Apprentissage par Renforcement Développemental

Matthieu Zimmer, Yann Boniface, Alain Dutech

UL/INRIA - LORIA, Nancy

Robotique et IA - PFIA'18

Loria
Laboratoire lorrain de recherche en informatique et ses applications

Plan

Intro
- Contexte et Motivation
- Fonction Valeur

A/C efficace
- Acteur-Critique
- Efficacité en Données

DevRL
- Augmentation sensori-motrice

Conclusion
- Il paraît qu'il faut ramener M. Message à la maison...

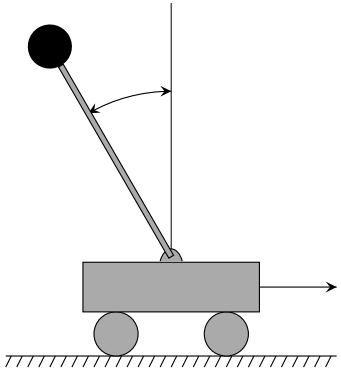
Contexte

Cadre et Difficultés

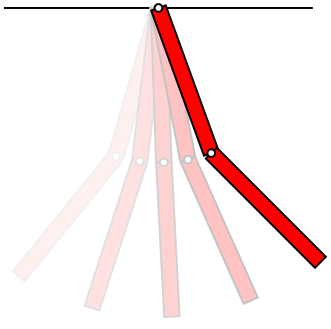
$$\frac{\partial s}{\partial t} = f(s,t) \qquad \text{(système dynamique)}$$

$$s_{t+1} = f(s_t) \qquad \text{(temps discret)}$$

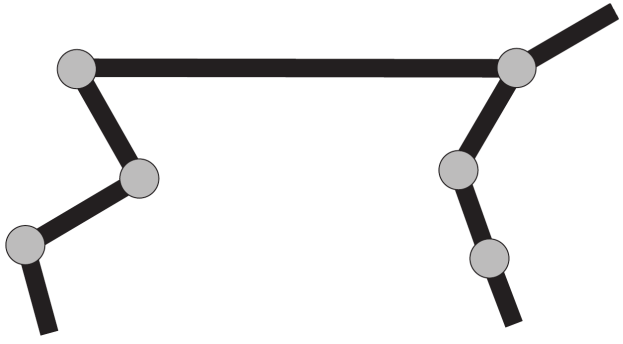
$$\begin{cases} s_{t+1} & = f(s_t, a_t) \\ a_t & = \pi(s_t) \end{cases} \qquad \text{(agent)}$$



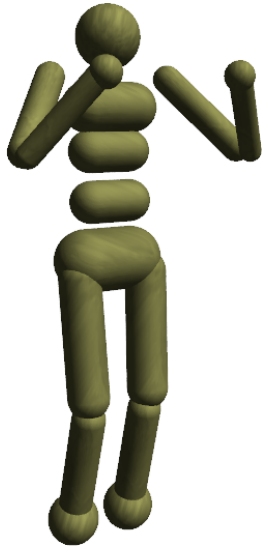
Acrobot



Cartpole



Half-Cheetah

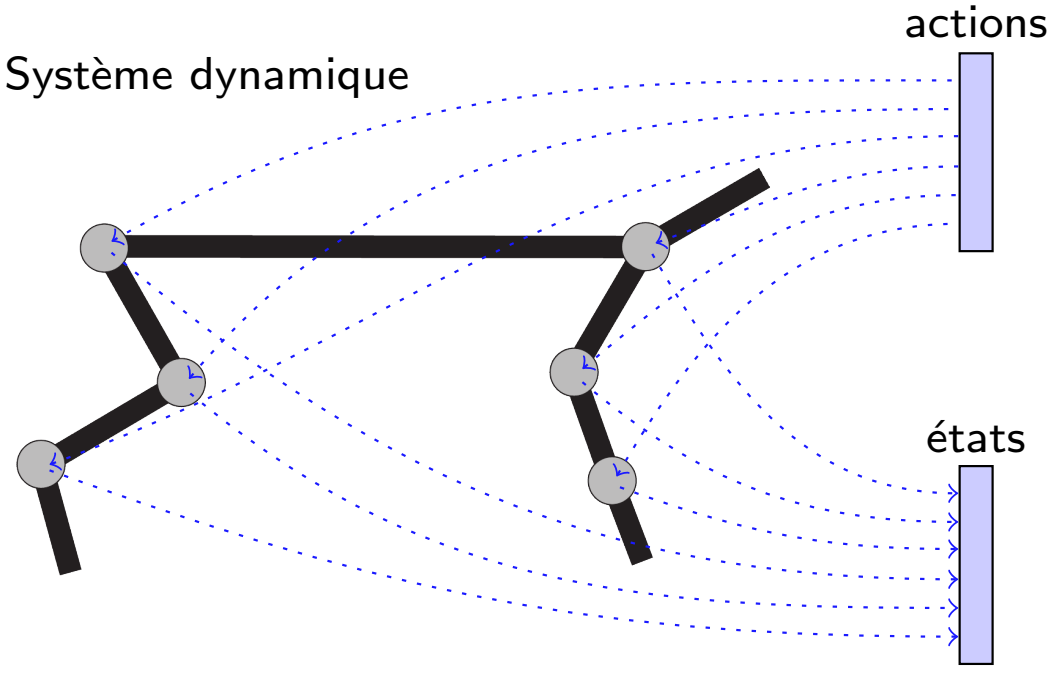


Humanoid

Cadre et Difficultés

$$\begin{cases} s_{t+1} &= f(s_t, a_t) \\ a_t &= \pi(s_t) \end{cases}$$ (agent)

$$\begin{cases} s_{t+1} &= T(s_t, a_t) \\ a_t &= \pi(s_t) \\ r_{t+1} &= R(s_t, a_t, s_{t+1}) \end{cases}$$ (MDP : $\mathcal{S}, \mathcal{A}, T, R$)



$$R(s,a) = \begin{cases} -1000 & \text{si } s \in \mathcal{S}^* \\ \frac{\dot{x}_4}{0.05} - 0.6 \cdot ||a||_2^2 & \text{sinon} \end{cases}$$

Cadre et Difficultés

$$\begin{cases} s_{t+1} &= f(s_t, a_t) \\ a_t &= \pi(s_t) \end{cases} \qquad \text{(agent)}$$

$$\begin{cases} s_{t+1} &= T(s_t, a_t) \\ a_t &= \pi(s_t) \\ r_{t+1} &= R(s_t, a_t, s_{t+1}) \end{cases} \qquad \text{(MDP : } \mathcal{S}, \mathcal{A}, T, R\text{)}$$

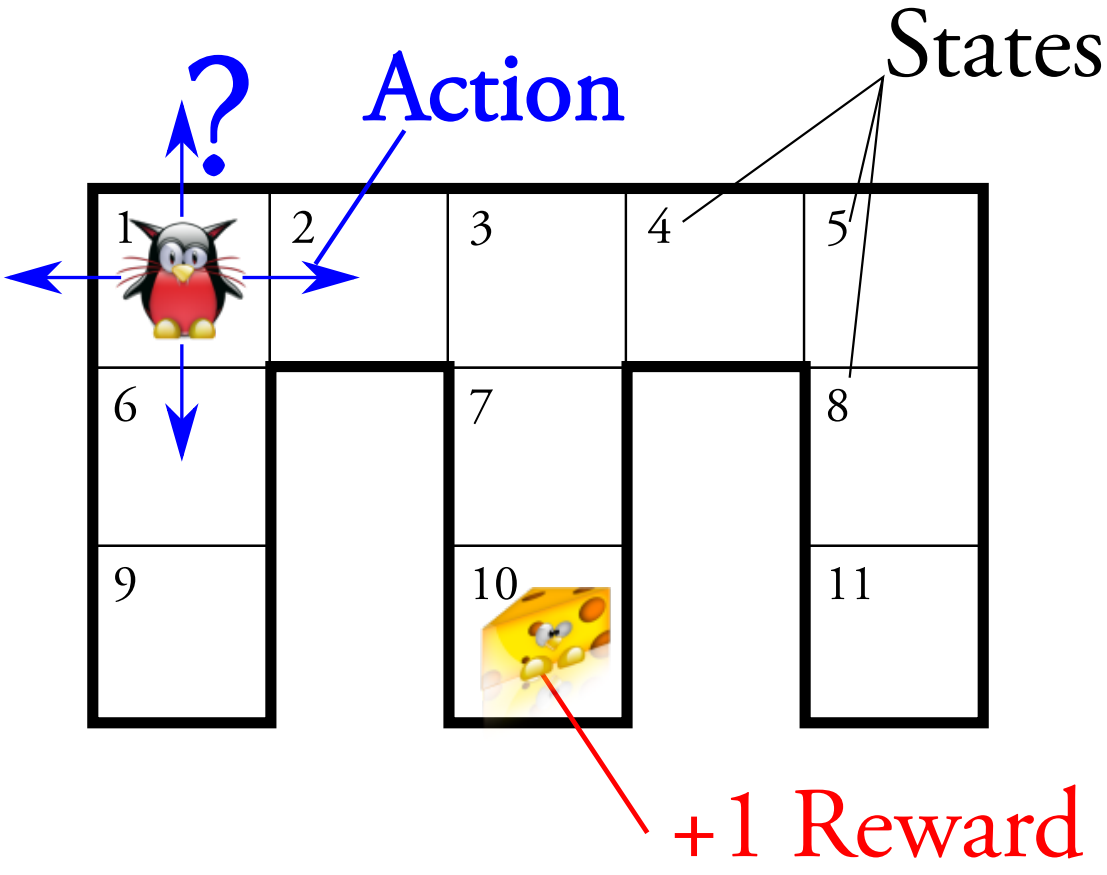
Problème d'optimisation

Trouver politique $\pi : \mathcal{S} \to \mathcal{A}$ qui maximise $J(\pi) = \mathbb{E}_\pi \left[\sum_{t=0}^{\infty} \gamma^t r_t \right]$

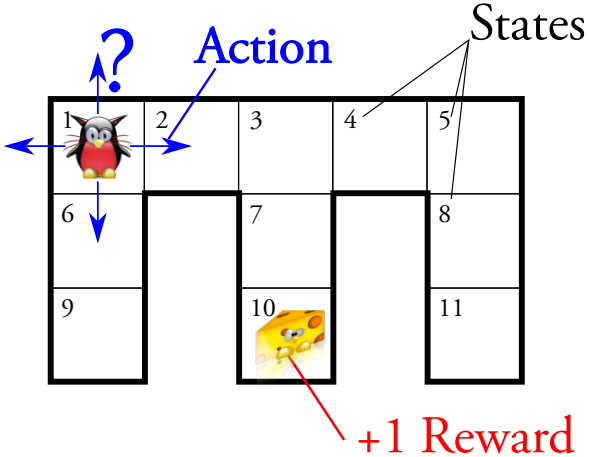
Difficultés

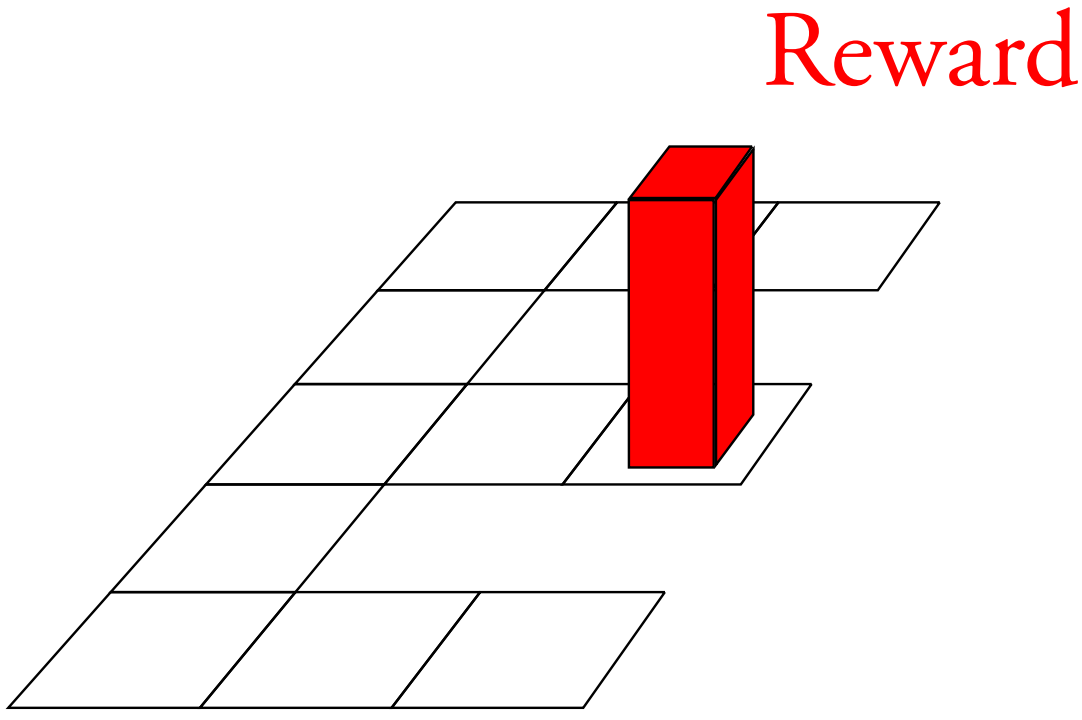
- T (transitions) et R (récompenses) inconnus de l'agent
- $\mathcal{S}$ (états) et $\mathcal{A}$ (actions) continus

Fonction Valeur



Fonction Valeur

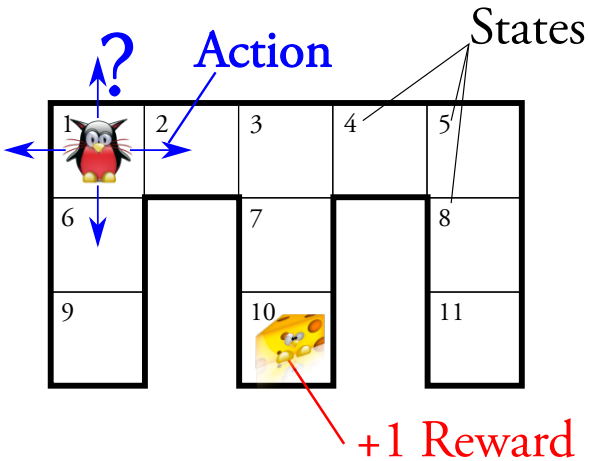


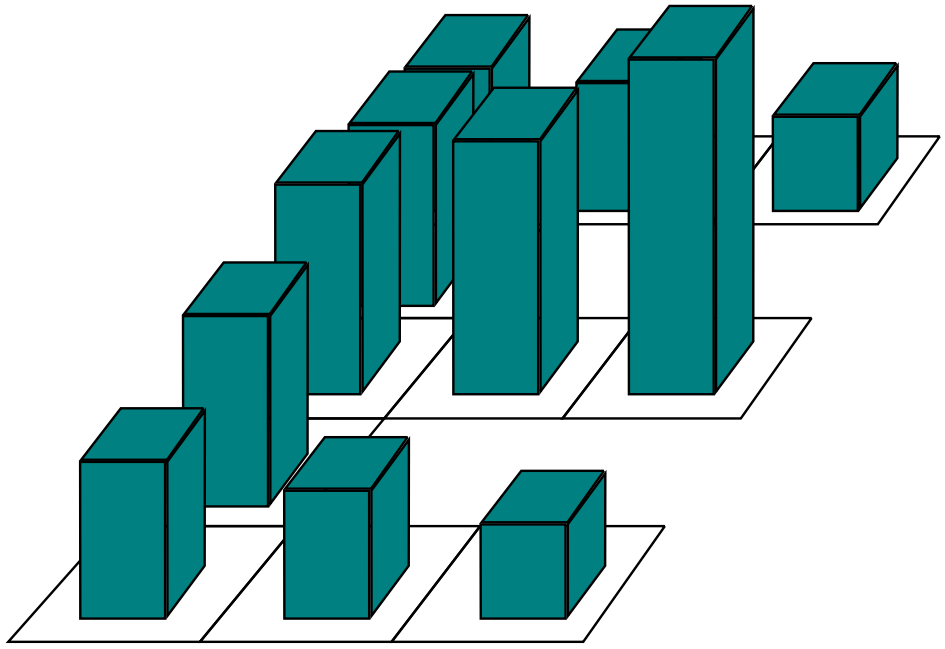


Criteria

Fonction Valeur : $V^{\pi}(s) = \mathbb{E}_{sim\pi}\left[\sum_{t=1}^{T} \gamma^t r_t | s_0 = s\right], \qquad \gamma \in [0, 1[$

Fonction Valeur





Criteria

Fonction Valeur : $V^{\pi}(s) = \mathbb{E}_{sim\pi}\left[\sum_{t=1}^{T} \gamma^t r_t | s_0 = s\right], \qquad \gamma \in [0, 1[$

Plan

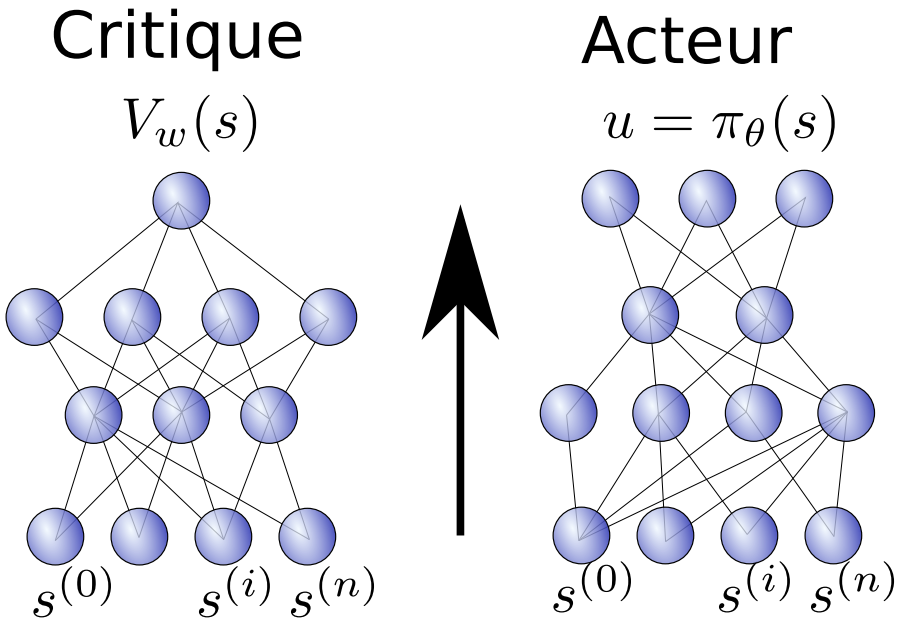
Intro
- Contexte et Motivation
- Fonction Valeur

A/C efficace
- Acteur-Critique
- Efficacité en Données

DevRL
- Augmentation sensori-motrice

Conclusion
- Il paraît qu'il faut ramener M. Message à la maison...

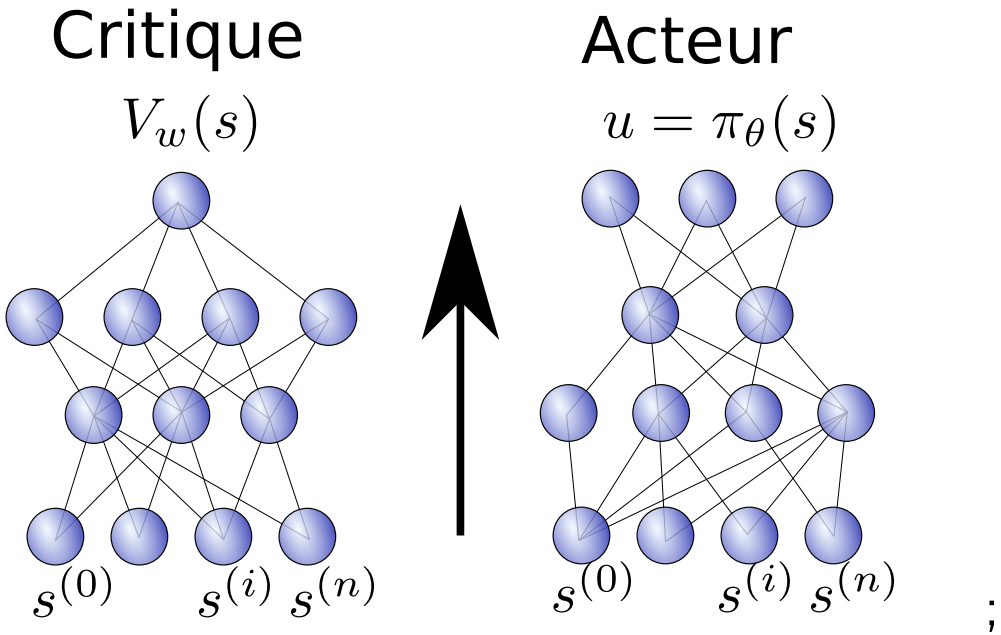
Architecture Acteur-Critique (Deep)



;

Architecture Acteur-Critique (Deep)

Olivier
+
Sigaud
+
youtube
=
2h26'



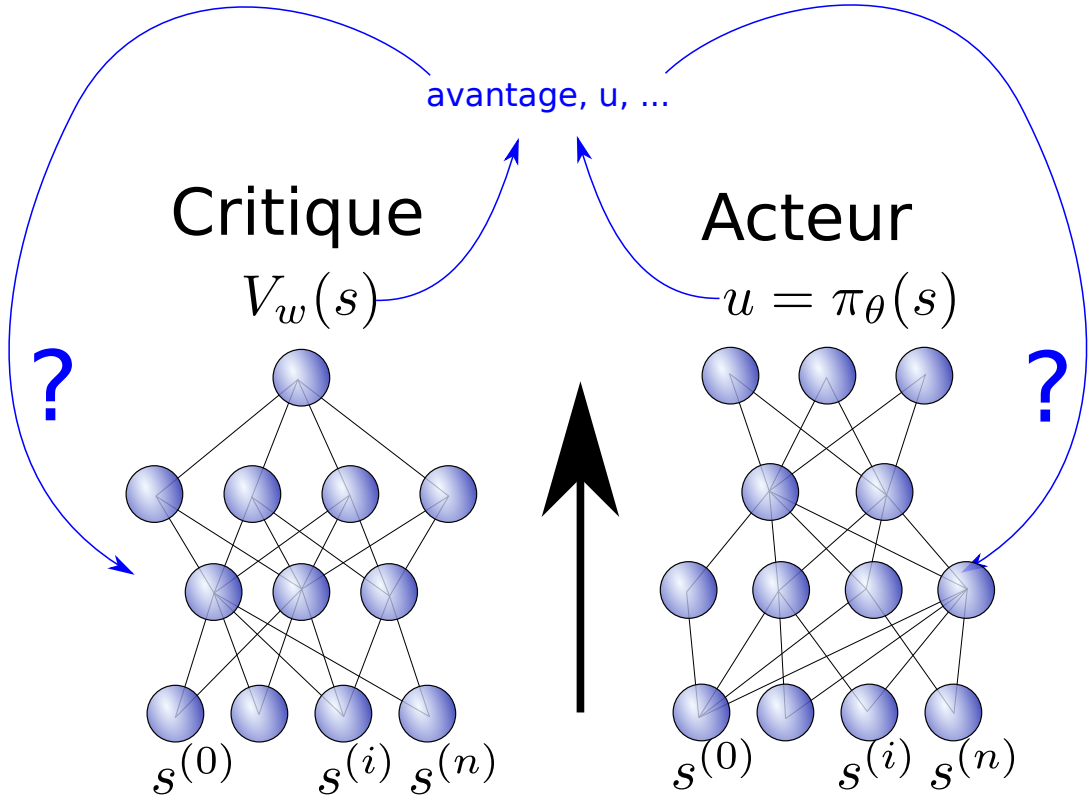
;

Architecture Acteur-Critique (Deep)

Exploration

$$\begin{cases} u_t \leftarrow \pi_\theta(s_t) \\ a_t \leftarrow \mu(u_t, \sigma) \end{cases}$$

μ stoch.



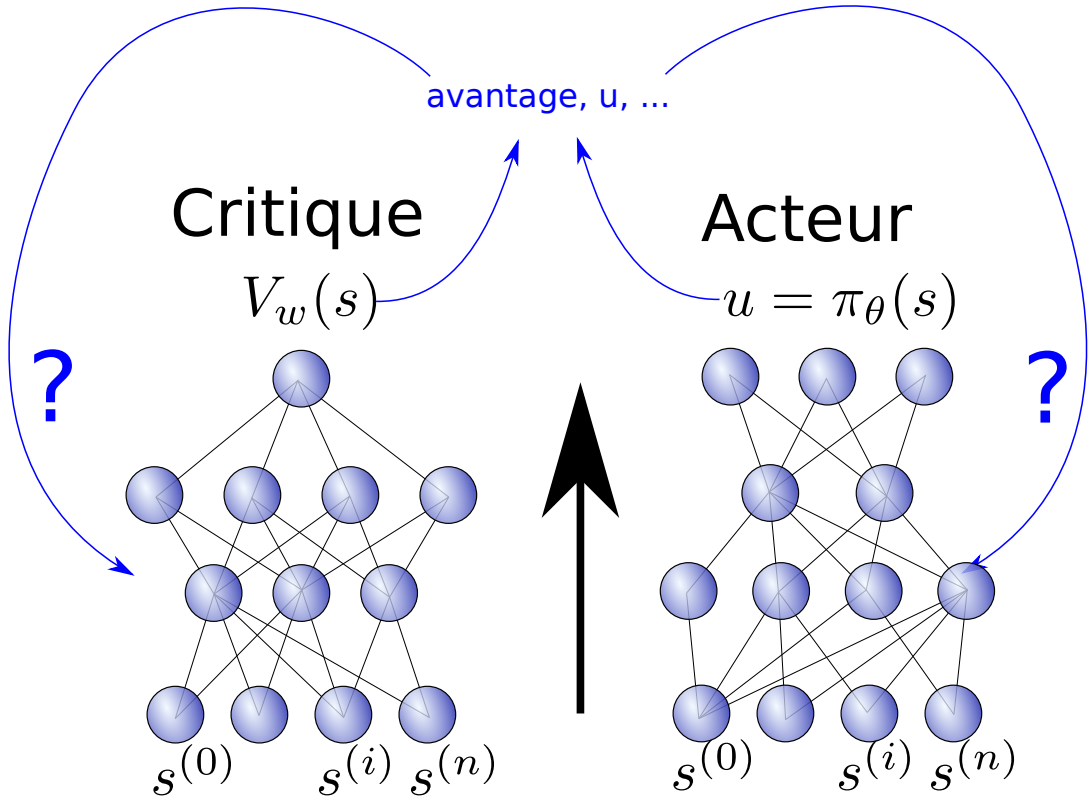
;

Architecture Acteur-Critique (Deep)

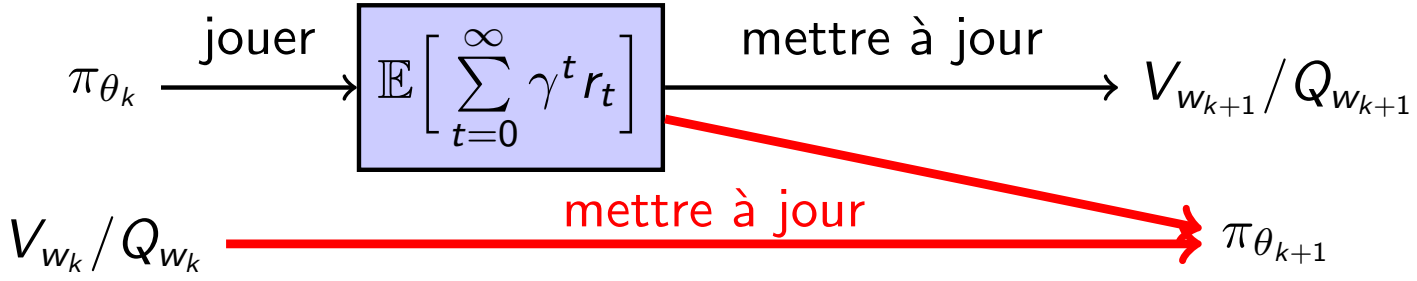
Exploration

$$\begin{cases} u_t \leftarrow \pi_\theta(s_t) \\ a_t \leftarrow \mu(u_t, \sigma) \end{cases}$$

μ stoch.

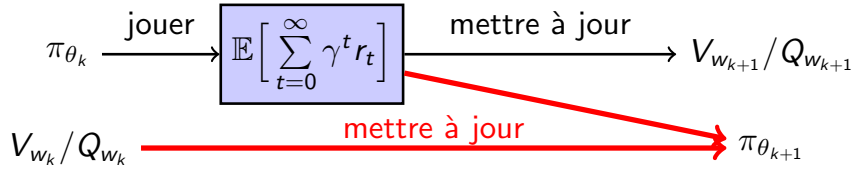


;



Améliorer une politique paramétrée déterministe

$$A^{\pi}(s,a) = Q^{\pi}(s,a) - V^{\pi}(s)$$



- SAC [Sutton et al., 1999]

$$A^{\mu}(s_t,a)\frac{\partial \log \mu_\theta(a|s_t)}{\partial\theta}\Big|_{asim\mu_\theta(\cdot|s_t)}$$

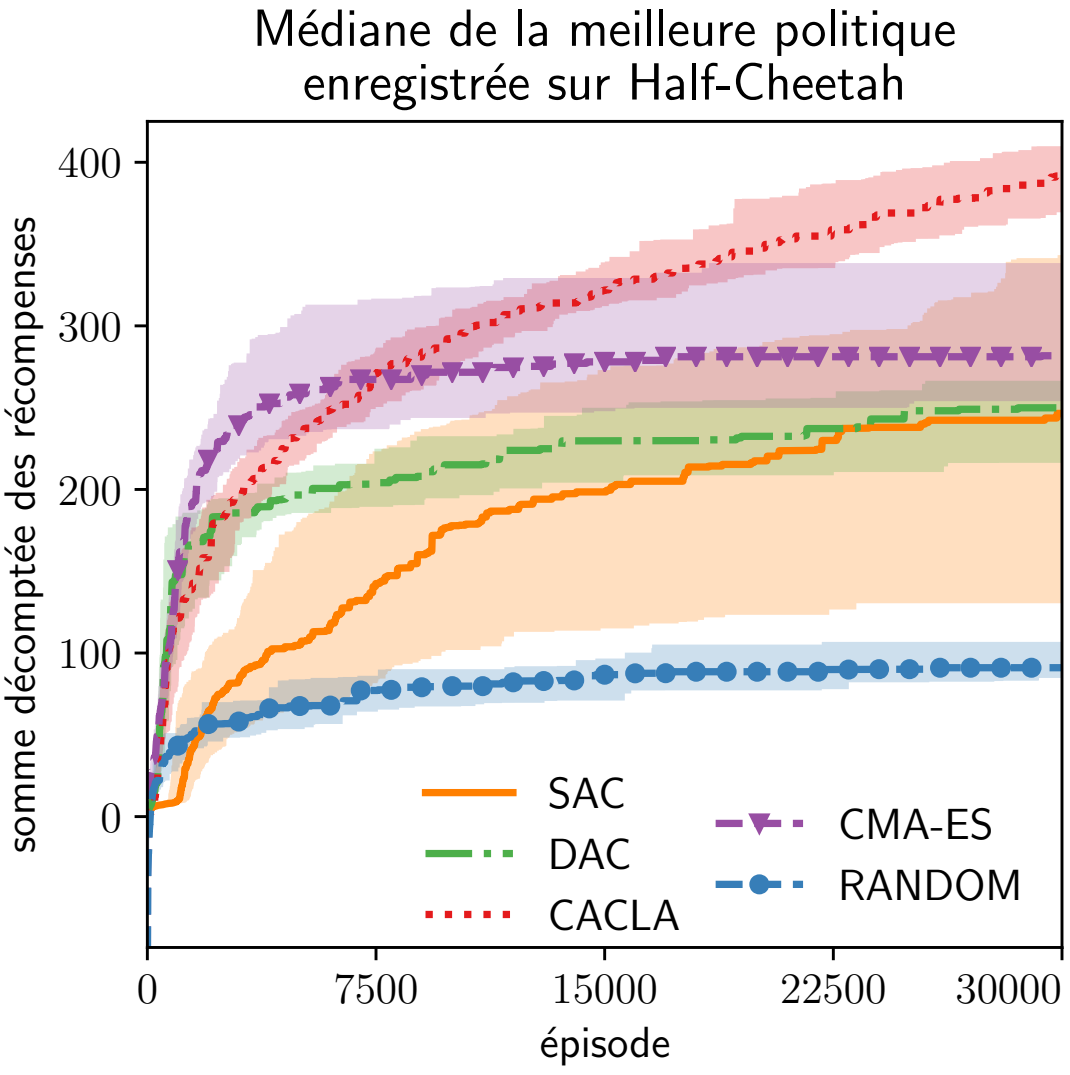
- DAC [Silver et al., 2014, Prokhorov and Wunsch, 1997]

$$\frac{\partial Q^{\mu}(s_t,a)}{\partial a}\frac{\partial \pi_\theta(s_t)}{\partial\theta}\Big|_{a=\pi_\theta(s_t)}$$

- CACLA [Van Hasselt and Wiering, 2007]

$$\underset{A^{\mu}(s_t,a_t)>0}{\mathbb{1}}\Big(a_t - \pi_\theta(s_t)\Big)\frac{\partial \pi_\theta(s_t)}{\partial\theta}\Big|_{a_t sim\mu_\theta(\cdot|s_t)}$$

Quelle est la meilleure *online* ?



acteur : $18 \times 50 \times 25 \times 6$

leaky ReLU (0.01) - TanH

paramètres : ˜4500

médiane sur 50 essais

ADAM (0, 0.999)

temps CPU :

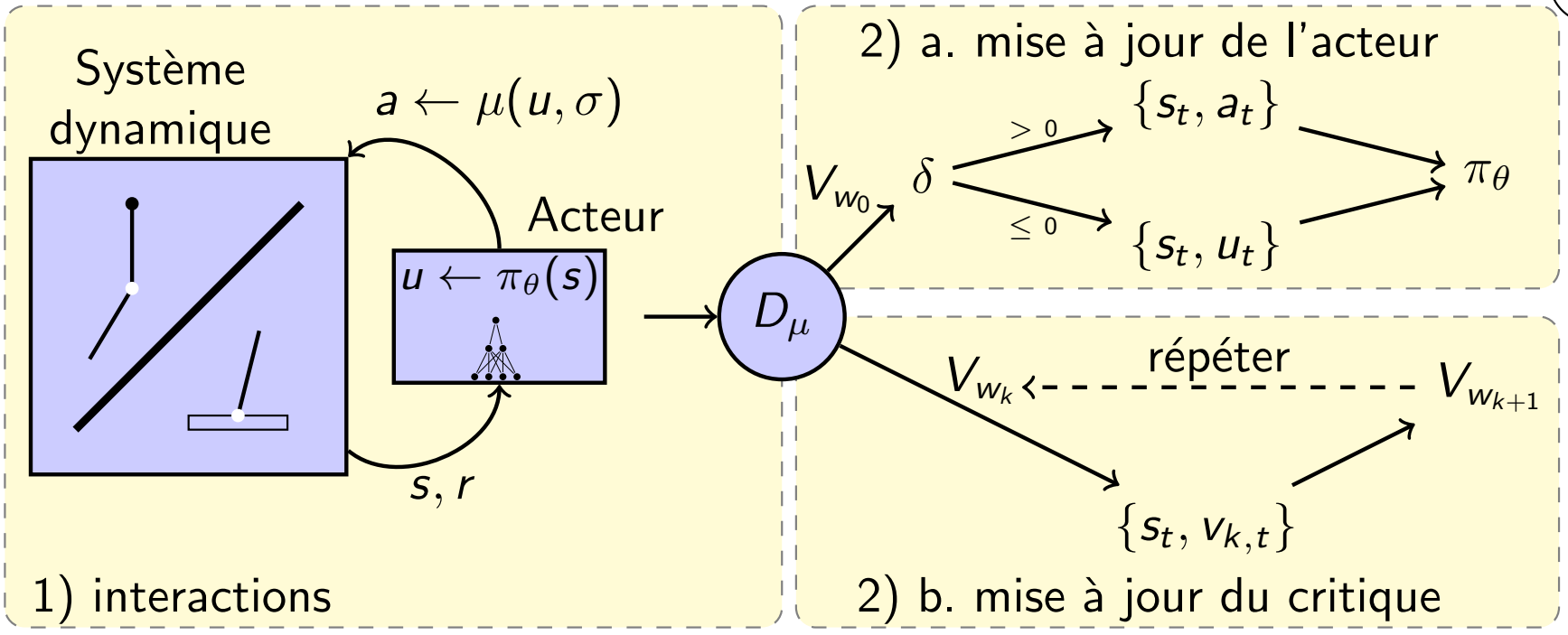
1h30, 2h10, 2h, 1h

μ : loi gaussienne tronquée ($\sigma = 0.1$)

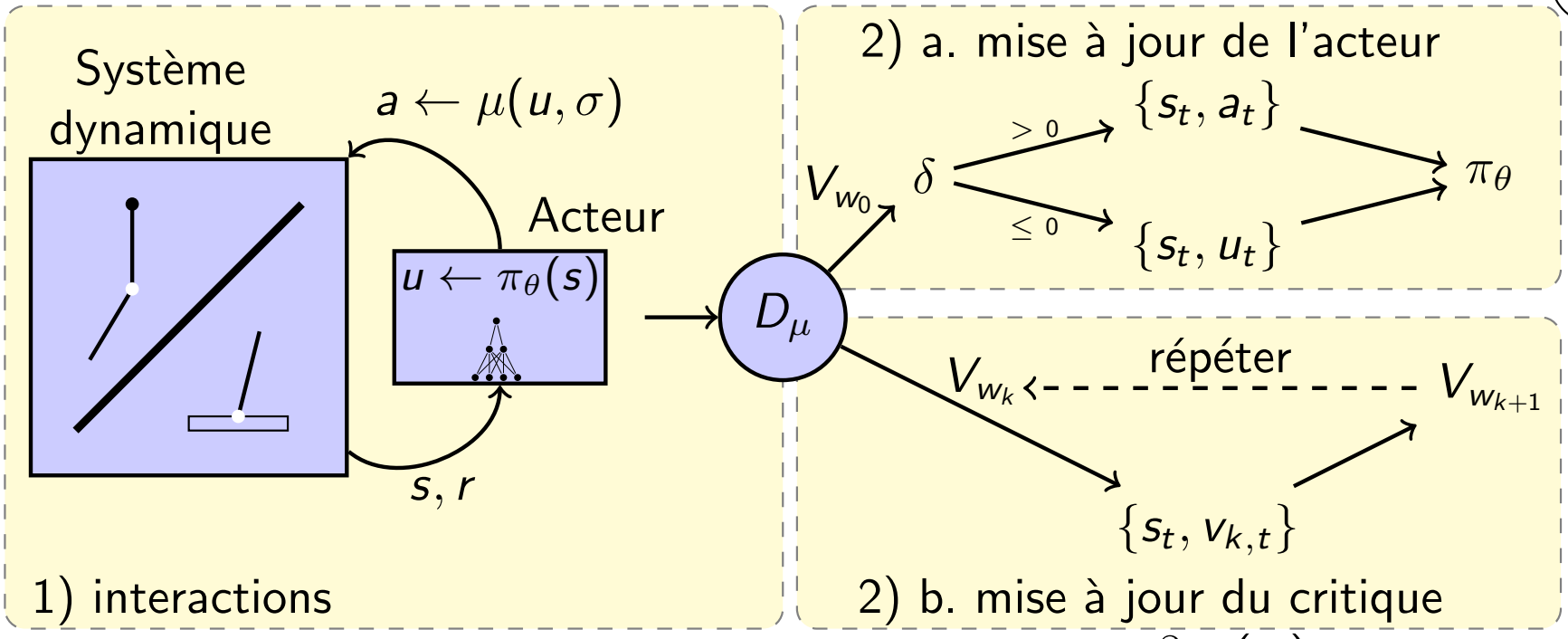
[Zimmer, 2018]

Résultats similaires obtenus par [Van Hasselt and Wiering, 2007] sur Acrobot et Cartpole.

Neural Fitted Actor-Critic



Neural Fitted Actor-Critic



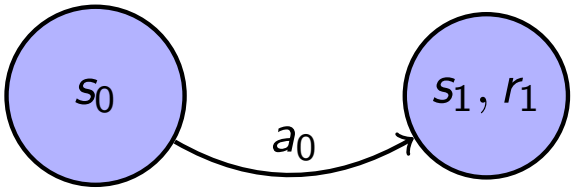
$$\Delta\theta = -\sum_{(s_t,a_t)\in\mathcal{D}_\mu} \mathbb{1}_{\hat{\delta}_t>0} \Big(\underbrace{\pi_\theta(s_t) - a_t}_{\text{rétropropagé}} \Big) \frac{\partial \pi_\theta(s_t)}{\partial\theta}$$

$$\Delta w_k = -\sum_{s_t\in\mathcal{D}_\mu} \Big(\underbrace{V_{w_k}(s_t) - v_{k,t}}_{\text{rétropropagé}} \Big) \frac{\partial V_{w_k}(s_t)}{\partial w_k}$$

ADAM [Kingma and Ba, 2015], Rprop [Igel and Hüsken, 2000], Caffe [Jia et al., 2014]

Dilemme Biais-Variance

NFAC(0)-V + Trace d'éligibilités $\rightarrow$ NFAC(λ)-V

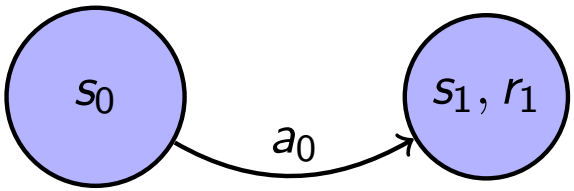


$$\hat{V}^{\pi}(s_0) \leftarrow r_1 + \gamma V_w^{\pi}(s_1)$$

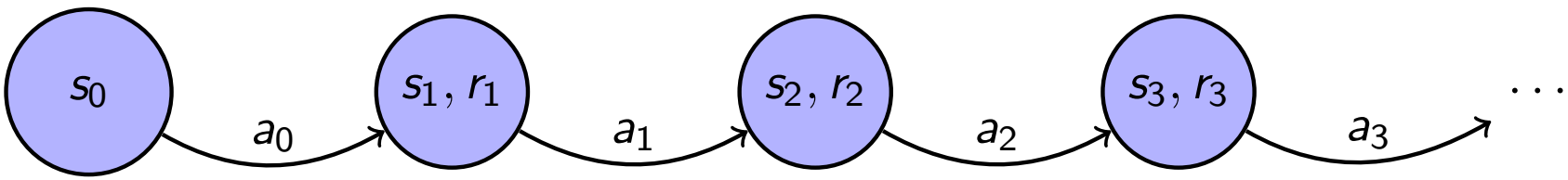
$$\hat{Q}^{\pi}(s_0, a_0) \leftarrow r_1 + \gamma Q_w^{\pi}\big(s_1, \pi(s_1)\big)$$

Dilemme Biais-Variance

NFAC(0)-V + Trace d'éligibilités $\rightarrow$ NFAC(λ)-V

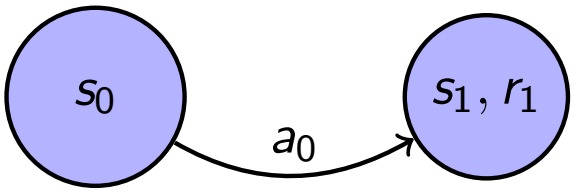


$$\hat{V}^{\pi}(s_0) \leftarrow r_1 + \gamma V_w^{\pi}(s_1)$$
$$\hat{Q}^{\pi}(s_0, a_0) \leftarrow r_1 + \gamma Q_w^{\pi}\big(s_1, \pi(s_1)\big)$$

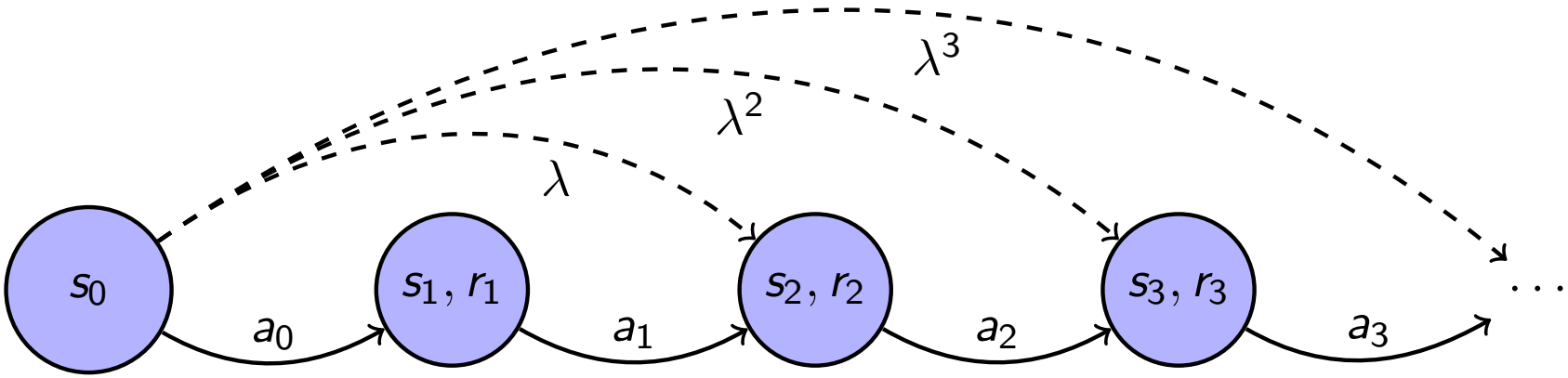


Dilemme Biais-Variance

NFAC(0)-V + Trace d'éligibilités → NFAC(λ)-V

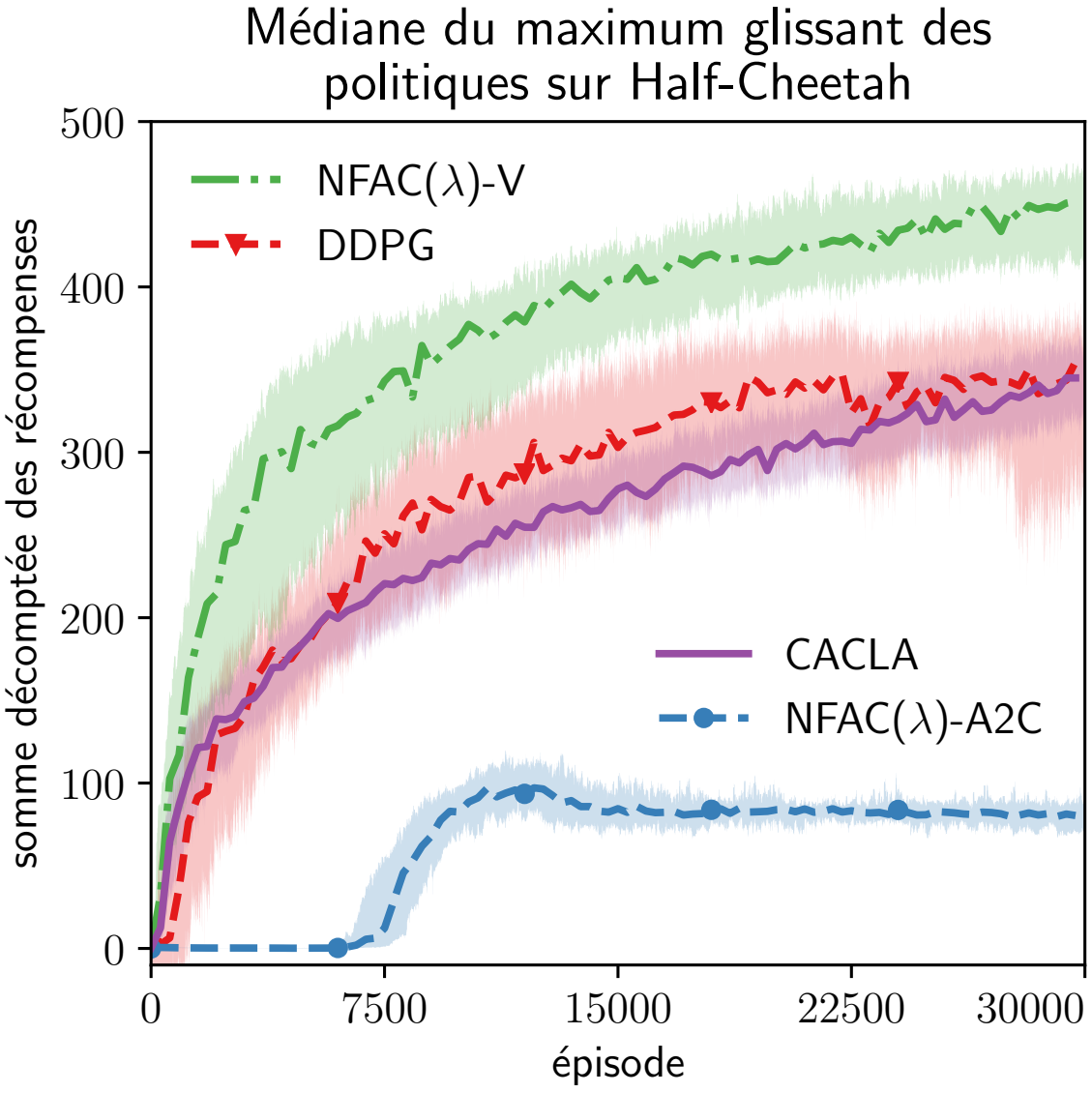


$$\hat{V}^{\pi}(s_0) \leftarrow r_1 + \gamma V_w^{\pi}(s_1)$$
$$\hat{Q}^{\pi}(s_0, a_0) \leftarrow r_1 + \gamma Q_w^{\pi}(s_1, \pi(s_1))$$



$$\hat{V}^{\pi}(s_0) \leftarrow V_w^{\pi}(s_0) + \sum_{n=0}^{\infty} (\gamma\lambda)^n \left(r_{n+1} + \gamma V_w^{\pi}(s_{n+1}) - V_w^{\pi}(s_n) \right)$$

Comparaisons expérimentales



$\lambda = 0.6$

acteur : $18 \times 50 \times 25 \times 6$

leaky ReLU (0.01) - TanH

paramètres : ˜4500

médiane sur 60 essais

ADAM (0, 0.999)

batch normalization (0.999)

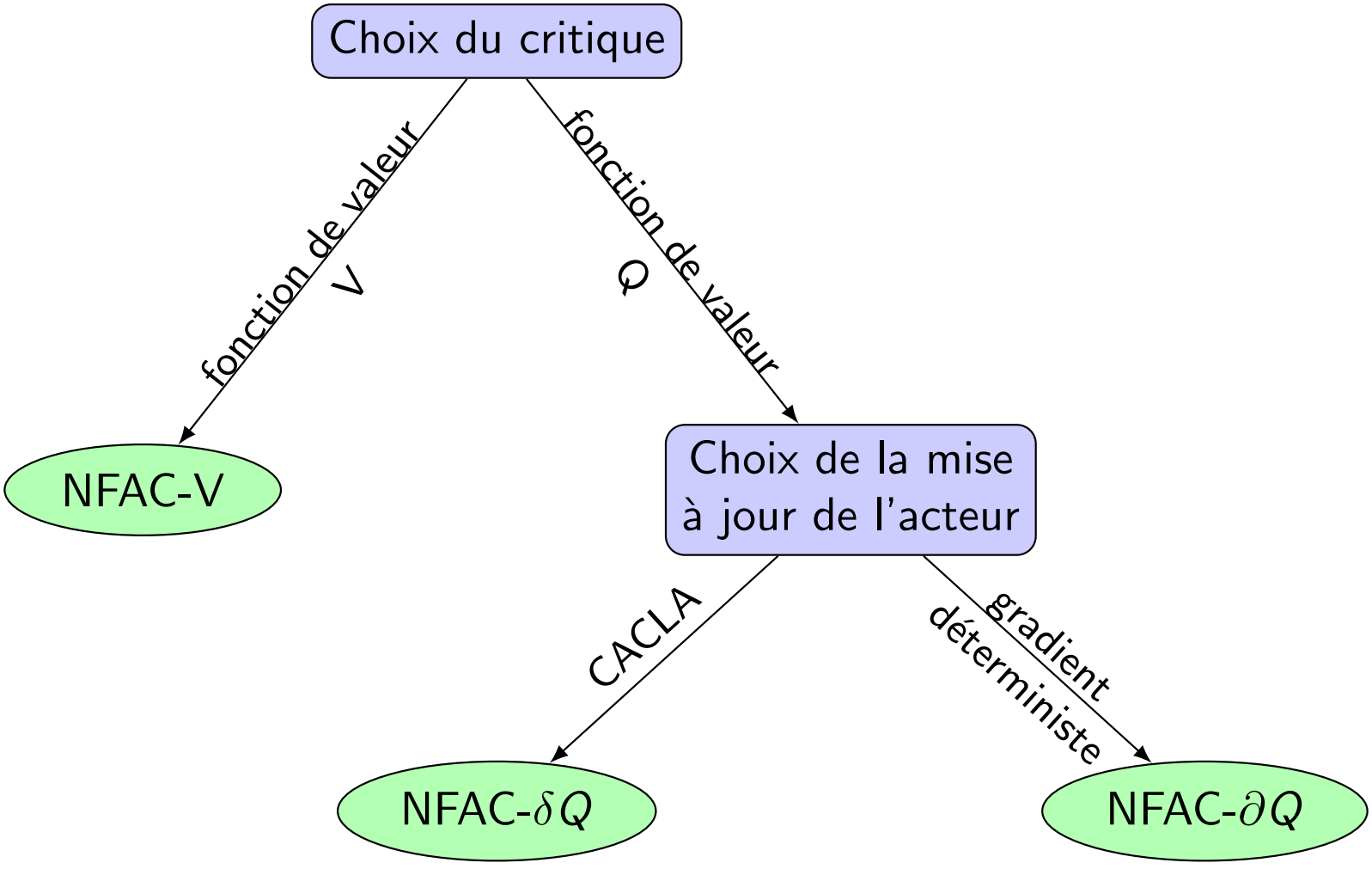
temps CPU :

1h30, 4h20, 2h, 1h20

μ : loi gaussienne tronquée ($\sigma = 0.1$)

[Zimmer, 2018]

Le cadre *Neural Fitted Actor-Critic*



Conclusion intermédiaire

- NFAC(λ)-V (on-policy CACLA) et DDPG (off-policy DAC) sont les plus efficaces
- CACLA avec Q (NFAC-δQ) peu efficace
- transitions off-policy

  $\rightarrow$ one-step Q

  $\rightarrow$ NFAC(λ)-V ne peut en bénéficier directement

Plan

Intro
- Contexte et Motivation
- Fonction Valeur

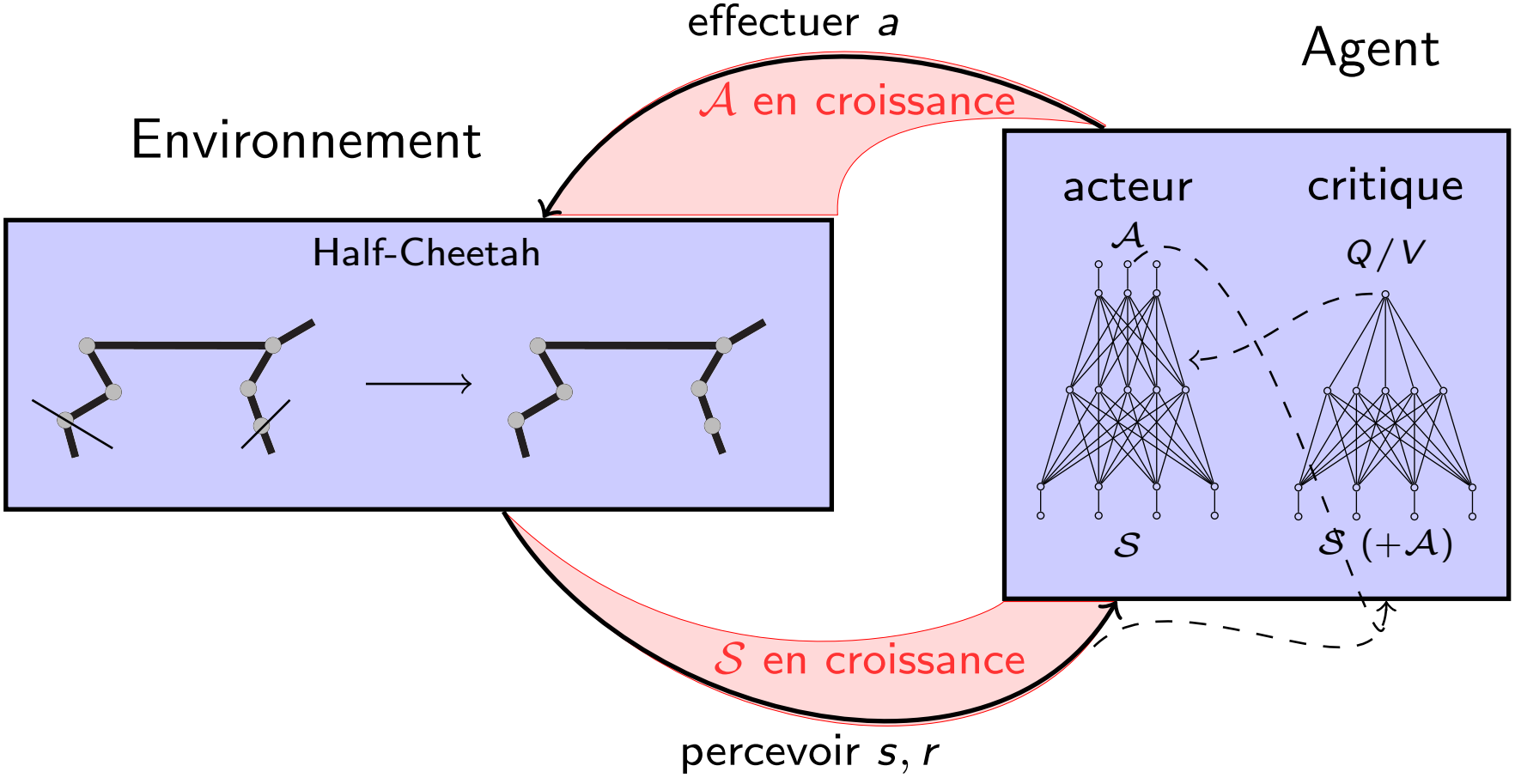
A/C efficace
- Acteur-Critique
- Efficacité en Données

DevRL
- Augmentation sensori-motrice

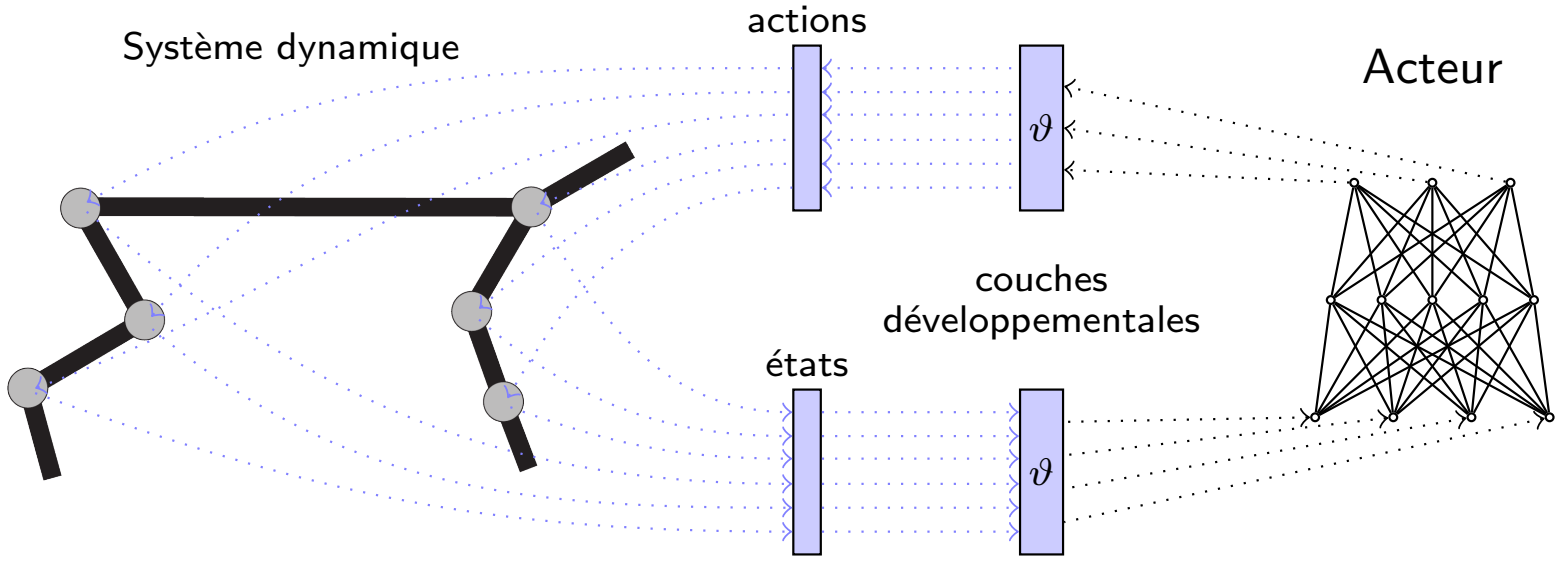
Conclusion
- Il paraît qu'il faut ramener M. Message à la maison...

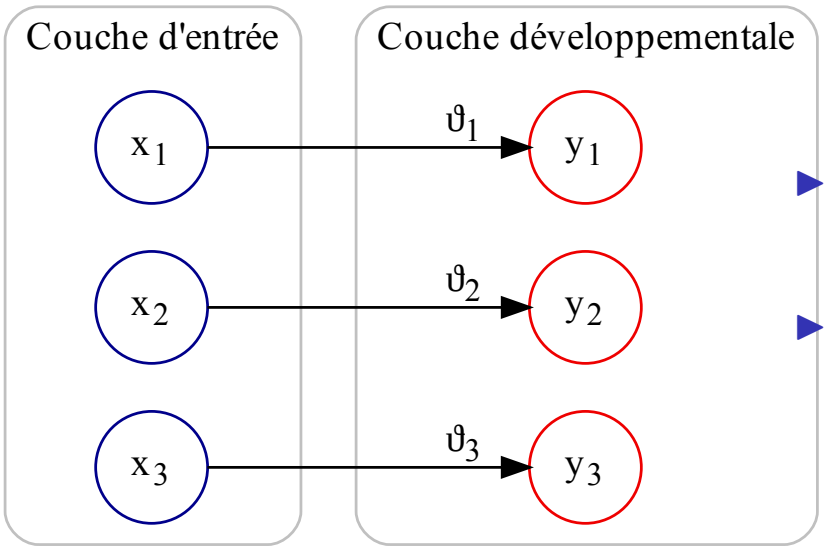
Augmentation progressive de l'espace sensorimoteur

- Transfer learning [Taylor and Stone, 2009]
- Curriculum learning [Bengio et al., 2009]



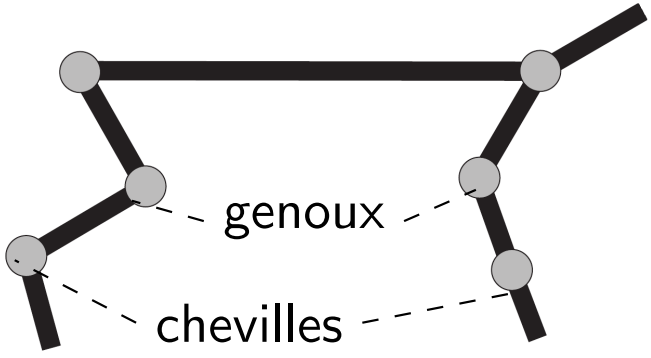
Solution : couches développementales

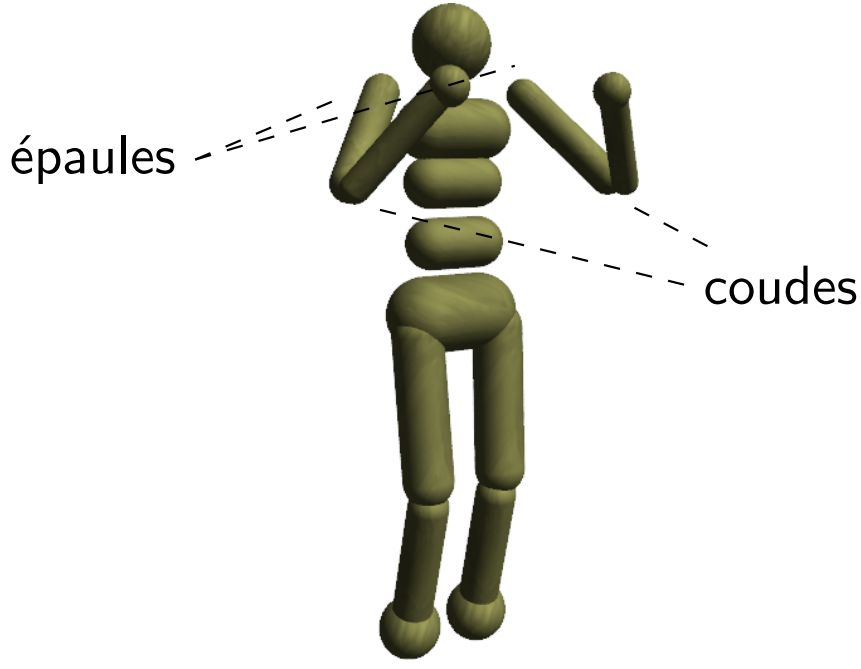




- déterministe $y_i = \begin{cases} x_i & \text{si } \vartheta_i \geq \text{seuil} \\ 0 & \text{sinon} \end{cases}$
- stochastique $\begin{cases} \mathbb{P}(y_i = x_i) = \vartheta_i & \text{si } x_i \neq 0 \\ y_i = 0 & \text{sinon} \end{cases}$

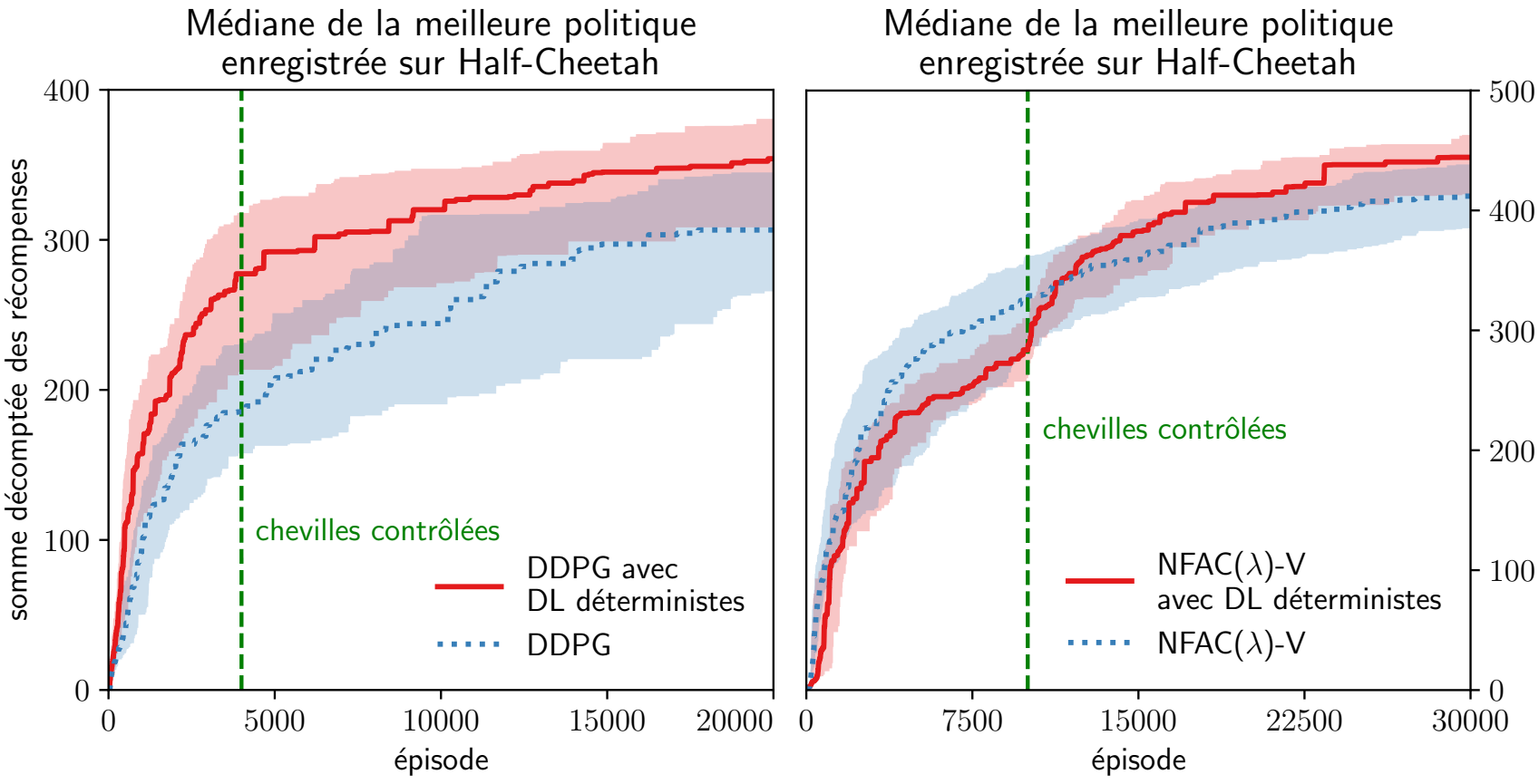
Contexte expérimental





	Offline	Exploration	Acteur	Critique	Off-policy
CMA-ES	oui	espace des paramètres	oui	non	non
NFAC(λ)-V	oui	espace des actions	oui	oui	non
DDPG	non	espace des actions	oui	oui	oui

Comparaisons expérimentales



médiane sur 40 essais

temps CPU : 4h50, 1h30

Conclusion intermédiaire

Exploration guidée de l'espace de recherche

- augmentation de l'espace sensorimoteur
- couches développementales
- fonctionne avec différents algorithmes et systèmes

Limitations

- systèmes dynamiques spécifiques
- actions par défaut
- transfert non perturbé
- récompenses définies premières phases

Plan

Intro
- Contexte et Motivation
- Fonction Valeur

A/C efficace
- Acteur-Critique
- Efficacité en Données

DevRL
- Augmentation sensori-motrice

Conclusion
- Il paraît qu'il faut ramener M. Message à la maison...

Aller plus loin

Apprentissage efficace en données

- Difficile avec approche neuronale
- $\rightsquigarrow$ méthodes **avec modèle** ?
  PILCO [Deisenroth and Rasmussen, 2011],
  Black-DROPS [Chatzilygeroudis and Mouret, 2018], ...

Approche Développementale pour Exploration

Augmentation sensorimoteur peut se combiner

- Reward-Shaping, Scaffholding, ...
- curriculum learning et "Source Task Creation"
  [Narvekar et al., 2016]
- Goal Exploration Processe [Colas et al., 2018]

Références I

Bengio, Y., Louradour, J., Collobert, R., and Weston, J. (2009).

Curriculum Learning.

In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, pages 41–48, New York, NY, USA. ACM.

Chatzilygeroudis, K. and Mouret, J.-B. (2018).

Using parameterized black-box priors to scale up model-based policy search for robotics.

In *IEEE International Conference on Robotics and Automation, ICRA'18*.

Colas, C., Sigaud, O., and Oudeyer, P. (2018).

GEP-PG : decoupling exploration and exploitation in deep reinforcement learning algorithms.

In *Journées Francophones de Planificatin, Décision, Apprentissage JFPDA'18*.

Références II

Deisenroth, M. and Rasmussen, C. (2011).

PILCO : A model-based and data-efficient approach to policy search.

In *Interna- tional Conference on Machine Learning (ICML 2011)*, pages 465–472, ?ew York, NY, USA.

Igel, C. and Hüsken, M. (2000).

Improving the Rprop learning algorithm.

In *International Symposium on Neural Computation*, pages 115–121.

Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., and Darrell, T. (2014).

Caffe : Convolutional Architecture for Fast Feature Embedding.

arXiv preprint arXiv :1408.5093.

Kingma, D. P. and Ba, J. L. (2015).

Adam : a Method for Stochastic Optimization.

International Conference on Learning Representations, pages 1–13.

Références III

Narvekar, S., Sinapov, J., Leonetti, M., and Stone, P. (2016).

Source Task Creation for Curriculum Learning.

In *Proceedings of the 15th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2016)*, Singapore.

Prokhorov, D. V. and Wunsch, D. C. (1997).

Adaptive critic designs.

IEEE Transactions on Neural Networks, 8(5) :997–1007.

Silver, D., Lever, G., Heess, N., Degris, T., Wierstra, D., and Riedmiller, M. (2014).

Deterministic Policy Gradient Algorithms.

Proceedings of the 31st International Conference on Machine Learning, pages 387–395.

Références IV

Sutton, R., Precup, D., and Singh, S. (1999).

Between mdps and semi-mdps : Learning, planning and representing knowledge at multiple temporal scales.

Artificial Intelligence, 112(1-2) :181–211.

Taylor, M. E. and Stone, P. (2009).

Transfer Learning for Reinforcement Learning Domains : A Survey.

Journal of Machine Learning Research, 10 :1633–1685.

Van Hasselt, H. and Wiering, M. A. (2007).

Reinforcement learning in continuous action spaces.

In *Proceedings of the IEEE Symposium on Approximate Dynamic Programming and Reinforcement Learning*, pages 272–279.

Zimmer, M. (2018).

Apprentissage par renforcement développemental.

PhD thesis, Université de Lorraine.