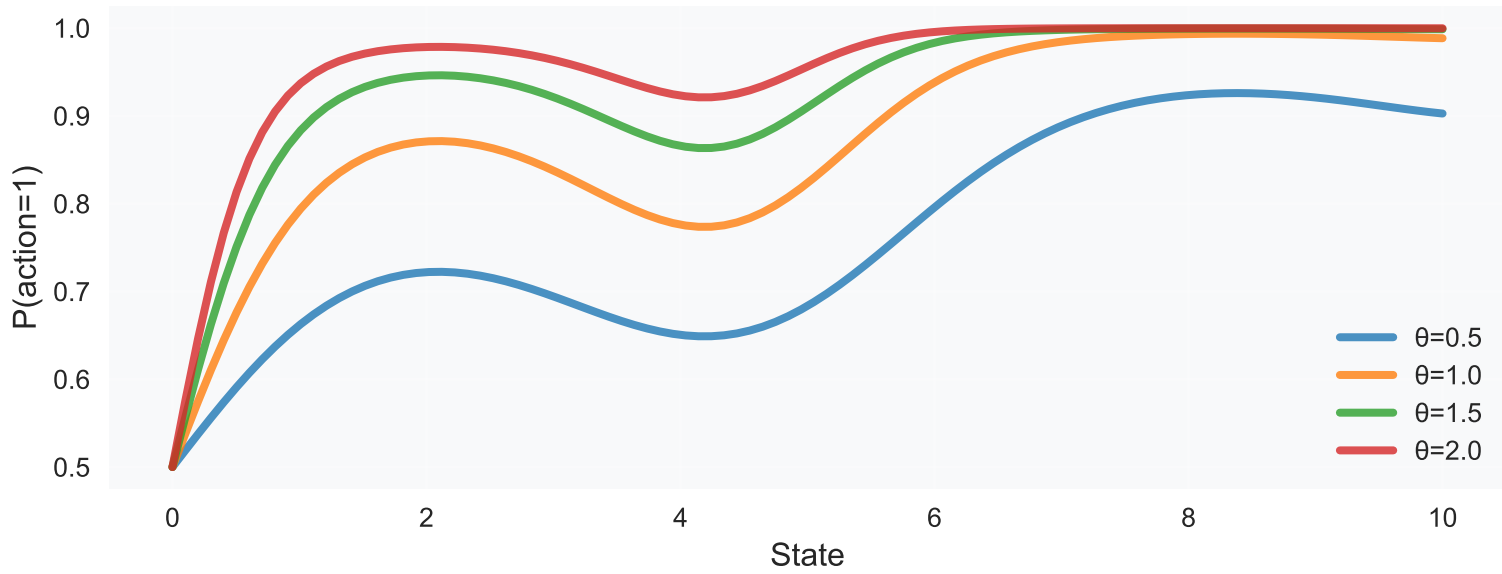
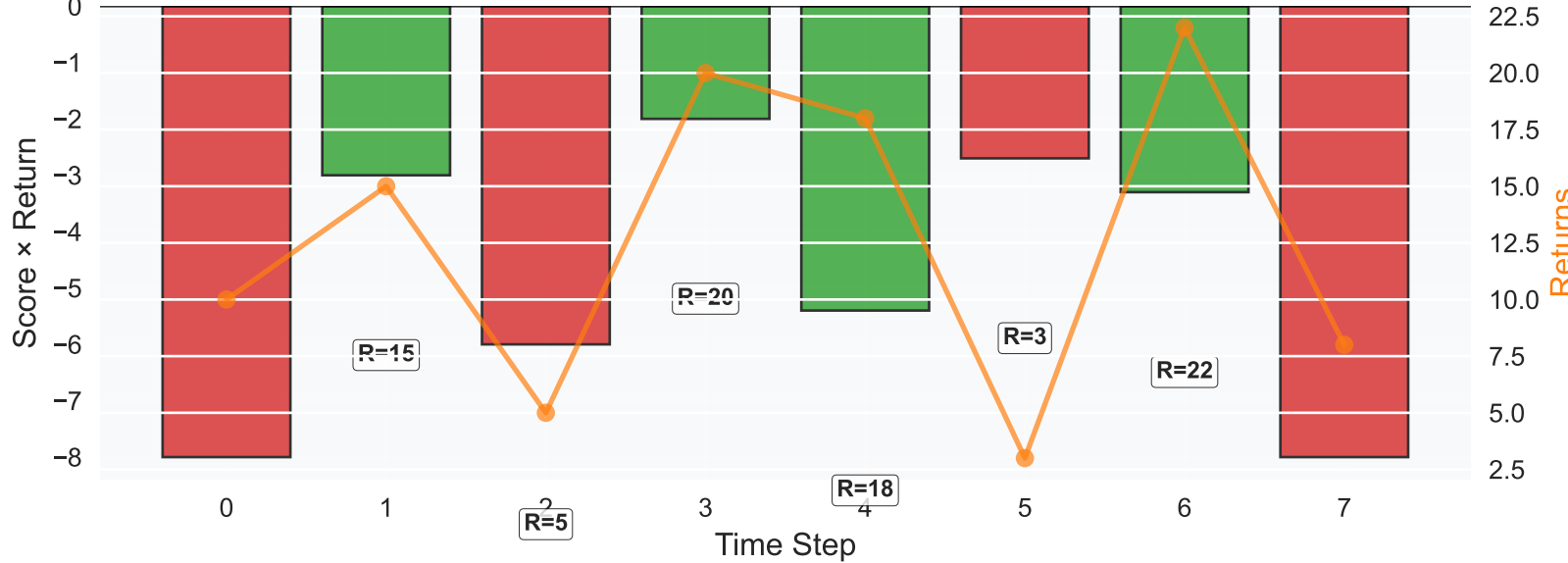


# Comprehensive Policy Gradient Intuition

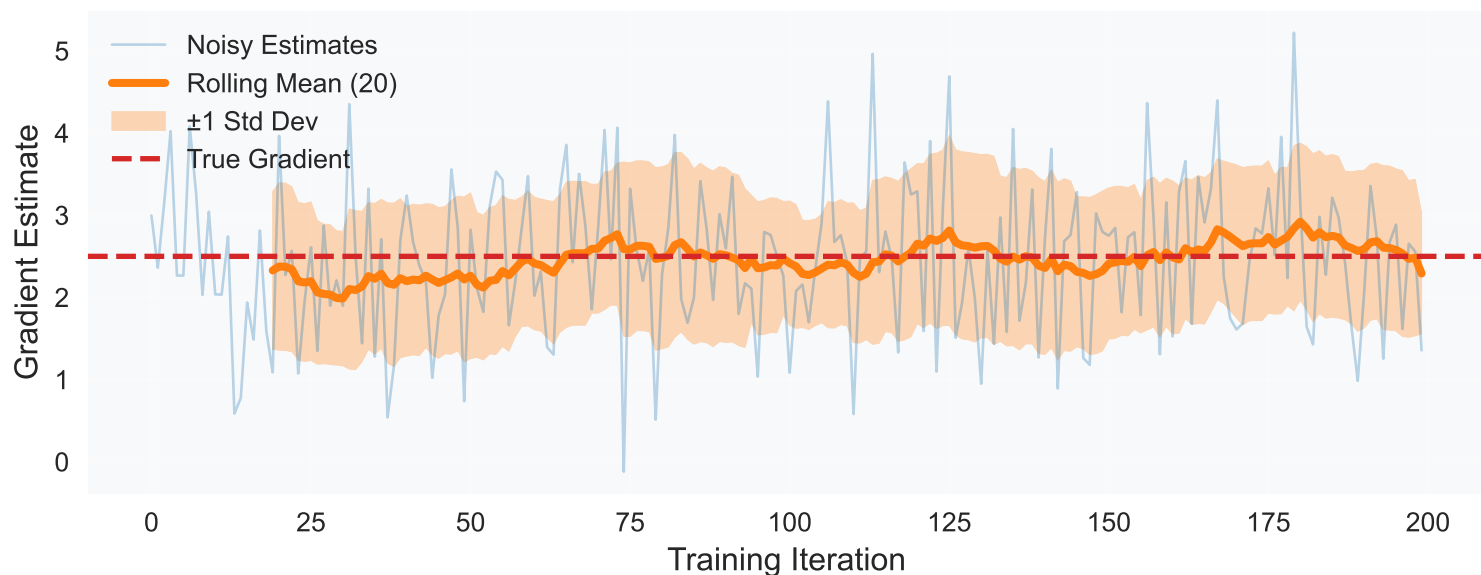
## Policy Parameterization: $\pi(a=1|s; \theta)$



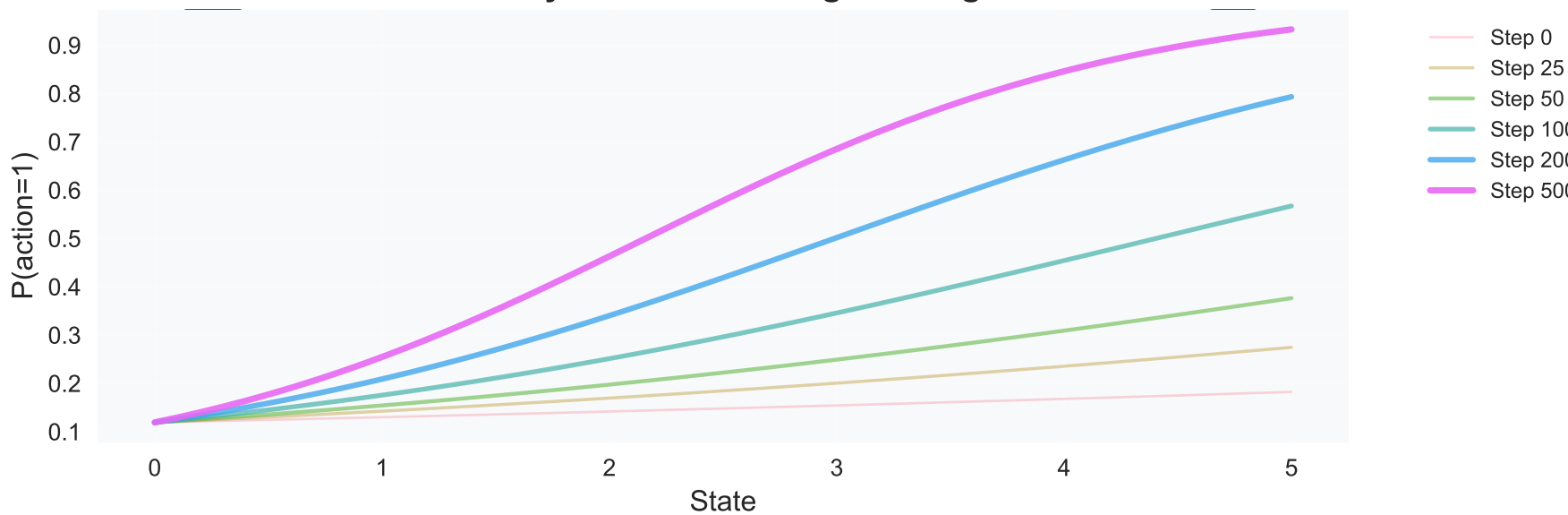
## Score Function: $\square \log \pi(a|s) \times \text{Return}$



## Gradient Estimation with Confidence Intervals



## Policy Evolution During Training



## Mathematical Foundation

### Policy Gradient Theorem Derivation:

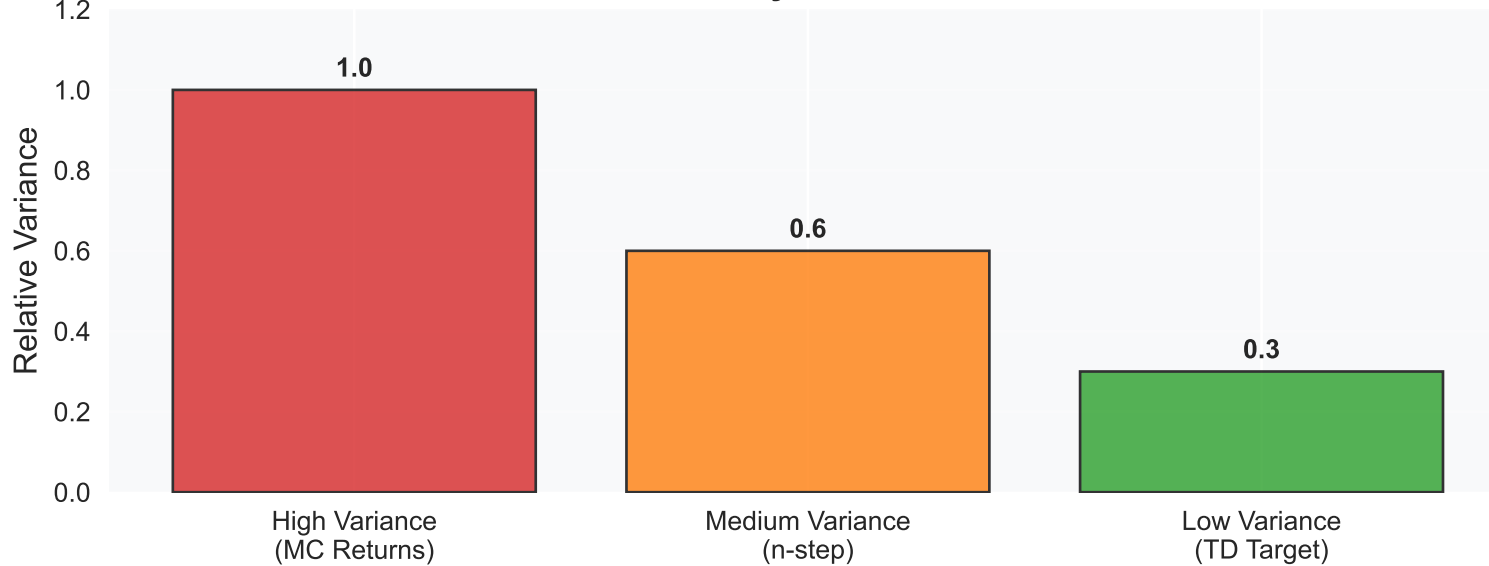
$$J(\theta) = E[\sum_{t=0}^{T-1} \square_{\theta} \log \pi_{\theta}(a_t|s_t) \cdot G_t]$$

### Where:

- $J(\theta)$ : Expected return
- $\square_{\theta} \log \pi_{\theta}$ : Score function
- $G_t$ : Return from time  $t$
- $E[\cdot]$ : Expectation over trajectories

$$\nabla_{\theta} J(\theta) = E_{\pi_{\theta}} \left[ \sum_{t=0}^{T-1} \nabla_{\theta} \log \pi_{\theta}(a_t|s_t) \cdot G_t \right]$$

## Variance in Policy Gradient Estimates



## Policy Gradient Algorithms Comparison

Algorithm	Variance	Bias	Sample Eff.	Stability
REINFORCE	High	Low	Low	Low
Actor-Critic	Medium	Medium	Medium	Medium
PPO	Low	Low	High	High
TRPO	Low	Low	High	Very High

## Policy Parameter Landscape

