澳門大學
UNIVERSIDADE DE MACAU
UNIVERSITY OF MACAU

Maximum diffusion reinforcement learning

Thomas A. Berrueta, Allison Pinosky, Todd D. Murphey

speaker: Xiongyi Li

Jan 12th 2024

Introduction

- Data is independent and identically distributed
- In RL, data collected sequentially are temporal correlated
- Solution: Sampling in random, experience replay
- Maximum entropy RL: maximize the entropy of an agent’s policy

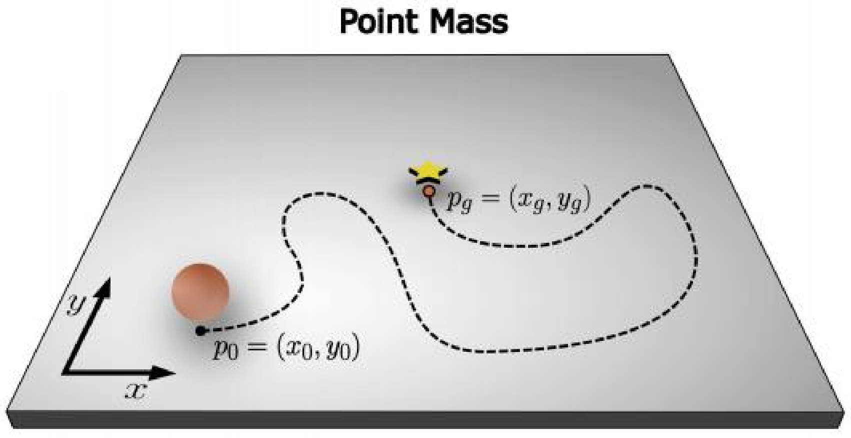
Maximum diffusion RL:

- Realizes statistics indistinguishable from i.i.d. sampling by exploiting the statistical mechanics of ergodic processes
- PROVE:
  - capable of single-shot learning regardless of how they are initialized
  - robust to random seeds and environmental stochasticity

Temporal correlations hinder performance

- Whether temporal correlations and their impact can be avoided depends on the properties of the underlying agent-environment dynamics.
- Completely destroying correlations between agent experiences requires the ability to discontinuously jump from state to state without continuity of experience.

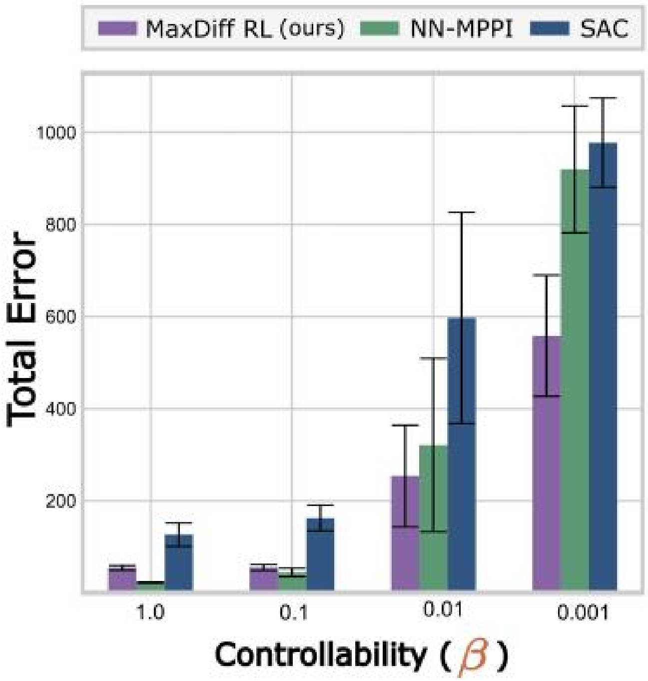
Task to evaluate deep RL algorithms

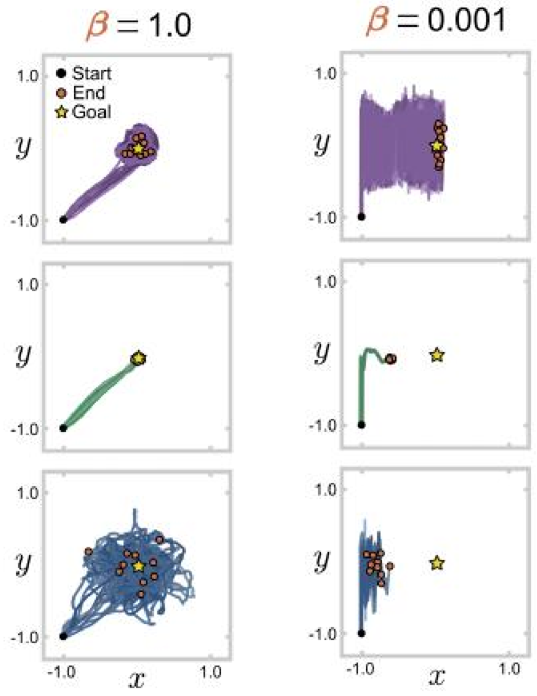


Dynamics: $\vec{x}_{t+1} = A\vec{x}_t + B\vec{u}_t$

$$A = \begin{bmatrix} 1 & 0 & \beta & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad B = \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 1 & 0 \\ 0 & 1 \end{bmatrix}$$

β = Controllability Parameter



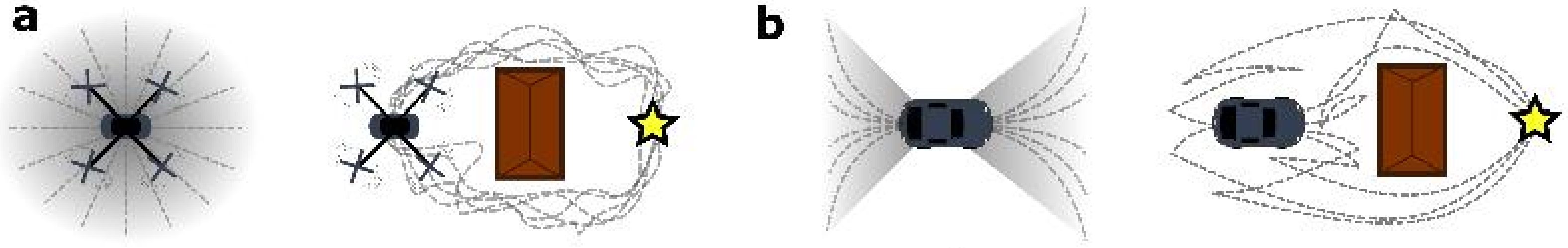


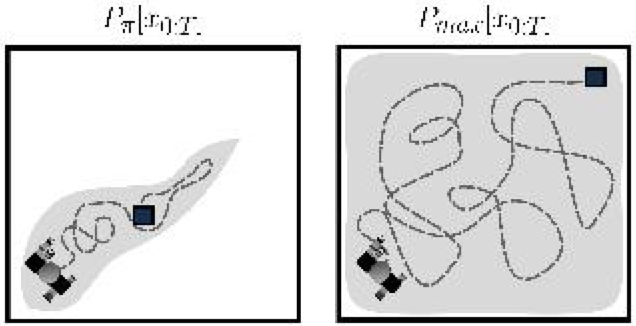
Maximum diffusion exploration and learning

- RL: take random actions to produce effective exploration
- MaxEnt RL: Maximize the entropy of a learned action distribution (policy)
- Propose: decorrelating agent experiences

What is the most decorrelated that agent experiences can be?

- Maximum entropy
  - trajectory distribution: $P_{max}[x(t)]$
  - optimal path distribution: $P_{max}[x(t)] = \frac{1}{Z}\exp\left[-\frac{1}{2}\int_{t_0}^{t}\dot{x}(\tau)^T\mathbf{C}^{-1}[x^*]\dot{x}(\tau)d\tau\right]$





Minimizing correlations among agent trajectories leads to diffusion-like exploration

- Maximally diffusive agent: satisfy optimal path distribution

$$P_{max}[x(t)] = \frac{1}{Z} \exp\left[-\frac{1}{2} \int_{t_0}^{t} \dot{x}(\tau)^T \mathbf{C}^{-1}[x^*] \dot{x}(\tau) d\tau \right]$$

- But agent can't realize maximally diffusive trajectories spontaneously
- Find a policy capable of satisfying maximally diffusive path statistics, which is the core of MaxDiff RL

KL

- KL divergence between:

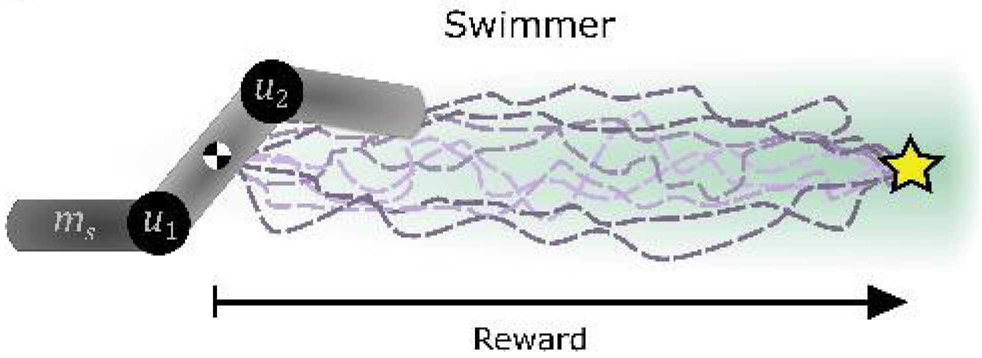
$$P_\pi[x_{0:T}, u_{0:T}] = \prod_{t=0}^{T-1} p(x_{t+1}|x_t, u_t)\pi(u_t|x_t)$$

$$P^r_{max}[x_{0:T}, u_{0:T}] = \prod_{t=0}^{T-1} p_{max}(x_{t+1}|x_t)e^{r(x_t,u_t)}$$

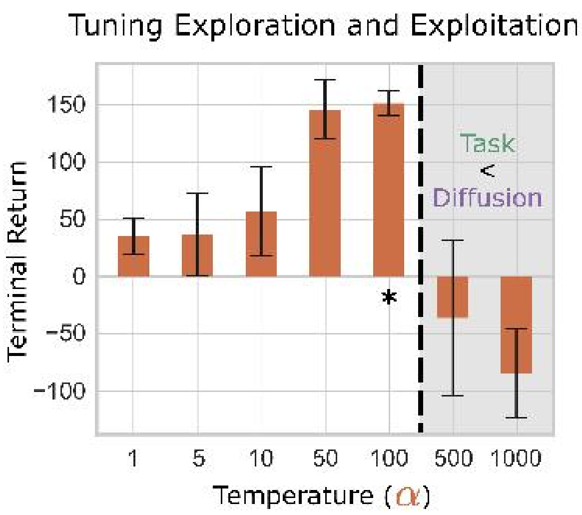
- Goal of MaxDiff RL
- SOC:

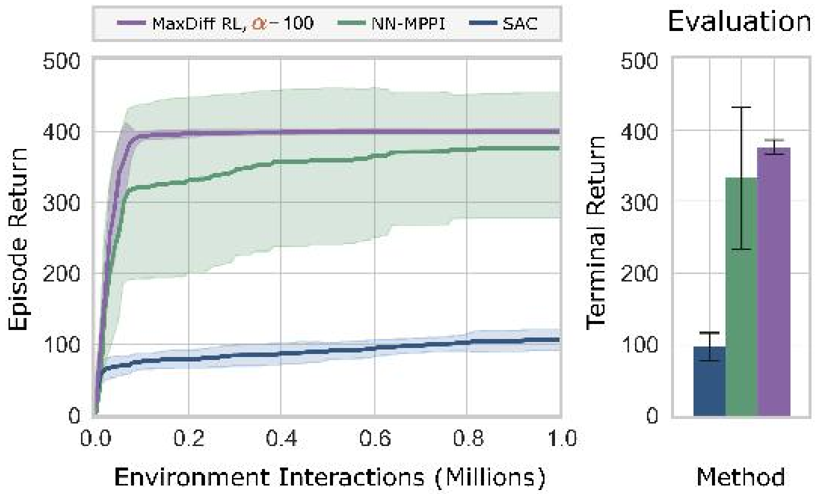
$$\operatorname*{argmax}_\pi E_{(x_{0:T},u_{0:T})\sim P_\pi}\left[\sum_{t=0}^{T-1} r(x_t, u_t) + \frac{\alpha}{2}\log\det \mathbf{C}[x_t]\right]$$

Robustness to initializations in ergodic agents

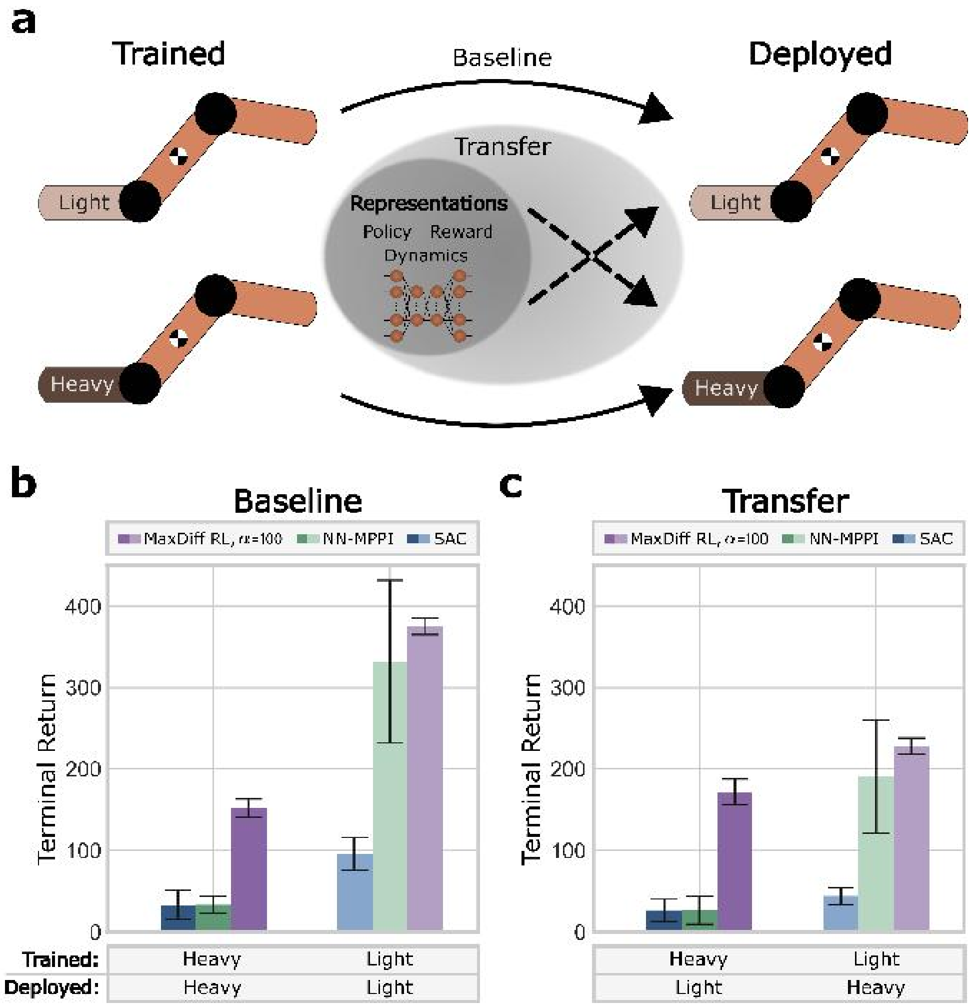


- Balance between achieving the task and diffusion
- Parameter: α



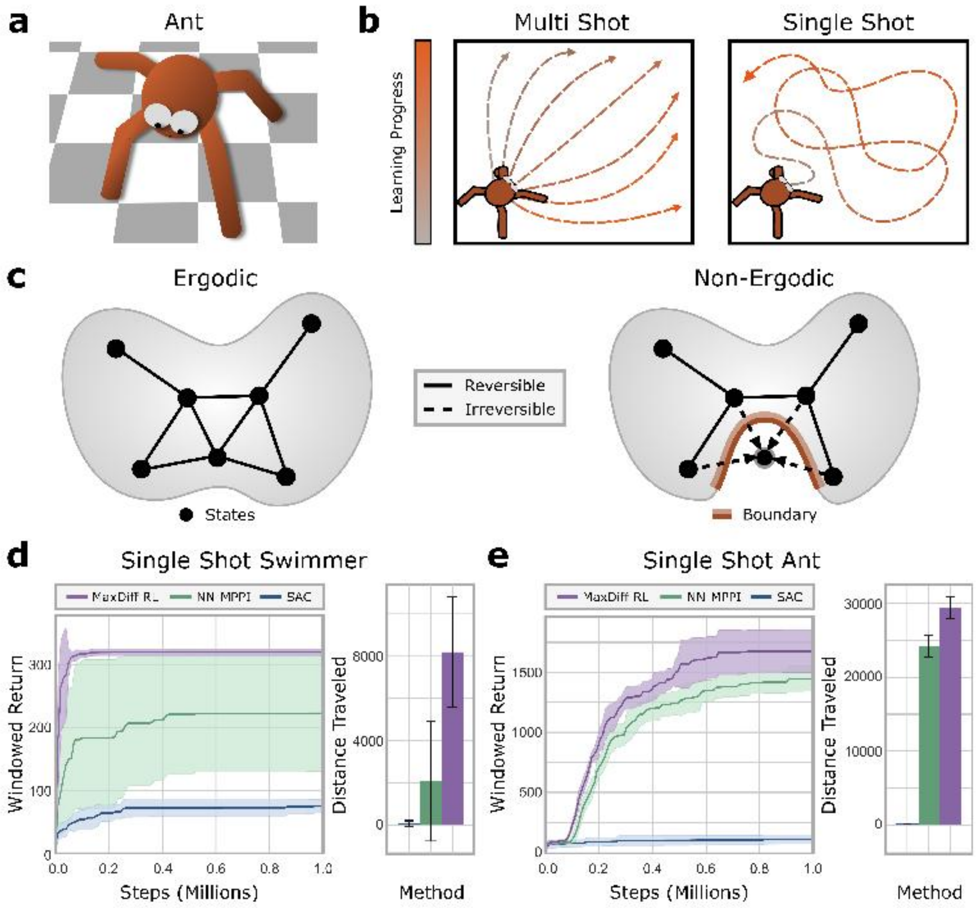


Zero-shot generalization across embodiments



- https://www.youtube.com/watch?v=eq6Fk-lp1i0&list=PLO5AGPa3klrCTSO-t7HZsVNQinHXFQmn9&index=3

Single-shot learning in ergodic agents



Thank You!

Avenida da Universidade, Taipa, Macau, China
Tel : (853) 8822 8833   Fax : (853) 8822 8822
Email : xiongyilee@outlook.com   Website : www.um.edu.mo

澳大