Preuves Cours 6 RLD

1 Preuve ACER

On souhaite montrer que l'astuce proposée dans ACER [1] (sous le nom de "Importance Weight Truncation with Bias Correction"), afin de contenir la variance pour le cadre des Off-Policy Policy Gradient, n'introduit pas biais dans les gradients considérés. Plus précisément, avec π_θ la politique apprise et μ la politique de comportement off-policy, il s'agit de montrer que :

$$\begin{aligned}
g &= \mathbb{E}_{s,a\sim\mu}\Big[\frac{\pi_\theta(a|s)}{\mu(a|s)} Q^\pi(s,a)\nabla_\theta \ln \pi_\theta(a|s)\Big] \\
&= \mathbb{E}_{s,a\sim\mu}\Big[\min(c,\omega_t(a))\big(Q^{\text{ret}}(s,a) - V_w(s)\big)\nabla_\theta \ln \pi_\theta(a|s) \\
&+ \mathbb{E}_{a'\sim\pi}\big[\max(0, \frac{\omega_t(a')-c}{\omega_t(a')})\big(Q_w(s,a') - V_w(s)\big)\nabla_\theta \ln \pi_\theta(a'|s)\big]\Big] \qquad ; \text{Avec } \omega_t(a) = \tfrac{\pi(a|S_t)}{\mu(a|S_t)}
\end{aligned}$$

Soit $\omega_t(a) = \frac{\pi(a|S_t)}{\mu(a|S_t)}$
Pour tout c, on a :

$$\begin{aligned}
g &= \mathbb{E}_{s,a\sim\mu}\Big[w_t(a)Q^\pi(s,a)\nabla_\theta \ln \pi_\theta(a|s)\Big] \\
&= \mathbb{E}_{s,a\sim\mu}\Big[(min(c,w_t(a)) + max(0, w_t(a)-c))Q^\pi(s,a)\nabla_\theta \ln \pi_\theta(a|s)\Big] \\
&= \mathbb{E}_{s,a\sim\mu}\Big[min(c,w_t(a))Q^\pi(s,a)\nabla_\theta \ln \pi_\theta(a|s) \\
&+ \mathbb{E}_{a'\sim\mu}\big[max(0, w_t(a')-c)Q^\pi(s,a')\nabla_\theta \ln \pi_\theta(a'|s)\big]\Big] \\
&= \mathbb{E}_{s,a\sim\mu}\Big[min(c,w_t(a))Q^\pi(s,a)\nabla_\theta \ln \pi_\theta(a|s) \\
&+ \mathbb{E}_{a'\sim\pi}\big[\frac{1}{w_t(a')}max(0, w_t(a')-c)Q^\pi(s,a')\nabla_\theta \ln \pi_\theta(a'|s)\big]\Big] \\
&= \mathbb{E}_{s,a\sim\mu}\Big[min(c,w_t(a))Q^\pi(s,a)\nabla_\theta \ln \pi_\theta(a|s) \\
&+ \mathbb{E}_{a'\sim\pi}\big[max(0, \frac{w_t(a')-c}{w_t(a')})Q^\pi(s,a')\nabla_\theta \ln \pi_\theta(a'|s)\big]\Big]
\end{aligned}$$

Pour la suite il suffit de noter que $V_w(s)$ ne dépend pas des actions samplées et donc ne biaise pas le gradient (baseline), et que Q^{ret} et Q_w sont centrés sur Q^π.

2 Preuves Off-Policy Policy Gradient

On montre que $J(\pi) = \mathbb{E}_{\tau\sim\pi_0}[\sum_{t=0}^T w_{0:T}\gamma^t r_t]$ avec $w_{0:T} = \prod_{i=0}^T \frac{\pi(a_t|s_t)}{\pi_0(a_t|s_t)}$, avec π_0 la politique d'échantillonnage, peut être estimée de manière équivalente par (Step-Wise Importance Sampling) :

$$R_\pi = \mathbb{E}_{\tau\sim\pi_0}[\sum_{t=0}^T w_{0:t}\gamma^t r_t]$$

Ce qui permet de réduire la variance de l'estimateur en remplaçant des poids d'Importance Sampling calculés sur des trajectoires entières par des poids ne considérant que le passé à chaque instant uniquement.

$$
\begin{aligned}
J(\pi) = \mathbb{E}_{\tau\sim\pi_0}[\sum_{t=0}^{T} w_{0:T}\gamma^t r_t] \\
&= \int \pi_0(\tau)\sum_{t=0}^{T} w_{0:T}\gamma^t r_t \; d\tau \\
&= \sum_{t=0}^{\infty}\int_\tau \pi_0(\tau) w_{0:\infty}\gamma^t r_t \; d\tau \\
&= \sum_{t=0}^{\infty}\int \pi_0(\tau_{0:t}) w_{0:t}\gamma^t r_t \int \pi_0(\tau_{t+1:\infty}|\tau_{0:t}) w_{t+1:\infty} \; d\tau_{t+1:\infty} d\tau_{0:t} \\
&= \sum_{t=0}^{\infty}\int \pi_0(\tau_{0:t}) w_{0:t}\gamma^t r_t \int \pi(\tau_{t+1:\infty}|\tau_{0:t}) \; d\tau_{t+1:\infty} d\tau_{0:t} \\
&= \sum_{t=0}^{\infty}\int \pi_0(\tau_{0:t}) w_{0:t}\gamma^t r_t \int \pi_0(\tau_{t+1:\infty}|\tau_{0:t}) \; d\tau_{t+1:\infty} d\tau_{0:t} \\
&\qquad (\text{car} \int \pi(\tau_{t+1:\infty}|\tau_{0:t}) \; d\tau_{t+1:\infty} = 1 = \int \pi_0(\tau_{t+1:\infty}|\tau_{0:t}) \; d\tau_{t+1:\infty}) \\
&= \sum_{t=0}^{\infty}\int \pi_0(\tau) w_{0:t}\gamma^t r_t d\tau \\
&= \int \pi_0(\tau)\sum_{t=0}^{T} w_{0:t}\gamma^t r_t \; d\tau \\
&= \mathbb{E}_{\tau\sim\pi_0}[\sum_{t=0}^{T} w_{0:t}\gamma^t r_t]
\end{aligned}
$$

Mais on a toujours des produits de ratios pouvant faire exploser la variance de l'estimateur. Pour aller plus loin, on souhaite montrer que, pour tout s', on a :

$$d^\pi(s') = \gamma \sum_s d^\pi(s) \sum_a \pi(a|s) P(s'|s,a) + (1-\gamma) P(s_0 = s')$$

avec $d^\pi(s) = (1-\gamma)\sum_{t=0}^{\infty}\gamma^t d_t^\pi(s_t = s)$ la distribution sur les états selon π.

On note $d_0(s) = P(s_0 = s)$ et $d_{\pi,t}(s) = d_t^\pi(s_t = s)$.

$$
\begin{aligned}
d^\pi(s) &= (1-\gamma)\sum_{t=0}^{\infty}\gamma^t d_{\pi,t}(s) \\
&= (1-\gamma)d_0(s) + (1-\gamma)\sum_{t=1}^{\infty}\gamma^t d_{\pi,t}(s) \\
&= (1-\gamma)d_0(s) + (1-\gamma)\gamma\sum_{t=0}^{\infty}\gamma^t d_{\pi,t+1}(s) \\
&= (1-\gamma)d_0(s) + (1-\gamma)\gamma\sum_{t=0}^{\infty}\gamma^t \sum_{s'} d_{\pi,t}(s')\sum_a \pi(a|s')P(s|s',a) \\
&= (1-\gamma)d_0(s) + \gamma\sum_{s'}\sum_a \pi(a|s')P(s|s',a)(1-\gamma)\sum_{t=0}^{\infty}\gamma^t d_{\pi,t}(s') \\
&= (1-\gamma)d_0(s) + \gamma\sum_{s'}\sum_a \pi(a|s')P(s|s',a)d^\pi(s') \\
&= (1-\gamma)d_0(s) + \gamma\sum_{s'} d^\pi(s')\sum_a \pi(a|s')P(s|s',a)
\end{aligned}
$$

Pour isoler le cas s_0, après avoir ajouté une transition fictive (s_{-1}, a_{-1}, s_0) de récompense 0 au début de chaque trajectoire (avec $s_{-1} \notin \mathcal{S}$ et $P(s_0 = s|s_{-1}, a_{-1}) = P(s_0 = s)$ pour tout $s \in \mathcal{S}$), on peut écrire de manière équivalente, pour tout $s' \neq s_{-1}$:

$$\tilde{d}^\pi(s') \propto \sum_s d^\pi(s)\sum_a \pi(a|s)P(s'|s,a) \text{ avec } \tilde{d}^\pi(s) = (1-\gamma)\sum_{t=-1}^{\infty}\gamma^{t+1}d_t^\pi(s_t = s)$$

On a $(1-\gamma)d_{-1}(s) = 0$ pour tout $s \in \mathcal{S}$. Donc pour tout $s \in \mathcal{S}$:

$$
\begin{aligned}
\tilde{d}^\pi(s) &= (1-\gamma)\sum_{t=-1}^{\infty}\gamma^{t+1}d_{\pi,t}(s) \\
&= (1-\gamma)d_{-1}(s) + (1-\gamma)\sum_{t=0}^{\infty}\gamma^{t+1}d_{\pi,t}(s) \\
&= (1-\gamma)\gamma\sum_{t=0}^{\infty}\gamma^t d_{\pi,t}(s) \\
&= (1-\gamma)\gamma\sum_{t=-1}^{\infty}\gamma^{t+1}d_{\pi,t+1}(s) \\
&= (1-\gamma)\gamma\sum_{t=-1}^{\infty}\gamma^{t+1}\sum_{s'} d_{\pi,t}(s')\sum_a \pi(a|s')P(s|s',a) \\
&= \gamma\sum_{s'}\sum_a \pi(a|s')P(s|s',a)(1-\gamma)\sum_{t=-1}^{\infty}\gamma^{t+1}d_{\pi,t}(s') \\
&= \gamma\sum_{s'}\sum_a \pi(a|s')P(s|s',a)\tilde{d}^\pi(s') \\
&= \gamma\sum_{s'}\tilde{d}^\pi(s')\sum_a \pi(a|s')P(s|s',a)
\end{aligned}
$$

Soit pour tout s' la contrainte :

$$\mathbb{E}_{(s,a)\sim\tilde{d}^{\pi_0}((s,a)|s')}[\Delta(w_\theta,s,a,s')|s']=0,\ \text{avec } \Delta(w_\theta,s,a,s')=w_\theta(s)\frac{\pi(a|s)}{\pi_0(a|s)}-w_\theta(s')$$

avec $\tilde{d}^{\pi_0}((s,a)|s')\propto\tilde{d}^{\pi_0}(s)\pi_0(a|s)P(s'|s,a)$.

On montre que lorsque cette contrainte est vérifiée, on a $w_\theta(s)=\frac{\tilde{d}^{\pi}(s)}{\tilde{d}^{\pi_0}(s)}$ et $\tilde{d}^{\pi}(s')\propto\sum_s d^{\pi}(s)\sum_a\pi(a|s)P(s'|s,a)$.

Soit $\tilde{d}^{\pi_0}((s,a)|s')\propto\tilde{d}^{\pi_0}(s)\pi_0(a|s)P(s'|s,a)$
Si pour tout s', on vérifie la contrainte :

$$\mathbb{E}_{(s,a)\sim\tilde{d}^{\pi_0}((s,a)|s')}[\Delta(w_\theta,s,a,s')|s']=0,\ \text{avec } \Delta(w_\theta,s,a,s')=w_\theta(s)\frac{\pi(a|s)}{\pi_0(a|s)}-w_\theta(s')$$

Alors par définition :

$$\sum_{(s,a)}\tilde{d}^{\pi_0}((s,a)|s')[w_\theta(s)\frac{\pi(a|s)}{\pi_0(a|s)}-w_\theta(s')]=0$$

ou encore :

$$w_\theta(s')=\sum_{(s,a)}\tilde{d}^{\pi_0}((s,a)|s')[w_\theta(s)\frac{\pi(a|s)}{\pi_0(a|s)}]$$

Or :

$$\tilde{d}^{\pi_0}((s,a)|s')=\frac{\tilde{d}^{\pi_0}(s)\pi_0(a|s)P(s'|s,a)}{\tilde{d}^{\pi_0}(s')}$$

On a donc :

$$\tilde{d}^{\pi_0}(s')w_\theta(s')=\sum_{(s,a)}\tilde{d}^{\pi_0}(s)\pi_0(a|s)P(s'|s,a)[w_\theta(s)\frac{\pi(a|s)}{\pi_0(a|s)}]$$

Ou :

$$\tilde{d}^{\pi_0}(s')w_\theta(s')=\sum_{(s,a)}\tilde{d}^{\pi_0}(s)P(s'|s,a)w_\theta(s)\pi(a|s)$$

La seule manière de satisfaire cette égalité pour un w_θ non dégénéré (i.e. $\exists s\in S$ tel que $\tilde{d}^{\pi_0}(s)>0\wedge w_\theta(s)>0$) est de prendre $w_\theta(s)=\frac{\tilde{d}^{\pi}(s)}{\tilde{d}^{\pi_0}(s)}$, ce qui donne la relation :

$$\tilde{d}^{\pi}(s')=\sum_s d^{\pi}(s)\sum_a\pi(a|s)P(s'|s,a)$$

Puisqu'on n'a pas la capacité de considérer les contraintes sur tous les états de manière efficace, [2] propose de les regrouper sous la contrainte :

$$\sum_{s'}d^{\pi_0}(s')\mathbb{E}_{(s,a)\sim\tilde{d}^{\pi_0}((s,a)|s')}[\Delta(w_\theta,s,a,s')|s']=0$$

Ce qui forme une condition nécessaire mais pas suffisante à l'obtention d'un w_θ efficace. Par contre, si on considère une valeur $f(s')$ associée à chaque contrainte, on peut se focaliser sur les contraintes violées. On peut alors vouloir garantir :

$$\max_f \sum_{s'} d^{\pi_0}(s')\mathbb{E}_{(s,a)\sim\tilde{d}^{\pi_0}((s,a)|s')}[\Delta(w_\theta, s, a, s')|s'] = 0$$

Ce qui correspond alors alors une condition suffisante pour garantir $w_\theta(s) = \frac{\tilde{d}^{\pi}(s)}{\tilde{d}^{\pi_0}(s)}$.
Cependant on ne possède généralement pas suffisamment de données pour bien estimer $\tilde{d}^{\pi_0}((s,a)|s')$ (et la maximisation de f est délicate dans le cas général). [2] limite alors la capacité de f à une certaine classe de fonctions : les fonctions de la boule unitaire d'un RKHS, ce qui revient à faire l'hypothèse que des états proches ont des ratios de distributions proches.

3 Preuves DPG

Pour une politique déterministe $\pi_\theta(s) = Dirac(\mu_\theta(s))$, on souhaite montrer que le gradient de $J(\theta) = \int P(s_0) \int \pi(a_0|s_0)Q(s_0, a_0)da_0ds_0$ s'écrit :

$$\nabla_\theta J(\theta) = \int_{\mathcal{S}} d^\pi(s)\nabla_a Q(s,a)_{|a=\mu_\theta(s)}\nabla_\theta\mu_\theta(s)ds$$

avec $d^\pi(s)$ la distribution discountée des futurs états :

$$d^\pi(s) = (1-\gamma)\sum_{t=0}^{\infty}\gamma^t P(s_t = s|\pi)$$

où $P(s_t = s|\pi)$ dénote la probabilité d'être dans l'état s à l'étape t d'une trajectoire en suivant la politique π.

Tout d'abord on note que, si on est dans le cas d'une politique déterministe $\pi_\theta(s) = Dirac(\mu_\theta(s))$, on a :

$$\begin{aligned} J(\theta) &= \int P(s_0) \int \pi(a_0|s_0)Q(s_0, a_0)da_0ds_0 \\ &= \int P(s_0)Q(s_0, \mu_\theta(s_0))ds_0 \end{aligned}$$

On a pour tout t :

$$Q(s_t, \mu_\theta(s_t)) = \int P(s_{t+1}|s_t, a_t = \mu_\theta(s_t))\Big(r_t + \gamma Q(s_{t+1}, \mu_\theta(s_{t+1}))\Big)ds_{t+1}$$

Pour tout s, on a alors :

$$\begin{aligned}
\nabla_\theta Q(s_t,\mu_\theta(s_t)) &= \int \nabla_\theta P(s_{t+1}|s_t,a_t=\mu_\theta(s_t))\Big(r_t+\gamma Q(s_{t+1},\mu_\theta(s_{t+1})\Big)ds_{t+1}\\
&+\int P(s_{t+1}|s_t,a_t=\mu_\theta(s_t))\gamma\nabla_\theta Q(s_{t+1},\mu_\theta(s_{t+1}))ds_{t+1}\\
&=\int \nabla_a P(s_{t+1}|s_t,a_t=a)_{|a=\mu_\theta(s_t)}\nabla_\theta\mu_\theta(s_t)\Big(r_t+\gamma Q(s_{t+1},\mu_\theta(s_{t+1})\Big)ds_{t+1}\\
&+\int P(s_{t+1}|s_t,a_t=\mu_\theta(s_t))\gamma\nabla_\theta Q(s_{t+1},\mu_\theta(s_{t+1}))ds_{t+1}\\
&=\nabla_\theta\mu_\theta(s_t)\nabla_a\left[\int P(s_{t+1}|s_t,a_t=a)\Big(r_t+\gamma Q(s_{t+1},\mu_\theta(s_{t+1})\Big)ds_{t+1}\right]_{|a=\mu_\theta(s_t)}\\
&+\int P(s_{t+1}|s_t,a_t=\mu_\theta(s_t))\gamma\nabla_\theta Q(s_{t+1},\mu_\theta(s_{t+1}))ds_{t+1}\\
&=\nabla_\theta\mu_\theta(s_t)\nabla_a Q(s_t,a)_{|a=\mu_\theta(s_t)}+\int P(s_{t+1}|s_t,a_t=\mu_\theta(s_t))\gamma\nabla_\theta Q(s_{t+1},\mu_\theta(s_{t+1}))ds_{t+1}
\end{aligned}$$

Selon la chain-rule, on a alors :

$$\begin{aligned}
\nabla_\theta J(\theta) &= \nabla_\theta\int P(s_0)Q(s_0,\mu_\theta(s_0))ds_0\\
&=\int P(s_0)\nabla_\theta Q(s_0,\mu_\theta(s_0))ds_0\\
&=\int_{\mathcal{S}} d^\pi(s)\nabla_a Q(s,a)_{|a=\mu_\theta(s)}\nabla_\theta\mu_\theta(s)ds
\end{aligned}$$

avec $d^\pi(s)=(1-\gamma)\sum_{t=0}^{\infty}\gamma^t P(s_t=s|\pi)$

4 Preuves Q-Prop

Selon une politique π_θ, on souhaite montrer que le gradient de $J(\theta)=\mathbb{E}_{d^\pi,\pi}\left[\hat{A}(s_t,a_t)\right]$, avec $\hat{A}(s,a)$ la fonction d'avantage de l'action a en l'état s, peut s'écrire de la manière non biaisée suivante :

$$\begin{aligned}
\nabla_\theta J(\theta) = \mathbb{E}_{d^\pi,\pi}&\left[\nabla_\theta\log\pi_\theta(a_t|s_t)\left(\hat{A}(s_t,a_t)-\nabla_a Q_w(s_t,a)|_{a=\mu_\theta(s_t)}(a_t-\mu_\theta(s_t))\right)\right]\\
&+\mathbb{E}_{d^\pi}\left[\nabla_a Q_w(s_t,a)|_{a=\mu_\theta(s_t)}\nabla_\theta\mu_\theta(s_t)\right]\ \text{avec}\ \mu_\theta(s_t)=\mathbb{E}_{a\sim\pi(a|s_t)}[a]
\end{aligned}$$

Pour toute estimée de la fonction d'avantage $\bar{A}(s_t,a_t)$, on a :

$$\begin{aligned}
\nabla_\theta J(\theta) &= \nabla_\theta\mathbb{E}_{d^\pi,\pi}\left[\hat{A}(s_t,a_t)\right]\\
&=\mathbb{E}_{d^\pi,\pi}\left[\nabla_\theta\log\pi_\theta(a_t|s_t)\hat{A}(s_t,a_t)\right]\\
&=\mathbb{E}_{d^\pi,\pi}\left[\nabla_\theta\log\pi_\theta(a_t|s_t)\Big(\hat{A}(s_t,a_t)-\bar{A}(s_t,a_t)\Big)\right]+\mathbb{E}_{d^\pi,\pi}\left[\nabla_\theta\log\pi_\theta(a_t|s_t)\bar{A}(s_t,a_t)\right]
\end{aligned}$$

Considérons pour tout s_t,a_t, l'estimateur : $\bar{A}(s_t,a_t)=\bar{Q}_w(s_t,a_t)-\mathbb{E}_{a\sim\pi(a|s_t)}[\bar{Q}_w(s_t,a)]$, avec $\bar{Q}_w(s_t,a_t)$ l'approximation de $Q_w(s_t,a_t)$ selon une expansion de Taylor d'ordre 1 au point $\mu_\theta(s_t)=$

$\mathbb{E}_{a\sim\pi(a|s_t)}[a]$. On a :

$$\begin{aligned}\bar{A}(s_t,a_t) &= \bar{Q}_w(s_t,a_t) - \mathbb{E}_{a\sim\pi(a|s_t)}[\bar{Q}_w(s_t,a)] \\ &= Q_w(s_t,\mu_\theta(s_t)) + \nabla_a Q_w(s_t,a)|_{a=\mu_\theta(s_t)}(a_t - \mu_\theta(s_t)) \\ &- \mathbb{E}_{a'\sim\pi(a'|s_t)}\left[Q_w(s_t,\mu_\theta(s_t)) + \nabla_a Q_w(s_t,a)|_{a=\mu_\theta(s_t)}(a' - \mu_\theta(s_t))\right] \\ &= \nabla_a Q_w(s_t,a)|_{a=\mu_\theta(s_t)}(a_t - \mu_\theta(s_t)) \\ &- \nabla_a Q_w(s_t,a)|_{a=\mu_\theta(s_t)}\mathbb{E}_{a'\sim\pi(a'|s_t)}\left[a' - \mu_\theta(s_t)\right] \\ &= \nabla_a Q_w(s_t,a)|_{a=\mu_\theta(s_t)}(a_t - \mu_\theta(s_t)) \\ &- \nabla_a Q_w(s_t,a)|_{a=\mu_\theta(s_t)}(\mu_\theta(s_t) - \mu_\theta(s_t)) \\ &= \nabla_a Q_w(s_t,a)|_{a=\mu_\theta(s_t)}(a_t - \mu_\theta(s_t))\end{aligned}$$

Or, dans ce cas on a :

$$\begin{aligned}\mathbb{E}_{d^\pi,\pi}\left[\nabla_\theta \log \pi_\theta(a_t|s_t)\bar{A}(s_t,a_t)\right] &= \mathbb{E}_{d^\pi,\pi}\left[\nabla_\theta \log \pi_\theta(a_t|s_t)\nabla_a Q_w(s_t,a)|_{a=\mu_\theta(s_t)}(a_t - \mu_\theta(s_t))\right] \\ &= \mathbb{E}_{d^\pi,\pi}\left[\nabla_\theta \log \pi_\theta(a_t|s_t)\nabla_a Q_w(s_t,a)|_{a=\mu_\theta(s_t)}a_t\right] \\ &- \mathbb{E}_{d^\pi,\pi}\left[\nabla_\theta \log \pi_\theta(a_t|s_t)\nabla_a Q_w(s_t,a)|_{a=\mu_\theta(s_t)}\mu_\theta(s_t)\right] \\ &= \mathbb{E}_{d^\pi}\left[\nabla_a Q_w(s_t,a)|_{a=\mu_\theta(s_t)}\mathbb{E}_\pi[\nabla_\theta \log \pi_\theta(a_t|s_t)a_t]\right] \\ &- \mathbb{E}_{d^\pi}\left[\nabla_a Q_w(s_t,a)|_{a=\mu_\theta(s_t)}\mu_\theta(s_t)\mathbb{E}_\pi[\nabla_\theta \log \pi_\theta(a_t|s_t)]\right] \\ &= \mathbb{E}_{d^\pi}\left[\nabla_a Q_w(s_t,a)|_{a=\mu_\theta(s_t)}\nabla_\theta\mathbb{E}_\pi[a_t]\right] \\ &- \mathbb{E}_{d^\pi}\left[\nabla_a Q_w(s_t,a)|_{a=\mu_\theta(s_t)}\mu_\theta(s_t)\mu_\theta(s_t)\nabla_\theta 1\right] \\ &= \mathbb{E}_{d^\pi}\left[\nabla_a Q_w(s_t,a)|_{a=\mu_\theta(s_t)}\nabla_\theta\mu_\theta(s_t)\right]\end{aligned}$$

On a donc bien :

$$\begin{aligned}\nabla_\theta J(\theta) = \mathbb{E}_{d^\pi,\pi}&\left[\nabla_\theta \log \pi_\theta(a_t|s_t)\left(\hat{A}(s_t,a_t) - \nabla_a Q_w(s_t,a)|_{a=\mu_\theta(s_t)}(a_t - \mu_\theta(s_t))\right)\right] \\ &+ \mathbb{E}_{d^\pi}\left[\nabla_a Q_w(s_t,a)|_{a=\mu_\theta(s_t)}\nabla_\theta\mu_\theta(s_t)\right] \text{ avec } \mu_\theta(s_t) = \mathbb{E}_{a\sim\pi(a|s_t)}[a]\end{aligned}$$

Références

[1] Z. Wang, V. Bapst, N. Heess, V. Mnih, R. Munos, K. Kavukcuoglu, and N. de Freitas, "Sample efficient actor-critic with experience replay," *arXiv preprint arXiv :1611.01224*, 2016.

[2] Q. Liu, L. Li, Z. Tang, and D. Zhou, "Breaking the curse of horizon : Infinite-horizon off-policy estimation," in *Advances in Neural Information Processing Systems*, 2018, pp. 5356–5366.