Apprentissage par renforcement

Cours 6: Off-Policy Policy Gradients

Sylvain Lamprier

UE RLD - Master DAC

2019

Off-Policy Policy Gradients

Toutes les méthodes PG type TRPO, contrôlant les déplacements dans l'espace des politiques, utilisent une forme de off-policy employant des ratios d'IS.
Mais elle restent on-policy :

- Elles ne peuvent se servir des anciennes trajectoires que dans une région de confiance restreinte (peu "sample-efficient")

Importance Sampling pour réutiliser d'anciennes trajectoires :

$$\nabla_\theta J^{IS}(\theta) \approx \frac{1}{M} \sum_{\tau^{(i)}} \frac{\pi_\theta(\tau^{(i)})}{\pi_{\theta^{(i)}}(\tau^{(i)})} \left[R(\tau^{(i)}) \nabla_\theta \log \pi_\theta(\tau^{(i)}) \right]$$

Problème : Très forte variance dès lors que $\pi_{\theta(i)}$ et π_θ trop différentes
Proposition : Weighted Importance Sampling pour contrôler la variance

$$J^{WIS}(\theta) = \frac{1}{Z} \sum_{\tau^{(i)}} w(\tau^{(i)}; \theta, \theta^{(i)}) R(\tau^{(i)})$$

avec : $w(\tau_i; \theta, \theta^{(i)}) = \frac{\pi_\theta(\tau^{(i)})}{\pi_{\theta(i)}(\tau^{(i)})}$ et $Z = \sum_{\tau^{(i)}} w(\tau^{(i)}; \theta, \theta^{(i)})$

- $\Rightarrow$ Objectif : limiter la variance en normalisant selon les écarts de politiques considérés
- $\Rightarrow$ Estimateur biaisé (mais consistant : biais qui s'annule asymptotiquement)

On peut alors prouver le gradient [Doe+19] :

$$\nabla_\theta J^{WIS}(\theta) \approx \frac{1}{Z} \sum_{\tau^{(i)}} \nabla_\theta w(\tau_i; \theta, \theta^{(i)}) \left[R(\tau^{(i)}) - J^{WIS}(\theta) \right]$$

Off-Policy Policy Gradients

... Mais encore une forte variance. Dans l'idéal, on aimerait se ramener à des rapports d'Importance Sampling sur des décisions uniques plutôt que sur des trajectoires entières.

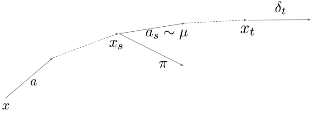
De nombreuses méthodes Off-Policy (Off-PAC [DWS12], RIS-Off-PAC [HC18], ACER [Wan+16], etc.) considèrent l'approximation suivante (avec d^β la distribution stationnaire des états selon la politique β) :

$$
\begin{aligned}
\nabla_\theta J(\theta) &= \nabla_\theta \mathbb{E}_{s\sim d^\beta}\Big[\sum_{a\in\mathcal{A}} Q^\pi(s,a)\pi_\theta(a|s)\Big] \\
&= \mathbb{E}_{s\sim d^\beta}\Big[\sum_{a\in\mathcal{A}} \big(Q^\pi(s,a)\nabla_\theta\pi_\theta(a|s) + {\color{red}\pi_\theta(a|s)\nabla_\theta Q^\pi(s,a)}\big)\Big] && \textbf{; Derivative product rule.} \\
&\overset{(i)}{\approx} \mathbb{E}_{s\sim d^\beta}\Big[\sum_{a\in\mathcal{A}} Q^\pi(s,a)\nabla_\theta\pi_\theta(a|s)\Big] && \textbf{; Ignore the red part : } {\color{red}\pi_\theta(a|s)\nabla_\theta Q^\pi(s,a)}. \\
&= \mathbb{E}_{s\sim d^\beta}\Big[\sum_{a\in\mathcal{A}} \beta(a|s)\frac{\pi_\theta(a|s)}{\beta(a|s)} Q^\pi(s,a)\frac{\nabla_\theta\pi_\theta(a|s)}{\pi_\theta(a|s)}\Big] \\
&= \mathbb{E}_\beta\Big[{\color{blue}\frac{\pi_\theta(a|s)}{\beta(a|s)}} Q^\pi(s,a)\nabla_\theta \ln \pi_\theta(a|s)\Big] && \textbf{; The blue part is the importance weight.}
\end{aligned}
$$

Bien que biaisé, [DWS12] montre que ce gradient permet de converger vers la politique optimale (localement) dans le cas tabulaire (gradient = 0 pour les mêmes optimas locaux).

Off-Policy Policy Gradients

Comment estimer efficacement Q^π en off-policy ?



Temporal Difference ok... :
$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha\delta_t$ avec $\delta_t = r_t + \gamma\mathbb{E}_{a\sim\pi} Q(s_{t+1}, a) - Q(s_t, a_t)$
... mais propagation (trop) lente des valeurs

Comment mettre en place un algorithme plus efficace "multi-steps learning" ? :

$$Q(s,a) \leftarrow Q(s,a) + \alpha\mathbb{E}_\mu\left[\sum_{t=0}^{k}\gamma^t(\prod_{i=0}^{t}c_i)\,(r_t + \gamma\mathbb{E}_\pi[Q(s_{t+1}, a_{t+1})] - Q(s_t, a_t))\,|s_0 = s, a_0 = a\right]$$

Avec $c_i = \lambda$, on a un algorithme $Q(\lambda)$ classique

Problème : la trajectoire dépend de μ plutôt que de π

[Har+16] montre qu'en utilisant $\lambda \leq \frac{1-\gamma}{\gamma||\mu-\pi||_1}$, on converge vers la vraie valeur Q^π.
Mais on ne connaît pas $||\mu - \pi||_1$ $\Rightarrow$ ne fonctionne que si μ et π très proches.

Off-Policy Policy Gradients

Multi-steps learning (forward view, version tabulaire) :

$Q(s,a) \leftarrow Q(s,a) + \alpha \mathbb{E}_\mu \left[\sum_{t=0}^{k} \gamma^t (\prod_{i=0}^{t} c_i) (r_t + \gamma \mathbb{E}_\pi [Q(s_{t+1}, a_{t+1})] - Q(s_t, a_t)) | s_0 = s, a_0 = a \right]$

Propositions :

- Importance Sampling : $c_i = \frac{\pi(a_i|s_i)}{\mu(a_i|s_i)}$
  - Non biaisé
  - Variance très élevée (possiblement infinie) $\Rightarrow$ Pas stable
- Tree backup TB(λ) [Pre00] : $c_i = \lambda \pi(a_i|s_i)$
  - Repondération des traces par leur probabilité dans la politique π
  - Atténuation exponentielle des importances des δ_t selon leur éloignement dans le temps
  - Fonctionne même pour des politiques μ et π éloignées
  - Coupures prématurées. Pas efficace lorsque π et μ sont proches
- Retrace [Mun+16] : $c_i = \lambda min(1, \frac{\pi(a_i|s_i)}{\mu(a_i|s_i)})$
  - Mix des approches précédentes
  - Pas d'explosion de la variance
  - Pas de coupure prématurée $(min(1, \frac{\pi(a_i|s_i)}{\mu(a_i|s_i)}) > \pi(a_i|s_i))$
  - Garanties de convergence

Dans le cas continu, pour une (sous-) trace τ, évaluation individuelle des valeurs :

$$Q_\tau^{ret}(s_t, a_t) = r_t + \gamma c_{t+1} [Q_\tau^{ret}(s_{t+1}, a_{t+1}) - Q_w(s_{t+1}, a_{t+1})] + \gamma \mathbb{E}_a Q_w(s_{t+1}, a)$$

avec $Q_w(s,a)$ appris selon moindres carrés avec Q^{ret}

Off-Policy Policy Gradients : ACER

ACER[Wan+16] utilise Retrace pour l'estimation de Q et considère le gradient :

$$g = \mathbb{E}_\mu \left[\frac{\pi_\theta(a|s)}{\mu(a|s)} Q^\pi(s,a) \nabla_\theta \ln \pi_\theta(a|s) \right]$$

$$= \mathbb{E}_\mu \Big[\min(c, \omega_t(a)) (Q^{\text{ret}}(s,a) - V_w(s)) \nabla_\theta \ln \pi_\theta(a|s)$$

$$+ \mathbb{E}_{a' \sim \pi} \left[\max(0, \frac{\omega_t(a') - c}{\omega_t(a')}) (Q_w(s,a') - V_w(s)) \nabla_\theta \ln \pi_\theta(a'|s) \right] \Big] \quad ; \textbf{ Avec } \omega_t(a) = \frac{\pi(a|S_t)}{\mu(a|S_t)}$$

Où :

- Le premier terme "clippe" le ratio d'importance sampling pour réduire la variance.
  - Utilise Q^{ret} sur les états-actions des traces collectées selon μ.
- Le second terme corrige le biais engendré par cet ajustement.
  - Activé seulement si $w_t(a) > c$. Borné par 1
  - Lorsque $w_t(a)$ est trop grand, on utilise alors des actions samplées selon π
  - On n'a pas de traces correspondant à ces actions : il est alors possible que l'on n'ait pas accès à des Q^{ret} correspondants
  - Utilisation d'une approximation réseau de neurones Q_w obtenue par minimisation moindre carrés selon Q^{ret}

Off-Policy Policy Gradients : ACER

ACER emploie une approche Trust-Region, mais contrairement à TRPO :

- Considère un pool de trajectoires issues de diverses politiques précédentes (pas seulement la dernière)
- Considère une contrainte KL avec une politique moyenne $\pi_{\theta_a}(.|s) = f(.|\phi_{\theta_a}(s))$, mise à jour régulièrement par : $\theta_a \leftarrow \alpha\theta_a + (1-\alpha)\theta$
- Calcul du gradient selon $\phi_\theta(s)$ plutôt que selon θ :

$$\hat{g}^{acer}(s) = \min(c, \omega(s,a))\big(Q^{\mathrm{ret}}(s,a) - V_w(s)\big)\nabla_{\phi_\theta(s)} \ln \pi_\theta(a|s) +$$
$$\mathbb{E}_{a'\sim\pi}\Big[\max(0, \frac{\omega(s,a')-c}{\omega(s,a')})\big(Q_w(s,a') - V_w(s)\big)\nabla_{\phi_\theta(s)} \ln \pi_\theta(a'|s)\Big]$$

- Recherche d'un vecteur proche de $\hat{g}^{acer}(s)$ qui n'implique pas un trop grand déplacement de politique selon s (recherche d'un vecteur le plus orthogonal à $k = \nabla_{\phi_\theta(s)} D_{KL}[f(.|\phi_{\theta_a}(s))||f(.|\phi_\theta(s))]$). Définition d'une contrainte linéaire pour tout s des trajectoires considérées :

$$z^* = \arg\min_z \frac{1}{2}||\hat{g}^{acer}(s) - z||_2^2$$
$$s.t. \nabla_{\phi_\theta(s)} D_{KL}[f(.|\phi_{\theta_a}(s))||f(.|\phi_\theta(s))]^T \; z \leq \delta$$

  dont la solution qui satisfait les KKT est : $z^* = \hat{g}^{acer}(s) - max(0; \frac{k^T\hat{g}^{acer}(s)-\delta}{||k||_2^2})k$

- Policy gradient considéré selon z^* (backpropagation chain-rule) : $z^*\nabla_\theta\phi_\theta(s)$

Off-Policy Policy Gradients : ACER

ACER est une méthode hybride on-policy/off-policy

- 1 passe On-Policy à chaque itération pour collecter de nouvelles trajectoires
- n passes Off-Policy pour exploiter les trajectoires collectées

Algorithm 1 ACER for discrete actions (master algorithm)

```
// Assume global shared parameter vectors θ and θ_v.
// Assume ratio of replay r.
repeat
    Call ACER on-policy, Algorithm 2.
    n ← Possion(r)
    for i ∈ {1, ··· , n} do
        Call ACER off-policy, Algorithm 2.
    end for
until Max iteration or time reached.
```

Off-Policy Policy Gradients : ACER

Algorithm 2 ACER for discrete actions

Reset gradients $d\theta \leftarrow 0$ and $d\theta_v \leftarrow 0$.
Initialize parameters $\theta' \leftarrow \theta$ and $\theta'_v \leftarrow \theta_v$.
if not On-Policy **then**
  Sample the trajectory $\{x_0, a_0, r_0, \mu(\cdot|x_0), \cdots, x_k, a_k, r_k, \mu(\cdot|x_k)\}$ from the replay memory.
else
  Get state x_0
end if
for $i \in \{0, \cdots, k\}$ **do**
  Compute $f(\cdot|\phi_{\theta'}(x_i))$, $Q_{\theta'_v}(x_i, \cdot)$ and $f(\cdot|\phi_{\theta_a}(x_i))$.
  if On-Policy **then**
    Perform a_i according to $f(\cdot|\phi_{\theta'}(x_i))$
    Receive reward r_i and new state x_{i+1}
    $\mu(\cdot|x_i) \leftarrow f(\cdot|\phi_{\theta'}(x_i))$
  end if
  $\bar{\rho}_i \leftarrow \min\left\{1, \frac{f(a_i|\phi_{\theta'}(x_i))}{\mu(a_i|x_i)}\right\}$.
end for

$$Q^{ret} \leftarrow \begin{cases} 0 & \text{for terminal } x_k \\ \sum_a Q_{\theta'_v}(x_k, a) f(a|\phi_{\theta'}(x_k)) & \text{otherwise} \end{cases}$$

for $i \in \{k-1, \cdots, 0\}$ **do**
  $Q^{ret} \leftarrow r_i + \gamma Q^{ret}$
  $V_i \leftarrow \sum_a Q_{\theta'_v}(x_i, a) f(a|\phi_{\theta'}(x_i))$
  Computing quantities needed for trust region updating:

$$\begin{aligned} g &\leftarrow \min\{c, \rho_i(a_i)\} \nabla_{\phi_{\theta'}(x_i)} \log f(a_i|\phi_{\theta'}(x_i))(Q^{ret} - V_i) \\ &\quad + \sum_a \left[1 - \frac{c}{\rho_i(a)}\right]_+ f(a|\phi_{\theta'}(x_i)) \nabla_{\phi_{\theta'}(x_i)} \log f(a|\phi_{\theta'}(x_i))(Q_{\theta'_v}(x_i, a_i) - V_i) \\ k &\leftarrow \nabla_{\phi_{\theta'}(x_i)} D_{KL}\left[f(\cdot|\phi_{\theta_a}(x_i)) \| f(\cdot|\phi_{\theta'}(x_i))\right] \end{aligned}$$

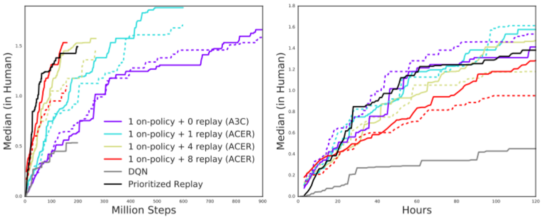
  Accumulate gradients wrt θ': $d\theta \leftarrow d\theta + \frac{\partial \phi_{\theta'}(x_i)}{\partial \theta'}\left(g - \max\left\{0, \frac{k^T g - \delta}{\|k\|_2^2}\right\} k\right)$
  Accumulate gradients wrt θ'_v: $d\theta_v \leftarrow d\theta_v + \nabla_{\theta'_v}(Q^{ret} - Q_{\theta'_v}(x_i, a))^2$
  Update Retrace target: $Q^{ret} \leftarrow \bar{\rho}_i \left(Q^{ret} - Q_{\theta'_v}(x_i, a_i)\right) + V_i$
end for
Perform asynchronous update of θ using $d\theta$ and of θ_v using $d\theta_v$.
Updating the average policy network: $\theta_a \leftarrow \alpha\theta_a + (1-\alpha)\theta$

Politique utilisée pour générer la trajectoire

Possiblement trajectoire incomplète

Retrace update

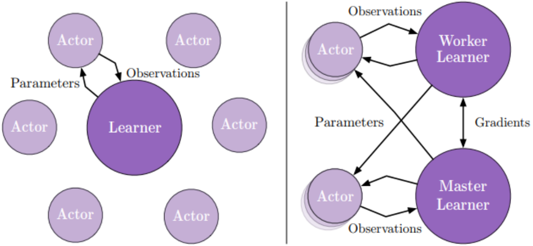
Off-Policy Policy Gradients : ACER



- Pas forcément plus rapide...
- ... Mais bien plus "sample efficient" !

Off-Policy Policy Gradients : Impala

Impala [Esp+18] : Similaire à ACER mais asynchrone

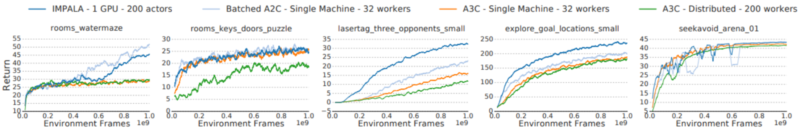


- A3C : Workers font un backward à la fin de chaque trajectoire et transmettent les gradients au master
- Impala : Workers Actor ne font qu'intéragir avec l'environnement et transmettent trajectoires au master qui s'occupe de l'apprentissage
  - Deux versions : une avec un seul learner, une avec plusieurs
  - Du fait du lag entre le moment où la trajectoire est samplée et la politique actuelle du learner, prise en compte de ratios IS comme dans ACER
  - Utilise V-trace plutôt que Retrace (équivalent mais travaille sur V plutôt que Q)

Off-Policy Policy Gradients : Impala

Architecture	CPUs	GPUs[1]	FPS[2]	
Single-Machine			Task 1	Task 2
A3C 32 workers	64	0	6.5K	9K
Batched A2C (sync step)	48	0	9K	5K
Batched A2C (sync step)	48	1	13K	5.5K
Batched A2C (sync traj.)	48	0	16K	17.5K
Batched A2C (dyn. batch)	48	1	16K	13K
IMPALA 48 actors	48	0	17K	20.5K
IMPALA (dyn. batch) 48 actors[3]	48	1	21K	24K
Distributed				
A3C	200	0	46K	50K
IMPALA	150	1	80K	
IMPALA (optimised)	375	1	200K	
IMPALA (optimised) batch 128	500	1	250K	

[1] Nvidia P100 [2] In frames/sec (4 times the agent steps due to action repeat). [3] Limited by amount of rendering possible on a single machine.



Off-Policy Policy Gradients

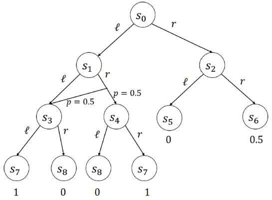
Gradient biaisé de [DWS12] ne permet pas d'assurer la convergence dans le cas général

Cas difficile pour les algos type Off-Pac :

Politique de collecte μ : $\mu(a|s) = 0.5$ pour tout (s,a)

Classe de Politiques cibles Π : $\{\pi_\alpha, \alpha \in [0,1]\}$, où α définit la probabilité de choisir *l* dans les états 1 et 2. Pour tous les autres on a $\pi(a = l|s) = 1$.

Selon α, on a les valeurs V suivantes : $V^{\pi_\alpha}(s_0) = V^{\pi_\alpha}(s_1) = \frac{1+\alpha}{2}$, $V^{\pi_\alpha}(s_2) = \frac{1-\alpha}{2}$, $V^{\pi_\alpha}(s_3) = 1$, $V^{\pi_\alpha}(s_4) = 0$.



$$\begin{aligned} g_{offpac} &= \mathbb{E}_{s\sim d^\mu(s), a\sim\mu(a|s)}[\frac{\nabla_\alpha \pi_\alpha(a|s)}{\mu(a|s)} Q^{\pi_\alpha}(s,a)] \\ &= d^\mu(s_1)(Q^{\pi_\alpha}(s_1, l) - Q^{\pi_\alpha}(s_1, r)) + d^\mu(s_2)(Q^{\pi_\alpha}(s_2, l) - Q^{\pi_\alpha}(s_2, r)) \\ &= 0.5(1 - 0.5) + 0.5(0 - 0.5) = 0 \end{aligned}$$

Gradient nul quelle que soit la valeur de α !

Off-Policy Policy Gradients

[Liu+18] propose de dépasser les limitations dues à l'approximation de [DWS12], en cherchant à caractériser $\frac{d^{\pi}}{d^{\pi_0}}$.

Pour l'évaluation d'une politique π à partir de traces collectées à partir d'une autre politique π_0, la quantité $J(\theta)$ estimée par (Trajectory-Wise Importance Sampling) :

$$R_\pi = \mathbb{E}_{\tau \sim \pi_0}[\sum_{t=0}^{T} w_{0:T} \gamma^t r_t] \text{ avec } w_{0:T} = \prod_{i=0}^{T} \frac{\pi(a_t|s_t)}{\pi_0(a_t|s_t)}$$

En notant que $w_{0:t} = \prod_{i=0}^{t} \frac{\pi(a_t|s_t)}{\pi_0(a_t|s_t)} = \mathbb{E}_{\tau \sim \pi_0}[w_{0:T}|\tau_{0:t}]$ (Rao-Backwellisation), on peut réduire la variance en considérant (Step-Wise Importance Sampling [Pre00]) :

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

[Liu+18] propose d'aller plus loin en considérant une formulation de $J(\theta)$ selon une distribution sur les etats :

$$J(\theta) = \mathbb{E}_{s \sim d^{\pi}(s), a \sim \pi(a|s), s' \sim P(s'|s,a)}[R(s,a,s')], \text{ avec } d^{\pi}(s) = (1-\gamma) \sum_{t=0}^{\infty} \gamma^t d_t^{\pi}(s_t = s)$$

Selon des traces collectées par π_0, on peut alors considérer l'estimateur :

$$R_\pi = \mathbb{E}_{s \sim d^{\pi_0}(s), a \sim \pi_0(a|s), s' \sim P(s'|s,a)}[w(s) \frac{\pi(a|s)}{\pi_0(a|s)} R(s,a,s')], \text{ avec } w(s) = \frac{d^{\pi}(s)}{d^{\pi_0}(s)}$$

Toute la difficulté réside alors dans l'estimation de $w(.)$

Off-Policy Policy Gradients

Toute la difficulté réside alors dans l'estimation de $w_\theta(s) = \frac{d^\pi(s)}{d^{\pi_0}(s)} \forall s$
On commence par noter que, pour tout s', on a :

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

Ou de manière équivalente, si on ajoute une transition fictive (de reward nul) (s_{-1}, a_{-1}, s_0) au temps -1 de chaque trajectoire :

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

Selon π_0, l'estimateur w_θ doit alors vérifier, pour tout $s' \in \mathcal{S}$:

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

Problème : on ne connaît pas $\tilde{d}^{\pi_0}((s,a)|s')$

- dans des MDP de grande taille ou continus, peu d'observations de s'

$\Rightarrow$ Utilisation d'une fonction $f(s')$ pour exploiter la structure de l'espace d'états

Off-Policy Policy Gradients

Selon π_0, l'estimateur w_θ doit alors vérifier, pour toute fonction f :

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

$$\text{avec } \Delta(w_\theta, s, a, s') = w_\theta(s) \frac{\pi(a|s)}{\pi_0(a|s)} - w_\theta(s')$$

Soit : $w^*_\theta = \arg\min_{w_\theta} \max_{f \in \mathcal{F}} L(w_\theta / z_\theta, f)^2$, avec $z_\theta = \mathbb{E}_{s \sim \tilde{d}^{\pi_0}} [w_\theta(s)]$

- Normalisation par z_θ pour éviter les solutions triviales $w_\theta(s) = 0$
- $L(w_\theta / z_\theta, f) = 0$, avec f quelconque : condition nécessaire mais pas suffisante
- $L(w_\theta / z_\theta, \arg\max_f L(w_\theta / z_\theta, f)) = 0$: condition suffisante
- On limite f à une famille de fonctions $\mathcal{F}$ qui permettent d'exploiter l'espace des états (peu d'observations de chaque (s, a) pour chaque s')

Off-Policy Policy Gradients

[Liu+18] considère la famille de fonctions dans la boule unitaire d'un RKHS (par exemple noyau gaussien) : $f(s) =< f, k(s,.) >_{\mathcal{H}}$, avec $||f||_{\mathcal{H}} \leq 1$. On a alors :

$$\begin{aligned} L(w_\theta, f) &= \mathbb{E}_{(s,a,s') \sim \tilde{d}^{\pi_0}((s,a,s')}[\Delta(w_\theta, s, a, s') < f, k(s',.) >_{\mathcal{H}}] \\ &=< f, \mathbb{E}_{(s,a,s') \sim \tilde{d}^{\pi_0}((s,a,s')}[\Delta(w_\theta, s, a, s') k(s',.)] >_{\mathcal{H}} \end{aligned}$$

Or pour $\mathcal{F} : \{f \in \mathcal{H} : ||f||_{\mathcal{H}} \leq 1\}$, $\max\limits_{f \in \mathcal{F}} < f, g >_{\mathcal{H}} = ||g||_{\mathcal{H}}$

On a alors une forme close pour :

$$\begin{aligned} \max_{f \in \mathcal{F}} L(w_\theta, f)^2 &= ||\mathbb{E}_{(s,a,s') \sim \tilde{d}^{\pi_0}((s,a,s')}[\Delta(w_\theta, s, a, s') k(s',.)]||^2_{\mathcal{H}} \\ &= \mathbb{E}_{(s,a,s') \sim \tilde{d}^{\pi_0}} \mathbb{E}_{(\bar{s},\bar{a},\bar{s}') \sim \tilde{d}^{\pi_0}} k(s', \bar{s}') \end{aligned}$$

Off-Policy Policy Gradients

Algorithm 2 Main Algorithm (Discounted Reward Case)

Input: Transition data $\mathcal{D} = \{s_t, a_t, s'_t, r_t\}_t$ from the behavior policy π_0; a target policy π for which we want to estimate the expected reward. Denote by $\beta_{\pi/\pi_0}(a|s) = \pi(a|s)/\pi_0(a|s)$. Discount factor $\gamma \in (0, 1]$.

Augment the data with dummy data $\{s_{-1}, a_{-1}, s'_{-1}, r_{-1}\}$ for which $r_{-1} = 0$, $s'_{-1} = s_0$ and $\Delta(w;\ s_{-1}, a_{-1}, s'_{-1}) := 1 - w(s_0)f(s_0)$. Add them to $\mathcal{D}$ to form an augment dataset $\tilde{\mathcal{D}}$.

Initial the density ratio $w(s) = w_\theta(s)$ to be a neural network parameterized by θ.

for iteration = 1, 2, ... **do**

Randomly choose a batch $\mathcal{M} \subseteq \{1, \ldots, n\}$ from the augmented transition data $\tilde{\mathcal{D}}$, by selecting time t with probability proportional to γ^{t+1}.

Update the parameter θ by $\theta \leftarrow \theta - \epsilon \nabla_\theta \hat{D}(w_\theta / z_{w_\theta})$, where

$$\hat{D}(w) = \frac{1}{|\mathcal{M}|} \sum_{i,j \in \mathcal{M}} \Delta(w, s_i, a_i, s'_i)\Delta(w, s_j, a_j, s'_j)k(s'_i, s'_j),$$

and z_{w_θ} is a normalization constant $z_{w_\theta} = \frac{1}{|\mathcal{M}|}\sum_{i\in\mathcal{M}} w_\theta(s_i)$.

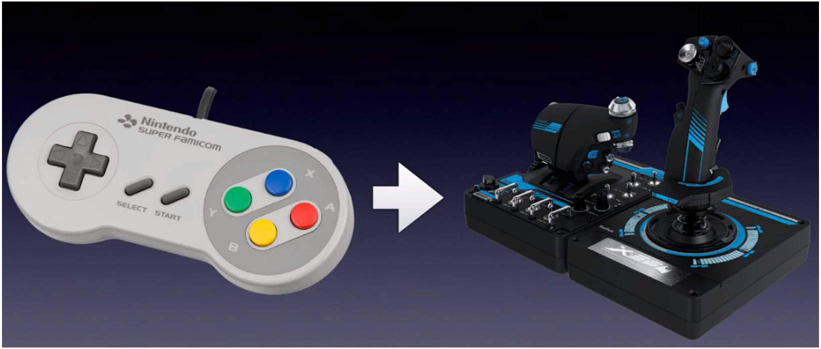
end for

Output: Estimate the expected reward of π by $\hat{R}_\pi = \sum_{i=1}^n v_i r_i / \sum_{i=1}^n v_i$, where $v_i = w_\theta(s_i)\beta_{\pi/\pi_0}(a_i, s_i)$.

[Liu+18] propose un algo de policy gradient basé sur cette estimation de w (+ une augmentation de données pour garantir $\pi(a|s) = 0$ pour tout s, a tel que $\mu(a|s) = 0$) :

- $\nabla_\theta J(\theta) = \mathbb{E}_{s\sim d^\mu, a\sim\mu} \frac{d^\pi(s)}{d^\mu(s)} \frac{\pi(a|s)}{\mu(a|s)} \nabla_\theta log\pi(a|s) Q^\pi$

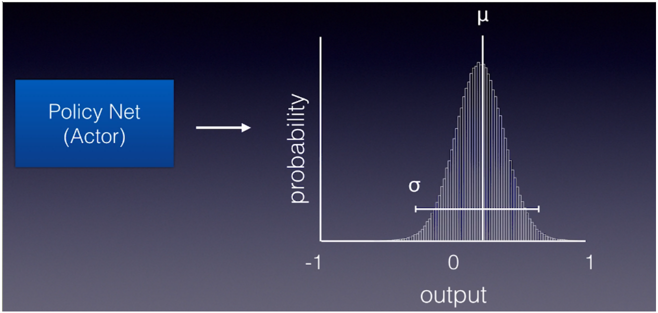
Actions Discrètes $\Rightarrow$ Actions Continues



Policy Gradients for Continuous Action Spaces

Discrete Action Space

Continuous Action Space



Policy Gradients for Continuous Action Spaces

Dans le contexte des problèmes à actions continues, on a $\mathcal{A}(s) \subseteq \mathbb{R}^d$ pour tout s, avec d la dimension des actions considérées.

On définit alors souvent la politique comme une gaussienne multivariée : $\pi_\theta(a|s) = \mathcal{N}(a; \mu_\theta(s), \Sigma_\theta(s))$, avec μ_θ un réseau produisant un vecteur de moyenne pour un état s. Deux possibilités pour $\Sigma_\theta(s)$:

- $\Sigma_\theta(s) = \alpha I$, où α est un vecteur de paramètres indépendants de s
- $\Sigma_\theta(s) = \sigma_\theta(s)^2 I$, où $\sigma_\theta(s)^2$ est réseau (de paramètres souvent partagés avec μ_θ) produisant un vecteur de variance pour l'état s

L'ensemble des algos PG présentés précédemment peuvent alors s'étendre au cas continu (en remplaçant les $\sum_a$ par des $\int_a$).

Notons néanmoins les trois méthodes populaires suivantes, spécifiques au cas des actions continues :

- DPG [Sil+14] : Deterministic Policy Gradient
- DDPG [Lil+15] : Deep Deterministic Policy Gradient
- QPROP [Gu+16] : Sample-Efficient Policy Gradient with an Off-Policy Critic

DPG : Deterministic Policy Gradient

Plutôt que considérer $a \sim \pi(a|s)$, DPG [Sil+14] définit $a = \mu(s)$ afin de limiter la variance des trajectoires.
Objectif :

$$J(\theta) = \int_{s_0} p(s_0) \int_{s_1} p(s_1|\mu(s_0))(r(s_0, \mu(s_0), s_1) + \gamma \int_{s_2} p(s_2|\mu(s_1))(r(s_1, \mu(s_1), s_2) + \gamma \int_{s_3} \dots$$
$$= \int_{\mathcal{S}} p(s_0) Q(s_0, \mu(s_0)) ds_0$$

Gradient de J selon les paramètres de la politique :

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

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

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

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

DPG : Deterministic Policy Gradient

Puisque le cas déterministe est simplement un cas spécial du cas stochastique (Dirac centré sur $\mu(.)$), il est possible d'utiliser tous les algos des PGs. Par exemple, un Actor-Critic (avec a_t l'action choisie à l'instant t) :

$$\begin{aligned}
\delta_t &= R_t + \gamma Q_w(s_{t+1}, a_{t+1}) - Q_w(s_t, a_t) && \text{; TD error in SARSA} \\
w_{t+1} &= w_t + \alpha_w \delta_t \nabla_w Q_w(s_t, a_t) \\
\theta_{t+1} &= \theta_t + \alpha_\theta \nabla_a Q_w(s_t, a_t) \nabla_\theta \mu_\theta(s)|_{a=\mu_\theta(s)} && \text{; Deterministic policy gradient theorem}
\end{aligned}$$

Cependant, on risque de s'enfermer rapidement dans des solutions sous-optimales. Afin d'éviter ce problème, DPG propose de considérer une version Off-Policy (avec remise à jour régulière de la politique β) :

$$\nabla_\theta J_\beta(\theta) = \mathbb{E}_{s \sim \rho^\beta}[\nabla_a Q_w(s, a) \nabla_\theta \mu_\theta(s)|_{a=\mu_\theta(s)}]$$

On note l'absence de ratio IS. Cela est dû à l'aspect déterministe de la méthode : on n'a plus de sampling selon la politique (on devrait néanmoins considérer un ratio IS sur les distributions des états pour être correct).

DDPG : Deep Deterministic Policy Gradient

Deep Deterministic Policy Gradient exploite les mécanismes d'Experience Replay et Target Network de DQN, en suivant les principes de DPG.

Algorithm 1 Deep Deterministic Policy Gradient

Input: initial policy parameters θ, Q-function parameters ϕ, empty replay buffer $\mathcal{D}$

Set target parameters equal to main parameters $\theta_{\text{targ}} \leftarrow \theta$, $\phi_{\text{targ}} \leftarrow \phi$

repeat

  Observe state s and select action $a = \text{clip}(\mu_\theta(s) + \epsilon, a_{Low}, a_{High})$, where $\epsilon \sim \mathcal{N}$

  Ajout d'un bruit gaussien pour exploration + clip pour rester dans des valeurs admissibles

  Execute a in the environment

  Observe next state s', reward r, and done signal d to indicate whether s' is terminal

  Store (s, a, r, s', d) in replay buffer $\mathcal{D}$

  If s' is terminal, reset environment state.

  if it's time to update **then**

    for however many updates **do**

      Randomly sample a batch of transitions, $B = \{(s, a, r, s', d)\}$ from $\mathcal{D}$

      Compute targets

$$y(r, s', d) = r + \gamma(1 - d) Q_{\phi_{\text{targ}}}(s', \mu_{\theta_{\text{targ}}}(s'))$$

Utilisation de réseaux cible (à la fois pour Q et pour μ) pour le calcul de la cible de Q

      Update Q-function by one step of gradient descent using

$$\nabla_\phi \frac{1}{|B|} \sum_{(s,a,r,s',d) \in B} \left(Q_\phi(s, a) - y(r, s', d)\right)^2$$

      Update policy by one step of gradient ascent using

$$\nabla_\theta \frac{1}{|B|} \sum_{s \in B} Q_\phi(s, \mu_\theta(s))$$

Gradient similaire à DPG, selon les transitions du batch

      Update target networks with

$$\phi_{\text{targ}} \leftarrow \rho \phi_{\text{targ}} + (1 - \rho)\phi$$
$$\theta_{\text{targ}} \leftarrow \rho \theta_{\text{targ}} + (1 - \rho)\theta$$

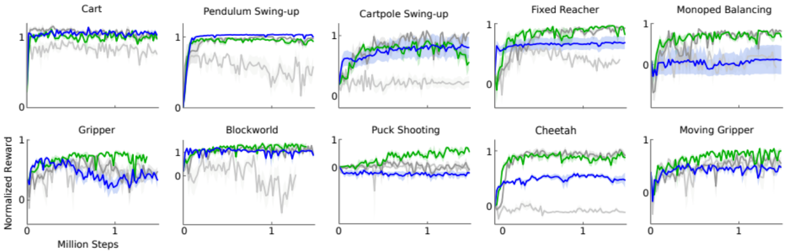
Mise à jour "soft" des paramètres des réseaux cible

    end for

  end if

until convergence

DDPG : Deep Deterministic Policy Gradient



- Gris clair : uniquement batch normalization sur les couches des réseaux
- Gris foncé : uniquement target networks
- Vert : avec target networks et batch normalization
- Bleu : comme vert mais avec pixels comme entrée (plutôt qu'une description spécifique au problème)

$\Rightarrow$ Target Networks essentiels pour stabiliser l'apprentissage

A noter les extensions :

- TD3 [FHM18] : Utilisation de deux réseaux Q pour éviter sa sur-estimation
- D4G [BM+18] : Prioritized Replay + Distributional Critic (e.g., $Q_w(s, a)$ issu d'une mixture de gaussiennes pour apprentissage plus stable)

Q-Prop : Policy Gradient with an Off-Policy Critic

Constat :

- Méthodes du type DDPG ne reposent pas sur des estimateurs de gradient type REINFORCE qui sont sujets à une forte variance
- Peuvent êtres apprises sur des données off-policy, ce qui les rend bien plus "sample-efficient"...
- ... Mais les estimateurs de PG qu'ils utilisent sont biaisés, ce qui rend difficile l'analyse de leur convergence et de leur stabilité

Proposition de Q-Prop : Allier les avantages des deux types de méthodes :

$$\begin{aligned}\nabla_\theta J(\theta) &= \mathbb{E}_{d^\pi,\pi}\left[\nabla_\theta \pi_\theta(a_t|s_t)\left(\hat{A}(s_t,a_t) - \bar{A}(s_t,a_t)\right)\right] \\ &\quad + \mathbb{E}_{d^\pi,\pi}\left[\nabla_\theta \pi_\theta(a_t|s_t)\bar{A}(s_t,a_t)\right] \\ &\quad \text{avec } \bar{A}(s_t,a_t) = A_w(s_t,\mu_\theta(s_t)) + \nabla_a A_w(s_t,a)|_{a=\mu_\theta(s_t)}(a_t - \mu_\theta(s_t)) \\ &\quad\quad = \nabla_a Q_w(s_t,a)|_{a=\mu_\theta(s_t)}(a_t - \mu_\theta(s_t)) \\ &\quad\quad \text{(expansion de Taylor d'ordre 1 en partant de } \mu_\theta(s_t) = \mathbb{E}_{a\sim\pi(a|s_t)}[a])\end{aligned}$$

Soit :

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

Q-Prop : Policy Gradient with an Off-Policy Critic

Après l'introduction d'un coefficient $\eta(s_t)$ qui ne biaise pas le gradient mais vise à réduire la variance, on a :

$$\nabla_\theta J(\theta) = \underbrace{\mathbb{E}_{d^\pi}\left[\eta(s_t)\nabla_a Q_w(s_t,a)|_{a=\mu_\theta(s_t)}\nabla_\theta\mu_\theta(s_t)\right]}_{\text{gradient moyen, équivalent à DDPG}}$$
$$+\underbrace{\mathbb{E}_{d^\pi,\pi}\left[\nabla_\theta\pi_\theta(a_t|s_t)\left(\hat{A}(s_t,a_t)-\eta(s_t)\nabla_a Q_w(s_t,a)|_{a=\mu_\theta(s_t)}(a_t-\mu_\theta(s_t))\right)\right]}_{\text{Prise en compte de l'écart entre les observations et l'estimateur d'avantage}}$$

- ⇒ En supposant que les anciennes politiques couvrent suffisamment l'espace des transitions, on peut évaluer la critique Off-Policy
- ⇒ Q-Prop se sert de cette critique Off-Policy uniquement comme une sorte de Baseline (le gradient ci-dessus est sans biais)

La version conservative de Q-Prop utilise : $\eta(s_t) = 1$ si $\hat{A}(s_t,a_t)\bar{A}(s_t,a_t) > 0$, 0 sinon. Cela permet de limiter les prises en compte de gradient selon la critique aux seuls exemples où l'observation d'avantage et son estimation sont au moins d'accord sur le signe (action à favoriser ou non).

Q-Prop : Policy Gradient with an Off-Policy Critic

Algorithm 1 Adaptive Q-Prop

Initialize w for critic Q_w, θ for stochastic policy π_θ, and replay buffer $\mathscr{R} \leftarrow \emptyset$.
repeat
  for $e = 1, \ldots, E$ **do** ▷ Collect E episodes of on-policy experience using π_θ
    $s_{0,e} \sim p(s_0)$
    for $t = 0, \ldots, T-1$ **do**
      $a_{t,e} \sim \pi_\theta(\cdot|s_{t,e})$, $s_{t+1,e} \sim p(\cdot|s_{t,e}, a_{t,e})$, $r_{t,e} = r(s_{t,e}, a_{t,e})$
  Add batch data $\mathscr{B} = \{s_{0:T,1:E}, a_{0:T-1,1:E}, r_{0:T-1,1:E}\}$ to replay buffer $\mathscr{R}$
  Take $E \cdot T$ gradient steps on Q_w using $\mathscr{R}$ and π_θ
  Fit $V_\phi(s_t)$ using $\mathscr{B}$ Off-Policy Critic
  Compute $\hat{A}_{t,e}$ using GAE(λ) and $\bar{A}_{t,e} = \nabla_a Q_w(s_t, a)|_{a=\mu_\theta(s_t)}(a_t - \mu_\theta(s_t))$.
  Set $\eta_{t,e}$
  Compute and center the learning signals $l_{t,e} = \hat{A}_{t,e} - \eta_{t,e}\bar{A}_{t,e}$ On-Policy Actor
  Compute $\nabla_\theta J(\theta) \approx \frac{1}{ET}\sum_e \sum_t \nabla_\theta \log \pi_\theta(a_{t,e}|s_{t,e}) l_{t,e} + \eta_{t,e} \nabla_a Q_w(s_{t,e}, a)|_{a=\mu_\theta(s_{t,e})} \nabla_\theta \mu_\theta(s_{t,e})$
  Take a gradient step on π_θ using $\nabla_\theta J(\theta)$, optionally with a trust-region constraint using $\mathscr{B}$
until π_θ converges.

Où Q_w est évalué selon le replay buffer avec :

$$w = \arg\min_w \mathbb{E}_{s_t \sim \rho_\beta(\cdot), a_t \sim \beta(\cdot|s_t)}[(r(s_t, a_t) + \gamma \mathbb{E}_\pi[Q'(s_{t+1}, a_{t+1})] - Q_w(s_t, a_t))^2].$$

- Q-Prop très efficace quand la critique a le temps d'être bien apprise entre chaque simulation (dans ce cas, bien meilleur que DDPG). Sinon, si simulations très rapides, équivalent à DDPG et moins bon que TRPO
- Q-Prop bien plus robuste à une mauvaise critique que les méthodes pures off-policy actor-critic telles que DDPG. Améliorations possibles en utilisant Retrace(λ)

Sources I

- Sergey Levine (UC Berkeley, Spring 2017)
- Daniel Takeshi : https://danieltakeshi.github.io/2017/03/28/going-deeper-into-reinforcement-learning-fundamentals-o
- Jonathan Hui : https://medium.com/@jonathan_hui/rl-deep-reinforcement-learning-series-833319a95530
- Lilian Weng : https://lilianweng.github.io/lil-log/2018/04/08/policy-gradient-algorithms.html
- Felix Yu : https ://flyyufelix.github.io/2017/10/12/dqn-vs-pg.html
- Joshua Achiam : http://rail.eecs.berkeley.edu/deeprlcourse-fa17/f17docs/lecture_13_advanced_pg.pdf

Sources II

- Nathan Ratliff : http://ipvs.informatik.uni-stuttgart.de/mlr/wp-content/uploads/2015/01/mathematics_for_intelligent_systems_lecture12_notes_I.pdf
- OpenAI : https://spinningup.openai.com/en/latest/algorithms/trpo.html

References I

[BM+18] Gabriel Barth-Maron et al. « Distributional Policy Gradients ». In : *International Conference on Learning Representations*. 2018.

[Doe+19] Andreas Doerr et al. « Trajectory-Based Off-Policy Deep Reinforcement Learning ». In : *arXiv preprint arXiv :1905.05710* (2019).

[DWS12] Thomas Degris, Martha White et Richard S Sutton. « Off-policy actor-critic ». In : *arXiv preprint arXiv :1205.4839* (2012).

[Esp+18] Lasse Espeholt et al. « IMPALA : Scalable Distributed Deep-RL with Importance Weighted Actor-Learner Architectures ». In : *CoRR* abs/1802.01561 (2018). arXiv : 1802.01561.

[FHM18] Scott Fujimoto, Herke van Hoof et David Meger. « Addressing Function Approximation Error in Actor-Critic Methods ». In : *CoRR* abs/1802.09477 (2018). arXiv : 1802.09477.

[Gu+16] Shixiang Gu et al. « Q-Prop : Sample-Efficient Policy Gradient with An Off-Policy Critic ». In : *CoRR* abs/1611.02247 (2016). arXiv : 1611.02247.

[Har+16] Anna Harutyunyan et al. « Q(\$$\lambda$\$) with Off-Policy Corrections ». In : *CoRR* abs/1602.04951 (2016). arXiv : 1602.04951.

References II

[HC18] Mahammad Humayoo et Xueqi Cheng. « Relative Importance Sampling For Off-Policy Actor-Critic in Deep Reinforcement Learning ». In : *arXiv preprint arXiv :1810.12558* (2018).

[Lil+15] Timothy P Lillicrap et al. « Continuous control with deep reinforcement learning ». In : *arXiv preprint arXiv :1509.02971* (2015).

[Liu+18] Qiang Liu et al. « Breaking the curse of horizon : Infinite-horizon off-policy estimation ». In : *Advances in Neural Information Processing Systems*. 2018, p. 5356-5366.

[Mun+16] Rémi Munos et al. « Safe and efficient off-policy reinforcement learning ». In : *Advances in Neural Information Processing Systems*. 2016, p. 1054-1062.

[Pre00] Doina Precup. « Eligibility traces for off-policy policy evaluation ». In : *Computer Science Department Faculty Publication Series* (2000), p. 80.

[Sil+14] David Silver et al. « Deterministic policy gradient algorithms ». In : 2014.

[Wan+16] Ziyu Wang et al. « Sample efficient actor-critic with experience replay ». In : *arXiv preprint arXiv :1611.01224* (2016).