Reward Shaping in Episodic Reinforcement Learning

Marek Grześ

University of Kent | Computing

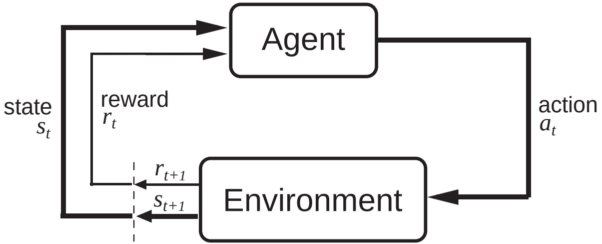
Canterbury, UK

AAMAS 2017

São Paulo, May 8–12

Motivating Reward Shaping

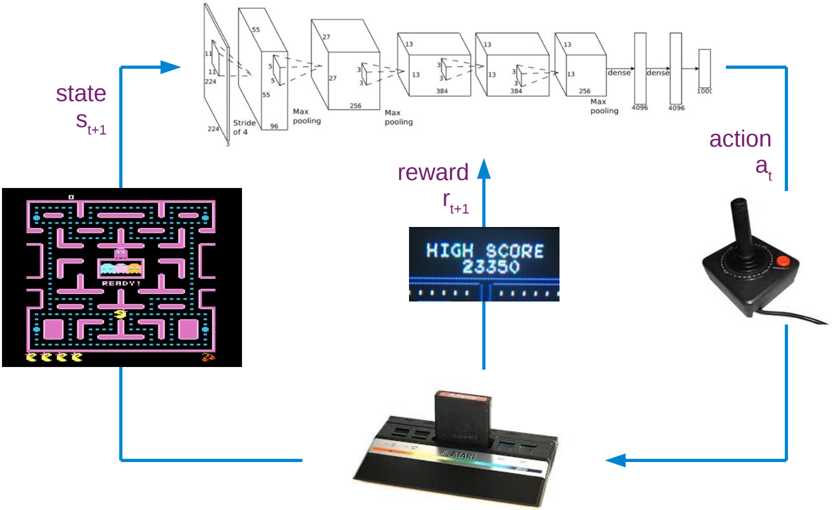
Reinforcement Learning



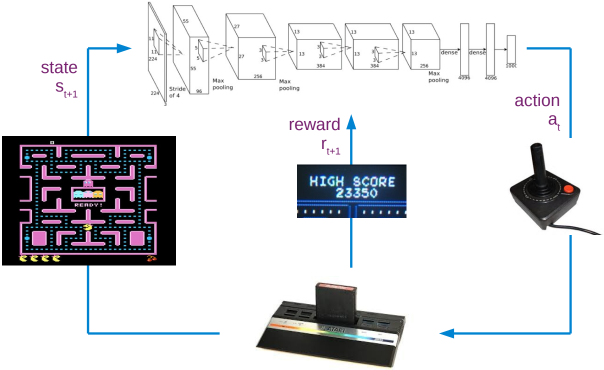
[Sutt 98]

Temporal credit assignment problem

Deep Reinforcement Learning

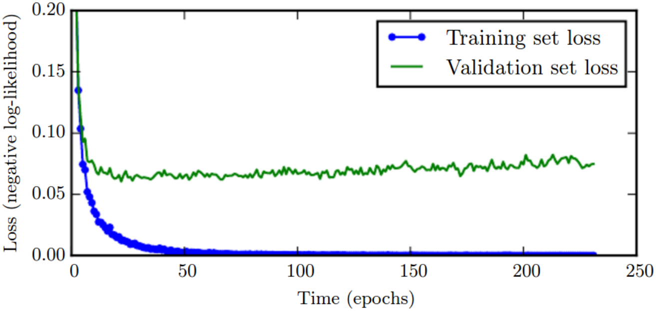


Challenges



- Temporal credit assignment problem
- In games, we can just generate more data for reinforcement learning
- However, 'more learning' in neural networks can be a challenge ... (see next slide)

Contradictory Objectives



http://www.deeplearningbook.org

- Easy to overfit
- Early stopping is a potential regulariser, but we need a lot of training to address the temporal-credit assignment problem
- Conclusion: It can be useful to mitigate the temporal credit assignment problem using reward shaping!

Reward Shaping

- $\langle s_t, a_t, s_{t+1}, r_{t+1} \rangle$
- r_{t+1} goes to Q-learning, SARSA, R-max etc.
- $r_{t+1} + F(s_t, a_t, s_{t+1})$
- where $F(s_t, a_t, s_{t+1}) = \gamma\Phi(s_{t+1}) - \Phi(s_t)$

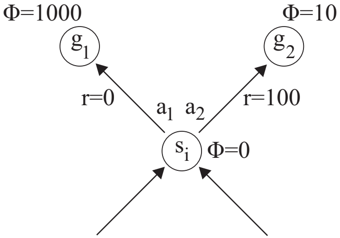
Policy Invariance under Reward Transformations

Potential-based reward shaping is necessary and sufficient to guarantee policy invariance [Ng 99]

Straightforward to show in infinite-horizon MDPs [Asmu 08]

Investigating episodic learning leads to new insights

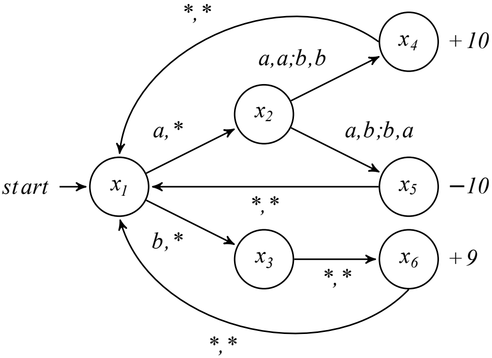
Problematic Example in Single-agent RL



[Grze 10]

- $F(s, goal) = 0$ in my PhD thesis
- [Ng 99] required $F(goal, \cdot) = 0$
- $\Phi(goal) = 0$ is what is necessary

Multi-agent Learning and Nash Equilibria



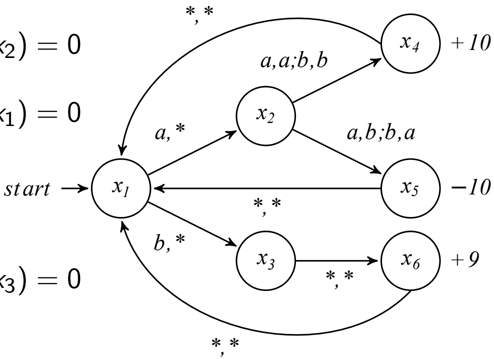
[Bout 99, Devl 11]

Multi-agent Learning and Nash Equilibria

$\Phi(x_2) = 0$

$\Phi(x_1) = 0$

$\Phi(x_3) = 0$



$\Phi(x_4) = 0$

$\Phi(x_5) = M$

$\Phi(x_6) = 0$

When M is sufficiently large,
we have a new Nash Equlibrium.

PAC-MDP Reinforcement Learning and R-max

Optimism in AI and Optimisation

- A*
- Branch-and-Bound
- R-max and optimistic potential functions [Asmu 08]

PAC-MDP Reinforcement Learning and R-max

Optimism in AI and Optimisation

- A*
- Branch-and-Bound
- R-max and optimistic potential functions [Asmu 08]

Sufficient conditions for R-max

- $\forall_{s \in Goals} \Phi(s) = 0$
- $\forall_{s \in Known} \Phi(s) = C$ where C is an arbitrary number
- $\forall_{s \in Unknown} \Phi(s) \geq 0$
- where $Goals \cap Known \cap Unknown = \varnothing$

MDP Planning: Infinite-horizon

- MDP solutions methods: linear programming
- $F(s, a, s') = \gamma\Phi(s') - \Phi(s)$
- The impact of reward shaping:

$$\sum_{s,a,s'} \lambda(s,a) T(s,a,s') F(s,a,s') = -\sum_{s'} \Phi(s')\mu(s')$$

MDP Planning: Finite-Horizon

$$\sum_{s \in S \setminus G} \sum_{a \in A} \sum_{s' \in S} \lambda(s, a) T(s, a, s') F(s, a, s')$$
$$= \sum_{s' \in G} \Phi(s') \Big[\sum_{s \in S \setminus G} \sum_{a \in A} \lambda(s, a) T(s, a, s') \Big]$$

References I

[Asmu 08] J. Asmuth, M. L. Littman, and R. Zinkov.
"Potential-based Shaping in Model-based Reinforcement Learning".
In: *Proceedings of AAAI*, 2008.

[Bout 99] C. Boutilier.
"Sequential Optimality and Coordination in Multiagent Systems".
In: *Proceedings of the International Joint Conferrence on Artificial Intelligencekue*, pp. 478–485, 1999.

[Devl 11] S. Devlin and D. Kudenko.
"Theoretical Considerations of Potential-Based Reward Shaping for Multi-Agent Systems".
In: *Proceedings of AAMAS*, 2011.

[Grze 10] M. Grzes.
Improving exploration in reinforcement learning through domain knowledge and parameter analysis.
PhD thesis, University of York, 2010.

[Ng 99] A. Y. Ng, D. Harada, and S. J. Russell.
"Policy Invariance Under Reward Transformations: Theory and Application to Reward Shaping".
In: *Proceedings of the 16th International Conference on Machine Learning*, pp. 278–287, 1999.

[Sutt 98] R. S. Sutton and A. G. Barto.
Reinforcement Learning: An Introduction.
MIT Press, 1998.