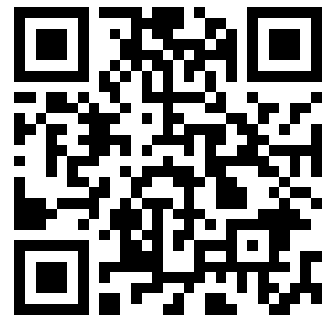
Deterministic Policy Gradient for Reinforcement Learning with Continuous Time and Space

Xin Guo, UC Berkeley
based on joint work with Ziheng Cheng (UC Berkeley)
and Yufei Zhang (Imperial College)

April 23, 2026, IMSI



arxiv

The talk is based on the following works

- 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, *Model-free sensitivity formula for McKean-Vlasov SDEs with application to policy gradient learning*, working paper.

Reinforcement Learning

- RL is a paradigm of learning through interaction with environment

(a) Go

deepseek R1

(b) Large Language Model

(c) Robotic control

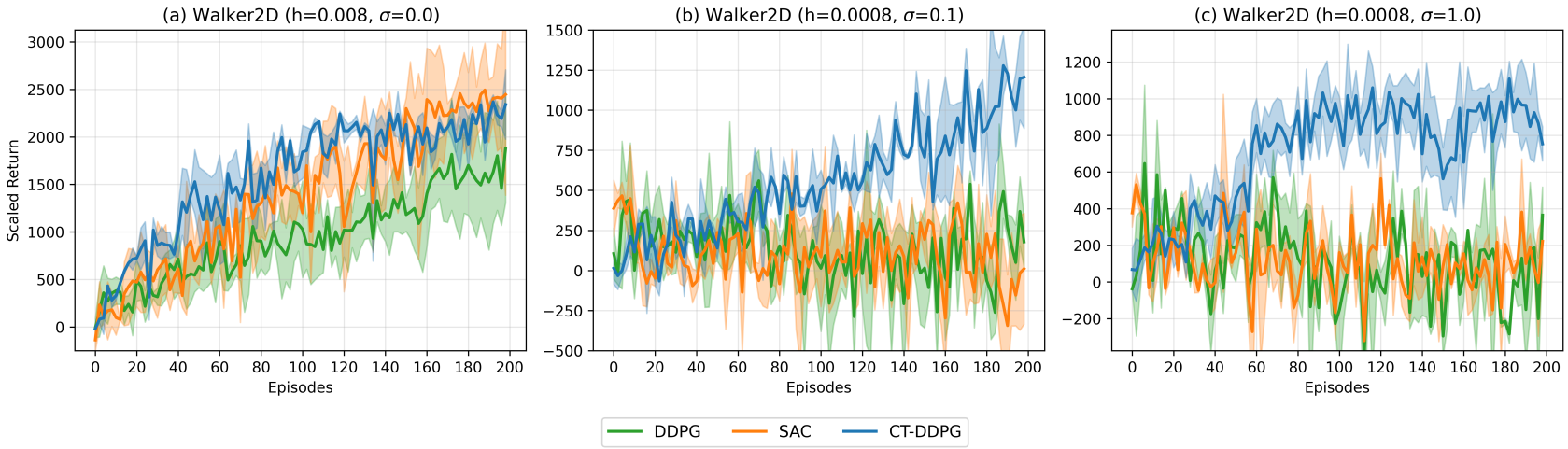
(d) Financial trading

Reinforcement Learning

- Significant development in discrete-time-space, while limited progress in continuous-time-space
- A natural approach for continuous-time-space RL is discretization

Bottleneck: inefficiency of discrete-time RL in continuous framework

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



Goal

Develop theoretical foundation and algorithms that achieves stability and efficiency for continuous-time-space RL

Problem formulation for continuous-time-space RL

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

$$\mathrm{d}X_t^{\boldsymbol{a}} = b(t, X_t^{\boldsymbol{a}}, a_t)\mathrm{d}t + \sigma(t, X_t^{\boldsymbol{a}}, a_t)\mathrm{d}W_t, \ t \in [0, T], \quad X_0^{\boldsymbol{a}} = x_0 \sim \nu$$

- reward:

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

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

Existing approach: exploratory formulation with stochastic policy [Wang, Zariphopoulou and Zhou (2020)]

Consider stochastic policies $\pi : [0, T] \times \mathbb{R}^d \to \mathcal{P}(A)$, and the corresponding exploratory state dynamics:

$$\mathrm{d}X_t^\pi = \tilde{b}^\pi(t, X_t^\pi)dt + \tilde{\sigma}^\pi(t, X_t^\pi)dW_t,$$

where

$$\tilde{b}^\pi(t, x) = \int_A b(t, x, a)\pi(da|t, x),$$

$$\tilde{\sigma}^\pi(t, x) = \sqrt{\int_A \sigma^2(t, x, a)\pi(da|t, x)}.$$

with the reward function

$$\tilde{r}^\pi(t, x) = \int_A r(t, x, a)\pi(a|t, x)da$$

and $\gamma > 0$ is a given parameter.

Stochastic policy for continuous-time-space

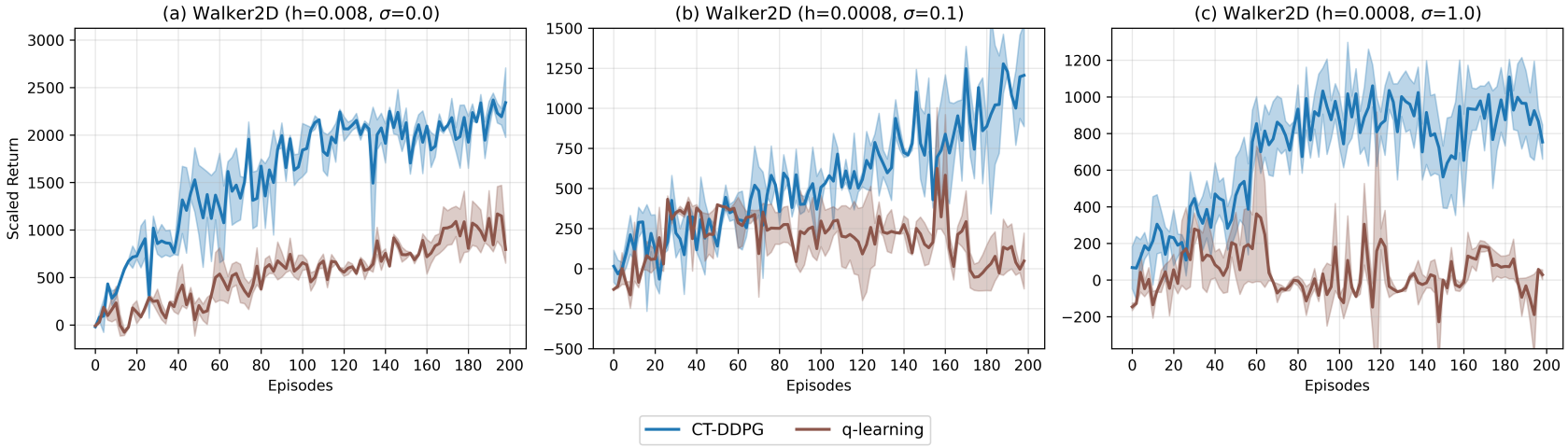
- Stochastic policy follows directly from discrete-time-space, where mixed strategies are necessary for algorithmic purpose
- Stochastic policy is natural with entropy regularization, which encourages exploration: the optimal policy is stochastic and has full support over the action space.
- Confirmation of optimal policy in LQ case being Gaussian
- Based on this framework, discrete-time RL algorithms are extended to continuous-time, q-learning [Jia and Zhou, 2023], policy gradient and actor-critic [Jia and Zhou, 2022], PPO [Zhao et al., 2023].

Drawback with stochastic policy for continuous-time-space

- Exploratory dynamics requires continuously sampling independent actions over continuous states, which is infeasible both theoretically and practically.
- (Optimal) exploratory policy is given by state-dependent Gibbs measures, which can be hard to sample.
- Frequent sampling makes actions discontinuous over time, in contrast to deterministic policy
- Recovering a deterministic policy from a learned stochastic one is challenging (except the LQ setting).

Drawback with stochastic policy for continuous-time-space

- Exploratory dynamics requires continuously sampling independent actions over continuous states, which is infeasible both theoretically and practically.
- (Optimal) exploratory policy is given by state-dependent Gibbs measures, which can be hard to sample.
- Frequent sampling makes actions discontinuous over time, in contrast to deterministic policy
- Recovering a deterministic policy from a learned stochastic one is challenging (except the LQ setting).



Is stochastic policy necessary?

- By classical control, when coefficients are given, then under mild conditions, the optimal control is given by

$$a_t^* = \mu^*(t, X_t^{\mu^*}), \quad \forall t \geq 0,$$

  - $\mu^* : [0, T] \times \mathbb{R}^d \to A$ is an (optimal) deterministic policy (a.k.a. feedback control).
  - μ^* is continuous/differentiable with proper regularity conditions.
- 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)).

This talk

Directly learn a deterministic policy

- Derives deterministic policy gradient (DPG) for continuous-time-space RL
- Establishes the martingale characterization of DPG
- Implements a new algorithm (CT-DDPG) with improved stability and faster convergence over existing one.

Table of Contents

General idea of deterministic policy

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 $\mu_\phi : [0, T] \times \mathbb{R}^n \to \mathcal{A}$

$$\mathrm{d}X_t^\phi = b(t, X_t^\phi, \mu_\phi(t, X_t^\phi))\mathrm{d}t + \sigma(t, X_t^\phi, \mu_\phi(t, X_t^\phi))\mathrm{d}W_t$$

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

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

Deterministic policy gradient (DPG) formula

Theorem (DPG Formula)

$$\partial_\phi V^\phi(t,x) = \mathbb{E}\left[\int_t^T e^{-\beta(s-t)}\partial_\phi \mu_\phi(s,X_s^\phi)^\top \partial_a A^\phi(s,X_s^\phi,\mu_\phi(s,X_s^\phi))\mathrm{d}s \,\middle|\, X_t^\phi = x\right],$$

where $A^\phi(t,x,a) := \mathcal{L}[V^\phi](t,x,a) + r(t,x,a)$, *with* $\mathcal{L}[\varphi](t,x,a) := \partial_t\varphi(t,x) - \beta\varphi(t,x) + b(t,x,a)^\top\partial_x\varphi(t,x) + \frac{1}{2}Tr(\Sigma(t,x,a)\partial_{xx}^2\varphi(t,x))$.

- A^ϕ is analogous to the advantage function in discrete-time DPG [Silver et al., 2014], and the q-function for stochastic policy gradient [Jia and Zhou, 2023].
- $A^{\Delta t,\phi}(t,x,a) := \dfrac{Q^{\Delta t,\phi}(t,x,a) - V^{\Delta t,\phi}(t,x)}{\Delta t}$ is the discrete-time advantage rate function.

How to learn A^ϕ:

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.*

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

where

$$\mathrm{d}X_s^{t,x,a} = b(s,X_s^{t,x,a},\alpha_s)\mathrm{d}s + \sigma(s,X_s^{t,x,a},\alpha_s)\mathrm{d}W_s, \quad X_t^{t,x,a} = x,$$

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

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

Important components in martingale characterization

(1) Martingale condition + local continuity:

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

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

Important components in martingale characterization

(1) Martingale condition + local continuity:

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

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

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

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

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

Important components in martingale characterization

(1) Martingale condition + local continuity:

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

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

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

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

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

Efficiency improvement over stochastic policy

- Exploring the neighborhood of current actions, instead of the entire action space
- Bellman condition:

$$\hat{q}(t, x, \mu_\phi(t, x)) = 0$$

  - Compared with the stochastic policy where the Bellman condition is replaced by

$$\mathbb{E}_{\pi(\mathrm{d}a|t,x)}[\hat{q}(t, x, a) - \gamma \log \pi(a|t, x)] = 0,$$

    which requires additional Monte Carlo steps to evaluate over the entire action space

Table of Contents

Policy evaluation

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

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

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

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

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

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

One-step (stochsatic) semi-gradient update is typically used in RL [Haarnoja et al., 2018, Tallec et al., 2019, Jia and Zhou, 2023]:

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

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

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* $0 < \underline{C} \leq \overline{C}$.

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

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

Stability of multi-step TD in CT-DDPG

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

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

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

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

Proposition (multi-step TD is more stable)

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

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

$$\overline{\lim_{h\to 0}} \mathrm{Var}(g_{\theta,h,\frac{\delta}{h}}) = \mathcal{O}(1).$$

Remarks

(1) Effect of $1/h$ scaling

- In multi-step TD, we omit the $1/h$ factor in contrast to one-step TD, crucial for preventing variance from blowing up
- If removed in one-step TD, then $G_{\theta,h}$ vanishes as $h \to 0$

Fundamental drawback of one-step TD in continuous-time RL

Remarks

(1) Effect of $1/h$ scaling

- In multi-step TD, we omit the $1/h$ factor in contrast to one-step TD, crucial for preventing variance from blowing up
- If removed in one-step TD, then $G_{\theta,h}$ vanishes as $h \to 0$

Fundamental drawback of one-step TD in continuous-time RL

(2) previous analysis of one-step TD

- Jia and Zhou [2022] considers the objective:

$$\min_{\theta} \frac{1}{h^2} \mathbb{E}_{\tilde{x}} \left(V_\theta(\tilde{x}_t) - r_t \cdot h - e^{-\beta h} V_\theta(\tilde{x}_{t+h}) \right)^2,$$

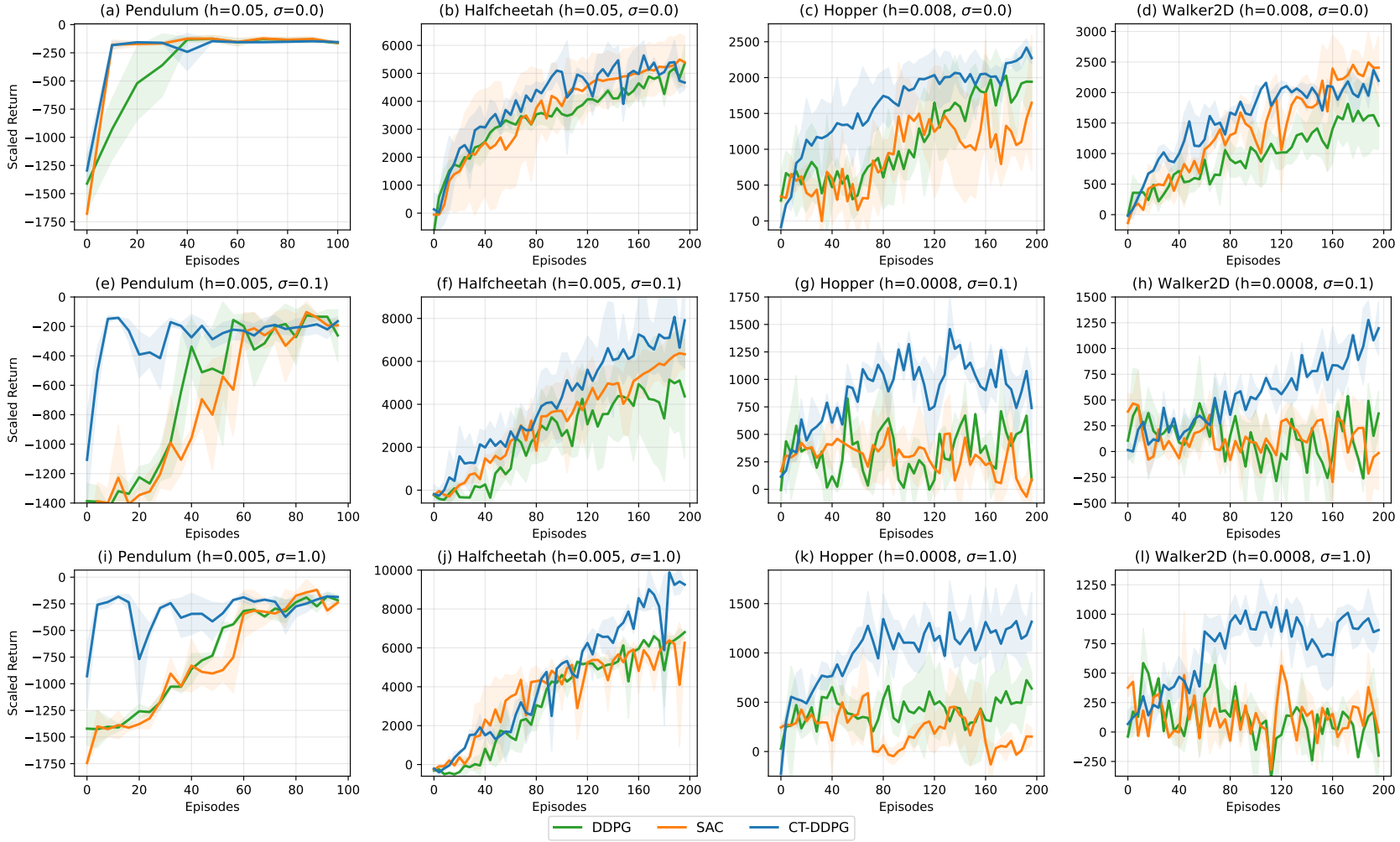
- The minimizer of the objective above does not converge to true value function as $h \to 0$
- not aligned with semi-gradient update in practice

Table of Contents

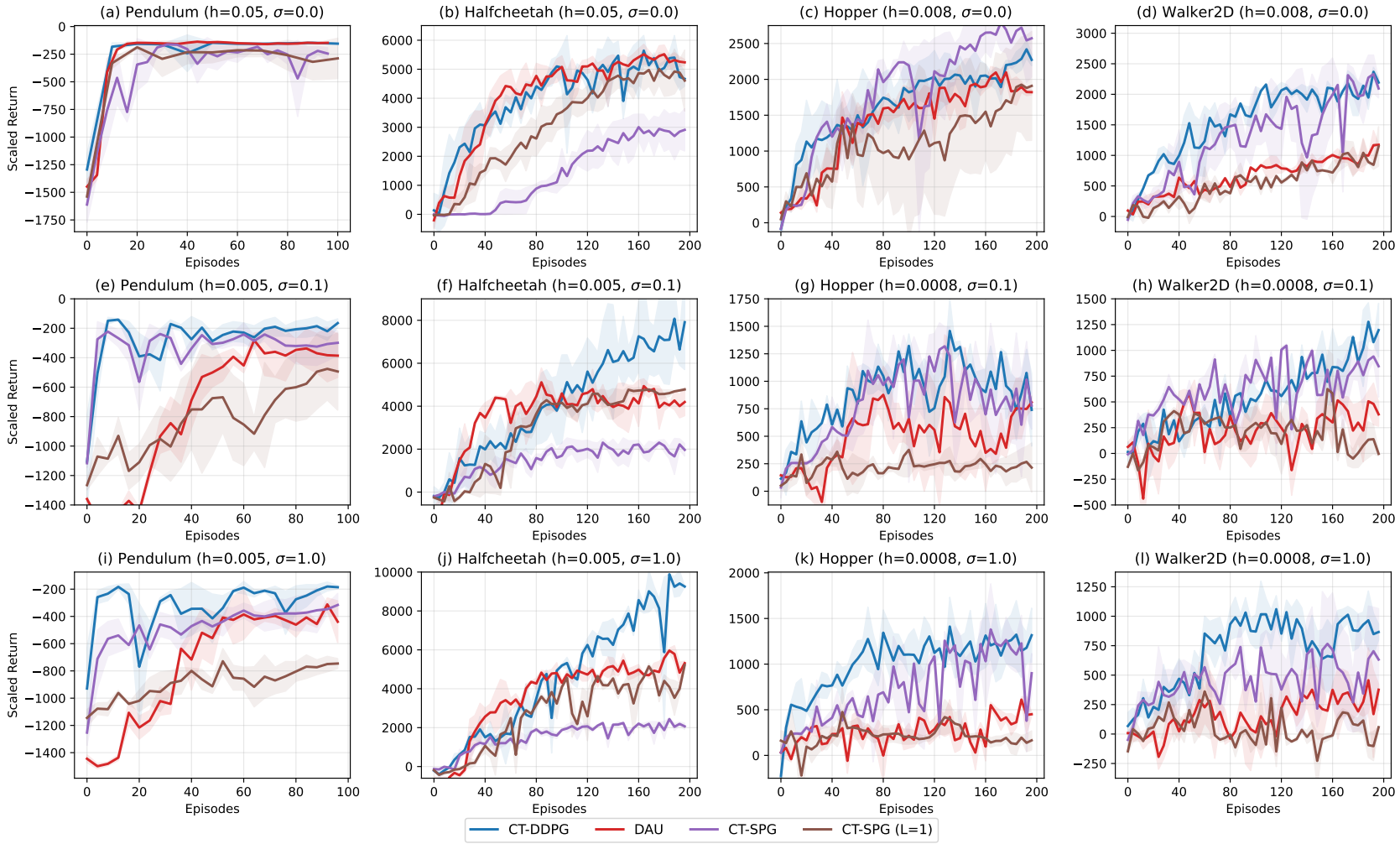
Experiments setup

- 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 [Silver et al., 2014], SAC [Haarnoja et al., 2018], and continuous-time algorithm with stochastic Gaussian policy and q-learning [Jia and Zhou, 2023]
  - 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



Conclusions

- A mathematical framework for model-free deterministic policy gradient (DPG) in continuous-time RL and learning of McKean-Vlasov control problems
- Martingale characterization shows DPG with improved efficiency over stochastic policy
  Bellman condition for stochastic policy is unimplementable in high-dim space and deep RL
- Multi-step TD is crucial for empirical success
  Theoretical insights of the failure of discrete-time RL in continuous and stochastic setting

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. Policy evaluation and temporal-difference learning in continuous time and space: A martingale approach. *Journal of Machine Learning Research*, 23 (154):1–55, 2022.

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.

References II

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.

Hanyang Zhao, Wenpin Tang, and David Yao. Policy optimization for continuous reinforcement learning. *Advances in Neural Information Processing Systems*, 36:13637–13663, 2023.

Extension to McKean-Vlasov Control

Consider the state dynamics

$$dX_s^{t,\nu,\phi} = b(s, X_s^{t,\nu,\phi}, \mathbb{P}_{X_s^{t,\nu,\phi}}, \phi)ds + \sigma(s, X_s^{t,\nu,\phi}, \mathbb{P}_{X_s^{t,\nu,\phi}}, \phi)dW_s,$$
$$X_t^{t,\nu,\phi} \sim \nu,$$

and the associated value function

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

For the application to McKean-Vlasov Control, for all $f \in \{b, \sigma, r\}$, $f^\phi(t, x, \nu, \phi) := f(t, x, \nu, \mu_\phi(t, x, \nu))$.

DPG in MFC

Theorem

$$\partial_\phi V^\phi(t,x,\nu) = \int_t^T \partial_\phi A[V^\phi](s, \mathbb{P}_{X_s^{t,\nu,\phi}}, \phi) ds,$$

where $A[\varphi](t,\nu,\phi) := \mathcal{L}^\phi[\varphi](t,\nu) + \langle r(t,\cdot,\nu,\phi),\nu\rangle$, *and*

$$\begin{aligned}\mathcal{L}^\theta[\varphi](t,\nu) :=& \partial_t\varphi(t,\nu) + \mathbb{E}_{\xi\sim\nu}\Big[b(t,\xi,\nu,\theta)\cdot\partial_\mu\varphi(t,\nu)(\xi) \\ &+ \frac{1}{2}\Sigma(t,\xi,\nu,\theta)\colon \partial_v\partial_\mu\varphi(t,\nu)(\xi)\Big].\end{aligned}$$

- A model-free characterization holds for $A[V^\phi](t,\nu,\phi)$.