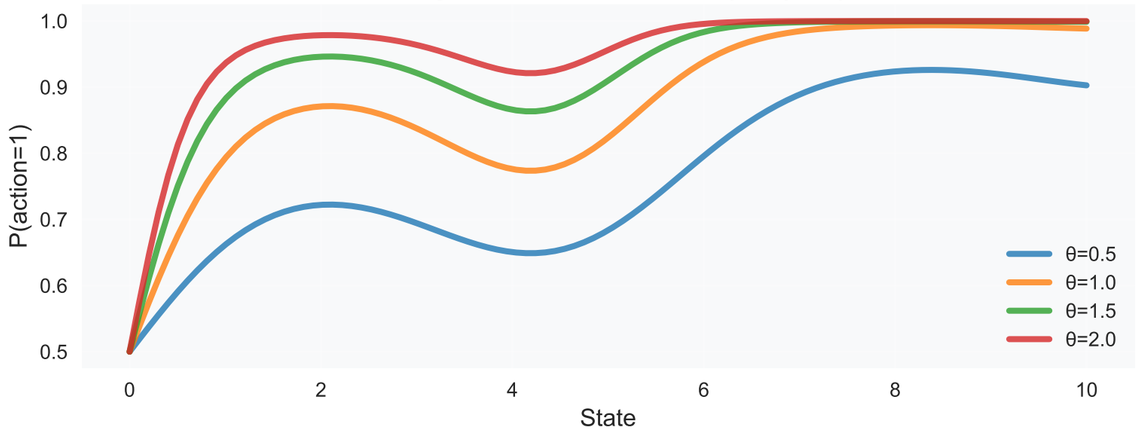
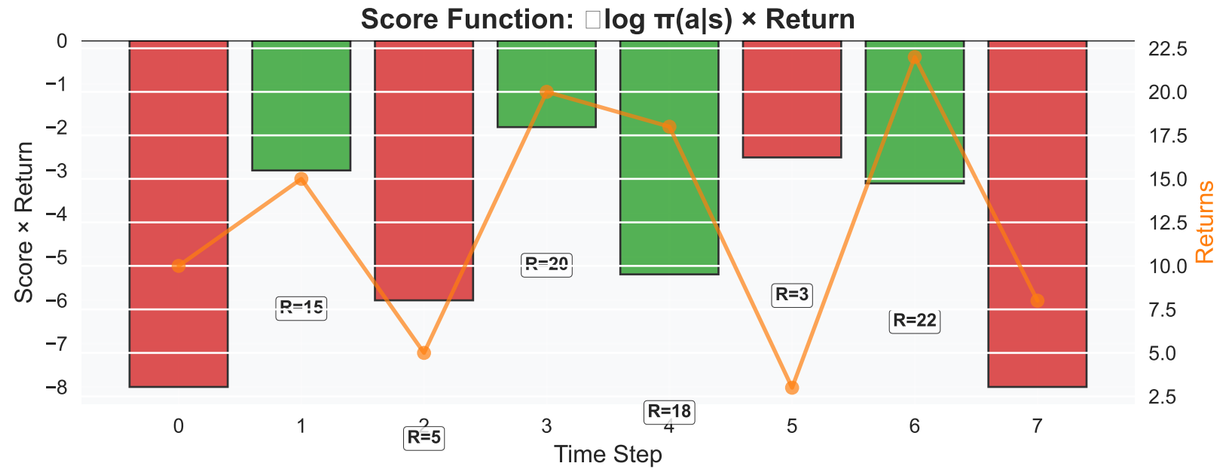
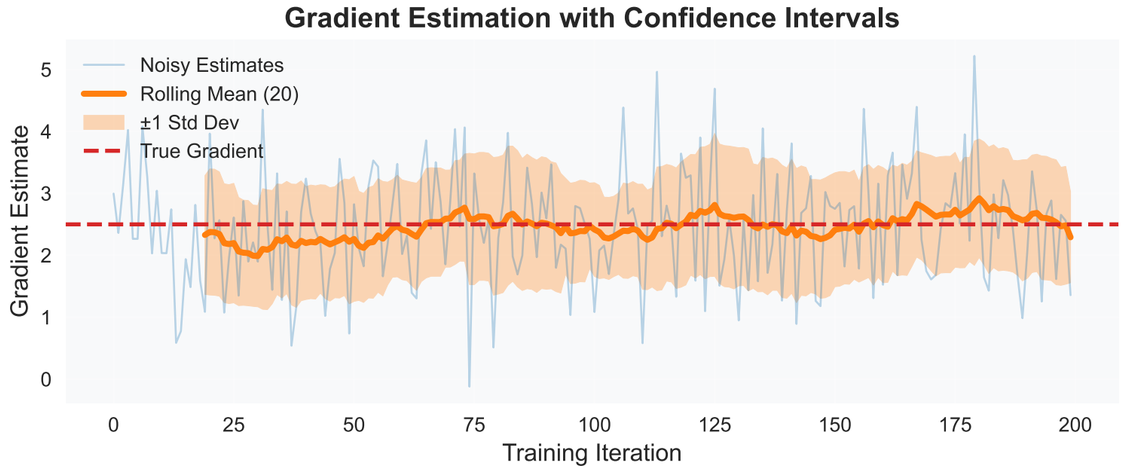
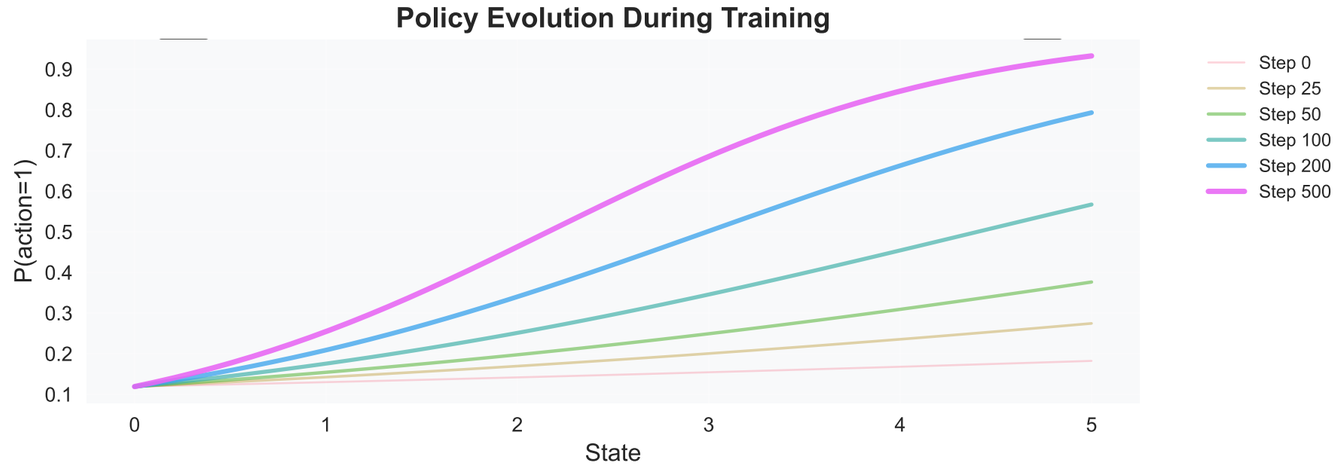
Comprehensive Policy Gradient Intuition









Mathematical Foundation

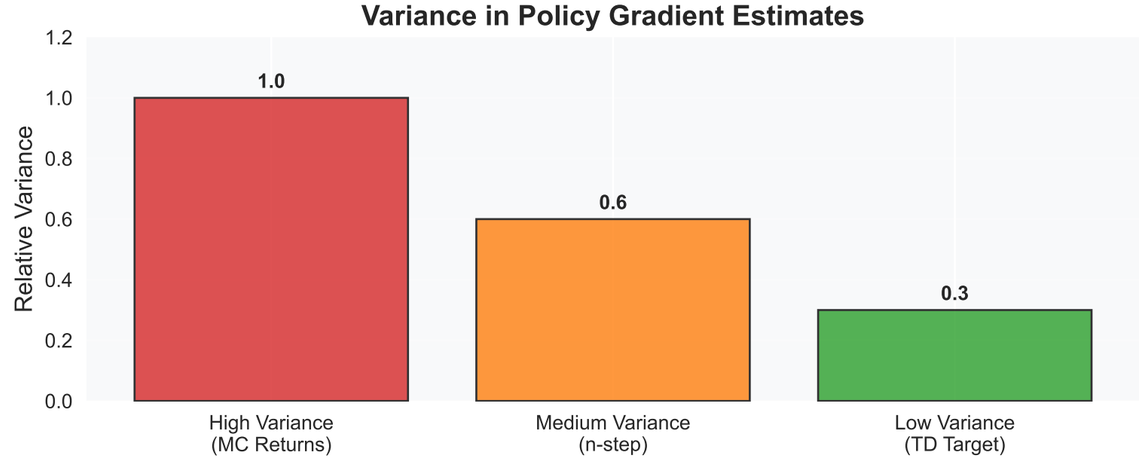
Policy Gradient Theorem Derivation:

$J(\theta) = E[\sum \nabla_\theta \log \pi_\theta(a_t|s_t) \cdot G_t]$

Where:

- J(θ): Expected return
- ∇_θ log π_θ: Score function
- G_t: Return from time t
- E[·]: Expectation over trajectories

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



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