RL 18: Self-Motivated Reinforcement Learning

Michael Herrmann

University of Edinburgh, School of Informatics

20/03/2015

Intrinsic Rewards

Questions

- How do we define a reward function?
- Where do rewards come from?
- Intrinsic or extrinsic rewards?
- Can an agent "learn" without rewards? What could it possibly learn?
- What actions are worth being explored?
- How can exploration be organised beyond purely random behaviour?

Potential answers

From earlier lectures:

- Actions with high-variance reward estimates require more exploration (MAB)
- Boltzmann exploration
- Model-based approaches (Dyna-$\mathcal{Q}$)
- Options (SMDPs)
- Policy search
- Multi-objective learning

Why is exploration a problem?

- Not if the number of states and actions is small and time horizon is short
- Exhaustive exploration may be impossible
- Frontier-based exploration becomes impractical in higher dimensions
- Reward signals may not reveal problem structure
- Early success may be misleading

Intrinsic motivation

Intrinsic motivation is defined as the doing of an activity for its inherent satisfaction rather than for some separable consequence. When intrinsically motivated, a person is moved to act for the fun or challenge entailed rather than because of external products, pressures, or rewards.

Ryan R. M., Deci E. L. (2000). Intrinsic and extrinsic motivations: Classic definitions and new directions. *Contemp. Educ. Psychol.* **25**, 54–67.

Intrinsic Motivation: Evolutionary Perspective

- Define a $\mathcal{Q}$-learning agents **A** with reward functions $r_{\mathbf{A}}$
- Do forever
  - set learning rate η and exploration rate ε
  - for $i = 1$ to N do
    - Generate a sample E_i from the environment $\mathcal{E}$
    - initialise $\mathcal{Q}$-function
    - Generate a history h_i over lifetime of the agent
    - Compute fitness $F(h_i)$
  - return average $\langle F(h_i) \rangle_i$
  - Select and reduplicate high fitness agents and modify $r_{\mathbf{A}}$

S. Singh, R.L. Lewis, A.G. Barto, and J. Sorg (2010) Intrinsically Motivated Reinforcement Learning: An Evolutionary Perspective. IEEE Transactions on Autonomous Mental Development **2**:2, 70-82.

What is interesting?

- "Interestingness" as a complement to utility could help shaping exploration strategies
- Agent could develop a sense of "curiosity", e.g.
  - counter-based: states that have not been visited
  - information-based, using a novelty detector
  - homeostatic
- or could observe secondary qualities of the learning process in a form of introspection, e.g.
  - learning time or slope of total reward average
  - robustness and generality
- Shape/evolve rewards signals as well as exploration strategies

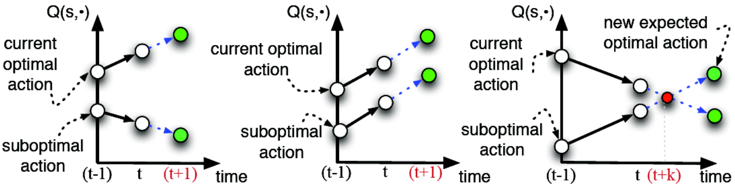
Intrinsic motivation (Barto, NIPS)

- "Sutton & Barto point out that one should not identify this RL agent with an entire animal or robot." (Barto, NIPS)
- External environment and internal environment
- A sophisticated system that should not have to be redesigned for different problem
- Learning a collection of reusable skills in order to generate a skill knowledge base
- Skills could be options (an option is not a sequence of actions; it is a closed-loop control rule, meaning that it is responsive to on-going state changes)

Barto

Chentanez, N., Barto, A. G., & Singh, S. P. (2004). Intrinsically motivated reinforcement learning. In Advances in neural information processing systems (pp. 1281-1288).

A Model-Free Algorithm for Efficient Exploration



Determine time until policy change

$$d\left(s,a\right)=\frac{1}{\alpha_M}\,\frac{\mathcal{Q}_t\left(s,a^*\right)-\mathcal{Q}_t\left(s,a\right)}{\delta_{s,a}\left(T_{s,a}\right)-\delta_{s,a^*}\left(T_{s,a^*}\right)}$$

$\delta_{s,a}\left(T_{s,a}\right)$ is the δ error for the last time (s,a) was updated.

α_M is the estimated slope of the expected reward

A Model-Free Algorithm for Efficient Exploration

Reward based on predicted future usefulness of an action

$$\tilde{r}(s,a) = \begin{cases} \exp\left(-\frac{d^2(s,a)}{\sigma}\right) & \text{if } |d(s,a) < \lambda| \\ -p & \text{otherwise} \end{cases}$$

$-p$ is a small penalty for stabilisation

d is the expected time until a policy change will occur (only in the image on the right we have $d < \infty$)

σ and λ define a prediction horizon

Da Silva, B. C., & Barto, A. G. (2012) TD-$\Delta\pi$: A Model-Free Algorithm for Efficient Exploration. 26th Conf. on Artificial Intelligence (AAAI-2012), Toronto, Ontario.

Da Silva, B. C., & Barto, A. G. (2012) TD-$\Delta\pi$: A Model-Free Algorithm for Efficient Exploration. 26th Conf. on Artificial Intelligence (AAAI-2012), Toronto, Ontario.

A Model-Free Algorithm for Efficient Exploration

For all (s, a): $\mathcal{Q}^0_{\text{exploit}}(s, a) \leftarrow 0$, $\mathcal{Q}^0_{\text{explore}}(s, a) \leftarrow 0$,
$\delta_{s,a}(0) \leftarrow 0$, $T_{s,a} \leftarrow 0$, visited $(s, a) \leftarrow$ False
For $t = 1, 2, \dots$ do
Choose action $a_t = \arg\max_b \mathcal{Q}^t_{\text{explore}}(s, b)$
observe reward r_t and next state, s_{t+1}
Choose action $a^*_t = \arg\max_b \mathcal{Q}^t_{\text{exploit}}(s, b)$
if not visited (s_t, a_t) or not visited (s_t, a^*_t) then $r(s_t, a_t) = 1$
else if $\left|\delta_{(s_t,a_t)}(T_{s_t,a_t}) - \delta_{(s_t,a^*_t)}\left(T_{s_t,a^*_t}\right)\right| < \kappa$ then $r(s_t, a_t) = -p$
else $r(s_t, a_t) = \tilde{r}(s_t, a_t)$ (see previous slide)
$\mathcal{Q}^{t+1}_{\text{exploit}}(s, a) \leftarrow L\left(s_t, a_t, r^M_t, s_{t+1}\mathcal{Q}^{t+1}_{\text{exploit}}(s, a), \rho_{\text{exploit}}\right)$
$\mathcal{Q}^{t+1}_{\text{explore}}(s, a) \leftarrow L\left(s_t, a_t, r_t, s_{t+1}\mathcal{Q}^{t+1}_{\text{explore}}(s, a), \rho_{\text{explore}}\right)$
$T_{s_t,a_t} \leftarrow t$, visited $(s_t, a_t) \leftarrow$ True
$\delta_{s_t,a_t}(t) \leftarrow \mathcal{Q}^{t+1}(s_t, a_t) - \mathcal{Q}^t(s_t, a_t)$

Da Silva, B. C., & Barto, A. G. (2012) TD-$\Delta\pi$: A Model-Free Algorithm for Efficient Exploration. 26th Conf. on Artificial Intelligence (AAAI-2012), Toronto, Ontario.

A Model-Free Algorithm for Efficient Exploration

Discussion

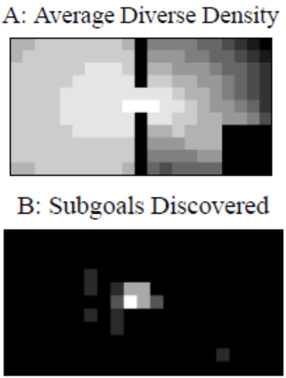
- Agents needs to be free to explore
- Restricted to discrete state and action spaces
- Performs poorly if many crossing are expected
- Linear approximation questionable as reward often saturates exponentially
- Smoothing and function approximation will be useful

Da Silva, B. C., & Barto, A. G. (2012) TD-$\Delta\pi$: A Model-Free Algorithm for Efficient Exploration. 26th Conf. on Artificial Intelligence (AAAI-2012), Toronto, Ontario.

Automatic Discovery of Subgoals

Algorithm

Init full trajectory database to $\emptyset$
For each trial
Interact with environment, learn using RL
Add observed full trajectory to database
Create pos. or neg. bag from state traj.
Search for diverse density peaks
For each peak concept c found
Update the running av. by $\bar{c} = \lambda(\bar{c} + 1)$
If $\bar{c}$ is above threshold
If c passes the static filter
Create a new option $o = \langle I, \pi, \beta \rangle$ for reaching concept c
Init I by examining trajectory database
Set $\beta(c) = 1$, $\beta(S - I) = 1$, $\beta(\cdot) = 0$ else
Init policy π using experience replay



McGovern, A., & Barto, A. G. (2001). Automatic discovery of subgoals in reinforcement learning using diverse density. Computer Science Dept. Faculty Publ. Series 8.

Early developmental AI

"Every time we teach a child something, we keep him from inventing it himself." (Piaget)

"An AI system can create and maintain knowledge only to the extent that it can verify that knowledge itself." (Sutton)

A. Turing ("Computing Machinery and Intelligence", 1950) "reckoned that it would be easier to write a program to simulate an infant's mind, rather than an adult's. The infant program could then be educated much like a human child, until it reached an adult level."

"The challenge here is to find a learning program which can continuously build on what it knows, to reach increasingly sophisticated levels of knowledge."

F. Guerin (2011) Learning Like Baby: A Survey of AI approaches. *The Knowledge Engineering Review* 26:02, 209-236.

Artificial Curiosity

- Additional rewards from the desire to improve the world model.
- Dynamic Curiosity and Boredom (Schmidhuber, 1991)
- Positive reward if the internal model fails to correctly predict the environment
- e.g. given a predictive model $M(x_t) = \hat{x}_{t+1}$ we can define intrinsic reward $r^{(2)} = 1$ if $|x_{t+1} - \hat{x}_{t+1}| > \vartheta$ and $r^{(2)} = 0$ otherwise, in addition to an extrinsic rewards signal $r^{(1)}$.
- Model is adapted in order to reduce prediction error while action are rewarded for having produced large prediction errors.

J. Schmidhuber (1991) A possibility for implementing curiosity and boredom in model-building neural controllers. In From Animals to Animats, 222–227, MIT Press.

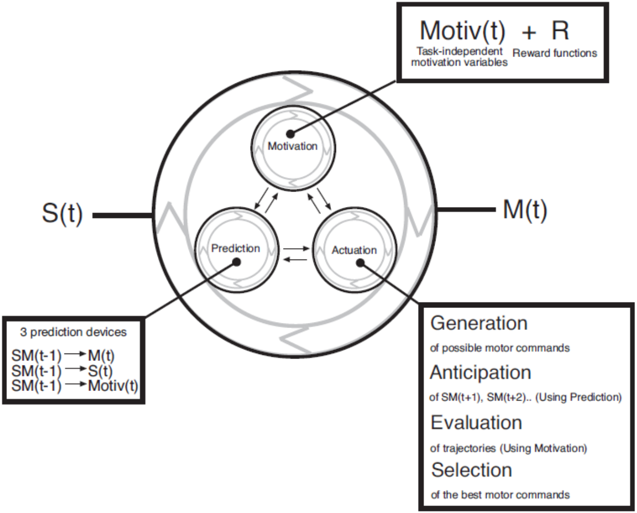
J. Schmidhuber, (2010). Formal theory of creativity, fun, and intrinsic motivation (1990–2010). IEEE Transact. Autonomous Mental Development 2(3), 230-247.

How can we define intrinsic motivation?

1. Knowledge based models
    - Comparisons between the predicted flow of sensorimotor values, (internal forward model) with the actual flow of values
    - Adaptive motivation: refers to mechanisms that assign different levels of interest to the same situation
2. Competence based models
    - Characterise the degree of performance/competence
    - Comparisons between self-generated goals and the extent to which they are reached in practice (internal inverse model)
    - Adaptive motivation
3. Morphological models
    - Measure immediate structural relationships among multiple sensorimotor channels
    - Fixed motivation

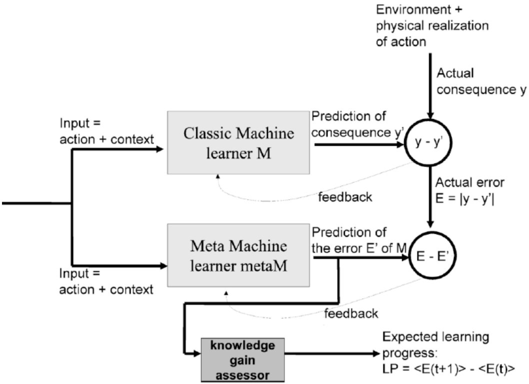
Oudeyer, P. Y., & Kaplan, F. (2008). How can we define intrinsic motivation?. Proc. 8th Int. Conf. on Epigenetic Robotics: Modeling cognitive development in robotic systems. Lund Univ. cognitive studies.

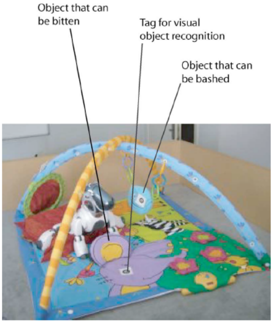
How can we define intrinsic motivation ?



Kaplan, F., & Oudeyer, P. Y. (2003). Motivational principles for visual know-how development. In C. G. Prince et al. 3. Int. Worksh. Epigen. Robotics, 73–80, Edinburgh, Scotland, Lund Univ. Cogn. Studies.

The playground experiment





Oudeyer, P. Y., Kaplan, F., & Hafner, V. V. (2007). Intrinsic motivation systems for autonomous mental development. IEEE Transact. Evol. Comput. **11**(2), 265-286.

Principles for self-motivation: Homeostasis etc.

- Homeostasis: Maintain state in a “viable” zone (W. B. Cannon, 1926; W. R. Ashby, 1948)
- Allostasis: achieving stability (homeostasis) through physiological or behavioral change (P. Sterling and J. Eyer, 1988)
- Heterostasis: Drive away from the habitual state (H. Selye, 1973)
- Homeokinesis: Self-organised behaviour aiming at predictable changes (R. Der, 1999)

Homeokinesis

- Aim at state transitions that are predictable $\Rightarrow$ model with minimal prediction error
- Aim at states where actions have an effect (or at actions that affect the state) $\Rightarrow$ sensitivity
- Playful behaviour as a compromise between these two conflicting goals
- Self-generated behaviours can be used as options for RL

LPZrobots (\http://robot.informatik.uni-leipzig.de)

Information gain and maximal learning progress

- Predictive model for state transitions $M\left(s_t\right) = \hat{s}_{t+1}$
- Self-evaluation of the model: Sliding average of prediction error
- Choose actions that minimise the 2nd derivative of the prediction error $\left\langle \left|\hat{s}_{t+1} - s_{t+1}\right|^2 \right\rangle)$
- Result: Agent follow a behaviour as long as it improves in learning. If the rate of the error reduction decays, agent is likely to move on to other behaviours

Actor-Critic: Heuristic balance model

Soft policies: how soft exactly? Use entropy.

Consider a game between Actor and critic:

- Actor aims at decrease $\langle H(\pi) \rangle_{\mu(s)}$ in order to get more reward Δr, i.e. the actor transfers entropy into reward. For a given entropy reduction prefer actions that increase Δr.
- Critic aims at increase $\langle H(\pi) \rangle_{\mu(s)}$ in order to explore, which may (or may not) result in a decrease of the reward
  For given entropy reduction prefer actions that decrease Δr least.

Act such as to keep the balance. Balance will obviously shift.

Conclusions

- Intrinsic rewards can
  - speed-up learning
  - generalise beyond known tasks
  - direct exploration
- Can be obtained from
  - From demonstration by inverse reinforcement learning
  - General principles related to homeostasis
  - Successful self-generated options
- Intrinsic rewards are essential in biological and psychological systems