

Preuves Cours 4 RLD

1 Preuve causalité

On souhaite montrer que les décisions à t n'affectent en rien les récompenses obtenues à t' , avec $t' < t$ (ce qu'on appelle causalité). Cela revient à montrer que :

$$\nabla_{\theta} J(\theta) = \sum_{\tau} \pi_{\theta}(\tau) \left[\sum_{t=0}^{|\tau|-1} \nabla_{\theta} \log \pi_{\theta}(a_t | s_t) \sum_{t'=t}^{|\tau|} r_{t'} \right]$$

avec r_t le reward obtenu selon $\mathcal{R}(s_t, a_t, s_{t+1})$ dans τ .

Cette relation simplifie grandement la formulation du gradient et réduit considérablement sa variance (du moins pour les décisions de fin de trajectoire).

On a :

$$\begin{aligned} \nabla_{\theta} J(\theta) &= \mathbb{E}_{\tau \sim \pi_{\theta}} [R(\tau) \nabla_{\theta} \log \pi_{\theta}(\tau)] \\ &= \sum_{\tau} \pi_{\theta}(\tau) \left[\left(\sum_{t=0}^{|\tau|-1} \nabla_{\theta} \log \pi_{\theta}(a_t | s_t) \right) \left(\sum_{t=0}^{|\tau|} r_t \right) \right] \\ &= \sum_{\tau} \pi_{\theta}(\tau) \left[\sum_{t=0}^{|\tau|-1} \nabla_{\theta} \log \pi_{\theta}(a_t | s_t) \sum_{t'=t}^{|\tau|} r_{t'} \right] + \nabla_{\theta} C(\theta) \end{aligned}$$

Notre problème revient alors à montrer que :

$$\nabla_{\theta} C(\theta) = \sum_{\tau} \pi_{\theta}(\tau) \left[\sum_{t=0}^{|\tau|-1} \nabla_{\theta} \log \pi_{\theta}(a_t | s_t) \sum_{t'=0}^{t-1} r_{t'} \right] = 0$$

Sans perte de généralité, on suppose que toutes les trajectoires font la même taille T . On a alors :

$$\nabla_{\theta} C(\theta) = \sum_{\tau} \pi_{\theta}(\tau) \left[\sum_{t=0}^{|\tau|-1} \nabla_{\theta} \log \pi_{\theta}(a_t | s_t) \sum_{t'=0}^{t-1} r_{t'} \right] = \sum_{t=0}^{T-1} \sum_{\tau} \pi_{\theta}(\tau) \nabla_{\theta} \log \pi_{\theta}(a_t | s_t) \sum_{t'=0}^{t-1} r_{t'}$$

On peut décomposer la trajectoire :

$$\begin{aligned} \nabla_{\theta} C(\theta) &= \sum_{t=0}^{T-1} \sum_{\tau_{0:t-1}} \pi_{\theta}(\tau_{0:t-1}) \sum_{\tau_t} \pi_{\theta}(\tau_t | \tau_{0:t-1}) \sum_{\tau_{t+1:T}} \pi_{\theta}(\tau_{t+1:T} | \tau_{0:t}) \nabla_{\theta} \log \pi_{\theta}(a_t | s_t) \sum_{t'=0}^{t-1} r_{t'} \\ &= \sum_{t=0}^{T-1} \sum_{\tau_{0:t-1}} \pi_{\theta}(\tau_{0:t-1}) \sum_{t'=0}^{t-1} r_{t'} \sum_{\tau_t} \pi_{\theta}(\tau_t | \tau_{0:t-1}) \nabla_{\theta} \log \pi_{\theta}(a_t | s_t) \sum_{\tau_{t+1:T}} \pi_{\theta}(\tau_{t+1:T} | \tau_{0:t}) \\ &= \sum_{t=0}^{T-1} \sum_{\tau_{0:t-1}} \pi_{\theta}(\tau_{0:t-1}) \sum_{t'=0}^{t-1} r_{t'} \sum_{\tau_t} \pi_{\theta}(\tau_t | \tau_{0:t-1}) \nabla_{\theta} \log \pi_{\theta}(a_t | s_t) \end{aligned}$$

avec $\tau_t = (s_t, a_t)$ et $\tau_{i:j} = ((s_i, a_i), \dots, (s_j, a_j))$.

Or on a : $\pi_{\theta}(\tau_t | \tau_{0:t-1}) = \pi_{\theta}(a_t | s_t) P(s_t | s_{t-1}, a_{t-1})$. Donc :

$$\begin{aligned} \pi_{\theta}(\tau_t | \tau_{0:t-1}) \nabla_{\theta} \log \pi_{\theta}(a_t | s_t) &= \pi_{\theta}(\tau_t | \tau_{0:t-1}) (\nabla_{\theta} \log \pi_{\theta}(a_t | s_t) + \nabla_{\theta} \log P(s_t | s_{t-1}, a_{t-1})) \\ &= \pi_{\theta}(\tau_t | \tau_{0:t-1}) \nabla_{\theta} \log \pi_{\theta}(\tau_t | \tau_{0:t-1}) \\ &= \nabla_{\theta} \pi_{\theta}(\tau_t | \tau_{0:t-1}) \end{aligned}$$

On a alors :

$$\sum_{\tau_t} \pi_{\theta}(\tau_t | \tau_{0:t-1}) \nabla_{\theta} \log \pi_{\theta}(a_t | s_t) = \sum_{\tau_t} \nabla_{\theta} \pi_{\theta}(\tau_t | \tau_{0:t-1}) = \nabla_{\theta} \sum_{\tau_t} \pi_{\theta}(\tau_t | \tau_{0:t-1}) = \nabla_{\theta} 1 = 0$$

On en conclut donc que $\nabla_{\theta} C(\theta) = 0$

2 Gradient Actor-Critic

On souhaite montrer que le Policy Gradient $\nabla_{\theta} J(\theta)$ peut s'écrire :

$$\nabla_{\theta} J(\theta) = \sum_{\tau} \pi_{\theta}(\tau) \left[\sum_{t=0}^{|\tau|-1} \nabla_{\theta} \log \pi_{\theta}(a_t | s_t) \gamma^t Q^{\pi}(s_t, a_t) \right]$$

Commençons par remarquer que :

$$\begin{aligned} J(\theta) &= \sum_{\tau} \pi(\tau) \left[\sum_{t=0}^{\infty} \gamma^t r_t \right] \\ &= \sum_{s_0} P(s_0) \sum_{a_0} \pi(a_0 | s_0) Q^{\pi}(s_0, a_0) \end{aligned}$$

On a alors :

$$\nabla_{\theta} J(\theta) = \sum_{s_0} P(s_0) \sum_{a_0} \pi(a_0 | s_0) \left(\nabla_{\theta} \log(\pi(a_0 | s_0)) Q^{\pi}(s_0, a_0) + \nabla_{\theta} Q^{\pi}(s_0, a_0) \right)$$

Or pour tout s_t, a_t :

$$\begin{aligned} \nabla_{\theta} Q^{\pi}(s_t, a_t) &= \nabla_{\theta} \left(\sum_{s_{t+1}} P(s_{t+1} | s_t, a_t) \left(r_t + \gamma \sum_{a_{t+1}} \pi(a_{t+1} | s_{t+1}) Q^{\pi}(s_{t+1}, a_{t+1}) \right) \right) \\ &= \sum_{s_{t+1}} P(s_{t+1} | s_t, a_t) \gamma \sum_{a_{t+1}} \pi(a_{t+1} | s_{t+1}) \left(\nabla_{\theta} \log(\pi(a_{t+1} | s_{t+1})) Q^{\pi}(s_{t+1}, a_{t+1}) + \nabla_{\theta} Q^{\pi}(s_{t+1}, a_{t+1}) \right) \end{aligned}$$

En rassemblant tous les $\nabla_{\theta} Q^{\pi}$, on a alors par chain-rule :

$$\nabla_{\theta} J(\theta) = \sum_{\tau} \pi_{\theta}(\tau) \left[\sum_{t=0}^{|\tau|-1} \nabla_{\theta} \log \pi_{\theta}(a_t | s_t) \gamma^t Q^{\pi}(s_t, a_t) \right]$$

3 Traces d'éligibilité pour GAE

On souhaite montrer que faire des mises à jour de la politique à chaque étape t de la trajectoire selon :

$$\theta \leftarrow \theta + \alpha \delta_t^V e_t$$

avec

$$\delta_t^V = r_t + \gamma V(s_{t+1}) - V(s_t)$$

et e_t la trace d'éligibilité définie comme :

$$\begin{aligned} e_0 &\leftarrow \nabla_0 \\ e_t &\leftarrow \lambda \gamma e_{t-1} + \nabla_t \end{aligned}$$

avec $\nabla_t := \nabla_{\theta} \log \pi_{\theta}(a_t | s_t)$, est équivalent (si travail sur une copie des paramètres) à faire une mise à jour globale en fin de trajectoire correspondant à :

$$\theta \leftarrow \theta + \alpha \hat{g}$$

avec

$$\begin{aligned}\hat{g} &= \sum_{t=0}^{\infty} \hat{A}_t^{GAE(\gamma,\lambda)} \nabla_t \\ &= \sum_{t=0}^{\infty} \nabla_t \sum_{l=0}^{\infty} (\gamma\lambda)^l \delta_{t+l}^V\end{aligned}$$

On a :

$$\begin{aligned}\hat{g} &= \nabla_0 \left(\delta_0^V + (\gamma\lambda)\delta_1^V + (\gamma\lambda)^2\delta_2^V + \dots \right) \\ &+ \nabla_1 \left(\delta_1^V + (\gamma\lambda)\delta_2^V + (\gamma\lambda)^2\delta_3^V + \dots \right) \\ &+ \nabla_2 \left(\delta_2^V + (\gamma\lambda)\delta_3^V + (\gamma\lambda)^2\delta_4^V + \dots \right) \\ &+ \dots\end{aligned}$$

En regroupant les δ^V on obtient :

$$\begin{aligned}\hat{g} &= \delta_0^V \left(\nabla_0 \right) \\ &+ \delta_1^V \left(\nabla_1 + (\gamma\lambda)\nabla_0 \right) \\ &+ \delta_2^V \left(\nabla_2 + (\gamma\lambda)\nabla_1 + (\gamma\lambda)^2\nabla_0 \right) \\ &+ \dots \\ &= \sum_{t=0}^{\infty} \delta_t^V e_t\end{aligned}$$

On a donc l'expression de \hat{g} sous la forme d'une somme dont chaque composante ne dépend que d'éléments obtenus avant t .

On peut alors mettre à jour θ à chaque étape t selon :

$$\theta \leftarrow \theta + \alpha \delta_t^V e_t$$

ce qui revient à la même chose que de considérer une mise à jour globale selon

$$\theta \leftarrow \theta + \alpha \hat{g} = \theta + \alpha \sum_{t=0}^{\infty} \delta_t^V e_t$$

4 Fonctions compatibles

Soit le gradient : $\hat{g} = \sum_{\tau} \pi_{\theta}(\tau) \left[\sum_{t=0}^{|\tau|-1} \gamma^t \nabla_{\theta} \log \pi_{\theta}(a_t | s_t) f_{\phi}(s_t, a_t) \right]$, avec f_{ϕ} une fonction $\mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$ de paramètres ϕ . On souhaite montrer qu'une condition suffisante pour rendre ce gradient non biaisé (i.e., $\hat{g} = \nabla_{\theta} J(\theta)$) est d'utiliser une fonction f_{ϕ} compatible, c'est à dire respectant les deux contraintes suivantes :

- Pour tout s et a : $\nabla_{\phi} f_{\phi}(s, a) = \frac{\nabla_{\theta} \pi_{\theta}(a|s)}{\pi_{\theta}(a|s)}$
- $\sum_{\tau} \pi_{\theta}(\tau) \left[\sum_{t=0}^{|\tau|-1} \gamma^t (Q^{\pi}(s_t, a_t) - f_{\phi}(s_t, a_t) - v_w(s_t)) \nabla_{\phi} f_{\phi}(s_t, a_t) \right] = 0$

avec $v_w(s)$ une fonction quelconque $\mathcal{S} \rightarrow \mathbb{R}$ de paramètres w .

Selon la seconde contrainte on a :

$$\sum_{\tau} \pi_{\theta}(\tau) \left[\sum_{t=0}^{|\tau|-1} \gamma^t (Q^{\pi}(s_t, a_t) - f_{\phi}(s_t, a_t) - v_w(s_t)) \nabla_{\phi} f_{\phi}(s_t, a_t) \right] = 0$$

or :

$$\begin{aligned} & \sum_{\tau} \pi_{\theta}(\tau) \left[\sum_{t=0}^{|\tau|-1} \gamma^t (Q^{\pi}(s_t, a_t) - f_{\phi}(s_t, a_t) - v_w(s_t)) \nabla_{\phi} f_{\phi}(s_t, a_t) \right] \\ &= \sum_{t=0}^{\infty} \gamma^t \sum_{\tau_0:t-1} \pi_{\theta}(\tau_0:t-1) \sum_{s_t \in \mathcal{S}} P(s_t | s_{t-1}, a_{t-1}) \sum_{a \in \mathcal{A}(s_t)} \pi_{\theta}(a | s_t) (Q^{\pi}(s_t, a) - f_{\phi}(s_t, a) - v_w(s_t)) \nabla_{\phi} f_{\phi}(s_t, a) \\ &= \sum_{t=0}^{\infty} \gamma^t \sum_{\tau} \pi_{\theta}(\tau) \sum_{a \in \mathcal{A}(s_t)} \pi_{\theta}(a | s_t) (Q^{\pi}(s_t, a) - f_{\phi}(s_t, a) - v_w(s_t)) \nabla_{\phi} f_{\phi}(s_t, a) \\ &= \sum_{\tau} \pi_{\theta}(\tau) \left[\sum_{t=0}^{\infty} \gamma^t \sum_{a \in \mathcal{A}(s_t)} \pi_{\theta}(a | s_t) (Q^{\pi}(s_t, a) - f_{\phi}(s_t, a) - v_w(s_t)) \nabla_{\phi} f_{\phi}(s_t, a) \right] \end{aligned}$$

On a donc :

$$\sum_{\tau} \pi_{\theta}(\tau) \left[\sum_{t=0}^{\infty} \gamma^t \sum_{a \in \mathcal{A}(s_t)} \pi_{\theta}(a | s_t) (Q^{\pi}(s_t, a) - f_{\phi}(s_t, a) - v_w(s_t)) \nabla_{\phi} f_{\phi}(s_t, a) \right] = 0$$

Si la première contrainte est respectée, on peut remplacer $\nabla_{\phi} f_{\phi}(s, a)$ par son expression $\frac{\nabla_{\theta} \pi_{\theta}(a | s)}{\pi_{\theta}(a | s)}$ dans cette équation :

$$\sum_{\tau} \pi_{\theta}(\tau) \left[\sum_{t=0}^{\infty} \gamma^t \sum_{a \in \mathcal{A}(s_t)} \nabla_{\theta} \pi_{\theta}(a | s_t) (Q^{\pi}(s_t, a) - f_{\phi}(s_t, a) - v_w(s_t)) \right] = 0$$

Ou encore :

$$\sum_{\tau} \pi_{\theta}(\tau) \left[\sum_{t=0}^{\infty} \gamma^t \sum_{a \in \mathcal{A}(s_t)} \nabla_{\theta} \pi_{\theta}(a | s_t) (Q^{\pi}(s_t, a) - f_{\phi}(s_t, a)) \right] = 0$$

Car :

$$\begin{aligned} \sum_{\tau} \pi_{\theta}(\tau) \left[\sum_{t=0}^{\infty} \gamma^t \sum_{a \in \mathcal{A}(s_t)} \nabla_{\theta} \pi_{\theta}(a | s_t) v_w(s_t) \right] &= \sum_{\tau} \pi_{\theta}(\tau) \left[\sum_{t=0}^{\infty} \gamma^t v_w(s_t) \sum_{a \in \mathcal{A}(s_t)} \nabla_{\theta} \pi_{\theta}(a | s_t) \right] \\ &= \sum_{\tau} \pi_{\theta}(\tau) \left[\sum_{t=0}^{\infty} \gamma^t v_w(s_t) \nabla_{\theta} 1 \right] = 0 \end{aligned}$$

On a alors :

$$\sum_{\tau} \pi_{\theta}(\tau) \left[\sum_{t=0}^{\infty} \gamma^t \sum_{a \in \mathcal{A}(s_t)} \nabla_{\theta} \pi_{\theta}(a | s_t) Q^{\pi}(s_t, a) \right] = \sum_{\tau} \pi_{\theta}(\tau) \left[\sum_{t=0}^{\infty} \gamma^t \sum_{a \in \mathcal{A}(s_t)} \nabla_{\theta} \pi_{\theta}(a | s_t) f_{\phi}(s_t, a) \right]$$

Et donc :

$$\begin{aligned}
\hat{g} &= \sum_{\tau} \pi_{\theta}(\tau) \left[\sum_{t=0}^{\infty} \gamma^t \nabla_{\theta} \log \pi_{\theta}(a_t | s_t) f_{\phi}(s_t, a_t) \right] = \sum_{\tau} \pi_{\theta}(\tau) \left[\sum_{t=0}^{\infty} \gamma^t \sum_{a \in \mathcal{A}(s_t)} \pi_{\theta}(a | s_t) \nabla_{\theta} \log \pi_{\theta}(a | s_t) f_{\phi}(s_t, a) \right] \\
&= \sum_{\tau} \pi_{\theta}(\tau) \left[\sum_{t=0}^{\infty} \gamma^t \sum_{a \in \mathcal{A}(s_t)} \nabla_{\theta} \pi_{\theta}(a | s_t) f_{\phi}(s_t, a) \right] \\
&= \sum_{\tau} \pi_{\theta}(\tau) \left[\sum_{t=0}^{\infty} \gamma^t \sum_{a \in \mathcal{A}(s_t)} \nabla_{\theta} \pi_{\theta}(a | s_t) Q^{\pi}(s_t, a) \right] \\
&= \sum_{\tau} \pi_{\theta}(\tau) \left[\sum_{t=0}^{\infty} \gamma^t \sum_{a \in \mathcal{A}(s_t)} \pi_{\theta}(a | s_t) \nabla_{\theta} \log \pi_{\theta}(a | s_t) Q^{\pi}(s_t, a) \right] \\
&= \sum_{\tau} \pi_{\theta}(\tau) \left[\sum_{t=0}^{\infty} \gamma^t \nabla_{\theta} \log \pi_{\theta}(a_t | s_t) Q^{\pi}(s_t, a_t) \right] \\
&= \nabla_{\theta} J(\theta)
\end{aligned}$$

Soit π définie selon une fonction softmax : $\pi_{\theta}(a|s) = \frac{e^{h_{\theta}(s,a)}}{\sum_{a' \in \mathcal{A}(s)} e^{h_{\theta}(s,a')}}$, avec $h_{\theta} : S \times A \rightarrow \mathbb{R}$

Montrons que pour respecter la lière condition, on peut prendre :

$$f_{\phi}(s, a) = \left[\nabla_{\theta} h_{\theta}(s, a) - \sum_{a' \in \mathcal{A}(s)} \pi_{\theta}(a' | s) \nabla_{\theta} h_{\theta}(s, a') \right]^T \phi$$

Soit $f_{\phi}(s, a) = \left[\nabla_{\theta} h_{\theta}(s, a) - \sum_{a' \in \mathcal{A}(s)} \pi_{\theta}(a' | s) \nabla_{\theta} h_{\theta}(s, a') \right]^T \phi$

On a alors :

$$\begin{aligned}
\nabla_{\phi} f_{\phi}(s, a) &= \nabla_{\theta} h_{\theta}(s, a) - \sum_{a' \in \mathcal{A}(s)} \pi_{\theta}(a' | s) \nabla_{\theta} h_{\theta}(s, a') \\
&= \nabla_{\theta} h_{\theta}(s, a) - \sum_{a' \in \mathcal{A}(s)} \frac{e^{h_{\theta}(s, a')}}{\sum_{b \in \mathcal{A}(s)} e^{h_{\theta}(s, b)}} \nabla_{\theta} h_{\theta}(s, a') \\
&= \nabla_{\theta} h_{\theta}(s, a) - \sum_{a' \in \mathcal{A}(s)} \frac{\nabla_{\theta} e^{h_{\theta}(s, a')}}{\sum_{b \in \mathcal{A}(s)} e^{h_{\theta}(s, b)}} \\
&= \nabla_{\theta} h_{\theta}(s, a) - \frac{\sum_{a' \in \mathcal{A}(s)} \nabla_{\theta} e^{h_{\theta}(s, a')}}{\sum_{a' \in \mathcal{A}(s)} e^{h_{\theta}(s, a')}} \\
&= \nabla_{\theta} h_{\theta}(s, a) - \frac{\nabla_{\theta} \sum_{a' \in \mathcal{A}(s)} e^{h_{\theta}(s, a')}}{\sum_{a' \in \mathcal{A}(s)} e^{h_{\theta}(s, a')}} \\
&= \nabla_{\theta} h_{\theta}(s, a) - \nabla_{\theta} \log \sum_{a' \in \mathcal{A}(s)} e^{h_{\theta}(s, a')} \\
&= \nabla_{\theta} [\log e^{h_{\theta}(s, a)} - \log \sum_{a' \in \mathcal{A}(s)} e^{h_{\theta}(s, a')}] \\
&= \nabla_{\theta} \left[\log \frac{e^{h_{\theta}(s, a)}}{\sum_{a' \in \mathcal{A}(s)} e^{h_{\theta}(s, a')}} \right] \\
&= \nabla_{\theta} \log \pi_{\theta}(s, a) \\
&= \frac{\nabla_{\theta} \pi_{\theta}(s, a)}{\pi_{\theta}(s, a)}
\end{aligned}$$

Une façon d'atteindre la seconde condition est de considérer le problème de minimisation suivant :

$$\min_{\phi, w} \mathbb{E}_{(s,a) \sim \pi_\theta} [(Q^\pi(s, a) - f_\phi(s, a) - v_w(s))^2]$$

On veut montrer que pour tout s , la fonction v_w qui minimise la variance de cette estimation de $Q^\pi(s, a)$ par $f_\phi(s, a) + v_w(s)$ pour les différentes actions a est égale à $V^\pi(s)$.

Soit pour tout s , la variance de l'estimateur $f_\phi(s, \cdot) + v_w(s)$ donnée par :

$$\epsilon_s = \sum_{a \in \mathcal{A}(s)} \pi_\theta(a|s) [Q^\pi(s, a) - f_\phi(s, a) - v_w(s)]^2$$

On a :

$$\frac{\partial \epsilon_s}{\partial v_w(s)} = -2 \sum_{a \in \mathcal{A}(s)} \pi_\theta(a|s) [Q^\pi(s, a) - f_\phi(s, a) - v_w(s)]$$

Et :

$$\frac{\partial^2 \epsilon_s}{\partial^2 v_w(s)} = 2$$

Puisque cette dérivée seconde est toujours positive on peut affirmer que ϵ_s est convexe en $v_w(s)$ et donc l'annulation de la dérivée $\frac{\partial \epsilon_s}{\partial v_w(s)}$ permet de minimiser ϵ_s .

On considère donc :

$$-2 \sum_{a \in \mathcal{A}(s)} \pi_\theta(a|s) [Q^\pi(s, a) - f_\phi(s, a) - v_w(s)] = 0$$

$$\sum_{a \in \mathcal{A}(s)} \pi_\theta(a|s) Q^\pi(s, a) - \sum_{a \in \mathcal{A}(s)} \pi_\theta(a|s) f_\phi(s, a) = v_w(s)$$

Or :

$$\begin{aligned} \sum_{a \in \mathcal{A}(s)} \pi_\theta(a|s) f_\phi(s, a) &= \sum_{a \in \mathcal{A}(s)} \pi_\theta(a|s) \left[\nabla_\theta h_\theta(s, a) - \sum_{a' \in \mathcal{A}(s)} \pi_\theta(a'|s) \nabla_\theta h_\theta(s, a') \right]^T \phi \\ &= \left[\sum_{a \in \mathcal{A}(s)} \pi_\theta(a|s) \nabla_\theta h_\theta(s, a) - \sum_{a \in \mathcal{A}(s)} \pi_\theta(a|s) \sum_{a' \in \mathcal{A}(s)} \pi_\theta(a'|s) \nabla_\theta h_\theta(s, a') \right]^T \phi \\ &= \left[\sum_{a \in \mathcal{A}(s)} \pi_\theta(a|s) \nabla_\theta h_\theta(s, a) - \sum_{a' \in \mathcal{A}(s)} \pi_\theta(a'|s) \nabla_\theta h_\theta(s, a') \right]^T \phi \\ &= [0]^T \phi \\ &= 0 \end{aligned}$$

On a donc :

$$\sum_{a \in \mathcal{A}(s)} \pi_\theta(a|s) Q^\pi(s, a) = v_w(s)$$

Soit : $v_w(s) = V^\pi(s)$

On peut alors voir f_w comme une fonction d'avantage : $A^\pi(s, a) = Q^\pi(s, a) - V^\pi(s)$, pouvant être obtenue par temporal difference.