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

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

$$d_\pi \doteq \lim_{t\to\infty} \Pr\{S_t = s\} \qquad \sum_s d_\pi(s) \sum_a \pi(a|s,\boldsymbol{\theta}) p(s'|s,a) = d_\pi(s')$$

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

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

$$\Delta\boldsymbol{\theta}_t \doteq \alpha \widehat{\frac{\partial r(\pi)}{\partial \boldsymbol{\theta}}} \doteq \alpha \widehat{\nabla r(\pi)}$$

$$\begin{aligned}
\nabla r(\pi) &= \sum_s d_\pi(s) \sum_a \tilde{q}_\pi(s,a) \nabla \pi(a|s,\boldsymbol{\theta}) \qquad \text{(the policy-gradient theorem)} \\
&= \sum_s d_\pi(s) \sum_a \Big(\tilde{q}_\pi(s,a) - v(s)\Big) \nabla \pi(a|s,\boldsymbol{\theta}) \qquad \text{(for any } v : \mathcal{S} \to \mathbb{R}) \\
&= \sum_s d_\pi(s) \sum_a \pi(a|s,\boldsymbol{\theta}) \Big(\tilde{q}_\pi(s,a) - v(s)\Big) \frac{\nabla \pi(a|s,\boldsymbol{\theta})}{\pi(a|s,\boldsymbol{\theta})} \\
&= \mathbb{E}\left[\Big(\tilde{q}_\pi(S_t, A_t) - v(S_t)\Big) \frac{\nabla \pi(A_t|S_t,\boldsymbol{\theta})}{\pi(A_t|S_t,\boldsymbol{\theta})} \;\middle|\; S_t \sim d_\pi, A_t \sim \pi(\cdot|S_t,\boldsymbol{\theta}) \right]
\end{aligned}$$

Forward view:

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

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

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

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

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

Re-arranging terms:

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

Summing both sides over s, weighted by $d_\pi(s)$:

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

Final, complete policy-gradient algorithm:

Initialize parameters of policy $\boldsymbol{\theta} \in \mathbb{R}^n$, and state-value function $\mathbf{w} \in \mathbb{R}^m$
Initialize eligibility traces $\mathbf{z}^{\boldsymbol{\theta}} \in \mathbb{R}^n$ and $\mathbf{z}^{\mathbf{w}} \in \mathbb{R}^m$ 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
$\delta \leftarrow R - \bar{R} + \hat{v}(S', \mathbf{w}) - \hat{v}(S, \mathbf{w})$
$\bar{R} \leftarrow \bar{R} + \alpha_1 \delta$
$\mathbf{z}^{\mathbf{w}} \leftarrow \lambda \mathbf{z}^{\mathbf{w}} + \nabla_{\mathbf{w}} \hat{v}(S, \mathbf{w})$
$\mathbf{w} \leftarrow \mathbf{w} + \alpha_2 \delta \mathbf{z}^{\mathbf{w}}$
$\mathbf{z}^{\boldsymbol{\theta}} \leftarrow \lambda \mathbf{z}^{\boldsymbol{\theta}} + \frac{\nabla \pi(A|S,\boldsymbol{\theta})}{\pi(A|S,\boldsymbol{\theta})}$
$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \alpha_3 \delta \mathbf{z}^{\boldsymbol{\theta}}$

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

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

$$\mu(s) \doteq \boldsymbol{\theta}_\mu^\top \boldsymbol{\phi}(s)$$

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

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

$$\boldsymbol{\theta} \doteq (\boldsymbol{\theta}_\mu^\top; \boldsymbol{\theta}_\sigma^\top)^\top$$

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

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