Métodos de gradiente de política

Fernando Lozano

Universidad de los Andes

25 de mayo de 2023

...previamente

...previamente

- Encontrar política óptima $\pi^* \ : \ v_{\pi_*}(s) \geq v_\pi(s) \ \forall s \in \mathcal{S}$.

...previamente

- Encontrar política óptima π^* : $v_{\pi_*}(s) \geq v_\pi(s)$ $\forall s \in \mathcal{S}$.
  - Estimar $q_\pi(s, a)$

...previamente

- Encontrar política óptima π^* : $v_{\pi_*}(s) \geq v_\pi(s)\ \forall s \in \mathcal{S}$.
  - Estimar $q_\pi(s, a) \rightarrow$

...previamente

- Encontrar política óptima π^* : $v_{\pi_*}(s) \geq v_\pi(s)\ \forall s \in \mathcal{S}$.
  - Estimar $q_\pi(s, a) \rightarrow$ mejorar π

...previamente

- Encontrar política óptima π^* : $v_{\pi_*}(s) \geq v_\pi(s)\ \forall s \in \mathcal{S}$.
  - Estimar $q_\pi(s,a) \to$ mejorar $\pi \to q_*(s,a)$

...previamente

- Encontrar política óptima $\pi^*\ :\ v_{\pi_*}(s) \geq v_\pi(s)\ \forall s \in \mathcal{S}$.
  - Estimar $q_\pi(s,a) \rightarrow$ mejorar $\pi \rightarrow q_*(s,a)$
  - π^* greedy con respecto a $q_*(s,a)$

...previamente

- Encontrar política óptima $\pi^*\ :\ v_{\pi_*}(s) \geq v_\pi(s)\ \forall s \in \mathcal{S}$.
  - Estimar $q_\pi(s,a) \rightarrow$ mejorar $\pi \rightarrow q_*(s,a)$
  - π^* greedy con respecto a $q_*(s,a)$
- Aproximación de funciones: $\hat{q}(s,a,\mathbf{w})$

...previamente

- Encontrar política óptima $\pi^* \; : \; v_{\pi_*}(s) \geq v_\pi(s) \; \forall s \in \mathcal{S}$.
  - Estimar $q_\pi(s,a) \rightarrow$ mejorar $\pi \rightarrow q_*(s,a)$
  - π^* greedy con respecto a $q_*(s,a)$
- Aproximación de funciones: $\hat{q}(s,a,\mathbf{w})$
  - π^* greedy con respecto a $\hat{q}(s,a,\mathbf{w})$.

...previamente

- Encontrar política óptima $\pi^* \ : \ v_{\pi_*}(s) \geq v_\pi(s) \ \forall s \in \mathcal{S}$.
  - Estimar $q_\pi(s,a) \rightarrow$ mejorar $\pi \rightarrow q_*(s,a)$
  - π^* greedy con respecto a $q_*(s,a)$
- Aproximación de funciones: $\hat{q}(s,a,\mathbf{w})$
  - π^* greedy con respecto a $\hat{q}(s,a,\mathbf{w})$.
- Para MDP finito, aproximación tabular, se puede asumir π^* determinística

...previamente

- Encontrar política óptima $\pi^*\ :\ v_{\pi_*}(s) \geq v_\pi(s)\ \forall s \in \mathcal{S}$.
  - Estimar $q_\pi(s,a) \to$ mejorar $\pi \to q_*(s,a)$
  - π^* greedy con respecto a $q_*(s,a)$
- Aproximación de funciones: $\hat{q}(s,a,\mathbf{w})$
  - π^* greedy con respecto a $\hat{q}(s,a,\mathbf{w})$.
- Para MDP finito, aproximación tabular, se puede asumir π^* determinística$\to$ ϵ-greedy.

...previamente

- Encontrar política óptima $\pi^*\ :\ v_{\pi_*}(s) \geq v_\pi(s)\ \forall s \in \mathcal{S}$.
  - Estimar $q_\pi(s,a) \rightarrow$ mejorar $\pi \rightarrow q_*(s,a)$
  - π^* greedy con respecto a $q_*(s,a)$
- Aproximación de funciones: $\hat{q}(s,a,\mathbf{w})$
  - π^* greedy con respecto a $\hat{q}(s,a,\mathbf{w})$.
- Para MDP finito, aproximación tabular, se puede asumir π^* determinística$\rightarrow$ ϵ-greedy.
- Esto no es necesariamente cierto con aproximación de funciones.

...previamente

- Encontrar política óptima $\pi^* \; : \; v_{\pi_*}(s) \geq v_\pi(s) \; \forall s \in \mathcal{S}$.
  - Estimar $q_\pi(s,a) \rightarrow$ mejorar $\pi \rightarrow q_*(s,a)$
  - π^* greedy con respecto a $q_*(s,a)$
- Aproximación de funciones: $\hat{q}(s,a,\mathbf{w})$
  - π^* greedy con respecto a $\hat{q}(s,a,\mathbf{w})$.
- Para MDP finito, aproximación tabular, se puede asumir π^* determinística$\rightarrow$ ϵ-greedy.
- Esto no es necesariamente cierto con aproximación de funciones.
  - Pensar en $\pi(a \mid s)$ como una función más general.

Métodos de política de gradiente

Métodos de política de gradiente

- Aprender directamente política parametrizada:

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

Métodos de política de gradiente

- Aprender directamente política parametrizada:

$$\pi(a \mid s, \boldsymbol{\theta}) = \mathbf{P}\left\{A_t = a \mid S_t = s, \boldsymbol{\theta}_t = \boldsymbol{\theta}\right\}$$

Métodos de política de gradiente

- Aprender directamente política parametrizada:

$$\pi(a \mid s, \boldsymbol{\theta}) = \mathbf{P}\left\{A_t = a \mid S_t = s, \boldsymbol{\theta}_t = \boldsymbol{\theta}\right\}$$

  - Diferenciable con respecto a $\boldsymbol{\theta}$.

Métodos de política de gradiente

- Aprender directamente política parametrizada:

$$\pi(a \mid s, \boldsymbol{\theta}) = \mathbf{P}\left\{A_t = a \mid S_t = s, \boldsymbol{\theta}_t = \boldsymbol{\theta}\right\}$$

  - Diferenciable con respecto a $\boldsymbol{\theta}$.
  - $0 < \pi(a \mid s, \boldsymbol{\theta}) < 1 \leftarrow$ mantener exploración.

Métodos de política de gradiente

- Aprender directamente política parametrizada:

$$\pi(a \mid s, \boldsymbol{\theta}) = \mathbf{P}\left\{A_t = a \mid S_t = s, \boldsymbol{\theta}_t = \boldsymbol{\theta}\right\}$$

  - Diferenciable con respecto a $\boldsymbol{\theta}$.
  - $0 < \pi(a \mid s, \boldsymbol{\theta}) < 1 \leftarrow$ mantener exploración.
- Puede ser más fácil aprender $\pi(a \mid s, \boldsymbol{\theta})$ que aprender $\hat{q}(s, \mathbf{w})$.

Métodos de política de gradiente

- Aprender directamente política parametrizada:

$$\pi(a \mid s, \boldsymbol{\theta}) = \mathbf{P}\left\{A_t = a \mid S_t = s, \boldsymbol{\theta}_t = \boldsymbol{\theta}\right\}$$

  - Diferenciable con respecto a $\boldsymbol{\theta}$.
  - $0 < \pi(a \mid s, \boldsymbol{\theta}) < 1 \leftarrow$ mantener exploración.
- Puede ser más fácil aprender $\pi(a \mid s, \boldsymbol{\theta})$ que aprender $\hat{q}(s, \mathbf{w})$.
- Se puede introducir conocimiento previo en $\pi(a \mid s, \boldsymbol{\theta})$.

Métodos de política de gradiente

- Aprender directamente política parametrizada:

$$\pi(a \mid s, \boldsymbol{\theta}) = \mathbf{P}\{A_t = a \mid S_t = s, \boldsymbol{\theta}_t = \boldsymbol{\theta}\}$$

  - Diferenciable con respecto a $\boldsymbol{\theta}$.
  - $0 < \pi(a \mid s, \boldsymbol{\theta}) < 1 \leftarrow$ mantener exploración.
- Puede ser más fácil aprender $\pi(a \mid s, \boldsymbol{\theta})$ que aprender $\hat{q}(s, \mathbf{w})$.
- Se puede introducir conocimiento previo en $\pi(a \mid s, \boldsymbol{\theta})$.
- Criterio a maximizar $J(\boldsymbol{\theta})$.

Métodos de política de gradiente

- Aprender directamente política parametrizada:

$$\pi(a \mid s, \boldsymbol{\theta}) = \mathbf{P}\left\{A_t = a \mid S_t = s, \boldsymbol{\theta}_t = \boldsymbol{\theta}\right\}$$

  - Diferenciable con respecto a $\boldsymbol{\theta}$.
  - $0 < \pi(a \mid s, \boldsymbol{\theta}) < 1 \leftarrow$ mantener exploración.
- Puede ser más fácil aprender $\pi(a \mid s, \boldsymbol{\theta})$ que aprender $\hat{q}(s, \mathbf{w})$.
- Se puede introducir conocimiento previo en $\pi(a \mid s, \boldsymbol{\theta})$.
- Criterio a maximizar $J(\boldsymbol{\theta})$.
- Ascenso de gradiente en $J(\boldsymbol{\theta})$:

Métodos de política de gradiente

- Aprender directamente política parametrizada:

$$\pi(a \mid s, \boldsymbol{\theta}) = \mathbf{P}\left\{A_t = a \mid S_t = s, \boldsymbol{\theta}_t = \boldsymbol{\theta}\right\}$$

    - Diferenciable con respecto a $\boldsymbol{\theta}$.
    - $0 < \pi(a \mid s, \boldsymbol{\theta}) < 1 \leftarrow$ mantener exploración.
- Puede ser más fácil aprender $\pi(a \mid s, \boldsymbol{\theta})$ que aprender $\hat{q}(s, \mathbf{w})$.
- Se puede introducir conocimiento previo en $\pi(a \mid s, \boldsymbol{\theta})$.
- Criterio a maximizar $J(\boldsymbol{\theta})$.
- Ascenso de gradiente en $J(\boldsymbol{\theta})$:

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

Métodos de política de gradiente

- Aprender directamente política parametrizada:

$$\pi(a \mid s, \boldsymbol{\theta}) = \mathbf{P}\left\{A_t = a \mid S_t = s, \boldsymbol{\theta}_t = \boldsymbol{\theta}\right\}$$

  - Diferenciable con respecto a $\boldsymbol{\theta}$.
  - $0 < \pi(a \mid s, \boldsymbol{\theta}) < 1 \leftarrow$ mantener exploración.
- Puede ser más fácil aprender $\pi(a \mid s, \boldsymbol{\theta})$ que aprender $\hat{q}(s, \mathbf{w})$.
- Se puede introducir conocimiento previo en $\pi(a \mid s, \boldsymbol{\theta})$.
- Criterio a maximizar $J(\boldsymbol{\theta})$.
- Ascenso de gradiente en $J(\boldsymbol{\theta})$:

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

  - $\widehat{\nabla J(\boldsymbol{\theta})}$ es estimativo estocástico de $\nabla J(\boldsymbol{\theta})$:

Métodos de política de gradiente

- Aprender directamente política parametrizada:

$$\pi(a \mid s, \boldsymbol{\theta}) = \mathbf{P}\left\{A_t = a \mid S_t = s, \boldsymbol{\theta}_t = \boldsymbol{\theta}\right\}$$

  - Diferenciable con respecto a $\boldsymbol{\theta}$.
  - $0 < \pi(a \mid s, \boldsymbol{\theta}) < 1 \leftarrow$ mantener exploración.
- Puede ser más fácil aprender $\pi(a \mid s, \boldsymbol{\theta})$ que aprender $\hat{q}(s, \mathbf{w})$.
- Se puede introducir conocimiento previo en $\pi(a \mid s, \boldsymbol{\theta})$.
- Criterio a maximizar $J(\boldsymbol{\theta})$.
- Ascenso de gradiente en $J(\boldsymbol{\theta})$:

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

  - $\widehat{\nabla J(\boldsymbol{\theta})}$ es estimativo estocástico de $\nabla J(\boldsymbol{\theta})$:

$$\mathbb{E}\left[\widehat{\nabla J(\boldsymbol{\theta})}\right] \approx \nabla J(\boldsymbol{\theta})$$

Métodos de política de gradiente

- Aprender directamente política parametrizada:

$$\pi(a \mid s, \boldsymbol{\theta}) = \mathbf{P}\left\{A_t = a \mid S_t = s, \boldsymbol{\theta}_t = \boldsymbol{\theta}\right\}$$

  - Diferenciable con respecto a $\boldsymbol{\theta}$.
  - $0 < \pi(a \mid s, \boldsymbol{\theta}) < 1 \leftarrow$ mantener exploración.
- Puede ser más fácil aprender $\pi(a \mid s, \boldsymbol{\theta})$ que aprender $\hat{q}(s, \mathbf{w})$.
- Se puede introducir conocimiento previo en $\pi(a \mid s, \boldsymbol{\theta})$.
- Criterio a maximizar $J(\boldsymbol{\theta})$.
- Ascenso de gradiente en $J(\boldsymbol{\theta})$:

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

  - $\widehat{\nabla J(\boldsymbol{\theta})}$ es estimativo estocástico de $\nabla J(\boldsymbol{\theta})$:

$$\mathbb{E}\left[\widehat{\nabla J(\boldsymbol{\theta})}\right] \approx \nabla J(\boldsymbol{\theta})$$

- Puede usar $\hat{v}(s, \mathbf{w})$

Métodos de política de gradiente

- Aprender directamente política parametrizada:

$$\pi(a \mid s, \boldsymbol{\theta}) = \mathbf{P}\{A_t = a \mid S_t = s, \boldsymbol{\theta}_t = \boldsymbol{\theta}\}$$

  - Diferenciable con respecto a $\boldsymbol{\theta}$.
  - $0 < \pi(a \mid s, \boldsymbol{\theta}) < 1 \leftarrow$ mantener exploración.
- Puede ser más fácil aprender $\pi(a \mid s, \boldsymbol{\theta})$ que aprender $\hat{q}(s, \mathbf{w})$.
- Se puede introducir conocimiento previo en $\pi(a \mid s, \boldsymbol{\theta})$.
- Criterio a maximizar $J(\boldsymbol{\theta})$.
- Ascenso de gradiente en $J(\boldsymbol{\theta})$:

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

  - $\widehat{\nabla J(\boldsymbol{\theta})}$ es estimativo estocástico de $\nabla J(\boldsymbol{\theta})$:

$$\mathbb{E}\left[\widehat{\nabla J(\boldsymbol{\theta})}\right] \approx \nabla J(\boldsymbol{\theta})$$

- Puede usar $\hat{v}(s, \mathbf{w}) \rightarrow$ métodos actor-crítico.

- Para $|\mathcal{A}| \ll$, usar preferencias $h(s, a, \boldsymbol{\theta}) \in \mathbb{R}$,

- Para $|\mathcal{A}| \ll$, usar preferencias $h(s, a, \boldsymbol{\theta}) \in \mathbb{R}$,
  - Lineal: $h(s, a, \boldsymbol{\theta}) = \boldsymbol{\theta}^T \mathbf{x}(s, a)$

- Para $|\mathcal{A}| \ll$, usar preferencias $h(s,a,\boldsymbol{\theta}) \in \mathbb{R}$,
  - Lineal: $h(s,a,\boldsymbol{\theta}) = \boldsymbol{\theta}^T \mathbf{x}(s,a)$
  - $h(s,a,\boldsymbol{\theta})$: Red neuronal con pesos $\boldsymbol{\theta}$.

- Para $|\mathcal{A}| \ll$, usar preferencias $h(s, a, \boldsymbol{\theta}) \in \mathbb{R}$,
  - Lineal: $h(s, a, \boldsymbol{\theta}) = \boldsymbol{\theta}^T \mathbf{x}(s, a)$
  - $h(s, a, \boldsymbol{\theta})$: Red neuronal con pesos $\boldsymbol{\theta}$.
- soft-max con preferencias:

$$\pi(a \mid s, \boldsymbol{\theta}) = \frac{e^{h(s,a,\boldsymbol{\theta})}}{\sum_{a'} e^{h(s,a',\boldsymbol{\theta})}}$$

- Para $|\mathcal{A}| \ll$, usar preferencias $h(s, a, \boldsymbol{\theta}) \in \mathbb{R}$,
  - Lineal: $h(s, a, \boldsymbol{\theta}) = \boldsymbol{\theta}^T \mathbf{x}(s, a)$
  - $h(s, a, \boldsymbol{\theta})$: Red neuronal con pesos $\boldsymbol{\theta}$.
- soft-max con preferencias:

$$\pi(a \mid s, \boldsymbol{\theta}) = \frac{e^{h(s,a,\boldsymbol{\theta})}}{\sum_{a'} e^{h(s,a',\boldsymbol{\theta})}}$$

- Puede aproximar política determinística, si para cada estado se tiene acción a' con

$$h(s, a', \boldsymbol{\theta}) \gg h(s, a, \boldsymbol{\theta})$$

- Para $|\mathcal{A}| \ll$, usar preferencias $h(s, a, \boldsymbol{\theta}) \in \mathbb{R}$,
  - Lineal: $h(s, a, \boldsymbol{\theta}) = \boldsymbol{\theta}^T \mathbf{x}(s, a)$
  - $h(s, a, \boldsymbol{\theta})$: Red neuronal con pesos $\boldsymbol{\theta}$.
- soft-max con preferencias:

$$\pi(a \mid s, \boldsymbol{\theta}) = \frac{e^{h(s,a,\boldsymbol{\theta})}}{\sum_{a'} e^{h(s,a',\boldsymbol{\theta})}}$$

- Puede aproximar política determinística, si para cada estado se tiene acción a' con

$$h(s, a', \boldsymbol{\theta}) \gg h(s, a, \boldsymbol{\theta})$$

- Por qué no usar $\hat{q}(s, a, \mathbf{w})$ directamente?

- Para $|\mathcal{A}| \ll$, usar preferencias $h(s, a, \boldsymbol{\theta}) \in \mathbb{R}$,
  - Lineal: $h(s, a, \boldsymbol{\theta}) = \boldsymbol{\theta}^T \mathbf{x}(s, a)$
  - $h(s, a, \boldsymbol{\theta})$: Red neuronal con pesos $\boldsymbol{\theta}$.
- soft-max con preferencias:

$$\pi(a \mid s, \boldsymbol{\theta}) = \frac{e^{h(s,a,\boldsymbol{\theta})}}{\sum_{a'} e^{h(s,a',\boldsymbol{\theta})}}$$

- Puede aproximar política determinística, si para cada estado se tiene acción a' con

$$h(s, a', \boldsymbol{\theta}) \gg h(s, a, \boldsymbol{\theta})$$

- Por qué no usar $\hat{q}(s, a, \mathbf{w})$ directamente?

$$\hat{q}(s, a, \mathbf{w}) \rightarrow q_*(s, a)$$

- Para $|\mathcal{A}| \ll$, usar preferencias $h(s, a, \boldsymbol{\theta}) \in \mathbb{R}$,
  - Lineal: $h(s, a, \boldsymbol{\theta}) = \boldsymbol{\theta}^T \mathbf{x}(s, a)$
  - $h(s, a, \boldsymbol{\theta})$: Red neuronal con pesos $\boldsymbol{\theta}$.
- soft-max con preferencias:

$$\pi(a \mid s, \boldsymbol{\theta}) = \frac{e^{h(s,a,\boldsymbol{\theta})}}{\sum_{a'} e^{h(s,a',\boldsymbol{\theta})}}$$

- Puede aproximar política determinística, si para cada estado se tiene acción a' con

$$h(s, a', \boldsymbol{\theta}) \gg h(s, a, \boldsymbol{\theta})$$

- Por qué no usar $\hat{q}(s, a, \mathbf{w})$ directamente?

$$\hat{q}(s, a, \mathbf{w}) \rightarrow q_*(s, a)$$

  - No necesariamente $\hat{q}(s, a', \mathbf{w}) \gg \hat{q}(s, a, \mathbf{w})$.

- Para $|\mathcal{A}| \ll$, usar preferencias $h(s, a, \boldsymbol{\theta}) \in \mathbb{R}$,
  - Lineal: $h(s, a, \boldsymbol{\theta}) = \boldsymbol{\theta}^T \mathbf{x}(s, a)$
  - $h(s, a, \boldsymbol{\theta})$: Red neuronal con pesos $\boldsymbol{\theta}$.
- soft-max con preferencias:

$$\pi(a \mid s, \boldsymbol{\theta}) = \frac{e^{h(s,a,\boldsymbol{\theta})}}{\sum_{a'} e^{h(s,a',\boldsymbol{\theta})}}$$

- Puede aproximar política determinística, si para cada estado se tiene acción a' con

$$h(s, a', \boldsymbol{\theta}) \gg h(s, a, \boldsymbol{\theta})$$

- Por qué no usar $\hat{q}(s, a, \mathbf{w})$ directamente?

$$\hat{q}(s, a, \mathbf{w}) \to q_*(s, a)$$

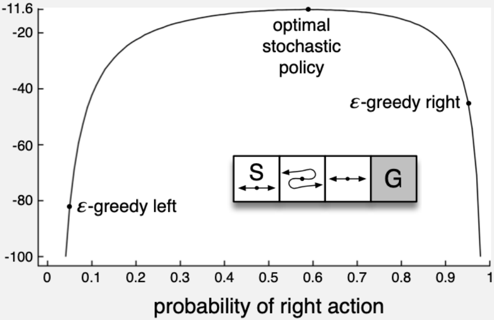
  - No necesariamente $\hat{q}(s, a', \mathbf{w}) \gg \hat{q}(s, a, \mathbf{w})$.
  - Maximizar $J(\theta) \to$ política óptima

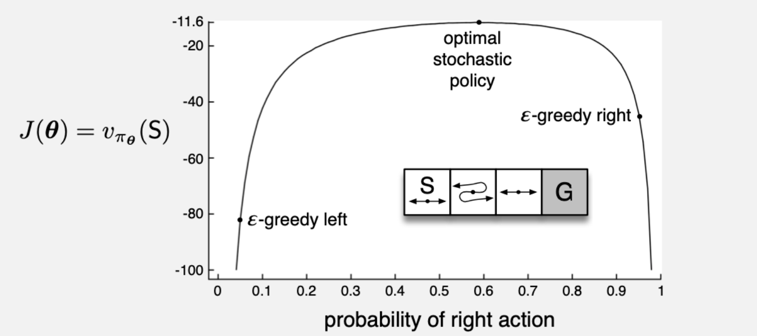
- En problemas con aproximación de funciones, la política óptima puede ser estocástica.

- En problemas con aproximación de funciones, la política óptima puede ser estocástica.
- Soft-max con preferencias puede aproximar probabilidades arbitrarias (con $\mathcal{H}_{\boldsymbol{\theta}}$ apropiada).

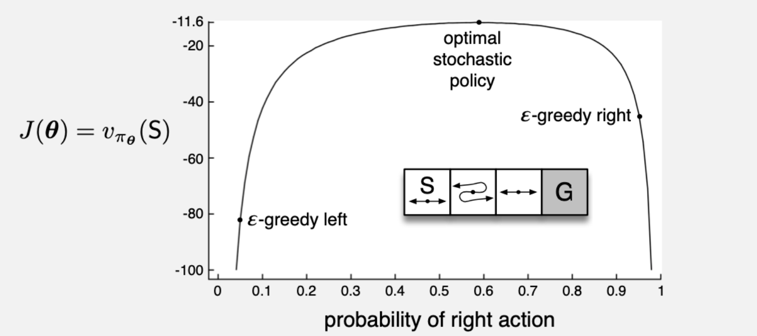
- En problemas con aproximación de funciones, la política óptima puede ser estocástica.
- Soft-max con preferencias puede aproximar probabilidades arbitrarias (con $\mathcal{H}_{\boldsymbol{\theta}}$ apropiada).
- Variación suave de las probabilidades con respecto a $\boldsymbol{\theta} \rightarrow$ garantías de convergencia.

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



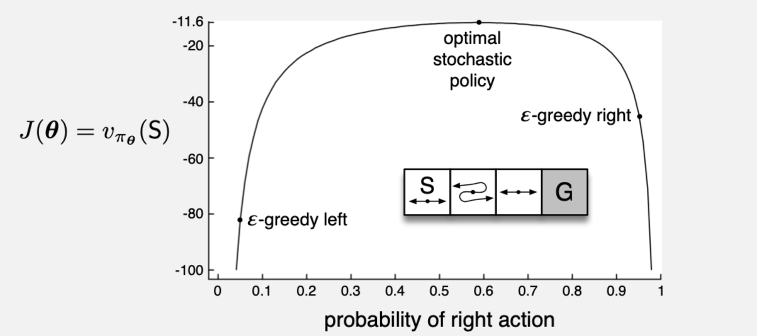


- $\mathbf{x}(s, \text{right}) = \begin{bmatrix} 1 & 0 \end{bmatrix}^T$, $\mathbf{x}(s, \text{left}) = \begin{bmatrix} 0 & 1 \end{bmatrix}^T$



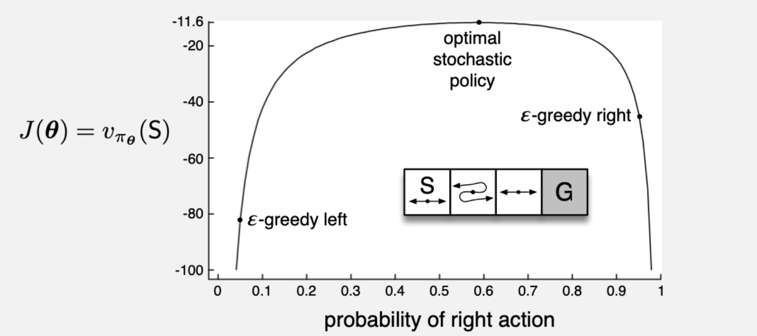
- $\mathbf{x}(s, \text{right}) = \begin{bmatrix}1 & 0\end{bmatrix}^T$, $\mathbf{x}(s, \text{left}) = \begin{bmatrix}0 & 1\end{bmatrix}^T$

$$\hat{q}(s, a, \mathbf{w}) = \mathbf{w}^T \mathbf{x}(s, a)$$



- $\mathbf{x}(s, \text{right}) = \begin{bmatrix}1 & 0\end{bmatrix}^T$, $\mathbf{x}(s, \text{left}) = \begin{bmatrix}0 & 1\end{bmatrix}^T$

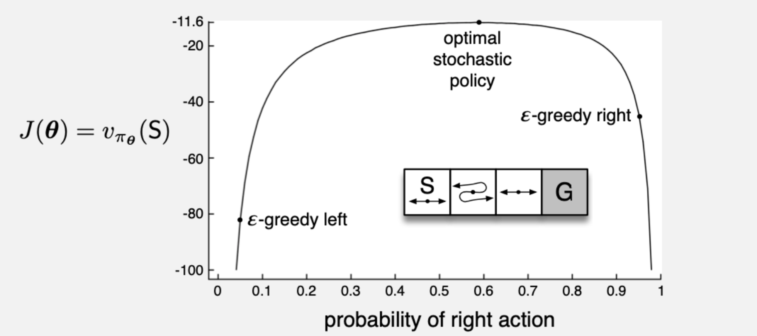
$$\hat{q}(s, a, \mathbf{w}) = \mathbf{w}^T \mathbf{x}(s, a) = \begin{cases} w_1 & a = \text{right} \\ w_2 & a = \text{left} \end{cases}$$



- $\mathbf{x}(s, \text{right}) = \begin{bmatrix} 1 & 0 \end{bmatrix}^T$, $\mathbf{x}(s, \text{left}) = \begin{bmatrix} 0 & 1 \end{bmatrix}^T$

$$\hat{q}(s, a, \mathbf{w}) = \mathbf{w}^T \mathbf{x}(s, a) = \begin{cases} w_1 & a = \text{right} \\ w_2 & a = \text{left} \end{cases}$$

- Política $\epsilon-$ greedy:



- $\mathbf{x}(s, \text{right}) = \begin{bmatrix} 1 & 0 \end{bmatrix}^T$, $\mathbf{x}(s, \text{left}) = \begin{bmatrix} 0 & 1 \end{bmatrix}^T$

$$\hat{q}(s, a, \mathbf{w}) = \mathbf{w}^T \mathbf{x}(s, a) = \begin{cases} w_1 & a = \text{right} \\ w_2 & a = \text{left} \end{cases}$$

- Política $\epsilon-$ greedy: $\pi(\text{left} \mid s) = 1 - \frac{\epsilon}{2}$, $\pi(\text{right} \mid s) = \frac{\epsilon}{2}$, o al revés.

Teorema de gradiente de política

Teorema de gradiente de política

- Criterio:

$$J(\boldsymbol{\theta}) \doteq v_{\pi_{\boldsymbol{\theta}}}(s_0)$$

Teorema de gradiente de política

- Criterio:

$$J(\boldsymbol{\theta}) \doteq v_{\pi_{\boldsymbol{\theta}}}(s_0)$$

- $J(\boldsymbol{\theta})$ depende de:

Teorema de gradiente de política

- Criterio:

$$J(\boldsymbol{\theta}) \doteq v_{\pi_{\boldsymbol{\theta}}}(s_0)$$

- $J(\boldsymbol{\theta})$ depende de:
  - Selección de acciones por $\pi(a \mid s, \boldsymbol{\theta})$

Teorema de gradiente de política

- Criterio:

$$J(\boldsymbol{\theta}) \doteq v_{\pi_{\boldsymbol{\theta}}}(s_0)$$

- $J(\boldsymbol{\theta})$ depende de:
  - Selección de acciones por $\pi(a \mid s, \boldsymbol{\theta}) \rightarrow$ se puede calcular.

Teorema de gradiente de política

- Criterio:

$$J(\boldsymbol{\theta}) \doteq v_{\pi_{\boldsymbol{\theta}}}(s_0)$$

- $J(\boldsymbol{\theta})$ depende de:
  - Selección de acciones por $\pi(a \mid s, \boldsymbol{\theta}) \rightarrow$ se puede calcular.
  - Estados en los que se toman las acciones

Teorema de gradiente de política

- Criterio:

$$J(\boldsymbol{\theta}) \doteq v_{\pi_{\boldsymbol{\theta}}}(s_0)$$

- $J(\boldsymbol{\theta})$ depende de:
  - Selección de acciones por $\pi(a \mid s, \boldsymbol{\theta}) \rightarrow$ se puede calcular.
  - Estados en los que se toman las acciones $\rightarrow$ no se puede calcular.

Teorema de gradiente de política

- Criterio:

$$J(\boldsymbol{\theta}) \doteq v_{\pi_{\boldsymbol{\theta}}}(s_0)$$

- $J(\boldsymbol{\theta})$ depende de:
  - Selección de acciones por $\pi(a \mid s, \boldsymbol{\theta})$ $\rightarrow$ se puede calcular.
  - Estados en los que se toman las acciones $\rightarrow$ no se puede calcular.

$$\nabla v_\pi(s) = \nabla \left[\sum_a \pi(a \mid s) q_\pi(s, a) \right]$$

$$
\begin{aligned}
\nabla v_\pi(s) &= \nabla \left[\sum_a \pi(a \mid s) q_\pi(s,a) \right] \\
&= \sum_a \left[\nabla \pi(a \mid s) q_\pi(s,a) + \pi(a \mid s) \nabla q_\pi(s,a) \right]
\end{aligned}
$$

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

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

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

$$
\begin{aligned}
\nabla v_\pi(s) &= \nabla \left[\sum_a \pi(a \mid s) q_\pi(s,a) \right] \\
&= \sum_a \left[\nabla \pi(a \mid s) q_\pi(s,a) + \pi(a \mid s) \nabla q_\pi(s,a) \right] \\
&= \sum_a \left[\nabla \pi(a \mid s) q_\pi(s,a) + \pi(a \mid s) \nabla \sum_{s',r} p(s', r \mid s, a)(r + v_\pi(s')) \right] \\
&= \sum_a \left[\nabla \pi(a \mid s) q_\pi(s,a) + \pi(a \mid s) \sum_{s'} p(s' \mid s, a) \color{red}{\nabla v_\pi(s')} \right] \\
&= \sum_a \Bigg[\nabla \pi(a \mid s) q_\pi(s,a) + \pi(a \mid s) \sum_{s'} p(s' \mid s, a) \\
& \color{red}{\sum_{a'} \nabla \pi(a' \mid s') q_\pi(s',a') + \pi(a' \mid s') \sum_{s''} p(s'' \mid s', a') \nabla v_\pi(s'') \Bigg]}
\end{aligned}
$$

$$
\begin{aligned}
\nabla v_\pi(s) &= \nabla \left[\sum_a \pi(a \mid s) q_\pi(s,a)\right] \\
&= \sum_a \left[\nabla \pi(a \mid s) q_\pi(s,a) + \pi(a \mid s) \nabla q_\pi(s,a)\right] \\
&= \sum_a \left[\nabla \pi(a \mid s) q_\pi(s,a) + \pi(a \mid s) \nabla \sum_{s',r} p(s', r \mid s, a)(r + v_\pi(s'))\right] \\
&= \sum_a \left[\nabla \pi(a \mid s) q_\pi(s,a) + \pi(a \mid s) \sum_{s'} p(s' \mid s, a) \nabla v_\pi(s')\right] \\
&= \sum_a \left[{\color{red}\nabla \pi(a \mid s) q_\pi(s,a)} + \pi(a \mid s) \sum_{s'} p(s' \mid s, a)\right. \\
& \left. \sum_{a'} {\color{red}\nabla \pi(a' \mid s') q_\pi(s',a')} + \pi(a' \mid s') \sum_{s''} p(s'' \mid s', a') \nabla v_\pi(s'')\right]
\end{aligned}
$$

$$
\begin{aligned}
\nabla v_\pi(s) &= \nabla \left[\sum_a \pi(a \mid s) q_\pi(s,a) \right] \\
&= \sum_a \left[\nabla \pi(a \mid s) q_\pi(s,a) + \pi(a \mid s) \nabla q_\pi(s,a) \right] \\
&= \sum_a \left[\nabla \pi(a \mid s) q_\pi(s,a) + \pi(a \mid s) \nabla \sum_{s',r} p(s', r \mid s, a)(r + v_\pi(s')) \right] \\
&= \sum_a \left[\nabla \pi(a \mid s) q_\pi(s,a) + \pi(a \mid s) \sum_{s'} p(s' \mid s, a) \nabla v_\pi(s') \right] \\
&= \sum_a \Bigg[{\color{red} \nabla \pi(a \mid s) q_\pi(s,a)} + \pi(a \mid s) \sum_{s'} p(s' \mid s, a) \\
&\qquad \sum_{a'} \nabla \pi(a' \mid s') q_\pi(s', a') + \pi(a' \mid s') \sum_{s''} p(s'' \mid s', a') \nabla v_\pi(s'') \Bigg]
\end{aligned}
$$

$$
\begin{aligned}
\nabla v_\pi(s) &= \nabla \left[\sum_a \pi(a \mid s) q_\pi(s,a) \right] \\
&= \sum_a \left[\nabla \pi(a \mid s) q_\pi(s,a) + \pi(a \mid s) \nabla q_\pi(s,a) \right] \\
&= \sum_a \left[\nabla \pi(a \mid s) q_\pi(s,a) + \pi(a \mid s) \nabla \sum_{s',r} p(s', r \mid s, a)(r + v_\pi(s')) \right] \\
&= \sum_a \left[\nabla \pi(a \mid s) q_\pi(s,a) + \pi(a \mid s) \sum_{s'} p(s' \mid s, a) \nabla v_\pi(s') \right] \\
&= \sum_a \left[\nabla \pi(a \mid s) q_\pi(s,a) + \pi(a \mid s) \sum_{s'} p(s' \mid s, a) \right. \\
&\left. \sum_{a'} \nabla \pi(a' \mid s') q_\pi(s', a') + \pi(a' \mid s') \sum_{s''} p(s'' \mid s', a') \nabla v_\pi(s'') \right]
\end{aligned}
$$

$$\nabla v_\pi(s) = \sum_a \Bigg[\nabla\pi(a \mid s) q_\pi(s,a) + \pi(a \mid s) \sum_{s'} p(s' \mid s, a)$$

$$\sum_{a'} \nabla\pi(a' \mid s') q_\pi(s', a') + \pi(s' \mid a') \sum_{s''} p(s'' \mid s', a') \nabla v_\pi(s'')\Bigg]$$

$$
\begin{aligned}
\nabla v_\pi(s) = \sum_a \Bigg[&\nabla \pi(a \mid s) q_\pi(s,a) + \pi(a \mid s) \sum_{s'} p(s' \mid s, a) \\
&\sum_{a'} \nabla \pi(a' \mid s') q_\pi(s', a') + \pi(s' \mid a') \sum_{s''} p(s'' \mid s', a') \nabla v_\pi(s'') \Bigg] \\
&= \sum_{x \in \mathcal{S}} \sum_{k=0}^{\infty} \mathbf{P}\left\{ s \to x, k, \pi \right\} \sum_a \nabla \pi(a \mid x) q_\pi(x, a)
\end{aligned}
$$

Distribución on-policy de tareas episódicas

- $h(s)$: probabilidad de comenzar el episodio en s.

Distribución on-policy de tareas episódicas

- $h(s)$: probabilidad de comenzar el episodio en s.
- $\eta(s)$ promedio de pasos en un episodio en s.

Distribución on-policy de tareas episódicas

- $h(s)$: probabilidad de comenzar el episodio en s.
- $\eta(s)$ promedio de pasos en un episodio en s.

$$\eta(s) =$$

Distribución on-policy de tareas episódicas

- $h(s)$: probabilidad de comenzar el episodio en s.
- $\eta(s)$ promedio de pasos en un episodio en s.

$$\eta(s) = h(s)$$

Distribución on-policy de tareas episódicas

- $h(s)$: probabilidad de comenzar el episodio en s.
- $\eta(s)$ promedio de pasos en un episodio en s.

$$\eta(s) = h(s) + \sum_{\bar{s}} \eta(\bar{s}) \sum_{a} \pi(a \mid \bar{s}) p(s \mid \bar{s}, a)$$

Distribución on-policy de tareas episódicas

- $h(s)$: probabilidad de comenzar el episodio en s.
- $\eta(s)$ promedio de pasos en un episodio en s.

$$\eta(s) = h(s) + \sum_{\bar{s}} \eta(\bar{s}) \sum_{a} \pi(a \mid \bar{s}) p(s \mid \bar{s}, a)$$

- Sistema de ecuaciones

Distribución on-policy de tareas episódicas

- $h(s)$: probabilidad de comenzar el episodio en s.
- $\eta(s)$ promedio de pasos en un episodio en s.

$$\eta(s) = h(s) + \sum_{\bar{s}} \eta(\bar{s}) \sum_{a} \pi(a \mid \bar{s}) p(s \mid \bar{s}, a)$$

- Sistema de ecuaciones $\rightarrow$ resolver para $\eta(s)$

Distribución on-policy de tareas episódicas

- $h(s)$: probabilidad de comenzar el episodio en s.
- $\eta(s)$ promedio de pasos en un episodio en s.

$$\eta(s) = h(s) + \sum_{\bar{s}} \eta(\bar{s}) \sum_{a} \pi(a \mid \bar{s}) p(s \mid \bar{s}, a)$$

- Sistema de ecuaciones $\rightarrow$ resolver para $\eta(s)$
- Distribución on-policy:

$$\mu(s) = \frac{\eta(s)}{\sum_{s'} \eta(s')}$$

$$\nabla v_\pi(s_0) = \sum_s \left(\sum_{k=0}^{\infty} \mathbf{P}\left\{s_0 \to s, k, \pi\right\} \right) \sum_a \nabla \pi(a \mid s) q_\pi(s, a)$$

$$
\begin{aligned}
\nabla v_\pi(s_0) &= \sum_s \left(\sum_{k=0}^{\infty} \mathbf{P}\left\{ s_0 \to s, k, \pi \right\} \right) \sum_a \nabla \pi(a \mid s) q_\pi(s, a) \\
&= \sum_s \eta(s) \sum_a \nabla \pi(a \mid s) q_\pi(s, a)
\end{aligned}
$$

$$
\begin{aligned}
\nabla v_\pi(s_0) &= \sum_s \left(\sum_{k=0}^{\infty} \mathbf{P}\left\{ s_0 \to s, k, \pi \right\} \right) \sum_a \nabla \pi(a \mid s) q_\pi(s, a) \\
&= \sum_s \eta(s) \sum_a \nabla \pi(a \mid s) q_\pi(s, a) \\
&= \sum_{s'} \eta(s') \sum_s \frac{\eta(s)}{\sum_{s'} \eta(s')} \sum_a \nabla \pi(a \mid s) q_\pi(s, a)
\end{aligned}
$$

$$
\begin{aligned}
\nabla v_\pi(s_0) &= \sum_s \left(\sum_{k=0}^{\infty} \mathbf{P}\left\{ s_0 \to s, k, \pi \right\} \right) \sum_a \nabla \pi(a \mid s) q_\pi(s, a) \\
&= \sum_s \eta(s) \sum_a \nabla \pi(a \mid s) q_\pi(s, a) \\
&= \sum_{s'} \eta(s') \sum_s \frac{\eta(s)}{\sum_{s'} \eta(s')} \sum_a \nabla \pi(a \mid s) q_\pi(s, a) \\
&= \sum_{s'} \eta(s') \sum_s \mu(s) \sum_a \nabla \pi(a \mid s) q_\pi(s, a)
\end{aligned}
$$

- Teorema de gradiente de política:

$$\nabla J(\boldsymbol{\theta}) \propto \sum_{s} \mu(s) \sum_{a} q_{\pi}(s, a) \nabla_{\boldsymbol{\theta}} \pi(a \mid s, \boldsymbol{\theta})$$

- Teorema de gradiente de política:

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

  - No depende de cambios en transiciones debidos a la política.

- Teorema de gradiente de política:

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

  - No depende de cambios en transiciones debidos a la política.
  - Constante de proporcionalidad:

- Teorema de gradiente de política:

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

  - No depende de cambios en transiciones debidos a la política.
  - Constante de proporcionalidad:
    - Duración promedio de episodios (caso episódico)

- Teorema de gradiente de política:

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

  - No depende de cambios en transiciones debidos a la política.
  - Constante de proporcionalidad:
    - Duración promedio de episodios (caso episódico)
    - 1 caso no episódico.

- Teorema de gradiente de política:

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

  - No depende de cambios en transiciones debidos a la política.
  - Constante de proporcionalidad:
    - Duración promedio de episodios (caso episódico)
    - 1 caso no episódico.
    - No importa

- Teorema de gradiente de política:

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

  - No depende de cambios en transiciones debidos a la política.
  - Constante de proporcionalidad:
    - Duración promedio de episodios (caso episódico)
    - 1 caso no episódico.
    - No importa$\rightarrow \alpha$

Gradiente de política Montecarlo

Gradiente de política Montecarlo

- Ascenso de gradiente estocástico:

Gradiente de política Montecarlo

- Ascensо de gradiente estocástico:
  - Muestras del gradiente (online) cuyo valor esperado es proporcional al gradiente real.

Gradiente de política Montecarlo

- Ascenso de gradiente estocástico:
  - Muestras del gradiente (online) cuyo valor esperado es proporcional al gradiente real.
- Si $\mu(s)$ es la distribución on-policy de π:

Gradiente de política Montecarlo

- Ascenso de gradiente estocástico:
  - Muestras del gradiente (online) cuyo valor esperado es proporcional al gradiente real.
- Si $\mu(s)$ es la distribución on-policy de π:

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

Gradiente de política Montecarlo

- Ascenso de gradiente estocástico:
  - Muestras del gradiente (online) cuyo valor esperado es proporcional al gradiente real.
- Si $\mu(s)$ es la distribución on-policy de π:

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

Gradiente de política Montecarlo

- Ascenso de gradiente estocástico:
  - Muestras del gradiente (online) cuyo valor esperado es proporcional al gradiente real.
- Si $\mu(s)$ es la distribución on-policy de π:

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

- Sugiere regla de actualización:

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t + \alpha \sum_a \hat{q}(S_t, a, \mathbf{w}) \nabla_{\boldsymbol{\theta}} \pi(a \mid S_t, \boldsymbol{\theta})$$

Gradiente de política Montecarlo

- Ascenso de gradiente estocástico:
  - Muestras del gradiente (online) cuyo valor esperado es proporcional al gradiente real.
- Si $\mu(s)$ es la distribución on-policy de π:

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

- Sugiere regla de actualización:

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t + \alpha \sum_a \hat{q}(S_t, a, \mathbf{w}) \nabla_{\boldsymbol{\theta}} \pi(a \mid S_t, \boldsymbol{\theta})$$

  - No totalmente on-line.

Gradiente de política Montecarlo

- Ascenso de gradiente estocástico:
  - Muestras del gradiente (online) cuyo valor esperado es proporcional al gradiente real.
- Si $\mu(s)$ es la distribución on-policy de π:

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

- Sugiere regla de actualización:

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t + \alpha \sum_a \hat{q}(S_t, a, \mathbf{w}) \nabla_{\boldsymbol{\theta}} \pi(a \mid S_t, \boldsymbol{\theta})$$

  - No totalmente on-line.
  - Requiere $\hat{q}(S_t, a, \mathbf{w})$

$$\nabla J(\boldsymbol{\theta}) \propto \mathbb{E}_\pi \left[\sum_a q_\pi(S_t, a) \nabla_{\boldsymbol{\theta}} \pi(a \mid S_t, \boldsymbol{\theta}) \right]$$

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

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

$$
\begin{aligned}
\nabla J(\boldsymbol{\theta}) &\propto \mathbb{E}_\pi \left[\sum_a q_\pi(S_t, a) \nabla_{\boldsymbol{\theta}} \pi(a \mid S_t, \boldsymbol{\theta}) \right] \\
&= \mathbb{E}_\pi \left[\sum_a \pi(a \mid S_t, \boldsymbol{\theta}) q_\pi(S_t, a) \frac{\nabla_{\boldsymbol{\theta}} \pi(a \mid S_t, \boldsymbol{\theta})}{\pi(a \mid S_t, \boldsymbol{\theta})} \right] \\
&= \mathbb{E}_\pi \left[q_\pi(S_t, A_t) \frac{\nabla_{\boldsymbol{\theta}} \pi(A_t \mid S_t, \boldsymbol{\theta})}{\pi(A_t \mid S_t, \boldsymbol{\theta})} \right] \\
&\underset{\mathbb{E}_\pi[G_t \mid S_t, A_t] = q_\pi(S_t, A_t)}{\underset{\uparrow}{=}} \mathbb{E}_\pi \left[G_t \frac{\nabla_{\boldsymbol{\theta}} \pi(A_t \mid S_t, \boldsymbol{\theta})}{\pi(A_t \mid S_t, \boldsymbol{\theta})} \right]
\end{aligned}
$$

- Actualización (REINFORCE):

- Actualización (REINFORCE):

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

- Actualización (REINFORCE):

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t + \alpha G_t \frac{\nabla_{\boldsymbol{\theta}} \pi(A_t \mid S_t, \boldsymbol{\theta})}{\pi(A_t \mid S_t, \boldsymbol{\theta})} = \boldsymbol{\theta}_t + \alpha G_t \nabla_{\boldsymbol{\theta}} \ln\left(\pi(A_t \mid S_t, \boldsymbol{\theta})\right)$$

- Actualización (REINFORCE):

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t + \alpha G_t \frac{\nabla_{\boldsymbol{\theta}} \pi(A_t \mid S_t, \boldsymbol{\theta})}{\pi(A_t \mid S_t, \boldsymbol{\theta})} = \boldsymbol{\theta}_t + \alpha G_t \nabla_{\boldsymbol{\theta}} \ln\left(\pi(A_t \mid S_t, \boldsymbol{\theta})\right)$$

  - Dirección en la que más crece la probabilidad de tomar acción A_t al volver al estado S_t.

- Actualización (REINFORCE):

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t + \alpha G_t \frac{\nabla_{\boldsymbol{\theta}} \pi(A_t \mid S_t, \boldsymbol{\theta})}{\pi(A_t \mid S_t, \boldsymbol{\theta})} = \boldsymbol{\theta}_t + \alpha G_t \nabla_{\boldsymbol{\theta}} \ln\left(\pi(A_t \mid S_t, \boldsymbol{\theta})\right)$$

  - Dirección en la que más crece la probabilidad de tomar acción A_t al volver al estado S_t.
  - Proporcional al retorno observado en la transición.

- Actualización (REINFORCE):

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t + \alpha G_t \frac{\nabla_{\boldsymbol{\theta}} \pi(A_t \mid S_t, \boldsymbol{\theta})}{\pi(A_t \mid S_t, \boldsymbol{\theta})} = \boldsymbol{\theta}_t + \alpha G_t \nabla_{\boldsymbol{\theta}} \ln\left(\pi(A_t \mid S_t, \boldsymbol{\theta})\right)$$

  - Dirección en la que más crece la probabilidad de tomar acción A_t al volver al estado S_t.
  - Proporcional al retorno observado en la transición.
  - Inversamente proporcional a la probabilidad de tomar acción A_t.

Usando preferencias lineales:

$$\nabla_{\boldsymbol{\theta}} \ln(\pi(a \mid s, \boldsymbol{\theta})) = \nabla_{\boldsymbol{\theta}} \ln \left(\frac{e^{\boldsymbol{\theta}^T \mathbf{x}(s,a)}}{\sum_{a'} e^{\boldsymbol{\theta}^T \mathbf{x}(s,a')}} \right)$$

Usando preferencias lineales:

$$\begin{aligned}\nabla_{\boldsymbol{\theta}} \ln(\pi(a \mid s, \boldsymbol{\theta})) &= \nabla_{\boldsymbol{\theta}} \ln\left(\frac{e^{\boldsymbol{\theta}^T \mathbf{x}(s,a)}}{\sum_{a'} e^{\boldsymbol{\theta}^T \mathbf{x}(s,a')}}\right) \\ &= \nabla_{\boldsymbol{\theta}} \ln\left(e^{\boldsymbol{\theta}^T \mathbf{x}(s,a)}\right) - \nabla_{\boldsymbol{\theta}} \ln\left(\sum_{a'} e^{\boldsymbol{\theta}^T \mathbf{x}(s,a')}\right)\end{aligned}$$

Usando preferencias lineales:

$$\begin{aligned}
\nabla_{\boldsymbol{\theta}} \ln(\pi(a \mid s, \boldsymbol{\theta})) &= \nabla_{\boldsymbol{\theta}} \ln\left(\frac{e^{\boldsymbol{\theta}^T \mathbf{x}(s,a)}}{\sum_{a'} e^{\boldsymbol{\theta}^T \mathbf{x}(s,a')}}\right) \\
&= \nabla_{\boldsymbol{\theta}} \ln\left(e^{\boldsymbol{\theta}^T \mathbf{x}(s,a)}\right) - \nabla_{\boldsymbol{\theta}} \ln\left(\sum_{a'} e^{\boldsymbol{\theta}^T \mathbf{x}(s,a')}\right) \\
&= \mathbf{x}(s, a)
\end{aligned}$$

Usando preferencias lineales:

$$
\begin{aligned}
\nabla_{\boldsymbol{\theta}} \ln(\pi(a \mid s, \boldsymbol{\theta})) &= \nabla_{\boldsymbol{\theta}} \ln \left(\frac{e^{\boldsymbol{\theta}^T \mathbf{x}(s,a)}}{\sum_{a'} e^{\boldsymbol{\theta}^T \mathbf{x}(s,a')}} \right) \\
&= \nabla_{\boldsymbol{\theta}} \ln \left(e^{\boldsymbol{\theta}^T \mathbf{x}(s,a)} \right) - \nabla_{\boldsymbol{\theta}} \ln \left(\sum_{a'} e^{\boldsymbol{\theta}^T \mathbf{x}(s,a')} \right) \\
&= \mathbf{x}(s,a) - \frac{\sum_{a'} \mathbf{x}(s,a') e^{\boldsymbol{\theta}^T \mathbf{x}(s,a')}}{\sum_{a'} e^{\boldsymbol{\theta}^T \mathbf{x}(s,a')}}
\end{aligned}
$$

Usando preferencias lineales:

$$
\begin{aligned}
\nabla_{\boldsymbol{\theta}} \ln(\pi(a \mid s, \boldsymbol{\theta})) &= \nabla_{\boldsymbol{\theta}} \ln\left(\frac{e^{\boldsymbol{\theta}^T \mathbf{x}(s,a)}}{\sum_{a'} e^{\boldsymbol{\theta}^T \mathbf{x}(s,a')}}\right) \\
&= \nabla_{\boldsymbol{\theta}} \ln\left(e^{\boldsymbol{\theta}^T \mathbf{x}(s,a)}\right) - \nabla_{\boldsymbol{\theta}} \ln\left(\sum_{a'} e^{\boldsymbol{\theta}^T \mathbf{x}(s,a')}\right) \\
&= \mathbf{x}(s,a) - \frac{\sum_{a'} \mathbf{x}(s,a') e^{\boldsymbol{\theta}^T \mathbf{x}(s,a')}}{\sum_{a'} e^{\boldsymbol{\theta}^T \mathbf{x}(s,a')}} \\
&= \mathbf{x}(s,a) - \sum_{a'} \pi(a' \mid s, \boldsymbol{\theta}) \mathbf{x}(s,a')
\end{aligned}
$$

Usando preferencias lineales:

$$\begin{aligned}
\nabla_{\boldsymbol{\theta}} \ln(\pi(a \mid s, \boldsymbol{\theta})) &= \nabla_{\boldsymbol{\theta}} \ln\left(\frac{e^{\boldsymbol{\theta}^T \mathbf{x}(s,a)}}{\sum_{a'} e^{\boldsymbol{\theta}^T \mathbf{x}(s,a')}}\right) \\
&= \nabla_{\boldsymbol{\theta}} \ln\left(e^{\boldsymbol{\theta}^T \mathbf{x}(s,a)}\right) - \nabla_{\boldsymbol{\theta}} \ln\left(\sum_{a'} e^{\boldsymbol{\theta}^T \mathbf{x}(s,a')}\right) \\
&= \mathbf{x}(s,a) - \frac{\sum_{a'} \mathbf{x}(s,a') e^{\boldsymbol{\theta}^T \mathbf{x}(s,a')}}{\sum_{a'} e^{\boldsymbol{\theta}^T \mathbf{x}(s,a')}} \\
&= \mathbf{x}(s,a) - \sum_{a'} \pi(a' \mid s, \boldsymbol{\theta}) \mathbf{x}(s,a') \\
&= \mathbf{x}(s,a) - \mathbb{E}_{\pi}\left[\mathbf{x}(s,a)\right]
\end{aligned}$$

- En general:

$$\nabla_{\boldsymbol{\theta}} \ln(\pi(a \mid s, \boldsymbol{\theta})) = \nabla_{\boldsymbol{\theta}} h(s, a, \boldsymbol{\theta}) - \mathbb{E}_{\pi}\left[\nabla_{\boldsymbol{\theta}} h(s, a, \boldsymbol{\theta})\right]$$

Monte Carlo Policy Gradient (REINFORCE)

Monte Carlo Policy Gradient (REINFORCE)

Require: $\pi(a \mid s, \boldsymbol{\theta})$, $\alpha > 0$

Monte Carlo Policy Gradient (REINFORCE)

Require: $\pi(a \mid s, \boldsymbol{\theta})$, $\alpha > 0$

Inicialice $\boldsymbol{\theta}$

Monte Carlo Policy Gradient (REINFORCE)

Require: $\pi(a \mid s, \boldsymbol{\theta})$, $\alpha > 0$
  Inicialice $\boldsymbol{\theta}$
  repeat

Monte Carlo Policy Gradient (REINFORCE)

Require: $\pi(a \mid s, \boldsymbol{\theta})$, $\alpha > 0$

Inicialice $\boldsymbol{\theta}$

repeat

Episodio $\pi(. \mid , \boldsymbol{\theta})\ :\ S_0, A_0, R_1, S_2, A_2, R_2, \ldots S_{T-1}, A_{T-1}, R_T,$

Monte Carlo Policy Gradient (REINFORCE)

Require: $\pi(a \mid s, \boldsymbol{\theta})$, $\alpha > 0$

Inicialice $\boldsymbol{\theta}$

repeat

Episodio $\pi(. \mid , \boldsymbol{\theta})\ :\ S_0, A_0, R_1, S_2, A_2, R_2, \ldots S_{T-1}, A_{T-1}, R_T,$

for $t = 0, 1, \ldots, T-1$ **do**

Monte Carlo Policy Gradient (REINFORCE)

Require: $\pi(a \mid s, \boldsymbol{\theta})$, $\alpha > 0$

  Inicialice $\boldsymbol{\theta}$

  repeat

    Episodio $\pi(. \mid , \boldsymbol{\theta})\ :\ S_0, A_0, R_1, S_2, A_2, R_2, \ldots S_{T-1}, A_{T-1}, R_T,$

    for $t = 0, 1, \ldots, T-1$ **do**

      $G \leftarrow \sum_{k=t+1}^{T} \gamma^{k-t-1} R_k$

Monte Carlo Policy Gradient (REINFORCE)

Require: $\pi(a \mid s, \boldsymbol{\theta})$, $\alpha > 0$

  Inicialice $\boldsymbol{\theta}$

  repeat

    Episodio $\pi(. \mid , \boldsymbol{\theta})\ :\ S_0, A_0, R_1, S_2, A_2, R_2, \ldots S_{T-1}, A_{T-1}, R_T,$

    for $t = 0, 1, \ldots, T-1$ **do**

      $G \leftarrow \sum_{k=t+1}^{T} \gamma^{k-t-1} R_k$

      $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \alpha\gamma^t G \nabla_{\boldsymbol{\theta}} \ln\left(\pi(A_t \mid S_t, \boldsymbol{\theta})\right)$

Monte Carlo Policy Gradient (REINFORCE)

Require: $\pi(a \mid s, \boldsymbol{\theta})$, $\alpha > 0$

  Inicialice $\boldsymbol{\theta}$

  repeat

    Episodio $\pi(. \mid, \boldsymbol{\theta})\ :\ S_0, A_0, R_1, S_2, A_2, R_2, \ldots S_{T-1}, A_{T-1}, R_T,$

    for $t = 0, 1, \ldots, T-1$ **do**

      $G \leftarrow \sum_{k=t+1}^{T} \gamma^{k-t-1} R_k$

      $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \alpha\gamma^t G \nabla_{\boldsymbol{\theta}} \ln\left(\pi(A_t \mid S_t, \boldsymbol{\theta})\right)$

    end for

Monte Carlo Policy Gradient (REINFORCE)

Require: $\pi(a \mid s, \boldsymbol{\theta})$, $\alpha > 0$

Inicialice $\boldsymbol{\theta}$

repeat

Episodio $\pi(. \mid , \boldsymbol{\theta})$: $S_0, A_0, R_1, S_2, A_2, R_2, \ldots S_{T-1}, A_{T-1}, R_T,$

for $t = 0, 1, \ldots, T-1$ **do**

$G \leftarrow \sum_{k=t+1}^{T} \gamma^{k-t-1} R_k$

$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \alpha\gamma^t G \nabla_{\boldsymbol{\theta}} \ln\left(\pi(A_t \mid S_t, \boldsymbol{\theta})\right)$

end for

until ∞

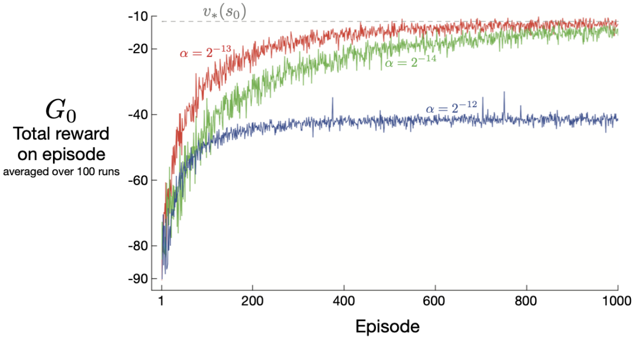


Figure 13.1: REINFORCE on the short-corridor gridworld (Example 13.1). With a good step size, the total reward per episode approaches the optimal value of the start state.

Baseline

Baseline

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

Baseline

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

- Si $b(s)$ no depende de a:

$$\sum_a (q_\pi(s,a) - b(s)) \nabla_{\boldsymbol{\theta}} \pi(a \mid s, \boldsymbol{\theta})$$

$$= \sum_a q_\pi(s,a) \nabla_{\boldsymbol{\theta}} \pi(a \mid s, \boldsymbol{\theta})$$

Baseline

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

- Si $b(s)$ no depende de a:

$$\sum_a (q_\pi(s,a) - b(s)) \nabla_{\boldsymbol{\theta}} \pi(a \mid s, \boldsymbol{\theta})$$

$$= \sum_a q_\pi(s,a) \nabla_{\boldsymbol{\theta}} \pi(a \mid s, \boldsymbol{\theta}) - b(s) \nabla_{\boldsymbol{\theta}} \cancelto{0}{\underbrace{\sum_a \pi(a \mid s, \boldsymbol{\theta})}_{1}}$$

- Actualización con baseline:

- Actualización con baseline:

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

- Actualización con baseline:

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

  - Baseline dependiente del estado.

- Actualización con baseline:

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

  - Baseline dependiente del estado.
  - Valor esperado de la actualización es el mismo.

- Actualización con baseline:

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

  - Baseline dependiente del estado.
  - Valor esperado de la actualización es el mismo.
  - Puede reducir varianza.

- Actualización con baseline:

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

  - Baseline dependiente del estado.
  - Valor esperado de la actualización es el mismo.
  - Puede reducir varianza.
- REINFORCE con baseline:

$$b(S_t) = \hat{v}(S_t, \mathbf{w})$$

REINFORCE con baseline

Require: $\pi(a \mid s, \boldsymbol{\theta})$, $\hat{v}(s, \mathbf{w})$, $\alpha_{\boldsymbol{\theta}}, \alpha_{\mathbf{w}} > 0$

REINFORCE con baseline

Require: $\pi(a \mid s, \boldsymbol{\theta})$, $\hat{v}(s, \mathbf{w})$, $\alpha_{\boldsymbol{\theta}}, \alpha_{\mathbf{w}} > 0$
  Inicialice $\boldsymbol{\theta}, \mathbf{w}$

REINFORCE con baseline

Require: $\pi(a \mid s, \boldsymbol{\theta})$, $\hat{v}(s, \mathbf{w})$, $\alpha_{\boldsymbol{\theta}}, \alpha_{\mathbf{w}} > 0$

Incialice $\boldsymbol{\theta}, \mathbf{w}$

repeat

REINFORCE con baseline

Require: $\pi(a \mid s, \boldsymbol{\theta})$, $\hat{v}(s, \mathbf{w})$, $\alpha_{\boldsymbol{\theta}}, \alpha_{\mathbf{w}} > 0$

Inicialice $\boldsymbol{\theta}, \mathbf{w}$

repeat

Episodio $\pi(. \mid , \boldsymbol{\theta}) \ : \ S_0, A_0, R_1, S_2, A_2, R_2, \ldots S_{T-1}, A_{T-1}, R_T,$

REINFORCE con baseline

Require: $\pi(a \mid s, \boldsymbol{\theta})$, $\hat{v}(s, \mathbf{w})$, $\alpha_{\boldsymbol{\theta}}, \alpha_{\mathbf{w}} > 0$

Inicialice $\boldsymbol{\theta}, \mathbf{w}$

repeat

Episodio $\pi(. \mid , \boldsymbol{\theta})$: $S_0, A_0, R_1, S_2, A_2, R_2, \ldots S_{T-1}, A_{T-1}, R_T,$

for $t = 0, 1, \ldots, T-1$ **do**

REINFORCE con baseline

Require: $\pi(a \mid s, \boldsymbol{\theta})$, $\hat{v}(s, \mathbf{w})$, $\alpha_{\boldsymbol{\theta}}, \alpha_{\mathbf{w}} > 0$
  Inicialice $\boldsymbol{\theta}, \mathbf{w}$
  repeat
    Episodio $\pi(. \mid , \boldsymbol{\theta})\ :\ S_0, A_0, R_1, S_2, A_2, R_2, \ldots S_{T-1}, A_{T-1}, R_T,$
    for $t = 0, 1, \ldots, T-1$ **do**
      $G \leftarrow \sum_{k=t+1}^{T} \gamma^{k-t-1} R_k$

REINFORCE con baseline

Require: $\pi(a \mid s, \boldsymbol{\theta})$, $\hat{v}(s, \mathbf{w})$, $\alpha_{\boldsymbol{\theta}}, \alpha_{\mathbf{w}} > 0$

Inicialice $\boldsymbol{\theta}, \mathbf{w}$

repeat

Episodio $\pi(. \mid, \boldsymbol{\theta})\ :\ S_0, A_0, R_1, S_2, A_2, R_2, \ldots S_{T-1}, A_{T-1}, R_T,$

for $t = 0, 1, \ldots, T-1$ **do**

$G \leftarrow \sum_{k=t+1}^{T} \gamma^{k-t-1} R_k$

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

REINFORCE con baseline

Require: $\pi(a \mid s, \boldsymbol{\theta})$, $\hat{v}(s, \mathbf{w})$, $\alpha_{\boldsymbol{\theta}}, \alpha_{\mathbf{w}} > 0$

Inicialice $\boldsymbol{\theta}, \mathbf{w}$

repeat

Episodio $\pi(. \mid , \boldsymbol{\theta})\ :\ S_0, A_0, R_1, S_2, A_2, R_2, \ldots S_{T-1}, A_{T-1}, R_T,$

for $t = 0, 1, \ldots, T-1$ **do**

$G \leftarrow \sum_{k=t+1}^{T} \gamma^{k-t-1} R_k$

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

$\mathbf{w} \leftarrow \mathbf{w} + \alpha_{\mathbf{w}} \delta \hat{v}(S_t, \mathbf{w})$

REINFORCE con baseline

Require: $\pi(a \mid s, \boldsymbol{\theta})$, $\hat{v}(s, \mathbf{w})$, $\alpha_{\boldsymbol{\theta}}, \alpha_{\mathbf{w}} > 0$
  Inicialice $\boldsymbol{\theta}, \mathbf{w}$
  repeat
    Episodio $\pi(. \mid , \boldsymbol{\theta})\ :\ S_0, A_0, R_1, S_2, A_2, R_2, \dots S_{T-1}, A_{T-1}, R_T,$
    for $t = 0, 1, \dots, T-1$ **do**
      $G \leftarrow \sum_{k=t+1}^{T} \gamma^{k-t-1} R_k$
      $\delta \leftarrow G_t - \hat{v}(S_t, \mathbf{w})$
      $\mathbf{w} \leftarrow \mathbf{w} + \alpha_{\mathbf{w}} \delta \hat{v}(S_t, \mathbf{w})$
      $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \alpha_{\boldsymbol{\theta}} \gamma^t \delta \nabla_{\boldsymbol{\theta}} \ln\left(\pi(A_t \mid S_t, \boldsymbol{\theta})\right)$

REINFORCE con baseline

Require: $\pi(a \mid s, \boldsymbol{\theta})$, $\hat{v}(s, \mathbf{w})$, $\alpha_{\boldsymbol{\theta}}, \alpha_{\mathbf{w}} > 0$

Inicialice $\boldsymbol{\theta}, \mathbf{w}$

repeat

Episodio $\pi(. \mid , \boldsymbol{\theta})$: $S_0, A_0, R_1, S_2, A_2, R_2, \ldots S_{T-1}, A_{T-1}, R_T,$

for $t = 0, 1, \ldots, T-1$ **do**

$G \leftarrow \sum_{k=t+1}^{T} \gamma^{k-t-1} R_k$

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

$\mathbf{w} \leftarrow \mathbf{w} + \alpha_{\mathbf{w}} \delta \hat{v}(S_t, \mathbf{w})$

$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \alpha_{\boldsymbol{\theta}} \gamma^t \delta \nabla_{\boldsymbol{\theta}} \ln \left(\pi(A_t \mid S_t, \boldsymbol{\theta}) \right)$

end for

REINFORCE con baseline

Require: $\pi(a \mid s, \boldsymbol{\theta})$, $\hat{v}(s, \mathbf{w})$, $\alpha_{\boldsymbol{\theta}}, \alpha_{\mathbf{w}} > 0$

Inicialice $\boldsymbol{\theta}, \mathbf{w}$

repeat

Episodio $\pi(. \mid, \boldsymbol{\theta})\ :\ S_0, A_0, R_1, S_2, A_2, R_2, \ldots S_{T-1}, A_{T-1}, R_T,$

for $t = 0, 1, \ldots, T-1$ **do**

$G \leftarrow \sum_{k=t+1}^{T} \gamma^{k-t-1} R_k$

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

$\mathbf{w} \leftarrow \mathbf{w} + \alpha_{\mathbf{w}} \delta \hat{v}(S_t, \mathbf{w})$

$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \alpha_{\boldsymbol{\theta}} \gamma^t \delta \nabla_{\boldsymbol{\theta}} \ln\left(\pi(A_t \mid S_t, \boldsymbol{\theta})\right)$

end for

until ∞

REINFORCE con baseline

Require: $\pi(a \mid s, \boldsymbol{\theta})$, $\hat{v}(s, \mathbf{w})$, $\alpha_{\boldsymbol{\theta}}, \alpha_{\mathbf{w}} > 0$

Inicialice $\boldsymbol{\theta}, \mathbf{w}$

repeat

  Episodio $\pi(. \mid , \boldsymbol{\theta})\ :\ S_0, A_0, R_1, S_2, A_2, R_2, \ldots S_{T-1}, A_{T-1}, R_T,$

  for $t = 0, 1, \ldots, T-1$ **do**

    $G \leftarrow \sum_{k=t+1}^{T} \gamma^{k-t-1} R_k$

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

    $\mathbf{w} \leftarrow \mathbf{w} + \alpha_{\mathbf{w}} \delta \hat{v}(S_t, \mathbf{w})$

    $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \alpha_{\boldsymbol{\theta}} \gamma^t \delta \nabla_{\boldsymbol{\theta}} \ln\left(\pi(A_t \mid S_t, \boldsymbol{\theta})\right)$

  end for

until ∞

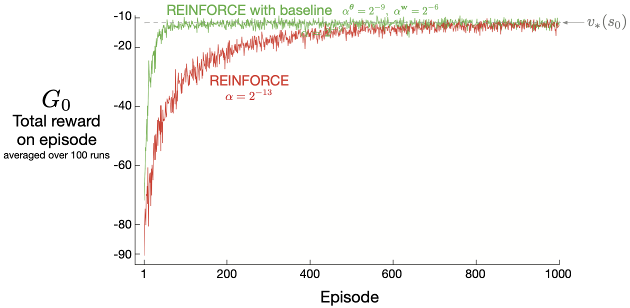


Figure 13.2: Adding a baseline to REINFORCE can make it learn much faster, as illustrated here on the short-corridor gridworld (Example 13.1). The step size used here for plain REINFORCE is that at which it performs best (to the nearest power of two; see Figure 13.1).

Métodos Actor-Crítico

Métodos Actor-Crítico

- Algoritmos de RL que separan función de valor y política:

Métodos Actor-Crítico

- Algoritmos de RL que separan función de valor y política:
  - Actor efectúa acciones

Métodos Actor-Crítico

- Algoritmos de RL que separan función de valor y política:
  - Actor efectúa acciones: $\pi(A_t \mid S_t, \boldsymbol{\theta})$

Métodos Actor-Crítico

- Algoritmos de RL que separan función de valor y política:
  - Actor efectúa acciones: $\pi(A_t \mid S_t, \boldsymbol{\theta})$
  - Crítico: Evalúa resultado de las acciones

Métodos Actor-Crítico

- Algoritmos de RL que separan función de valor y política:
  - Actor efectúa acciones: $\pi(A_t \mid S_t, \boldsymbol{\theta})$
  - Crítico: Evalúa resultado de las acciones: $\hat{v}(S_t, \mathbf{w})$

Métodos Actor-Crítico

- Algoritmos de RL que separan función de valor y política:
  - Actor efectúa acciones: $\pi(A_t \mid S_t, \boldsymbol{\theta})$
  - Crítico: Evalúa resultado de las acciones: $\hat{v}(S_t, \mathbf{w})$
- Incorporan bootstrapping:

Métodos Actor-Crítico

- Algoritmos de RL que separan función de valor y política:
  - Actor efectúa acciones: $\pi(A_t \mid S_t, \boldsymbol{\theta})$
  - Crítico: Evalúa resultado de las acciones: $\hat{v}(S_t, \mathbf{w})$
- Incorporan bootstrapping:
  - Reduce varianza.

Métodos Actor-Crítico

- Algoritmos de RL que separan función de valor y política:
  - Actor efectúa acciones: $\pi(A_t \mid S_t, \boldsymbol{\theta})$
  - Crítico: Evalúa resultado de las acciones: $\hat{v}(S_t, \mathbf{w})$
- Incorporan bootstrapping:
  - Reduce varianza.
  - Introduce sesgo.

Métodos Actor-Crítico

- Algoritmos de RL que separan función de valor y política:
  - Actor efectúa acciones: $\pi(A_t \mid S_t, \boldsymbol{\theta})$
  - Crítico: Evalúa resultado de las acciones: $\hat{v}(S_t, \mathbf{w})$
- Incorporan bootstrapping:
  - Reduce varianza.
  - Introduce sesgo.
- Actor-crítico de un paso:

Métodos Actor-Crítico

- Algoritmos de RL que separan función de valor y política:
  - Actor efectúa acciones: $\pi(A_t \mid S_t, \boldsymbol{\theta})$
  - Crítico: Evalúa resultado de las acciones: $\hat{v}(S_t, \mathbf{w})$
- Incorporan bootstrapping:
  - Reduce varianza.
  - Introduce sesgo.
- Actor-crítico de un paso:

$$\boldsymbol{\theta}_{t+1} \doteq \boldsymbol{\theta}_t + \alpha(G_{t:t+1} - \hat{v}(S_t, \mathbf{w}))\nabla_{\boldsymbol{\theta}} \ln\left(\pi(A_t \mid S_t, \boldsymbol{\theta})\right)$$

Métodos Actor-Crítico

- Algoritmos de RL que separan función de valor y política:
  - Actor efectúa acciones: $\pi(A_t \mid S_t, \boldsymbol{\theta})$
  - Crítico: Evalúa resultado de las acciones: $\hat{v}(S_t, \mathbf{w})$
- Incorporan bootstrapping:
  - Reduce varianza.
  - Introduce sesgo.
- Actor-crítico de un paso:

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

Métodos Actor-Crítico

- Algoritmos de RL que separan función de valor y política:
  - Actor efectúa acciones: $\pi(A_t \mid S_t, \boldsymbol{\theta})$
  - Crítico: Evalúa resultado de las acciones: $\hat{v}(S_t, \mathbf{w})$
- Incorporan bootstrapping:
  - Reduce varianza.
  - Introduce sesgo.
- Actor-crítico de un paso:

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

Actor Crítico de un paso- Episódico

Require: $\pi(a \mid s, \boldsymbol{\theta})$, $\hat{v}(s, \mathbf{w})$, $\alpha_{\boldsymbol{\theta}}, \alpha_{\mathbf{w}} > 0$

Actor Crítico de un paso- Episódico

Require: $\pi(a \mid s, \boldsymbol{\theta})$, $\hat{v}(s, \mathbf{w})$, $\alpha_{\boldsymbol{\theta}}, \alpha_{\mathbf{w}} > 0$

Inicialice $\boldsymbol{\theta}, \mathbf{w}$

Actor Crítico de un paso- Episódico

Require: $\pi(a \mid s, \boldsymbol{\theta})$, $\hat{v}(s, \mathbf{w})$, $\alpha_{\boldsymbol{\theta}}, \alpha_{\mathbf{w}} > 0$

Inicialice $\boldsymbol{\theta}, \mathbf{w}$

repeat

Actor Crítico de un paso- Episódico

Require: $\pi(a \mid s, \boldsymbol{\theta})$, $\hat{v}(s, \mathbf{w})$, $\alpha_{\boldsymbol{\theta}}, \alpha_{\mathbf{w}} > 0$

Incialice $\boldsymbol{\theta}, \mathbf{w}$

repeat

$k = 0$

Actor Crítico de un paso- Episódico

Require: $\pi(a \mid s, \boldsymbol{\theta})$, $\hat{v}(s, \mathbf{w})$, $\alpha_{\boldsymbol{\theta}}, \alpha_{\mathbf{w}} > 0$

Inicialice $\boldsymbol{\theta}, \mathbf{w}$

repeat

$k = 0$

Inicialice S

Actor Crítico de un paso- Episódico

Require: $\pi(a \mid s, \boldsymbol{\theta})$, $\hat{v}(s, \mathbf{w})$, $\alpha_{\boldsymbol{\theta}}, \alpha_{\mathbf{w}} > 0$

  Inicialice $\boldsymbol{\theta}, \mathbf{w}$

  repeat

    $k = 0$

    Inicialice S

    $A \sim \pi(. \mid S, \boldsymbol{\theta})$

Actor Crítico de un paso- Episódico

Require: $\pi(a \mid s, \boldsymbol{\theta})$, $\hat{v}(s, \mathbf{w})$, $\alpha_{\boldsymbol{\theta}}, \alpha_{\mathbf{w}} > 0$

Inicialice $\boldsymbol{\theta}, \mathbf{w}$

repeat

$k = 0$

Inicialice S

$A \sim \pi(. \mid S, \boldsymbol{\theta})$

while S no terminal **do**

Actor Crítico de un paso- Episódico

Require: $\pi(a \mid s, \boldsymbol{\theta})$, $\hat{v}(s, \mathbf{w})$, $\alpha_{\boldsymbol{\theta}}, \alpha_{\mathbf{w}} > 0$
  Inicialice $\boldsymbol{\theta}, \mathbf{w}$
  repeat
    $k = 0$
    Inicialice S
    $A \sim \pi(. \mid S, \boldsymbol{\theta})$
    while S no terminal **do**
      Tome Acción A, observe S', R

Actor Crítico de un paso- Episódico

Require: $\pi(a \mid s, \boldsymbol{\theta})$, $\hat{v}(s, \mathbf{w})$, $\alpha_{\boldsymbol{\theta}}, \alpha_{\mathbf{w}} > 0$

Inicialice $\boldsymbol{\theta}, \mathbf{w}$

repeat

$k = 0$

Inicialice S

$A \sim \pi(. \mid S, \boldsymbol{\theta})$

while S no terminal **do**

Tome Acción A, observe S', R

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

Actor Crítico de un paso- Episódico

Require: $\pi(a \mid s, \boldsymbol{\theta})$, $\hat{v}(s, \mathbf{w})$, $\alpha_{\boldsymbol{\theta}}, \alpha_{\mathbf{w}} > 0$

Inicialice $\boldsymbol{\theta}, \mathbf{w}$

repeat

$k = 0$

Inicialice S

$A \sim \pi(. \mid S, \boldsymbol{\theta})$

while S no terminal **do**

Tome Acción A, observe S', R

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

$\mathbf{w} \leftarrow \mathbf{w} + \alpha_{\mathbf{w}} \delta \hat{v}(S, \mathbf{w})$

Actor Crítico de un paso- Episódico

Require: $\pi(a \mid s, \boldsymbol{\theta})$, $\hat{v}(s, \mathbf{w})$, $\alpha_{\boldsymbol{\theta}}, \alpha_{\mathbf{w}} > 0$

Inicialice $\boldsymbol{\theta}, \mathbf{w}$

repeat

$k = 0$

Inicialice S

$A \sim \pi(. \mid S, \boldsymbol{\theta})$

while S no terminal **do**

Tome Acción A, observe S', R

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

$\mathbf{w} \leftarrow \mathbf{w} + \alpha_{\mathbf{w}} \delta \hat{v}(S, \mathbf{w})$

$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \alpha_{\boldsymbol{\theta}} \gamma^k \delta \nabla_{\boldsymbol{\theta}} \ln \left(\pi(A_t \mid S_t, \boldsymbol{\theta}) \right)$

Actor Crítico de un paso- Episódico

Require: $\pi(a \mid s, \boldsymbol{\theta})$, $\hat{v}(s, \mathbf{w})$, $\alpha_{\boldsymbol{\theta}}, \alpha_{\mathbf{w}} > 0$

Inicialice $\boldsymbol{\theta}, \mathbf{w}$

repeat

$k = 0$

Inicialice S

$A \sim \pi(. \mid S, \boldsymbol{\theta})$

while S no terminal **do**

Tome Acción A, observe S', R

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

$\mathbf{w} \leftarrow \mathbf{w} + \alpha_{\mathbf{w}} \delta \hat{v}(S, \mathbf{w})$

$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \alpha_{\boldsymbol{\theta}} \gamma^k \delta \nabla_{\boldsymbol{\theta}} \ln\left(\pi(A_t \mid S_t, \boldsymbol{\theta})\right)$

$k \leftarrow k + 1$

Actor Crítico de un paso- Episódico

Require: $\pi(a \mid s, \boldsymbol{\theta})$, $\hat{v}(s, \mathbf{w})$, $\alpha_{\boldsymbol{\theta}}, \alpha_{\mathbf{w}} > 0$

Inicialice $\boldsymbol{\theta}, \mathbf{w}$

repeat

$k = 0$

Inicialice S

$A \sim \pi(. \mid S, \boldsymbol{\theta})$

while S no terminal **do**

Tome Acción A, observe S', R

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

$\mathbf{w} \leftarrow \mathbf{w} + \alpha_{\mathbf{w}} \delta \hat{v}(S, \mathbf{w})$

$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \alpha_{\boldsymbol{\theta}} \gamma^k \delta \nabla_{\boldsymbol{\theta}} \ln\left(\pi(A_t \mid S_t, \boldsymbol{\theta})\right)$

$k \leftarrow k + 1$

$S \leftarrow S'$

Actor Crítico de un paso- Episódico

Require: $\pi(a \mid s, \boldsymbol{\theta})$, $\hat{v}(s, \mathbf{w})$, $\alpha_{\boldsymbol{\theta}}, \alpha_{\mathbf{w}} > 0$

Inicialice $\boldsymbol{\theta}, \mathbf{w}$

repeat

$k = 0$

Inicialice S

$A \sim \pi(. \mid S, \boldsymbol{\theta})$

while S no terminal **do**

Tome Acción A, observe S', R

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

$\mathbf{w} \leftarrow \mathbf{w} + \alpha_{\mathbf{w}} \delta \hat{v}(S, \mathbf{w})$

$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \alpha_{\boldsymbol{\theta}} \gamma^k \delta \nabla_{\boldsymbol{\theta}} \ln \left(\pi(A_t \mid S_t, \boldsymbol{\theta})\right)$

$k \leftarrow k + 1$

$S \leftarrow S'$

end while

Actor Crítico de un paso- Episódico

Require: $\pi(a \mid s, \boldsymbol{\theta})$, $\hat{v}(s, \mathbf{w})$, $\alpha_{\boldsymbol{\theta}}, \alpha_{\mathbf{w}} > 0$

  Inicialice $\boldsymbol{\theta}, \mathbf{w}$

  repeat

    $k = 0$

    Inicialice S

    $A \sim \pi(. \mid S, \boldsymbol{\theta})$

    while S no terminal **do**

      Tome Acción A, observe S', R

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

      $\mathbf{w} \leftarrow \mathbf{w} + \alpha_{\mathbf{w}} \delta \hat{v}(S, \mathbf{w})$

      $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \alpha_{\boldsymbol{\theta}} \gamma^k \delta \nabla_{\boldsymbol{\theta}} \ln\left(\pi(A_t \mid S_t, \boldsymbol{\theta})\right)$

      $k \leftarrow k + 1$

      $S \leftarrow S'$

    end while

  until ∞

Actor Crítico de un paso- Episódico

Require: $\pi(a \mid s, \boldsymbol{\theta})$, $\hat{v}(s, \mathbf{w})$, $\alpha_{\boldsymbol{\theta}}, \alpha_{\mathbf{w}} > 0$

Inicialice $\boldsymbol{\theta}, \mathbf{w}$

repeat

$k = 0$

Inicialice S

$A \sim \pi(. \mid S, \boldsymbol{\theta})$

while S no terminal **do**

Tome Acción A, observe S', R

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

$\mathbf{w} \leftarrow \mathbf{w} + \alpha_{\mathbf{w}} \delta \hat{v}(S, \mathbf{w})$

$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \alpha_{\boldsymbol{\theta}} \gamma^k \delta \nabla_{\boldsymbol{\theta}} \ln \left(\pi(A_t \mid S_t, \boldsymbol{\theta})\right)$

$k \leftarrow k + 1$

$S \leftarrow S'$

end while

until ∞

Problemas no episódicos

$$J(\boldsymbol{\theta}) = r(\pi) \doteq \lim_{h\to\infty} \frac{1}{h} \sum_{t=1}^{h} \mathbb{E}\left[R_t \;:\; S_0, A_{0:t-1} \sim \pi\right]$$

Problemas no episódicos

$$
\begin{aligned}
J(\boldsymbol{\theta}) = r(\pi) &\doteq \lim_{h\to\infty} \frac{1}{h}\sum_{t=1}^{h} \mathbb{E}\left[R_t \;:\; S_0, A_{0:t-1} \sim \pi\right] \\
&= \sum_{s} \mu_\pi(s) \sum_{a} \pi(a \mid s) \sum_{s',r} p(s', r \mid s, a) r
\end{aligned}
$$

Problemas no episódicos

$$
\begin{aligned}
J(\boldsymbol{\theta}) = r(\pi) &\doteq \lim_{h\to\infty} \frac{1}{h} \sum_{t=1}^{h} \mathbb{E}\left[R_t \;:\; S_0, A_{0:t-1} \sim \pi\right] \\
&= \sum_{s} \mu_\pi(s) \sum_{a} \pi(a \mid s) \sum_{s',r} p(s', r \mid s, a) r
\end{aligned}
$$

- $\mu_\pi(s)$ es la distribución de estado estable, independiente de S_0:

$$
\mu_\pi(s) = \lim_{t\to\infty} \mathbf{P}\left\{S_t = s \mid A_{0:t-1} \sim \pi\right\}
$$

(MDP ergódico)

Problemas no episódicos

$$
\begin{aligned}
J(\boldsymbol{\theta}) = r(\pi) &\doteq \lim_{h\to\infty} \frac{1}{h} \sum_{t=1}^{h} \mathbb{E}\left[R_t \;:\; S_0, A_{0:t-1} \sim \pi\right] \\
&= \sum_s \mu_\pi(s) \sum_a \pi(a \mid s) \sum_{s',r} p(s', r \mid s, a) r
\end{aligned}
$$

- $\mu_\pi(s)$ es la distribución de estado estable, independiente de S_0:

$$
\mu_\pi(s) = \lim_{t\to\infty} \mathbf{P}\left\{S_t = s \mid A_{0:t-1} \sim \pi\right\}
$$

(MDP ergódico)

- Satisface:

$$
\sum_s \mu_\pi(s) \sum_a \pi(a \mid s) p(s' \mid s, a) = \mu_\pi(s')
$$

- Retorno diferencial:

$$G_t \doteq R_{t+1} - r(\pi) + R_{t+2} - r(\pi) + R_{t+3} - r(\pi) + \cdots + \ldots$$

- Retorno diferencial:

$$G_t \doteq R_{t+1} - r(\pi) + R_{t+2} - r(\pi) + R_{t+3} - r(\pi) + \cdots + \ldots$$

- Función de valor de acción

$$q_\pi(s, a) \doteq \mathbb{E}_\pi \left[G_t \mid S_t = s, A_t = a \right]$$

- Retorno diferencial:

$$G_t \doteq R_{t+1} - r(\pi) + R_{t+2} - r(\pi) + R_{t+3} - r(\pi) + \cdots + \ldots$$

- Función de valor de acción

$$q_\pi(s,a) \doteq \mathbb{E}_\pi \left[G_t \mid S_t = s, A_t = a \right]$$

- Criterio a maximizar: $J(\boldsymbol{\theta}) = r(\pi)$

- Retorno diferencial:

$$G_t \doteq R_{t+1} - r(\pi) + R_{t+2} - r(\pi) + R_{t+3} - r(\pi) + \cdots + \ldots$$

- Función de valor de acción

$$q_\pi(s,a) \doteq \mathbb{E}_\pi\left[G_t \mid S_t = s, A_t = a\right]$$

- Criterio a maximizar: $J(\boldsymbol{\theta}) = r(\pi)$
- Teorema de gradiente de política:

$$\nabla J(\boldsymbol{\theta}) = \sum_s \mu_\pi(s) \sum_a q_\pi(s,a) \nabla_{\boldsymbol{\theta}} \pi(a \mid s, \boldsymbol{\theta})$$

Acciones contínuas

Acciones contínuas

- $\pi(a \mid s, \boldsymbol{\theta}) \rightarrow$ función de densidad de probabilidad.

Acciones contínuas

- $\pi(a \mid s, \boldsymbol{\theta}) \rightarrow$ función de densidad de probabilidad.
- Por ejemplo, si $a \in \mathbb{R}$:

$$\pi(a \mid s, [\boldsymbol{\theta}_\mu, \boldsymbol{\theta}_\sigma]^T) = \frac{1}{\sqrt{2\pi}\sigma(s, \boldsymbol{\theta}_\sigma)} \exp\left(-\frac{(a - \mu(s, \boldsymbol{\theta}_\mu))^2}{2\sigma(s, \boldsymbol{\theta}_\sigma)}\right)^2$$

Acciones contínuas

- $\pi(a \mid s, \boldsymbol{\theta}) \rightarrow$ función de densidad de probabilidad.
- Por ejemplo, si $a \in \mathbb{R}$:

$$\pi(a \mid s, [\boldsymbol{\theta}_\mu, \boldsymbol{\theta}_\sigma]^T) = \frac{1}{\sqrt{2\pi}\sigma(s, \boldsymbol{\theta}_\sigma)} \exp\left(-\frac{(a - \mu(s, \boldsymbol{\theta}_\mu))^2}{2\sigma(s, \boldsymbol{\theta}_\sigma)}\right)^2$$

  - Si $\mu(s, \boldsymbol{\theta}) = \boldsymbol{\theta}_\mu^T \mathbf{x}_\mu(s)$ y $\sigma(s, \boldsymbol{\theta}) = \exp\left(\boldsymbol{\theta}_\sigma^T \mathbf{x}_\sigma(s)\right)$

Acciones contínuas

- $\pi(a \mid s, \boldsymbol{\theta}) \rightarrow$ función de densidad de probabilidad.
- Por ejemplo, si $a \in \mathbb{R}$:

$$\pi(a \mid s, [\boldsymbol{\theta}_\mu, \boldsymbol{\theta}_\sigma]^T) = \frac{1}{\sqrt{2\pi}\sigma(s, \boldsymbol{\theta}_\sigma)} \exp\left(-\frac{(a - \mu(s, \boldsymbol{\theta}_\mu))^2}{2\sigma(s, \boldsymbol{\theta}_\sigma)}\right)^2$$

  - Si $\mu(s, \boldsymbol{\theta}) = \boldsymbol{\theta}_\mu^T \mathbf{x}_\mu(s)$ y $\sigma(s, \boldsymbol{\theta}) = \exp\left(\boldsymbol{\theta}_\sigma^T \mathbf{x}_\sigma(s)\right)$

$$\nabla \ln(\pi(a \mid s, [\boldsymbol{\theta}_\mu, \boldsymbol{\theta}_\sigma]^T)) = \begin{bmatrix} \frac{1}{\sigma(s,\boldsymbol{\theta})^2}(a - \mu(s, \boldsymbol{\theta}))\mathbf{x}_\mu(s)) \\ \left(\frac{(a-\mu(s,\boldsymbol{\theta}))^2}{\sigma(s,\boldsymbol{\theta})^2} - 1\right)\mathbf{x}_\sigma(s) \end{bmatrix}$$