Preuves Cours 5 RLD

1 Preuve variance Importance Sampling

Soit l'estimateur IS de $f(x)$ selon P, en échantillonnant de Q :

$$\bar{f} = \frac{1}{N} \sum_{x \sim Q(x)} w(x) f(x) \approx \mathbb{E}_{x\sim Q}\Big[w(x)f(x)\Big] = \mathbb{E}_{x\sim P}\Big[f(x)\Big]$$

avec N le nombre d'échantillons issus de la distribution Q et $w(x) = \frac{P(x)}{Q(x)}$ le poids d'Importance Sampling associé.

On montre que la variance de cet estimateur selon P est donnée par :

$$Var[\bar{f}] = \frac{1}{N}\Big(\mathbb{E}_{x\sim Q}\Big[\big(\frac{P(x)}{Q(x)}f(x)\big)^2\Big] - \mathbb{E}_{x\sim P}\Big[f(x)\Big]^2\Big)$$

Tout simplement, si les échantillons de X issus de Q sont i.i.d., on a :

$$Var[\bar{f}] = \frac{1}{N} Var[w(X)f(X)]$$

Or :

$$Var[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2$$

On a donc :

$$Var[w(X)f(X)] = \mathbb{E}_{x\sim Q}\Big[\big(w(x)f(x)\big)^2\Big] - \mathbb{E}_{x\sim Q}\Big[w(X)f(x)\Big]^2\Big)$$

Sachant que $w(X) = \frac{P(x)}{Q(x)}$, on obtient :

$$Var[\bar{f}] = \frac{1}{N}\Big(\mathbb{E}_{x\sim Q}\Big[\big(\frac{P(x)}{Q(x)}f(x)\big)^2\Big] - \mathbb{E}_{x\sim P}\Big[f(x)\Big]^2\Big)$$

2 Preuve Policy Performance Bound

On souhaite tout d'abord montrer que la performance relative d'une politique π' par rapport à une autre π s'écrit :

$$J(\pi') = J(\pi) + \mathbb{E}_{\tau\sim\pi'}[\sum_{t=0}^{\infty} \gamma^t A^\pi(s_t, a_t)]$$

avec A^π la fonction d'avantage selon π :

$$A^\pi(s_t, a_t) = Q^\pi(s_t, a_t) - V^\pi(s_t)$$

$$\begin{aligned}
\mathbb{E}_{\tau\sim\pi'}[\sum_{t=0}^{\infty} \gamma^t A^\pi(s_t, a_t)] &= \mathbb{E}_{\tau\sim\pi'}[\sum_{t=0}^{\infty} \gamma^t (r_t + \gamma V^\pi(s_{t+1}) - V^\pi(s_t)] \\
&= \mathbb{E}_{\tau\sim\pi'}[-V^\pi(s_0) + \sum_{t=0}^{\infty} \gamma^t r_t]
\end{aligned}$$

car :

$$\sum_{t=0}^{\infty}\gamma^t(\gamma V^\pi(s_{t+1}) - V^\pi(s_t) = \gamma V^\pi(s_1) - V^\pi(s_0) + \gamma^2 V^\pi(s_2) - \gamma V^\pi(s_1) + \gamma^3 V^\pi(s_3) - \gamma^2 V^\pi(s_2)$$
$$+ \gamma^4 V^\pi(s_4) - \gamma^3 V^\pi(s_3) + ... + \gamma^{|T|} V^\pi(s_{|T|}) - \gamma^{|T|-1} V^\pi(s_{|T|-1})$$
$$= -V^\pi(s_0)$$

sachant que $V^\pi(s_{|T|}) = 0$.

On a donc :

$$\begin{aligned}\mathbb{E}_{\tau\sim\pi'}[\sum_{t=0}^{\infty}\gamma^t A^\pi(s_t,a_t)] &= -\mathbb{E}_{s_0}[V^\pi(s_0)] + \mathbb{E}_{\tau\sim\pi'}[\sum_{t=0}^{\infty}\gamma^t r_t] \\ &= -J(\pi) + J(\pi')\end{aligned}$$

Soit la distribution discountée des futurs états :

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

On montre qu'on peut alors ré-écrire $J(\pi') - J(\pi)$ sous la forme d'une espérance sur les états :

$$J(\pi') - J(\pi) = \frac{1}{1-\gamma}\mathbb{E}_{s\sim d^{\pi'}(s), a\sim\pi(a|s)}[\frac{\pi'(a|s)}{\pi(a|s)}A^\pi(s,a)]$$

$$\begin{aligned}\mathbb{E}_{\tau\sim\pi'}[\sum_{t=0}^{\infty}\gamma^t A^\pi(s_t,a_t)] &= \sum_{t=0}^{\infty}\int_s P(s_t = s|\pi')\sum_a \pi'(a|s)\gamma^t A^\pi(s,a) \\ &= \int_s \sum_{t=0}^{\infty}\gamma^t P(s_t = s|\pi')\sum_a \pi'(a|s)A^\pi(s,a) \\ &= \frac{1}{1-\gamma}\mathbb{E}_{s\sim d^{\pi'}(s)}\Big[\sum_a \pi'(a|s)A^\pi(s,a)\Big]\end{aligned}$$

car $d^\pi(s) = (1-\gamma)\sum_{t=0}^{\infty}\gamma^t P(s_t = s|\pi)$, avec $(1-\gamma)$ le facteur de normalisation de cette distribution stationnaire (car $\sum_{t=0}^{\infty}\gamma^t = \frac{1}{1-\gamma}$).
On a donc :

$$\begin{aligned}J(\pi') - J(\pi) &= \mathbb{E}_{\tau\sim\pi'}[\sum_{t=0}^{\infty}\gamma^t A^\pi(s_t,a_t)] \\ &= \frac{1}{1-\gamma}\mathbb{E}_{s\sim d^{\pi'}(s)}\Big[\sum_a \pi'(a|s)A^\pi(s,a)\Big] \\ &= \frac{1}{1-\gamma}\mathbb{E}_{s\sim d^{\pi'}(s), a\sim\pi'(a|s)}\Big[A^\pi(s,a)\Big] \\ &= \frac{1}{1-\gamma}\mathbb{E}_{s\sim d^{\pi'}(s), a\sim\pi(a|s)}[\frac{\pi'(a|s)}{\pi(a|s)}A^\pi(s,a)]\end{aligned}$$

On souhaite montrer :

$$J(\pi') - J(\pi) \geq L^\pi(\pi') - C\ D_{KL}^{max}(\pi||\pi')$$

avec $C = 4\epsilon\gamma/(1-\gamma)^2$, $\epsilon = \max_{s,a}|A^\pi(s,a)|$, $L^\pi(\pi') = \frac{1}{1-\gamma}\mathbb{E}_{s\sim d^\pi(s), a\sim\pi(a|s)}[\frac{\pi'(a|s)}{\pi(a|s)}A^\pi(s,a)]$ et $D_{KL}^{max}(\pi||\pi')$ la divergence de Kullback-Leibler maximale entre π' et π prise sur l'ensemble des états.

Pour cela, on commence par donner la définition suivante :

Définition : (π, π') est une paire de politiques α-couplées si elle définit une distribution jointe $(a, a')|s$ telle que $p(a \neq a'|s) \leq \alpha$ pour tout s.

Puis on montre tout d'abord le lemme suivant, avec $\bar{A}(s) = \mathbb{E}_{a\sim\pi'(a|s)}[A_\pi(s,a)]$:

Lemme 1 : Soient π et π' deux politiques α-couplées. On a alors pour tout s :

$$|\bar{A}(s)| \leq 2\alpha \max_{s',a}|A_\pi(s',a)|$$

$$\begin{aligned}
\bar{A}(s) &= \mathbb{E}_{a\sim\pi'(a|s)}[A_\pi(s,a)] \\
&= \mathbb{E}_{(a,a')\sim(\pi(a|s),\pi'(a'|s))}[A_\pi(s,a') - A_\pi(s,a)]
\end{aligned}$$

car on sait que, selon la définition de la fonction d'avantage, $\mathbb{E}_{a\sim\pi(a|s)}[A_\pi(s,a)] = 0$.

Donc, en séparant les cas on a :

$$\begin{aligned}
\bar{A}(s) = & P(a = a'|s)\mathbb{E}_{(a,a')\sim(\pi(a|s),\pi'(a'|s))|a=a'}[A_\pi(s,a') - A_\pi(s,a)] \\
& + P(a \neq a'|s)\mathbb{E}_{(a,a')\sim(\pi(a|s),\pi'(a'|s))|a\neq a'}[A_\pi(s,a') - A_\pi(s,a)] \\
& = P(a = a'|s) \times 0 \\
& + P(a \neq a'|s)\mathbb{E}_{(a,a')\sim(\pi(a|s),\pi'(a'|s))|a\neq a'}[A_\pi(s,a') - A_\pi(s,a)] \\
& \leq \alpha \times \mathbb{E}_{(a,a')\sim(\pi(a|s),\pi'(a'|s))|a\neq a'}[A_\pi(s,a') - A_\pi(s,a)] \\
& \leq \alpha \times 2\max_{s',a}|A_\pi(s',a)|
\end{aligned}$$

On montre ensuite le lemme suivant :

Lemme 2 : Soit (π, π') une paire de politiques α-couplées. Alors pour tout temps t :

$$|\mathbb{E}_{s\sim P(s_t=s|\pi')}\left[\bar{A}(s)\right] - \mathbb{E}_{s\sim P(s_t=s|\pi)}\left[\bar{A}(s)\right]| \leq 4\alpha(1-(1-\alpha)^t)\max_{s,a}|A_\pi(s,a)|$$

Soit n_t le nombre de fois où $a_i \neq a'_i$ pour tout $i < t$, avec a_i et a'_i issus de π et π' jusqu'à t.

On a pour tout π et tout t :

$$\mathbb{E}_{s\sim P(s_t=s|\pi)}\left[\bar{A}(s)\right] = P(n_t = 0)\mathbb{E}_{s\sim P(s_t=s|\pi,n_t=0)}\left[\bar{A}(s)\right] + P(n_t > 0)\mathbb{E}_{s\sim P(s_t=s|\pi,n_t>0)}\left[\bar{A}(s)\right]$$

Notons que : $\mathbb{E}_{s\sim P(s_t=s|\pi',n_t=0)}\left[\bar{A}(s)\right] = \mathbb{E}_{s\sim P(s_t=s|\pi,n_t=0)}\left[\bar{A}(s)\right]$ (car on a choisi les mêmes actions avec les deux politiques jusqu'à t).

On a alors :

$$\mathbb{E}_{s\sim P(s_t=s|\pi')}\Big[\bar{A}(s)\Big] - \mathbb{E}_{s\sim P(s_t=s|\pi)}\Big[\bar{A}(s)\Big] = \\ P(n_t>0)\Big(\mathbb{E}_{s\sim P(s_t=s|\pi',n_t>0)}\Big[\bar{A}(s)\Big] - \mathbb{E}_{s\sim P(s_t=s|\pi,n_t>0)}\Big[\bar{A}(s)\Big]\Big)$$

Par définition, pour tout a_t et a'_t issus de π et π' jusqu'à t, $P(a_t = a'_t) \geq 1-\alpha$,
donc $P(n_t = 0) \geq (1-\alpha)^t$ et donc $P(n_t > 0) \leq 1-(1-\alpha)^t$

Enfin, notons que :

$$|\mathbb{E}_{s\sim P(s_t=s|\pi',n_t>0)}\Big[\bar{A}(s)\Big] - \mathbb{E}_{s\sim P(s_t=s|\pi,n_t>0)}\Big[\bar{A}(s)\Big]| \leq |\mathbb{E}_{s\sim P(s_t=s|\pi',n_t>0)}\Big[\bar{A}(s)\Big]| + |\mathbb{E}_{s\sim P(s_t=s|\pi,n_t>0)}\Big[\bar{A}(s)\Big]| \\ \leq 4\alpha \max_{s',a} |A_\pi(s',a)|$$

selon le lemme 1 (car $|\mathbb{E}_{s\sim P}\Big[\bar{A}(s)\Big]| \leq \max_s |\bar{A}(s)| \leq \max_{s,a} |A_\pi(s,a)|$ pour toute distribution P).
On a donc :

$$|\mathbb{E}_{s\sim P(s_t=s|\pi')}\Big[\bar{A}(s)\Big] - \mathbb{E}_{s\sim P(s_t=s|\pi)}\Big[\bar{A}(s)\Big]| \leq 4\alpha(1-(1-\alpha)^t) \max_{s,a} |A_\pi(s,a)|$$

Montrons maintenant que :

$$J(\pi') - J(\pi) \geq L^\pi(\pi') - \frac{4\epsilon\gamma\alpha^2}{(1-\gamma)^2}$$

avec $\epsilon = \max_{s,a} |A^\pi(s,a)|$

On a d'une part :

$$\begin{aligned} L^\pi(\pi') &= \frac{1}{1-\gamma}\mathbb{E}_{s\sim d^\pi(s),a\sim\pi(a|s)}[\frac{\pi'(a|s)}{\pi(a|s)}A^\pi(s,a)] \\ &= \int_s \sum_{t=0}^{\infty} \gamma^t P(s_t=s|\pi) \sum_a \pi(a|s)\frac{\pi'(a|s)}{\pi(a|s)}A^\pi(s,a) \\ &= \sum_{t=0}^{\infty} \int_s P(s_t=s|\pi) \sum_a \pi(a|s)\gamma^t \frac{\pi'(a|s)}{\pi(a|s)}A^\pi(s,a) \\ &= \mathbb{E}_{\tau\sim\pi}[\sum_{t=0}^{\infty} \gamma^t \frac{\pi'(a_t|s_t)}{\pi(a_t|s_t)}A^\pi(s_t,a_t)] \end{aligned}$$

D'autre part, daprès la performance relative des politiques, on sait que (voir ci-dessus) :

$$J(\pi') - J(\pi) = \mathbb{E}_{\tau\sim\pi'}[\sum_{t=0}^{\infty} \gamma^t A^\pi(s_t,a_t)]$$

Sachant que $\bar{A}(s) = \mathbb{E}_{a\sim\pi'(a|s)}[A_\pi(s,a)]$, on a alors :

$$\begin{aligned}
|J(\pi') - J(\pi) - L^\pi(\pi')| &= |\mathbb{E}_{\tau\sim\pi'}[\sum_{t=0}^{\infty} \gamma^t A^\pi(s_t,a_t)] - \mathbb{E}_{\tau\sim\pi}[\sum_{t=0}^{\infty} \gamma^t \frac{\pi'(a_t|s_t)}{\pi(a_t|s_t)} A^\pi(s_t,a_t)]| \\
&= |\sum_{t=0}^{\infty} \gamma^t \mathbb{E}_{\tau\sim\pi'}[\bar{A}(s_t)] - \sum_{t=0}^{\infty} \gamma^t \mathbb{E}_{\tau\sim\pi}[\bar{A}(s_t)]| \\
&\leq |\sum_{t=0}^{\infty} \gamma^t \mathbb{E}_{\tau\sim\pi'}[\bar{A}(s_t)]| + |\sum_{t=0}^{\infty} \gamma^t \mathbb{E}_{\tau\sim\pi}[\bar{A}(s_t)]| \\
&\leq \sum_{t=0}^{\infty} \gamma^t (4\epsilon\alpha(1-(1-\alpha)^t)) \\
&= 4\epsilon\alpha\Big(\sum_{t=0}^{\infty} \gamma^t - \sum_{t=0}^{\infty} (\gamma - \gamma\alpha)^t\Big) \\
&= 4\epsilon\alpha\Big(\frac{-1}{\gamma-1} + \frac{1}{\gamma-\gamma\alpha-1}\Big) \\
&= 4\epsilon\alpha\Big(\frac{1}{(1-\gamma)} - \frac{1}{(1-\gamma(1-\alpha))}\Big) \\
&= 4\epsilon\alpha\Big(\frac{\gamma\alpha}{(1-\gamma)(1-\gamma(1-\alpha))}\Big) \\
&\leq 4\epsilon\alpha\Big(\frac{\gamma\alpha}{(1-\gamma)^2}\Big)
\end{aligned}$$

où la seconde inégalité provient de l'application du lemme 2 sur chacun des deux termes de la somme et où on utilise ensuite que $\sum_{t=0}^{n-1} x^t = \frac{x^n-1}{x-1}$ et que $\gamma < 1$ et $\alpha \in [0;1]$.

Donc $-J(\pi') + J(\pi) + L^\pi(\pi') \leq \frac{4\epsilon\gamma\alpha^2}{(1-\gamma)^2}$ et alors $J(\pi') - J(\pi) \geq L^\pi(\pi') - \frac{4\epsilon\gamma\alpha^2}{(1-\gamma)^2}$.

On peut alors s'appuyer sur ce résultat pour montrer enfin que :

$$J(\pi') - J(\pi) \geq L^\pi(\pi') - C\ D_{KL}^{max}(\pi||\pi')$$

Soient P_X et P_Y deux distributions telles que la divergence de variation totale $D_{TV}(P_X||P_Y) = \alpha$. Alors il existe une distribution jointe de (X,Y), de marginales P_X et P_Y, telle que $X \neq Y$ avec une probabilité α.

Si on prend $\alpha = \max_s D_{TV}(\pi(.|s)||\pi'(.|s))$, alors on assure que (π,π') est une paire de politiques α-couplées (car dans ce cas, on a $P(a\sim\pi(a|s) \neq a\sim\pi'(a|s)) \leq \alpha$ pour tout s).

On a donc $J(\pi') - J(\pi) \geq L^\pi(\pi') - C\ D_{TV}^{max}(\pi||\pi')^2$,
avec $C = \frac{4\epsilon\gamma}{(1-\gamma)^2}$ et $D_{TV}^{max}(\pi||\pi') = \max_s D_{TV}(\pi(.|s)||\pi'(.|s))$.

On sait que $D_{TV}(P||Q)^2 \leq D_{KL}(P||Q)$ pour toutes distributions P et Q.
On a alors $D_{TV}^{max}(\pi||\pi')^2 \leq D_{KL}^{max}(\pi||\pi')$, avec $D_{KL}^{max}(\pi||\pi') = \max_s D_{KL}(\pi(.|s)||\pi'(.|s))$

On a donc bien :

$$J(\pi') - J(\pi) \geq L^\pi(\pi') - C\ D_{KL}^{max}(\pi||\pi')$$

A noter que l'on aurait pu prendre la KL dans l'autre sens sachant que la D_{TV} est symétrique.

On a aussi
$$J(\pi') - J(\pi) \geq L^{\pi}(\pi') - C\ D_{KL}^{max}(\pi'||\pi)$$

3 Preuves TRPO

On veut optimiser un problème de la forme :

$$\pi_{k+1} = \arg\max_{\pi'} \mathcal{L}_{\pi_k}(\pi') \quad s.t. \quad \bar{D}_{KL}(\pi_k||\pi') \leq \delta$$

La méthode des gradients naturels propose de s'intéresser à l'expansion de Taylor de second ordre :

$$f(\theta) \approx f(\theta_k) + \nabla_\theta f(\theta)^T_{|\theta_k}(\theta - \theta_k) + \frac{1}{2}(\theta - \theta_k)^T \nabla^2_\theta f(\theta)_{|\theta_k}(\theta - \theta_k)$$

Appliquée à notre problème, on montre que cela donne :

$$\mathcal{L}_{\theta_k}(\theta) \approx g^T(\theta - \theta_k) \text{ avec } g = \nabla_\theta \mathcal{L}_{\theta_k}(\theta)|_{\theta_k}$$

$$\bar{D}_{KL}(\theta_k||\theta) \approx \frac{1}{2}(\theta - \theta_k)^T F(\theta - \theta_k) \text{ avec } F = \nabla^2_\theta \bar{D}_{KL}(\theta_k||\theta)|_{\theta_k}$$

Selon l'expansion de Taylor de second ordre on a :

$$\bar{D}_{KL}(\theta_k||\theta) \approx \bar{D}_{KL}(\theta_k||\theta_k) + \nabla_\theta \bar{D}_{KL}(\theta_k||\theta)^T_{|\theta_k}(\theta - \theta_k) + \frac{1}{2}(\theta - \theta_k)^T \nabla^2_\theta \bar{D}_{KL}(\theta_k||\theta)_{|\theta_k}(\theta - \theta_k)$$

On a :

$$\begin{aligned}\nabla_\theta \bar{D}_{KL}(\theta_k||\theta) &= \nabla_\theta \int P(x|\theta_k)\log P(x|\theta_k) - P(x|\theta_k)\log P(x|\theta)dx \\ &= -\int P(x|\theta_k)\nabla_\theta \log P(x|\theta)dx\end{aligned}$$

Pris en θ_k, cela donne :

$$\begin{aligned}\nabla_\theta \bar{D}_{KL}(\theta_k||\theta)_{|\theta_k} &= -\int P(x|\theta_k)\nabla_\theta \log P(x|\theta)_{|\theta_k} dx \\ &= -\int P(x|\theta_k)\frac{\nabla_\theta P(x|\theta)_{|\theta_k}}{P(x|\theta_k)}dx \\ &= -\nabla_\theta \int P(x|\theta)dx_{|\theta_k} = \nabla_\theta 1 = 0\end{aligned}$$

D'autre part :

$$\begin{aligned}\nabla^2_\theta \bar{D}_{KL}(\theta_k||\theta) &= -\nabla_\theta \int P(x|\theta_k)\nabla_\theta \log P(x|\theta)dx \\ &= -\int P(x|\theta_k)\nabla^2_\theta \log P(x|\theta)dx \\ &= -\int P(x|\theta_k)\nabla_\theta \Big[\frac{\nabla_\theta P(x|\theta)}{P(x|\theta)}\Big]dx\end{aligned}$$

Pris en θ_k, cela donne :

$$\begin{aligned}
\nabla^2_\theta \bar{D}_{KL}(\theta_k||\theta)_{|\theta_k} &= -\int P(x|\theta_k)\nabla_\theta\Big[\frac{\nabla_\theta P(x|\theta)}{P(x|\theta)}\Big]_{|\theta_k} dx \\
&= -\int P(x|\theta_k)\frac{\nabla_\theta[\nabla_\theta P(x|\theta)]_{|\theta_k} P(x|\theta_k) - \nabla_\theta P(x|\theta)_{|\theta_k}\nabla_\theta P(x|\theta)^T_{|\theta_k}}{P(x|\theta_k)^2} dx \\
&= -\int \nabla_\theta[\nabla_\theta P(x|\theta)]_{|\theta_k} dx + \int \frac{\nabla_\theta P(x|\theta)_{|\theta_k}\nabla_\theta P(x|\theta)^T_{|\theta_k}}{P(x|\theta_k)} dx \\
&= -\nabla^2_\theta \int P(x|\theta) dx_{|\theta_k} + \int P(x|\theta_k)\nabla_\theta \log P(x|\theta)_{|\theta_k}\nabla_\theta \log P(x|\theta)^T_{|\theta_k} dx \\
&= -\nabla^2_\theta 1 + \int P(x|\theta_k)\nabla_\theta \log P(x|\theta)_{|\theta_k}\nabla_\theta \log P(x|\theta)^T_{|\theta_k} dx \\
&= \int P(x|\theta_k)\nabla_\theta \log P(x|\theta)_{|\theta_k}\nabla_\theta \log P(x|\theta)^T_{|\theta_k} dx \\
&= \mathbb{E}_{x\sim P(x|\theta_k)}\Big[\nabla_\theta \log P(x|\theta)_{|\theta_k}\nabla_\theta \log P(x|\theta)^T_{|\theta_k}\Big]
\end{aligned}$$

On note F la matrice de Fisher obtenue : $F = \mathbb{E}_{x\sim P(x|\theta_k)}\Big[\nabla_\theta \log P(x|\theta)_{|\theta_k}\nabla_\theta \log P(x|\theta)^T_{|\theta_k}\Big]$.
On a donc bien :

$$\bar{D}_{KL}(\theta_k||\theta) \approx \frac{1}{2}(\theta-\theta_k)^T F(\theta-\theta_k) \text{ avec } F = \nabla^2_\theta \bar{D}_{KL}(\theta_k||\theta)_{|\theta_k}$$

car $\bar{D}_{KL}(\theta_k||\theta_k) = 0$ et $\nabla_\theta \bar{D}_{KL}(\theta_k||\theta)_{|\theta_k} = 0$.

D'autre part on a : $\mathcal{L}_{\theta_k}(\theta_k) = 0$ et $\nabla^2_\theta \mathcal{L}_{\theta_k}(\theta)_{|\theta_k}$ insignifiant par rapport à $\nabla_\theta \mathcal{L}_{\theta_k}(\theta)_{|\theta_k}$;

On note qu'en prenant la KL dans l'autre sens, on obtiendrait le même résultat.

Selon l'expansion de Taylor de second ordre on a :

$$\bar{D}_{KL}(\theta||\theta_k) \approx \bar{D}_{KL}(\theta_k||\theta_k) + \nabla_\theta \bar{D}_{KL}(\theta||\theta_k)^T_{|\theta_k}(\theta-\theta_k) + \frac{1}{2}(\theta-\theta_k)^T\nabla^2_\theta \bar{D}_{KL}(\theta||\theta_k)_{|\theta_k}(\theta-\theta_k)$$

On a :

$$\begin{aligned}
\nabla_\theta \bar{D}_{KL}(\theta||\theta_k) &= \nabla_\theta \int P(x|\theta)\log P(x|\theta) - P(x|\theta)\log P(x|\theta_k) dx \\
&= \int \nabla_\theta P(x|\theta)\log P(x|\theta) + P(x|\theta)\nabla_\theta \log P(x|\theta) - \nabla_\theta P(x|\theta)\log P(x|\theta_k) dx
\end{aligned}$$

Pris en θ_k, cela donne :

$$\begin{aligned}
\nabla_\theta \bar{D}_{KL}(\theta||\theta_k)_{|\theta_k} &= \int \nabla_\theta P(x|\theta)_{|\theta_k}\log P(x|\theta_k) + P(x|\theta_k)\nabla_\theta \log P(x|\theta)_{|\theta_k} - \nabla_\theta P(x|\theta)_{|\theta_k}\log P(x|\theta_k) dx \\
&= \int P(x|\theta_k)\nabla_\theta \log P(x|\theta)_{|\theta_k} dx \\
&= \int \nabla_\theta P(x|\theta)_{|\theta_k} dx \\
&= \nabla_\theta \int P(x|\theta) dx_{|\theta_k} = \nabla_\theta 1 = 0
\end{aligned}$$

D'autre part :

$$\begin{aligned}\nabla^2_\theta \bar{D}_{KL}(\theta||\theta_k) &= \nabla_\theta \int \nabla_\theta P(x|\theta)\log P(x|\theta) + P(x|\theta)\nabla_\theta \log P(x|\theta) - \nabla_\theta P(x|\theta)\log P(x|\theta_k)dx \\ &= \int \nabla^2_\theta P(x|\theta)\log P(x|\theta) + \nabla_\theta P(x|\theta)\nabla_\theta \log P(x|\theta)^T \\ &+ \quad \nabla^2_\theta P(x|\theta) - \nabla^2_\theta P(x|\theta)\log P(x|\theta_k)dx\end{aligned}$$

Pris en θ_k, cela donne :

$$\begin{aligned}\nabla^2_\theta \bar{D}_{KL}(\theta||\theta_k)_{|\theta_k} &= \int \nabla^2_\theta P(x|\theta)_{|\theta_k}\log P(x|\theta_k) + \nabla_\theta P(x|\theta)_{|\theta_k}\nabla_\theta \log P(x|\theta)^T_{|\theta_k} \\ &+ \quad \nabla^2_\theta P(x|\theta)_{|\theta_k} - \nabla^2_\theta P(x|\theta)_{|\theta_k}\log P(x|\theta_k)dx \\ &= \int \nabla_\theta P(x|\theta)_{|\theta_k}\nabla_\theta \log P(x|\theta)^T_{|\theta_k}dx + \int \nabla^2_\theta P(x|\theta)_{|\theta_k}dx \\ &= \int \nabla_\theta P(x|\theta)_{|\theta_k}\nabla_\theta \log P(x|\theta)^T_{|\theta_k}dx \\ &= \int P(x|\theta_k)\nabla_\theta \log P(x|\theta)_{|\theta_k}\nabla_\theta \log P(x|\theta)^T_{|\theta_k}dx \\ &= \mathbb{E}_{x\sim P(x|\theta_k)}\left[\nabla_\theta \log P(x|\theta)_{|\theta_k}\nabla_\theta \log P(x|\theta)^T_{|\theta_k}\right]\end{aligned}$$

On note F la matrice de Fisher obtenue : $F = \mathbb{E}_{x\sim P(x|\theta_k)}\left[\nabla_\theta \log P(x|\theta)_{|\theta_k}\nabla_\theta \log P(x|\theta)^T_{|\theta_k}\right]$.
On a donc bien :

$$\bar{D}_{KL}(\theta||\theta_k) \approx \frac{1}{2}(\theta-\theta_k)^T F(\theta-\theta_k) \text{ avec } F = \nabla^2_\theta \bar{D}_{KL}(\theta||\theta_k)|_{\theta_k}$$

car $\bar{D}_{KL}(\theta_k||\theta_k) = 0$ et $\nabla_\theta \bar{D}_{KL}(\theta||\theta_k)|_{\theta_k} = 0$.

D'autre part on a : $\mathcal{L}_{\theta_k}(\theta_k) = 0$ et $\nabla^2_\theta \mathcal{L}_{\theta_k}(\theta)|_{\theta_k}$ insignifiant par rapport à $\nabla_\theta \mathcal{L}_{\theta_k}(\theta)|_{\theta_k}$;

Soit le Lagrangien de notre problème considérant les expansions de Taylor précédentes et une contrainte d'inégalité sur la KL (plutôt que l'inégalité $KL \leq \delta$) :

$$L(\theta,\lambda) = g^T(\theta-\theta_k) + \lambda\left(\frac{1}{2}(\theta-\theta_k)^T F(\theta-\theta_k) - \delta\right)$$

Selon les KKT, on a l'optimum : $\begin{cases}\nabla_\theta L(\theta,\lambda) = 0 \\ \nabla_\lambda L(\theta,\lambda) = 0\end{cases}$

En exploitant ces conditions, on vise à montrer que la règle de mise à jour suivante correspond au mouvement des paramètres optimal à chaque étape de l'optimisation selon notre problème de maximisation sous contrainte :

$$\theta_{k+1} = \theta_k + \beta F^{-1} g$$

avec $\beta = -\frac{1}{\lambda} = \sqrt{\frac{2\delta}{g^T F^{-1} g}}$ correspondant à un pas de gradient.

$$L(\theta,\lambda) = g^T(\theta-\theta_k) + \lambda\left(\frac{1}{2}(\theta-\theta_k)^T F(\theta-\theta_k) - \delta\right)$$

On a : $\nabla_\theta L(\theta,\lambda) = g + \lambda F(\theta-\theta_k)$

Selon les KKT on a à l'optimum $\nabla_\theta L(\theta, \lambda)$. Donc :

$$\theta = \theta_k - \frac{1}{\lambda} F^{-1} g$$

On a alors, en remplaçant dans le lagrangien :

$$\begin{aligned} L(\theta, \lambda) &= g^T \Big(\theta_k - \frac{1}{\lambda} F^{-1} g - \theta_k\Big) - \lambda\delta + \frac{\lambda}{2} \Big(\theta_k - \frac{1}{\lambda} F^{-1} g - \theta_k\Big)^T F \Big(\theta_k - \frac{1}{\lambda} g^T F^{-1} g - \theta_k\Big) \\ &= -\frac{1}{\lambda} g^T F^{-1} g - \lambda\delta + \frac{1}{2\lambda} g^T F^{-1} g \\ &= -\frac{1}{2\lambda} g^T F^{-1} g - \lambda\delta \end{aligned}$$

Il en vient que :

$$\nabla_\lambda L(\theta, \lambda) = \frac{1}{2\lambda^2} g^T F^{-1} g - \delta$$

Selon les KKT on a à l'optimum $\nabla_\lambda L(\theta, \lambda)$. Donc :

$$\lambda^2 = \frac{1}{2\delta} g^T F^{-1} g$$

Le terme de droite correspond à une matrice semi-définie positive, on peut donc considérer les deux solutions suivantes :

$$\lambda_1 = \sqrt{\frac{g^T F^{-1} g}{2\delta}} \text{ ou } \lambda_2 = -\sqrt{\frac{g^T F^{-1} g}{2\delta}}$$

La première de ces solutions mène à $\theta = \theta_k - \sqrt{\frac{2\delta}{g^T F^{-1} g}} F^{-1} g$ et au lagrangien :

$$L1 = g^T (\theta_k - \sqrt{\frac{2\delta}{g^T F^{-1} g}} F^{-1} g - \theta_k) + \sqrt{\frac{g^T F^{-1} g}{2\delta}} \Big(\frac{1}{2} (\theta_k - \sqrt{\frac{2\delta}{g^T F^{-1} g}} F^{-1} g - \theta_k)^T F (\theta_k - \sqrt{\frac{2\delta}{g^T F^{-1} g}} F^{-1} g - \theta_k) - \delta\Big)$$

Or :

$$\begin{aligned} \frac{1}{2} (\theta_k - \sqrt{\frac{2\delta}{g^T F^{-1} g}} F^{-1} g - \theta_k)^T F (\theta_k - \sqrt{\frac{2\delta}{g^T F^{-1} g}} F^{-1} g - \theta_k) - \delta &= \frac{1}{2} (\sqrt{\frac{2\delta}{g^T F^{-1} g}} F^{-1} g)^T F (\sqrt{\frac{2\delta}{g^T F^{-1} g}} F^{-1} g) - \delta \\ &= \delta - \delta = 0 \end{aligned}$$

On a alors $L1 = -\sqrt{\frac{2\delta}{g^T F^{-1} g}} g^T F^{-1} g \leq 0$

D'un autre côté, selon λ_2, on obtient $\theta = \theta_k + \sqrt{\frac{2\delta}{g^T F^{-1} g}} F^{-1} g$ et le lagrangien

$$L2 = \sqrt{\frac{2\delta}{g^T F^{-1} g}} g^T F^{-1} g \geq 0$$

L'objectif étant de maximiser selon θ, on prend donc $\lambda = \lambda_2$.

Selon l'expression de θ correspondant, on obtient alors la règle de mise à jour optimale :

$$\theta_{k+1} = \theta_k + \beta F^{-1} g$$

avec $\beta = -\frac{1}{\lambda} = \sqrt{\frac{2\delta}{g^T F^{-1} g}}$

4 Preuves Gradient Naturel avec Fonctions Compatibles

Plutôt que d'utiliser une estimation de A^π pour le gradient, on peut utiliser une approximation f_ϕ, avec f_ϕ une fonction compatible : $f_\phi(s,a) = (\nabla_\theta log \pi_\theta(a|s))^T \phi$

Si on a :

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

Alors on peut montrer que :

$$\nabla_\theta J(\theta) = F\phi$$

Ou encore :

$$\tilde{\nabla}_\theta J(\theta) = \phi$$

On a :

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

et : $f_\phi(s,a) = (\nabla_\theta log \pi_\theta(a|s))^T \phi$
Alors :

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

Donc :

$$\begin{aligned}
\nabla_\theta J(\theta) &= \sum_\tau \pi_\theta(\tau) \left[\sum_{t=0}^{|\tau|-1} \nabla_\theta \log \pi_\theta(a_t|s_t) f_\phi(s_t,a_t) \right] \\
&= \sum_\tau \pi_\theta(\tau) \left[\sum_{t=0}^{|\tau|-1} \nabla_\theta \log \pi_\theta(a_t|s_t) \nabla_\theta \log \pi_\theta(a_t|s_t)^T \phi \right] \\
&= \sum_\tau \pi_\theta(\tau) \left[\sum_{t=0}^{|\tau|-1} \nabla_\theta \log \pi_\theta(a_t|s_t) \nabla_\theta \log \pi_\theta(a_t|s_t)^T \right] \phi \\
&= F\phi
\end{aligned}$$

Et donc $\tilde{\nabla}_\theta J(\theta) = F^{-1} \nabla_\theta J(\theta) = \phi$

Une version incrémentale de TRPO avec fonction compatible pourrait considérer la mise à jour après chaque observation (s_t, a_t) :

$$\theta \leftarrow \theta + \beta_i \frac{\sqrt{2\delta}}{|f_\phi(s_t, a_t)|} \quad \phi$$

La mise à jour de TRPO est :

$$\theta \leftarrow \theta + \beta_i \sqrt{\frac{2\delta}{g^T F^{-1} g}} F^{-1} g$$

avec β_i un pas d'apprentissage < 1.
Or on a avec les fonction compatibles : $F^{-1} \nabla_\theta J(\theta) = \phi$ et donc :

$$\begin{aligned}
\nabla_\theta J(\theta)^T F^{-1} \nabla_\theta J(\theta) &= \nabla_\theta J(\theta)^T \phi \\
&= \sum_\tau \pi_\theta(\tau) \left[\sum_{t=0}^{|\tau|-1} \nabla_\theta \log \pi_\theta(a_t|s_t) f_\phi(s_t, a_t) \right]^T \phi \\
&= \sum_\tau \pi_\theta(\tau) \left[\sum_{t=0}^{|\tau|-1} \nabla_\theta \log \pi_\theta(a_t|s_t)^T \phi f_\phi(s_t, a_t) \right] \\
&= \sum_\tau \pi_\theta(\tau) \left[\sum_{t=0}^{|\tau|-1} f_\phi(s_t, a_t)^2 \right]
\end{aligned}$$

On peut donc mettre à jour après chaque observation (s_t, a_t) selon :

$$\theta \leftarrow \theta + \beta_i \frac{\sqrt{2\delta}}{|f_\phi(s_t, a_t)|} \quad \phi$$