Policy-gradient methods

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
   - update by gradient ascent in performance
   - includes *actor-critic methods*, which learn *both* value and policy parameters

The old approach:
Action-value methods

- The *value of an action in a state given a policy* is the expected future reward starting from the state taking that first action, then following the policy thereafter

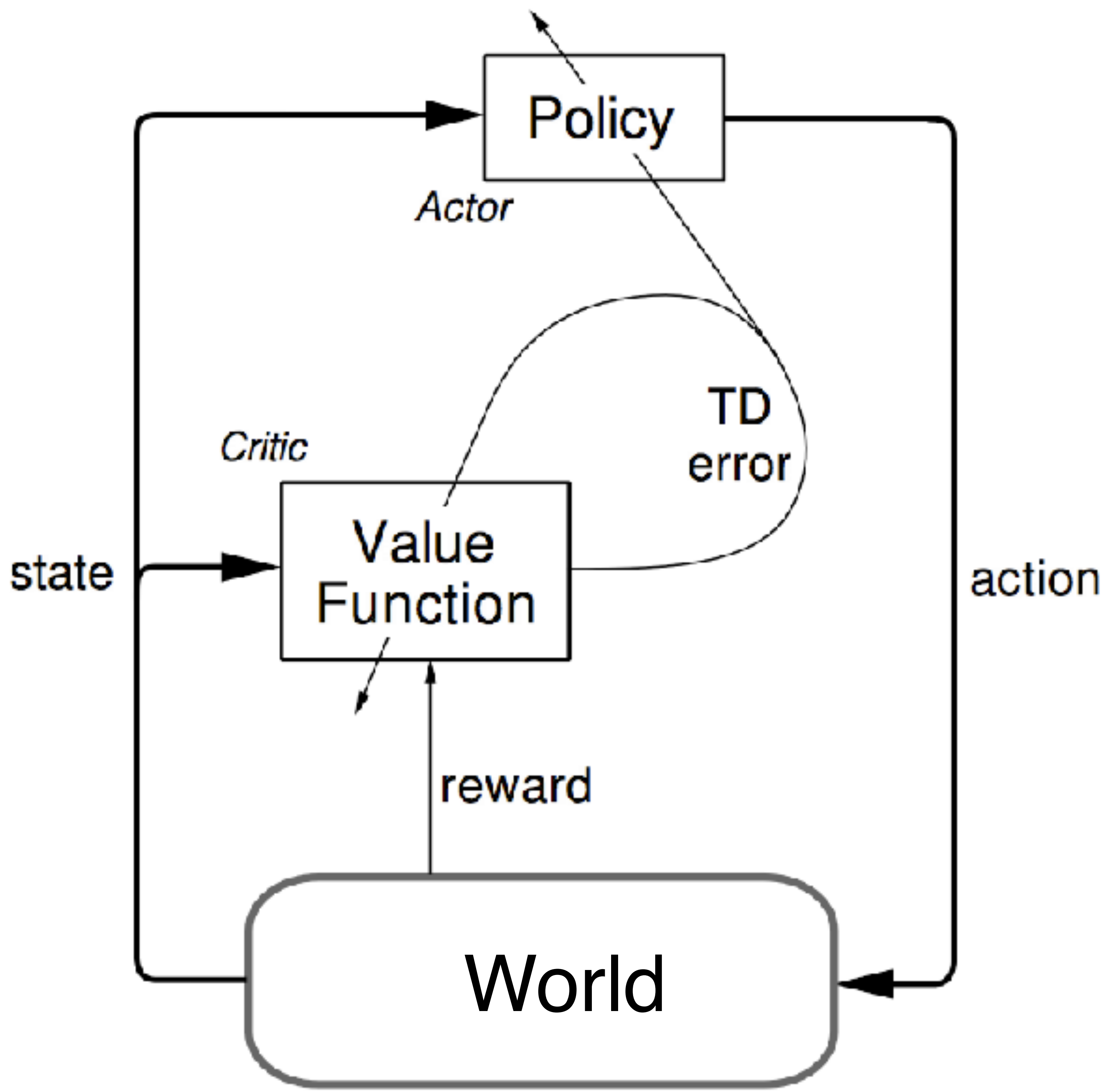
$$q_\pi(s,a) = \mathbb{E}\left[\sum_{t=1}^{\infty} \gamma^{t-1} R_t \middle| S_0 = s, A_0 = a\right]$$

- Policy: pick the max most of the time

$$A_t = \arg\max_a \hat{Q}_t(S_t, a)$$

but sometimes pick at random (ε-greedy)

Actor-critic architecture



Why approximate policies rather than values?

- In many problems, the policy is simpler to approximate than the value function
- In many 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)
- To better relate to biology

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

We first saw this in Chapter 2, with the

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

Gradient-bandit algorithms on the 10-armed testbed

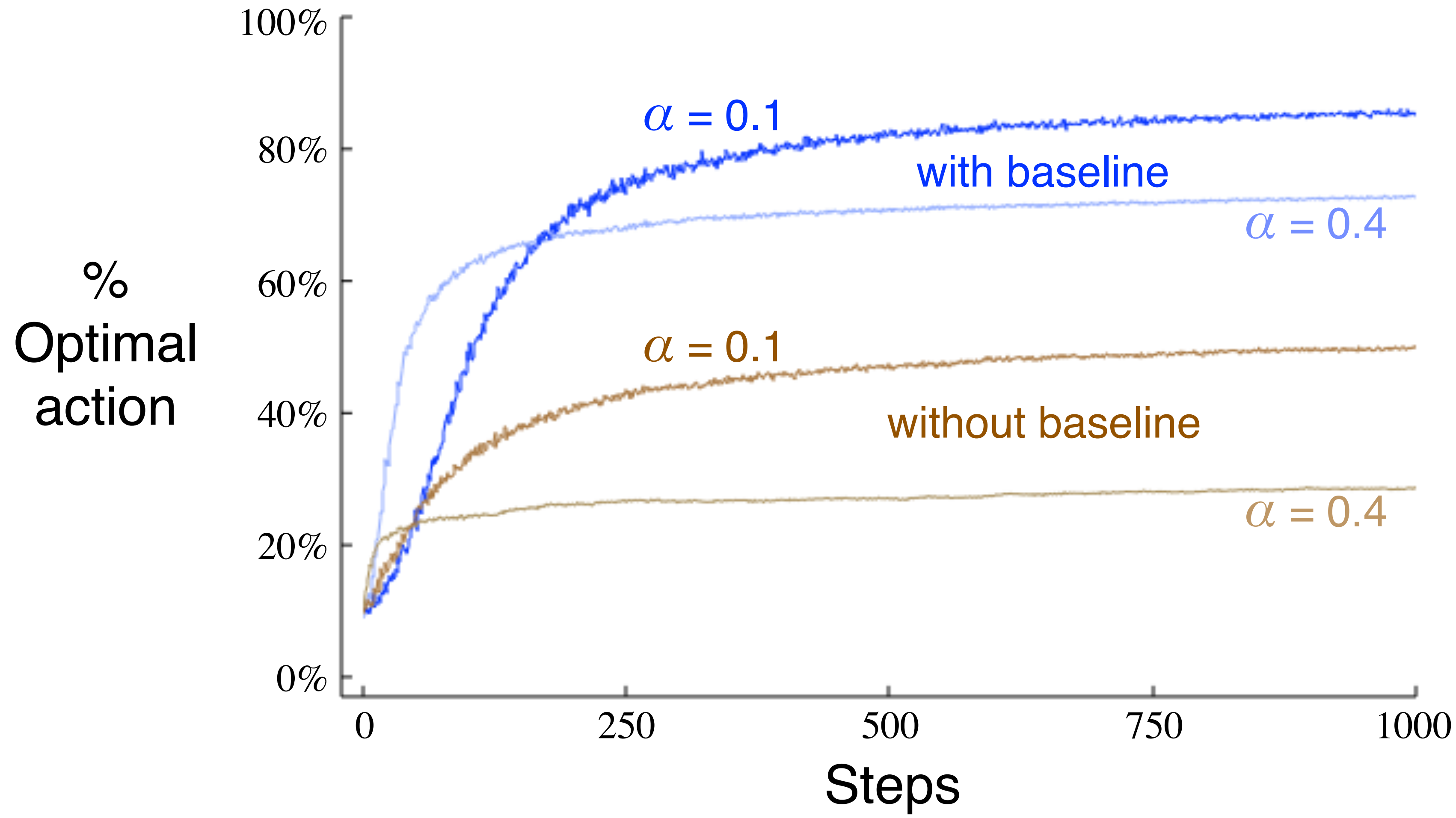


Figure 2.6: Average performance of the gradient-bandit algorithm with and without a reward baseline on the 10-armed testbed when the $q_*(a)$ are chosen to be near $+4$ rather than near zero.

eg, linear-exponential policies (discrete actions)

- The "preference" for action a in state s is linear in θ and a state-action feature vector $\phi(s,a)$

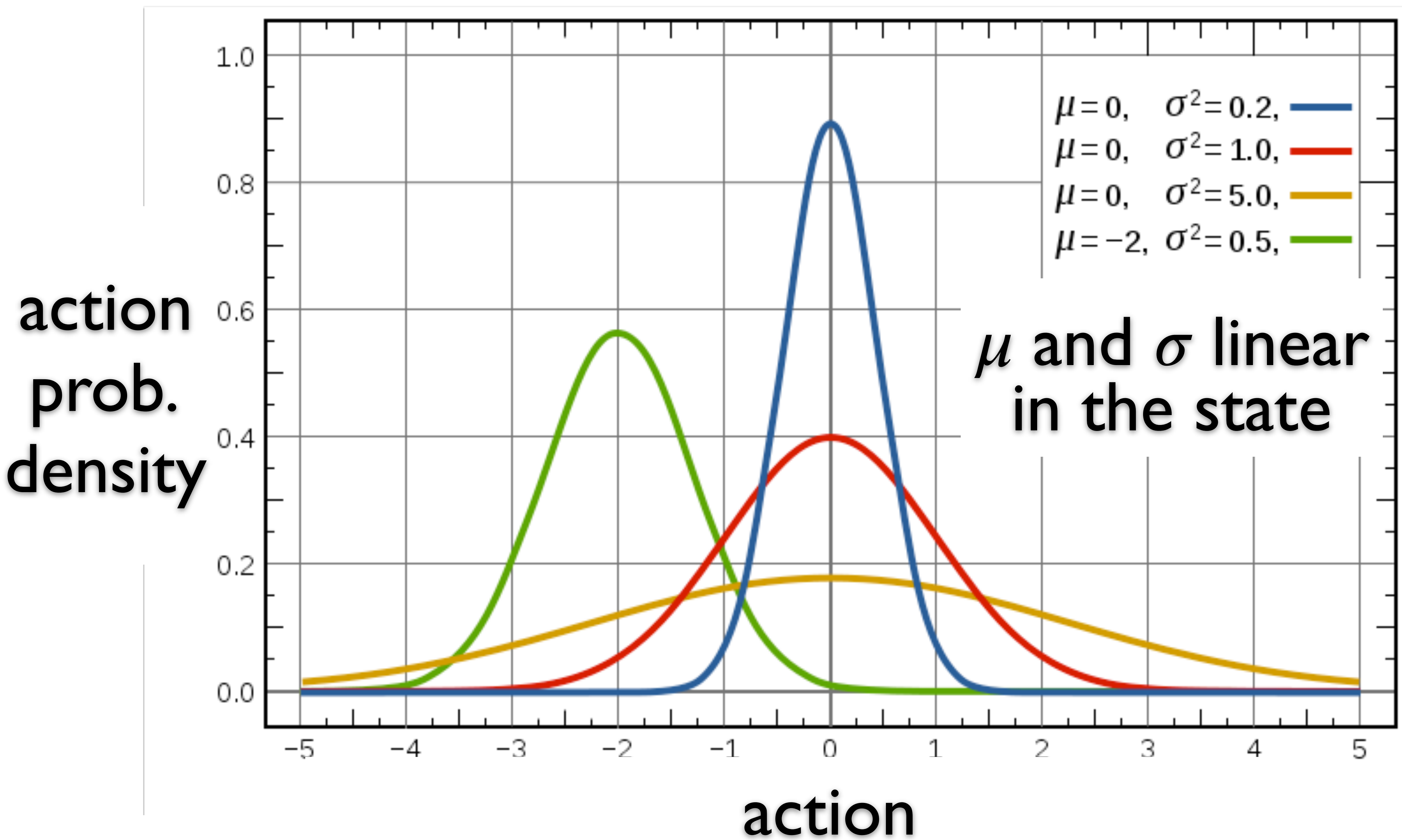
- The probability of action a in state s is exponential in its preference

$$\pi(a|s,\boldsymbol{\theta}) \doteq \frac{\exp(\boldsymbol{\theta}^\top \boldsymbol{\phi}(s,a))}{\sum_b \exp(\boldsymbol{\theta}^\top \boldsymbol{\phi}(s,b))}$$

- Corresponding *eligibility function:*

$$\frac{\nabla \pi(a|s,\boldsymbol{\theta})}{\pi(a|s,\boldsymbol{\theta})} = \boldsymbol{\phi}(s,a) - \sum_b \pi(b|s,\boldsymbol{\theta})\boldsymbol{\phi}(s,b)$$

eg, linear-gaussian policies (continuous actions)



eg, linear-gaussian policies (continuous actions)

- The mean and std. dev. for the action taken in state s are linear and linear-exponential in

$$\boldsymbol{\theta} \doteq (\boldsymbol{\theta}_\mu^\top; \boldsymbol{\theta}_\sigma^\top)^\top \qquad \mu(s) \doteq \boldsymbol{\theta}_\mu^\top \boldsymbol{\phi}(s) \qquad \sigma(s) \doteq \exp(\boldsymbol{\theta}_\sigma^\top \boldsymbol{\phi}(s))$$

- The probability density function for the action taken in state s is gaussian

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

Gaussian eligibility functions

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

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

Policy-gradient setup

Given a policy parameterization:

$$\pi(a|s,\boldsymbol{\theta}) \qquad\qquad \frac{\nabla_{\boldsymbol{\theta}}\pi(a|s,\boldsymbol{\theta})}{\pi(a|s,\boldsymbol{\theta})} = \nabla_{\boldsymbol{\theta}}\log\pi(a|s,\boldsymbol{\theta})$$

And objective:

$$\eta(\boldsymbol{\theta}) \doteq v_{\pi_{\boldsymbol{\theta}}}(S_0) \text{ (or average reward)}$$

Approximate stochastic gradient ascent:

$$\boldsymbol{\theta}_{t+1} \doteq \boldsymbol{\theta}_t + \alpha\widehat{\nabla\eta(\boldsymbol{\theta}_t)}$$

Typically, based on the Policy-Gradient Theorem:

$$\nabla\eta(\boldsymbol{\theta}) = \sum_s d_\pi(s)\sum_a q_\pi(s,a)\nabla_{\boldsymbol{\theta}}\pi(a|s,\boldsymbol{\theta})$$

Proof of the Policy-Gradient Theorem
(from the 2nd Edition)

$$
\begin{aligned}
\nabla v_\pi(s) &= \nabla \left[\sum_a \pi(a|s) q_\pi(s,a) \right], \quad \forall s \in \mathcal{S} && \text{(Exercise 3.11)} \\
&= \sum_a \Big[\nabla \pi(a|s) q_\pi(s,a) + \pi(a|s) \nabla q_\pi(s,a) \Big] && \text{(product rule)} \\
&= \sum_a \Big[\nabla \pi(a|s) q_\pi(s,a) + \pi(a|s) \nabla \sum_{s',r} p(s',r|s,a) \big(r + \gamma v_\pi(s')\big) \Big] \\
& && \text{(Exercise 3.12 and Equation 3.8)} \\
&= \sum_a \Big[\nabla \pi(a|s) q_\pi(s,a) + \pi(a|s) \sum_{s'} \gamma p(s'|s,a) \nabla v_\pi(s') \Big] \\
& && \text{(Eq. 3.10)} \\
&= \sum_a \Big[\nabla \pi(a|s) q_\pi(s,a) + \pi(a|s) \sum_{s'} \gamma p(s'|s,a) && \text{(unrolling)} \\
&\qquad \sum_{a'} \big[\nabla \pi(a'|s') q_\pi(s',a') + \pi(a'|s') \sum_{s''} \gamma p(s''|s',a') \nabla v_\pi(s'') \big] \Big] \\
&= \sum_{x \in \mathcal{S}} \sum_{k=0}^{\infty} \gamma^k \Pr(s \to x, k, \pi) \sum_a \nabla \pi(a|x) q_\pi(x,a),
\end{aligned}
$$

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

$$
\begin{aligned}
\nabla \eta(\boldsymbol{\theta}) &= \nabla v_\pi(s_0) \\
&= \sum_s \sum_{k=0}^{\infty} \gamma^k \Pr(s_0 \to s, k, \pi) \sum_a \nabla \pi(a|s) q_\pi(s,a) \\
&= \sum_s d_\pi(s) \sum_a \nabla \pi(a|s) q_\pi(s,a). \qquad \text{Q.E.D.}
\end{aligned}
$$

Deriving REINFORCE from the PGT

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

Thus

$$
\boldsymbol{\theta}_{t+1} \triangleq \boldsymbol{\theta}_t + \alpha \widehat{\nabla \eta(\boldsymbol{\theta}_t)} \triangleq \boldsymbol{\theta}_t + \alpha \gamma^t G_t \frac{\nabla_{\boldsymbol{\theta}} \pi(A_t|S_t,\boldsymbol{\theta})}{\pi(A_t|S_t,\boldsymbol{\theta})}
$$

REINFORCE with baseline

Policy-gradient theorem with baseline:

any function of state, not action

$$\nabla \eta(\boldsymbol{\theta}) = \sum_s d_\pi(s) \sum_a q_\pi(s,a) \nabla_{\boldsymbol{\theta}} \pi(a|s,\boldsymbol{\theta})$$

$$= \sum_s d_\pi(s) \sum_a \Big(q_\pi(s,a) - b(s) \Big) \nabla_{\boldsymbol{\theta}} \pi(a|s,\boldsymbol{\theta})$$

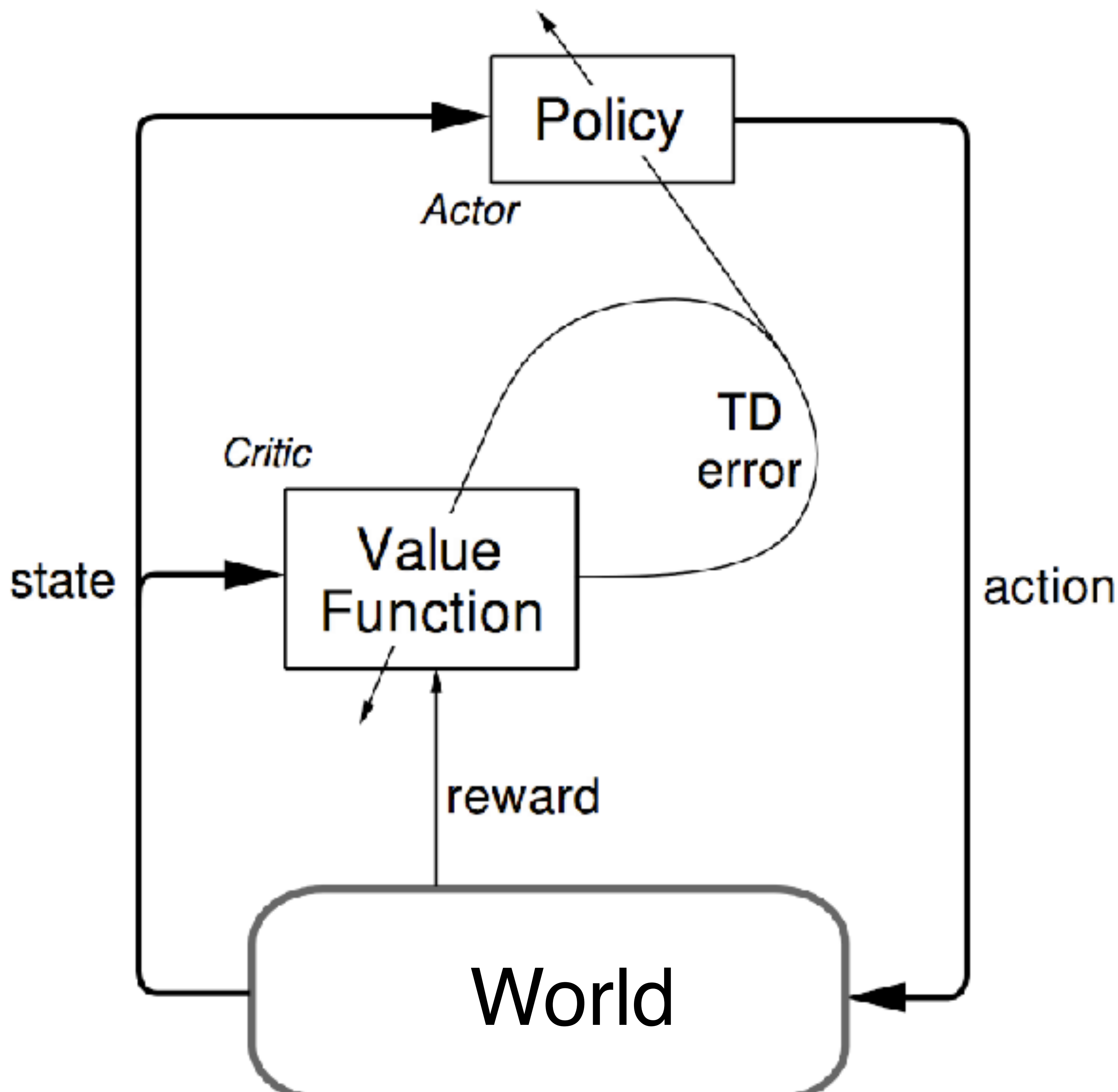
Because

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

Thus

$$\boldsymbol{\theta}_{t+1} \triangleq \boldsymbol{\theta}_t + \alpha \gamma^t \Big(G_t - b(S_t) \Big) \frac{\nabla_{\boldsymbol{\theta}} \pi(A_t|S_t,\boldsymbol{\theta})}{\pi(A_t|S_t,\boldsymbol{\theta})} \qquad \text{e.g., } b(s) = \hat{v}(s,\mathbf{w})$$

Actor-critic architecture



Actor-Critic methods

REINFORCE with baseline:

$$\boldsymbol{\theta}_{t+1} \triangleq \boldsymbol{\theta}_t + \overset{\gamma^t}{\underset{\wedge}{\alpha}}\Big(G_t - b(S_t)\Big)\frac{\nabla_{\boldsymbol{\theta}}\pi(A_t|S_t,\boldsymbol{\theta})}{\pi(A_t|S_t,\boldsymbol{\theta})}$$

Actor-Critic method:

$$\begin{aligned}\boldsymbol{\theta}_{t+1} &\triangleq \boldsymbol{\theta}_t + \overset{\gamma^t}{\underset{\wedge}{\alpha}}\Big(G_t^{(1)} - \hat{v}(S_t,\mathbf{w})\Big)\frac{\nabla_{\boldsymbol{\theta}}\pi(A_t|S_t,\boldsymbol{\theta})}{\pi(A_t|S_t,\boldsymbol{\theta})}\\ &= \boldsymbol{\theta}_t + \overset{\gamma^t}{\underset{\wedge}{\alpha}}\Big(R_{t+1} \overset{-\bar{R}_t}{\underset{\wedge}{\vphantom{x}}} + \gamma\hat{v}(S_{t+1},\mathbf{w}) - \hat{v}(S_t,\mathbf{w})\Big)\frac{\nabla_{\boldsymbol{\theta}}\pi(A_t|S_t,\boldsymbol{\theta})}{\pi(A_t|S_t,\boldsymbol{\theta})}\end{aligned}$$

We should never discount when approximating policies!

γ is ok it there is a start state/distribution

Average reward setting

- All rewards are compared to the average reward

$$q_\pi(s,a) = \mathbb{E}\left[\sum_{t=1}^{\infty} R_t - \bar{r}(\pi) \middle| S_0 = s, A_0 = a\right]$$

- where

$$\bar{r}(\pi) = \lim_{t\to\infty} \frac{1}{t}\mathbb{E}\left[R_1 + R_2 + \cdots + R_t \mid A_{0:t-1} \sim \pi\right]$$

- and we learn an approximation

$$\bar{R}_t \approx \bar{r}(\pi_t)$$

The average-reward setting

- Maximize the reward rate (reward per step):

$$r(\pi) \doteq \lim_{n\to\infty} \frac{1}{n}\sum_{t=1}^{n} \mathbb{E}_\pi[R_t] = \sum_s d_\pi(s) \sum_a \pi(a|s) \sum_{s',r} p(s',r|s,a) r$$

where $d_\pi(s) \doteq \lim_{t\to\infty} \mathbb{P}\{S_t = s\}$

- Learn to approximate $r(\pi)$ and new "differential" values, in which all rewards are compared to the reward rate:

$$\tilde{v}_\pi(s) = \sum_{k=1}^{\infty} \mathbb{E}_\pi[R_{t+k} - r(\pi) \mid S_t = s]$$

$$\tilde{q}_\pi(s,a) = \sum_{k=1}^{\infty} \mathbb{E}_\pi[R_{t+k} - r(\pi) \mid S_t = s, A_t = a]$$

Average-reward Q-learning (R-learning)

Initialize $\bar{R}$ and $Q(s, a)$, for all s, a, arbitrarily
Repeat forever:
    $S \leftarrow$ current state
    Choose action A in S using behavior policy (e.g., ϵ-greedy)
    Take action A, observe R, S'
    $\delta \leftarrow R - \bar{R} + \max_a Q(S', a) - Q(S, A)$
    $Q(S, A) \leftarrow Q(S, A) + \alpha\delta$
    If $Q(S, A) = \max_a Q(S, a)$, then:
        $\bar{R} \leftarrow \bar{R} + \beta\delta$

Policy-gradient setup

parameterized policies $\pi(a|s,\boldsymbol{\theta}) \doteq \Pr\{A_t = a \mid S_t = s\}$

average-reward objective $r(\pi) \doteq \lim_{n\to\infty} \frac{1}{n}\sum_{t=1}^{n} \mathbb{E}_\pi[R_t] = \sum_s d_\pi(s) \sum_a \pi(a|s) \sum_{s',r} p(s',r|s,a)r$

steady-state distribution $d_\pi \doteq \lim_{t\to\infty} \Pr\{S_t = s\}$

differential state-value fn $\tilde{v}_\pi(s) \doteq \sum_{k=1}^{\infty} \mathbb{E}_\pi[R_{t+k} - r(\pi) \mid S_t = s]$

differential action-value fn $\tilde{q}_\pi(s,a) \doteq \sum_{k=1}^{\infty} \mathbb{E}_\pi[R_{t+k} - r(\pi) \mid S_t = s, A_t = a]$

stochastic gradient ascent $\Delta\boldsymbol{\theta}_t \approx \alpha \frac{\partial r(\pi)}{\partial \boldsymbol{\theta}} \doteq \alpha \nabla r(\pi)$

stochastic gradient ascent $$\Delta\boldsymbol{\theta}_t \approx \alpha\frac{\partial r(\pi)}{\partial\boldsymbol{\theta}} \doteq \alpha\nabla r(\pi)$$

stochastic gradient ascent $$\Delta\boldsymbol{\theta}_t \approx \alpha\frac{\partial r(\pi)}{\partial\boldsymbol{\theta}} \doteq \alpha\nabla r(\pi)$$

policy-gradient theorem $$\nabla r(\pi) = \sum_s d_\pi(s)\sum_a \tilde{q}_\pi(s,a)\nabla\pi(a|s,\boldsymbol{\theta})$$

stochastic gradient ascent $$\Delta \boldsymbol{\theta}_t \approx \alpha \frac{\partial r(\pi)}{\partial \boldsymbol{\theta}} \doteq \alpha \nabla r(\pi)$$

policy-gradient theorem $$\nabla r(\pi) = \sum_s d_\pi(s) \sum_a \tilde{q}_\pi(s,a) \nabla \pi(a|s,\boldsymbol{\theta})$$

$$= \mathbb{E}\left[\Big(\tilde{q}_\pi(S_t, A_t) - v(S_t) \Big) \frac{\nabla \pi(A_t|S_t,\boldsymbol{\theta})}{\pi(A_t|S_t)} \;\middle|\; S_t \sim d_\pi, A_t \sim \pi(\cdot|S_t,\boldsymbol{\theta}) \right]$$

$$= \mathbb{E}\left[\Big(\tilde{G}_t^\lambda - \hat{v}(S_t,\mathbf{w}) \Big) \frac{\nabla \pi(A_t|S_t,\boldsymbol{\theta})}{\pi(A_t|S_t)} \;\middle|\; S_t \sim d_\pi, A_{t:\infty} \sim \pi \right]$$

$$\approx \Big(\tilde{G}_t^\lambda - \hat{v}(S_t,\mathbf{w}) \Big) \frac{\nabla \pi(A_t|S_t,\boldsymbol{\theta})}{\pi(A_t|S_t)} \qquad \text{(by sampling under } \pi\text{)}$$

$$\boldsymbol{\theta}_{t+1} \doteq \boldsymbol{\theta}_t + \alpha \Big(\tilde{G}_t^\lambda - \hat{v}(S_t,\mathbf{w}) \Big) \frac{\nabla \pi(A_t|S_t,\boldsymbol{\theta})}{\pi(A_t|S_t)}$$

e.g., in the one-step linear case:

$$= \boldsymbol{\theta}_t + \alpha \Big(R_{t+1} - \bar{R}_t + \mathbf{w}_t^\top \boldsymbol{\phi}_{t+1} - \mathbf{w}_t^\top \boldsymbol{\phi}_t \Big) \frac{\nabla \pi(A_t|S_t,\boldsymbol{\theta})}{\pi(A_t|S_t)}$$

Deriving the policy-gradient theorem: $\nabla r(\pi) = \sum_s d_\pi(s) \sum_a \tilde{q}_\pi(s,a) \nabla \pi(a|s,\boldsymbol{\theta})$:

$$\begin{aligned}
\nabla \tilde{v}_\pi(s) &= \nabla \sum_a \pi(a|s,\boldsymbol{\theta}) \tilde{q}_\pi(s,a) \\
&= \sum_a \Big[\nabla \pi(a|s,\boldsymbol{\theta}) \tilde{q}_\pi(s,a) + \pi(a|s,\boldsymbol{\theta}) \nabla \tilde{q}_\pi(s,a) \Big] \\
&= \sum_a \Big[\nabla \pi(a|s,\boldsymbol{\theta}) \tilde{q}_\pi(s,a) + \pi(a|s,\boldsymbol{\theta}) \nabla \sum_{s',r} p(s',r|s,a) \big[r - r(\pi) + \tilde{v}_\pi(s') \big] \Big] \\
&= \sum_a \Big[\nabla \pi(a|s,\boldsymbol{\theta}) \tilde{q}_\pi(s,a) + \pi(a|s,\boldsymbol{\theta}) \Big[-\nabla r(\pi) + \sum_{s',r} p(s'|s,a) \nabla \tilde{v}_\pi(s') \Big] \Big]
\end{aligned}$$

$$\therefore \nabla r(\pi) = \sum_a \Big[\nabla \pi(a|s,\boldsymbol{\theta}) \tilde{q}_\pi(s,a) + \pi(a|s,\boldsymbol{\theta}) \sum_{s'} p(s'|s,a) \nabla \tilde{v}_\pi(s') \Big] - \nabla \tilde{v}_\pi(s)$$

$$\therefore \nabla r(\pi) = \sum_a \Big[\nabla \pi(a|s,\boldsymbol{\theta}) \tilde{q}_\pi(s,a) + \pi(a|s,\boldsymbol{\theta}) \sum_{s'} p(s'|s,a) \nabla \tilde{v}_\pi(s') \Big] - \nabla \tilde{v}_\pi(s)$$

$$\begin{aligned}
\therefore \sum_s d_\pi(s) \nabla r(\pi) &= \sum_s d_\pi(s) \sum_a \nabla \pi(a|s,\boldsymbol{\theta}) \tilde{q}_\pi(s,a) \\
&\quad + \sum_s d_\pi(s) \sum_a \pi(a|s,\boldsymbol{\theta}) \sum_{s'} p(s'|s,a) \nabla \tilde{v}_\pi(s') - \sum_s d_\pi(s) \nabla \tilde{v}_\pi(s) \\
&= \sum_s d_\pi(s) \sum_a \nabla \pi(a|s,\boldsymbol{\theta}) \tilde{q}_\pi(s,a) \\
&\quad + \sum_{s'} \sum_s d_\pi(s) \sum_a \pi(a|s,\boldsymbol{\theta}) p(s'|s,a) \nabla \tilde{v}_\pi(s') - \sum_s d_\pi(s) \nabla \tilde{v}_\pi(s)
\end{aligned}$$

$$\nabla r(\pi) = \sum_s d_\pi(s) \sum_a \nabla \pi(a|s,\boldsymbol{\theta}) \tilde{q}_\pi(s,a)$$

Complete PG algorithm

Initialize parameters of policy $\boldsymbol{\theta} \in \mathbb{R}^n$, and state-value function $\mathbf{w} \in \mathbb{R}^m$

Initialize eligibility traces $\mathbf{e}^{\boldsymbol{\theta}} \in \mathbb{R}^n$ and $\mathbf{e}^{\mathbf{w}} \in \mathbb{R}^m$ to $\mathbf{0}$

Initialize $\bar{R} = 0$

On each step, in state S:

- Choose A according to $\pi(\cdot|S, \boldsymbol{\theta})$
- Take action A, observe S', R
- $\delta \leftarrow R - \bar{R} + \hat{v}(S', \mathbf{w}) - \hat{v}(S, \mathbf{w})$ — form TD error from critic
- $\bar{R} \leftarrow \bar{R} + \alpha^{\boldsymbol{\theta}} \delta$ — update average reward estimate
- $\mathbf{e}^{\mathbf{w}} \leftarrow \lambda \mathbf{e}^{\mathbf{w}} + \nabla_{\mathbf{w}} \hat{v}(S, \mathbf{w})$ — update eligibility trace for critic
- $\mathbf{w} \leftarrow \mathbf{w} + \alpha^{\mathbf{w}} \delta \mathbf{e}^{\mathbf{w}}$ — update critic parameters
- $\mathbf{e}^{\boldsymbol{\theta}} \leftarrow \lambda \mathbf{e}^{\boldsymbol{\theta}} + \frac{\nabla \pi(A|S,\boldsymbol{\theta})}{\pi(A|S,\boldsymbol{\theta})}$ — update eligibility trace for actor
- $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \alpha^{\boldsymbol{\theta}} \delta \mathbf{e}^{\boldsymbol{\theta}}$ — update actor parameters

The generality of the policy-gradient strategy

- Can be applied whenever we can compute the effect of parameter changes on the action probabilities,
- E.g., has been applied to spiking neuron models
- There are many possibilities other than linear-exponential and linear-gaussian, e.g., mixture of random, argmax, and fixed-width gaussian; learn the mixing weights, drift/diffusion models
- Can be applied whenever we can compute the effect of parameter changes on the action probabilities, $\nabla \pi(A_t|S_t, \boldsymbol{\theta})$