How to Improve Policy?

- For deterministic policy and deterministic environment:
  - Let $\rho(\theta)$ be the policy value – expected reward-to-go when π_θ is executed.
  - If $\rho(\theta)$ is differentiable: Take a step in the direction of the policy gradient vector $\nabla_\theta \rho(\theta)$ – Look for the local optimum
- For stochastic environment and/or policy $\pi_\theta(s, a)$:
  - Obtain an unbiased estimate of the gradient at θ, $\nabla_\theta \rho(\theta)$ directly from results of trials executed at θ

Use gradient ascent or stochastic gradient ascent