RL: Policy Gradient

How do we decide what to do?

- Emotions/Intuition

$$V_t(s) \qquad Q_t(s,a)$$

- Thinking

$$S_{t+1} = M(S_t, A_t, \theta)$$

- Reflexes/Habits

$$A_t = \pi(S_t, \theta)$$

Policy Approximation

We want to learn this directly!

$$\pi(a|s, \boldsymbol{\theta})$$

- 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

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:

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

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

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

How can we learn $\pi(a|s, \boldsymbol{\theta})$?

How can we learn $\pi(a|s, \boldsymbol{\theta})$?

- Directly from Experience?

- From V and Q?

- From a World-Model M(S, A) = S' ?

How can we learn $\pi(a|s, \boldsymbol{\theta})$?

- Directly from Experience?
  - REINFORCE
- From V and Q?
  - Actor Critic Algorithms
  - Deterministic Policy Gradient (DPG)
- From a World-Model M(S, A) = S' ?

Parametrizing π, how do we write π as a neural net?

- For discrete actions?

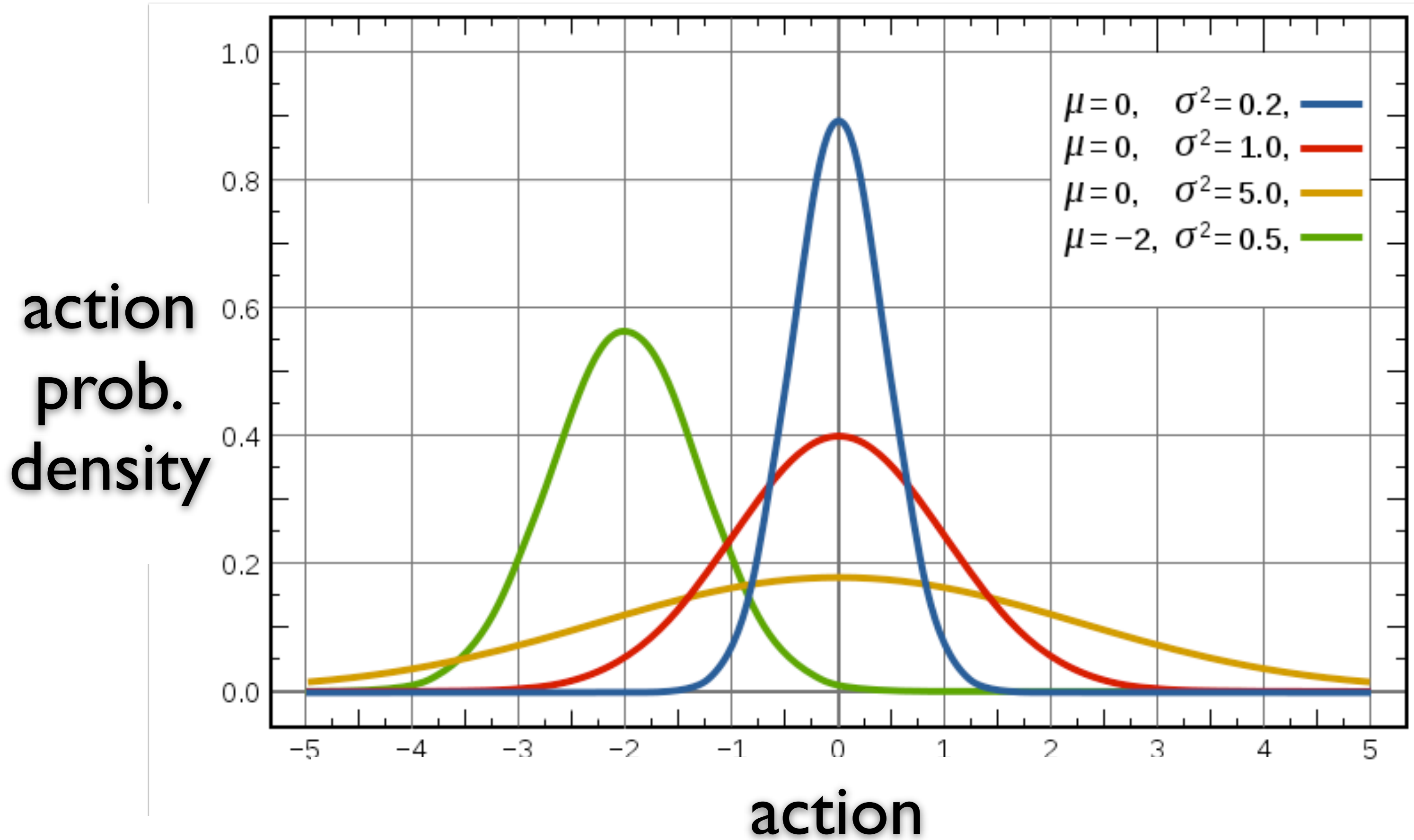
- For continuous actions?

Typical example - **Deep Softmax Policies** for **discrete actions**:

$$\pi(A_i \mid S) = \frac{\exp(\phi(A_i, S))}{\sum_j \exp(\phi(A_j, S))}$$

where ϕ is a neural network,
or any other function approximation parametrize by some weights.

Typical example - **Gaussian Policies** for **continuous actions**:



Typical example - **Gaussian Policies** for **continuous actions**:

$$\mu(S), \sigma(S) = \phi(S)$$

where ϕ is a neural network,
or any other function approximation
parametrize by some weights.

$$\pi(A \mid S) = \mathcal{N}\left(\mu(S), \sigma(S)\right)$$

Typical example - **Gaussian Policies** for **continuous actions**:

These are vectors if the action
has more than 1 dim,
Example: the torques for
4 different motors.

$$\mu(S), \sigma(S) = \phi(S)$$

where ϕ is a neural network,
or any other function approximation
parametrize by some weights.

$$\pi(A \mid S) = \mathcal{N}\left(\mu(S), \sigma(S)\right)$$

Typical example - **Gaussian Policies** for **continuous actions**:

$$\mu(S), \sigma(S) = \phi(S)$$

where ϕ is a neural network,
or any other function approximation
parametrize by some weights.

$$\pi(A \mid S) = \mathcal{N}\left(\mu(S), \sigma(S)\right)$$

Act by sampling from the distribution:

$$A = \mu(S) + \sigma(S)\epsilon, \qquad \epsilon \sim \mathcal{N}(0,1)$$

REINFORCE ALGORITHM

Only π

~~V~~, ~~Q~~, ~~M~~

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:

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

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

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

Policy Gradient

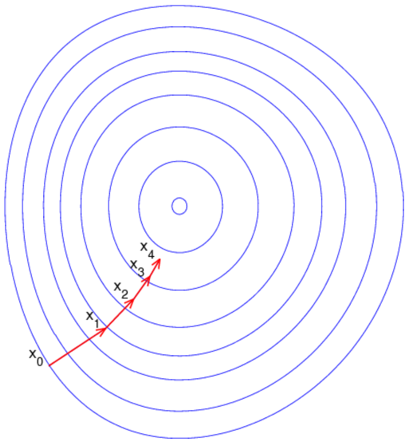
- Idea: ascent the gradient of the objective $J(\boldsymbol{\theta})$

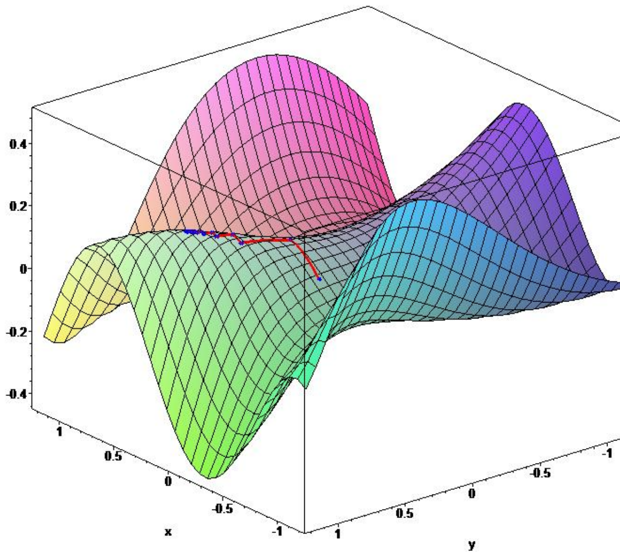
$$\Delta\boldsymbol{\theta} = \alpha \nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta})$$

- Where $\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta})$ is the **policy gradient**

$$\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) = \begin{pmatrix} \frac{\partial J(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_1} \\ \vdots \\ \frac{\partial J(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_n} \end{pmatrix}$$

- and α is a step-size parameter
- Stochastic policies help ensure $J(\boldsymbol{\theta})$ is smooth (typically/mostly)





Contextual Bandits Policy Gradient

- Consider a one-step case (a contextual bandit) such that $\boldsymbol{J(\theta) = \mathbb{E}_{\pi_\theta}[R(S, A)]}$.
  (Expectation is over d (states) and π (actions))
  (For now, d does **not** depend on π)
- We cannot sample R_{t+1} and then take a gradient:
  R_{t+1} is just a number and does not depend on $\boldsymbol{\theta}$!
- Instead, we use the identity:

$$\nabla_{\boldsymbol{\theta}} \mathbb{E}_{\pi_{\boldsymbol{\theta}}}[R(S, A)] = \mathbb{E}_{\pi_{\boldsymbol{\theta}}}[R(S, A) \nabla_{\boldsymbol{\theta}} \log \pi(A|S)] .$$

  (Proof on next slide)
- The right-hand side gives an expected gradient that can be sampled
- Also known as REINFORCE (Williams, 1992)

The score function trick

Let $r_{sa} = \mathbb{E}\left[R(S, A) \mid S = s, A = s\right]$

$$
\begin{aligned}
\nabla_{\boldsymbol{\theta}} \mathbb{E}_{\pi_{\boldsymbol{\theta}}}[R(S, A)] &= \nabla_{\boldsymbol{\theta}} \sum_s d(s) \sum_a \pi_{\boldsymbol{\theta}}(a|s)\, r_{sa} \\
&= \sum_s d(s) \sum_a r_{sa}\, \nabla_{\boldsymbol{\theta}} \pi_{\boldsymbol{\theta}}(a|s) \\
&= \sum_s d(s) \sum_a r_{sa}\, \pi_{\boldsymbol{\theta}}(a|s) \frac{\nabla_{\boldsymbol{\theta}} \pi_{\boldsymbol{\theta}}(a|s)}{\pi_{\boldsymbol{\theta}}(a|s)} \\
&= \sum_s d(s) \sum_a \pi_{\boldsymbol{\theta}}(a|s)\, r_{sa}\, \nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(a|s) \\
&= \mathbb{E}_{d, \pi_{\boldsymbol{\theta}}}[R(S, A)\, \nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(A|S)]
\end{aligned}
$$

Policy Gradient Theorem

- The policy gradient approach also applies to (multi-step) MDPs
- Replaces reward R with long-term return G_t or value $q_\pi(s, a)$
- There are actually two policy gradient theorems (Sutton et al., 2000):

average return per episode & **average reward per step**

Policy gradient theorem (episodic)

Theorem

For any differentiable policy $\pi_{\boldsymbol{\theta}}(s, a)$, let d_0 be the starting distribution over states in which we begin an episode. Then, the policy gradient of $J(\boldsymbol{\theta}) = \mathbb{E}\left[G_0 \mid S_0 \sim d_0\right]$ is

$$\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) = \mathbb{E}_{\pi_{\boldsymbol{\theta}}}\left[\sum_{t=0}^{T} \gamma^t q_{\pi_{\boldsymbol{\theta}}}(S_t, A_t) \nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(A_t | S_t) \mid S_0 \sim d_0\right]$$

where

$$\begin{aligned} q_\pi(s, a) &= \mathbb{E}_\pi[G_t \mid S_t = s, A_t = a] \\ &= \mathbb{E}_\pi[R_{t+1} + \gamma q_\pi(S_{t+1}, A_{t+1}) \mid S_t = s, A_t = a] \end{aligned}$$

Notice this is the return and not the reward, G not r!

Policy gradient theorem (episodic)

Theorem

For any differentiable policy $\pi_{\boldsymbol{\theta}}(s,a)$, let d_0 be the starting distribution over states in which we begin an episode. Then, the policy gradient of $J(\boldsymbol{\theta}) = \mathbb{E}\left[G_0 \mid S_0 \sim d_0\right]$ is

$$\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) = \mathbb{E}_{\pi_{\boldsymbol{\theta}}}\left[\sum_{t=0}^{T} \gamma^t q_{\pi_{\boldsymbol{\theta}}}(S_t, A_t) \nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(A_t|S_t) \mid S_0 \sim d_0\right]$$

where

$$\begin{aligned} q_\pi(s,a) &= \mathbb{E}_\pi[G_t \mid S_t = s, A_t = a] \\ &= \mathbb{E}_\pi[R_{t+1} + \gamma q_\pi(S_{t+1}, A_{t+1}) \mid S_t = s, A_t = a] \end{aligned}$$

Important "Trick" / Identity

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

Important "Trick" / Identity

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

Or written in a different way:

$$\mathbb{E}\left(b(s)\, \nabla_\theta \log(\pi(a\,|\,s, \theta))\right) = \sum_{s,a} b(s) p(s) \pi(a\,|\,s, \theta)\, \nabla_\theta \log(\pi(a\,|\,s, \theta))$$

Important "Trick" / Identity

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

Or written in a different way:

$$\mathbb{E}\left(b(s)\nabla_\theta \log(\pi(a\,|\,s,\theta))\right) = \sum_{s,a} b(s)p(s)\pi(a\,|\,s,\theta)\nabla_\theta \log(\pi(a\,|\,s,\theta))$$

$$= \sum_{s,a} b(s)p(s)\pi(a\,|\,s,\theta)\frac{\nabla_\theta \pi(a\,|\,s,\theta)}{\pi(a\,|\,s,\theta)}$$

Important "Trick" / Identity

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

Or written in a different way:

$$\mathbb{E}\left(b(s)\nabla_\theta \log(\pi(a\,|\,s,\theta))\right) = \sum_{s,a} b(s)p(s)\pi(a\,|\,s,\theta)\nabla_\theta \log(\pi(a\,|\,s,\theta))$$

$$= \sum_{s,a} b(s)p(s)\pi(a\,|\,s,\theta)\frac{\nabla_\theta \pi(a\,|\,s,\theta)}{\pi(a\,|\,s,\theta)}$$

$$= \sum_{s,a} b(s)p(s)\nabla_\theta \pi(a\,|\,s,\theta)$$

Important "Trick" / Identity

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

Or written in a different way:

$$\mathbb{E}\left(b(s)\nabla_\theta \log(\pi(a\,|\,s,\theta))\right) = \sum_{s,a} b(s)p(s)\pi(a\,|\,s,\theta)\nabla_\theta \log(\pi(a\,|\,s,\theta))$$

$$= \sum_{s,a} b(s)p(s)\pi(a\,|\,s,\theta)\frac{\nabla_\theta \pi(a\,|\,s,\theta)}{\pi(a\,|\,s,\theta)}$$

$$= \sum_{s,a} b(s)p(s)\nabla_\theta \pi(a\,|\,s,\theta)$$

$$= 0$$

Policy gradient theorem (episodic)

Theorem

For any differentiable policy $\pi_{\boldsymbol{\theta}}(s, a)$, let d_0 be the starting distribution over states in which we begin an episode. Then, the policy gradient of $J(\boldsymbol{\theta}) = \mathbb{E}\left[G_0 \mid S_0 \sim d_0\right]$ is

$$\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) = \mathbb{E}_{\pi_{\boldsymbol{\theta}}}\left[\sum_{t=0}^{T} \gamma^t q_{\pi_{\boldsymbol{\theta}}}(S_t, A_t) \nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(A_t | S_t) \mid S_0 \sim d_0\right]$$

where

$$\begin{aligned} q_\pi(s, a) &= \mathbb{E}_\pi[G_t \mid S_t = s, A_t = a] \\ &= \mathbb{E}_\pi[R_{t+1} + \gamma q_\pi(S_{t+1}, A_{t+1}) \mid S_t = s, A_t = a] \end{aligned}$$

Episodic policy gradient theorem — proof (1/3)

- Consider trajectory $\tau = S_0, A_0, R_1, S_1, A_1, R_1, S_2, \ldots$ with return $G(\tau) = \sum_i \gamma^i R_i$

$$\nabla_\theta J_\theta(\pi) = \nabla_\theta \mathbb{E}\left[G(\tau)\right] = \nabla_\theta \sum_\tau G(\tau) p(\tau)$$

Episodic policy gradient theorem — proof (1/3)

- Consider trajectory $\tau = S_0, A_0, R_1, S_1, A_1, R_1, S_2, \ldots$ with return $G(\tau) = \sum_i \gamma^i R_i$

$$
\begin{aligned}
\nabla_\theta J_\theta(\pi) = \nabla_\theta \mathbb{E}\left[G(\tau)\right] &= \nabla_\theta \sum_\tau G(\tau) p(\tau) \\
&= \sum_\tau G(\tau) \nabla_\theta p(\tau)
\end{aligned}
$$

Episodic policy gradient theorem — proof (1/3)

- Consider trajectory $\tau = S_0, A_0, R_1, S_1, A_1, R_1, S_2, \ldots$ with return $G(\tau) = \sum_i \gamma^i R_i$

$$\nabla_\theta J_\theta(\pi) = \nabla_\theta \mathbb{E}\left[G(\tau)\right] = \nabla_\theta \sum_\tau G(\tau) p(\tau)$$

$$= \sum_\tau G(\tau) \nabla_\theta p(\tau)$$

$$= \sum_\tau G(\tau) p(\tau) \frac{\nabla_\theta p(\tau)}{p(\tau)}$$

Episodic policy gradient theorem — proof (1/3)

- Consider trajectory $\tau = S_0, A_0, R_1, S_1, A_1, R_1, S_2, \ldots$ with return $G(\tau) = \sum_i \gamma^i R_i$

$$
\begin{aligned}
\nabla_\theta J_\theta(\pi) = \nabla_\theta \mathbb{E}\left[G(\tau)\right] &= \nabla_\theta \sum_\tau G(\tau) p(\tau) \\
&= \sum_\tau G(\tau)\, \nabla_\theta p(\tau) \\
&= \sum_\tau G(\tau) p(\tau) \frac{\nabla_\theta p(\tau)}{p(\tau)} \\
&= \sum_\tau p(\tau) G(\tau)\, \nabla_\theta \log(p(\tau))
\end{aligned}
$$

Episodic policy gradient theorem — proof (1/3)

- Consider trajectory $\tau = S_0, A_0, R_1, S_1, A_1, R_1, S_2, \ldots$ with return $G(\tau) = \sum_i \gamma^i R_i$

$$\begin{aligned}
\nabla_\theta J_\theta(\pi) = \nabla_\theta \mathbb{E}\left[G(\tau)\right] &= \nabla_\theta \sum_\tau G(\tau) p(\tau) \\
&= \sum_\tau G(\tau)\, \nabla_\theta p(\tau) \\
&= \sum_\tau G(\tau) p(\tau) \frac{\nabla_\theta p(\tau)}{p(\tau)} \\
&= \sum_\tau p(\tau) G(\tau)\, \nabla_\theta \log(p(\tau)) \\
&= \mathbb{E}\left[G(\tau)\, \nabla_\theta \log(p(\tau))\right]
\end{aligned}$$

Episodic policy gradient theorem — proof (1/3)

- Consider trajectory $\tau = S_0, A_0, R_1, S_1, A_1, R_1, S_2, \ldots$ with return $G(\tau) = \sum_i \gamma^i R_i$

$$\nabla_{\boldsymbol{\theta}} J_{\boldsymbol{\theta}}(\pi) = \nabla_{\boldsymbol{\theta}} \mathbb{E}\left[G(\tau)\right] = \mathbb{E}\left[G(\tau) \nabla_{\boldsymbol{\theta}} \log p(\tau)\right]$$

Episodic policy gradient theorem — proof (2/3)

- Consider trajectory $\tau = S_0, A_0, R_1, S_1, A_1, R_1, S_2, \ldots$ with return $G(\tau) = \sum_i \gamma^i R_i$

$$\nabla_{\boldsymbol{\theta}} J_{\boldsymbol{\theta}}(\pi) = \nabla_{\boldsymbol{\theta}} \mathbb{E}\left[G(\tau)\right] = \mathbb{E}\left[G(\tau) \nabla_{\boldsymbol{\theta}} \log p(\tau)\right]$$

$$\nabla_{\boldsymbol{\theta}} \log p(\tau) =$$

Episodic policy gradient theorem — proof (2/3)

- Consider trajectory $\tau = S_0, A_0, R_1, S_1, A_1, R_1, S_2, \ldots$ with return $G(\tau) = \sum_i \gamma^i R_i$

$$\nabla_{\boldsymbol{\theta}} J_{\boldsymbol{\theta}}(\pi) = \nabla_{\boldsymbol{\theta}} \mathbb{E}\left[G(\tau)\right] = \mathbb{E}\left[G(\tau) \nabla_{\boldsymbol{\theta}} \log p(\tau)\right]$$

$$\nabla_{\boldsymbol{\theta}} \log p(\tau) = \nabla_{\boldsymbol{\theta}} \log \left[p(S_0)\pi(A_0|S_0)p(S_1|S_0, A_0)\pi(A_1|S_1) \cdots \right]$$

Episodic policy gradient theorem — proof (2/3)

- Consider trajectory $\tau = S_0, A_0, R_1, S_1, A_1, R_1, S_2, \ldots$ with return $G(\tau) = \sum_i \gamma^i R_i$

$$\nabla_{\boldsymbol{\theta}} J_{\boldsymbol{\theta}}(\pi) = \nabla_{\boldsymbol{\theta}} \mathbb{E}\left[G(\tau)\right] = \mathbb{E}\left[G(\tau) \nabla_{\boldsymbol{\theta}} \log p(\tau)\right]$$

$$\begin{aligned}
\nabla_{\boldsymbol{\theta}} \log p(\tau) &= \nabla_{\boldsymbol{\theta}} \log \left[p(S_0)\pi(A_0|S_0)p(S_1|S_0, A_0)\pi(A_1|S_1)\cdots \right] \\
&= \nabla_{\boldsymbol{\theta}} \left[\log p(S_0) + \log \pi(A_0|S_0) + \log p(S_1|S_0, A_0) + \log \pi(A_1|S_1) + \cdots \right]
\end{aligned}$$

Episodic policy gradient theorem — proof (2/3)

- Consider trajectory $\tau = S_0, A_0, R_1, S_1, A_1, R_1, S_2, \ldots$ with return $G(\tau) = \sum_i \gamma^i R_i$

$$\nabla_{\boldsymbol{\theta}} J_{\boldsymbol{\theta}}(\pi) = \nabla_{\boldsymbol{\theta}} \mathbb{E}\left[G(\tau)\right] = \mathbb{E}\left[G(\tau)\nabla_{\boldsymbol{\theta}} \log p(\tau)\right]$$

$$\begin{aligned}
\nabla_{\boldsymbol{\theta}} \log p(\tau) &= \nabla_{\boldsymbol{\theta}} \log \left[p(S_0)\pi(A_0|S_0)p(S_1|S_0, A_0)\pi(A_1|S_1) \cdots \right] \\
&= \nabla_{\boldsymbol{\theta}} \left[\log p(S_0) + \log \pi(A_0|S_0) + \log p(S_1|S_0, A_0) + \log \pi(A_1|S_1) + \cdots \right] \\
&= \nabla_{\boldsymbol{\theta}} \left[\log \pi(A_0|S_0) + \log \pi(A_1|S_1) + \cdots \right]
\end{aligned}$$

Episodic policy gradient theorem — proof (2/3)

- Consider trajectory $\tau = S_0, A_0, R_1, S_1, A_1, R_1, S_2, \ldots$ with return $G(\tau) = \sum_i \gamma^i R_i$

$$\nabla_{\boldsymbol{\theta}} J_{\boldsymbol{\theta}}(\pi) = \nabla_{\boldsymbol{\theta}} \mathbb{E}\left[G(\tau)\right] = \mathbb{E}\left[G(\tau)\nabla_{\boldsymbol{\theta}} \log p(\tau)\right]$$

$$\begin{aligned}
\nabla_{\boldsymbol{\theta}} \log p(\tau) &= \nabla_{\boldsymbol{\theta}} \log \left[p(S_0)\pi(A_0|S_0)p(S_1|S_0, A_0)\pi(A_1|S_1)\cdots\right] \\
&= \nabla_{\boldsymbol{\theta}}\left[\log p(S_0) + \log \pi(A_0|S_0) + \log p(S_1|S_0, A_0) + \log \pi(A_1|S_1) + \cdots\right] \\
&= \nabla_{\boldsymbol{\theta}}\left[\log \pi(A_0|S_0) + \log \pi(A_1|S_1) + \cdots\right]
\end{aligned}$$

So:

$$\nabla_{\boldsymbol{\theta}} J_{\boldsymbol{\theta}}(\pi) = \mathbb{E}_\pi[G(\tau)\nabla_{\boldsymbol{\theta}} \sum_{t=0}^{T} \log \pi(A_t|S_t)]$$

Episodic policy gradient theorem — proof (3/3)

$$\nabla_{\boldsymbol{\theta}} J_{\boldsymbol{\theta}}(\pi) = \mathbb{E}_{\pi}[G(\tau) \sum_{t=0}^{T} \nabla_{\boldsymbol{\theta}} \log \pi(A_t | S_t)]$$

Episodic policy gradient theorem — proof (3/3)

$$
\begin{aligned}
\nabla_{\boldsymbol{\theta}} J_{\boldsymbol{\theta}}(\pi) &= \mathbb{E}_{\pi}[G(\tau) \sum_{t=0}^{T} \nabla_{\boldsymbol{\theta}} \log \pi(A_t | S_t)] \\
&= \mathbb{E}_{\pi}[\sum_{t=0}^{T} G(\tau) \nabla_{\boldsymbol{\theta}} \log \pi(A_t | S_t)]
\end{aligned}
$$

Episodic policy gradient theorem — proof (3/3)

$$\nabla_{\boldsymbol{\theta}} J_{\boldsymbol{\theta}}(\pi) = \mathbb{E}_{\pi}[G(\tau) \sum_{t=0}^{T} \nabla_{\boldsymbol{\theta}} \log \pi(A_t | S_t)]$$

$$= \mathbb{E}_{\pi}[\sum_{t=0}^{T} G(\tau) \nabla_{\boldsymbol{\theta}} \log \pi(A_t | S_t)]$$

$$= \mathbb{E}_{\pi}[\sum_{t=0}^{T} \left(\sum_{k=0}^{T} \gamma^k R_{k+1} \right) \nabla_{\boldsymbol{\theta}} \log \pi(A_t | S_t)]$$

Episodic policy gradient theorem — proof (3/3)

$$\nabla_{\boldsymbol{\theta}} J_{\boldsymbol{\theta}}(\pi) = \mathbb{E}_\pi[G(\tau)\sum_{t=0}^{T} \nabla_{\boldsymbol{\theta}} \log \pi(A_t|S_t)]$$

$$= \mathbb{E}_\pi[\sum_{t=0}^{T} G(\tau) \nabla_{\boldsymbol{\theta}} \log \pi(A_t|S_t)]$$

$$= \mathbb{E}_\pi[\sum_{t=0}^{T} \left(\sum_{k=0}^{T} \gamma^k R_{k+1}\right) \nabla_{\boldsymbol{\theta}} \log \pi(A_t|S_t)]$$

$$= \mathbb{E}_\pi[\sum_{t=0}^{T} \left(\sum_{k=\boldsymbol{t}}^{T} \gamma^k R_{k+1}\right) \nabla_{\boldsymbol{\theta}} \log \pi(A_t|S_t)]$$

Why?

Episodic policy gradient theorem — proof (3/3)

$$\nabla_{\boldsymbol{\theta}} J_{\boldsymbol{\theta}}(\pi) = \mathbb{E}_\pi[G(\tau) \sum_{t=0}^{T} \nabla_{\boldsymbol{\theta}} \log \pi(A_t|S_t)]$$

$$= \mathbb{E}_\pi[\sum_{t=0}^{T} G(\tau) \nabla_{\boldsymbol{\theta}} \log \pi(A_t|S_t)]$$

$$= \mathbb{E}_\pi[\sum_{t=0}^{T} \left(\sum_{k=0}^{T} \gamma^k R_{k+1} \right) \nabla_{\boldsymbol{\theta}} \log \pi(A_t|S_t)]$$

$$= \mathbb{E}_\pi[\sum_{t=0}^{T} \left(\sum_{k=\boldsymbol{t}}^{T} \gamma^k R_{k+1} \right) \nabla_{\boldsymbol{\theta}} \log \pi(A_t|S_t)]$$

Why?

Important "Trick" / Identity

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

Episodic policy gradient theorem — proof (3/3)

$$
\begin{aligned}
\nabla_{\boldsymbol{\theta}} J_{\boldsymbol{\theta}}(\pi) &= \mathbb{E}_\pi[G(\tau) \sum_{t=0}^{T} \nabla_{\boldsymbol{\theta}} \log \pi(A_t | S_t)] \\
&= \mathbb{E}_\pi[\sum_{t=0}^{T} G(\tau) \nabla_{\boldsymbol{\theta}} \log \pi(A_t | S_t)] \\
&= \mathbb{E}_\pi[\sum_{t=0}^{T} \left(\sum_{k=0}^{T} \gamma^k R_{k+1} \right) \nabla_{\boldsymbol{\theta}} \log \pi(A_t | S_t)] \\
&= \mathbb{E}_\pi[\sum_{t=0}^{T} \left(\sum_{k=\boldsymbol{t}}^{T} \gamma^k R_{k+1} \right) \nabla_{\boldsymbol{\theta}} \log \pi(A_t | S_t)]
\end{aligned}
$$

Episodic policy gradient theorem — proof (3/3)

$$
\begin{aligned}
\nabla_{\boldsymbol{\theta}} J_{\boldsymbol{\theta}}(\pi) &= \mathbb{E}_\pi[G(\tau) \sum_{t=0}^{T} \nabla_{\boldsymbol{\theta}} \log \pi(A_t | S_t)] \\
&= \mathbb{E}_\pi[\sum_{t=0}^{T} G(\tau) \nabla_{\boldsymbol{\theta}} \log \pi(A_t | S_t)] \\
&= \mathbb{E}_\pi[\sum_{t=0}^{T} \left(\sum_{k=0}^{T} \gamma^k R_{k+1} \right) \nabla_{\boldsymbol{\theta}} \log \pi(A_t | S_t)] \\
&= \mathbb{E}_\pi[\sum_{t=0}^{T} \left(\sum_{k=\boldsymbol{t}}^{T} \gamma^k R_{k+1} \right) \nabla_{\boldsymbol{\theta}} \log \pi(A_t | S_t)] \\
&= \mathbb{E}_\pi[\sum_{t=0}^{T} \left(\gamma^t \sum_{k=t}^{T} \gamma^{k-t} R_{k+1} \right) \nabla_{\boldsymbol{\theta}} \log \pi(A_t | S_t)]
\end{aligned}
$$

Episodic policy gradient theorem — proof (3/3)

$$
\begin{aligned}
\nabla_{\boldsymbol{\theta}} J_{\boldsymbol{\theta}}(\pi) &= \mathbb{E}_\pi[G(\tau)\sum_{t=0}^{T}\nabla_{\boldsymbol{\theta}}\log\pi(A_t|S_t)] \\
&= \mathbb{E}_\pi[\sum_{t=0}^{T} G(\tau)\nabla_{\boldsymbol{\theta}}\log\pi(A_t|S_t)] \\
&= \mathbb{E}_\pi[\sum_{t=0}^{T}\left(\sum_{k=0}^{T}\gamma^k R_{k+1}\right)\nabla_{\boldsymbol{\theta}}\log\pi(A_t|S_t)] \\
&= \mathbb{E}_\pi[\sum_{t=0}^{T}\left(\sum_{k=t}^{T}\gamma^k R_{k+1}\right)\nabla_{\boldsymbol{\theta}}\log\pi(A_t|S_t)] \\
&= \mathbb{E}_\pi[\sum_{t=0}^{T}\left(\gamma^t\sum_{k=t}^{T}\gamma^{k-t} R_{k+1}\right)\nabla_{\boldsymbol{\theta}}\log\pi(A_t|S_t)] \\
&= \mathbb{E}_\pi[\sum_{t=0}^{T}\left(\gamma^t G_t\right)\nabla_{\boldsymbol{\theta}}\log\pi(A_t|S_t)] \qquad = \mathbb{E}_\pi[\sum_{t=0}^{T}\gamma^t q_\pi(S_t, A_t)\nabla_{\boldsymbol{\theta}}\log\pi(A_t|S_t)]
\end{aligned}
$$

Episodic policy gradients algorithm

$$\nabla_{\boldsymbol{\theta}} J_{\boldsymbol{\theta}}(\pi) = \mathbb{E}_\pi[\sum_{t=0}^{T} \gamma^t q_\pi(S_t, A_t) \nabla_{\boldsymbol{\theta}} \log \pi(A_t|S_t)]$$

- We can sample this, given a whole episode
- Typically, people pull out the sum, and split up this into separate gradients, e.g.,

$$\Delta\boldsymbol{\theta}_t = \gamma^t G_t \nabla_{\boldsymbol{\theta}} \log \pi(A_t|S_t)$$

  such that $\mathbb{E}_\pi[\sum_t \Delta\boldsymbol{\theta}_t] = \nabla_{\boldsymbol{\theta}} J_{\boldsymbol{\theta}}(\pi)$
- Typically, people ignore the γ^t term, use $\Delta\boldsymbol{\theta}_t = G_t \nabla_{\boldsymbol{\theta}} \log \pi(A_t|S_t)$
- This is actually okay-ish — we just partially pretend on each step that we could have started an episode in that state instead. Or if we use γ=1, this is also ok.
  (alternatively, view it as a slightly biased gradient)

REINFORCE (Monte-Carlo)

$$\nabla_{\boldsymbol{\theta}} J_{\boldsymbol{\theta}}(\pi) = \mathbb{E}_{\pi}[\sum_{t=0}^{T} (\gamma^t G_t) \nabla_{\boldsymbol{\theta}} \log \pi(A_t | S_t)]$$

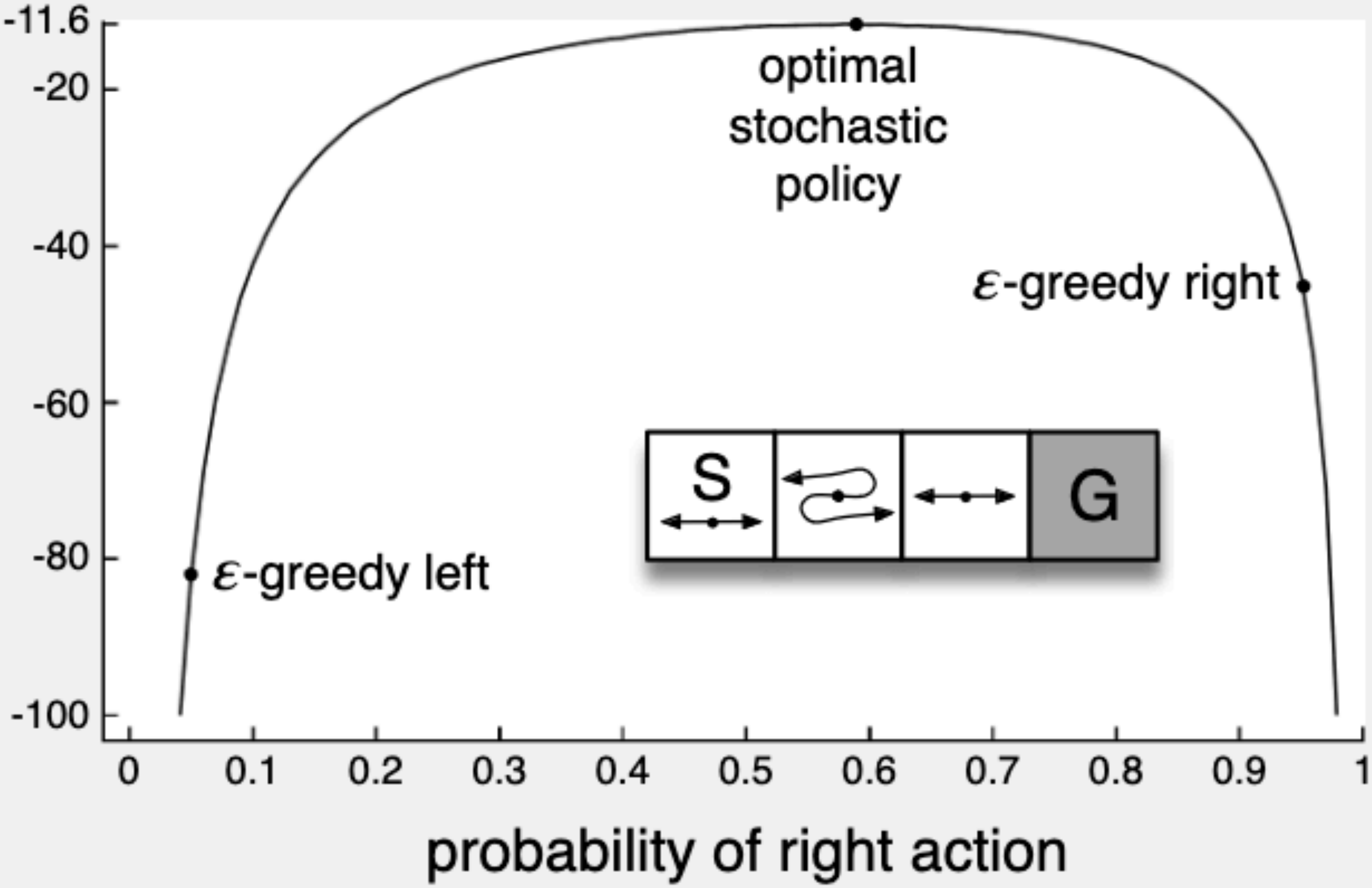
REINFORCE: Monte-Carlo Policy-Gradient Control (episodic) for π_*

Input: a differentiable policy parameterization $\pi(a|s, \boldsymbol{\theta})$
Algorithm parameter: step size $\alpha > 0$
Initialize policy parameter $\boldsymbol{\theta} \in \mathbb{R}^{d'}$ (e.g., to $\mathbf{0}$)

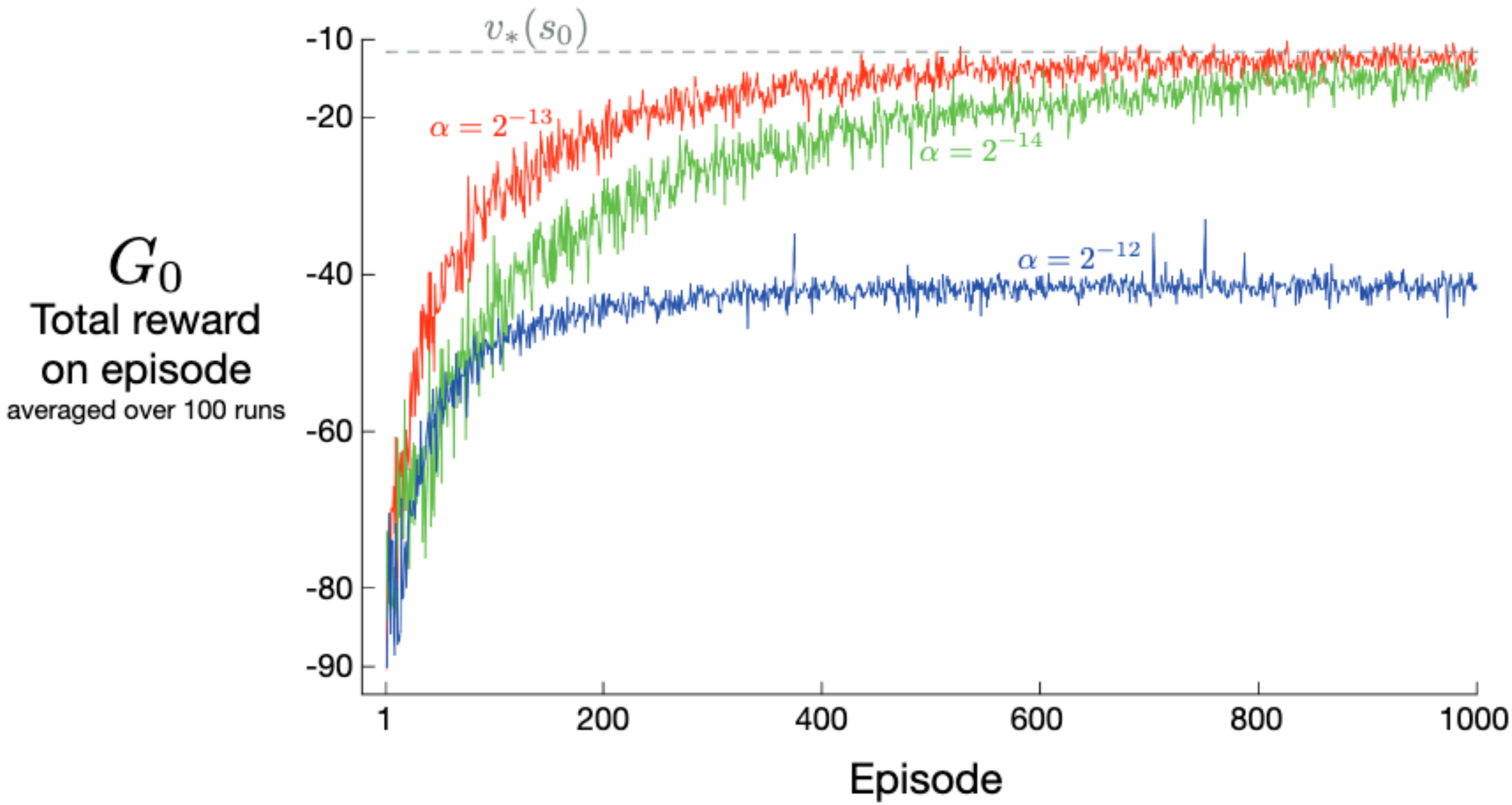
Loop forever (for each episode):
    Generate an episode $S_0, A_0, R_1, \ldots, S_{T-1}, A_{T-1}, R_T$, following $\pi(\cdot|\cdot, \boldsymbol{\theta})$
    Loop for each step of the episode $t = 0, 1, \ldots, T-1$:
        $G \leftarrow \sum_{k=t+1}^{T} \gamma^{k-t-1} R_k$ $\qquad (G_t)$
        $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \alpha \gamma^t G \nabla \ln \pi(A_t | S_t, \boldsymbol{\theta})$

Example: REINFORCE

$$J(\boldsymbol{\theta}) = v_{\pi_{\boldsymbol{\theta}}}(\mathsf{S})$$



Example: REINFORCE



Improvements to REINFORCE

$$\nabla_{\boldsymbol{\theta}} J_{\boldsymbol{\theta}}(\pi) = \mathbb{E}_{\pi}[\sum_{t=0}^{T} \left(\gamma^{t} G_{t}\right) \nabla_{\boldsymbol{\theta}} \log \pi(A_{t} | S_{t})]$$

- Can we use our "trick" $\mathbb{E}\left(b(s)\nabla_{\theta}\log(\pi(a\,|\,s,\theta))\right) = 0$ to improve REINFORCE?

Improvements to REINFORCE

$$\nabla_{\boldsymbol{\theta}} J_{\boldsymbol{\theta}}(\pi) = \mathbb{E}_{\pi}[\sum_{t=0}^{T} \left(\gamma^{t} G_{t}\right) \nabla_{\boldsymbol{\theta}} \log \pi(A_{t} | S_{t})]$$

- Can we use our "trick" $\mathbb{E}\left(b(s)\nabla_\theta \log(\pi(a\,|\,s,\theta))\right) = 0$ to improve REINFORCE?

$$\nabla_\theta J_\theta(\pi) = \mathbb{E}\left[\sum_{t=0}^{T} \gamma^t \left(G_t - \bar{G}\right) \nabla_\theta \log(\pi)\right]$$

Improvements to REINFORCE

$$\nabla_{\boldsymbol{\theta}} J_{\boldsymbol{\theta}}(\pi) = \mathbb{E}_{\pi}[\sum_{t=0}^{T} \left(\gamma^t G_t\right) \nabla_{\boldsymbol{\theta}} \log \pi(A_t | S_t)]$$

- Can we use our "trick" $\mathbb{E}\left(b(s)\nabla_\theta \log(\pi(a\,|\,s,\theta))\right) = 0$ to improve REINFORCE?

$$\nabla_\theta J_\theta(\pi) = \mathbb{E}\left[\sum_{t=0}^{T} \gamma^t \left(G_t - \bar{G}\right) \nabla_\theta \log(\pi)\right]$$

$$\nabla_\theta J_\theta(\pi) = \mathbb{E}\left[\sum_{t=0}^{T} \gamma^t \left(q_\pi(S_t, A_t) - v_\pi(S_t)\right) \nabla_\theta \log(\pi)\right]$$

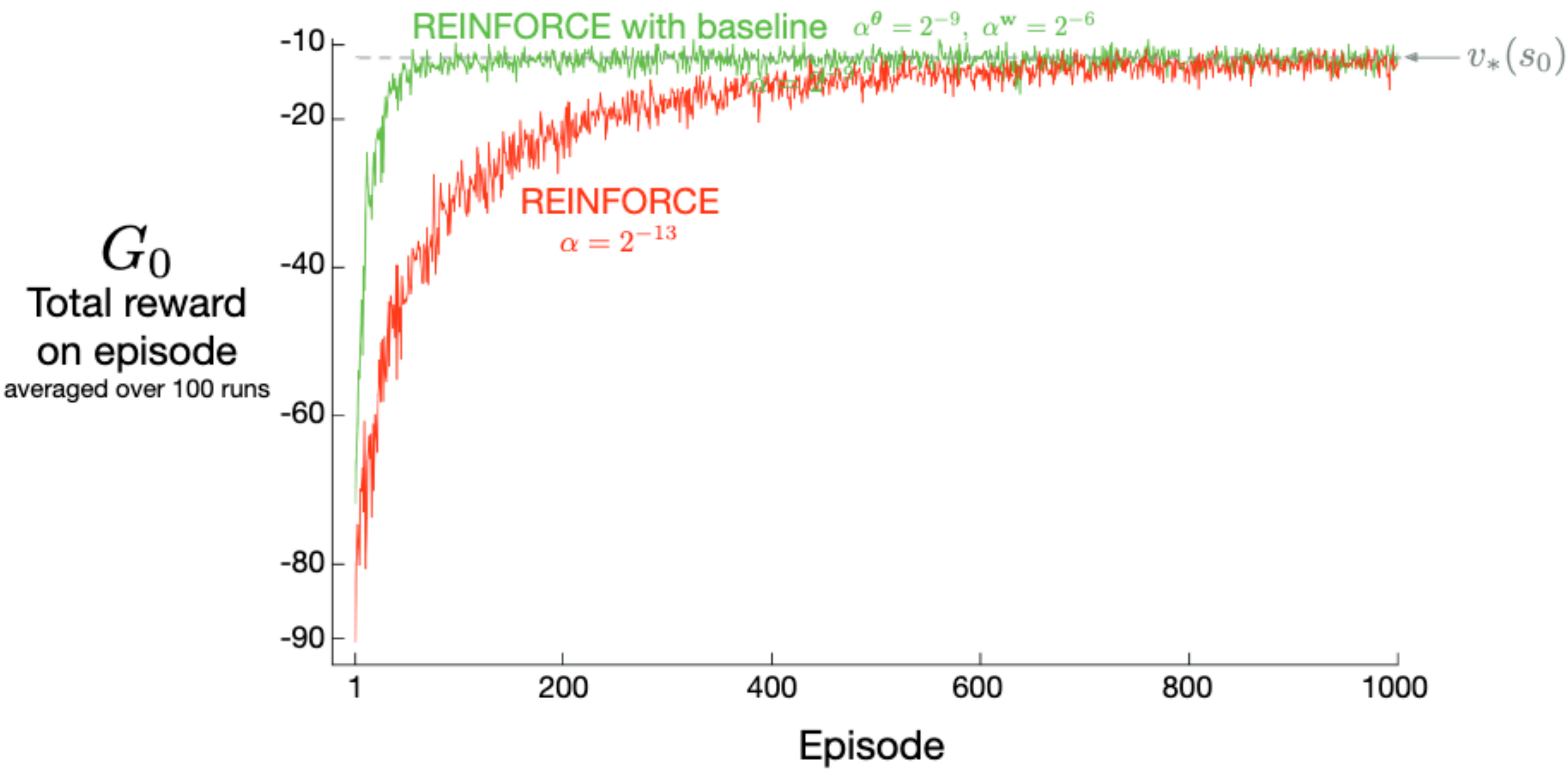
REINFORCE with baseline:

REINFORCE with Baseline (episodic), for estimating $\pi_{\boldsymbol{\theta}} \approx \pi_*$

Input: a differentiable policy parameterization $\pi(a|s, \boldsymbol{\theta})$
Input: a differentiable state-value function parameterization $\hat{v}(s,\mathbf{w})$
Algorithm parameters: step sizes $\alpha^{\boldsymbol{\theta}} > 0$, $\alpha^{\mathbf{w}} > 0$
Initialize policy parameter $\boldsymbol{\theta} \in \mathbb{R}^{d'}$ and state-value weights $\mathbf{w} \in \mathbb{R}^{d}$ (e.g., to $\mathbf{0}$)

Loop forever (for each episode):
    Generate an episode $S_0, A_0, R_1, \ldots, S_{T-1}, A_{T-1}, R_T$, following $\pi(\cdot|\cdot, \boldsymbol{\theta})$
    Loop for each step of the episode $t = 0, 1, \ldots, T-1$:
        $G \leftarrow \sum_{k=t+1}^{T} \gamma^{k-t-1} R_k$ $\quad (G_t)$
        $\delta \leftarrow G - \hat{v}(S_t,\mathbf{w})$
        $\mathbf{w} \leftarrow \mathbf{w} + \alpha^{\mathbf{w}} \delta \nabla \hat{v}(S_t,\mathbf{w})$
        $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \alpha^{\boldsymbol{\theta}} \gamma^t \delta \nabla \ln \pi(A_t|S_t, \boldsymbol{\theta})$

REINFORCE with baseline:



Actor-Critic Algorithms

- ACTOR: policy π
- CRITIC: value fct V (or Q)