“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
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 \( \times \) actions:

\[
\begin{align*}
x_0 & \xrightarrow{\pi} u_0 \Longrightarrow r_1, \\
x_1 & \xrightarrow{\pi} u_1 \Longrightarrow \cdots u_{T-1} \Longrightarrow r_T, x_T
\end{align*}
\]
Principles of “Direct Policy Search”

- Policy $\pi$ is **parametrized** (by $w$)
- 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.

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

Objective function

\[
J^{T}_\gamma(w) = \mathbb{E}_x \left[ V^{T}_\gamma(x) \right]
\] (2)

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

Linked:

\[
\bar{J}(w) = \mu^\top (I - \gamma P) J^\infty
\] (4)

with \( \mu \) static distribution and \( P \) transition probabilities.
Outline

- Context: RL and direct policy search
- Direct gradient: No Critic
- eNAC: Compatible Critic
- EM: MC Critic
- Summary, discussion
Direct policy gradient \cite{Kimura:1997,Baxter:2001}

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

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

**Unbiased** estimation if regenerative process/episode

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

**Biased** on-line estimate

- Each step: $z_{t+1} = \beta z_t + \frac{\nabla_w \pi(x_t, u_t, w)}{\pi(x_t, u_t, w)}$, $w_{t+1} = w_t + \alpha_t r_t z_t$
- return $w_T$
Why a biased estimate? \( (\mu^T \nabla_w (J^\infty_\beta)) \neq \nabla_w (\mu^T J^\infty_\beta) \)

\[
\nabla_w \bar{J} = (1 - \beta) \nabla_w (\mu^T) J^\infty_\beta + \beta \mu^T \nabla_w (P) J^\infty_\beta 
\]

Left term vanishes

\[
\lim_{\beta \to 1} (1 - \beta) \nabla_w (\mu^T) J^\infty_\beta = 0
\]

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

Convergence of estimate to

\[
(1 - \beta) \mu^T \nabla_w (J^\infty_\beta) = \beta \mu^T \nabla_w (P) J^\infty_\beta
\]
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

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

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

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

\[
\tilde{\nabla} \bar{J} = \left( \mathbb{E}_{x \sim \mu, u \sim \pi} \left[ \nabla_w \log \pi(x, u) \nabla_w \log \pi(x, u)^\top \right] \right)^{-1} \nabla_w \bar{J} \tag{10}
\]

Fisher’s Information Matrix
"Easy" computation of natural gradient

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

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

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

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

($\theta^*$ independent of $b(.)$, $V^\pi(x)$ minimizes variance once $\theta^*$ 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 $\theta$)

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

is a "simple" regression problem. (See Matthieu G.)

$$w_{n+1} = w_n + \alpha \theta^*$$ (12)
eNAC : Compatible Critic

- must use a special critic
- estimation of $V \rightsquigarrow \text{LSTD}(\lambda)$, ...
- 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 \cite{Kober2008, Kormushev2010}

Estimation

\[
\log \bar{J}(w) = \log \int \frac{\mu(\tau; w)}{\mu(\tau; w')} \mu(\tau; w') R(\tau) d\tau \\
\geq \int \mu(\tau; w) R(\tau) \log \frac{\mu(\tau; w')}{\mu(\tau; w)} d\tau + K \tag{14}
\]

\[
\propto -D(\mu(\tau; w) R(\tau) \| \mu(\tau; w')) = l(w, w') \tag{15}
\]

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

Maximization

\begin{itemize}
  \item Analytical solution to \( \arg\max_{w'} \nabla_{w'} l(w, w') \)
  \item Policy with \textbf{exponential distribution} function
  \item \( i.e. \ \pi(x, u; w) = \mathcal{N}(w^\top \phi(x, u); \Sigma(x) = [\epsilon]) \)
\end{itemize}

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

\[ \log \hat{J}(w_{n+1}) \]

\[ l(w_n, w_{n+1}) \]

\[ \log \hat{J}(w_n) = l(w_n, w_n) \]

\[ l(w_n, w') \]

\[ w_n \quad w' \quad w_{n+1} \]
▶ 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)

- kinesthetic teaching
- parameters adaptation around shown example

- EM
- ... or Dynamic Movement Primitives [Schaal et al., 2007]?
- ... or Importance sampling?
Your turn :o)
Références I

Infinite-horizon policy-gradient estimation.

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

Reinforcement learning in POMDPs with function approximation.
In Proc. of the Fourteenth Int. Conf. on Machine Learning (ICML '97), pages 152–160.


Références III

*Markov Decision Processes : discrete stochastic dynamic programming*.  
John Wiley & Sons, Inc. New York, NY.

Dynamics systems vs. optimal control–a unifying view.  

*Reinforcement Learning*.  
Reinforcement Learning

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

$\leadsto$ 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.
- $E_{x,u\sim\pi} \left[ \sum_{t=1}^{\infty} \gamma^t r_t \middle| x_0 = x, u_0 = u \right]$ 
- Find the optimal policy $\pi$. 
  $\pi : \mathcal{X} \rightarrow \mathcal{U}$

$\Rightarrow$ 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.

$$Q(s_1, .) = 4.40$$

$$Q(s_0, a) = 3.09$$

$$Q(s_0, b) = 3.92$$

$\sum_{t=1}^{\infty} \gamma^t r_t | x_0 = x, u_0 = u$}

Find the optimal policy $\pi$.

$\pi : \mathcal{X} \longrightarrow \mathcal{U}$

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):

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

[Puterman, 1994], [Sutton and Barto, 1998], [Groupe PDMIA, 2008], ...
No Critic with continuous state and action $\neq$ Kimura or Baxter.

- $u = w^T \Phi(x) + \mathcal{N}(0, \Sigma)$
- $\Phi$ of dimension 512
- $z_k = (1 - \beta)z_k + \beta \frac{\nabla_w \pi(x_k, u_k; w_k)}{\pi(x_k, u_k; w_k)}$
- $w_{k+1} = w_k + \alpha r_k z_k$
Cumulative cost

"damage" on the third joint effector
Motor response

(a)

(b)

(c)

(d)