2019.

Extension to McKean-Vlasov Control

Consider the state dynamics

$$\begin{aligned} \dot{X}_s^{t,\nu} &= b(s, X_s^{t,\nu}, \mathbb{P}_{X_s^{t,\nu}}) ds + \sigma(s, X_s^{t,\nu}, \mathbb{P}_{X_s^{t,\nu}}) dW_s, \\ X_t^{t,\nu} &\sim \nu, \end{aligned}$$

and the associated value function

$$\mathbb{E} \left[\int_t^T r(s, X_s^{t,\nu}, \mathbb{P}_{X_s^{t,\nu}}) ds + g(X_T^{t,\nu}, \mathbb{P}_{X_T^{t,\nu}}) \right].$$

Dynamic programming principle for MFC

- MFC is time-inconsistent
 - ▶ Strict controls/without learning: Pham and Wei (2016, 2017), Bayraktar, Cosso, Pham (2018), Wu and Zhang (2019), Djete, Possamai, and Tan (2022), Talbi, Touzi, and Zhang (2023)
 - ▶ Relaxed control for discrete-time MFC with learning: Gu, G., Wei, and Xu (2023)

DPP for discrete-time MFC with learning (Gu, G. Wei, and Xu, (2023))

Idea:

- ▶ Identify the “correct” Q function: with a wrong form of Q function, Q-table will converge to different values with different initial population distribution
- ▶ Need to work with the correct “state-action” space
- ▶ Similar to the MFC theory, “lift” the state-action space into their probability measure space

Q function on the probability measure space

- ▶ $Q(s, a)$ No $Q(\mu, h)$ Yes
- ▶ Define Integrated Q (IQ) function for MFC

$$Q(\mu, h) = \sup_{\pi \in \Pi_1} Q^\pi(\mu, h), \quad \text{for any } \mu \in \mathcal{P}(\mathcal{S}), \quad h \in \mathcal{H},$$

$$Q^\pi(\mu, h) = \mathbb{E}^\pi \left[\sum_{t=0}^{\infty} \gamma^t r(s_t, \mu_t, a_t) \mid s_0 \sim \mu, a_0 \sim h(s_0), a_t \sim \pi_t(s_t, \mu_t) \right].$$

Bellman Equation for IQ Function (Gu, G., Wei, Xu, 2023)

For any $\mu \in \mathcal{P}(\mathcal{S})$ and $h \in \mathcal{H}$,

$$Q(\mu, h) = \hat{r}(\mu, h) + \gamma \sup_{h' \in \mathcal{H}} Q(\Phi(\mu, h), h').$$

- ▶ $\mathcal{H} := \{h : \mathcal{S} \rightarrow \mathcal{P}(\mathcal{A})\}$: set of local policies $\underbrace{\mathcal{P}(\mathcal{A}) \times \dots \times \mathcal{P}(\mathcal{A})}_{|\mathcal{S}|}$
- ▶ Aggregated reward: $\hat{r}(\mu, h) := \sum_{s,a} \mu(s) h(s, a) r(s, a, \mu)$
- ▶ Aggregated dynamics
 - ▶ $\Phi(\mu, h) := \sum_{s,a} P(s, \mu, a) \mu(s) h(s, a)$: aggregated dynamics
 - ▶ $\mu_{t+1} = \Phi(\mu_t, h)$: distribution at time $t + 1$, flow property
 - ▶ $\mathcal{H}$ is key for DPP of the IQ function

- ▶ Let $X^{t,\xi,\theta'}$ satisfy for all $s \in [t, T]$,

$$dX_s^{t,\xi,\theta'} = b(s, X_s^{t,\xi,\theta'}, \mathbb{P}_{X_s^{t,\xi,\theta'}, \theta'})ds + \sigma(s, X_s^{t,\xi,\theta'}, \mathbb{P}_{X_s^{t,\xi,\theta'}, \theta'})dW_s,$$

- ▶ $X_t^{t,\xi,\theta'} = \xi$, with $\xi \in L^2(\mathcal{F}_t; \mathbb{R}^n)$ having the law μ
- ▶ Denoting $\mathbb{P}_r^{t,\mu,\theta'} = \mathbb{P}_{X_r^{t,\xi,\theta'}}$

Key characterization for MFC

Key characterization for MFC

Let $\theta \in \mathbb{R}^k$, let $\hat{V} \in C^{1,2}([0, T] \times \mathcal{P}_2(\mathbb{R}^n))$, and $\hat{q} \in C([0, T] \times \mathcal{P}_2(\mathbb{R}^n) \times \mathbb{R}^k)$. If for all $(t, \mu) \in [0, T] \times \mathcal{P}_2(\mathbb{R}^n)$,

$$\hat{V}(T, \mu) = \bar{g}(\mu), \quad \hat{q}(t, \mu, \theta) = 0,$$

and there exists a neighborhood $\mathcal{O}_{t,\mu}(\theta) \subset \mathbb{R}^k$ of θ such that for all $\theta' \in \mathcal{O}_{t,\mu}(\theta)$ and $s \in [t, T]$,

$$\begin{aligned} & e^{-\beta s} \hat{V}(s, \mathbb{P}_s^{t,\mu,\theta'}) - e^{-\beta t} \hat{V}(t, \mu) \\ & + \int_t^s e^{-\beta u} [\bar{r}(u, \mathbb{P}_u^{t,\mu,\theta'}, \theta') - \hat{q}(u, \mathbb{P}_u^{t,\mu,\theta'}, \theta')] du = 0, \end{aligned}$$

Then for all $(t, \mu, \theta') \in [0, T] \times \mathcal{P}_2(\mathbb{R}^n) \times \mathcal{O}_{t,\mu}(\theta)$,

$$\hat{V}(t, \mu) = V^\theta(t, \mu), \quad \hat{q}(t, \mu, \theta') = A[V^\theta](t, \mu, \theta').$$

Learning via local policy in continuous-time MFC

- ▶ The characterization is from the flow property and the stability of the law of McKean–Vlasov system with respect to perturbations of the initial condition and model and policy parameters
- ▶ Similar to $\mathcal{H}$ for the IQ function of discrete-time MFC

$$h^\varphi(t, x, \mu, \theta) = h(t, x, \varphi_\theta(t, x, \mu), (id, \varphi_\theta(t, \cdot, \mu))_{\#}\mu).$$

- ▶ $(id, \varphi(t, \cdot, \mu))_{\#}\mu \in \mathcal{P}_2(\mathbb{R}^n \times \mathbb{R}^d)$ is the induced state-action measure such that $\forall B \in \mathcal{B}(\mathbb{R}^n \times \mathbb{R}^d)$,

$$[(id, \varphi(t, \cdot, \mu))_{\#}\mu](B) = \mu(\{x \in \mathbb{R}^n \mid (x, \varphi(t, x, \mu)) \in B\})$$

- ▶ Martingale characterization and learning algorithms for decoupled version of $\hat{V}(t, \mu)$ and $q(t, \mu, \theta)$, analogous to the single-agent RL

Relevant References

- ▶ Z. Cheng, X. Guo, and Y. Zhang, *Deterministic policy gradient for reinforcement learning with continuous time and space*, arXiv:2509.23711, 2025.
- ▶ Z. Cheng, X. Guo, H. Pham, and Y. Zhang, *Deterministic actor-critic learning for extended McKean-Vlasov control problems*, preprint.