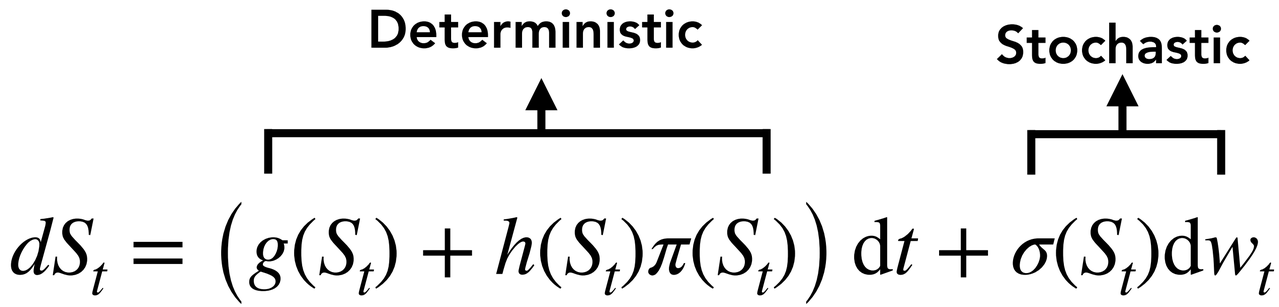
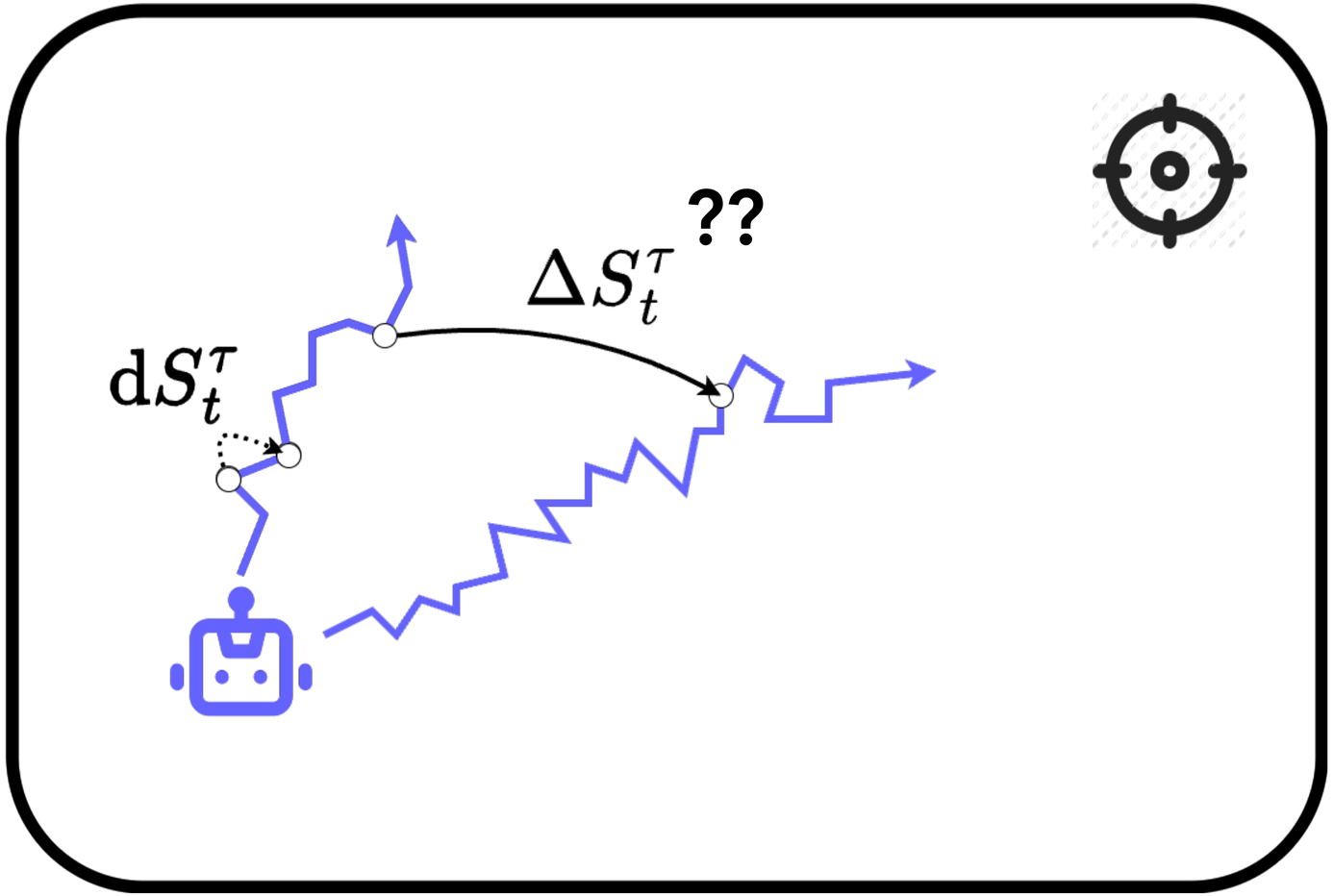
Learning Dynamics: Stochastic Transitions

Stochastic Control



Learning Dynamics: Stochastic Transitions

Continuous Time Actor Critic



$$J^\tau$$

Learning Dynamics: Stochastic Transitions

Theoretical Model: Key Challenges

- A sequence of discrete time MDPs that converges to a continuous time MDP
- Actor and critic learning simultaneously

Stochastic Transitions

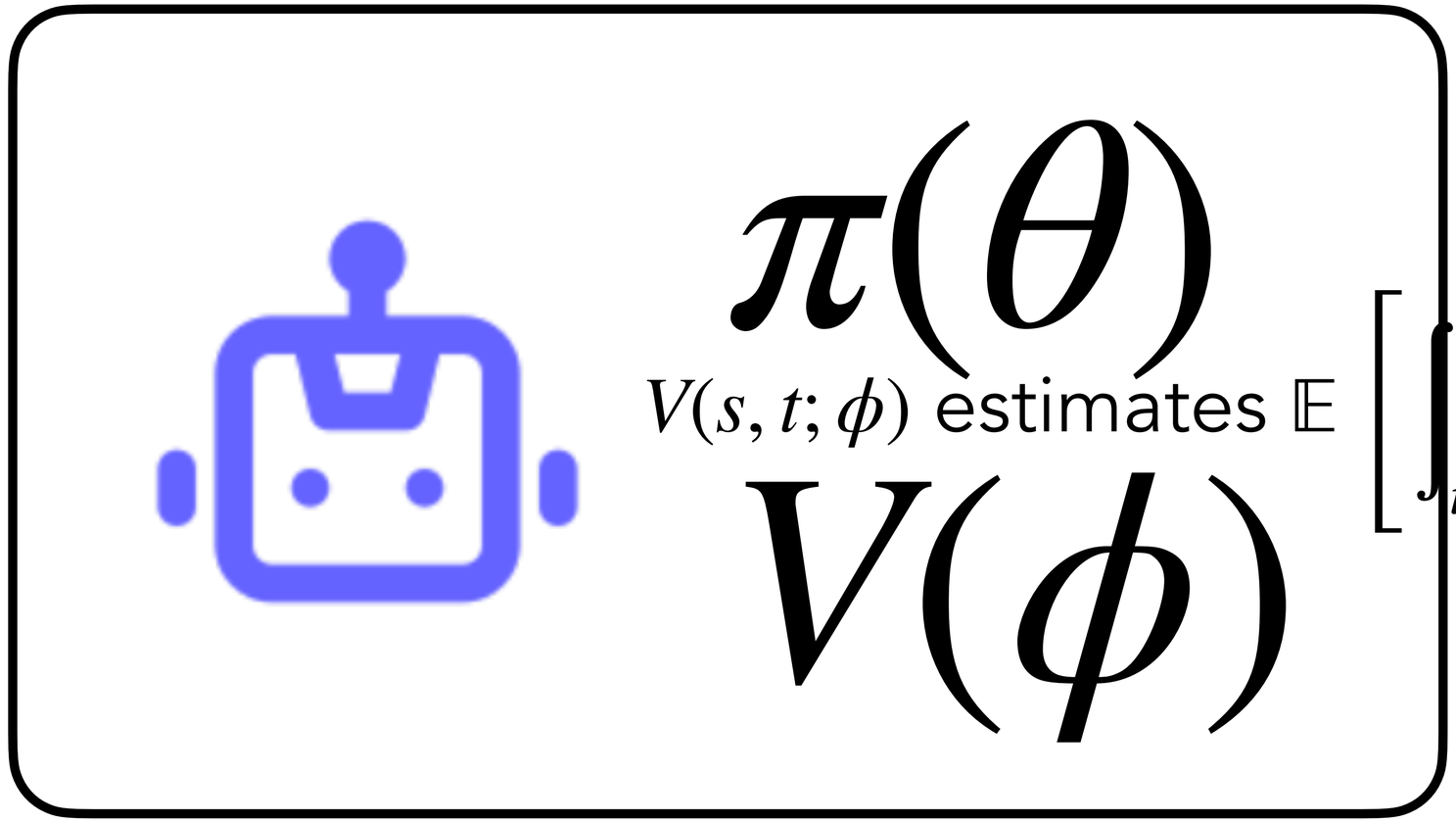
A Discrete Time Model that Converges to Continuous Time MDPs

Discrete Time MDP

Continuous Time MDP

Challenge: Exploration

Actor and Critic Learning Simultaneously



Actor and Critic Learning Simultaneously

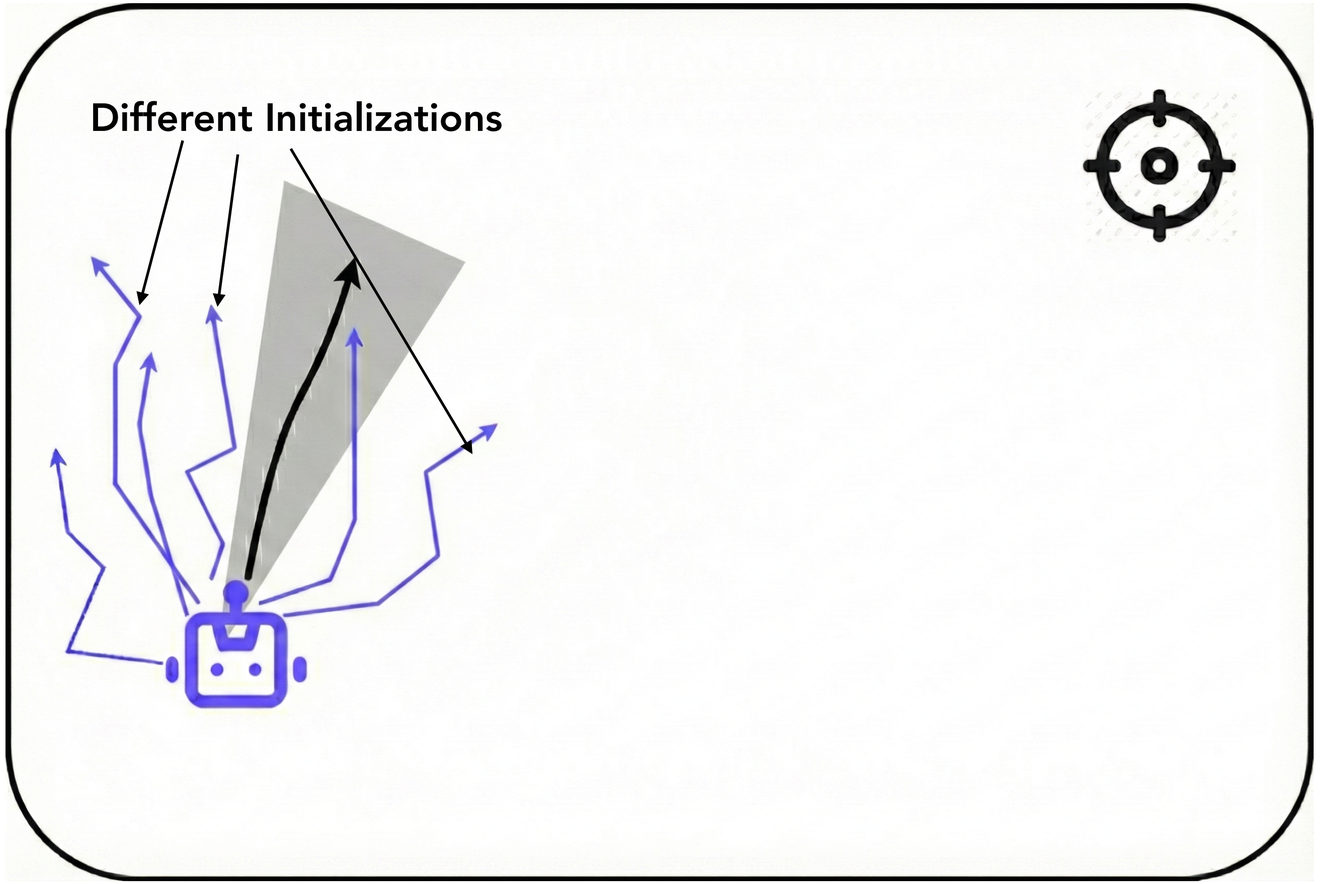
Online Episodic Actor-Critic (Jia and Zhou 2022)

Critic Semi-Gradient: $$\int_0^T e^{-\beta t} \frac{\partial V(S_t^\tau, t; \phi)}{\partial \phi} \delta_t \mathrm{d}t$$

Actor Semi-Gradient: $$\int_0^T e^{-\beta t} \frac{\partial \pi(S_t^\tau; \theta)}{\partial \theta} \delta_t \mathrm{d}t$$

$$\delta_t = \left[\partial_t V(S_t^\tau, t; \phi) + r(S_t^\tau) - \beta V(S_t^\tau, l; \phi) \right]$$

Actor and Critic Learning Simultaneously



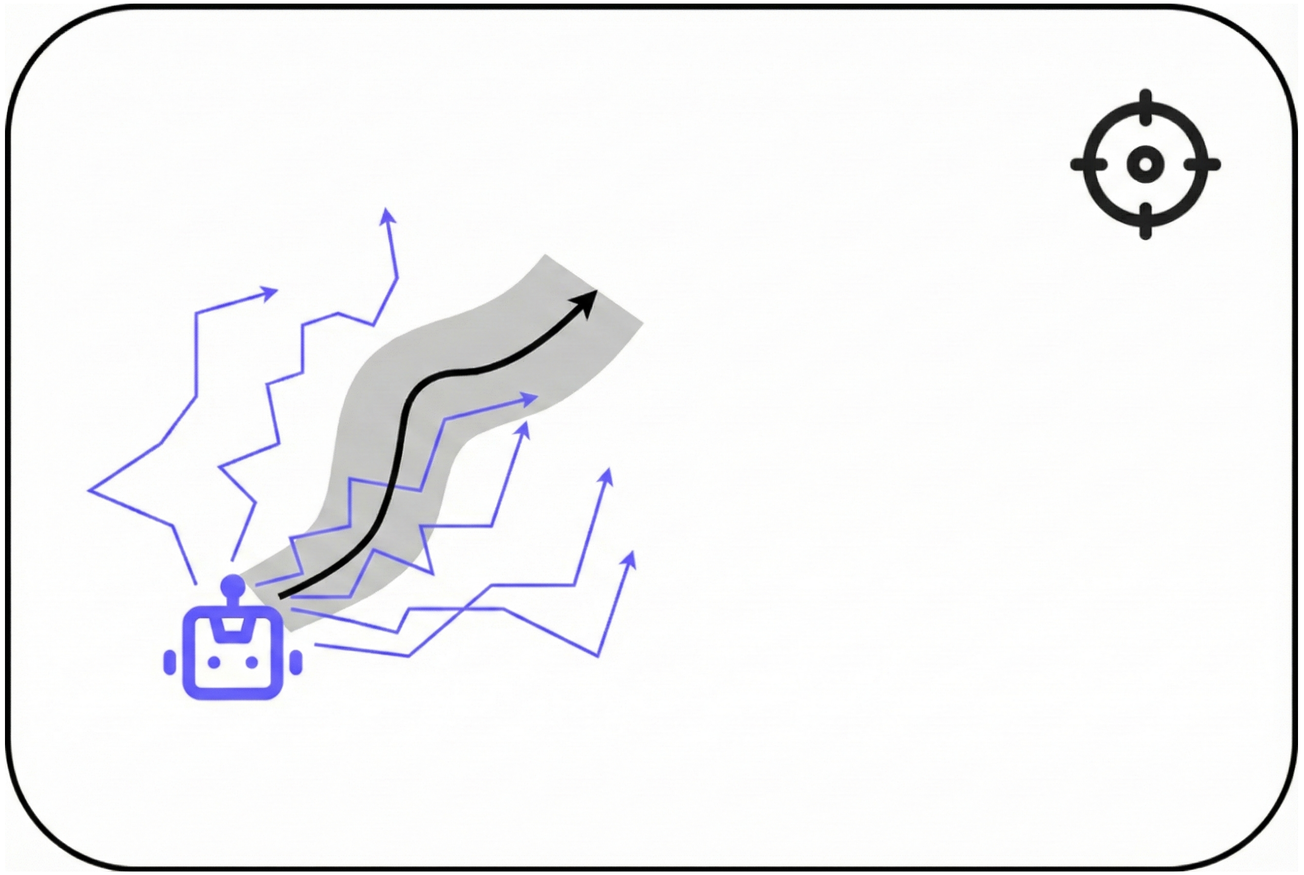
$$S_t^0, A_t^0, R_t^0$$

$$\downarrow$$

$$\phi_V^1 \leftarrow \phi_V^0 + \eta \mathcal{G}(\theta_\pi^0, \phi_V^0)$$

$$\theta_\pi^1 \leftarrow \theta_\pi^0 + \eta \mathbb{G}(\theta_\pi^0, \phi_V^0)$$

Actor and Critic Learning Simultaneously



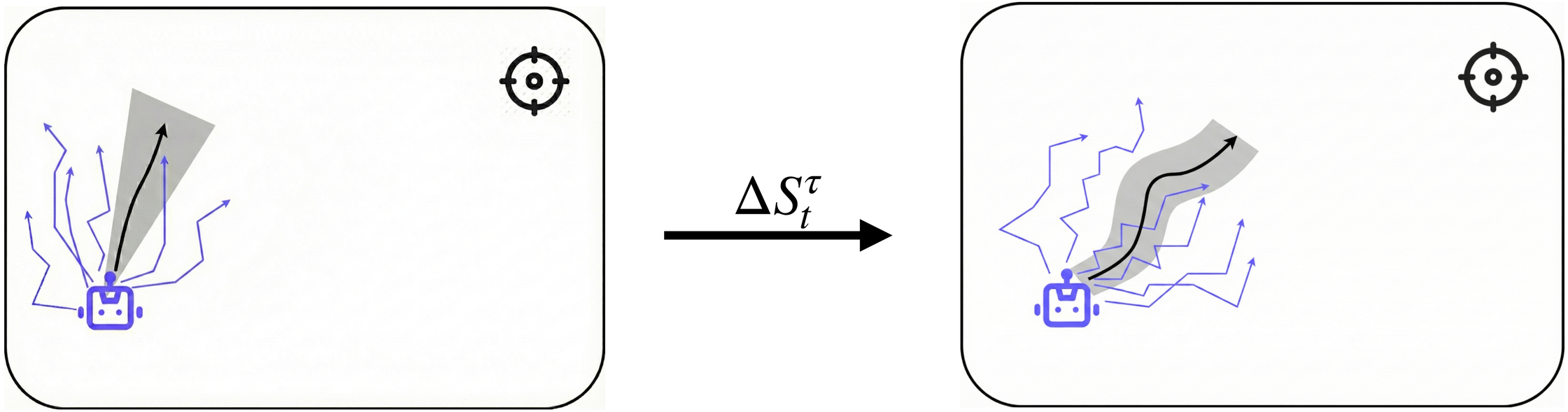
$$S_t^1, A_t^1, R_t^1$$

$$\downarrow$$

$$\phi_V^2 \leftarrow \phi_V^1 + \eta \mathcal{G}(\theta_\pi^1, \phi_V^1)$$

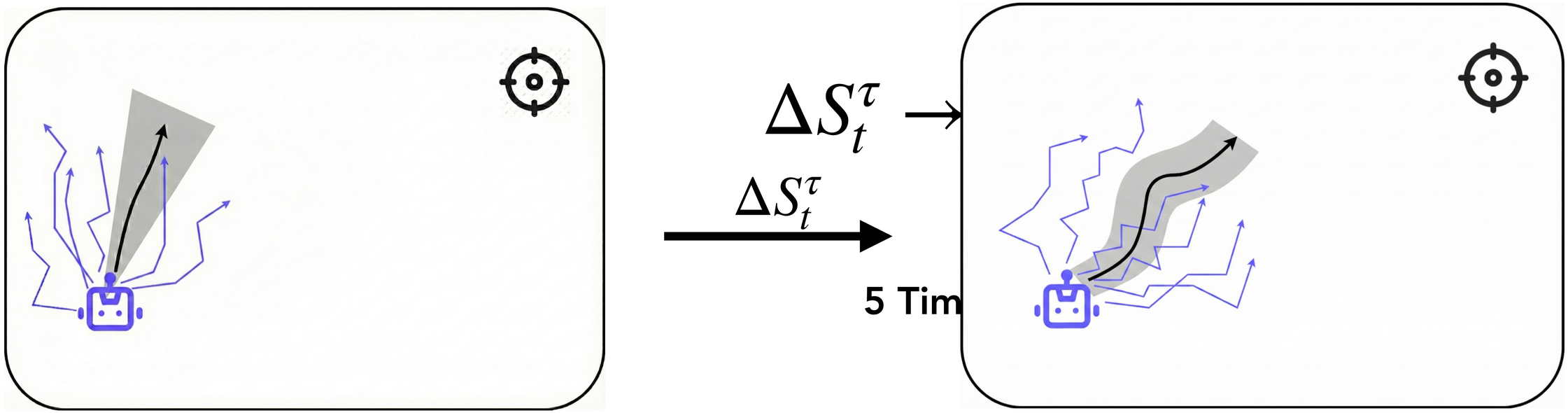
$$\theta_\pi^2 \leftarrow \theta_\pi^1 + \eta \mathbb{G}(\theta_\pi^1, \phi_V^1)$$

Learning Dynamics: Stochastic Transitions



Learning Dynamics: Stochastic Transitions

Main Result



ICLR 2026, Under review

Learning Dynamics: Stochastic Transitions

Main Result

$$A_t^\tau = \pi(S_t^\tau; \theta^\tau)$$

$$\bar{A}_t^\tau = \partial_s \pi(s; \theta^\tau)\,|_{s=S_t^\tau}$$

$$V_t^\tau = V(S_t^\tau, t; \phi^\tau)$$

$$\bar{V}_t^\tau = \partial_t V(S_t^\tau, t; \phi^\tau)$$

$$J^\tau$$

Empirical Theoretical Match

Main Result

Infinte Width Neural Network (Non-Linear)

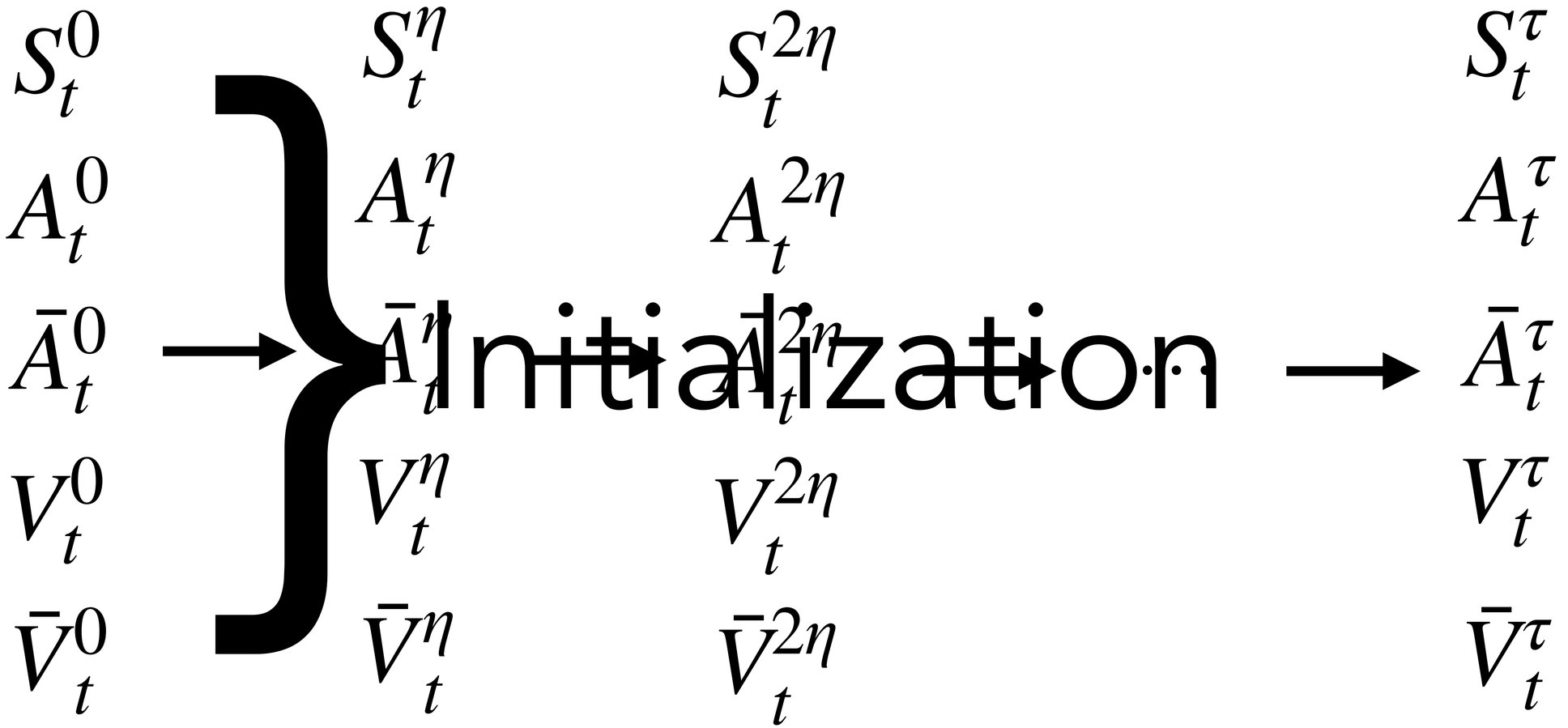
$$dS_t = (GS_t + H\,\boxed{\pi(S_t)})\mathrm{d}t + \sigma \mathrm{d}w_t$$

$$r(S_t) = -QS_t^2$$

(The theory is a lot more general for non-linear dynamics)

Empirical Theoretical Match

Main Result



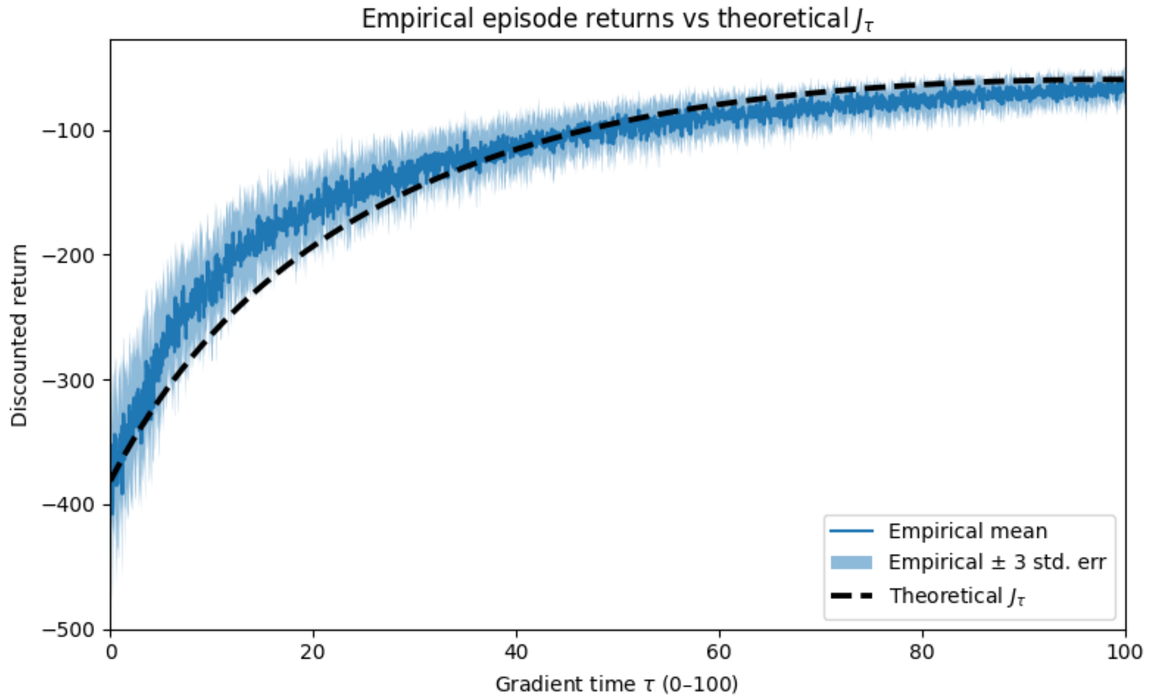
Empirical Theoretical Match

Main Result

$$J^0 \longrightarrow J^\eta \longrightarrow J^{2\eta} \longrightarrow \cdots \longrightarrow J^\tau$$

Empirical Theoretical Match

Main Result



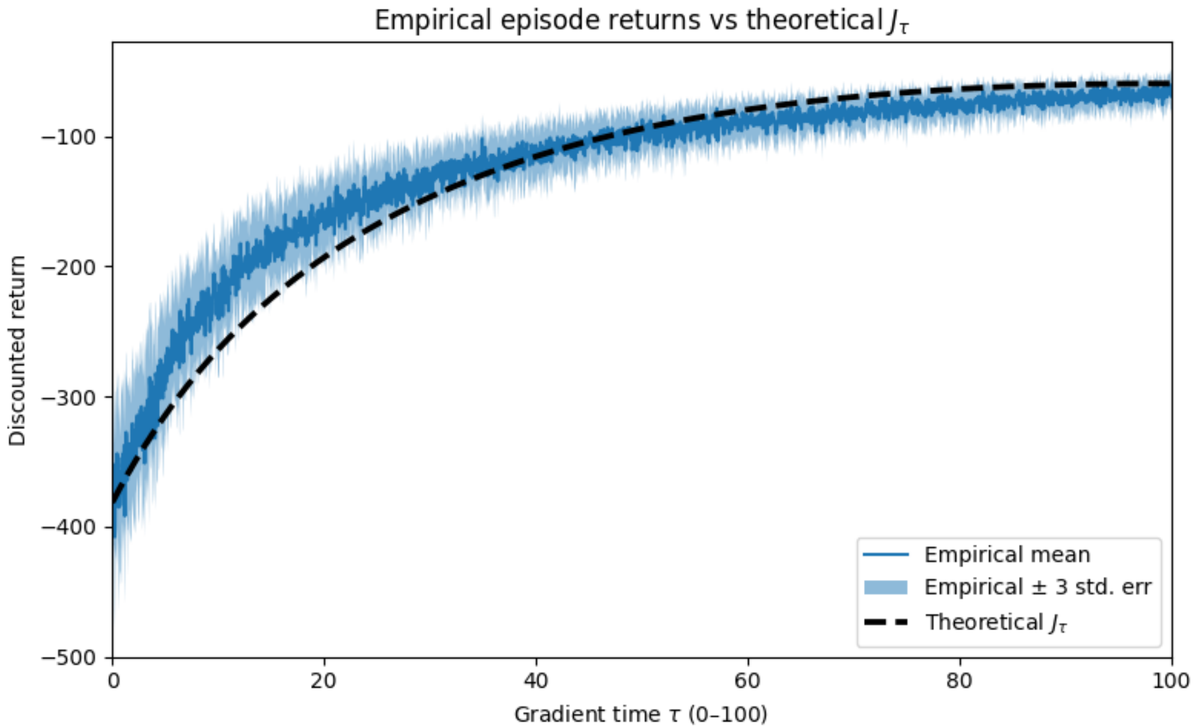
Empirical Theoretical Match

Main Result

- Same reward formulation: $r(s) = -500s^2$, Discount: $\beta = 0.1$
- Same deterministic drift: $g(s) + h(s)a$ and diffusion coefficient: σ
- Learning rate fixed: $1/\sqrt{n}$ where n is the width
- No "oracle" or features assumed all learning from scratch
- Empirical learning updates are as described: $\int_0^T e^{-\beta l} \frac{\partial v(\tilde{s}_l^{\pi^\theta}, l; \phi)}{\partial \phi} \delta_t \mathrm{d}l, \int_0^T e^{-\beta l} \frac{\partial \pi(\tilde{s}_l^{\pi^\theta}; \theta)}{\partial \theta} \delta_t \mathrm{d}l$
- Online, Episodic
- (No replay buffer, no target network, no LR fine-tuning, no gradient clipping)

Empirical Theoretical Match

Main Result



$$J^{\tau}$$