

Policy Gradient Methods

Chris Amato

Northeastern University

with some slides from Rob Platt and U Alberta

Announcements

- DQN assignment due Wednesday (midnight)
- Policy gradient assignment out tonight
- Lecture schedule

Deep policy gradient (A3C)



Approaches to control

1. Previous approach: *Action-value methods*:
 - learn the value of each action
 - pick the max (usually)
2. New approach: *Policy-gradient methods*:
 - learn the parameters of a stochastic policy (or a continuous-action policy)
 - update by gradient ascent in performance
 - includes *actor-critic methods*, which learn *both* value and policy parameters

Why approximate policies rather than values?

- In many problems, the policy is simpler to approximate than the value function
- In some problems, the optimal policy is stochastic
 - e.g., bluffing, POMDPs
- To enable smoother change in policies
- To avoid a search on every step (the max)

Policy Approximation

- Policy = a function from state to action
 - How does the agent select actions?
 - In such a way that it can be affected by learning?
 - In such a way as to assure exploration?
- Approximation: there are too many states and/or actions to represent all policies
 - To handle large/continuous action spaces
- Any parameterization is ok as long as it is differentiable (we are going to use gradient methods)

Remember: Gradient-bandit algorithm

- Store action preferences $H_t(a)$ rather than action-value estimates $Q_t(a)$
- Instead of ε -greedy, pick actions by an exponential soft-max:

$$\Pr\{A_t = a\} \doteq \frac{e^{H_t(a)}}{\sum_{b=1}^k e^{H_t(b)}} \doteq \pi_t(a)$$

- Also store the sample average of rewards as \bar{R}_t
- Then update:

$$H_{t+1}(a) = H_t(a) + \alpha(R_t - \bar{R}_t) \left(\mathbf{1}_{a=A_t} - \pi_t(a) \right)$$

$$\frac{\partial \pi_t(A_t)}{\partial H_t(a)} / \pi_t(A_t)$$

$$H_{t+1}(a) = H_t(a) + \alpha(R_t - \bar{R}_t) \left(\mathbf{1}_{a=A_t} - \pi_t(a) \right)$$

1 or 0, depending on whether
the predicate (subscript) is true

Linear-exponential policies (discrete actions)

- E.g., can use linear function approximation for “preference” h for action a in state s . Linear in θ and a state-action feature vector

$$h(s, a, \theta) = \theta^\top \mathbf{x}(s, a)$$

- The probability of each action is exponential in its preference

$$\pi(a|s, \theta) = \frac{\exp(h(s, a, \theta))}{\sum_b \exp(h(s, b, \theta))}$$

- Corresponding *eligibility function*:

$$\nabla_\theta \ln \pi(a|s, \theta) = \mathbf{x}(s, a) - \sum_b \pi(b|s, \theta) \mathbf{x}(s, b)$$

Linear-exponential policies (discrete actions)

- E.g., can use “preference”

These could also be Q-values, but need to ‘harden’ distribution over time to converge to deterministic policy or in θ and a state-action feature vector

$$h(s, a, \theta) = \theta^\top \mathbf{x}(s, a)$$

- The probability of each action is exponential in its preference

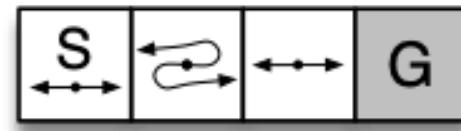
$$\pi(a|s, \theta) = \frac{\exp(h(s, a, \theta))}{\sum_b \exp(h(s, b, \theta))}$$

- Corresponding *eligibility function*:

$$\nabla_\theta \ln \pi(a|s, \theta) = \mathbf{x}(s, a) - \sum_b \pi(b|s, \theta) \mathbf{x}(s, b)$$

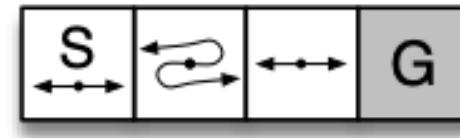
Example of the need for stochastic policies

- Reward: -1 in all states except terminal
- Two actions: left and right
 - Actions deterministically go to corresponding state
 - Action results reversed in middle state
- One terminal state (G)
- Function approximation with:
 - $x(s,\text{right})=[1,0]$
 - $x(s,\text{left})=[0,1]$



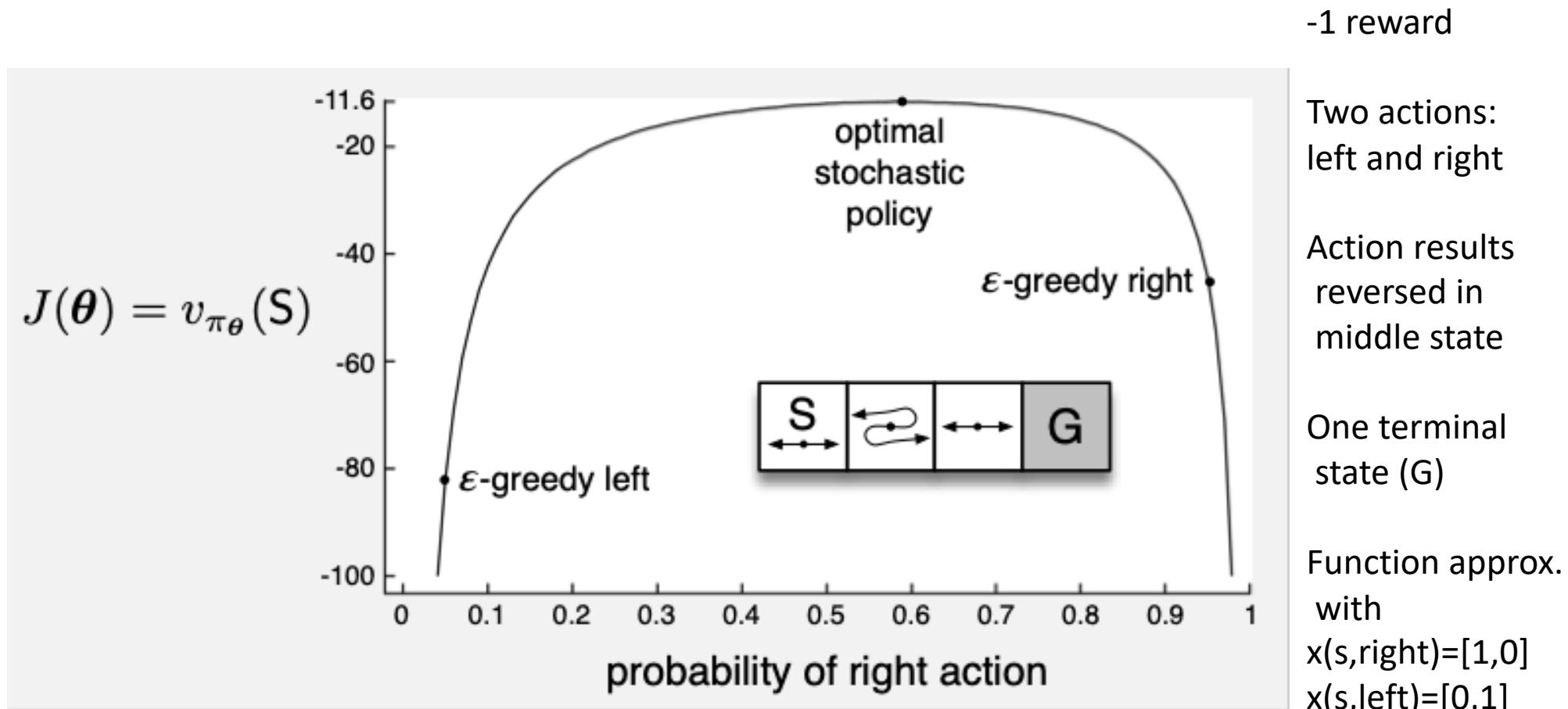
Think-pair-share

- Reward: -1 in all states except terminal
- Two actions: left and right
 - Actions deterministically go to corresponding state
 - Action results reversed in middle state
- One terminal state (G)
- Function approximation with:
 - $x(s,\text{right})=[1,0]$
 - $x(s,\text{left})=[0,1]$



What would an action-value method do?
What is the best deterministic policy?
What is the best stochastic policy?

Example of the need for stochastic policies



Neither ϵ -greedy policy is good

Optimal performance requires acting with particular probabilities

Policy gradient setup

Given a policy parameterization:

$$\pi(a|s, \theta)$$

And objective:

$$J(\theta) \doteq v_{\pi_\theta}(s_0) \text{ (or average reward)}$$

Approximate stochastic gradient ascent:

$$\theta_{t+1} \doteq \theta_t + \alpha \widehat{\nabla J(\theta_t)}$$

Typically, based on the Policy-Gradient Theorem:

$$\nabla J(\theta) \propto \sum_s \mu(s) \sum_a q_\pi(s, a) \nabla_\theta \pi(a|s, \theta)$$


on-policy distribution

Policy gradient setup

Given a policy parameterization:

$$\pi(a|s, \theta)$$

Stochastic policy that is parameterized by θ

And objective:

$$J(\theta) \doteq v_{\pi_\theta}(s_0) \text{ (or average reward)}$$

Approximate stochastic gradient ascent:

$$\theta_{t+1} \doteq \theta_t + \alpha \widehat{\nabla J(\theta_t)}$$

Typically, based on the Policy-Gradient Theorem:

$$\nabla J(\theta) \propto \sum_s \mu(s) \sum_a q_\pi(s, a) \nabla_\theta \pi(a|s, \theta)$$

on-policy distribution

Policy gradient setup

Given a policy parameterization:

$$\pi(a|s, \theta)$$

Assume an initial state for now

And objective:

$$J(\theta) \doteq v_{\pi_\theta}(s_0) \text{ (or average reward)}$$

Approximate stochastic gradient ascent:

$$\theta_{t+1} \doteq \theta_t + \alpha \widehat{\nabla J(\theta_t)}$$

Typically, based on the Policy-Gradient Theorem:

$$\nabla J(\theta) \propto \sum_s \mu(s) \sum_a q_\pi(s, a) \nabla_\theta \pi(a|s, \theta)$$

on-policy distribution

Policy gradient setup

Given a policy parameterization:

$$\pi(a|s, \theta)$$

Similar to discounted return
in continuing case

And objective:

$$J(\theta) \doteq v_{\pi_\theta}(s_0) \text{ (or average reward)}$$

Approximate stochastic gradient ascent:

$$\theta_{t+1} \doteq \theta_t + \alpha \widehat{\nabla J(\theta_t)}$$

Typically, based on the Policy-Gradient Theorem:

$$\nabla J(\theta) \propto \sum_s \mu(s) \sum_a q_\pi(s, a) \nabla_\theta \pi(a|s, \theta)$$

on-policy distribution

Policy gradient setup

Given a policy parameterization:

$$\pi(a|s, \theta)$$

And objective:

$$J(\theta) \doteq v_{\pi_\theta}(s_0) \text{ (or average reward)}$$

Approximate stochastic gradient ascent:

$$\theta_{t+1} \doteq \theta_t + \alpha \widehat{\nabla J(\theta_t)}$$

Directly maximize objective
with gradient ascent

Typically, based on the Policy-Gradient Theorem:

$$\nabla J(\theta) \propto \sum_s \mu(s) \sum_a q_\pi(s, a) \nabla_\theta \pi(a|s, \theta)$$

on-policy distribution

Policy gradient setup

Given a policy parameterization:

$$\pi(a|s, \theta)$$

And objective:

$$J(\theta) \doteq v_{\pi_\theta}(s_0) \text{ (or average reward)}$$

Approximate stochastic gradient ascent:

$$\theta_{t+1} \doteq \theta_t + \alpha \widehat{\nabla J(\theta_t)}$$

Like before, we are approximating the gradient with samples

Typically, based on the Policy-Gradient Theorem:

$$\nabla J(\theta) \propto \sum_s \mu(s) \sum_a q_\pi(s, a) \nabla_\theta \pi(a|s, \theta)$$

on-policy distribution

Policy gradient setup

Given a policy parameterization:

$$\pi(a|s, \theta)$$

And objective:

$$J(\theta) \doteq v_{\pi_\theta}(s_0) \text{ (or average reward)}$$

Approximate stochastic gradient ascent:

$$\theta_{t+1} \doteq \theta_t + \alpha \widehat{\nabla J(\theta_t)}$$

We'll talk about this next

Typically, based on the Policy-Gradient Theorem:

$$\nabla J(\theta) \propto \sum_s \mu(s) \sum_a q_\pi(s, a) \nabla_\theta \pi(a|s, \theta)$$

on-policy distribution

The policy gradient theorem

- Just like we approximated the value function and updated its parameters with gradient decent, we can do the same thing with the policy parameterization
- Policy now changes smoothly (unlike in action-value methods)
- Can't guarantee the policy improves values at all states, but (remember) we really care about states in proportion to their visits
- But this depends on the policy that is changing
- The policy gradient theorem gives a way to calculate the gradient of the performance objective without the derivative of the state distribution:

$$\nabla J(\boldsymbol{\theta}) \propto \sum_s \mu(s) \sum_a q_{\pi}(s, a) \nabla \pi(a|s, \boldsymbol{\theta})$$

Proof of the Policy Gradient Theorem (episodic case)

With just elementary calculus and re-arranging of terms, we can prove the policy gradient theorem from first principles. To keep the notation simple, we leave it implicit in all cases that π is a function of θ , and all gradients are also implicitly with respect to θ . First note that the gradient of the state-value function can be written in terms of the action-value function as

$$\nabla v_\pi(s) = \nabla \left[\sum_a \pi(a|s) q_\pi(s, a) \right], \quad \text{for all } s \in \mathcal{S} \quad (\text{Exercise 3.16})$$

$$= \sum_a \left[\nabla \pi(a|s) q_\pi(s, a) + \pi(a|s) \nabla q_\pi(s, a) \right] \quad (\text{product rule of calculus})$$

$$= \sum_a \left[\nabla \pi(a|s) q_\pi(s, a) + \pi(a|s) \nabla \sum_{s', r} p(s', r|s, a) (r + v_\pi(s')) \right] \\ (\text{Exercise 3.17 and Equation 3.2})$$

$$= \sum_a \left[\nabla \pi(a|s) q_\pi(s, a) + \pi(a|s) \sum_{s'} p(s'|s, a) \nabla v_\pi(s') \right] \quad (\text{Eq. 3.4})$$

$$= \sum_a \left[\nabla \pi(a|s) q_\pi(s, a) + \pi(a|s) \sum_{s'} p(s'|s, a) \right.$$

$$\left. \sum_{a'} [\nabla \pi(a'|s') q_\pi(s', a') + \pi(a'|s') \sum_{s''} p(s''|s', a') \nabla v_\pi(s'')] \right]$$

$$= \sum_{x \in \mathcal{S}} \sum_{k=0}^{\infty} \Pr(s \rightarrow x, k, \pi) \sum_a \nabla \pi(a|x) q_\pi(x, a),$$

Proof of the Policy Gradient Theorem (episodic case)

With just elementary calculus and re-arranging of terms, we can prove the policy gradient theorem from first principles. To keep the notation simple, we leave it implicit in all cases that π is a function of θ , and all gradients are also implicitly with respect to θ . First note that the gradient of the state-value function can be written in terms of the action-value function as

$$\nabla v_\pi(s) = \nabla \left[\sum_a \pi(a|s) q_\pi(s, a) \right], \quad \text{for all } s \in \mathcal{S} \quad (\text{Exercise 3.16})$$

$$= \sum_a \left[\nabla \pi(a|s) q_\pi(s, a) + \pi(a|s) \nabla q_\pi(s, a) \right] \quad (\text{product rule of calculus})$$

$$= \sum_a \left[\nabla \pi(a|s) q_\pi(s, a) + \pi(a|s) \nabla \sum_{s', r} p(s', r|s, a) (r + v_\pi(s')) \right]$$

(Exercise 3.7 and Equation 3.2)

$$= \sum_a \left[\nabla \pi(a|s) q_\pi(s, a) + \pi(a|s) \sum_{s'} p(s'|s, a) \nabla v_\pi(s') \right] \quad (\text{Eq. 3.4})$$

$$= \sum_a \left[\nabla \pi(a|s) q_\pi(s, a) + \pi(a|s) \sum_{s'} p(s'|s, a) \nabla v_\pi(s') \right]$$

Just replacing q with
it's expectation

$$\sum_{a'} \left[\nabla \pi(a'|s') q_\pi(s', a') + \pi(a'|s') \sum_{s''} p(s''|s', a') \nabla v_\pi(s'') \right]$$

$$= \sum_{x \in \mathcal{S}} \sum_{k=0}^{\infty} \Pr(s \rightarrow x, k, \pi) \sum_a \nabla \pi(a|x) q_\pi(x, a),$$

Proof of the Policy Gradient Theorem (episodic case)

With just elementary calculus and re-arranging of terms, we can prove the policy gradient theorem from first principles. To keep the notation simple, we leave it implicit in all cases that π is a function of θ , and all gradients are also implicitly with respect to θ . First note that the gradient of the state-value function can be written in terms of the action-value function as

$$\nabla v_\pi(s) = \nabla \left[\sum_a \pi(a|s) q_\pi(s, a) \right], \quad \text{for all } s \in \mathcal{S} \quad (\text{Exercise 3.16})$$

$$= \sum_a \left[\nabla \pi(a|s) q_\pi(s, a) + \pi(a|s) \nabla q_\pi(s, a) \right] \quad (\text{product rule of calculus})$$

$$= \sum_a \left[\nabla \pi(a|s) q_\pi(s, a) + \pi(a|s) \nabla \sum_{s', r} p(s', r|s, a) (r + v_\pi(s')) \right] \quad (\text{Exercise 3.17 and Equation 3.2})$$

$$= \sum_a \left[\nabla \pi(a|s) q_\pi(s, a) + \pi(a|s) \sum_{s'} p(s'|s, a) \nabla v_\pi(s') \right] \quad (\text{Eq. 3.4})$$

$$= \sum_a \left[\nabla \pi(a|s) q_\pi(s, a) + \pi(a|s) \sum_{s'} p(s'|s, a) \right] \quad (\text{unrolling})$$

$$\sum_{a'} [\nabla \pi(a'|s') q_\pi(s', a') + \pi(a'|s')] \quad \text{Summing out r}$$

$$= \sum_{x \in \mathcal{S}} \sum_{k=0}^{\infty} \Pr(s \rightarrow x, k, \pi) \sum_a \nabla \pi(a|x) q_\pi(x, a),$$

Proof of the Policy Gradient Theorem (episodic case)

With just elementary calculus and re-arranging of terms, we can prove the policy gradient theorem from first principles. To keep the notation simple, we leave it implicit in all cases that π is a function of θ , and all gradients are also implicitly with respect to θ . First note that the gradient of the state-value function can be written in terms of the action-value function as

$$\nabla v_\pi(s) = \nabla \left[\sum_a \pi(a|s) q_\pi(s, a) \right], \quad \text{for all } s \in \mathcal{S} \quad (\text{Exercise 3.16})$$

$$= \sum_a \left[\nabla \pi(a|s) q_\pi(s, a) + \pi(a|s) \nabla q_\pi(s, a) \right] \quad (\text{product rule of calculus})$$

$$= \sum_a \left[\nabla \pi(a|s) q_\pi(s, a) + \pi(a|s) \nabla \sum_{s', r} p(s', r|s, a) (r + v_\pi(s')) \right] \quad (\text{Exercise 3.17 and Equation 3.2})$$

$$= \sum_a \left[\nabla \pi(a|s) q_\pi(s, a) + \pi(a|s) \sum_{s'} p(s'|s, a) \nabla v_\pi(s') \right] \quad (\text{Eq. 3.4})$$

$$= \sum_a \left[\nabla \pi(a|s) q_\pi(s, a) + \pi(a|s) \sum_{s'} p(s'|s, a) \right. \quad (\text{unrolling})$$

This is what we had
on the previous step

$$\left. \sum_{a'} \left[\nabla \pi(a'|s') q_\pi(s', a') + \pi(a'|s') \sum_{s''} p(s''|s', a') \nabla v_\pi(s'') \right] \right]$$

$$= \sum_{x \in \mathcal{S}} \sum_{k=0}^{\infty} \Pr(s \rightarrow x, k, \pi) \sum_a \nabla \pi(a|x) q_\pi(x, a),$$

Proof of the Policy Gradient Theorem (episodic case)

With just elementary calculus and re-arranging of terms, we can prove the policy gradient theorem from first principles. To keep the notation simple, we leave it implicit in all cases that π is a function of θ , and all gradients are also implicitly with respect to θ . First note that the gradient of the state-value function can be written in terms of the action-value function as

$$\nabla v_\pi(s) = \nabla \left[\sum_a \pi(a|s) q_\pi(s, a) \right], \quad \text{for all } s \in \mathcal{S} \quad (\text{Exercise 3.16})$$

$$= \sum_a \left[\nabla \pi(a|s) q_\pi(s, a) + \pi(a|s) \nabla q_\pi(s, a) \right] \quad (\text{product rule of calculus})$$

$$= \sum_a \left[\nabla \pi(a|s) q_\pi(s, a) + \pi(a|s) \nabla \sum_{s', r} p(s', r|s, a) (r + v_\pi(s')) \right] \quad (\text{Exercise 3.17 and Equation 3.2})$$

$$= \sum_a \left[\nabla \pi(a|s) q_\pi(s, a) + \pi(a|s) \sum_{s'} p(s'|s, a) \nabla v_\pi(s') \right] \quad (\text{Eq. 3.4})$$

$$= \sum_a \left[\nabla \pi(a|s) q_\pi(s, a) + \pi(a|s) \sum_{s'} p(s'|s, a) \right] \quad (\text{unrolling})$$

$P(x, t_k | s, \pi)$

$$\sum' \left[\nabla \pi(a'|s') q_\pi(s', a') + \pi(a'|s') \sum_{s''} p(s''|s', a') \nabla v_\pi(s'') \right]$$

$$= \sum_{x \in \mathcal{S}} \sum_{k=0}^{\infty} \Pr(s \rightarrow x, k, \pi) \sum_a \nabla \pi(a|x) q_\pi(x, a), \quad \text{Repeated unrolling}$$

Policy gradient theorem (cont.)

$$= \sum_{x \in \mathcal{S}} \sum_{k=0}^{\infty} \Pr(s \rightarrow x, k, \pi) \sum_a \nabla \pi(a|x) q_{\pi}(x, a),$$

after repeated unrolling, where $\Pr(s \rightarrow x, k, \pi)$ is the probability of transitioning from state s to state x in k steps under policy π . It is then immediate that

$$\begin{aligned} \nabla J(\boldsymbol{\theta}) &= \nabla v_{\pi}(s_0) \\ &= \sum_s \left(\sum_{k=0}^{\infty} \Pr(s_0 \rightarrow s, k, \pi) \right) \sum_a \nabla \pi(a|s) q_{\pi}(s, a) \\ &= \sum_s \eta(s) \sum_a \nabla \pi(a|s) q_{\pi}(s, a) && \text{(box page 199)} \\ &= \sum_{s'} \eta(s') \sum_s \frac{\eta(s)}{\sum_{s'} \eta(s')} \sum_a \nabla \pi(a|s) q_{\pi}(s, a) \\ &= \sum_{s'} \eta(s') \sum_s \mu(s) \sum_a \nabla \pi(a|s) q_{\pi}(s, a) && \text{(Eq. 9.3)} \\ &\propto \sum_s \mu(s) \sum_a \nabla \pi(a|s) q_{\pi}(s, a) && \text{(Q.E.D.)} \end{aligned}$$

Policy gradient theorem (cont.)

$$= \sum_{x \in \mathcal{S}} \sum_{k=0}^{\infty} \Pr(s \rightarrow x, k, \pi) \sum_a \nabla \pi(a|x) q_{\pi}(x, a),$$

after repeated unrolling, where $\Pr(s \rightarrow x, k, \pi)$ is the probability of transitioning from state s to state x in k steps under policy π . It is then immediate that

$$\nabla J(\theta) = \nabla v_{\pi}(s_0)$$

Just changing notation

$$= \sum_s \left(\sum_{k=0}^{\infty} \Pr(s_0 \rightarrow s, k, \pi) \right) \sum_a \nabla \pi(a|s) q_{\pi}(s, a)$$

$$= \sum_s \eta(s) \sum_a \nabla \pi(a|s) q_{\pi}(s, a) \quad (\text{box page 199})$$

$$= \sum_{s'} \eta(s') \sum_s \frac{\eta(s)}{\sum_{s'} \eta(s')} \sum_a \nabla \pi(a|s) q_{\pi}(s, a)$$

$$= \sum_{s'} \eta(s') \sum_s \mu(s) \sum_a \nabla \pi(a|s) q_{\pi}(s, a) \quad (\text{Eq. 9.3})$$

$$\propto \sum_s \mu(s) \sum_a \nabla \pi(a|s) q_{\pi}(s, a) \quad (\text{Q.E.D.})$$

Policy gradient theorem (cont.)

$$= \sum_{x \in \mathcal{S}} \sum_{k=0}^{\infty} \Pr(s \rightarrow x, k, \pi) \sum_a \nabla \pi(a|x) q_{\pi}(x, a),$$

after repeated unrolling, where $\Pr(s \rightarrow x, k, \pi)$ is the probability of transitioning from state s to state x in k steps under policy π . It is then immediate that

$$\begin{aligned} \nabla J(\boldsymbol{\theta}) &= \nabla v_{\pi}(s_0) \\ &= \sum_s \left(\sum_{k=0}^{\infty} \Pr(s_0 \rightarrow s, k, \pi) \right) \sum_a \nabla \pi(a|s) q_{\pi}(s, a) \\ &= \sum_s \eta(s) \sum_a \nabla \pi(a|s) q_{\pi}(s, a) \end{aligned} \tag{box page 199}$$

$$= \sum_{s'} \eta(s') \sum_s \frac{\eta(s)}{\sum_{s'} \eta(s')} \sum_a \nabla \pi(a|s) q_{\pi}(s, a)$$

Avg number of steps in s

$$= \sum_{s'} \eta(s') \sum_s \mu(s) \sum_a \nabla \pi(a|s) q_{\pi}(s, a) \tag{Eq. 9.3}$$

$$\propto \sum_s \mu(s) \sum_a \nabla \pi(a|s) q_{\pi}(s, a) \tag{Q.E.D.}$$

Policy gradient theorem (cont.)

$$= \sum_{x \in \mathcal{S}} \sum_{k=0}^{\infty} \Pr(s \rightarrow x, k, \pi) \sum_a \nabla \pi(a|x) q_{\pi}(x, a),$$

after repeated unrolling, where $\Pr(s \rightarrow x, k, \pi)$ is the probability of transitioning from state s to state x in k steps under policy π . It is then immediate that

$$\begin{aligned} \nabla J(\boldsymbol{\theta}) &= \nabla v_{\pi}(s_0) \\ &= \sum_s \left(\sum_{k=0}^{\infty} \Pr(s_0 \rightarrow s, k, \pi) \right) \sum_a \nabla \pi(a|s) q_{\pi}(s, a) \\ &= \sum_s \eta(s) \sum_a \nabla \pi(a|s) q_{\pi}(s, a) && \text{(box page 199)} \\ &= \sum_{s'} \eta(s') \sum_s \frac{\eta(s)}{\sum_{s'} \eta(s')} \sum_a \nabla \pi(a|s) q_{\pi}(s, a) \\ &= \sum_{s'} \eta(s') \sum_s \mu(s) \sum_a \nabla \pi(a|s) q_{\pi}(s, a) && \text{(Eq. 9.3)} \\ &\propto \sum_s \mu(s) \sum_a \nabla \pi(a|s) q_{\pi}(s, a) && \text{(Q.E.D.)} \end{aligned}$$

Policy gradient theorem (cont.)

$$= \sum_{x \in \mathcal{S}} \sum_{k=0}^{\infty} \Pr(s \rightarrow x, k, \pi) \sum_a \nabla \pi(a|x) q_{\pi}(x, a),$$

after repeated unrolling, where $\Pr(s \rightarrow x, k, \pi)$ is the probability of transitioning from state s to state x in k steps under policy π . It is then immediate that

$$\begin{aligned} \nabla J(\boldsymbol{\theta}) &= \nabla v_{\pi}(s_0) \\ &= \sum_s \left(\sum_{k=0}^{\infty} \Pr(s_0 \rightarrow s, k, \pi) \right) \sum_a \nabla \pi(a|s) q_{\pi}(s, a) \\ &= \sum_s \eta(s) \sum_a \nabla \pi(a|s) q_{\pi}(s, a) \quad (\text{box page 199}) \\ &= \sum_{s'} \eta(s') \sum_s \frac{\eta(s)}{\sum_{s'} \eta(s')} \sum_a \nabla \pi(a|s) q_{\pi}(s, a) \quad \text{On-policy distribution} \\ &= \sum_{s'} \eta(s') \sum_s \mu(s) \sum_a \nabla \pi(a|s) q_{\pi}(s, a) \quad (\text{Eq. 9.3}) \\ &\propto \sum_s \mu(s) \sum_a \nabla \pi(a|s) q_{\pi}(s, a) \quad (\text{Q.E.D.}) \end{aligned}$$

Policy gradient theorem (cont.)

$$= \sum_{x \in \mathcal{S}} \sum_{k=0}^{\infty} \Pr(s \rightarrow x, k, \pi) \sum_a \nabla \pi(a|x) q_{\pi}(x, a),$$

after repeated unrolling, where $\Pr(s \rightarrow x, k, \pi)$ is the probability of transitioning from state s to state x in k steps under policy π . It is then immediate that

$$\begin{aligned} \nabla J(\boldsymbol{\theta}) &= \nabla v_{\pi}(s_0) \\ &= \sum_s \left(\sum_{k=0}^{\infty} \Pr(s_0 \rightarrow s, k, \pi) \right) \sum_a \nabla \pi(a|s) q_{\pi}(s, a) \\ &= \sum_s \eta(s) \sum_a \nabla \pi(a|s) q_{\pi}(s, a) \quad (\text{box page 199}) \\ &= \sum_{s'} \eta(s') \sum_s \frac{\eta(s)}{\sum_{s'} \eta(s')} \sum_a \nabla \pi(a|s) q_{\pi}(s, a) \\ &= \sum_{s'} \eta(s') \underbrace{\sum_s \mu(s)}_{s} \sum_a \nabla \pi(a|s) q_{\pi}(s, a) \quad (\text{Eq. 9.3}) \\ &\propto \sum_s \mu(s) \sum_a \nabla \pi(a|s) q_{\pi}(s, a) \end{aligned}$$

A constant

(Q.E.D.)

REINFORCE

- How do we actually use the policy gradient theorem?
- Need to estimate the on-policy distribution $\mu(s)$ and the q-values $q_\pi(s, a)$
- Like usual, we'll do this by (Monte Carlo) sampling

Deriving REINFORCE from the PGT

$$\begin{aligned}\nabla J(\boldsymbol{\theta}) &\propto \sum_s \mu(s) \sum_a q_{\pi}(s, a) \nabla_{\boldsymbol{\theta}} \pi(a|s, \boldsymbol{\theta}) \\ &= \mathbb{E}_{\pi} \left[\sum_a q_{\pi}(S_t, a) \nabla_{\boldsymbol{\theta}} \pi(a|S_t, \boldsymbol{\theta}) \right]\end{aligned}$$

Using expectation notation:
we will visit states with the
on-policy distribution when
we follow the policy

Deriving REINFORCE from the PGT

$$\nabla J(\boldsymbol{\theta}) \propto \sum_s \mu(s) \sum_a q_\pi(s, a) \nabla_{\boldsymbol{\theta}} \pi(a|s, \boldsymbol{\theta})$$

$$= \mathbb{E}_\pi \left[\sum_a q_\pi(S_t, a) \nabla_{\boldsymbol{\theta}} \pi(a|S_t, \boldsymbol{\theta}) \right]$$

$$\begin{aligned} &= \mathbb{E}_\pi \left[\sum_a \pi(a|S_t, \boldsymbol{\theta}) q_\pi(S_t, a) \frac{\nabla \pi(a|S_t, \boldsymbol{\theta})}{\pi(a|S_t, \boldsymbol{\theta})} \right] \\ &= \mathbb{E}_\pi \left[q_\pi(S_t, A_t) \frac{\nabla \pi(A_t|S_t, \boldsymbol{\theta})}{\pi(A_t|S_t, \boldsymbol{\theta})} \right] \quad (\text{replacing } a \text{ by the sample } A_t \sim \pi) \\ &= \mathbb{E}_\pi \left[G_t \frac{\nabla \pi(A_t|S_t, \boldsymbol{\theta})}{\pi(A_t|S_t, \boldsymbol{\theta})} \right], \quad (\text{because } \mathbb{E}_\pi[G_t|S_t, A_t] = q_\pi(S_t, A_t)) \end{aligned}$$

Multiplying through

Deriving REINFORCE from the PGT

$$\begin{aligned}\nabla J(\boldsymbol{\theta}) &\propto \sum_s \mu(s) \sum_a q_\pi(s, a) \nabla_{\boldsymbol{\theta}} \pi(a|s, \boldsymbol{\theta}) \\&= \mathbb{E}_\pi \left[\sum_a q_\pi(S_t, a) \nabla_{\boldsymbol{\theta}} \pi(a|S_t, \boldsymbol{\theta}) \right] \\&= \mathbb{E}_\pi \left[\sum_a \pi(a|S_t, \boldsymbol{\theta}) q_\pi(S_t, a) \frac{\nabla \pi(a|S_t, \boldsymbol{\theta})}{\pi(a|S_t, \boldsymbol{\theta})} \right] \\&= \mathbb{E}_\pi \left[q_\pi(S_t, A_t) \frac{\nabla \pi(A_t|S_t, \boldsymbol{\theta})}{\pi(A_t|S_t, \boldsymbol{\theta})} \right] \quad (\text{replacing } a \text{ by the sample } A_t \sim \pi) \\&= \mathbb{E}_\pi \left[G_t \frac{\nabla \pi(A_t|S_t, \boldsymbol{\theta})}{\pi(A_t|S_t, \boldsymbol{\theta})} \right], \quad (\text{because } \mathbb{E}_\pi[G_t|S_t, A_t] = q_\pi(S_t, A_t))\end{aligned}$$

Including this as part
of the expectation

Deriving REINFORCE from the PGT

$$\begin{aligned}\nabla J(\boldsymbol{\theta}) &\propto \sum_s \mu(s) \sum_a q_\pi(s, a) \nabla_{\boldsymbol{\theta}} \pi(a|s, \boldsymbol{\theta}) \\&= \mathbb{E}_\pi \left[\sum_a q_\pi(S_t, a) \nabla_{\boldsymbol{\theta}} \pi(a|S_t, \boldsymbol{\theta}) \right] \\&= \mathbb{E}_\pi \left[\sum_a \pi(a|S_t, \boldsymbol{\theta}) q_\pi(S_t, a) \frac{\nabla \pi(a|S_t, \boldsymbol{\theta})}{\pi(a|S_t, \boldsymbol{\theta})} \right] \\&= \mathbb{E}_\pi \left[q_\pi(S_t, A_t) \frac{\nabla \pi(A_t|S_t, \boldsymbol{\theta})}{\pi(A_t|S_t, \boldsymbol{\theta})} \right] \quad (\text{replacing } a \text{ by the sample } A_t \sim \pi) \\&= \mathbb{E}_\pi \left[G_t \frac{\nabla \pi(A_t|S_t, \boldsymbol{\theta})}{\pi(A_t|S_t, \boldsymbol{\theta})} \right], \quad (\text{because } \mathbb{E}_\pi[G_t|S_t, A_t] = q_\pi(S_t, A_t))\end{aligned}$$

Estimating q with MC returns

Deriving REINFORCE from the PGT

$$\begin{aligned}\nabla J(\boldsymbol{\theta}) &\propto \sum_s \mu(s) \sum_a q_\pi(s, a) \nabla_{\boldsymbol{\theta}} \pi(a|s, \boldsymbol{\theta}) \\&= \mathbb{E}_\pi \left[\sum_a q_\pi(S_t, a) \nabla_{\boldsymbol{\theta}} \pi(a|S_t, \boldsymbol{\theta}) \right] \\&= \mathbb{E}_\pi \left[\sum_a \pi(a|S_t, \boldsymbol{\theta}) q_\pi(S_t, a) \frac{\nabla \pi(a|S_t, \boldsymbol{\theta})}{\pi(a|S_t, \boldsymbol{\theta})} \right] \\&= \mathbb{E}_\pi \left[q_\pi(S_t, A_t) \frac{\nabla \pi(A_t|S_t, \boldsymbol{\theta})}{\pi(A_t|S_t, \boldsymbol{\theta})} \right] \quad (\text{replacing } a \text{ by the sample } A_t \sim \pi) \\&= \mathbb{E}_\pi \left[G_t \frac{\nabla \pi(A_t|S_t, \boldsymbol{\theta})}{\pi(A_t|S_t, \boldsymbol{\theta})} \right], \quad (\text{because } \mathbb{E}_\pi[G_t|S_t, A_t] = q_\pi(S_t, A_t))\end{aligned}$$

Thus

SGD version

$$\boldsymbol{\theta}_{t+1} \doteq \boldsymbol{\theta}_t + \alpha G_t \frac{\nabla_{\boldsymbol{\theta}} \pi(A_t|S_t, \boldsymbol{\theta}_t)}{\pi(A_t|S_t, \boldsymbol{\theta}_t)}$$

Deriving REINFORCE from the PGT

$$\begin{aligned}\nabla J(\boldsymbol{\theta}) &\propto \sum_s \mu(s) \sum_a q_\pi(s, a) \nabla_{\boldsymbol{\theta}} \pi(a|s, \boldsymbol{\theta}) \\&= \mathbb{E}_\pi \left[\sum_a q_\pi(S_t, a) \nabla_{\boldsymbol{\theta}} \pi(a|S_t, \boldsymbol{\theta}) \right] \\&= \mathbb{E}_\pi \left[\sum_a \pi(a|S_t, \boldsymbol{\theta}) q_\pi(S_t, a) \frac{\nabla \pi(a|S_t, \boldsymbol{\theta})}{\pi(a|S_t, \boldsymbol{\theta})} \right] \\&= \mathbb{E}_\pi \left[q_\pi(S_t, A_t) \frac{\nabla \pi(A_t|S_t, \boldsymbol{\theta})}{\pi(A_t|S_t, \boldsymbol{\theta})} \right] \quad (\text{replacing } a \text{ by the sample } A_t \sim \pi) \\&= \mathbb{E}_\pi \left[G_t \frac{\nabla \pi(A_t|S_t, \boldsymbol{\theta})}{\pi(A_t|S_t, \boldsymbol{\theta})} \right], \quad (\text{because } \mathbb{E}_\pi[G_t|S_t, A_t] = q_\pi(S_t, A_t))\end{aligned}$$

Thus

$$\boldsymbol{\theta}_{t+1} \doteq \boldsymbol{\theta}_t + \alpha G_t \frac{\nabla_{\boldsymbol{\theta}} \pi(A_t|S_t, \boldsymbol{\theta}_t)}{\pi(A_t|S_t, \boldsymbol{\theta}_t)}$$

The REINFORCE update

Deriving REINFORCE from the PGT

$$\begin{aligned}\nabla J(\boldsymbol{\theta}) &\propto \sum_s \mu(s) \sum_a q_\pi(s, a) \nabla_{\boldsymbol{\theta}} \pi(a|s, \boldsymbol{\theta}) \\&= \mathbb{E}_\pi \left[\sum_a q_\pi(S_t, a) \nabla_{\boldsymbol{\theta}} \pi(a|S_t, \boldsymbol{\theta}) \right] \\&= \mathbb{E}_\pi \left[\sum_a \pi(a|S_t, \boldsymbol{\theta}) q_\pi(S_t, a) \frac{\nabla \pi(a|S_t, \boldsymbol{\theta})}{\pi(a|S_t, \boldsymbol{\theta})} \right] \\&= \mathbb{E}_\pi \left[q_\pi(S_t, A_t) \frac{\nabla \pi(A_t|S_t, \boldsymbol{\theta})}{\pi(A_t|S_t, \boldsymbol{\theta})} \right] \quad (\text{replacing } a \text{ by the sample } A_t \sim \pi) \\&= \mathbb{E}_\pi \left[G_t \frac{\nabla \pi(A_t|S_t, \boldsymbol{\theta})}{\pi(A_t|S_t, \boldsymbol{\theta})} \right], \quad (\text{because } \mathbb{E}_\pi[G_t|S_t, A_t] = q_\pi(S_t, A_t))\end{aligned}$$

Thus

Parameters increase
with return

$$\boldsymbol{\theta}_{t+1} \doteq \boldsymbol{\theta}_t + \alpha G_t \frac{\nabla_{\boldsymbol{\theta}} \pi(A_t|S_t, \boldsymbol{\theta}_t)}{\pi(A_t|S_t, \boldsymbol{\theta}_t)}$$

Deriving REINFORCE from the PGT

$$\begin{aligned}\nabla J(\boldsymbol{\theta}) &\propto \sum_s \mu(s) \sum_a q_\pi(s, a) \nabla_{\boldsymbol{\theta}} \pi(a|s, \boldsymbol{\theta}) \\&= \mathbb{E}_\pi \left[\sum_a q_\pi(S_t, a) \nabla_{\boldsymbol{\theta}} \pi(a|S_t, \boldsymbol{\theta}) \right] \\&= \mathbb{E}_\pi \left[\sum_a \pi(a|S_t, \boldsymbol{\theta}) q_\pi(S_t, a) \frac{\nabla \pi(a|S_t, \boldsymbol{\theta})}{\pi(a|S_t, \boldsymbol{\theta})} \right] \\&= \mathbb{E}_\pi \left[q_\pi(S_t, A_t) \frac{\nabla \pi(A_t|S_t, \boldsymbol{\theta})}{\pi(A_t|S_t, \boldsymbol{\theta})} \right] \quad (\text{replacing } a \text{ by the sample } A_t \sim \pi) \\&= \mathbb{E}_\pi \left[G_t \frac{\nabla \pi(A_t|S_t, \boldsymbol{\theta})}{\pi(A_t|S_t, \boldsymbol{\theta})} \right], \quad (\text{because } \mathbb{E}_\pi[G_t|S_t, A_t] = q_\pi(S_t, A_t))\end{aligned}$$

Thus

$$\boldsymbol{\theta}_{t+1} \doteq \boldsymbol{\theta}_t + \alpha G_t \frac{\nabla_{\boldsymbol{\theta}} \pi(A_t|S_t, \boldsymbol{\theta}_t)}{\pi(A_t|S_t, \boldsymbol{\theta}_t)}$$

Towards repeating A_t in S_t
adjusting magnitude based
on probability

REINFORCE

REINFORCE: Monte-Carlo Policy-Gradient Control (episodic) for π_*

Input: a differentiable policy parameterization $\pi(a|s, \theta)$

Algorithm parameter: step size $\alpha > 0$

Initialize policy parameter $\theta \in \mathbb{R}^{d'}$ (e.g., to $\mathbf{0}$)

Loop forever (for each episode):

Generate an episode $S_0, A_0, R_1, \dots, S_{T-1}, A_{T-1}, R_T$, following $\pi(\cdot| \cdot, \theta)$

Loop for each step of the episode $t = 0, 1, \dots, T - 1$:

$$\begin{aligned} G &\leftarrow \sum_{k=t+1}^T \gamma^{k-t-1} R_k \\ \theta &\leftarrow \theta + \alpha \gamma^t G \nabla \ln \pi(A_t | S_t, \theta) \end{aligned} \tag{G_t}$$

- What's different from what is presented before?
- Is it still correct?

REINFORCE

REINFORCE: Monte-Carlo Policy-Gradient Control (episodic) for π_*

Input: a differentiable policy parameterization $\pi(a|s, \theta)$

Algorithm parameter: step size $\alpha > 0$

Initialize policy parameter $\theta \in \mathbb{R}^{d'}$ (e.g., to $\mathbf{0}$)

Loop forever (for each episode):

Generate an episode $S_0, A_0, R_1, \dots, S_{T-1}, A_{T-1}, R_T$, following $\pi(\cdot| \cdot, \theta)$

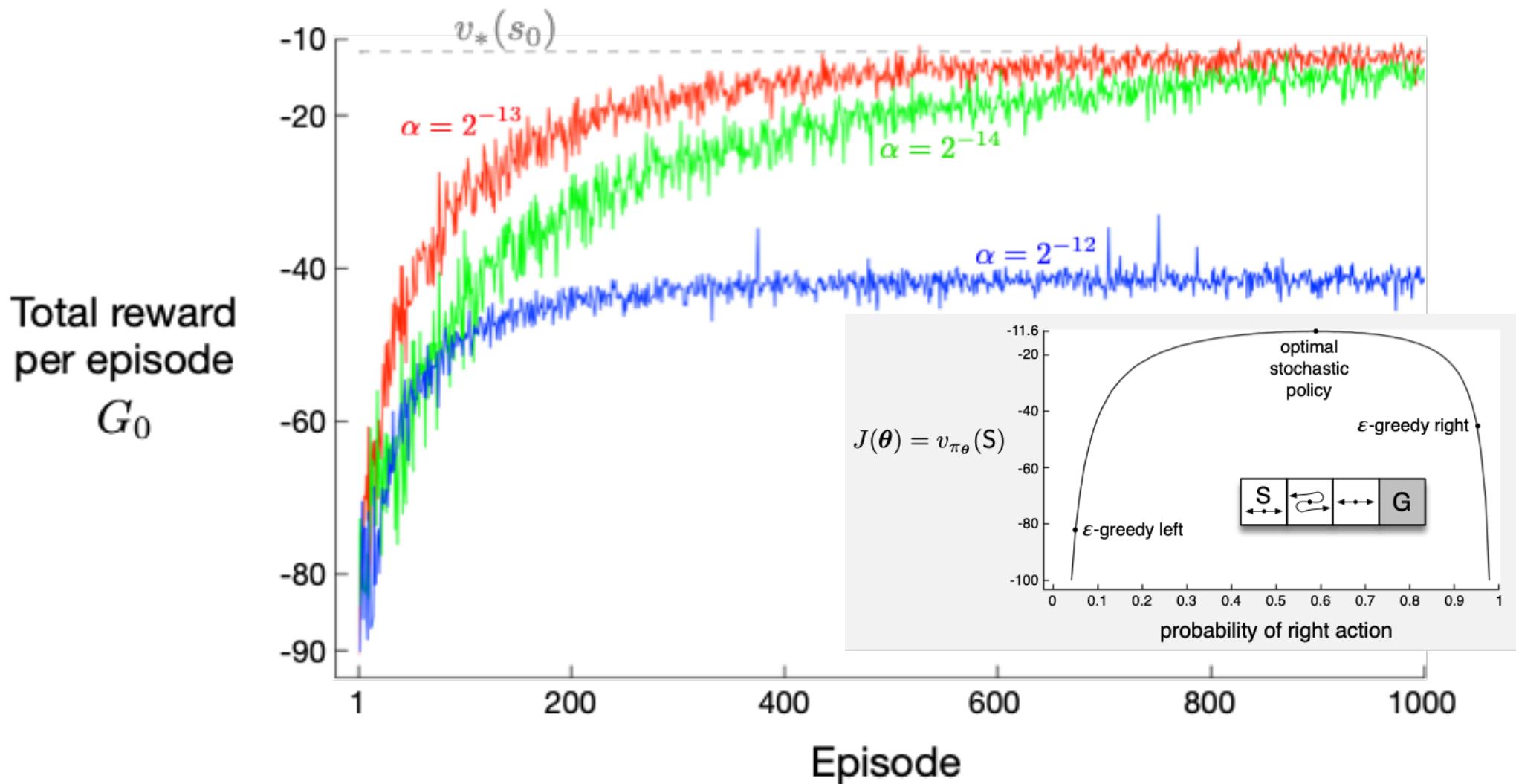
Loop for each step of the episode $t = 0, 1, \dots, T - 1$:

$$\begin{aligned} G &\leftarrow \sum_{k=t+1}^T \gamma^{k-t-1} R_k \\ \theta &\leftarrow \theta + \alpha \gamma^t G \nabla \ln \pi(A_t | S_t, \theta) \end{aligned} \tag{G_t}$$

$$\nabla \ln x = \frac{\nabla x}{x}$$

- $\nabla \ln \pi(A_t | S_t, \theta_t)$ called **eligibility vector** in book

REINFORCE finds the right probability



REINFORCE with baseline

- As we know, Monte Carlo estimates have high variance
- What did we do in the gradient bandit case?

REINFORCE with baseline

Policy-gradient theorem with baseline:

$$\begin{aligned}\nabla J(\boldsymbol{\theta}) &\propto \sum_s \mu(s) \sum_a q_\pi(s, a) \nabla_{\boldsymbol{\theta}} \pi(a|s, \boldsymbol{\theta}) \\ &= \sum_s \mu(s) \sum_a \left(q_\pi(s, a) - b(s) \right) \nabla_{\boldsymbol{\theta}} \pi(a|s, \boldsymbol{\theta})\end{aligned}$$

any function of state, not action

REINFORCE with baseline

Policy-gradient theorem with baseline:

Why?

$$\begin{aligned}\nabla J(\boldsymbol{\theta}) &\propto \sum_s \mu(s) \sum_a q_\pi(s, a) \nabla_{\boldsymbol{\theta}} \pi(a|s, \boldsymbol{\theta}) \\ &= \sum_s \mu(s) \sum_a \left(q_\pi(s, a) - b(s) \right) \nabla_{\boldsymbol{\theta}} \pi(a|s, \boldsymbol{\theta})\end{aligned}$$

any function of state, not action

REINFORCE with baseline

Policy-gradient theorem with baseline:

$$\begin{aligned}\nabla J(\boldsymbol{\theta}) &\propto \sum_s \mu(s) \sum_a q_\pi(s, a) \nabla_{\boldsymbol{\theta}} \pi(a|s, \boldsymbol{\theta}) \\ &= \sum_s \mu(s) \sum_a \left(q_\pi(s, a) - b(s) \right) \nabla_{\boldsymbol{\theta}} \pi(a|s, \boldsymbol{\theta})\end{aligned}$$

any function of state, not action

Because

$$\sum_a b(s) \nabla_{\boldsymbol{\theta}} \pi(a|s, \boldsymbol{\theta}) = b(s) \nabla_{\boldsymbol{\theta}} \sum_a \pi(a|s, \boldsymbol{\theta}) = b(s) \nabla_{\boldsymbol{\theta}} 1 = 0 \quad \forall s \in \mathcal{S}.$$

REINFORCE with baseline

Policy-gradient theorem with baseline:

$$\begin{aligned}\nabla J(\boldsymbol{\theta}) &\propto \sum_s \mu(s) \sum_a q_\pi(s, a) \nabla_{\boldsymbol{\theta}} \pi(a|s, \boldsymbol{\theta}) \\ &= \sum_s \mu(s) \sum_a \left(q_\pi(s, a) - b(s) \right) \nabla_{\boldsymbol{\theta}} \pi(a|s, \boldsymbol{\theta})\end{aligned}$$

any function of state, not action

Because

$$\sum_a b(s) \nabla_{\boldsymbol{\theta}} \pi(a|s, \boldsymbol{\theta}) = b(s) \nabla_{\boldsymbol{\theta}} \sum_a \pi(a|s, \boldsymbol{\theta}) = b(s) \nabla_{\boldsymbol{\theta}} 1 = 0 \quad \forall s \in \mathcal{S}$$

Thus

$$\boldsymbol{\theta}_{t+1} \doteq \boldsymbol{\theta}_t + \alpha \left(G_t - b(S_t) \right) \frac{\nabla_{\boldsymbol{\theta}} \pi(A_t|S_t, \boldsymbol{\theta}_t)}{\pi(A_t|S_t, \boldsymbol{\theta}_t)}$$

e.g., $b(s) = \hat{v}(s, \mathbf{w})$

REINFORCE with Baseline (episodic)

Input: a differentiable policy parameterization $\pi(a|s, \boldsymbol{\theta})$

Input: a differentiable state-value parameterization $\hat{v}(s, \mathbf{w})$

Parameters: step sizes $\alpha > 0, \beta > 0$

Initialize policy parameter $\boldsymbol{\theta} \in \mathbb{R}^{d'}$ and state-value weights $\mathbf{w} \in \mathbb{R}^d$

Repeat forever:

 Generate an episode $S_0, A_0, R_1, \dots, S_{T-1}, A_{T-1}, R_T$, following $\pi(\cdot| \cdot, \boldsymbol{\theta})$

 For each step of the episode $t = 0, \dots, T - 1$:

$G_t \leftarrow$ return from step t

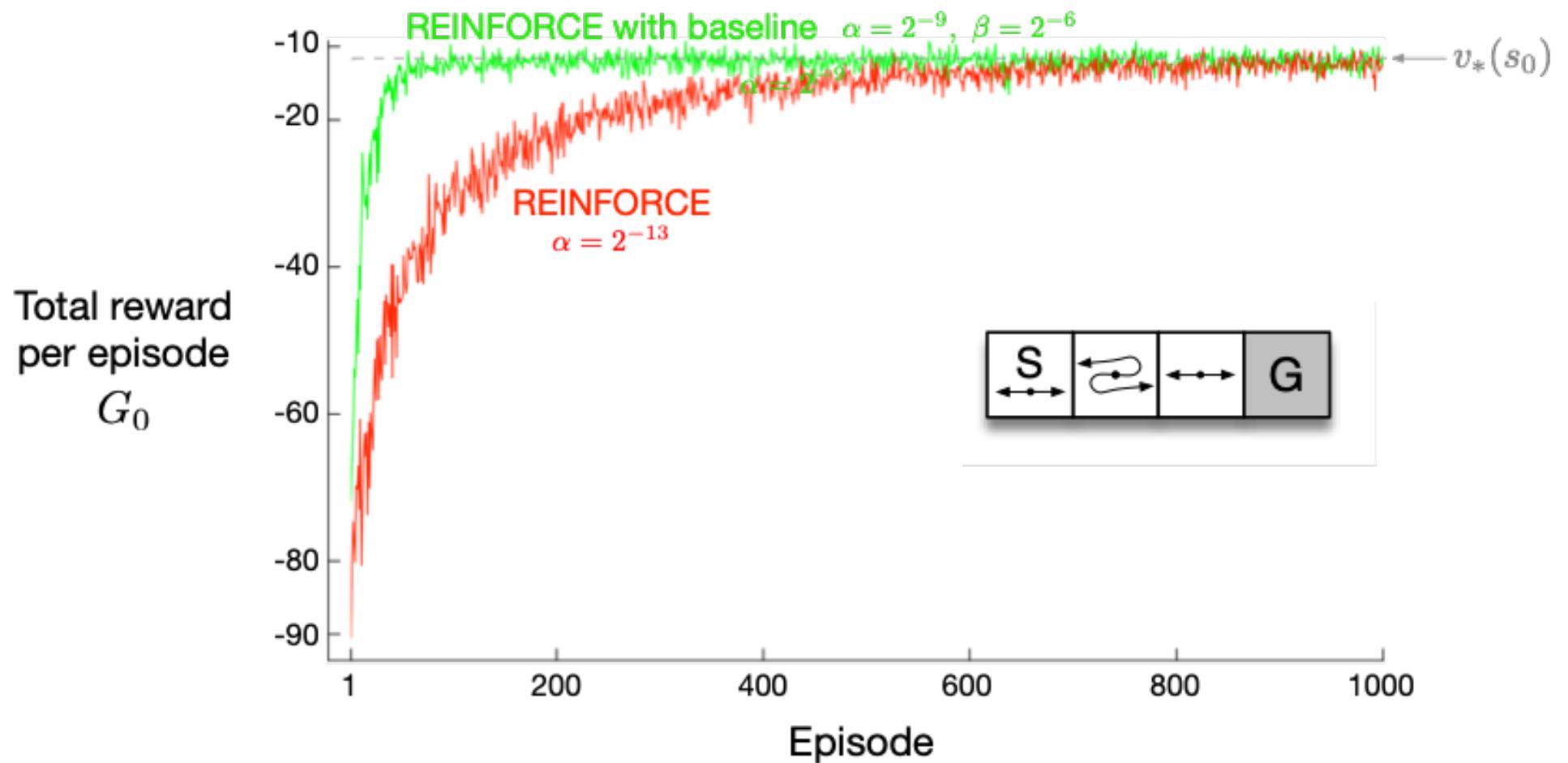
$\delta \leftarrow G_t - \hat{v}(S_t, \mathbf{w})$

$\mathbf{w} \leftarrow \mathbf{w} + \beta \gamma^t \delta \nabla_{\mathbf{w}} \hat{v}(S_t, \mathbf{w})$

$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \alpha \gamma^t \delta \nabla_{\boldsymbol{\theta}} \ln \pi(A_t | S_t, \boldsymbol{\theta})$

$$\text{Again } \nabla_{\boldsymbol{\theta}} \ln \pi(a|s, \boldsymbol{\theta}) = \frac{\nabla_{\boldsymbol{\theta}} \pi(a|s, \boldsymbol{\theta})}{\pi(a|s, \boldsymbol{\theta})}$$

REINFORCE with baseline is faster



Actor-critic methods

- How is the state-value estimate used in REINFORCE? Does it use bootstrapping?
- Actor-critic methods use bootstrapping
 - As a result they are biased, but often learn faster than MC methods
 - Replace the MC return with a bootstrapped return (and still use a state-value baseline)

Actor-Critic methods

REINFORCE with baseline:

$$\boldsymbol{\theta}_{t+1} \doteq \boldsymbol{\theta}_t + \alpha \left(G_t - b(S_t) \right) \frac{\nabla_{\boldsymbol{\theta}} \pi(A_t | S_t, \boldsymbol{\theta}_t)}{\pi(A_t | S_t, \boldsymbol{\theta}_t)}$$

Actor-Critic method (TD(0) version):

$$\begin{aligned}\boldsymbol{\theta}_{t+1} &\doteq \boldsymbol{\theta}_t + \alpha \left(G_{t:t+1} - \hat{v}(S_t, \mathbf{w}) \right) \frac{\nabla_{\boldsymbol{\theta}} \pi(A_t | S_t, \boldsymbol{\theta}_t)}{\pi(A_t | S_t, \boldsymbol{\theta}_t)} \\ &= \boldsymbol{\theta}_t + \alpha \left(R_{t+1} + \gamma \hat{v}(S_{t+1}, \mathbf{w}) - \hat{v}(S_t, \mathbf{w}) \right) \frac{\nabla_{\boldsymbol{\theta}} \pi(A_t | S_t, \boldsymbol{\theta}_t)}{\pi(A_t | S_t, \boldsymbol{\theta}_t)} \\ &= \boldsymbol{\theta}_t + \alpha \delta_t \frac{\nabla_{\boldsymbol{\theta}} \pi(A_t | S_t, \boldsymbol{\theta}_t)}{\pi(A_t | S_t, \boldsymbol{\theta}_t)}.\end{aligned}$$

Actor-Critic (TD(0) version)

One-step Actor–Critic (episodic), for estimating $\pi_\theta \approx \pi_*$

Input: a differentiable policy parameterization $\pi(a|s, \theta)$

Input: a differentiable state-value function parameterization $\hat{v}(s, \mathbf{w})$

Parameters: step sizes $\alpha^\theta > 0$, $\alpha^w > 0$

Initialize policy parameter $\theta \in \mathbb{R}^{d'}$ and state-value weights $\mathbf{w} \in \mathbb{R}^d$ (e.g., to $\mathbf{0}$)

Loop forever (for each episode):

Initialize S (first state of episode)

What is *I* doing?

Loop while S is not terminal (for each time step):

$$A \sim \pi(\cdot | S, \theta)$$

Take action A , observe S', R

$$\delta \leftarrow R + \gamma \hat{v}(S', \mathbf{w}) - \hat{v}(S, \mathbf{w})$$

(if S' is terminal, then $\hat{v}(S', \mathbf{w}) \doteq 0$)

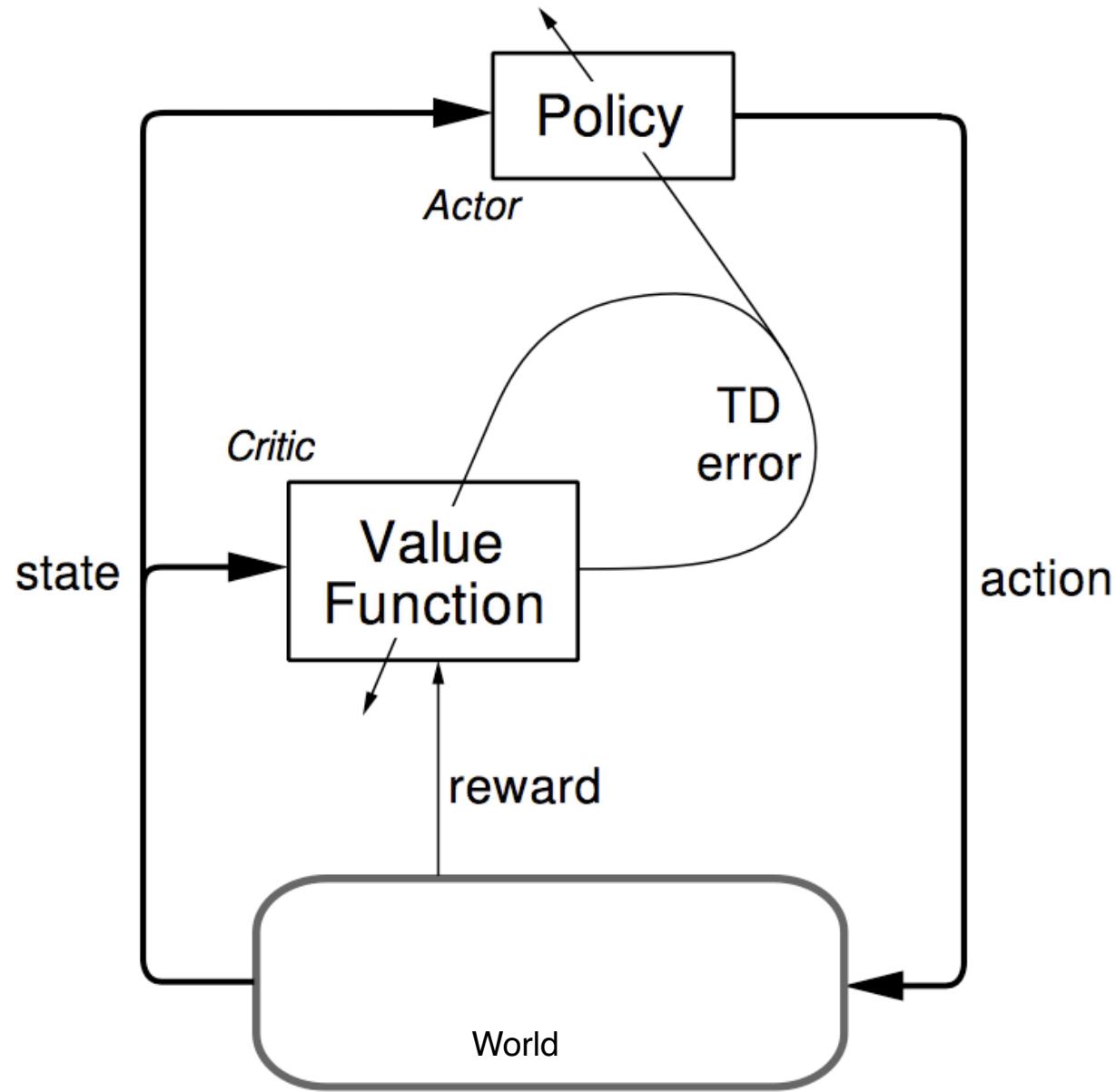
$$\mathbf{w} \leftarrow \mathbf{w} + \alpha^{\mathbf{w}} I \delta \nabla \hat{v}(S, \mathbf{w})$$

$$\theta \leftarrow \theta + \alpha^\theta I \delta \nabla \ln \pi(A|S, \theta)$$

$$I \leftarrow \gamma I$$

$$S \leftarrow S'$$

Actor-critic architecture



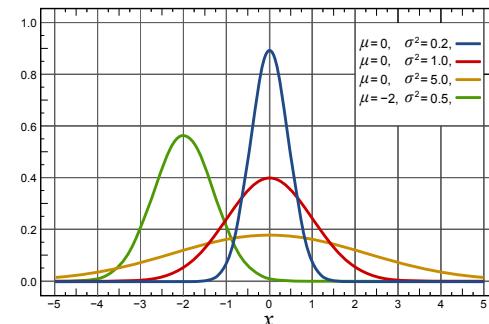
Continuous actions

- Policy gradient methods can also learn parameterizations of continuous action policies
 - Either as parameters for a continuous probability distribution (rather than a discrete one)
 - Or directly considering parameters of the continuous action space

Actions chosen by a Gaussian

- Actions as real numbers
- Gaussian given by mean and standard deviation

$$p(x) \doteq \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x - \mu)^2}{2\sigma^2}\right)$$



- The policy can be defined with a Gaussian

$$\pi(a|s, \theta) \doteq \frac{1}{\sigma(s, \theta)\sqrt{2\pi}} \exp\left(-\frac{(a - \mu(s, \theta))^2}{2\sigma(s, \theta)^2}\right)$$

- Need to learn

$$\mu : \mathcal{S} \times \mathbb{R}^{d'} \rightarrow \mathbb{R} \quad \sigma : \mathcal{S} \times \mathbb{R}^{d'} \rightarrow \mathbb{R}^+$$

Actions chosen by a Gaussian

- The policy can be defined with a Gaussian

$$\pi(a|s, \boldsymbol{\theta}) \doteq \frac{1}{\sigma(s, \boldsymbol{\theta})\sqrt{2\pi}} \exp\left(-\frac{(a - \mu(s, \boldsymbol{\theta}))^2}{2\sigma(s, \boldsymbol{\theta})^2}\right)$$

- Need to learn

$$\mu : \mathcal{S} \times \mathbb{R}^{d'} \rightarrow \mathbb{R} \quad \sigma : \mathcal{S} \times \mathbb{R}^{d'} \rightarrow \mathbb{R}^+$$

- Maybe we want a linear representation for μ and an exponential for σ

$$\mu(s, \boldsymbol{\theta}) \doteq \boldsymbol{\theta}_\mu^\top \mathbf{x}_\mu(s) \quad \text{and} \quad \sigma(s, \boldsymbol{\theta}) \doteq \exp\left(\boldsymbol{\theta}_\sigma^\top \mathbf{x}_\sigma(s)\right)$$

- We can use the same algorithms to learn these parameters

Gaussian eligibility functions

$$\frac{\nabla_{\theta_\mu} \pi(a|s, \theta)}{\pi(a|s, \theta)} = \frac{1}{\sigma(s, \theta)^2} (a - \mu(s, \theta)) \mathbf{x}_\mu(s)$$

$$\frac{\nabla_{\theta_\sigma} \pi(a|s, \theta)}{\pi(a|s, \theta)} = \left(\frac{(a - \mu(s, \theta))^2}{\sigma(s, \theta)^2} - 1 \right) \mathbf{x}_\sigma(s)$$

Summary

- Policy gradient methods allow learning of policies (as well as value functions)
- More naturally scale to large and continuous action spaces than action-value methods
- Like other gradient methods only have locally optimal convergence guarantees

Next time

- Deep policy gradient methods