# Démarche

Mehdi Mounsif

6 mars 2018

## 1 Récap : The day before

- Lecture sur les stratégies d'évolution [4], [2]. Les auteurs supportent les algorithmes génétiques basés sur gradient ascent (resemble à REINFORCE) et expliquent que ceux-ci sont compatibles avec RL et que les deux classes pourraient mutuellement bénéficier de leur spécificités.
- Recherche d'articles sur GANs
- Test du robot 3 joints.

## 2 Learning without a reward function

Dans [3], les auteurs proposent une méthode de RL qui permet d'apprendre en se détachant de la fonction de récompense. La méthode fait appel au principe d'entropie maximum et à la théorie de l'information. La brique de base est dénommée *skill*. Chaque *skill*est en fait une politique $\pi$ qui doit être aussi différent que possible des autres *skills*(dans le sens où les états visités sous l'influence de ce *skill*sont les plus différents possibles que ceux visités avec un autre *skill*).

### 2.1 Principe

Deux phases :

- Exploration non-supervisée : Exploration de l'espace, pas de récompense.
- Stage supervisé : RL classique. L'agent reçoit une récompense et doit la maximiser.

Fonctionnement n'est pas sans évoquer [1]. En effet, on peut comparer la partie non-supervisée à du peaufinage de talent(farming). Les points clés sont :

- Les *skills*doivent se distinguer au niveau des états (au plutôt, on doit pouvoir distinguer les *skills*)
- L'exploration est encouragée en maximisant l'entropie entre les *skills*

Formellement, on chercher à maximiser l'information entre les *skills*et les états $MI(s, z)$, où $z$ correspond à un *skill*. Egalement, on minimise l'information entre les *skills*et les actions en fonction de l'état $MI(a, z|s)$. On veut pour finir maximiser $\mathcal{H}(a|s)$. Soit :

$$\mathcal{F}(\theta) = MI(s, z) + \mathcal{H}(a|s) - MI(a, z|s)$$

Que l'on peut écrire :

$$\mathcal{F}(\theta) = \mathcal{H}(a, z|s)\mathcal{H}(z) - \mathcal{H}(z|s)$$

A finir. J'ai envoyé un mail à Ben Eysenbach `eysenbachbe@google.com` (GoogleBrains) pour demander des clarifications sur les *skills*.

# 3 GANs

- KL divergence : mesure la divergence entre deux distributions $p$ et $q$. La distance est nulle si $p(x) = q(x)$ sur tout l'espace. KL est asymétrique. En effet, on note que si $p(x)$ tend vers 0 et que $q(x)$ est non nulle, alors, $q$ est imperceptible

$$D_{KL}(p||q) = \int_x p(x) \log \frac{p(x)}{q(x)} dx$$

- Jensen-Shannon divergence. Symétrique et plus lisse.

$$D_{JS}(p||q) = \frac{1}{2} D_{KL}(p||\frac{p+q}{2}) + \frac{1}{2} D_{KL}(q||\frac{p+q}{2})$$

- Le mieux est encore d'utiliser Wasserstein distance.

Pour D : Si les données sont *vraies*, on veut maximiser $\mathbb{E}_{x \sim p(x)}[\log(D(x))]$. Si les données sont fausses, ie issues de G, alors on veut que $D(G(Z))$ soit proche de 0, ie. maximiser $\mathbb{E}_{x \sim p(x)}[1 - \log(D(x))]$

# 4 DPPG - PPO

Toujours pas de progrès

# Références


[1] Andrychowicz, M., Wolski, F., Ray, A., Schneider, J., Fong, R., Welinder, P., McGrew, B., Tobin, J., Abbeel, P., and Zaremba, W. Hindsight Experience Replay. *ArXiv e-prints* (jul 2017).

[2] Chrabaszcz, P., Loshchilov, I., and Hutter, F. Back to Basics : Benchmarking Canonical Evolution Strategies for Playing Atari. *ArXiv e-prints* (Feb. 2018).

[3] Eysenbach, B., Gupta, A., Ibarz, J., and Levine, S. Diversity is All You Need : Learning Skills without a Reward Function. *ArXiv e-prints* (Feb. 2018).

[4] Salimans, T., Ho, J., Chen, X., Sidor, S., and Sutskever, I. Evolution Strategies as a Scalable Alternative to Reinforcement Learning. *ArXiv e-prints* (mar 2017).