Borrowing From the Future

—— Addressing Double Sampling in Model-free Control

Yuhua Zhu - Stanford University

Joint work with
Zach Izzo - Stanford University
Lexing Ying - Stanford University

Double Sampling problem

Borrow From the Future Algorithm

Numerical experiments

Markov Decision Process (MDP)

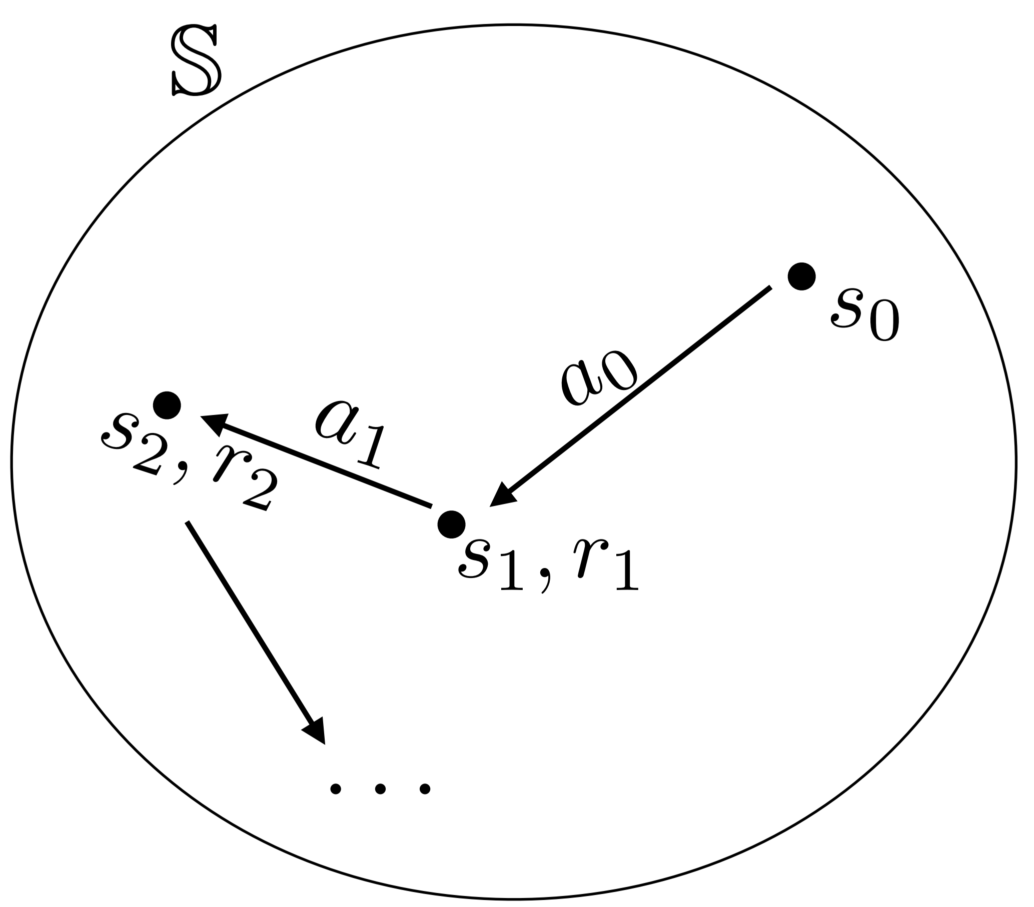
A discrete time stochastic process modeling decision making

MDP

- **State space:** $\mathbb{S} \subset \mathbb{R}^{d_s}$ is a compact set
- **Action space:** $a \in \mathbb{A}$
- **Transition matrix:**

$$\mathbb{P}_a(s, s') = \Pr(s_{m+1} = s' | s_m = s, a_m = a)$$

- **Immediate reward:** $r(s, a)$



- **Policy:** $\pi(s)$ **specifies the action at state s.**

Given a policy, MDP generates a trajectory $\{(s_t, a_t, r_t)\}_{t \geq 0}$.

Value function and Bellman operator

- **State-action value function** $Q^\pi(s,a)$**:**
  The expected discounted cumulative reward starting from state s **and action** a **if policy** π **is applied.**

given a policy — discount factor $\in (0,1)$ — Start at s with action a

$$Q^{\pi}(s,a) = \mathbb{E}\left[r(s_0,s_1) + \gamma r(s_1,s_2) + \cdots + \gamma^t r(s_t,s_{t+1}) + \cdots \middle| (s_0,a_0) = (s,a)\right].$$

Goal of Reinforcement Learning: find the best policy that maximizes the return

$$Q^*(s,a) = \max_{\pi} Q^{\pi}(s,a)$$

The state-action value function under the optimal policy satisfies the optimal Bellman equation:

$$Q^* = \mathbb{T}^* Q^* \longrightarrow$$ Q^* is the fixed point of $\mathbb{T}^*$

$$\mathbb{T}^* Q(s,a) = R(s) + \gamma \mathbb{E}[\max_{a'} Q(s_1,a') | (s_0,a_0) = (s,a)]$$

Optimization problem in model-free control

Based on the **contractive property** of the Bellman operator $\mathbb{T}^*$:

$$Q_{k+1} = \mathbb{T}^* Q_k \rightarrow Q^*$$

Iterative methods, such as Q learning, DQN are all based on the contractive property of the Bellman operator.

However,

- When the state space is large, computational cost is large.
- When the discount factor close to 1, the convergence rate is slow.

Function Approximation

Consider parameterized form $Q_\theta(s, a)$: → **No longer contractive**

Another approach: **Fixed point problem** → **Optimization problem**

$$\min_\theta \; \frac{1}{2} \mathbb{E}[(Q - \mathbb{T}^* Q)^2]$$

- **The expressive of nonlinear functions, such as DNN**
- **Less computational cost for continuous state space**
- **More stable than variants of Q-learning methods**

However, There is double sampling problem in this formulation.

Model-free RL and Double Sampling Problem

$$\min_{\theta} \frac{1}{2}\mathbb{E}[(Q - \mathbb{T}^* Q)^2]$$

with a trajectory $\{s_t\}_{t=0}^{T}$ generated from an underlying transition dynamics

$$s_{t+1} = s_t + \alpha(s_t, a_t)\epsilon + \sqrt{\epsilon} Z_t, Z_t \sim N(0,1)$$

Model-free RL: → **unknown !**

Only a trajectory is available in model-free RL!

Double Sampling Problem

Gradient of the objective function: $\mathbb{E}[(Q - \mathbb{T}^* Q)\nabla_\theta (Q - \mathbb{T}^* Q)]$

$$\mathbb{E}[(Q - R - \gamma \mathbb{E}[\max_a Q(s_{t+1}, a) | s_t, a_t])\nabla_\theta (Q - R - \gamma \mathbb{E}[\max_a Q(s_{t+1}, a) | s_t, a_t])]$$

Two independent expectations on the next state

Unbiased gradient: $(Q(s_t, a_t) - R_t - \gamma \max_a Q(s_{t+1}, a))\nabla_\theta (Q(s_t, a_t) - R_t - \gamma \max_a Q(s'_{t+1}, a))$

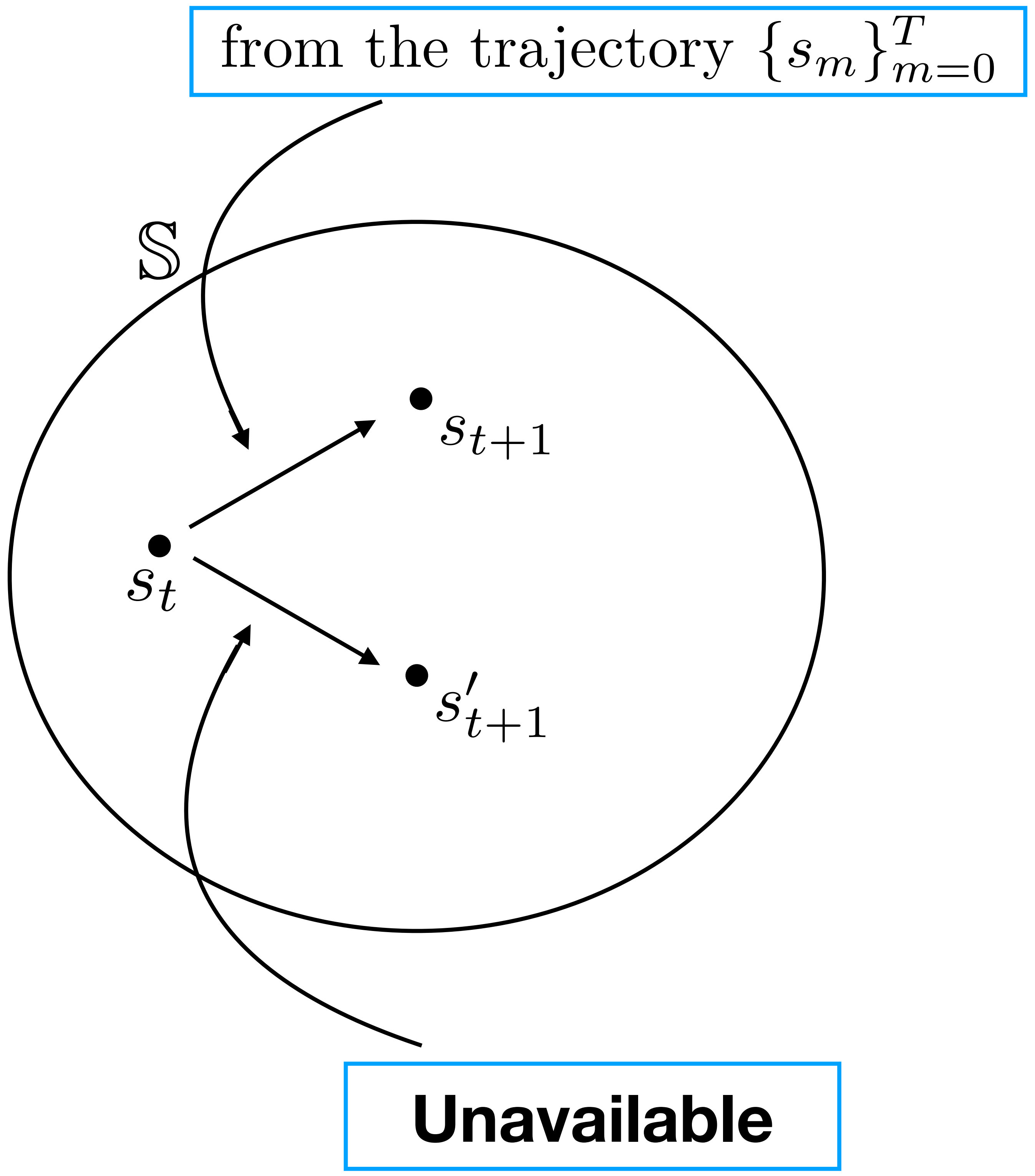
Two independent samples for the next state

Double Sampling Problem

$$\min_{\theta} \frac{1}{2}\mathbb{E}[(Q - \mathbb{T}^* Q)^2]$$

Unbiased gradient: $(Q(s_t, a_t) - R_t - \gamma \max_a Q(s_{t+1}, a))\nabla_\theta (Q(s_t, a_t) - R_t - \gamma \max_a Q(s'_{t+1}, a))$

Two independent samples for the next state



Model-free RL:

Only the trajectory $\{s_t\}_{t=0}^T$ under the given policy is available !

- Trajectory is not recorded because of the high dimensionality.
- Hard to simulate exactly from the current state again.

Double Sampling problem

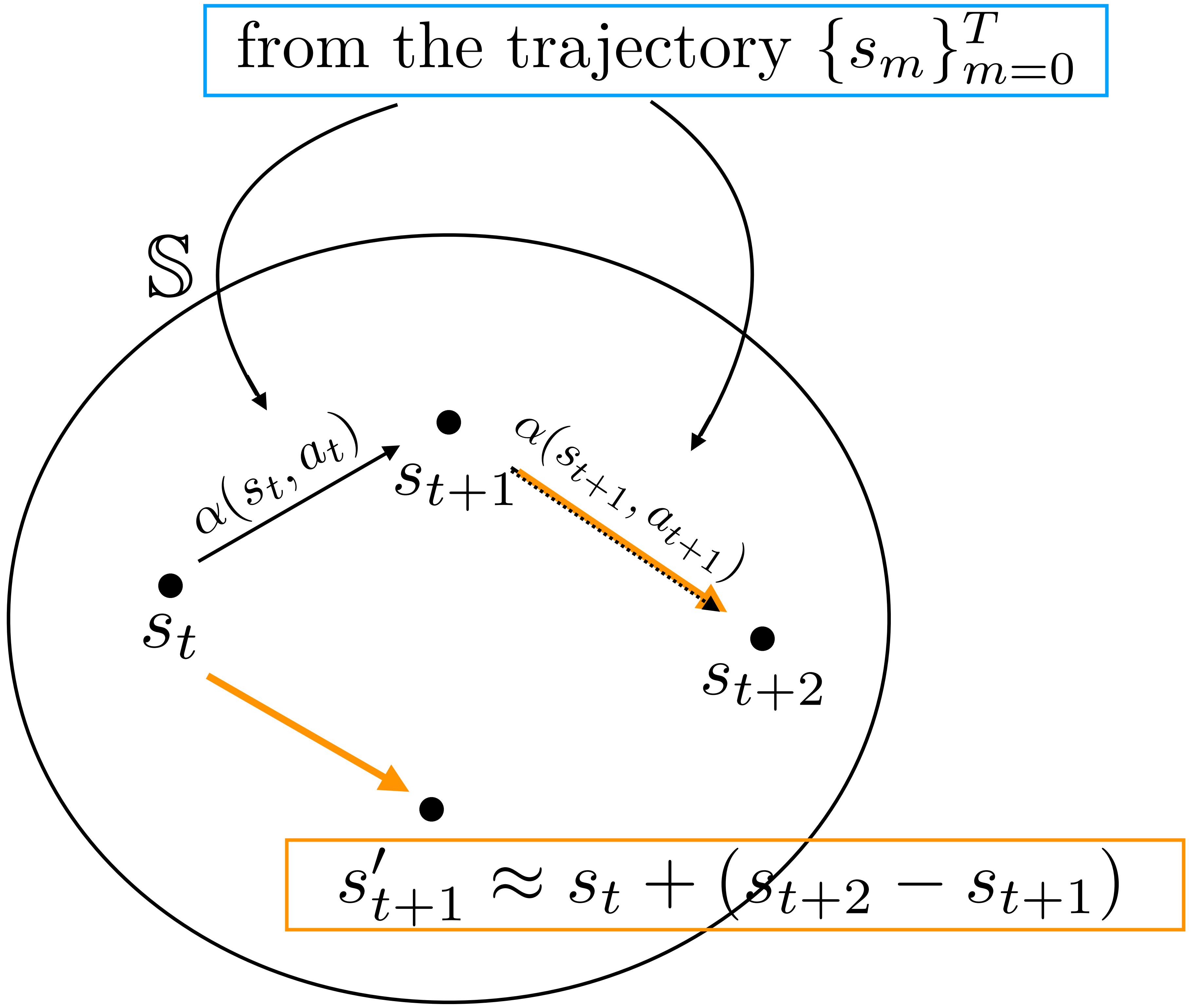
Borrowing From the Future

Numerical experiments

Borrowing From the Future

Unbiased gradient: $(Q(s_t, a_t) - R_t - \gamma \max_a Q(s_{t+1}, a))\nabla_\theta(Q(s_t, a_t) - R_t - \gamma \max_a Q(s'_{t+1}, a))$

The underlying transition: $s_{t+1} = s_t + \alpha(s_t, a_t)\epsilon + \sqrt{\epsilon} Z_t, Z_t \sim N(0, 1)$



Good approximation when the drift term is sufficiently smooth.

Borrow extra randomness from the future.

BFF model-free control

Unbiased gradient: $(Q(s_t, a_t) - R_t - \gamma \max_a Q(s_{t+1}, a))\nabla_\theta (Q(s_t, a_t) - R_t - \gamma \max_a Q(s'_{t+1}, a))$

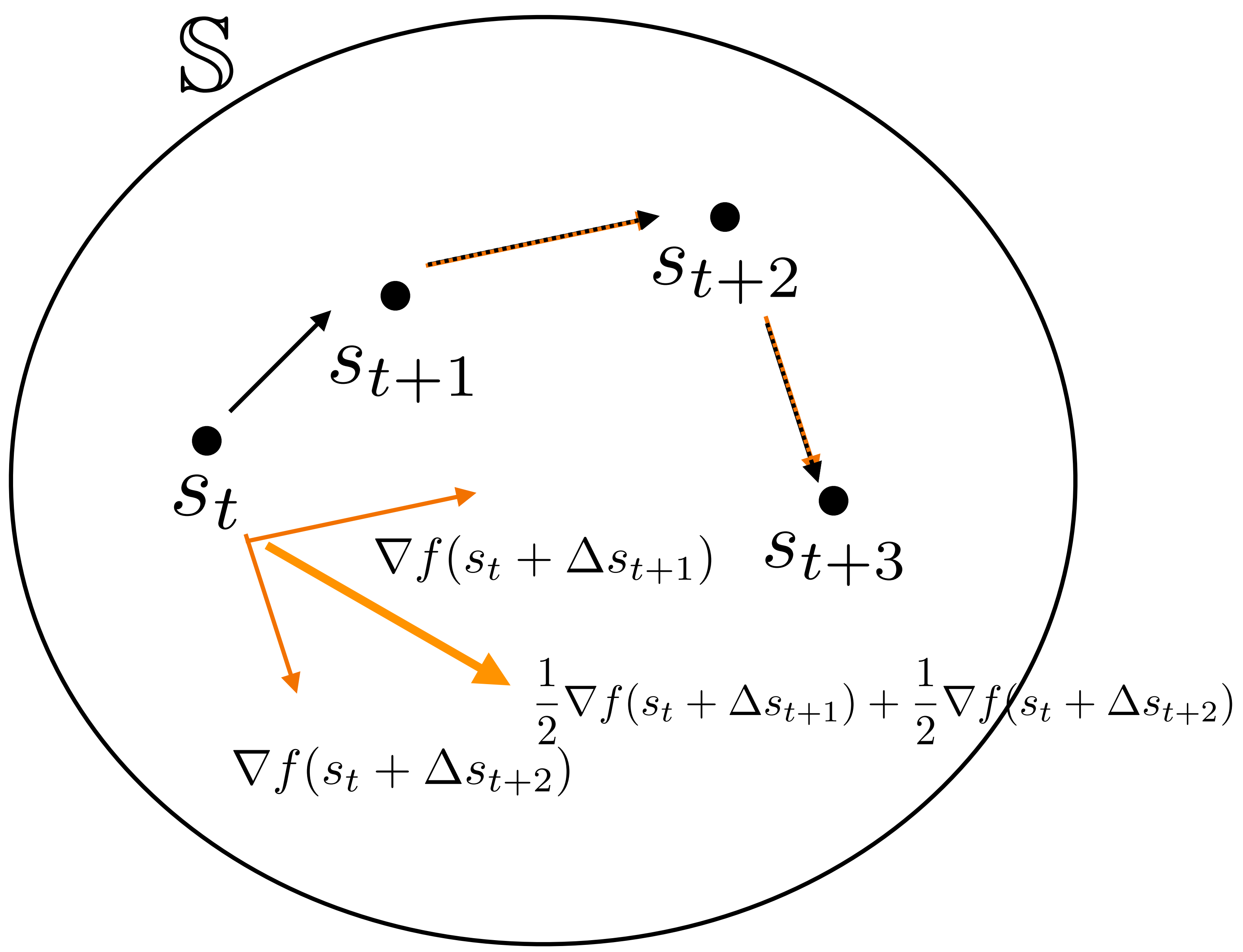
Unbiased SGD: $\theta_{k+1} = \theta_k - \tau f(s_t, s_{t+1}; \theta_k) \nabla_\theta f(s_t, s'_{t+1}; \theta_k)$

where $f(s_t, s_{t+1}; \theta) = Q(s_t, a_t) - R(s_t) - \gamma \max_a Q(s_{t+1}, a')$

BFF:

$s_t + \Delta s_{t+1},$
where $\Delta s_{t+1} = s_{t+2} - s_{t+1}$

nBFF



More generally, $\theta_{k+1} = \theta_k - \tau f(s_{t+1}) \sum_{i=1}^{n} w_i \nabla_\theta f(s_t + \Delta s_{t+i})$. with $\sum_{i=1}^{n} w_i = 1$

Theoretical results

$$\min_{\theta} \; \mathbb{E}\left[\frac{1}{2}\delta^2\right]$$

$$\text{where } \delta = Q - \mathbb{T}^* Q = Q(s_t, a_t) - R_t - \mathbb{E}[\max_a Q(s_{t+1}, a) | s_t, a_t]$$

$$\text{with underlying transition dynamics: } \; s_{t+1} = s_t + \alpha(s_t, a_t)\epsilon + \sqrt{\epsilon} Z_t, Z_t \sim N(0,1)$$

Assumption:

State space $\mathbb{S}$ and action space $\mathbb{A}$ can be embedded into a compact set.

Learning rate η is small.

The underlying dynamics change slowly w.r.t. actions: $\|\alpha(s, a_1) - \alpha(s, a_2)\| \leq C$.

Thm [Z-Izzo-Ying]

$$\| \text{ (p.d.f of BFF)} - \text{(p.d.f of unbiased SGD) } \|$$

$$\leq C_1 e^{-C_2 t} + O\left(\epsilon \sqrt{\mathbb{E}[\delta_*^2]}\right) \sqrt{1 - e^{-C_2 t}}$$

$\mathbb{E}[\delta_\star^2] = \min_\theta \mathbb{E}[\delta^2]$ is the smallest Bellman residual that the unbiased SGD can achieve

Double Sampling problem

Borrow From the Future Algorithm

Numerical experiments

Continuous state space

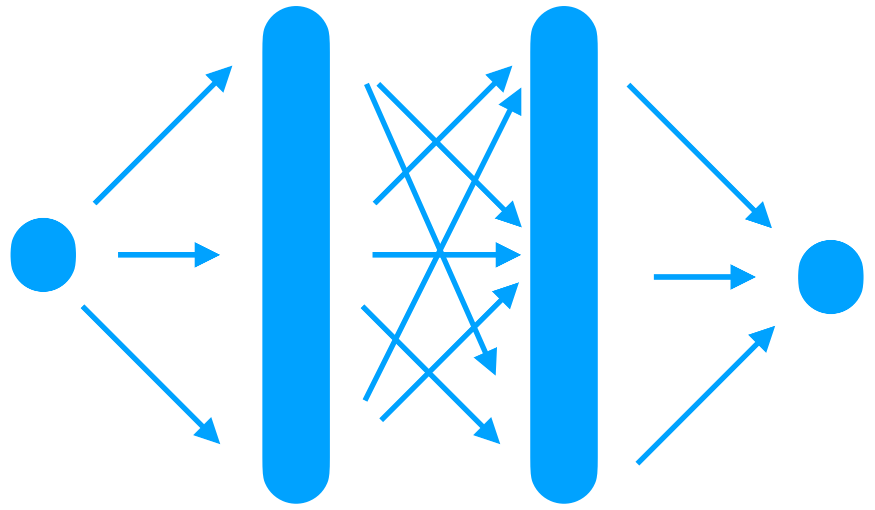
Underlying transition probability:

$s_{t+1} = s_t + a_t\epsilon + \sigma Z_t\sqrt{\epsilon},$

$a_t \in \mathbb{A} = \{\pm 1\},\ \epsilon = \frac{2\pi}{32},\ \sigma = 0.2.$

The reward function is $r(s_{t+1}, s_t, a_t) = \sin(s_{t+1}) + 1.$

$Q_\theta(s, a)$ is approximated by a 3-layer NN

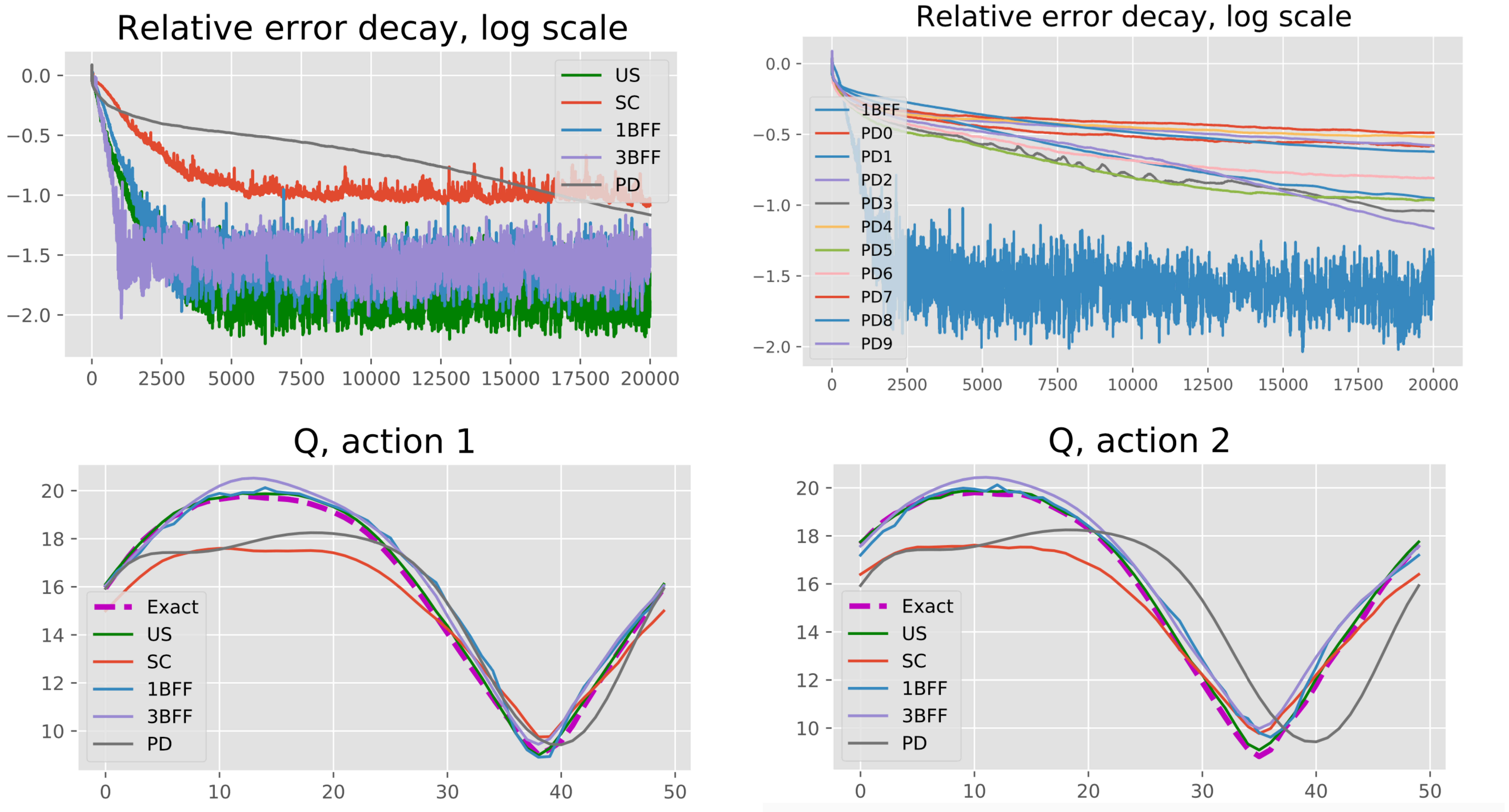


Compared BFF with:

- Uncorrelated sampling: $f(s_{t+1})\nabla f(s'_{t+1})$ ⟶ Unbiased SGD, but unrealistic!
- Sample Cloning: $f(s_{t+1})\nabla f(s_{t+1})$ ⟶ Commonly used biased SGD in practice, but less accurate than BFF.
- Primal-Dual: $\min_\theta \delta(\theta)^2 = \min_\theta \max_\omega \delta(\theta)y(\omega) - \frac{1}{2}y(\omega)^2$ GTD: Sutton (2008); SBEED: Dai et al. (2018)

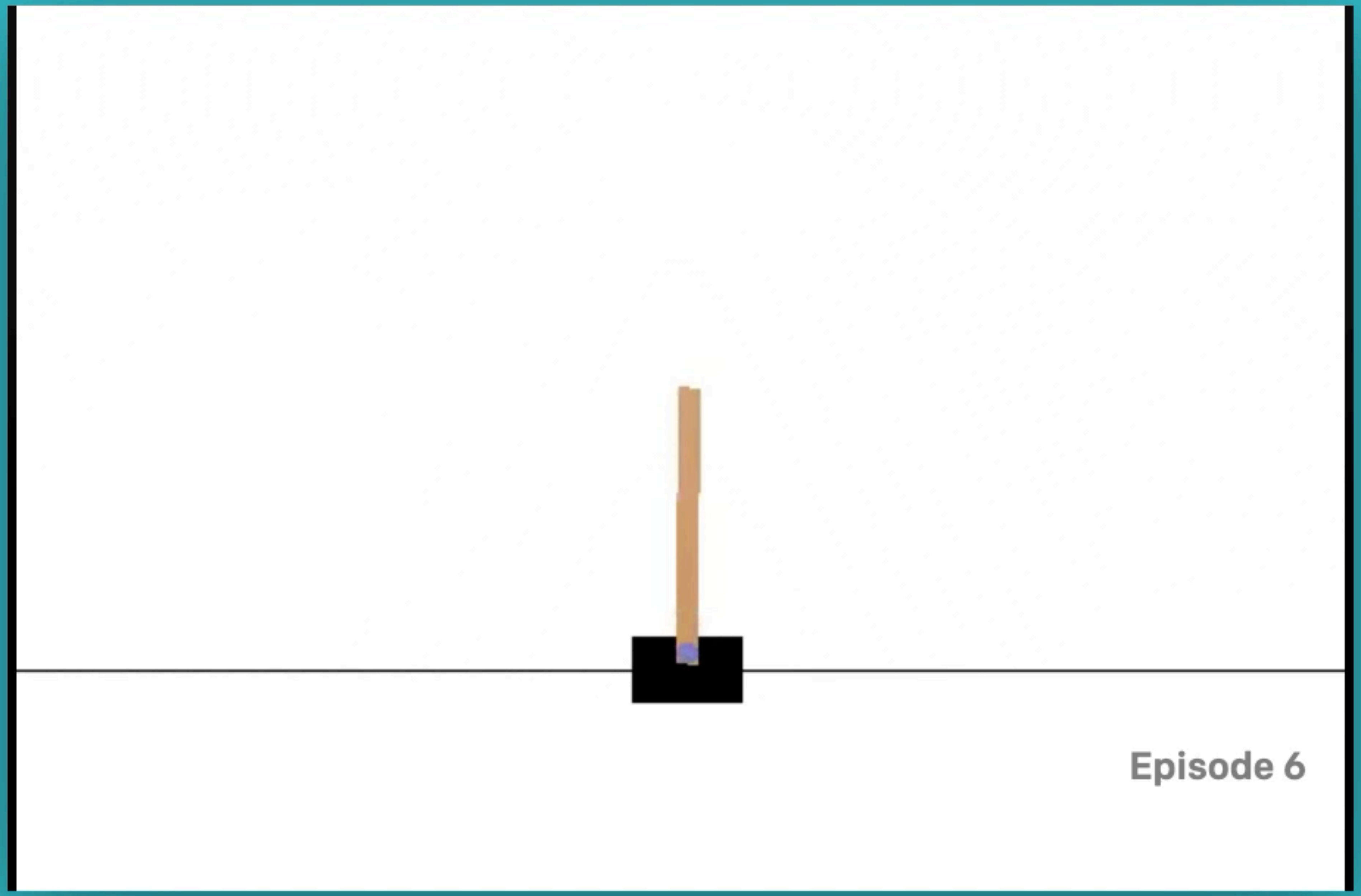
  ⟶ Not stable when the max is taken over non concave function

Q-control



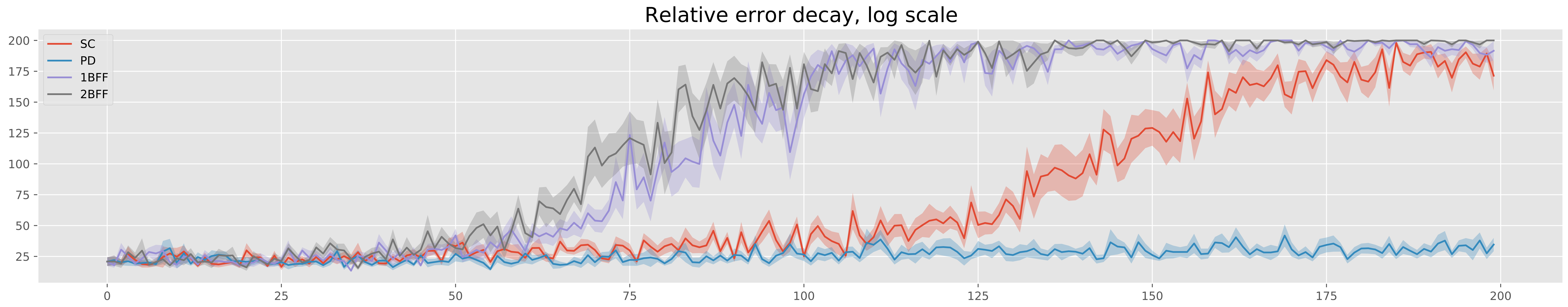
Relative error decay, log scale
US
SC
1BFF
3BFF
PD
Relative error decay, log scale
1BFF
PD0
PD1
PD2
PD3
PD4
PD5
PD6
PD7
PD8
PD9
Q, action 1
Exact
US
SC
1BFF
3BFF
PD
Q, action 2
Exact
US
SC
1BFF
3BFF
PD

Cartpole from Open AI Gym



RandomAgent on CartPole-v1

Cartpole



Relative error decay, log scale
SC
PD
1BFF
2BFF
200
175
150
125
100
75
50
25
0
25
50
75
100
125
150
175
200

Summary

- We propose a new algorithm BFF to alleviate the double sampling problem in the model-free control.
- BFF has an advantage over other BRM algorithms for model-free RL, especially for problems with continuous state spaces and smooth underlying dynamics.
- We prove that the difference between the BFF algorithm and the unbiased SGD first decays exponentially and eventually stabilizes at an error of $O(\delta_\star \epsilon)$, where $\delta_\star$ is the smallest Bellman residual that unbiased SGD can achieve.

Thanks!