

Continuous-time Reinforcement Learning from discrete-time data

Yuhua Zhu - University of California, Los Angeles

- UQ for plasma dynamics
- Sensitivity analysis based on Hypocoercivity

- I listened talks by
- Jean-Pierre Fouque
 - Halil Mete Soner
 - Sebastian Jaimungal
 - Mathieu Lauriere
 -



PhD thesis

Postdoc

2023 summer
ICIAM

- Anisotropic CBO for high-dimensional nonconvex optimization
- Reinforce under MD

- FedCBO

arXiv > math > arXiv:2405.12535

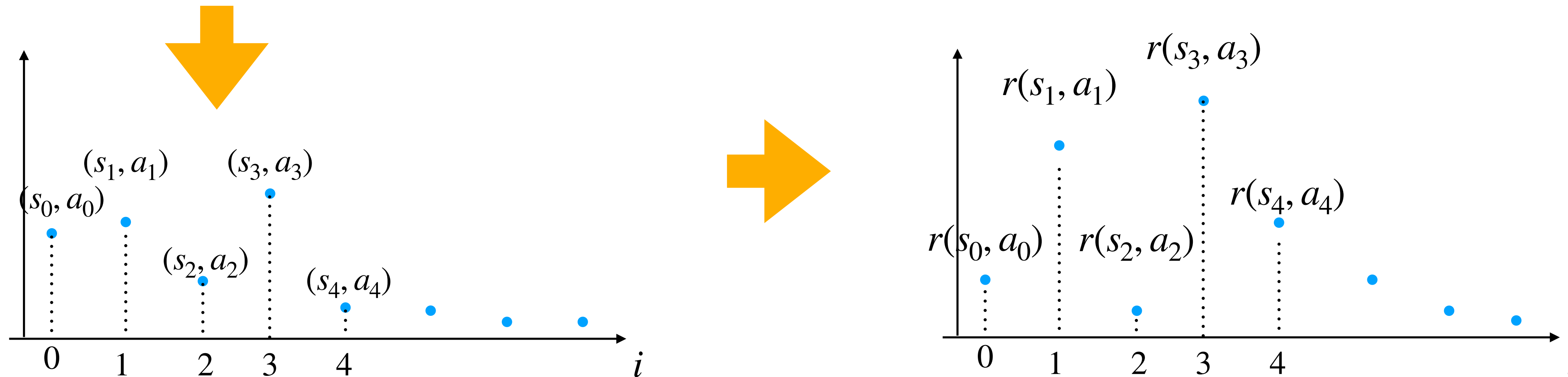
Mathematics > Optimization and Control

[Submitted on 21 May 2024 (v1), last revised 20 Feb 2026 (this version, v4)]

PhiBE: A PDE-based Bellman Equation for Continuous Time Policy Evaluation

Yuhua Zhu

RL in the digital world (a.k.a. Markov Decision Process)



$$V^*(s) = \max_{a(s)} \mathbb{E} \left[\sum_{i=0}^{\infty} \gamma^i r(s_i, a_i) \mid s_0 = s \right]$$

discounted factor $\gamma \in (0, 1)$

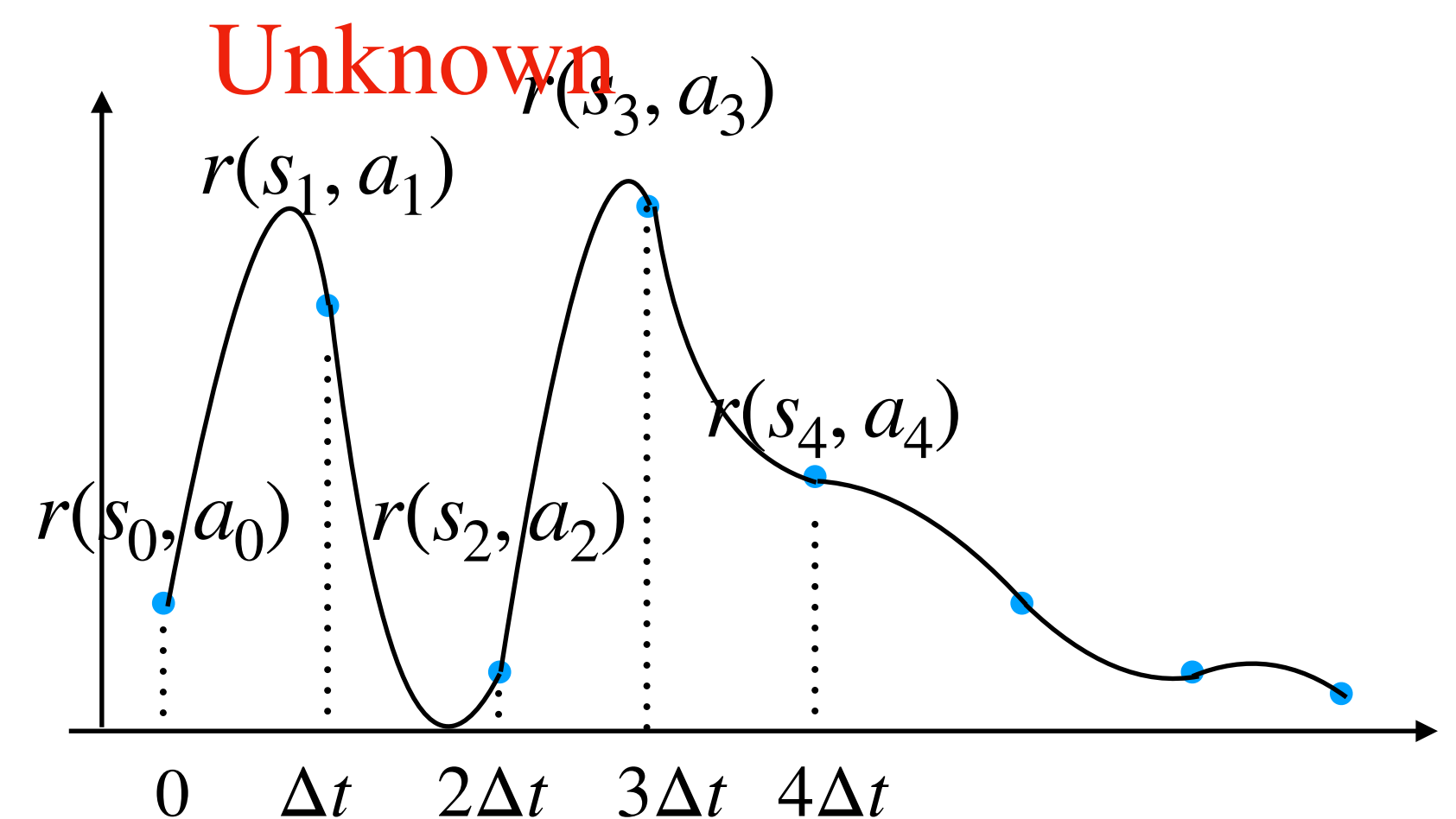
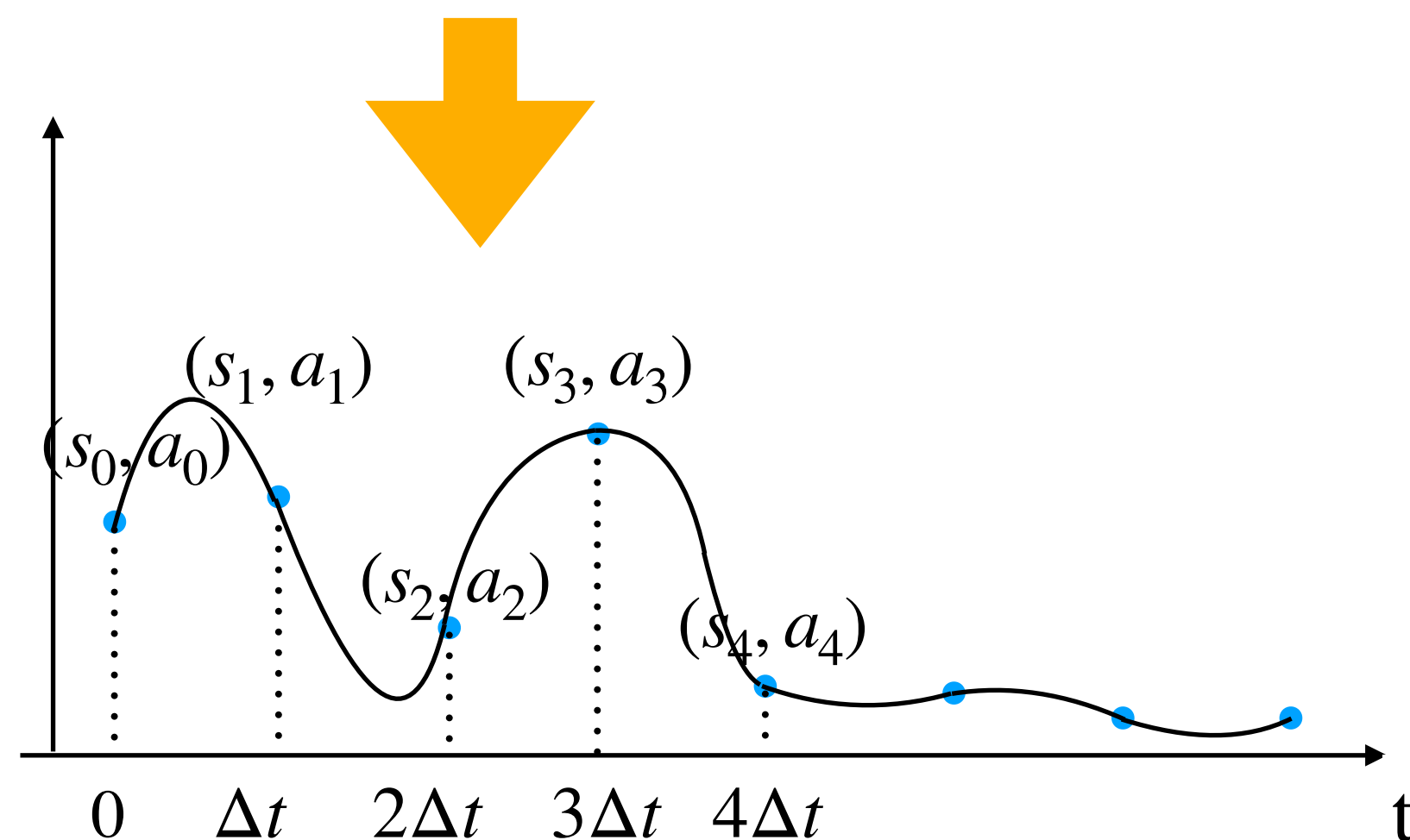
More difficult if $\gamma \rightarrow 1$

RL in the physical world



$$s_0, a_0 \longrightarrow s_1, a_1 \longrightarrow \dots$$

The underlying dynamics: $ds_t = b(s_t, a_t)dt + \sigma(s_t, a_t)dB_t$



Goal:

$$\max_{a_t = \pi(s_t)}$$

$$V^*(s) = \mathbb{E}_{a \sim \pi} \left[\int_0^\infty e^{-\beta t} r(s_t, a_t) dt \mid s_0 = s \right]$$

discounted constant $\beta > 0$

More difficult if $\beta \rightarrow 0$

RL in the digital world

RL in the physical world

Given discrete-time data $\{s_i, a_i, r_i, s'_i\}_{i=1}^n$

Goal:

$$\max_{a_i=\pi(s_i)} \mathbb{E} \left[\sum_{i=0}^{\infty} \gamma^i r(s_{i\Delta t}, a_{i\Delta t}) \mid s_0 = s \right] \longrightarrow \max_{a_t=\pi(s_t)} \mathbb{E} \left[\int_0^{\infty} e^{-\beta t} r(s_t, a_t) dt \mid s_0 = s \right]$$

The goal is more challenging.



Transition dynamics $s_{(i+1)\Delta t} \sim \rho(s' \mid s_{i\Delta t}, a_{i\Delta t})$

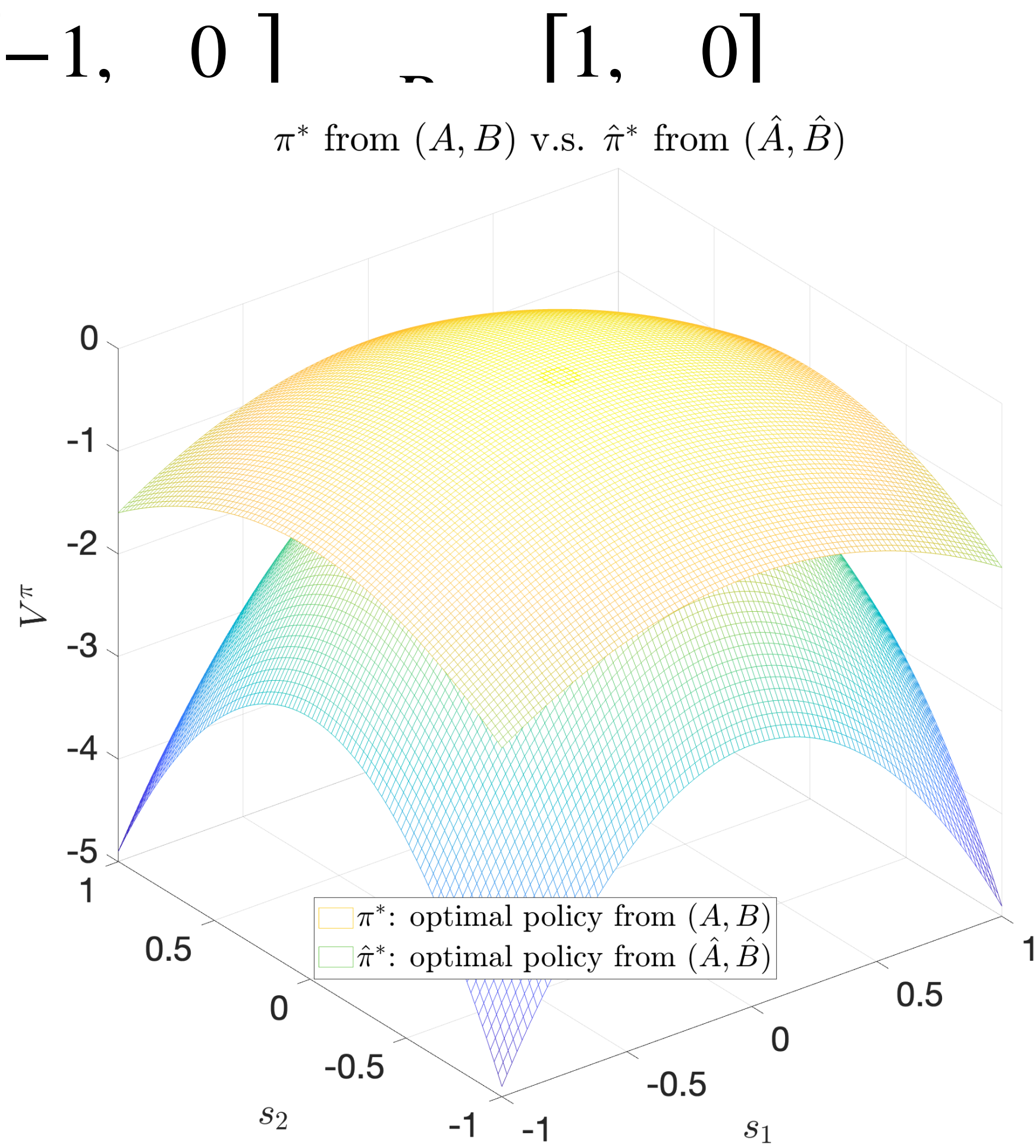
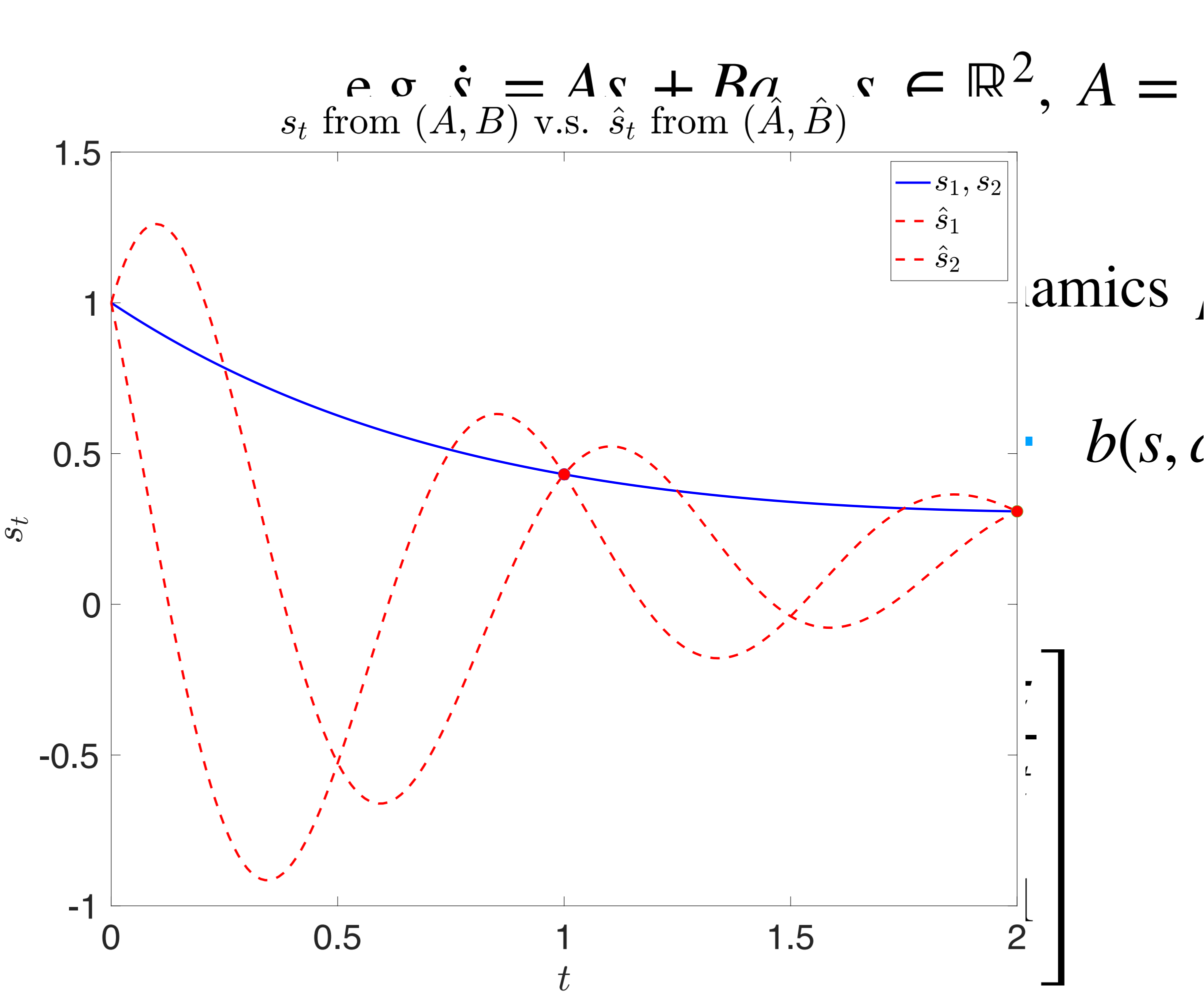
Underlying
dynamics

$$\forall \rho(s' \mid s, a), \text{ usually nonsmooth} \longrightarrow ds_t = b(s_t, a_t) dt + \sigma(s_t, a_t) dB_t$$

The transition dynamics are more smooth.

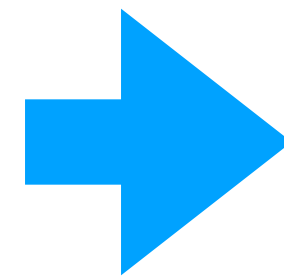


Unidentifiable



RL Perspective: Why bother studying continuous-time RL?

Discrete-time data



why not treat it as a discrete-time RL?

Cons



Pros

- Model-free
 - One can skip the identification of the dynamics and directly compute $V^*(s)$
- Many well-developed RL algorithms

How RL Approximates Optimal Control Problem

$$V^*(s) = \max_{a_t = \pi(s_t)} \mathbb{E} \left[\int_0^\infty e^{-\beta t} r(s_t, a_t) dt \mid s_0 = s \right]$$

$$\tilde{V}^*(s) = \max_{\pi(s)} \mathbb{E} \left[\sum_{i=0}^{\infty} (e^{-\beta \Delta t})^i r(s_{i\Delta t}, a_{i\Delta t}) \Delta t \mid s_0 = s \right]$$

(Discretized) Optimal-Bellman Equation (BE):

$$\tilde{V}^*(s) = \max_a \left\{ r_{\Delta t}(s, a) \Delta t + \gamma \mathbb{E} [\tilde{V}^*(s_{\Delta t}) \mid s_0 = s, a_0 = a] \right\}$$

Numerical integral

Bellman Optimality Principle

MDP / RL
approximation

Optimal-BE

$\tilde{V}_{\Delta t}^*(s)$ satisfies the discretized **optimal Bellman Equation** (optimal-BE):

$$\tilde{V}_{\Delta t}^*(s) = \max_a \left\{ \tilde{r}_{\Delta t}(s, a) + \gamma \mathbb{E} \left[\tilde{V}_{\Delta t}^*(s_{t+\Delta t}) \mid s_t = s, a_\tau = a, \tau \in [t, t + \Delta t) \right] \right\}$$

$\gamma = e^{-\beta \Delta t}$, $\tilde{r}_{\Delta t}(s, a) = r(s, a) \Delta t$.
 $s_{t+\Delta t} \sim \rho_{b, \sigma}(s', \Delta t \mid s_t, a_t)$

Model-free algorithms Q-learning, TD, PPO, AC are all derived from the **optimal-BE**

Two important features about optimal-BE

- The dynamics information is hidden in the expectation
 - ▶ The BE is the same for different dynamics $(b(s, a), \sigma(s, a))$
- Only depends on the current state and next state
 - ▶ Easy to plug in data directly

Is it the optimal tool for continuous-time RL?

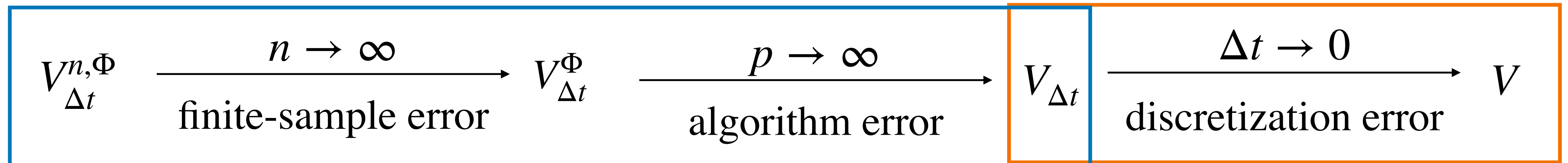
Challenge I: Error bound

[Puterman-2014]

[Asadi-Misra-Littman-2018]

$$\|V_{\Delta t}^*\| \lesssim (1 - \gamma)^{-1}, \quad \|\nabla V_{\Delta t}^*\| \lesssim (1 - \gamma)^{-1} \sim O(\Delta t^{-1})$$

Adapt to the continuous-time setting $\gamma = e^{-\beta\Delta t}$



?

Policy Evaluation [Lazaric - Ghavamzadeh - Munos - 2012]

$$\text{Sample error: } \lesssim (1 - \gamma)^{-2} \sim \beta^{-2} \Delta t^{-2} \quad \text{Model error } \lesssim (1 - \gamma^2)^{-1/2} \min_{\theta} \|\Phi^{\top} \theta - V_{\Delta t}\| \sim (\beta \Delta t)^{-1/2}$$

Challenge II: Numerics

$$\max_{\pi(s)} \quad \mathbb{E} \left[\int_0^\infty e^{-\beta t} (s_t^\top Q s_t + a_t^\top R a_t) \middle| s_0 = s \right]$$

$$\text{s.t.} \quad ds_t = (\cancel{As_t} + \cancel{Ba_t}) dt + \cancel{\sigma} dB_t \quad s, a \in \mathbb{R}^d, A, B, Q, R \in \mathbb{R}^{d \times d}$$

$\beta \rightarrow 0$: more difficult

Problem setting

- We do not know the underlying dynamics is linear
- We only have access to the discrete-time dynamics:

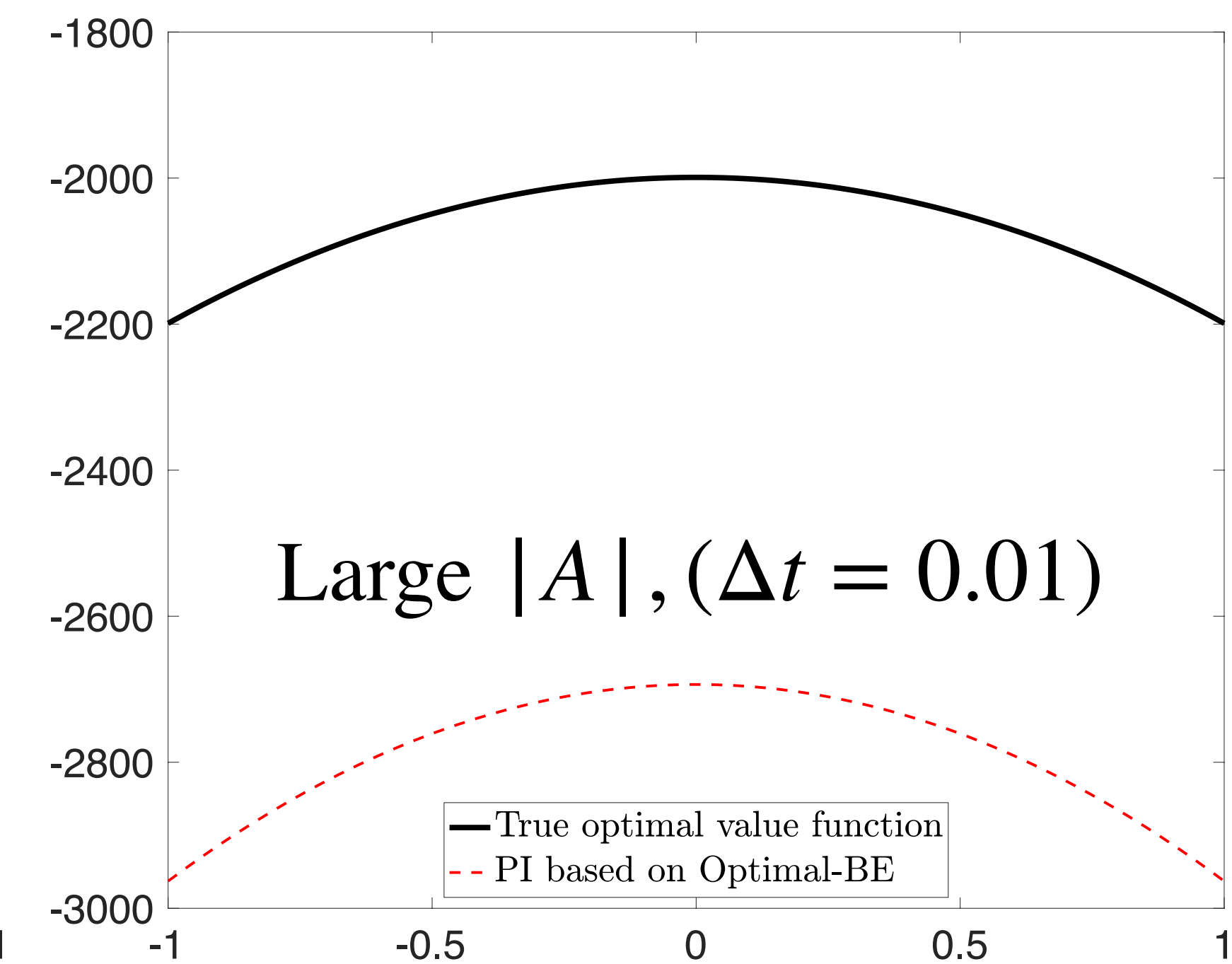
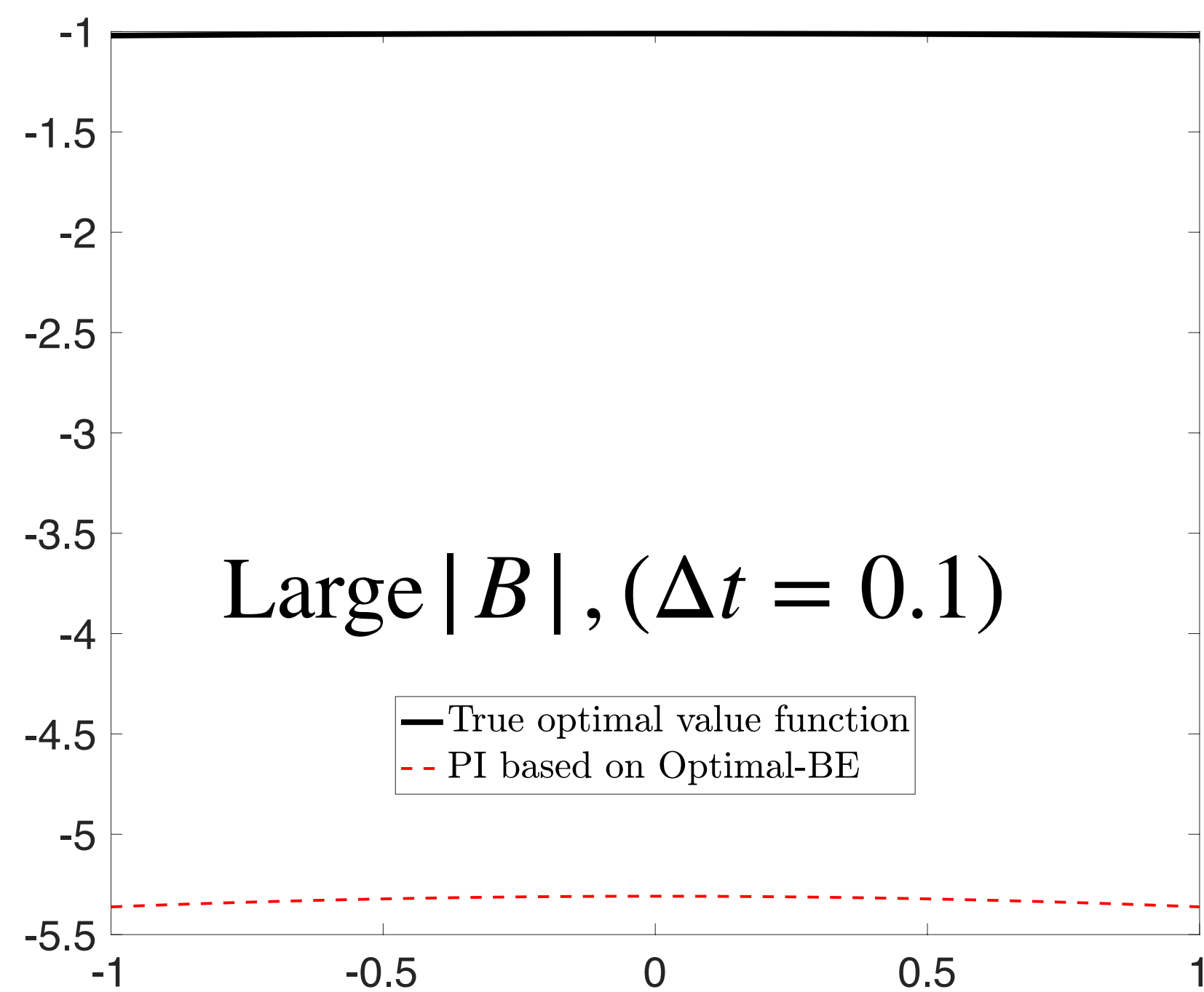
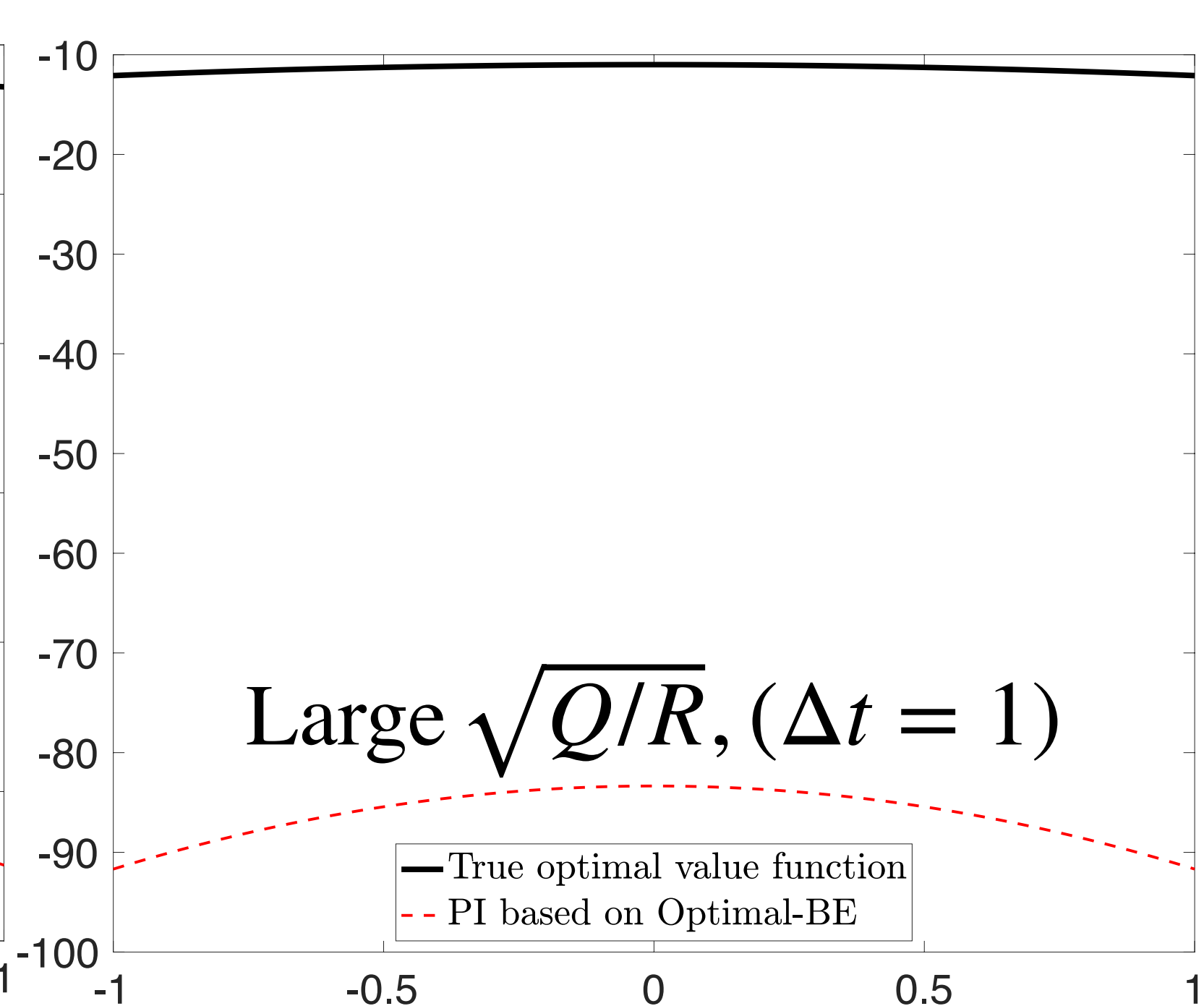
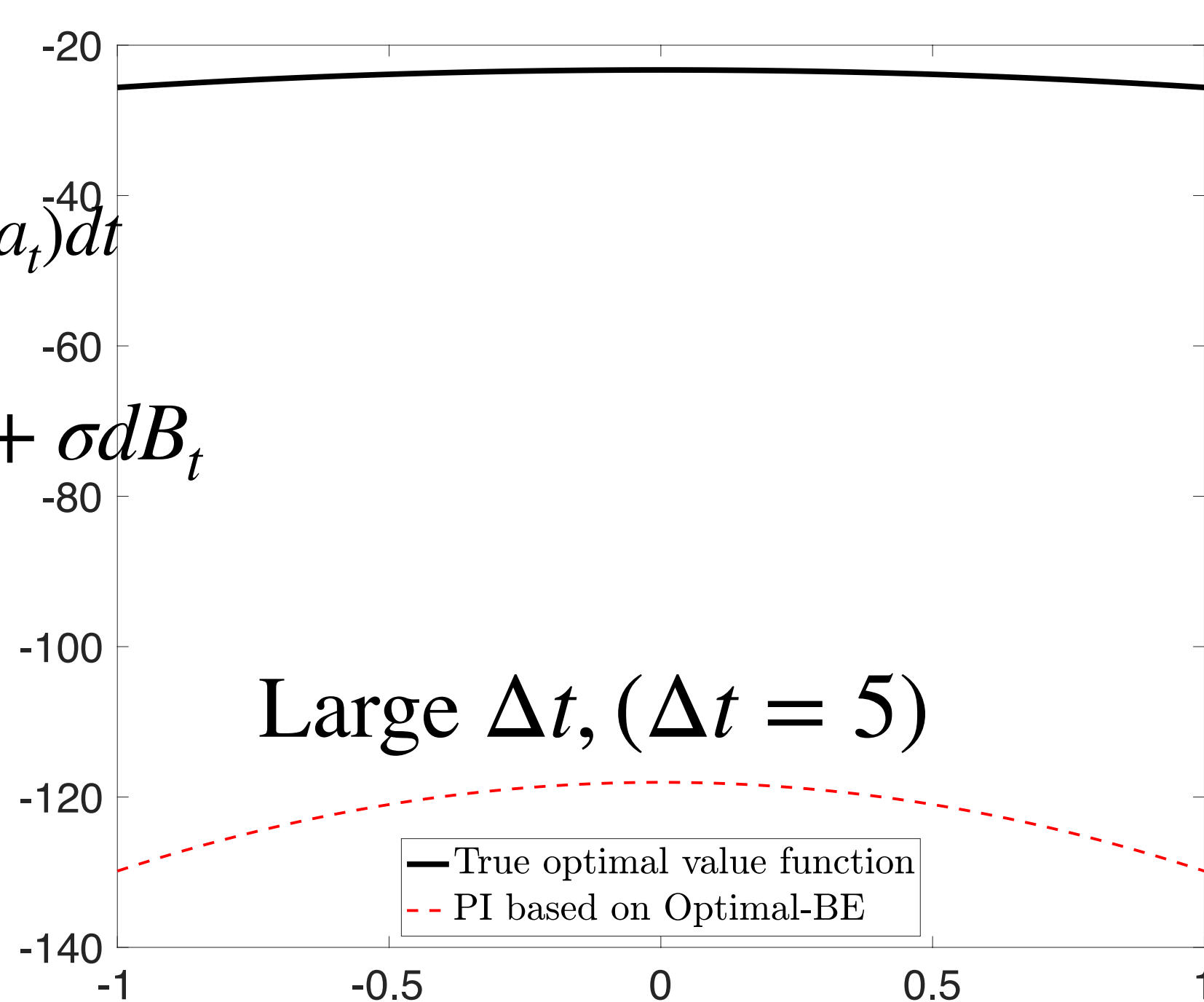
$$(s_i, a_i, r_i, s'_i)_{i=0}^I$$

$$\begin{aligned} & \max_{\pi(s)} \int_0^\infty e^{-\beta t} (s_t^\top Q s_t + a_t^\top R a_t) dt \\ \text{s.t. } & ds_t = (A s_t + B a_t) dt + \sigma dB_t \end{aligned}$$

$$\beta \in (0, 0.1)$$

$$A = 100, B = 1, \sigma = 1$$

$$Q = -10, R = -0.1$$



10^3 data points

LSTD
+
Policy Iteration

Why RL can be improved?

$$\tilde{V}^*(s) = \max_a \left\{ r_{\Delta t}(s, a) \Delta t + \gamma \mathbb{E} \left[\tilde{V}^*(s_{\Delta t}) \mid s_0 = s, a_0 = a \right] \right\}$$

The optimal-BE only uses $p_{\Delta t}(s' \mid s, a)$ $\longleftrightarrow$ Gap $ds_t = b(s_t, a) dt + \sigma(s_t, a) dB_t$

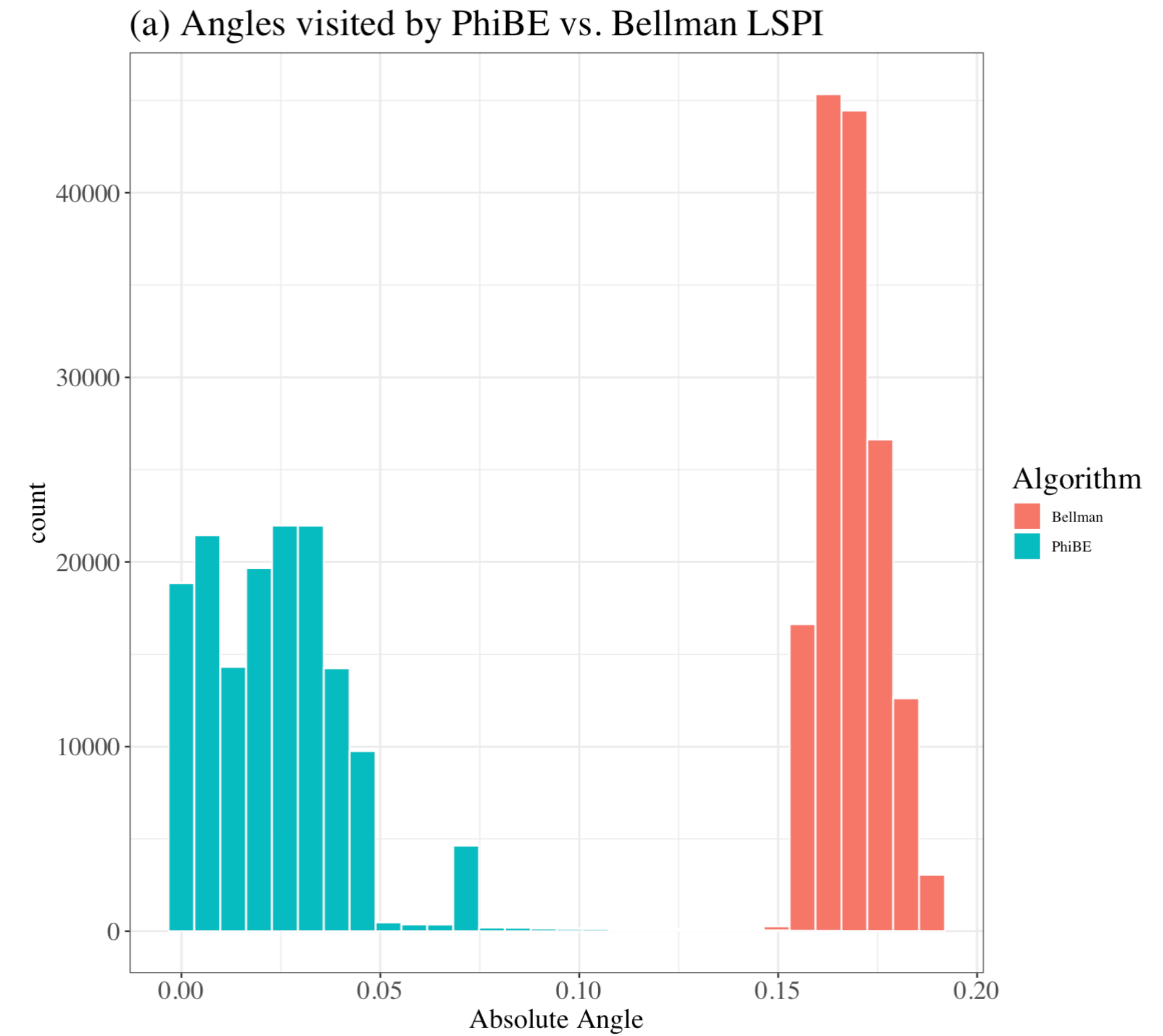
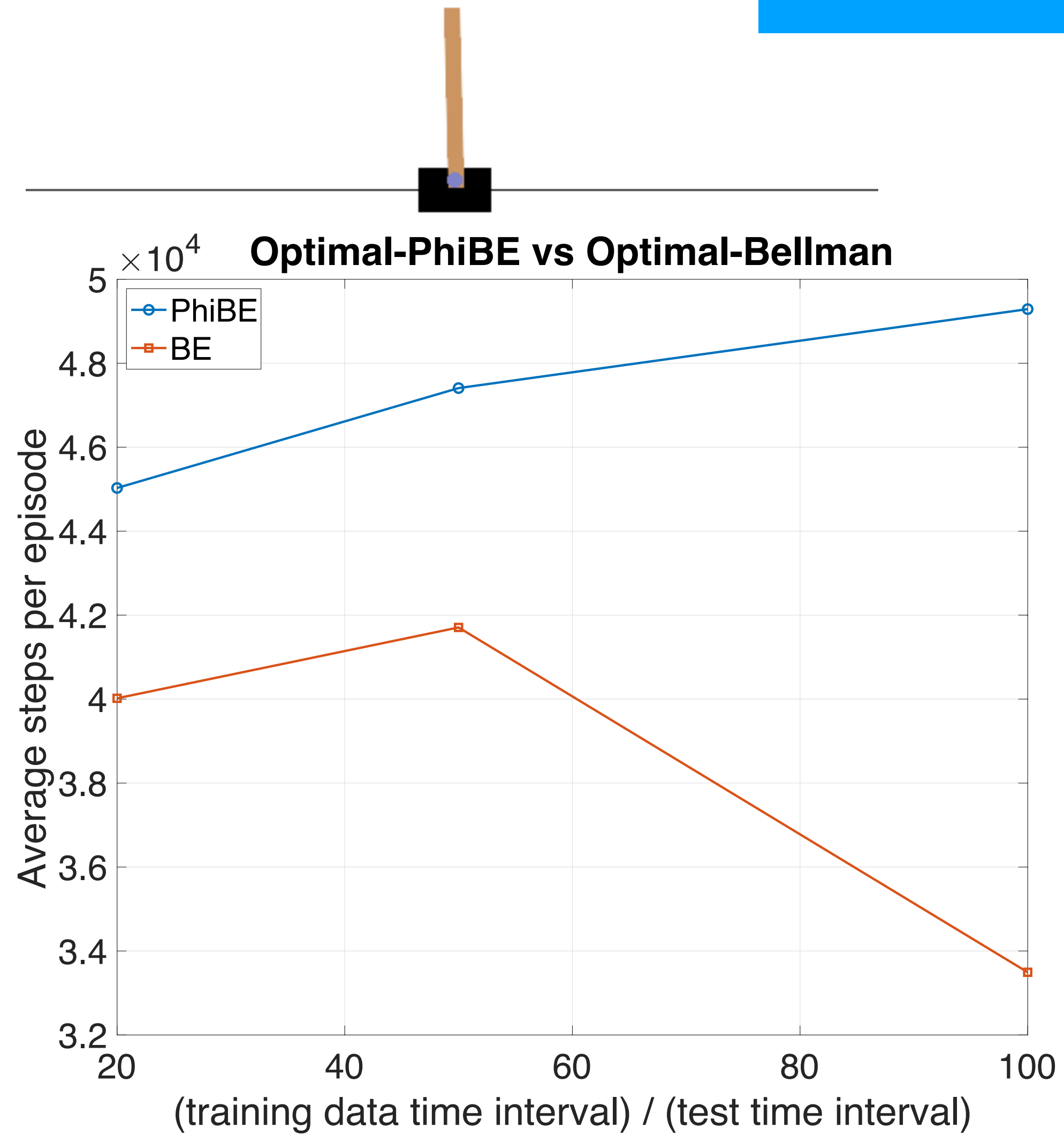
It can be derived from any continuous-time dynamics, or even discrete-time dynamics

The smooth information in continuous-time is not used in optimal-BE/MDP framework.

Optimal-BE is not ideal for CTRL

- MDP is designed for discrete-time decision-making process
- The current theorem is not considering this structure
- There is nowhere in the MDP framework that one can embed this continuous-time information.

A benchmark RL problem





RL framework



RL in the physical world

Challenge I: Error bound



RL framework



Apply RL framework
in the physical world

policy evaluation

Classical RL analysis:
The smoothness
structure is not utilized

[Puterman-2014], [Asadi-Misra-Littman-2018]

$$\|V_{\Delta t}^*\|, \quad \|\nabla V_{\Delta t}^*\| \lesssim O(\Delta t^{-1})$$

[Pham-Zhang-Z.-2026] Regularized HJB

$$\|V_{\Delta t}^*\|, \|\nabla V_{\Delta t}^*\| \lesssim C \quad \text{Discretization error: } \|V^*(s) - V^{\tilde{\pi}^*}(s)\|_{L^\infty} \lesssim O(\Delta t^{1/(N+1)} \log(\Delta t))$$

Using the ellipticity
of the infinitesimal
generator

Challenge I: Error bound



RL framework



Apply RL framework
in the physical world

policy evaluation

Classical RL analysis:
The smoothness
structure is not utilized

e.g. [Lazaric - Ghavamzadeh - Munos - 2012]

$$\text{Sample error: } \lesssim (\beta \Delta t)^{-2} \quad \text{Model error } \lesssim (\beta \Delta t)^{-1/2} \min_{\theta} \|\Phi^{\top} \theta - V_{\Delta t}\|$$

[Mou-Z.-2024]

$$\|V_{\Delta t}^{n,\Phi} - V\| \lesssim \beta^{-2} \Delta t^{-1/2} + \beta^{-1} \min_{\theta} \|\Phi^{\top} \theta - V_{\Delta t}\|_{H^1} + \beta^{-1} \Delta t$$

Using the ellipticity
of the infinitesimal
generator

Discretization error

Challenge II: Numerics



RL framework

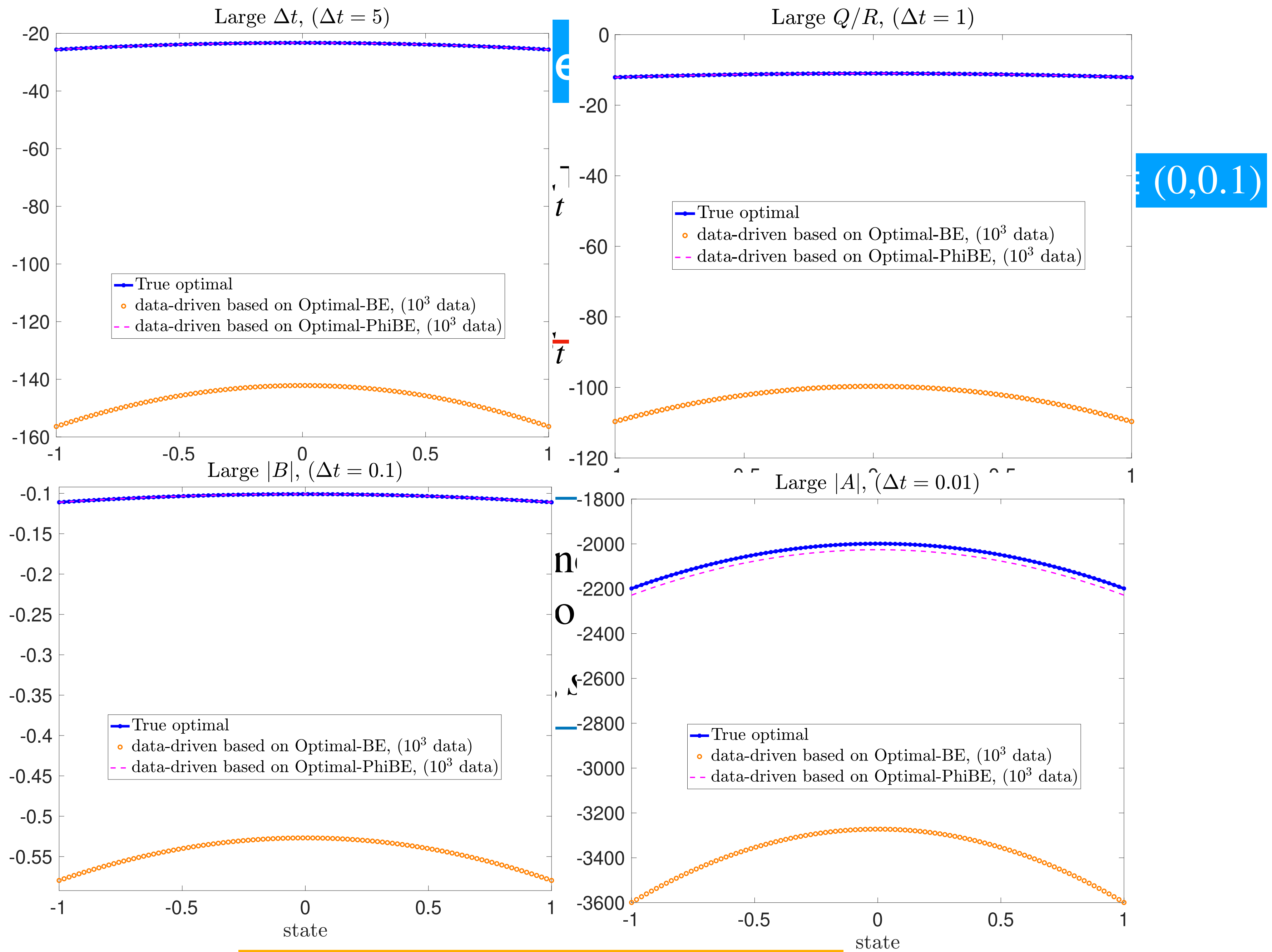
Go beyond the
MDP framework



**Find the optimal tool for
RL in the physical world**

Wishlist

- ◎ Utilize the underlying SDE structure
- ◎ Only based on the discrete-time transition dynamics $\rho_{\Delta_t}(s' | s, a)$, cannot based on the continuous-time dynamics $b(s, a), \sigma(s, a)$
- ◎ Can be applied to one-step transition data $\{s_i, a_i, r_i, s'_i\}_{i=1}^n$
- ◎ Model-free algorithm
- ◎ With minimum modification to the existing RL algorithms



With the same computational cost!

Optimal-PhiBE

Taking second-order Taylor expansion of Optimal-BE

$$\text{Optimal-BE } \tilde{V}^*(s) = \max_a \left\{ r_{\Delta t}(s, a) \Delta t + \gamma \mathbb{E} \left[\tilde{V}^*(s_{t+\Delta t}) \mid s_t = s, a_t = a \right] \right\}$$
$$\mathbb{E} \left[\tilde{V}^*(s_t) + (s_{t+\Delta t} - s_t) \cdot \nabla \tilde{V}^*(s_t) + \frac{1}{2} (s_{t+\Delta t} - s_t)^\top \nabla^2 \tilde{V}^*(s_t) (s_{t+\Delta t} - s_t) \mid s_t = s, a_t = a \right]$$

[Z.-2024], [Z.-Zhang-Zhang-2025]

$$\text{Optimal-PhiBE: } \beta \hat{V}^*(s) = \max_a \left\{ r(s, a) + \mathbb{E} \left[\frac{s_{t+\Delta t} - s_t}{\Delta t} \mid s, a \right] \cdot \nabla_s \hat{V}^*(s) + \mathbb{E} \left[\frac{(s_{t+\Delta t} - s_t)(s_{t+\Delta t} - s_t)^\top}{\Delta t} \mid s, a \right] : \nabla^2 \hat{V}^*(s) \right\}$$

Wishlist

✓ Model-free algorithm

- ▶ Same formulation for different dynamics
- ▶ Only depends on the current state and next state —> Can be applied to one-step transition data $\{s_i, a_i, r_i, s'_i\}_{i=1}^n$

Optimal-PhiBE

[Z.-2024], [Z.-Zhang-Zhang-2025]

$$\text{optimal-PhiBE: } \beta \hat{V}^*(s) = \max_a \left\{ r(s, a) + \mathbb{E} \left[\frac{s_{t+\Delta t} - s_t}{\Delta t} \mid s, a \right] \cdot \nabla_s \hat{V}^*(s) + \frac{1}{2} \mathbb{E} \left[\frac{(s_{t+\Delta t} - s_t)(s_{t+\Delta t} - s_t)^\top}{\Delta t} \mid s, a \right] : \nabla^2 \hat{V}^*(s) \right\}$$

$$\beta \hat{V}^*(s) = \max_a \left\{ r(s, a) + \hat{b}(s, a) \cdot \nabla_s \hat{V}^*(s) + \frac{1}{2} \hat{\Sigma}(s, a) : \nabla^2 \hat{V}^*(s) \right\}$$

Hamilton-Jacobi-Bellman Equation

$$\beta V^*(s) = \max_a \left\{ r(s, a) + b(s, a) \cdot \nabla V^*(s) + \Sigma(s, a) : \nabla^2 V^*(s) \right\}$$

- Similarity: PDE, containing continuous-time information
- Difference: Only continuous information, has to estimate the dynamics first

Wishlist

- ✓ Utilize the underlying SDE structure
- ✓ Only based on the discrete-time transition dynamics $\rho_{\Delta t}(s' \mid s, a)$, cannot based on the continuous-time dynamics $b(s, a), \sigma(s, a)$

Optimal-PhiBE Is Fundamentally Different From Numerical Scheme for HJB

Numerical Scheme

- Goal: Solve $\mathcal{L}_{b,\Sigma} V^* = 0$ with known b, Σ
 - Discretize $\mathcal{L}_{b,\Sigma} V^* = 0$ uniformly in space
 - $V^*(s_0), V^*(s_0 + h), V^*(s_0 + 2h)$

Optimal-PhiBE

- Goal: Solve $\mathcal{L}_{b,\Sigma} V^* = 0$ with unknown b, Σ

Step 1: approximate the unknown operator $\mathcal{L}_{b,\sigma}$ using discrete-time information

$$\mathcal{L}_{b,\sigma} \xrightarrow{\text{Optimal-PhiBE}} \mathcal{L}_{\hat{b},\hat{\Sigma}}$$

Step 2: Solve $\mathcal{L}_{\hat{b},\hat{\Sigma}} \hat{V}^* = 0$ using $s_0, s_{\Delta t}, s_{2\Delta t}, \dots$

$$s_0, s_{\Delta t}, s_{2\Delta t}, \dots \xrightarrow{\text{Model-free}} \hat{V}^*(s), \hat{\pi}^*(s)$$

Do not explicitly estimate $\hat{b}, \hat{\Sigma}$

What are the theoretical guarantees for PhiBE?

Policy evaluation

Bellman equation

[Mou-Z.-2025]

$$\|V_{\Delta t}^{n,\Phi} - V\| \lesssim \|\mathcal{L}_{b,\Sigma} r\|_{\infty} (\beta^{-2} \Delta t^{-1/2} + \beta^{-1} \min_{\theta} \|\Phi^{\top} \theta - V_{\Delta t}\|_{H^1} + \beta^{-1} \Delta t)$$

$$\mathcal{L}_{b,\Sigma} = b(s) \cdot \nabla + \frac{1}{2} \Sigma(s) : \nabla^2$$

[Z.-2024]

PhiBE

$$\|V_{\Delta t}^{n,\Phi} - V\| \lesssim \|\mathcal{L}_{b,\Sigma} b\|_{\infty} (\beta^{-2} \Delta t^{-1/2} + \beta^{-1} \min_{\theta} \|\Phi^{\top} \theta - V_{\Delta t}\|_{H^1} + \beta^{-2} \Delta t)$$

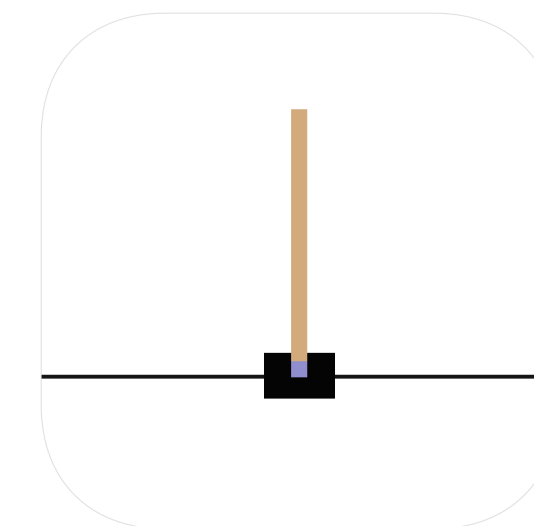
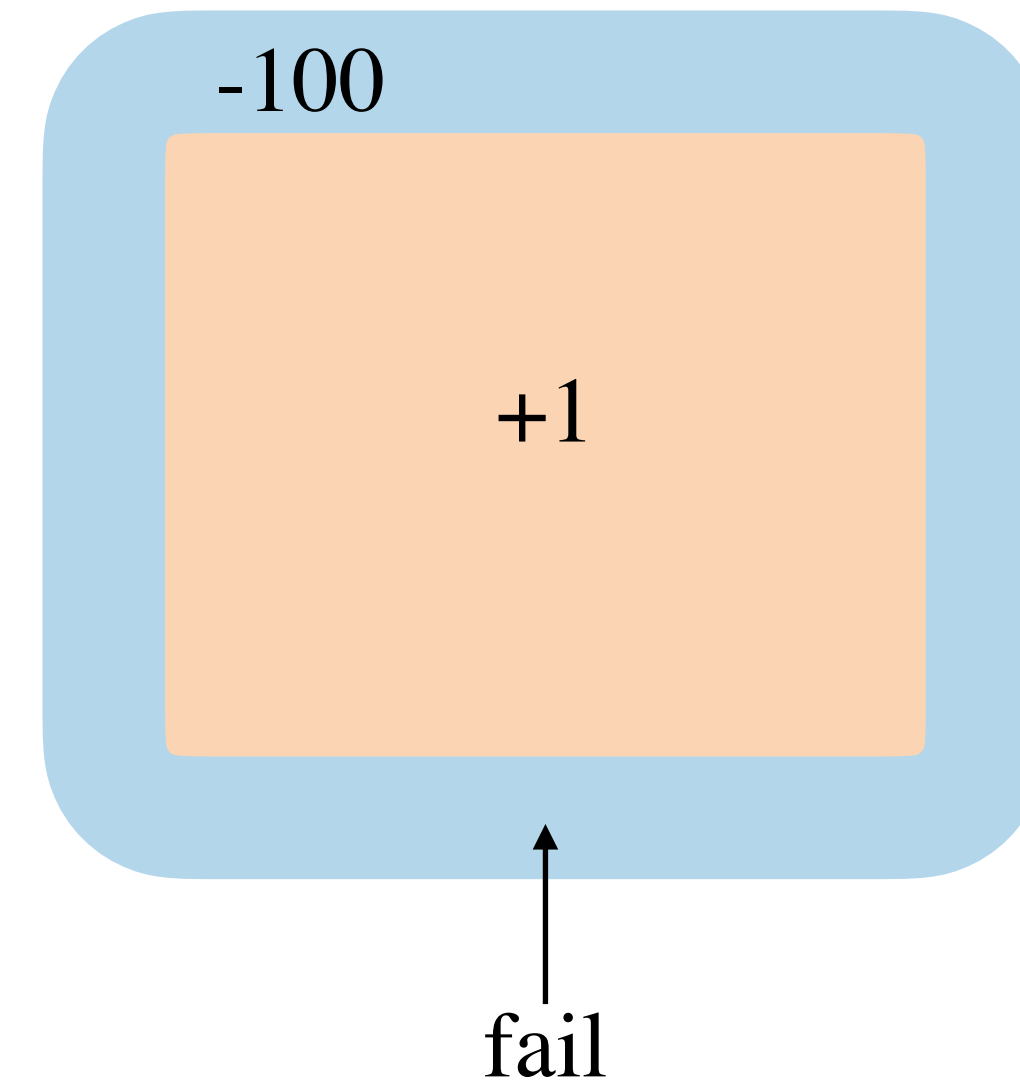
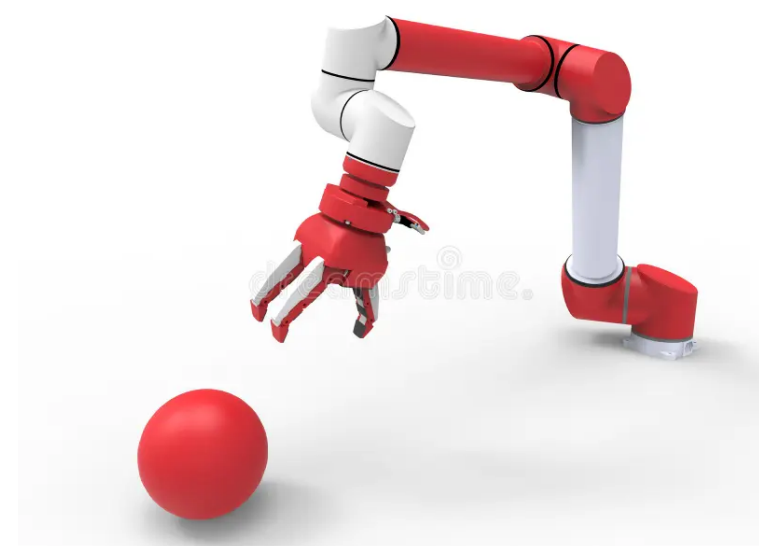
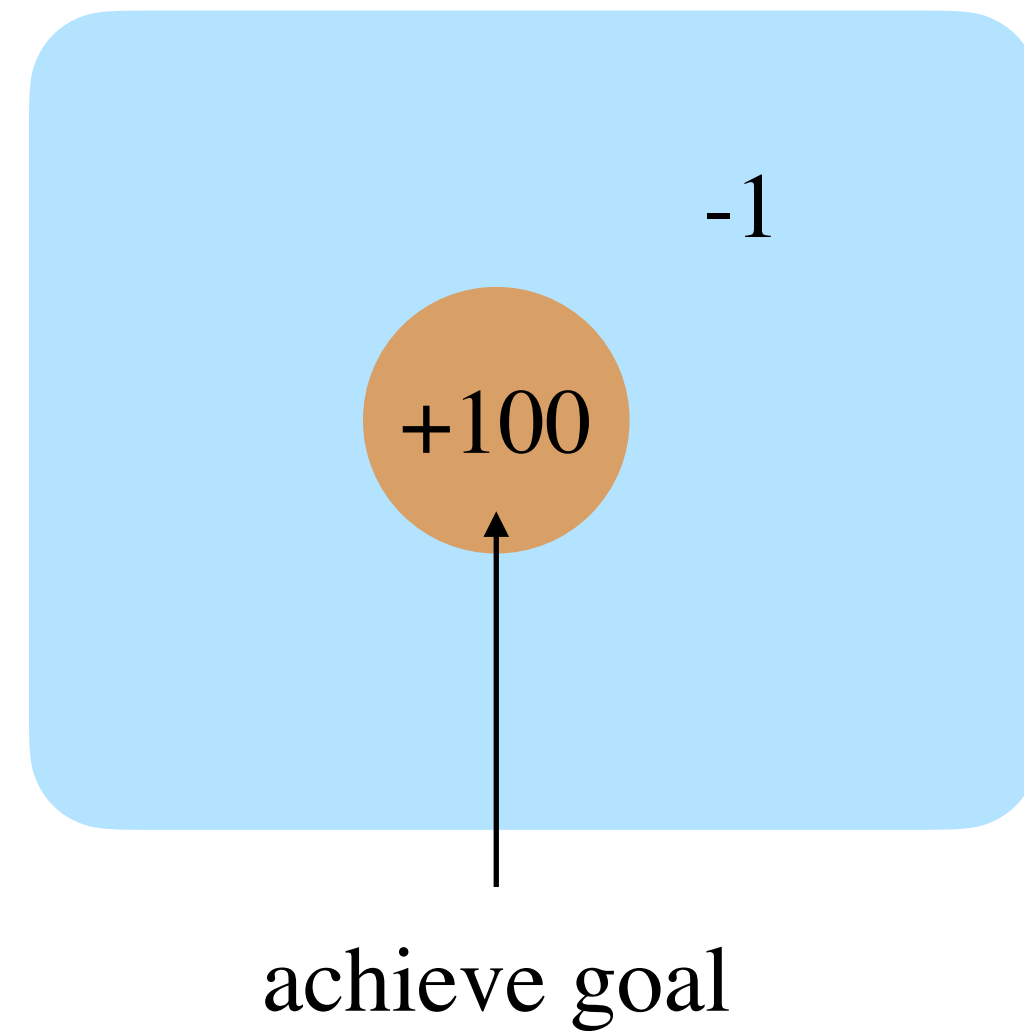
Trade-off between discretization error and sample error

More frequent data observation is not necessarily beneficial.

- Less sensitive to reward oscillation
- The natural dynamics evolves slowly

Common in RL applications

How does reward look like in RL applications?



The reward often varies a lot in the state space, but not necessarily in the action space.

The natural dynamics usually evolves slowly.

Theoretical Guarantees in general deterministic dynamics

Stochastic Optimal-PhiBE [Pham-Zhang-Z.-26]

regularized Optimal BE $\|V^*(s) - V^{\tilde{\pi}^*}(s)\|_{L^\infty} \lesssim O(\Delta t^{1/(N+1)} \log(\Delta t))$

Stochastic Optimal-PhiBE [Z.-Zhang-Zhang-25]

Optimal-PhiBE: $\|V^*(s) - V^{\hat{\pi}^*}(s)\|_\infty \lesssim O(\Delta t)$

Stochastic LQR

[Z.-Zhang-Zhang-25]

Let $\tilde{\pi}^*(s) = \tilde{K}s$, $\hat{\pi}^*(s) = \hat{K}s$ be the optimal policy from Optimal-BE and Optimal-PhiBE, then

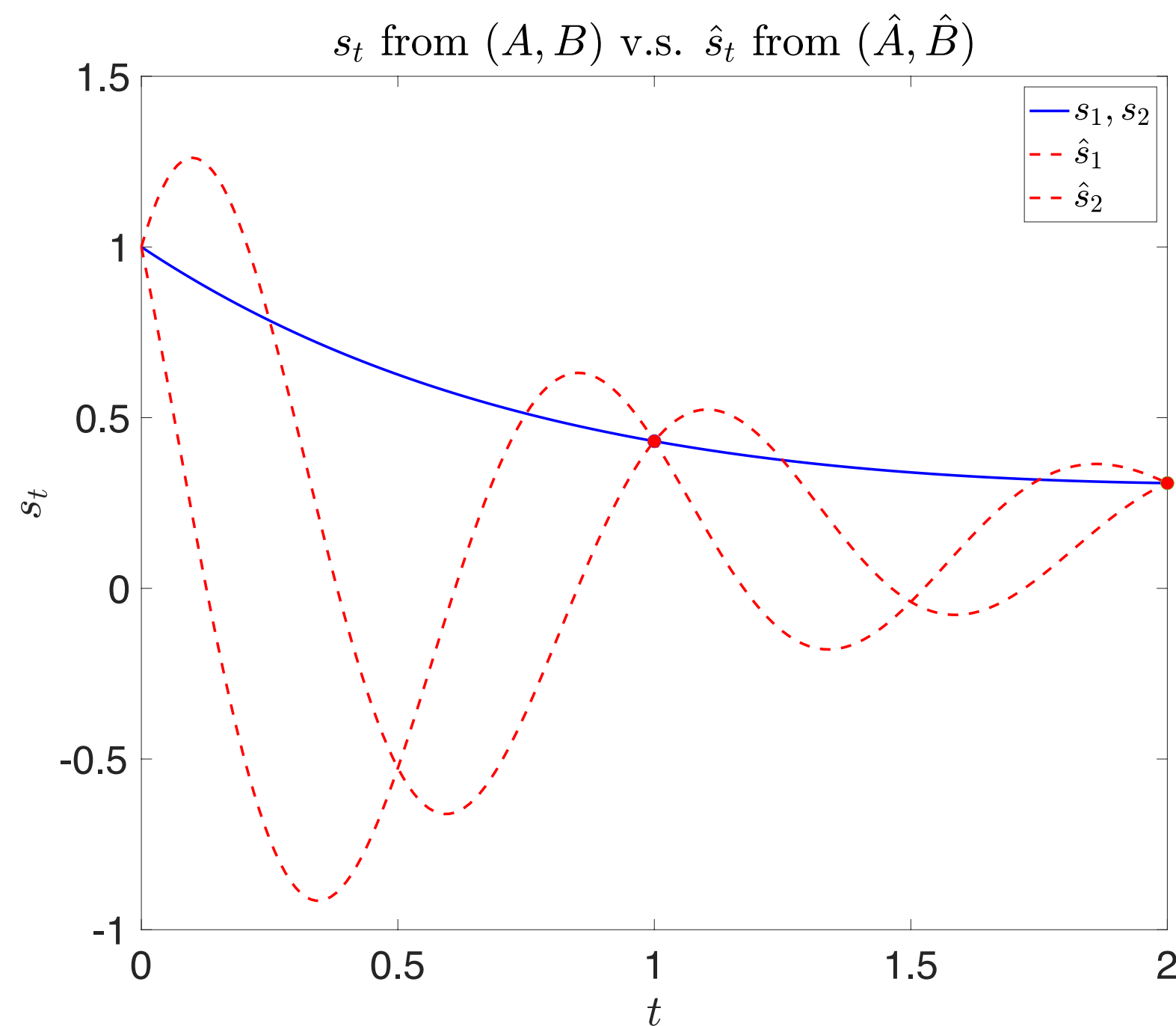
$$|\tilde{K} - K| \lesssim \left[|B| \left(\sqrt{\frac{Q}{R}} + \frac{|A|}{|B|} \right)^2 + |A - \beta/2| \frac{|A|}{|B|} \right] \Delta t$$
$$|\hat{K} - K| \begin{cases} \leq \frac{\beta |A|}{|B|} \Delta t, & \text{when } \beta > 0 \\ = 0, & \text{when } \beta = 0 \end{cases}$$

When $\beta = 0$, PhiBE will give the accurate optimal control $\pi^*(s)$ with only discrete-time information

Unidentifiability of the continuous-time dynamics from discrete-time data

$$\text{e.g. } \dot{s}_t = As_t + Ba_t, \quad s_t \in \mathbb{R}^2, \quad A = \begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix}, \quad B = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

Given discrete-time transition dynamics $p_{\Delta t}(s, a) = e^{-\Delta t}s - (e^{-\Delta t} - 1)a + b(s, a)$ is linear



The dynamics are unidentifiable,
BUT the LQR optimal policy is
identifiable with discrete-time
information when $\beta = 0$!

Summary

Optimal-BE: $\tilde{V}_{\Delta t}^*(s) = \max_a \left\{ \tilde{r}_{\Delta t}(s, a) + \gamma \mathbb{E} \left[\tilde{V}_{\Delta t}^*(s_{t+\Delta t}) \mid s_t = s, a_\tau = a, \tau \in [t, t + \Delta t) \right] \right\}$ ✗

Optimal-PhiBE: $\beta \hat{V}^*(s) = \max_a \left\{ r(s, a) + \hat{b}(s, a) \cdot \nabla_s \hat{V}(s) + \frac{1}{2} \hat{\Sigma}(s, a) : \nabla_s^2 V^*(s) \right\}$ ✓

$$\hat{b}(s, a) = \mathbb{E} \left[\frac{s_{\Delta t} - s_0}{\Delta t} \mid s_0 = s, a_\tau = a, \tau \in [0, \Delta t) \right], \quad \hat{\Sigma}(s, a) = \mathbb{E} \left[\frac{(s_{\Delta t} - s_0)(s_{\Delta t} - s_0)^\top}{\Delta t} \mid s_0 = s, a_\tau = a, \tau \in [0, \Delta t) \right]$$

BUT, how to solve it with discrete-time data?

Wishlist

✓ With minimum modification to the existing RL algorithms

Temporal Difference

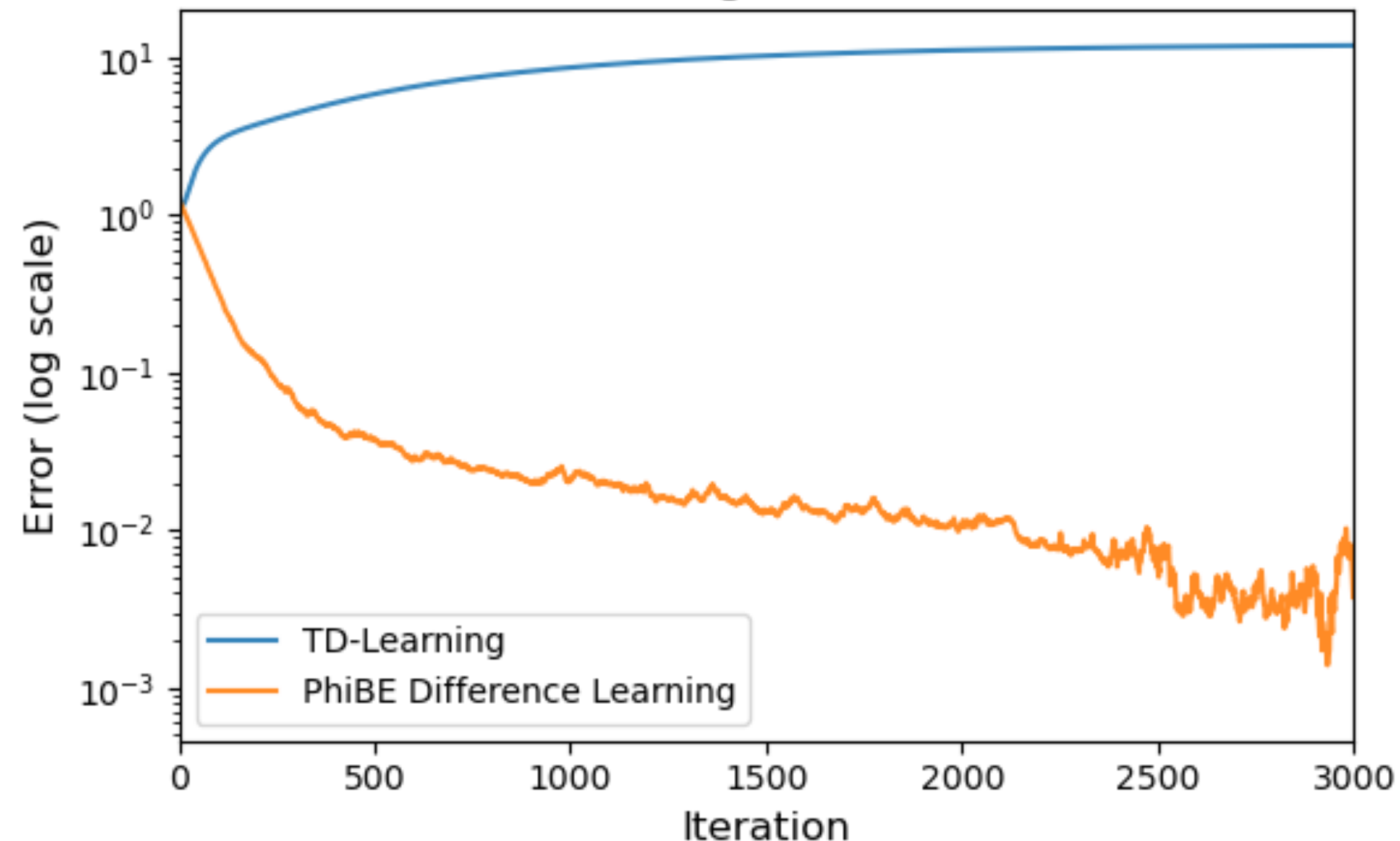
$$\delta_i = r_i \Delta t + \gamma V_{\theta_i}(s'_i) - V_{\theta_i}(s_i) \longrightarrow \delta_i = r_i + \frac{s'_i - s_i}{\Delta t} \cdot \nabla_s V_{\theta_i}(s_i) + \frac{(s'_i - s_i)(s'_i - s_i)^\top}{\Delta t} : \nabla_s^2 V_{\theta_i}(s_i) - \beta V_{\theta_i}(s_i)$$

(discrete-time) Bellman residual

PhiBE residual

$$\theta_{i+1} = \theta_i + \alpha_i \delta_i \nabla_{\theta} V(s_i)$$

Semilog Error Plot



Error analysis

$$\|V^\pi - \hat{V}_\theta\| \lesssim \Delta t + \min_{\theta} \|V^\pi - \theta^\top \Phi\| + \frac{1}{\sqrt{n\Delta t}}$$



Yutong Ren, Yuhua Zhu, PhiBE Difference for continuous-time policy evaluation. In preparation

Yutong Ren

Algorithm: Q-learning with **off-policy** data

Q-learning

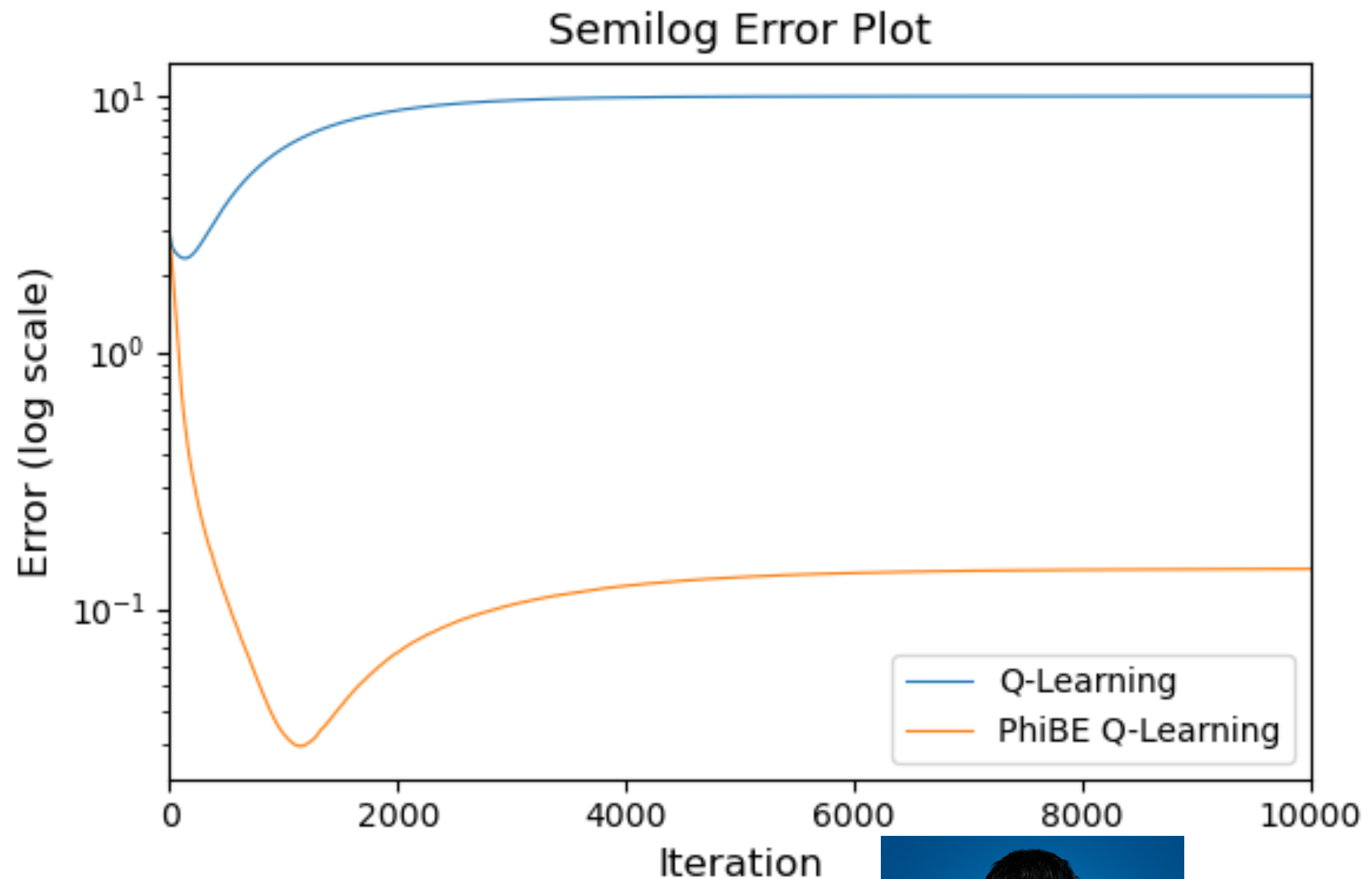
Given off policy data $\{s_i, a_i, r_i, s'_i\}$

$$\delta_i = r_i \Delta t + \gamma \max_{a'} Q_{\theta_k}(s'_i, a') - Q_{\theta_k}(s_i, a_i)$$



$$\delta_i = r_i + \frac{s'_i - s_i}{\Delta t} \cdot \nabla_s \max_{a'} Q_{\theta_k}(s'_i, a') + (1 - \beta) \max_{a'} Q_{\theta_k}(s_i, a_i) - Q(s_i, a_i)$$

$$\theta_{k+1} = \theta_k + \frac{\alpha_i}{n} \sum_i \delta_i \nabla_{\theta} Q(s_i, a_i)$$

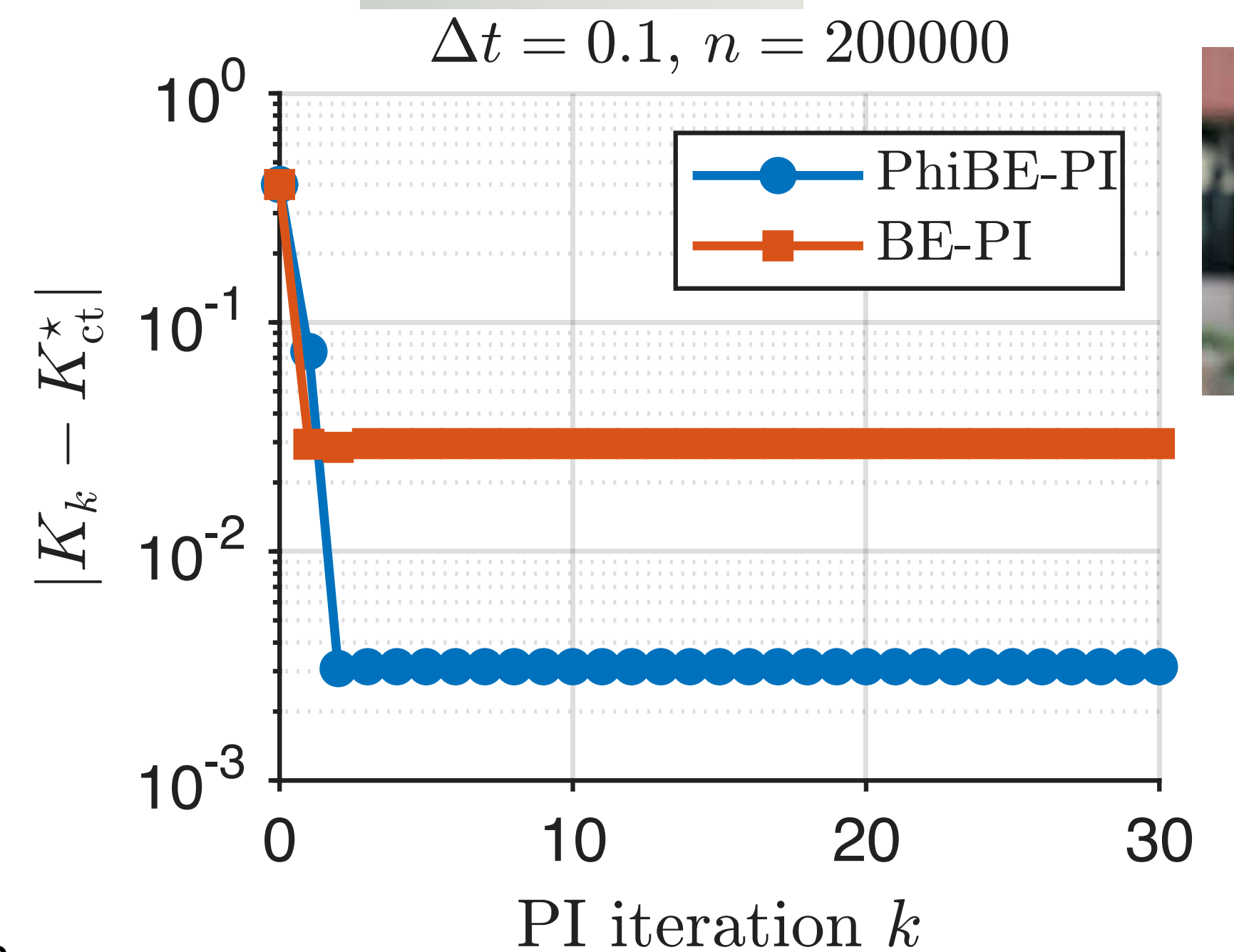
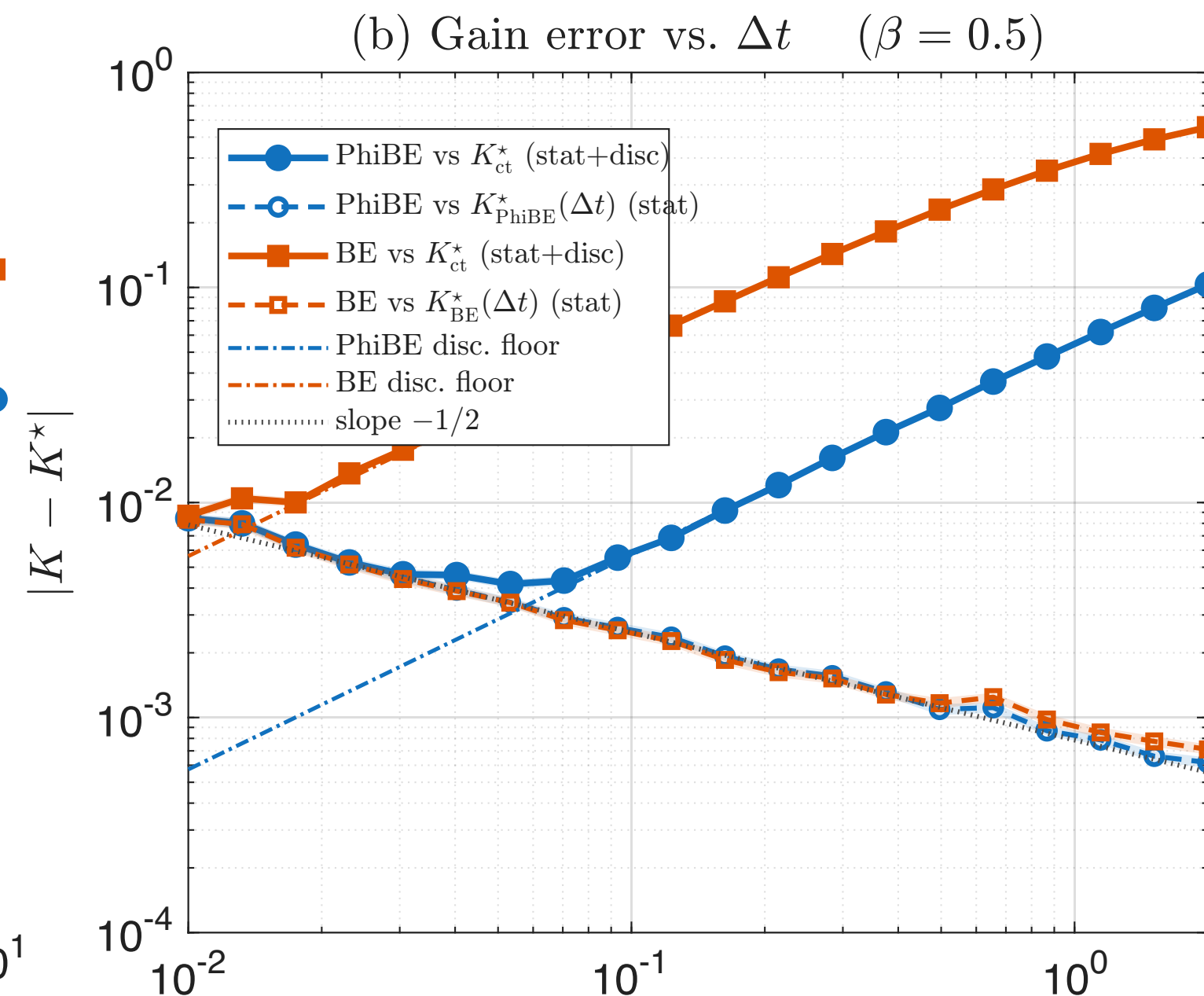
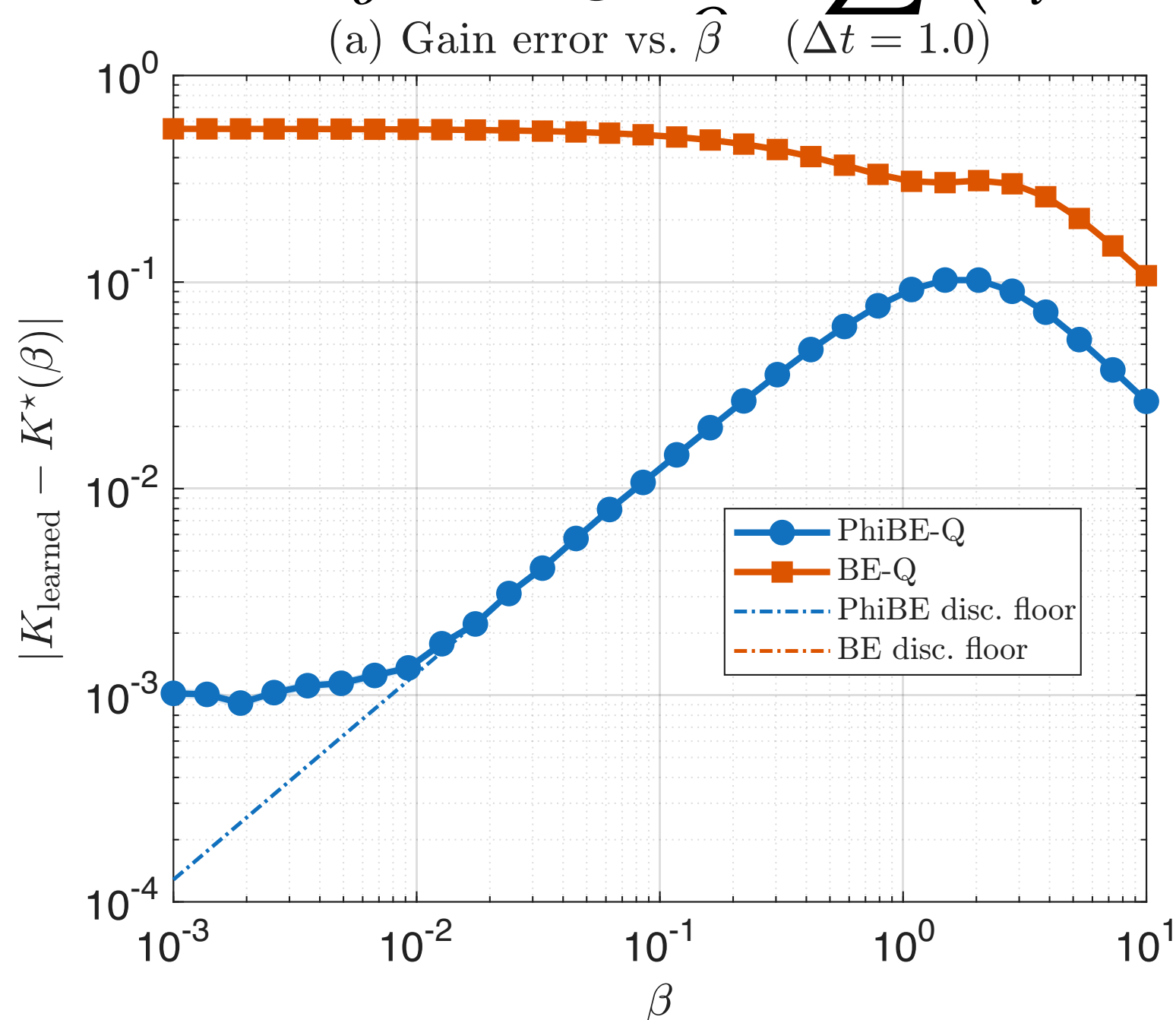


Yutong Ren

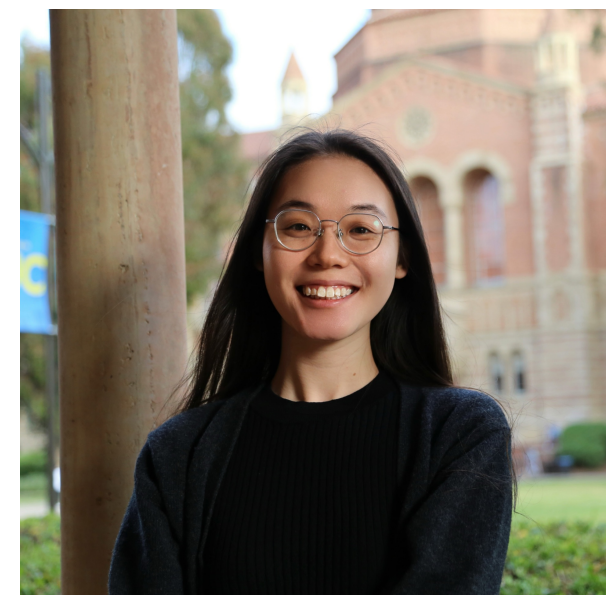
Algorithm: Policy Iteration with off-policy data

Given off policy data $\{s_i, a_i, r_i, s'_i\}_i$

$$\text{Solve } Q_\theta = \arg \min \sum (r_i \Delta t + \gamma E_{a \sim \pi^k} [Q_\theta(s'_i, a)] - Q_\theta(s'_i, a)) \nabla_\theta Q(s_i, a_i)$$



$$\|V^\star - V^{\pi_N}\| \lesssim O(\Delta t) + \frac{1}{\sqrt{\beta}} + \frac{\text{cov}}{\beta} \left[\epsilon_{model} + O \left(\alpha_\Phi(\beta)^{-2} \sqrt{\frac{1}{n\Delta t}} \right) \right].$$



Chaorui Yao



Mean-field Optimal Control



Erhan Bayraktar

2025/03



Martin Hernandez

2025/11



Qinxin Yan

2026/1

Problem Setting

Goal $\pi^*(\cdot | s, \mu) = \arg \max_{\pi} V^{\pi}$

$$V^{\pi}(\mu) = \mathbb{E} \left[\int_0^{\infty} e^{-\beta t} r_{\lambda}^{\pi}(\mu_t) dt \mid \mu_0 = \mu \right] \quad r_{\lambda}^{\pi}(\mu) = \mathbb{E}_{a \sim \pi(\cdot | s, \mu), s \sim \mu(\cdot)} [r(s, \mu, a) - \lambda \log \pi(a | s, \mu)]$$

$$ds_t = b^{\pi}(s_t, \mu_t) dt + \sigma^{\pi}(s_t, \mu_t) dB_t, \quad \mu_t = \text{Law}(s_t).$$

$$b^{\pi}(s, \mu) = \int_A \cancel{b(s, \mu, a)} \pi(da | s, \mu), \quad \Sigma^{\pi}(s, \mu) = \int_A \cancel{\sigma(x, \mu, a)} \pi(da | s, \mu).$$

$$\{s_i, \mu_i, a_i, r_i, s'_i\}_{i=1}^n$$

$$\bullet s'_i \sim \rho_{\Delta t}(s' | s_i, \mu_i, a_i), \quad \Delta t > 0$$

• One step transition data

• $\mu_i \rightarrow E_{s \sim \mu_i}[\Phi(s)]$ (with cylindrical basis functions)

RL approximation

$$\pi_{RL}^*(\mu) = \arg \max_{\pi} V_{RL}^{\pi}(\mu) = \arg \max_{\pi} \mathbb{E} \left[\sum_{i=0}^{\infty} \gamma^i r_{i\Delta t} dt \middle| \mu_0 = \mu \right], \quad \gamma = e^{-\beta\Delta t}$$

$$s'_{(i+1)\Delta t} \sim \rho_{\Delta t}(s' | s_{i\Delta t}, \mu_{i\Delta t}, a_{i\Delta t})$$

Ignore the SDE structure between the discrete-time data

PhiBE

$$\pi_{PH}^*(\mu) = \arg \max_{\pi} V_{PH}^{\pi}(\mu) = \arg \max_{\pi} \mathbb{E} \left[\int_0^{\infty} e^{-\beta t} r(\hat{\mu}_t) dt \middle| \mu_0 = \mu \right]$$

$$ds_t = \hat{b}^{\pi}(s_t, \mu_t) dt + \hat{\sigma}^{\pi}(s_t, \mu_t) dB_t, \quad \mu_t = \text{Law}(s_t).$$

$$\hat{b}(s, \mu, a) = \mathbb{E}_{s' \sim \rho_{\Delta t}} \left[\frac{s' - s}{\Delta t} \middle| s, \mu, a \right] \quad \hat{\sigma} \hat{\sigma}^{\top} = \hat{\Sigma}(s, \mu, a) = \mathbb{E}_{s' \sim \rho_{\Delta t}} \left[\frac{(s' - s)(s' - s)^{\top}}{\Delta t} \middle| s, \mu, a \right]$$

How good is the PhiBE approximation?

General Dynamics

[Bayraktar - Hernandez - Yan - Z. - 26]

$$d(\pi_{Ph}^*, \pi^*) = \|V^{\pi_{Ph}} - V^*\| \lesssim O(\Delta t)$$

LQ-MFC

[Bayraktar - Hernandez - Yan - Z. - 26]

$$\|V^{\pi_{Ph}^*} - V^*\| \lesssim O(\Delta t^2)$$

$$\text{When } \beta = 0: \mathbb{E}_{s \sim \mu}[\pi_{Ph}^*(s, \mu)] = \mathbb{E}_{s \sim \mu}[\pi^*(s, \mu)]$$

How to solve PhiBE-MFC?

Actor Critic

Collect data under π_ω

$$\{s_i, a_i \sim \pi_\omega, r_i, s'_i\}_{i=1}^n$$

Critic:

Approximate $V^\pi(\mu)$

Actor:

$$\omega \leftarrow \omega + \eta \nabla_\omega V^{\pi_\omega}$$

- [Frikha-Germain-Lauriere-Pham-Song-2025]
 - $G(\mu) = \nabla_\omega V^\pi$ satisfies a new linear HJB equation:
$$\beta G = \hat{r}(V^\pi, \nabla_\theta b, \nabla_\theta \Sigma) + \mathcal{L}_{b, \Sigma} G$$
 - Another critic step before the actor update
 - In general, $\nabla_\theta b, \nabla_\theta \Sigma$ is not computable when the dynamics are unknown

We construct an explicit expression of $\nabla_\omega V^\pi$ in terms of V^π

Policy Gradient in MFC

Instantaneous Advantage Function

$$q^\pi(\mu, s, a) := r(s, \mu, a) - \lambda \log \pi(a \mid s, \mu) + b(s, \mu, a) \cdot \partial_\mu V^\pi(\mu)(s) + \frac{1}{2} \Sigma(s, \mu, a) : D_s \partial_\mu V^\pi(\mu)(s).$$

- [Wei-Yu-25] [Ren-Wei-Yu-Zhou-26]

At each iteration

- Compute $q^h(s, \mu; \pi) = \mathbb{E}_{a \sim h(\cdot \mid s, \mu)}[q^\pi(s, \mu, a)]$ for $\forall h(\cdot \mid s, \mu) \in B_\delta(\pi)$
- Requires $\exp\left(\frac{1}{\lambda} q_\theta(\mu, s, a)\right) = 1$ for $\forall s, \mu$

Policy Gradient in MFC

$$J(\varepsilon) := V^{\pi^\varepsilon}(\mu). \quad \pi^\varepsilon(da \mid s, \mu) := \pi(da \mid s, \mu) + \varepsilon \varphi(da \mid s, \mu) = (1 + \varepsilon \psi(s, \mu, a)) \pi(da \mid s, \mu),$$

The Gâteaux gradient

$$\left. \frac{d}{d\varepsilon} J(\varepsilon) \right|_{\varepsilon=0} = \mathbb{E}_{(\mu, s) \sim \rho_{\mu_0}^\pi} \left[\int_{\mathcal{A}} q^\pi(\mu, s, a) \varphi(da \mid s, \mu) \right].$$

The parametric gradient

$$\nabla_\omega J(\omega) = \mathbb{E}_{\substack{(\mu, s) \sim \rho_{\mu_0}^\pi \\ a \sim \pi(\cdot \mid s, \mu)}} \left[q^\pi(\mu, s, a) \nabla_\omega \log \pi(a \mid s, \mu) \right]$$

$\rho_{\mu_0}^\pi(s, \mu)$: discounted occupancy measure under policy π with initial distribution μ_0

$$\rho_{\mu_0}^\pi(s) = \frac{1}{\beta} \int_0^\infty e^{-\beta t} \mu_t(s) dt \quad \rho_{\mu_0}^\pi(\mu) : \rho_{\mu_0}^\pi(B) = \frac{1}{\beta} \int_0^\infty e^{-\beta t} \mathbf{1}_B(\mu_t) dt, \quad \forall B \in \mathcal{B}(\mathcal{P}_2(\mathbb{R}^d))$$

The policy gradient update remains robust under biased distribution

How to solve PhiBE-MFC?

Actor Critic

Collect data under π_ω

$$\{s_i, a_i \sim \pi_\omega, r_i, s'_i\}_{i=1}^n$$

Critic:

Approximate $V^\pi(\mu)$

Actor:

$$\omega \leftarrow \omega + \eta \nabla_\omega J(\omega)$$

Mean-Field-PhiBE $\nabla_\omega \hat{J}(\omega) = \mathbb{E}_{\substack{(\mu, s) \sim \rho_{\mu_0}^\pi \\ a \sim \pi(\cdot | s, \mu)}} \left[\hat{q}^\pi(\mu, s, a) \nabla_\omega \log p_\pi(a | s, \mu) \right]$

$$\hat{q}^\pi(\mu, s, a) := r(s, \mu, a) - \lambda \log \pi(a | s, \mu) + \hat{b}(s, \mu, a) \cdot \partial_\mu V^\pi(\mu)(s) + \frac{1}{2} \hat{\Sigma}(s, \mu, a) : D_s \partial_\mu V^\pi(\mu)(s).$$

$$q_i = \hat{q}^\pi(\mu_i, s_i, a_i) := r_i - \lambda \log \pi(a_i | s_i, \mu_i) + \frac{s'_i - s_i}{\Delta t} \cdot \partial_\mu V^\pi(\mu)(s_i) + \frac{1}{2\Delta t} (s'_i - s_i)^\top D_s \partial_\mu V^\pi(\mu)(s_i) (s'_i - s_i).$$

$$\nabla_\omega J(\omega) \approx \sum_{i=1}^n q_i \log \pi(a_i | s_i, \mu_i)$$

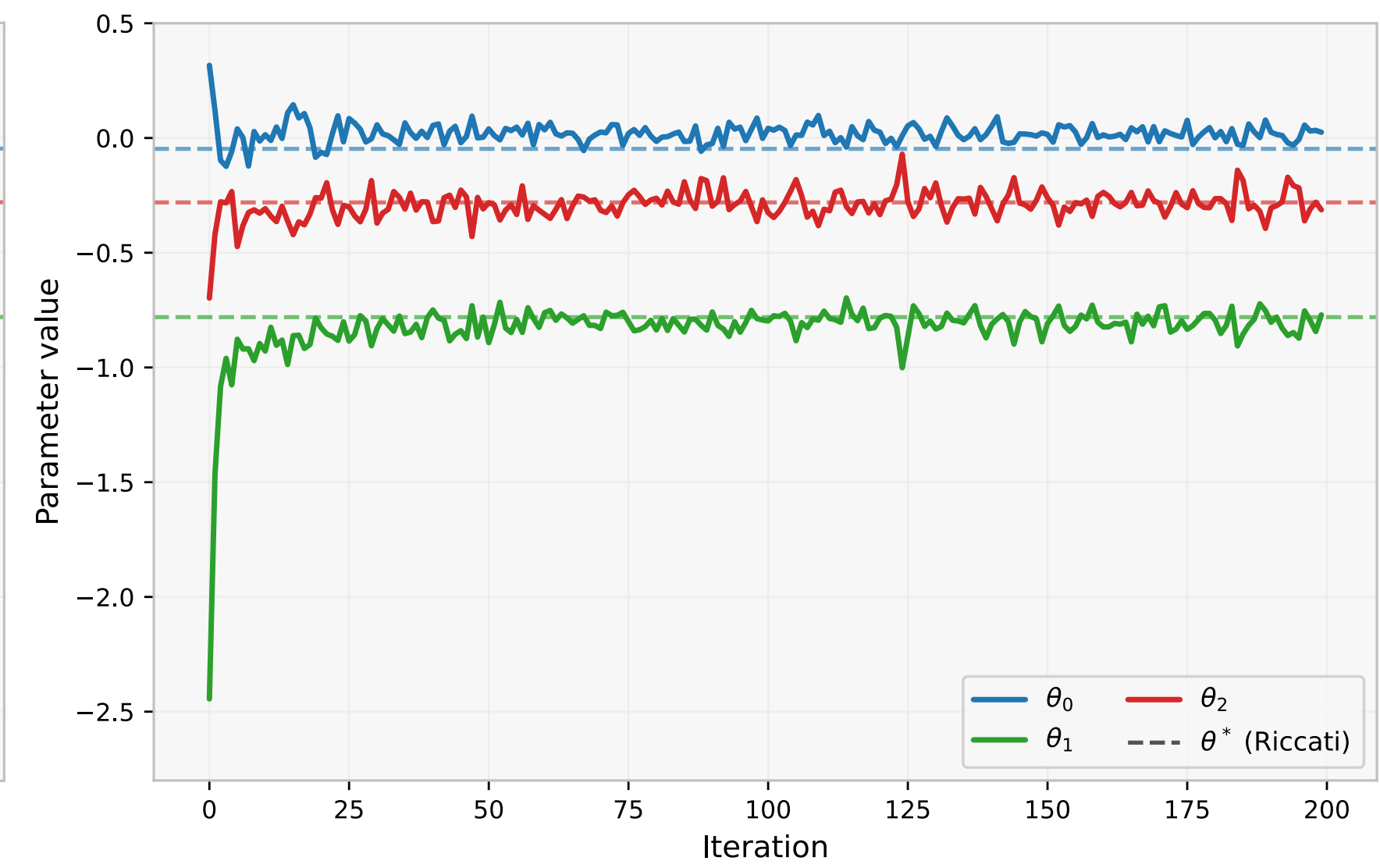
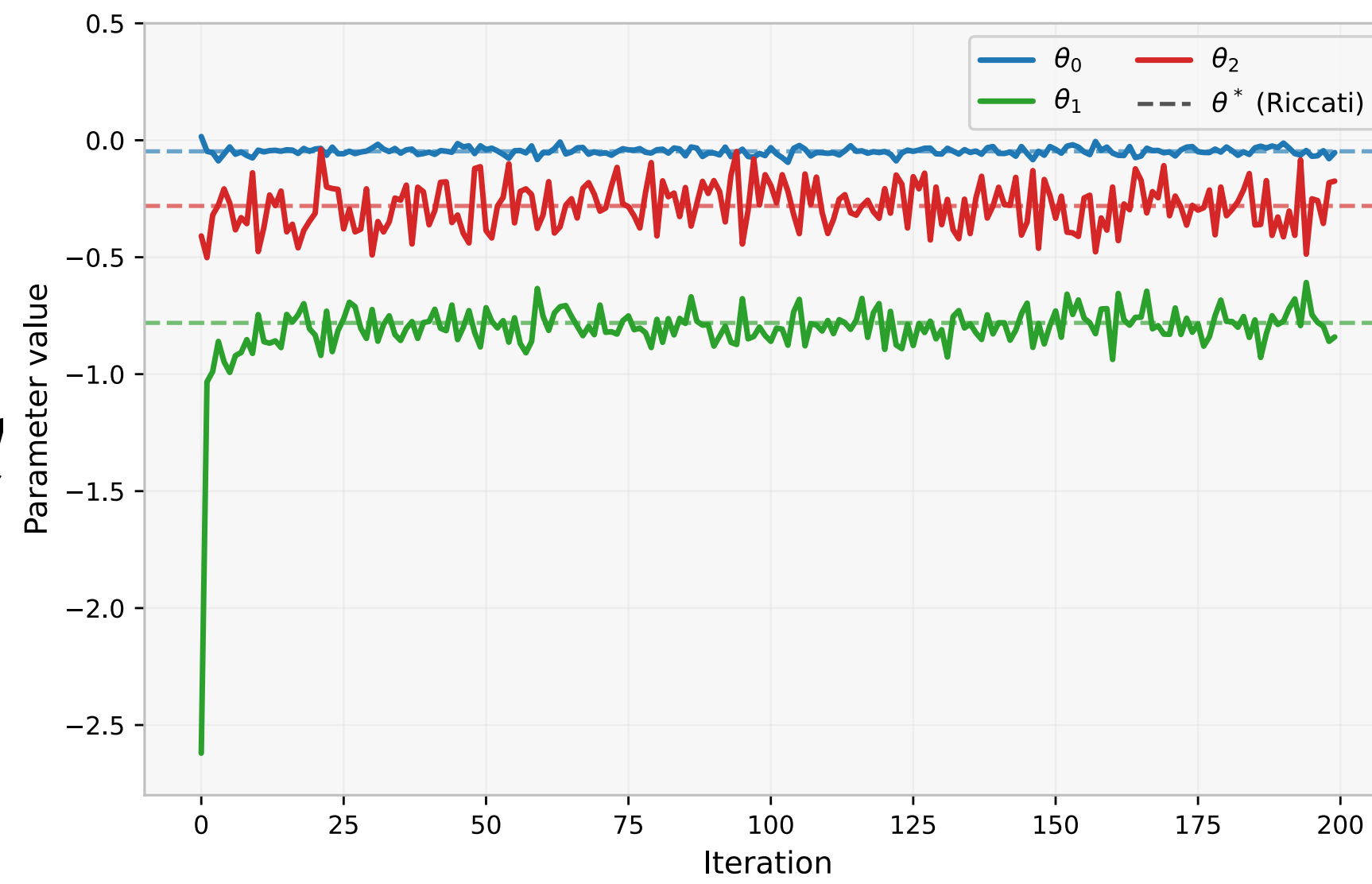
- Keep the McKean–Vlasov structure of the problem
- Model-free: same algorithm for any b, σ
- Computationally efficient

Known dynamics

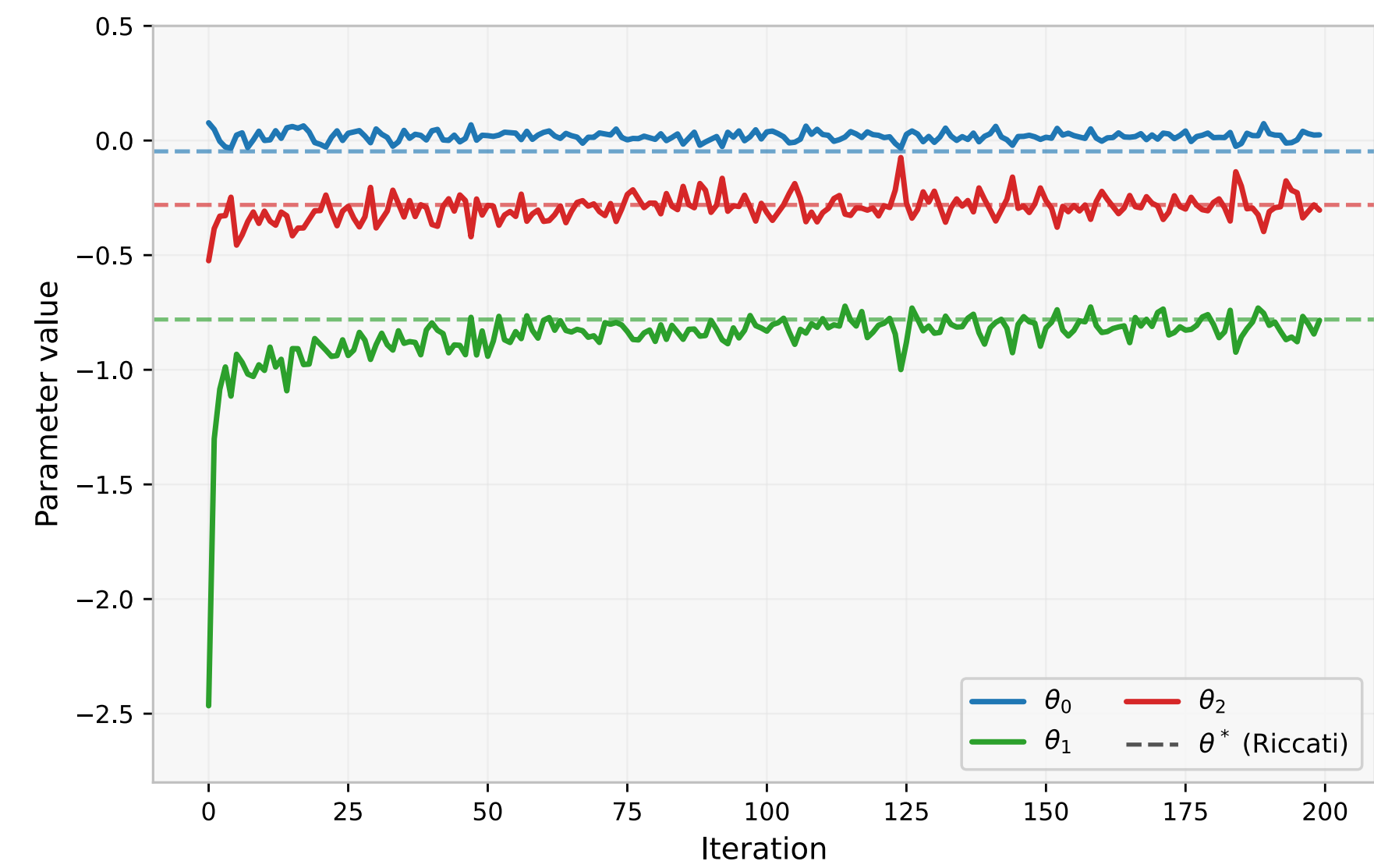
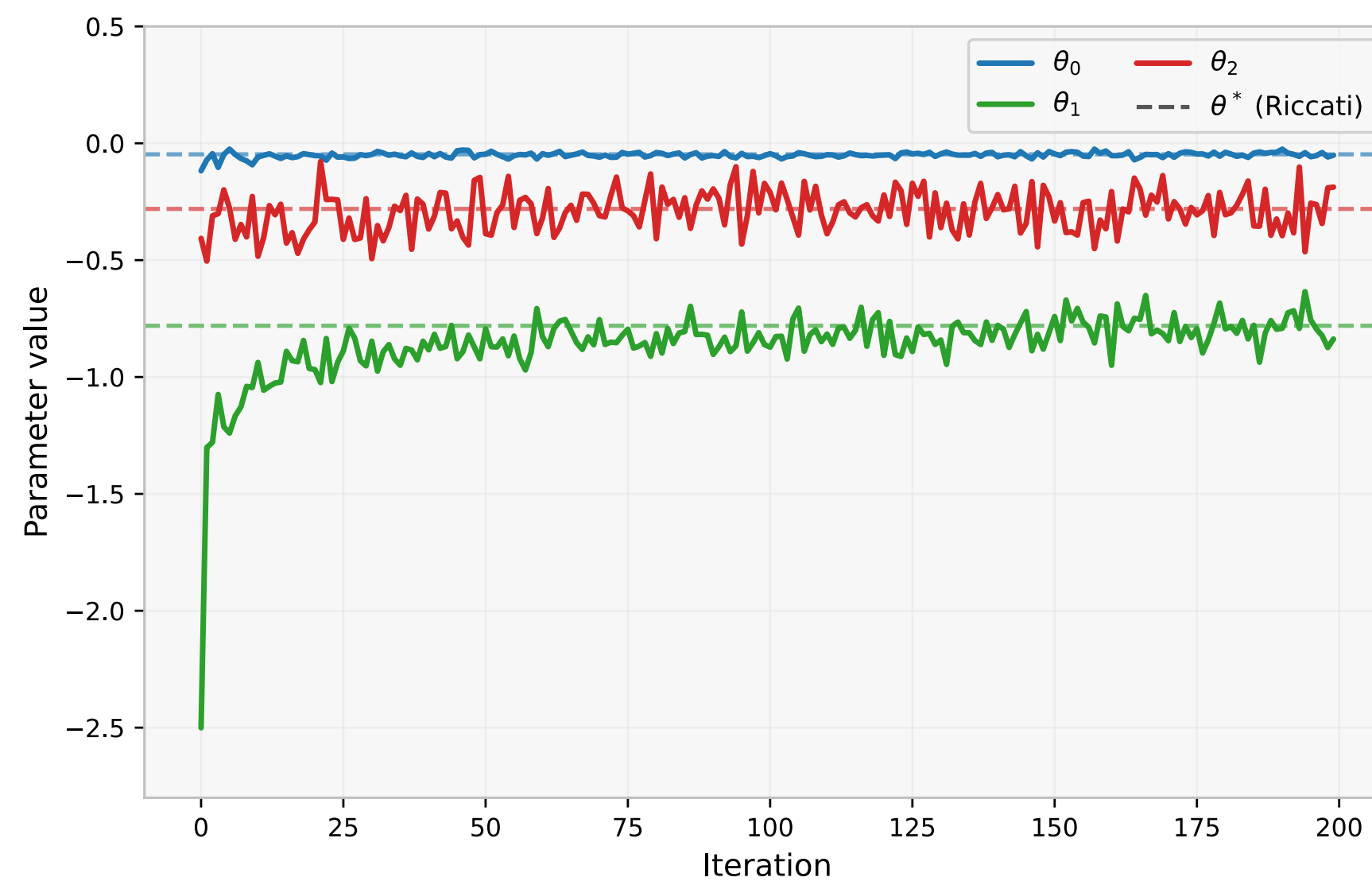
MF-LQ

Unknown dynamics

$$(s, \mu) \sim \rho_{\mu_0}^{\pi} d$$



Biased measure



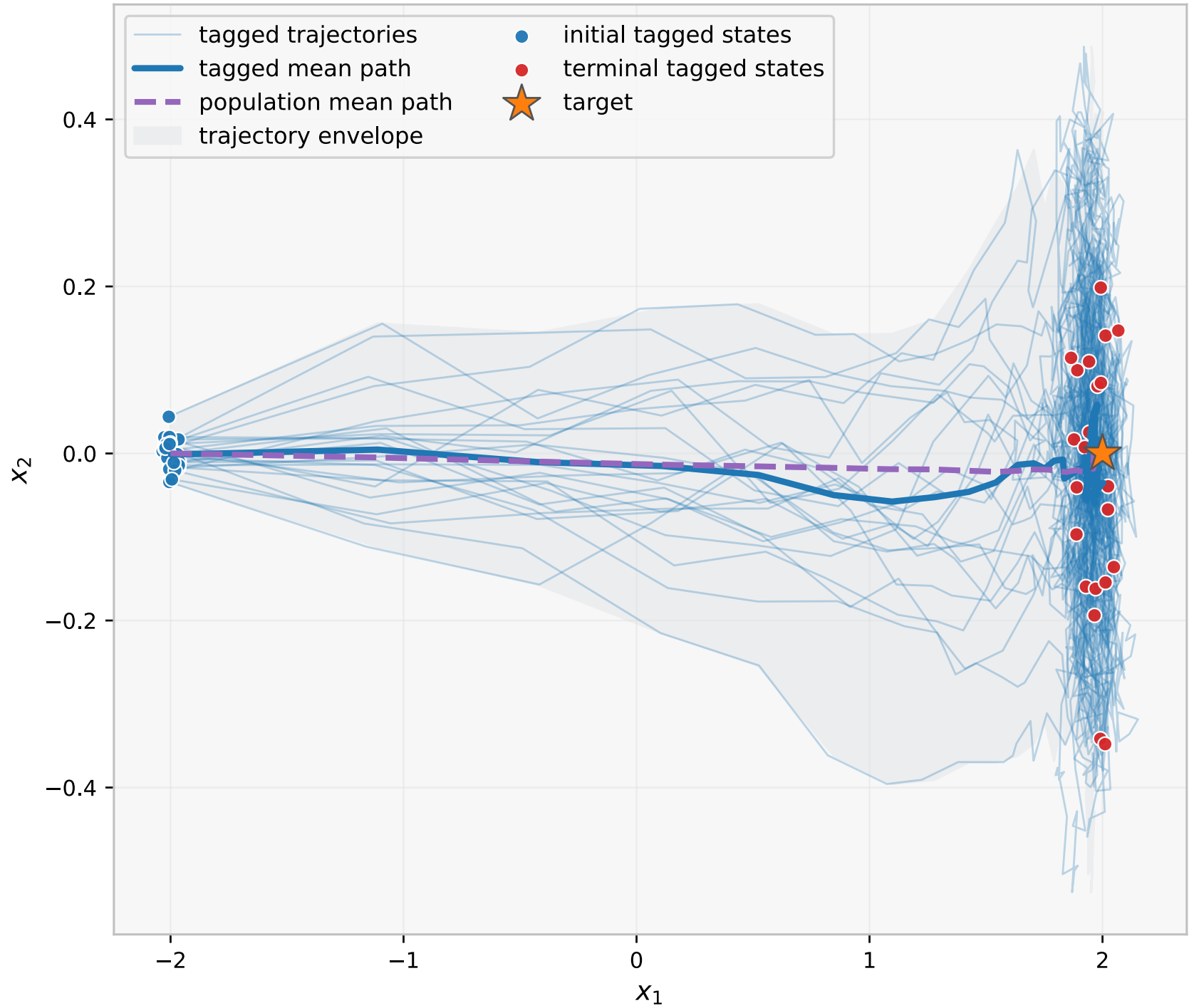
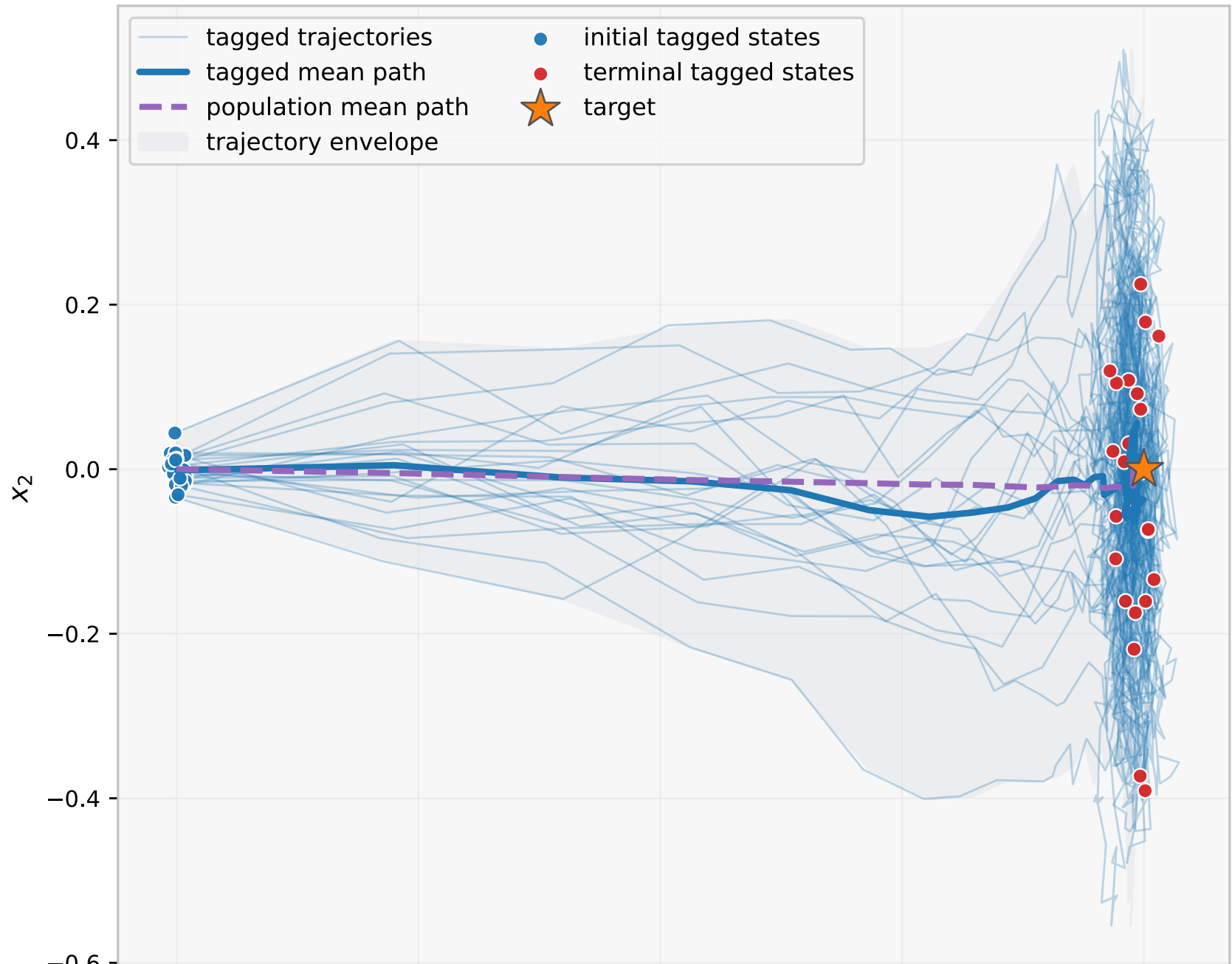
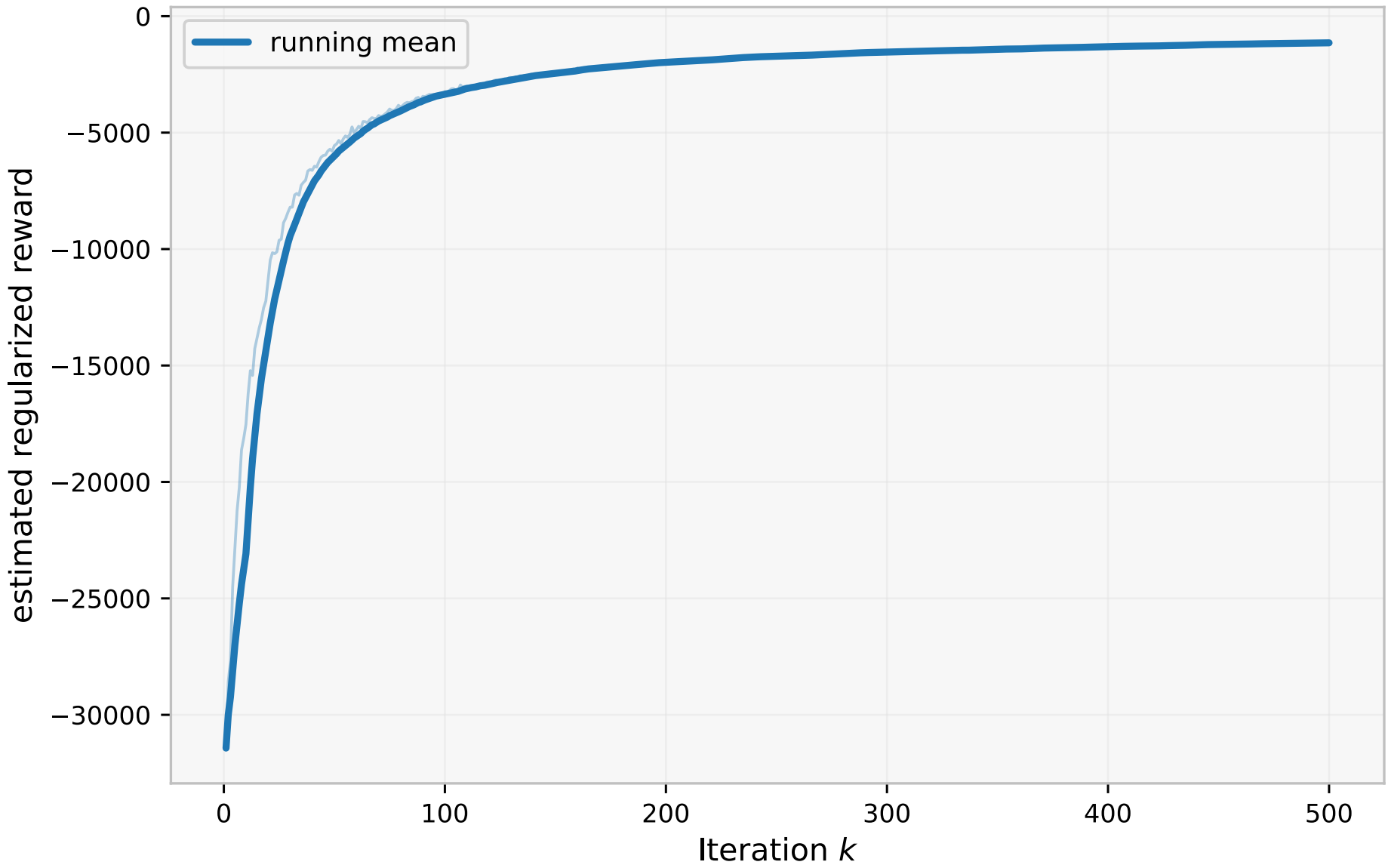
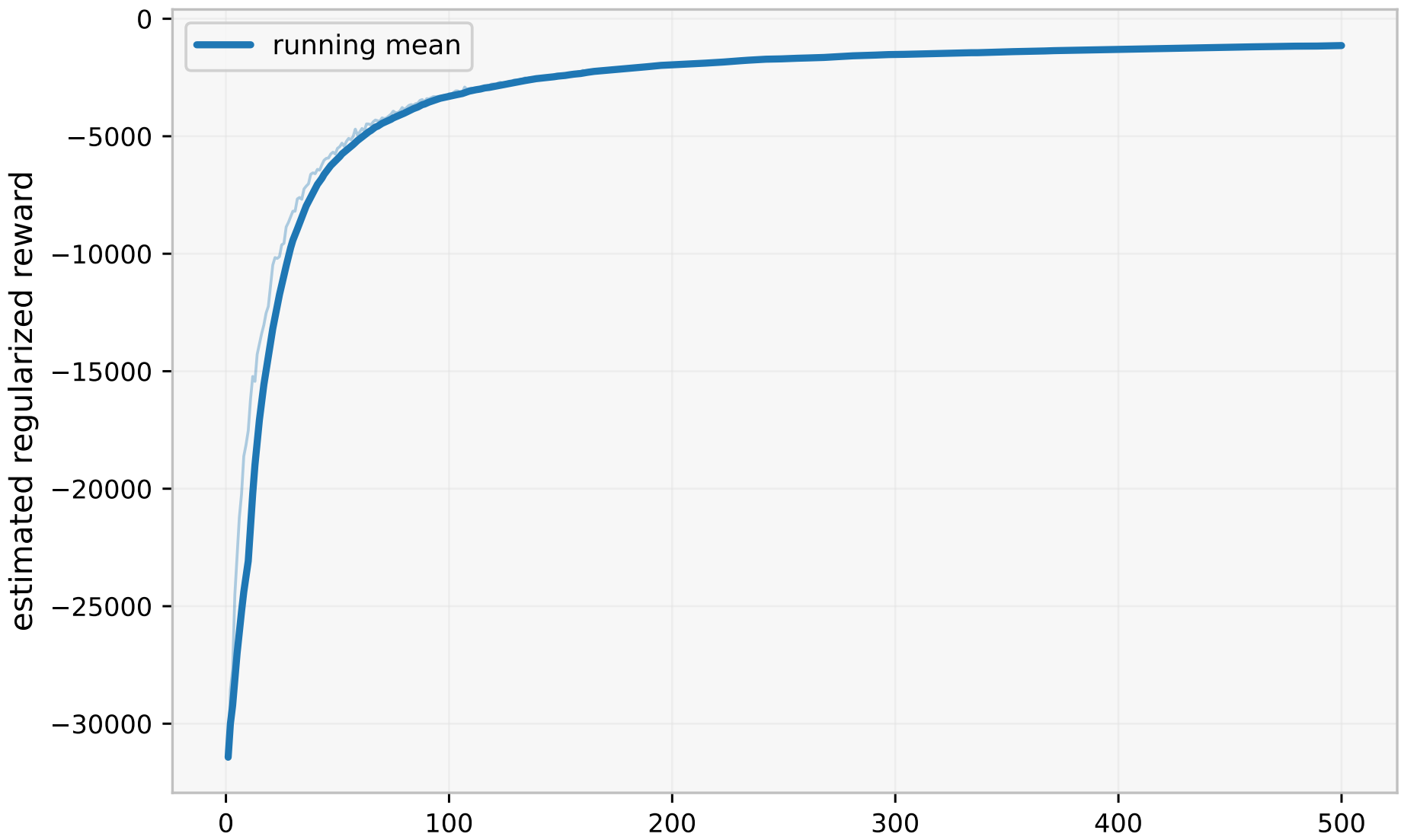
Crowd-aversion Tra

$$r(s, \mu, a) = - \left(\right.$$

$$(s, \mu) \sim \rho_{\mu_0}^{\pi}$$

$$\sigma = 0.2$$

Biased measure



$$\left. \right) \mu(ds'),$$

Related papers

- Yuhua Zhu, 2024. PhiBE: A PDE-based Bellman Equation for Continuous-Time Policy Evaluation
- Wenlong Mou, Yuhua Zhu, 2024. On Bellman equations for continuous-time policy evaluation: high-order discretization and function approximation
- Yuhua Zhu, Yuming Zhang, Haoyu Zhang, 2025. Optimal-PhiBE: A PDE-based Model-free framework for Continuous-Time Reinforcement Learning
- Yuhua Zhu, Yutong Ren, 2026. PhiBE-Q-Learning: Bridging Off-Policy Reinforcement Learning and Continuous-Time Control
- Huyen Pham, Yuming Zhang, Yuhua Zhu, 2026. Discretization error from regularized Reinforcement Learning to continuous-time stochastic control
- Erhan Bayraktar, Martin Hernandez, Qinxin Yan, Yuhua ZHU, 2026. Policy Gradient for Continuous-Time Mean-Field Control
- Erhan Bayraktar, Martin Hernandez, Qinxin Yan, Yuhua ZHU, 2026. Mean-Field PhiBE: Continuous-Time Mean-Field Reinforcement Learning from Discrete-Time Data
-