Expérimentations et Réflexions autour de l'Apprentissage par Renforcement Développemental

Sém. Apprentissage Développemental, Lyon.

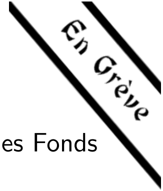
Alain Dutech

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

16 déc 2014

En Grève

En grève !? Mais pourquoi ??



- Le Ministère ne finance pas assez les Universités

⇝ LRU + RCE ⇒ A la charge de l'Université de trouver des Fonds

- UL (comme ailleurs) : Le Président n'assume pas :o(

⇝ Obei au Ministère et Management Autocratique

⇝ "Fait redescendre" sur pôles et collégium

⇝ Gel des postes, Refus solution alternative

⇝ "Fusion" ou Mutualisation "forcée"

⇝ (à terme ... droits d'inscriptions)

Difficilement mais sûrement, lutte et solidarité se mettent en place.

Outline

En Grève

3

Experiment

- **Context : RL, Developmental approach**
- Learning Architectures
- Robotic experiments

Discussion

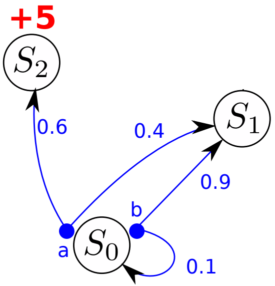
- What is “Developmental” ?
- Problems and Questions

Reinforcement Learning

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

En Grève

4



- States $\mathcal{S}$, Actions $\mathcal{A}$, probabilistic transitions.
- $E_{s,a\sim\pi}\left[\sum_{t=1}^{\infty}\gamma^t r_t | s_0 = s, a_0 = a\right]$
- Find the optimal policy π.

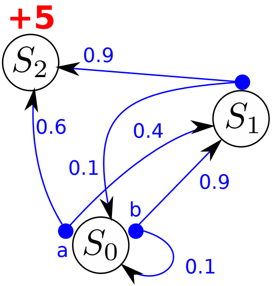
$\rightsquigarrow$ Action a or b in S_0 ?

Reinforcement Learning

En Grève

4

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

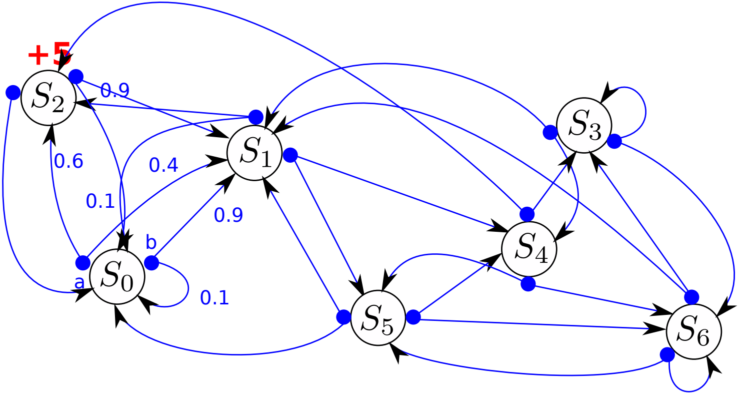


- States $\mathcal{S}$, Actions $\mathcal{A}$, probabilistic transitions.
- $E_{s,a\sim\pi}\left[\sum_{t=1}^{\infty}\gamma^t r_t | s_0 = s, a_0 = a\right]$
- Find the optimal policy π.

$\rightsquigarrow$ Action a or b in S_0 ?

Reinforcement Learning

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



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

$$Q^*(s,a) \quad = \quad \mathbf{r(s,a)} + \gamma \sum_{s' \in \mathcal{S}} \mathbf{p(s'|a,s)} \max_{a' \in \mathcal{A}} [Q^*(s',a')]$$

Q-Learning [Watkins, 1989]

En Grève

5

States $\mathcal{S}$, Actions $\mathcal{A}$, probabilistic transitions.

Stochastic approximation of the Q-values of an optimal policy.

1. For a given state s
2. Choose and apply exploratory action a
3. Environment gives back new state s' and reward r
4. Update

$$Q(s,a) \quad \longleftarrow \quad Q(s,a) + \alpha[r + \gamma \max_{a' \in \mathcal{A}} Q(s',a') - Q(s,a)]$$

5. Goto (1) with $s \longleftarrow s'$

RL problems with huge state × action space

Typical with robots for example.

$$Q(s,a) \longleftarrow Q(s,a) + \alpha[r + \gamma \max_{a' \in \mathcal{A}} Q(s', a') - Q(s,a)] \quad ⑥$$

1. Continuous environment. RL easier with discrete states *and actions*.
   $\rightsquigarrow$ approximation
2. Costly experiences. Time, energy to try (s, a) vs algorithms with huge itearation needs.
   $\rightsquigarrow$ Re-use.
3. Harmfull experiences. Robot destruction ?
   $\rightsquigarrow$ Special low-level behavior.
4. Sparse Reward. Non-zero reward is difficult to get.
   $\rightsquigarrow$ Eligibility traces.
5. Rich environment. Too many “states” to consider.
   $\rightsquigarrow$ Factorization, aggregation
6. Partial observability. No guarantee for learning.
   $\rightsquigarrow$ (State-extension to an MDP), incremental.

En Grève

RL problems with huge state $\times$ action space

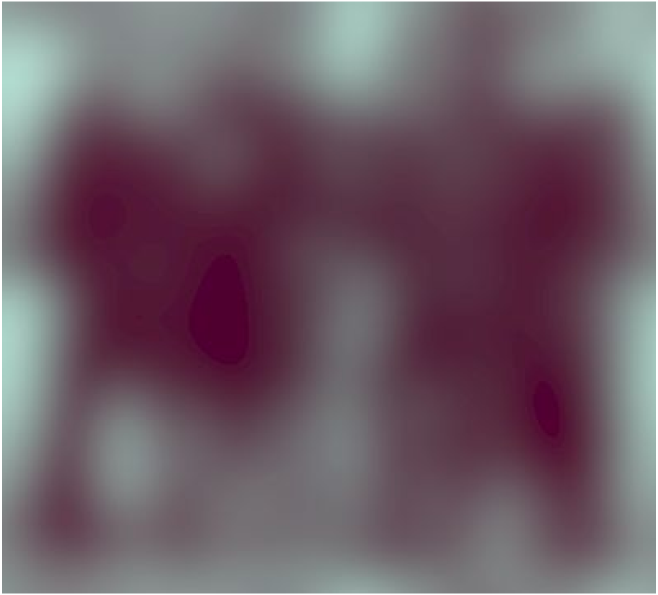
Typical with robots for example.

$$Q(s,a) \longleftarrow Q(s,a) + \alpha[r + \gamma \max_{a' \in \mathcal{A}} Q(s',a') - Q(s,a)]$$

6

1. **Continuous environment**. RL easier with discrete states *and actions*.
   $\rightsquigarrow$ approximation
2. Costly experiences. Time, energy to try (s,a) vs algorithms with huge itearation needs.
   $\rightsquigarrow$ Re-use.
3. Harmfull experiences. Robot destruction ?
   $\rightsquigarrow$ Special low-level behavior.
4. **Sparse Reward**. Non-zero reward is difficult to get.
   $\rightsquigarrow$ Eligibility traces.
5. **Rich environment**. Too many “states” to consider.
   $\rightsquigarrow$ Factorization, aggregation
6. Partial observability. No guarantee for learning.
   $\rightsquigarrow$ (State-extension to an MDP), incremental.

A developmental point of view



A developmental point of view

En Grève

A developmental point of view

7

Developmental Reinforcement Learning

$\rightsquigarrow$ **Ease the exploration of state $\times$ action space**

En Grève

Concept

Robot's task, perceptual and motor skills **increase** when learned behavior becomes **more efficient**.
(Only one aspect of Developmental Robotic, see [Lungarella et al., 2003]).

Developmental Reinforcement Learning

$\rightsquigarrow$ **Ease the exploration of state $\times$ action space**

En Grève

8

Concept

Robot's task, perceptual and motor skills **increase** when learned behavior becomes **more efficient**.
(Only one aspect of Developmental Robotic, see [Lungarella et al., 2003]).

New problems arise:

- (Learn approximation of $Q(s, a)$).
- How to deal with increased number of actions ?
- How to deal with increased number of states ?
- How to changes goals, sometimes drastically ?
- How to "transfer" learned behavior to more complex tasks ?

Outline

Experiment

- Context : RL, Developmental approach
- **Learning Architecture**
- Robotic experiments

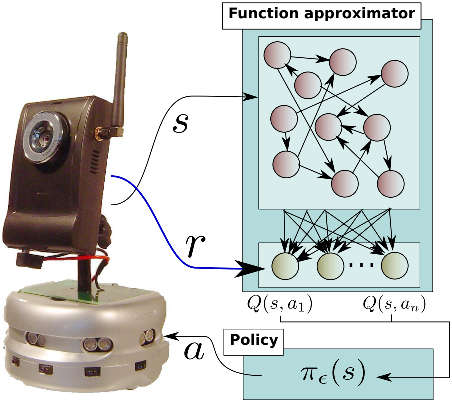
Discussion

- What is "Developmental" ?
- Problems and Questions

Function approximation

Dynamic Self-Organizing Map and a linear readout

En Grève



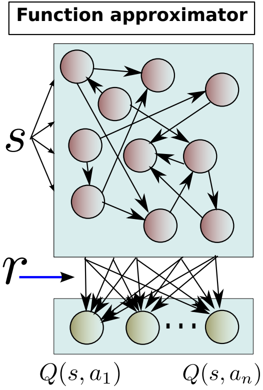
"Supervised learning" using Bellman error:
$\Delta Q_{\mathrm{NET}}(s,a) = \alpha[r + \gamma \max_{a'} Q_{\mathrm{NET}}(s',a') - Q_{\mathrm{NET}}(s,a)]$

Dynamic Self-Organizing Maps (DSOM)

[Rougier and Boniface, 2011]

En Grève

11



- Present input v
- Winner neuron w_w

$$s \longleftarrow \underset{i \in \mathcal{N}}{\operatorname{argmin}}(||v - w_i||)$$

- Learn "lateral" connexions towards v

$$\delta w_i \longleftarrow \epsilon_D ||v - w_i|| h_\eta(i, w_s, v)(v - w_i)$$

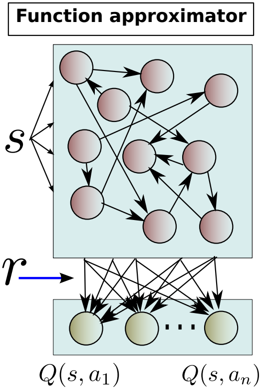
- Readout

$$\mathrm{out}_i \longleftarrow \sum \omega_i . w_i$$

- Learning readout weights

$$\delta \omega_i \longleftarrow -\epsilon_L . \mathrm{error} . w_i$$

Grow sensori-motor space



Increase input dimension

- Start with *n* input neurons (where *n* is the maximum input dimension)
- At start, some input are cloned.
- Then, inputs are discriminated

Increase nb of actions

- Add an output neuron
- Init weights
  $\rightsquigarrow$ random, copy

Outline

Experiment

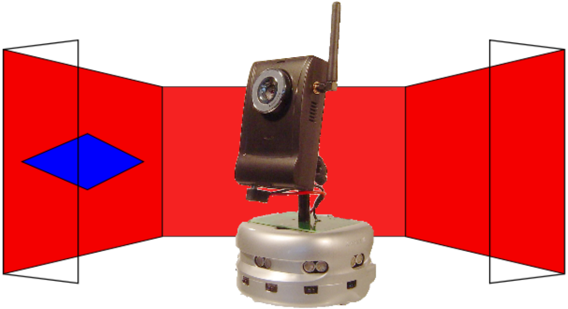
- Context : RL, Developmental approach
- Learning Architectures
- **Robotic experiments**

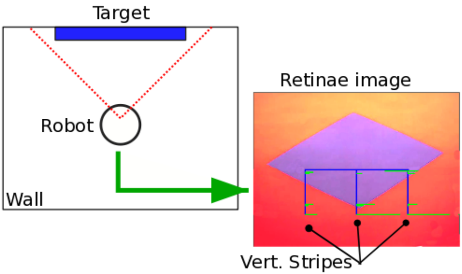
Discussion

- What is “Developmental” ?
- Problems and Questions

Robotic setting

En Grève

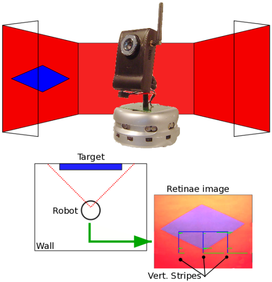




En Grève

Task details

15



- One Khepera3 robot
- **Continuous** perception
  $\rightsquigarrow$ camera and blue stripes sensor
- Actions are **discrete**
  - 3 actions: Stop, Left, Right
  - 5 actions: Stop, Left, Right, slowLeft, slowRight
- Reward linked to the blue levels in sensor
  ex: $0.7B1 + B2 + 0.7B3 \geq 1.2$
- Real 2.000 samples (used and re-used) or Simulator

Learning Scenarii

Learning

- Generate (ϵ_p) or use transition $(s, a : s', r)$
- unsupervised DSOM : ϵ_D, elasticity η
- supervised LIN : ϵ_L, α

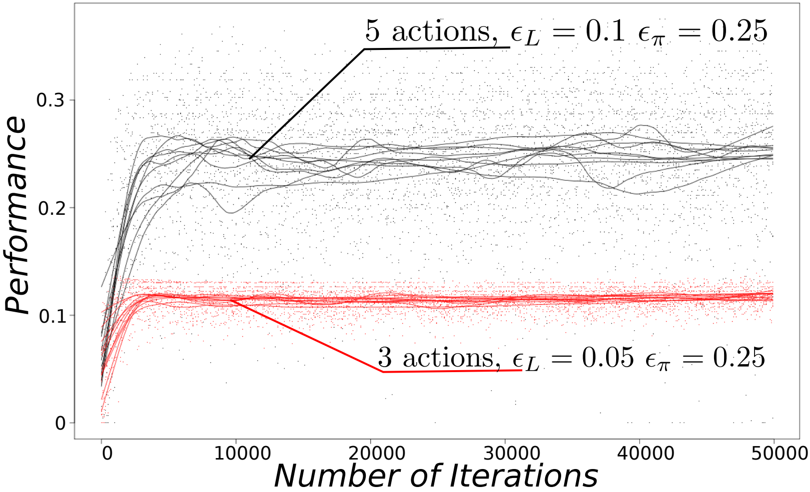
Evaluation

- Periodically
- No learning
- Mean reward : $\frac{1}{N}\sum_{t=1}^{N} r_t$

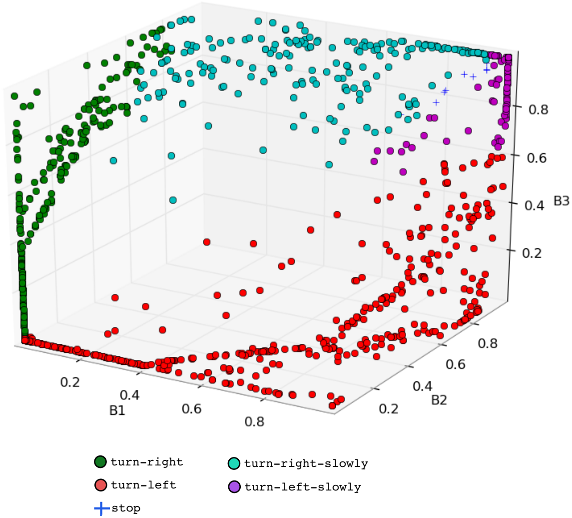
Visual evaluation of policy

Better performances with 5 actions

En Grève

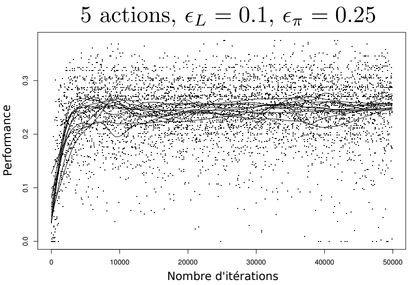


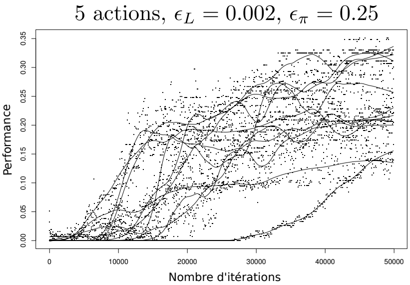
Learned policy, 5 actions



Variability vs. Speed

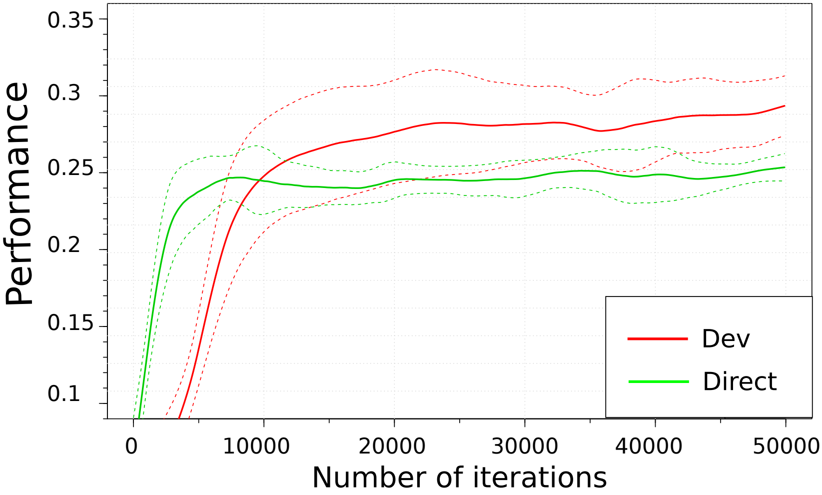
En Grève





Developmental learning, 3 then 5 actions

En Grève



Direct: 5 actions / DevRL: 2 actions added $N_B = 5000$ (ϵ_L 0.1 to 0.01)

Outline

En Grève

Experiment

- Context : RL, Developmental approach
- Learning Architectures
- Robotic experiments

Discussion

- **What is “Developmental” ?**
- Problems and Questions

What is “Developmental” ?

En Grève

Autonomous All-Life-Long Learning Agent

$\rightsquigarrow$ Interactions Body-Brain-Environment
$\rightsquigarrow$ Coupling Development-Learning

- Agent Development

- No “exogen” intervention

- Build on Previous Acquired “Behaviors” :

What is "Developmental" ?

En Grève

22

Autonomous All-Life-Long Learning Agent

$\rightsquigarrow$ Interactions Body-Brain-Environment
$\rightsquigarrow$ Coupling Development-Learning

- Agent Development
  - "Only" actions, (perceptions)
  - Body !!
- No "exogen" intervention
  - Intrinsic Motivations vs "external" reward signal !
- Build on Previous Acquired "Behaviors" :
  - What starting substrate ?
  - "Knowledge" Transfer

Facets of Developmental Robotics [Lungarella et al., 2003]

1. Development is an incremental process
2. Development as a set of constraints
3. Development as a self-organizing process
4. Degrees of freedom and motor activity
5. Self-exploratory activity
6. Spontaneous activity
7. Anticipatory movements and early abilities
8. Categorization and sensorimotor co-ordination
9. Neuromodulation, values and neural plasticity
10. Social interaction

En Grève

Outline

Experiment

- Context : RL, Developmental approach
- Learning Architectures
- Robotic experiments

Discussion

- What is “Developmental” ?
- **Problems and Questions**

Problems and Question

En Grève

25

- Body !!
  - Simulated vs Real
  - Richness (Sensors, DOF), but very poor compared to “us”
  - “Growing” Maturation
- Architecture, Learning MechanismS
- “Sum” of Individually Tested Mechanisms ??
- Multiple Time Scales : individual and generations
  - How to care and interact with agents during all that time?
- Changing Motivations
  - How to create new motivations ?
  - Priority of motivations ?

Main Question

“Just” a question of “right Integration” or Still Lack Brook’s “Juice” [**?**, **?**] ?

En Grève

Focussing on the Experiment

- Limits of Reinforcement Learning in “low level” Robotics
- Original(s) architecture for Developmental approach
  - Dynamic Self-Organizing Map : online life-long learning
  - Linear regression readout
  - Compatible with growing actions (and perceptions)
- NEED MORE RESULTS ... (especially for Perceptions)
- Memorize (replay) useful transitions, sequence of tasks...

Short-term and long-term work

En Grève

27

- scenario for Perception growing.
- parameters influence, especially on stability
- more complexe setting $\rightsquigarrow$ different tasks to learn
- select and memorize useful transitions
- recurrent DSOM (???) for sequence learning in non-Markov (??)

$\rightsquigarrow$ PhD of Matthieu ZIMMER...

En Grève

A vous :o)

Références I

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

Lungarella, M., Metta, G., Pfeifer, R., and Sandini, G. (2003).
Developmental robotics: a survey.
Connection Science, 15(4):151–190.

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

Rougier, N. P. and Boniface, Y. (2011).
Dynamic Self-Organising Map.
Neurocomputing, 74(11):1840–1847.

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

Références II

En Grève

30

Watkins, C. (1989).
Learning from delayed rewards.
PhD thesis, King's College of Cambridge, UK.