"Actor-Critic" Reinforcement learning

GDR Robotique & Neurosciences

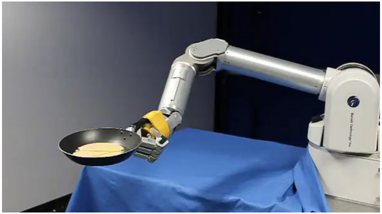
Alain Dutech

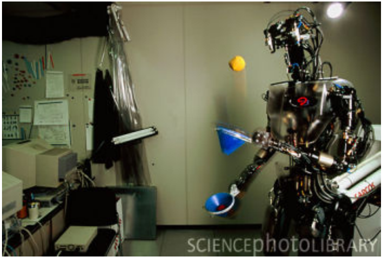
Equipe MAIA - LORIA - INRIA
Nancy, France
Web : http://maia.loria.fr
Mail : Alain.Dutech@loria.fr

Paris - 06/09/2012

Learn to do

2

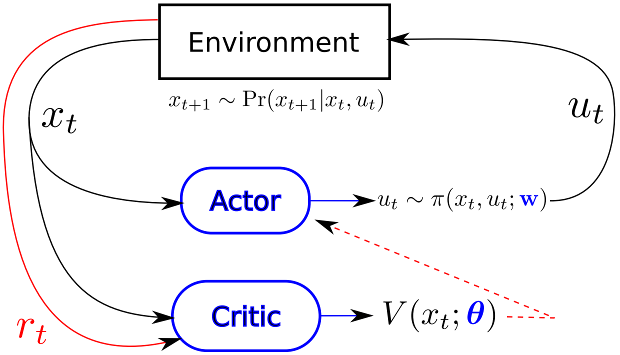




Outline

- **Context : RL and direct policy search**
- Direct Gradient : No Critic
- eNAC : Compatible Critic
- EM : MC Critic
- Summary, discussion

Actor-Critic architecture



Sequence of state × actions :

$$x_0 \xrightarrow{\pi} u_0 \Longrightarrow r_1, x_1 \xrightarrow{\pi} u_1 \Longrightarrow \cdots u_{T-1} \Longrightarrow r_T, x_T$$

Principles of "Direct Policy Search"

- Policy π is **parametrized** (by $\mathbf{w}$)
- Every policy can, theoretically, be valued (V or J or ...)
- Learn/Adapt : **modify** the parameters to increase value

but

- How to modify the parameters ? (**Gradient** or **EM**)
- When is a **critic** needed ?
- Beware of large sensorimotor space, local minima, ...

Common language

Value function.

$$V_\gamma^T(x;\mathbf{w}) = \mathbb{E}_{(r_t)}\left[\sum_{t=1}^{T}\gamma^t r_t | x_0 = x, \pi\right] \tag{1}$$

Objective function

$$J_\gamma^T(\mathbf{w}) = \mathbb{E}_x\left[V_\gamma^T(x)\right] \tag{2}$$

$$\bar{J}(\mathbf{w}) = \lim_{T\to\infty}\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T} r_t\right] \tag{3}$$

Linked :

$$\bar{J}(\mathbf{w}) = \boldsymbol{\mu}^\top(I-\gamma\mathbf{P})J_\gamma^\infty \tag{4}$$

with $\boldsymbol{\mu}$ static distribution and $\mathbf{P}$ transition probabilities.

Outline

- Context : RL and direct policy search
- **Direct gradient : No Critic**
- eNAC : Compatible Critic
- EM : MC Critic
- Summary, discussion

Direct policy gradient [Kimura et al., 1997], [Baxter and Bartlett, 2001]

Gradient of the objective function (also with J_γ^T).

$$\nabla_{\mathbf{w}} \bar{J} = r(x) \frac{\nabla_{\mathbf{w}} \mu(x; \pi, w)}{\mu(x; \pi, w)} \quad (5)$$

Unbiased estimation if regenerative process/episode

- Each step : $z_{t+1} = z_t + \frac{\nabla_{\mathbf{w}} \pi(x_t, u_t, \mathbf{w})}{\pi(x_t, u_t, \mathbf{w})}$, $v_{t+1} = v_t + r_t$
- End of episode : $\Delta_{j+1} = \Delta_j + v_t z_t / \mathrm{length}$
- return Δ_N / N

Biased on-line estimate

- Each step : $z_{t+1} = \beta z_t + \frac{\nabla_{\mathbf{w}} \pi(x_t, u_t, \mathbf{w})}{\pi(x_t, u_t, \mathbf{w})}$, $w_{t+1} = w_t + \alpha_t . r_t . z_t$
- return w_T

Why a biased estimate ? $(\boldsymbol{\mu}^\top \nabla_{\mathbf{w}}(J_\beta^\infty) \neq \nabla_{\mathbf{w}}(\boldsymbol{\mu}^\top J_\beta^\infty)$

$$\nabla_{\mathbf{w}} \bar{J} = (1-\beta)\nabla_{\mathbf{w}}(\boldsymbol{\mu}^\top) J_\beta^\infty + \beta \boldsymbol{\mu}^\top \nabla_{\mathbf{w}}(\mathbf{P}) J_\beta^\infty \tag{6}$$

Left term vanishes

$$\lim_{\beta \to 1} (1-\beta)\nabla_{\mathbf{w}}(\boldsymbol{\mu}^\top) J_\beta^\infty = 0 \tag{7}$$

but right term can have large variance when $\beta \to 1$.

Convergence of estimate to

$$(1-\beta)\boldsymbol{\mu}^\top \nabla_{\mathbf{w}}(J_\beta^\infty) = \beta \boldsymbol{\mu}^\top \nabla_{\mathbf{w}}(\mathbf{P}) J_\beta^\infty \tag{8}$$

No critic

- direct estimation of gradient
- simple algorithms
- application to POMDPs, NN versions.

- lots of samples
- prone to local optimum

Outline

- Context : RL and direct policy search
- Direct Gradient : No Critic
- **eNAC : Compatible Critic**
- EM : MC Critic
- Summary, discussion

eNAC [Peters and Schaal, 2008]

Using baseline function for the gradient

$$\nabla_{\mathbf{w}} \bar{J}(\mathbf{w}) = \mathbb{E}_{x\sim\mu,u\sim\pi}\left[\nabla_{\mathbf{w}} \log \pi(x,u;w)[Q^{\pi}(x,u) - b(x)]\right] \qquad (9)$$

Would like to use **Natural Gradient** $\tilde{\nabla}\bar{J}$

- sure convergence to local optimum
- fastest convergence and can sometimes prevent premature convergence
- independant of parametrization of the policy
- less samples to be correctly evaluated

$$\tilde{\nabla}\bar{J} = \underbrace{\left(\mathbb{E}_{x\sim\mu,u\sim\pi}\left[\nabla_{\mathbf{w}} \log \pi(x,u) \nabla_{\mathbf{w}} \log \pi(x,u)^{\top}\right]\right)}_{\text{Fisher's Information Matrix}}^{-1} \nabla_{\mathbf{w}} \bar{J} \qquad (10)$$

"Easy" computation of natural gradient

Compatible Q-value approximation : $\nabla_{\boldsymbol{\theta}} \hat{Q}^{\pi}_{\boldsymbol{\theta}}(x,u) = \nabla_{\mathbf{w}} \log \pi(x,u)$
(for example $\hat{Q}^{\pi}_{\boldsymbol{\theta}}(x,u) = \boldsymbol{\theta}^{\top} \nabla_{\mathbf{w}} \log \pi(x,u)$).

Then, the best solution $\boldsymbol{\theta}^*$ for the advantage function

$$\boldsymbol{\theta}^* = \operatorname*{argmin}_{\boldsymbol{\theta}} \mathbb{E}_{x\sim\mu,u\sim\pi}\left[Q^{\pi}(x,u) - V^{\pi}(x) - \hat{Q}^{\pi}_{\boldsymbol{\theta}}(x,u)\right]^2 \qquad (11)$$

is the natural gradient : $\nabla_{\mathbf{w}} \bar{J}(\mathbf{w}) = \boldsymbol{\theta}^*$

($\boldsymbol{\theta}^*$ independant of $b(.)$, $V^{\pi}(x)$ minimizes variance once $\boldsymbol{\theta}*$ found)

Episodic estimation of the advantage function

Advantage function $A^\pi(x,u) + V^\pi(x) = r(x,u) + \gamma\mathbb{E}[V^\pi(x)]$
Then, along one trajectory

$$\sum_{t=0}^{N-1} \gamma^t A^\pi(x_t,u_t) = -V^\pi(x_0) + \sum_{t=0}^{N-1} \gamma^t r(x_t,u_t) + \gamma_N V^\pi(x_N)$$

So, if enough trajectories (compared to size of $\boldsymbol{\theta}$)

$$\sum_{t=0}^{N-1} \gamma_t \nabla_w \log \pi(x_t,u_t)^\top \boldsymbol{\theta} + V_0 = \sum_{t=0}^{N-1} \gamma^t r(x_t,u_t)$$

is a “simple” regression problem. (See Matthieu G.)

$$w_{n+1} = w_n + \alpha.\boldsymbol{\theta}^* \quad (12)$$

eNAC : Compatible Critic

- must use a special critic
- estimation of $V \rightsquigarrow$ LSTD(λ), ...
- successes in robotic movements

- good choice of learning parameters and basis function
- ((we had difficulties to get it working))

Outline

- Context : RL and direct policy search
- Direct Gradient : No Critic
- eNAC : Compatible Critic
- **EM : MC Critic**
- Summary, discussion

The EM approach [Kober and Peters, 2008], [Kormushev et al., 2010]

Estimation

$$\log \bar{J}(\mathbf{w}) = \log \int_\tau \frac{\mu(\tau; w)}{\mu(\tau; w)} \mu(\tau; w') R(\tau) d\tau \qquad (13)$$

$$\geq \int_\tau \mu(\tau; w) R(\tau) \log \frac{\mu(\tau; w')}{\mu(\tau; w)} d\tau + K \qquad (14)$$

$$\propto -D(\mu(\tau; w) R(\tau) || \mu(\tau; w')) = l(w, w') \qquad (15)$$

$\nabla_{w'} l(w, w') = \mathbb{E}_w \left[\sum_{t=1}^{T} \nabla_{w'} \pi(x, u; w') Q^\pi(x, u) \right]$

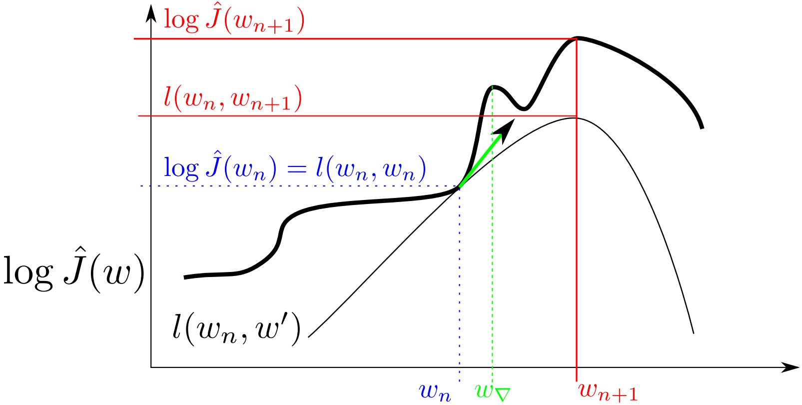
Maximization

- Analytical solution to $\mathrm{argmax}_{w'} \nabla_{w'} l(w, w')$
- Policy with **exponential distribution** function
- *i.e.* $\pi(x, u; w) = \mathcal{N}(\mathbf{w}^\top \phi(x, u); \mathbf{\Sigma}(x) = [\epsilon]_{ij})$

$\rightsquigarrow$

$$w_{n+1} = w_n + \mathbb{E}\left[\sum_{t=1}^{T} \epsilon_t Q^\pi(x, u, t) \right] / \mathbb{E}\left[\sum_{t=1}^{T} Q^\pi(x, u, t) \right]$$

Illustration



EM

- no learning coefficient
- should be able to avoid some local maxima
- should use less samples

- no tried
- succes also because of "Movement Dynamic Primitive" ? [Schaal et al., 2007]

Outline

- Context : RL and direct policy search
- Direct Gradient : No Critic
- eNAC : Compatible Critic
- EM : MC Critic
- **Summary, discussion**

Conclusion

- Nothing original, just review of works
- Quite technical (sorry)

- kinethetic teaching
- parameters adaptation around shown example

- EM
- ... or Dynamic Movement Primitives [Schaal et al., 2007] ?
- ... or Importance sampling ?

Your turn :o)

Références I

Baxter, J. and Bartlett, P. (2001).
Infinite-horizon policy-gradient estimation.
Journal of Artificial Intelligence Research, 15 :319–350.

Groupe PDMIA (2008).
Processus Décisionnels de Markov en Intelligence Artificielle. (Edité par Olivier Buffet et Olivier Sigaud), volume 1 & 2.
Lavoisier - Hermes Science Publications.

Kimura, H., Miyazaki, K., and Kobayashi, K. (1997).
Reinforcement learning in POMDPs with function approximation.
In *Proc. of the Fourteenth Int. Conf. on Machine Learning (ICML'97)*, pages 152–160.

Références II

Kober, J. and Peters, J. (2008).
Policy search for motor primitives in robotics.
In *Advances in Neural Information Processing Systems, NIPS'08*.

Kormushev, P., Calinon, S., and Caldwell, D. G. (2010).
Robot motor skill coordination with EM-based reinforcement learning.
In *Proc. IEEE/RSJ Intl Conf. on Intelligent Robots and Systems (IROS)*, pages 3232–3237.

Peters, J. and Schaal, S. (2008).
Natural actor-critic.
Neurocomputing, 71(7-9) :1180–1190.

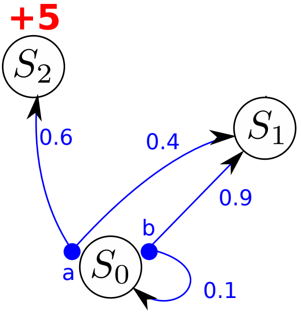
Références III

Puterman, M. (1994).
Markov Decision Processes : discrete stochastic dynamic programming.
John Wiley & Sons, Inc. New York, NY.

Schaal, S., Mohajerian, P., and Ijspeert, A. (2007).
Dynamics systems vs. optimal control–a unifying view.
Progress in brain research, 165 :425–445.

Sutton, R. and Barto, A. (1998).
Reinforcement Learning.
Bradford Book, MIT Press, Cambridge, MA.

Reinforcement Learning

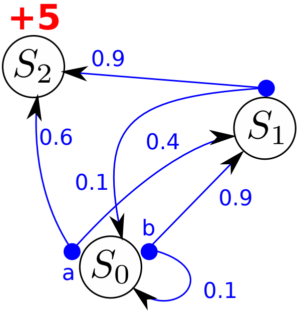


- States $\mathcal{X}$, Actions $\mathcal{U}$, probabilistic transitions.
- $E_{x,u\sim\pi}\left[\sum_{t=1}^{\infty}\gamma^t r_t | x_0 = x, u_0 = u\right]$
- Find the optimal policy π.
  $\pi : \mathcal{X} \longrightarrow \mathcal{U}$

$\rightsquigarrow$ Action 'a' or 'b' in S_0 ?

[Puterman, 1994], [Sutton and Barto, 1998], [Groupe PDMIA, 2008], ...

Reinforcement Learning

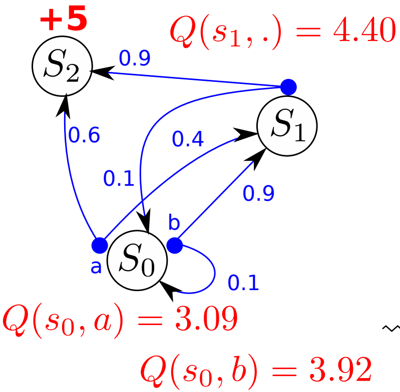


- States $\mathcal{X}$, Actions $\mathcal{U}$, probabilistic transitions.
- $E_{x,u\sim\pi}\left[\sum_{t=1}^{\infty}\gamma^t r_t | x_0 = x, u_0 = u\right]$
- Find the optimal policy π.
  $\pi : \mathcal{X} \longrightarrow \mathcal{U}$

$\rightsquigarrow$ Action 'a' or 'b' in S_0 ?

[Puterman, 1994], [Sutton and Barto, 1998], [Groupe PDMIA, 2008], ...

Reinforcement Learning

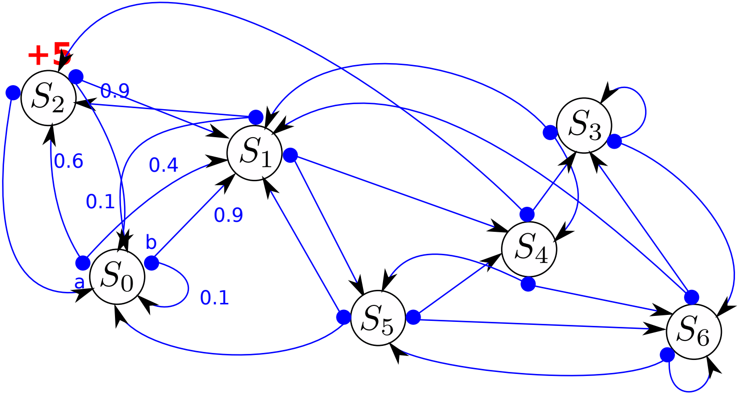


- States $\mathcal{X}$, Actions $\mathcal{U}$, probabilistic transitions.
- $E_{x,u\sim\pi}\left[\sum_{t=1}^{\infty} \gamma^t r_t | x_0 = x, u_0 = u\right]$
- Find the optimal policy π.
  $\pi : \mathcal{X} \longrightarrow \mathcal{U}$

$\rightsquigarrow$ Action 'a' or 'b' in S_0 ?

[Puterman, 1994], [Sutton and Barto, 1998], [Groupe PDMIA, 2008], ...

Reinforcement Learning



Compute directly the **optimal** value function (as a solution to) :

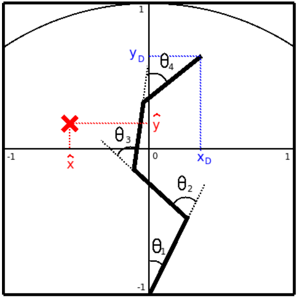
$$Q^*(x,u) \quad = \quad \mathbf{r(x,u)} + \gamma \sum_{x' \in \mathcal{X}} \mathbf{p(x'|u,x)} \max_{u' \in \mathcal{U}} [Q^*(x',u')]$$

[Puterman, 1994], [Sutton and Barto, 1998], [Groupe PDMIA, 2008], ...

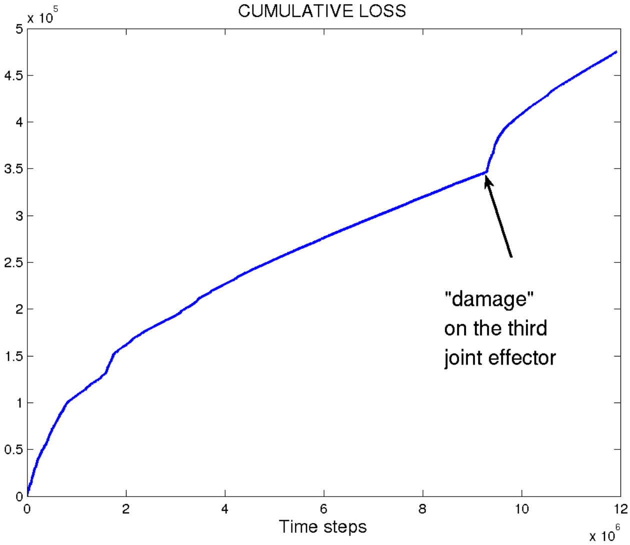
Neurocontroler with Gaussian Noise

No Critic with continuous state and action $\neq$ Kimura or Baxter.

- $u = \mathbf{w}^\top \mathbf{\Phi}(x) + \mathcal{N}(0, \Sigma)$
- $\mathbf{\Phi}$ of dimension 512
- $z_k = (1-\beta) z_k + \beta \frac{\nabla_{\mathbf{w}} \pi(x_k, u_k; \mathbf{w}_k)}{\pi(x_k, u_k; \mathbf{w}_k)}$
- $\mathbf{w}_{k+1} = \mathbf{w}_k + \alpha . r_k . z_k$



Cumulative cost



Motor response