University of Glasgow

A WORLD TOP 100 UNIVERSITY

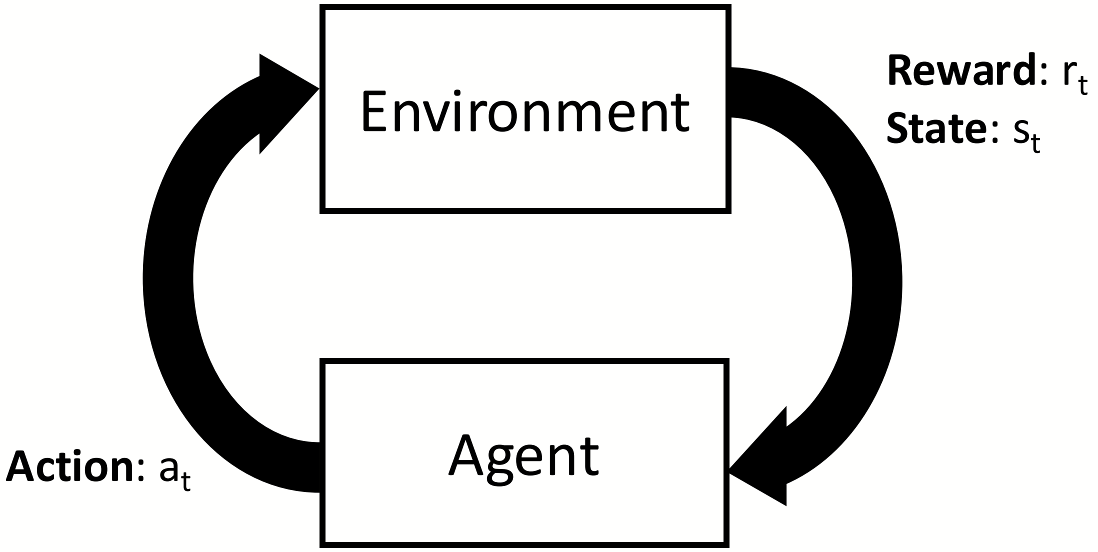
Enhancing Robustness in Deep Reinforcement Learning: A Lyapunov Exponent Approach

Rory Young, Nicolas Pugeault

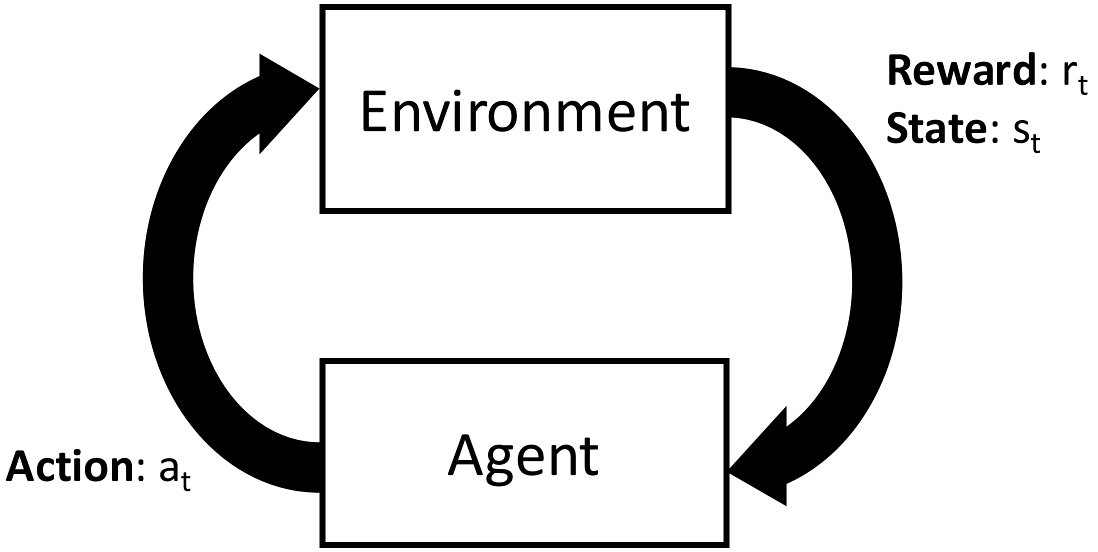
WORLD CHANGING GLASGOW

THE TIMES THE SUNDAY TIMES GOOD UNIVERSITY GUIDE 2024 SCOTTISH UNIVERSITY OF THE YEAR

Reinforcement Learning



Reinforcement Learning



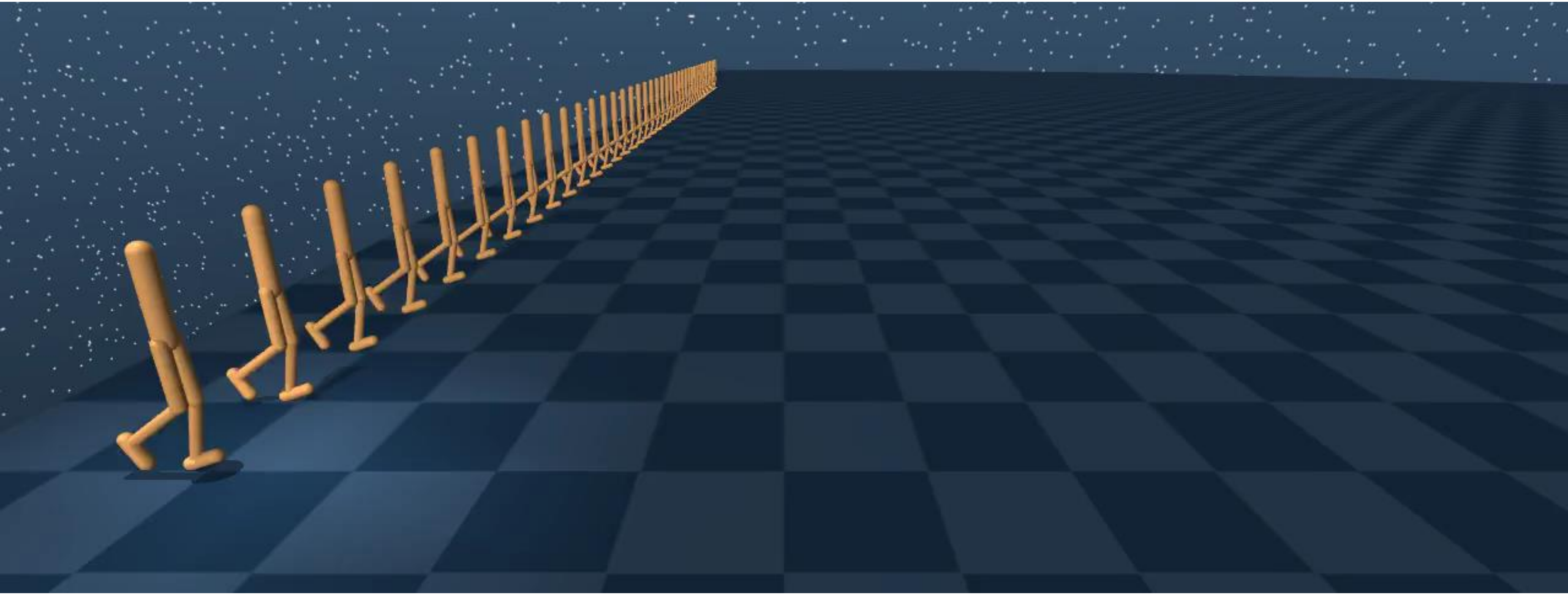
Produce a policy:

$$a_t = \pi_\theta(s_t)$$

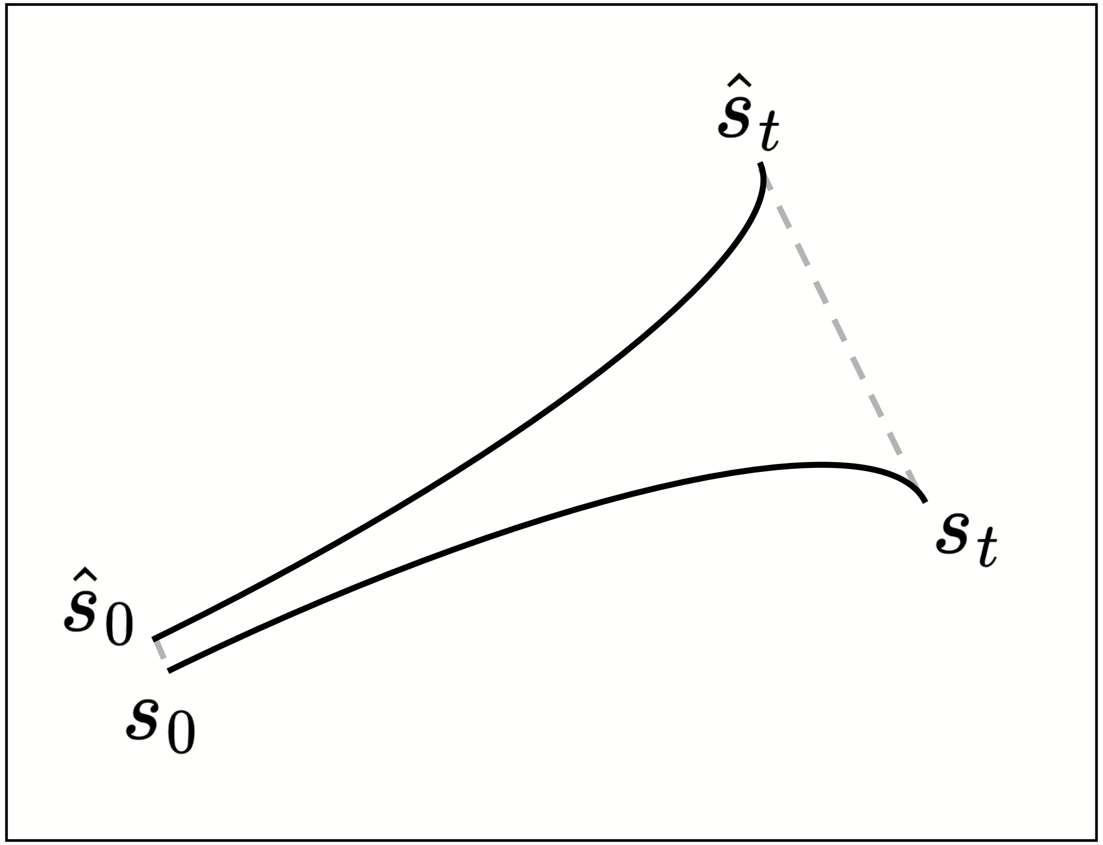
which maximises the expected sum of discounted rewards:

$$\mathbb{E}_{s_0 \sim \rho_0} \left[\sum_{t=0}^{\infty} \gamma^t \times r(s_t, a_t) \right]$$

State Dynamics



Maximal Lyapunov Exponent (λ_1)

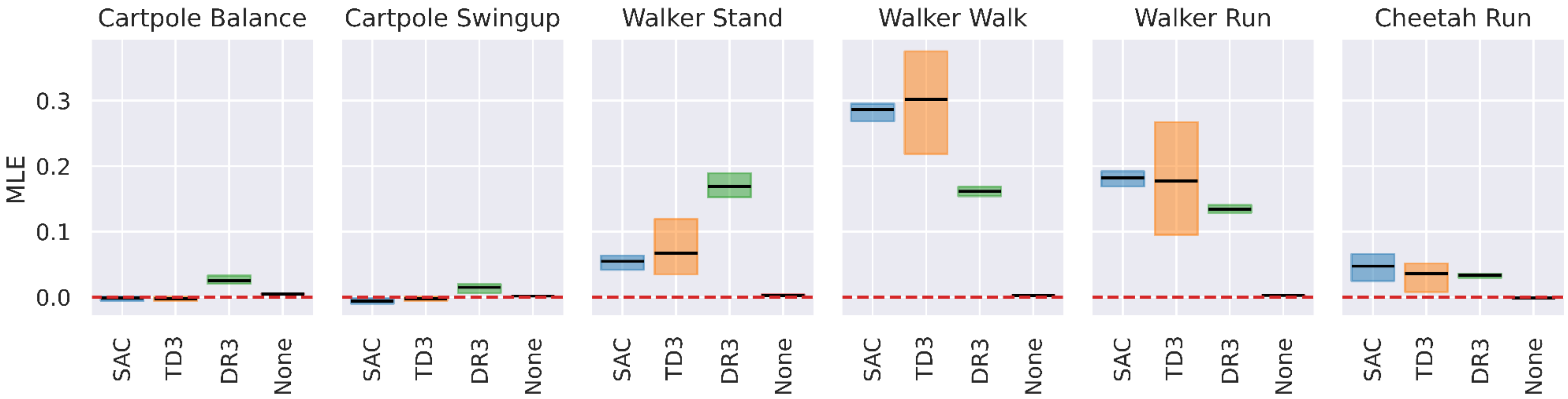


$$|s_t - \hat{s}_t| \approx |s_0 - \hat{s}_0| \times e^{\lambda_1 t}$$

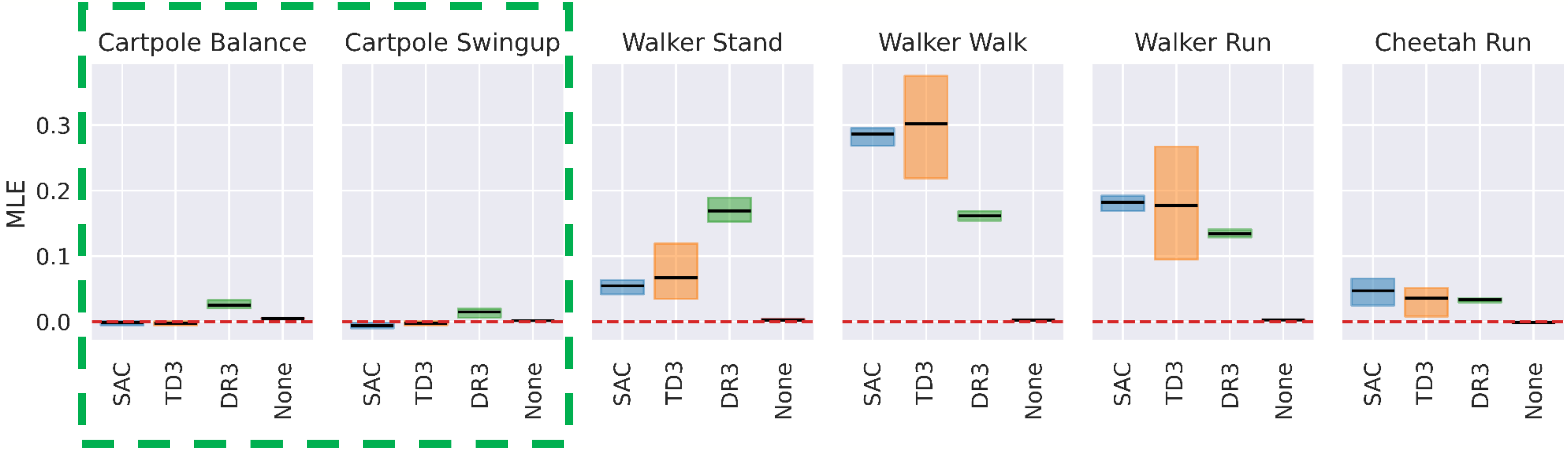
$$\lambda_1 = \lim_{t \to \infty} \lim_{\hat{s}_0 \to s_0} \frac{1}{t} \ln \left(\frac{|s_t - \hat{s}_t|}{|s_0 - \hat{s}_0|} \right)$$

λ_1	Dynamics
-	Stable
+	Chaotic

System Stability



System Stability

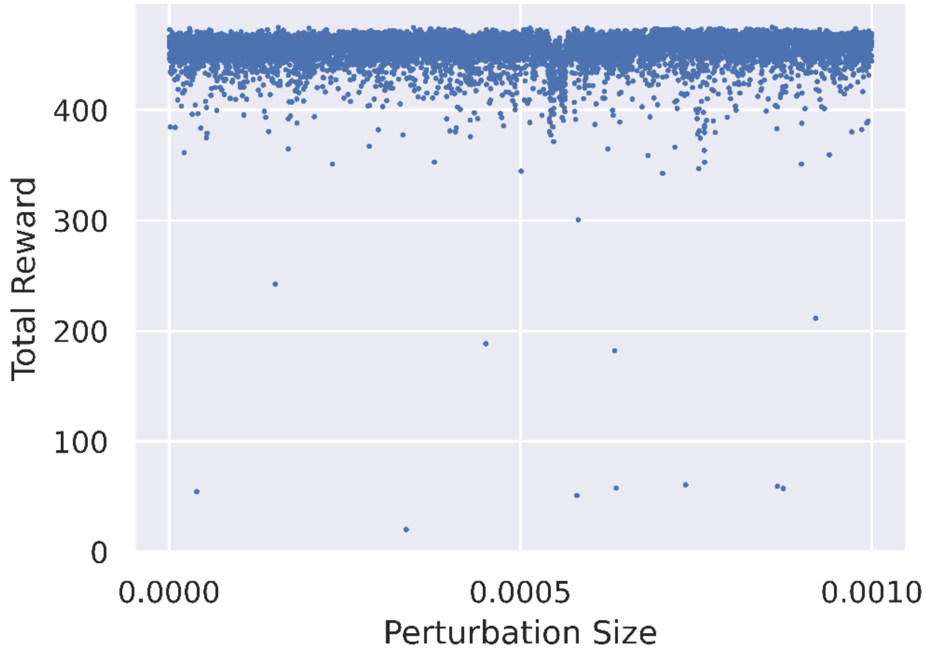


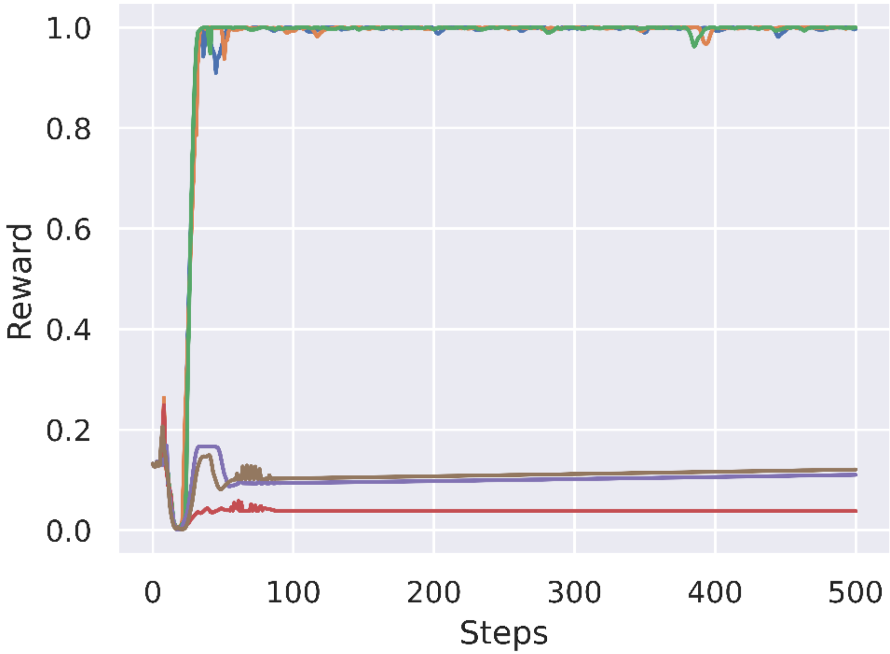
System Stability

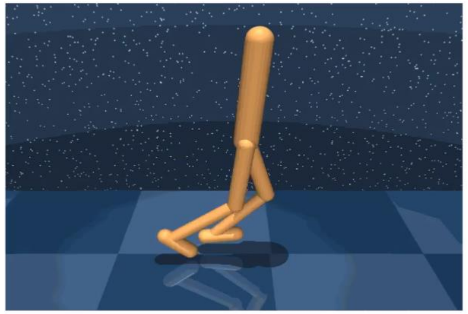
System Stability

It is impossible to accurately predict the long-term state dynamics given an approximate observation.

Reward stability

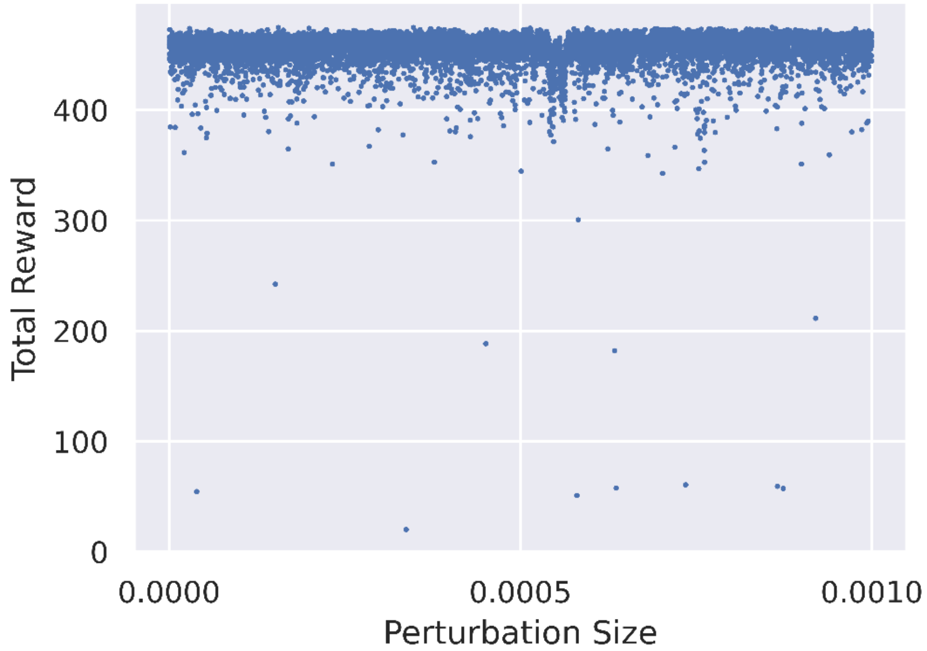


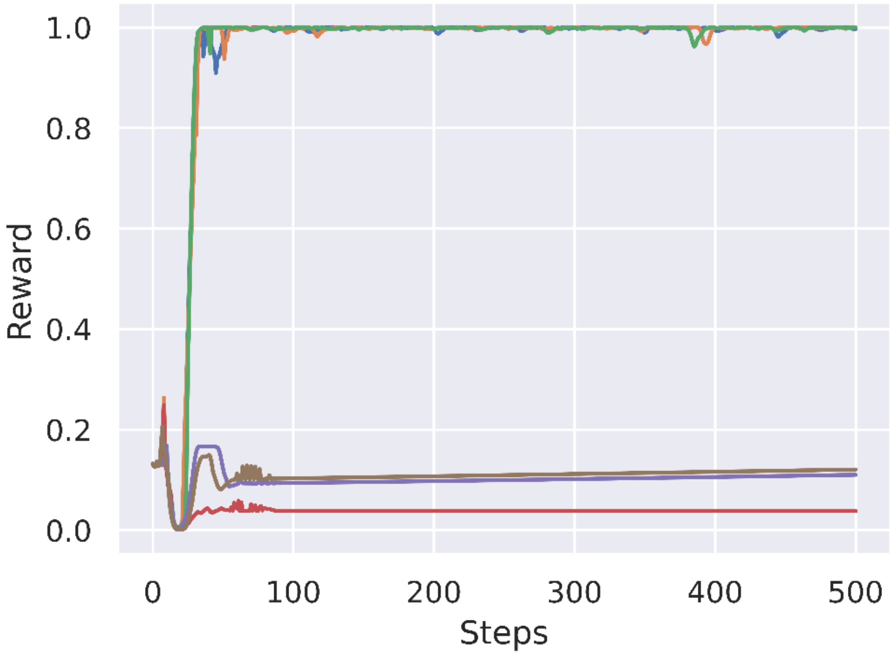


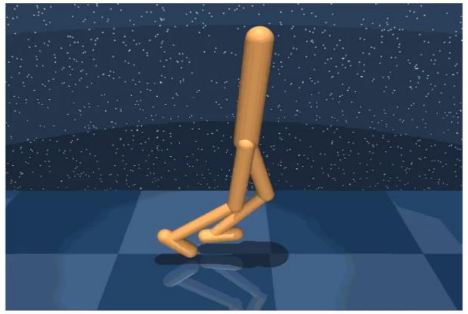


Task: Walker Walk
Agent: SAC

Reward stability



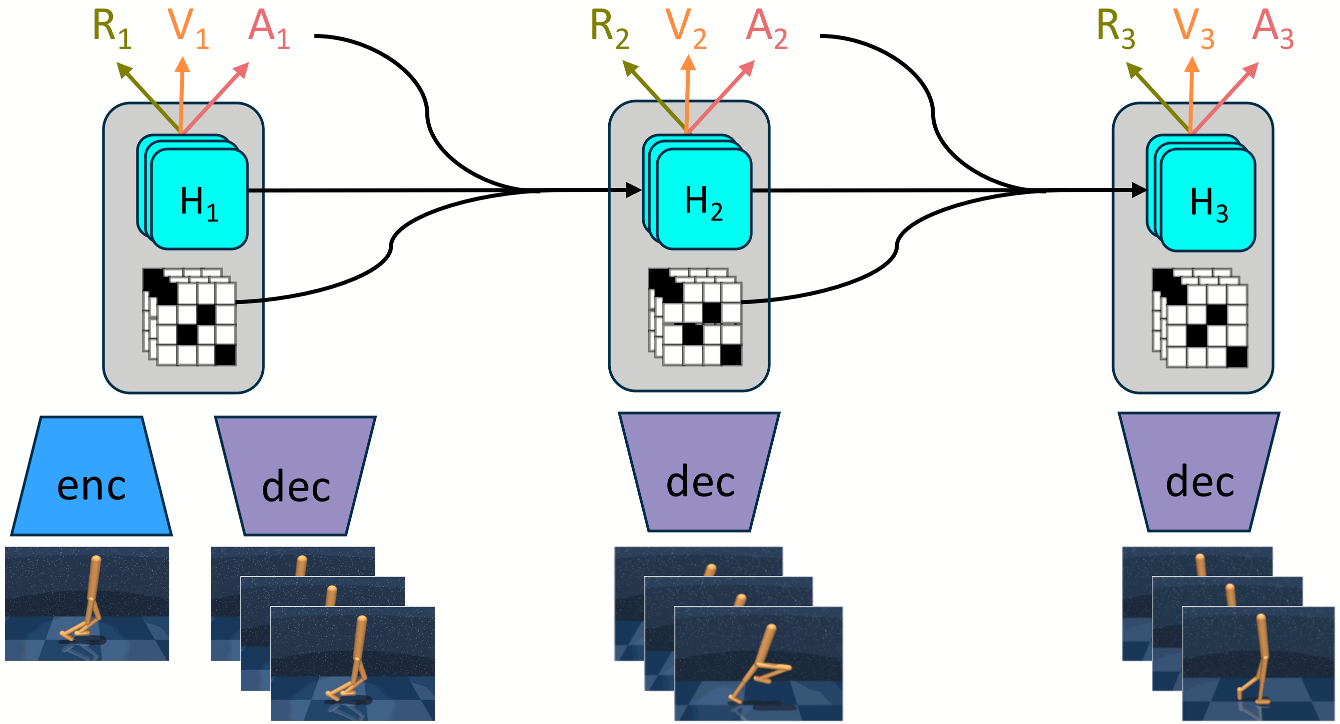




Task: Walker Walk
Agent: SAC

Adversarial methods can leverage this instability to significantly decrease performance with a single attack.

Maximal Lyapunov Exponent Regularisation

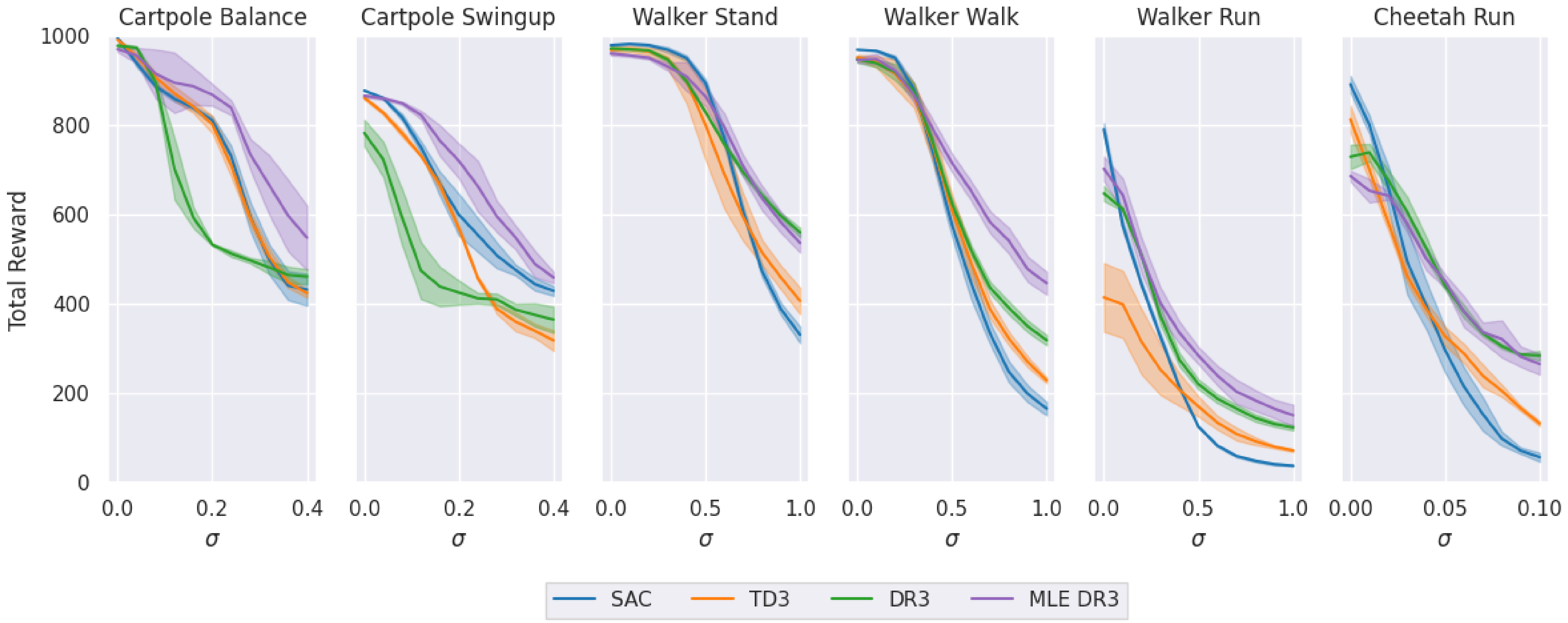


$$\mathcal{L}(\theta) \doteq \underbrace{-\sum_{t=1}^{T}\left(\mathrm{sg}\left(\frac{R_t^{\lambda} - v_\phi(s_t)}{\max(1, S)}\right)\log \pi_\theta(a_t|s_t) + \eta \mathrm{H}\left[\pi_\theta(a_t|s_t)\right]\right)}_{\text{Dreamer V3}} + \underbrace{\sum_{t=1}^{T}\left(\underset{L}{\mathrm{Var}}(S_t) + \underset{L}{\mathrm{Var}}(H_t)\right)}_{\text{MLE Regularisation}}$$

Results

Environment	Reward DR3	Reward MLE DR3	MLE DR3	MLE MLE DR3
Pointmass	869.5	**880.5**	0.0326	**-0.0275**
Cartpole Balance	**978.6**	970.5	0.0249	**0.0231**
Cartpole Swingup	781.4	**866.4**	**0.0149**	0.0235
Walker Stand	**973.0**	961.6	0.1688	**0.0654**
Walker Walk	948.6	**950.7**	0.1614	**0.1405**
Walker Run	646.3	**698.4**	0.1345	**0.1106**
Cheetah Run	**737.7**	675.2	0.0337	**0.0283**

Results


Cartpole Balance
Cartpole Swingup
Walker Stand
Walker Walk
Walker Run
Cheetah Run
Total Reward
1000
800
600
400
200
0
0.0
0.2
0.4
0.5
1.0
0.00
0.05
0.10
σ
SAC
TD3
DR3
MLE DR3

Summary

1. Deep reinforcement learning policies can produce chaotic state and reward trajectories in continuous control tasks.

Summary

1. Deep reinforcement learning policies can produce chaotic state and reward trajectories in continuous control tasks.

2. Chaotic systems are highly sensitive to initial conditions, so it is impossible to accurately predict the long-term state dynamics given a noisy observation.

Summary

1. Deep reinforcement learning policies can produce chaotic state and reward trajectories in continuous control tasks.

2. Chaotic systems are highly sensitive to initial conditions, so it is impossible to accurately predict the long-term state dynamics given a noisy observation.

3. This instability can substantially decrease overall performance with a single state perturbation.

Summary

1. Deep reinforcement learning policies can produce chaotic state and reward trajectories in continuous control tasks.

2. Chaotic systems are highly sensitive to initial conditions, so it is impossible to accurately predict the long-term state dynamics given a noisy observation.

3. This instability can substantially decrease overall performance with a single state perturbation.

4. To improve the stability of the control interaction, we propose Maximal Lyapunov Exponent regularisation for Dreamer V3.

R.Young.4@research.gla.ac.uk