HUMBOLDT-UNIVERSITÄT ZU BERLIN

Reinforcement Learning

An Introduction/Overview

Dennis Wagner

Humboldt University of Berlin

Overview

Reinforcement Learning

Overview

Reinforcement Learning

Overview

Reinforcement Learning

Overview

Overview

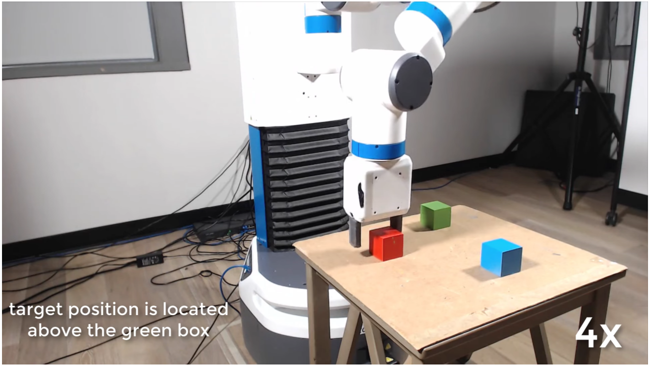
Reinforcement Learning

Introduction

Examples of Reinforcement Learning Problems

Industrial Robotics

Industrial robots learning to manipulate objects.

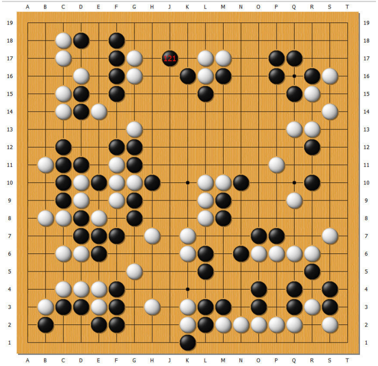


Andrychowicz, M., Wolski, F., Ray, A., Schneider, J., Fong, R., Welinder, P., McGrew, B., Tobin, J., Abbeel, O.P. and Zaremba, W., 2017. Hindsight experience replay. In Advances in Neural Information Processing Systems (pp. 5048-5058).

Examples of Reinforcement Learning Problems

Games

Teaching a computer program to play complex games.



Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A. and Chen, Y., 2017. Mastering the game of go without human knowledge. Nature, 550(7676), p.354.

Examples of Reinforcement Learning Problems

Games

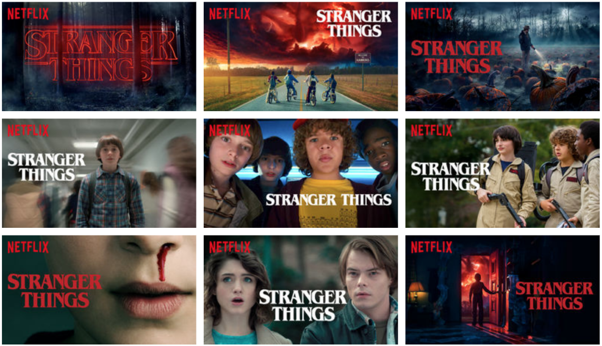
Teaching a computer program to play complex games.

Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D. and Riedmiller, M., 2013. Playing atari with deep reinforcement learning. arXiv preprint arXiv:1312.5602.

Examples of Reinforcement Learning Problems

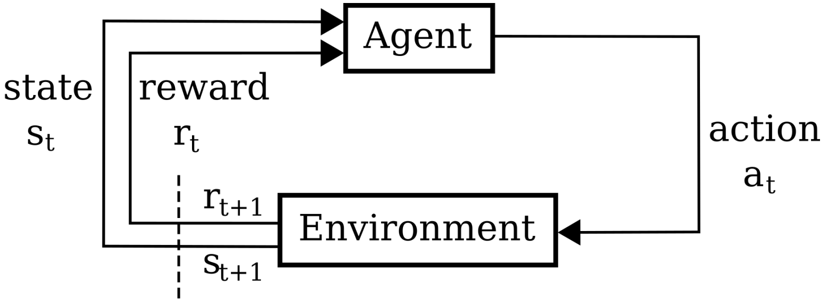
Recommender Systems

Modelling interactions with the user as an MDP.



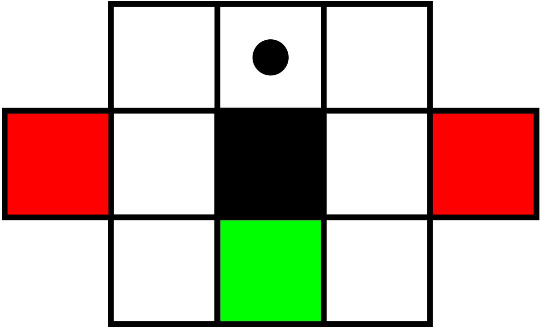
Zhao, X., Zhang, L., Ding, Z., Yin, D., Zhao, Y. and Tang, J., 2017. Deep Reinforcement Learning for List-wise Recommendations. arXiv preprint arXiv:1801.00209.

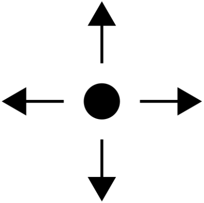
The general setting



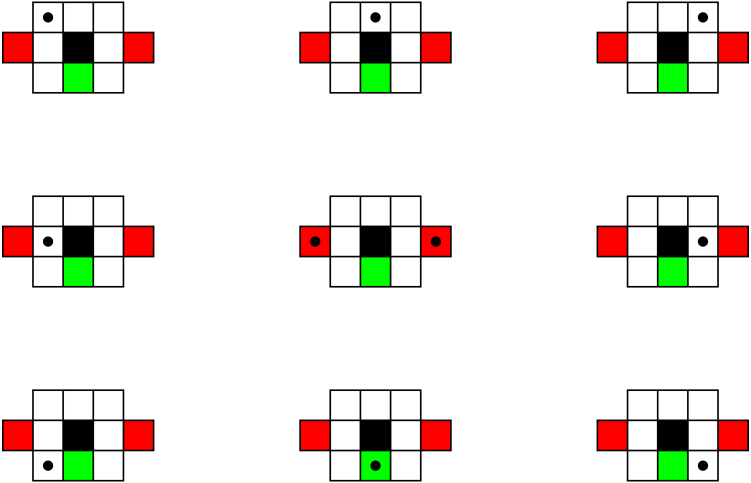
Markov Decision Processes

Running example - Gridworld

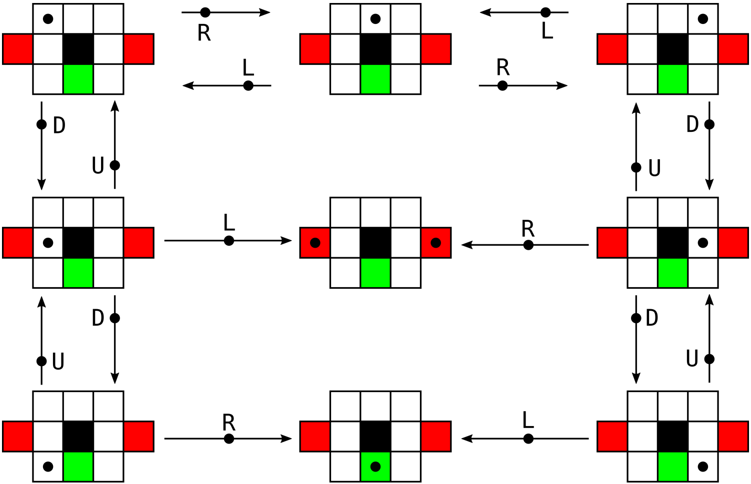




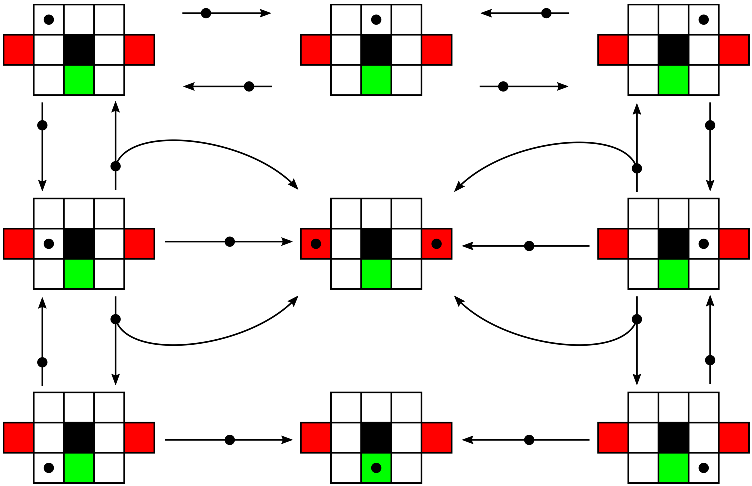
Elements of the RL Model - Gridworld



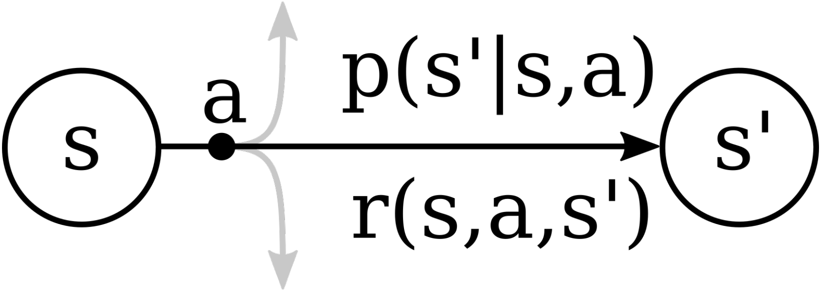
Elements of the RL Model - Gridworld



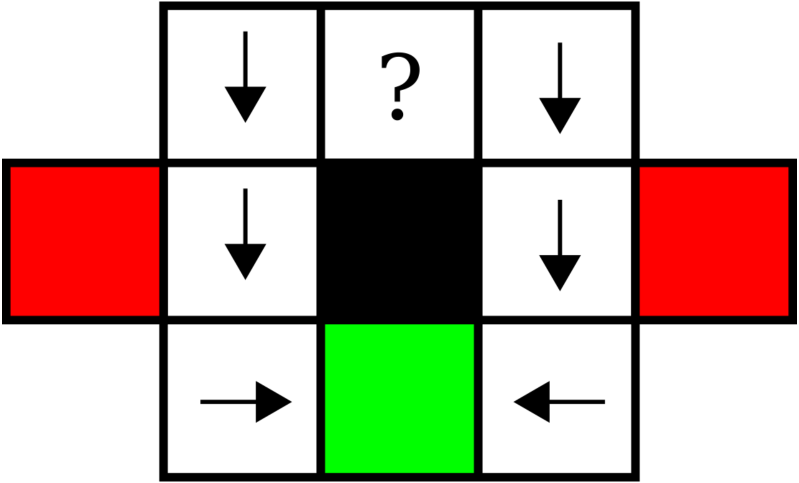
Elements of the RL Model - Gridworld



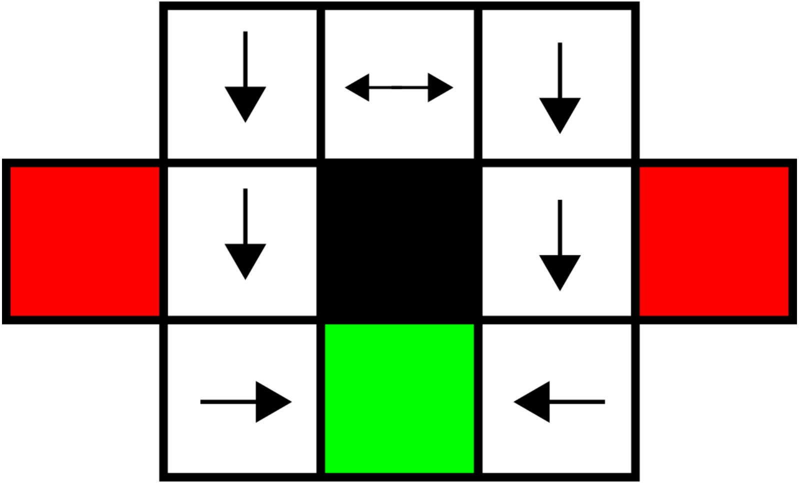
Elements of the RL Model - MDP



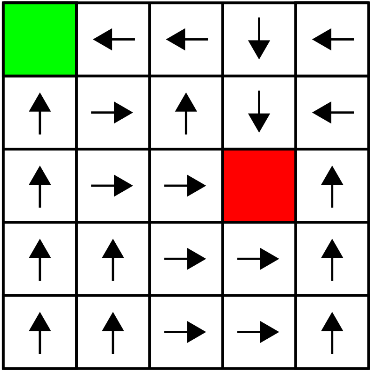
Elements of the RL Model - policy



Elements of the RL Model - policy



Example: Gridworld



Example: Gridworld

	←	←	↓	←
↑	→	↑	↓	←
↑	→	→		↑
↑	↑	→	→	↑
↑	↑	→	→	↑

	1	0.8	-1.2	-1.4
1	0.4	0.6	-1	-1.2
0.8	-1.2	-1		-1.4
0.6	-1.4	-2	-1.8	-1.6
?	-1.6	-2.2	-2	-1.8

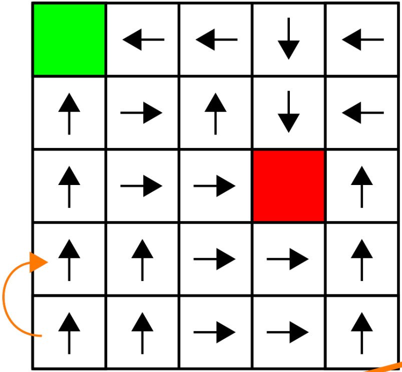
Example: Gridworld

	←	←	↓	←
↑	→	↑	↓	←
↑	→	→		↑
↑	↑	→	→	↑
↑	↑	→	→	↑

	1	0.8	-1.2	-1.4
1	0.4	0.6	-1	-1.2
0.8	-1.2	-1		-1.4
0.6	-1.4	-2	-1.8	-1.6
?	-1.6	-2.2	-2	-1.8

$$V^{\pi}(s) =$$

Example: Gridworld



	1	0.8	-1.2	-1.4
1	0.4	0.6	-1	-1.2
0.8	-1.2	-1		-1.4
0.6	-1.4	-2	-1.8	-1.6
?	-1.6	-2.2	-2	-1.8

$$V^{\pi}(s) = -0.2$$

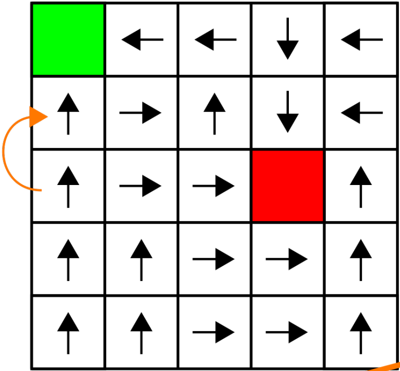
Example: Gridworld

	←	←	↓	←
↑	→	↑	↓	←
↑	→	→		↑
↑	↑	→	→	↑
↑	↑	→	→	↑

	1	0.8	-1.2	-1.4
1	0.4	0.6	-1	-1.2
0.8	-1.2	-1		-1.4
0.6	-1.4	-2	-1.8	-1.6
?	-1.6	-2.2	-2	-1.8

$$V^{\pi}(s) = -0.2 - 0.2$$

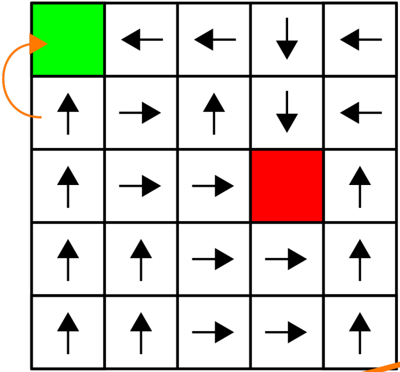
Example: Gridworld



	1	0.8	-1.2	-1.4
1	0.4	0.6	-1	-1.2
0.8	-1.2	-1		-1.4
0.6	-1.4	-2	-1.8	-1.6
?	-1.6	-2.2	-2	-1.8

$$V^\pi(s) = -0.2 - 0.2 - 0.2$$

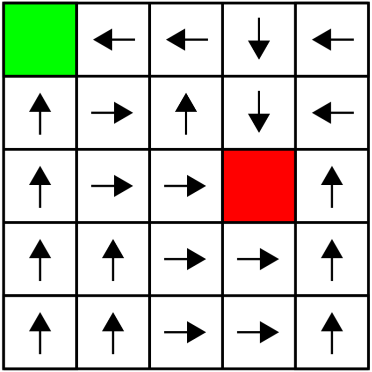
Example: Gridworld



	1	0.8	-1.2	-1.4
1	0.4	0.6	-1	-1.2
0.8	-1.2	-1		-1.4
0.6	-1.4	-2	-1.8	-1.6
?	-1.6	-2.2	-2	-1.8

$$V^{\pi}(s) = -0.2 - 0.2 - 0.2 + 1$$

Example: Gridworld



	1	0.8	-1.2	-1.4
1	0.4	0.6	-1	-1.2
0.8	-1.2	-1		-1.4
0.6	-1.4	-2	-1.8	-1.6
0.4	-1.6	-2.2	-2	-1.8

Example: Gridworld (2)

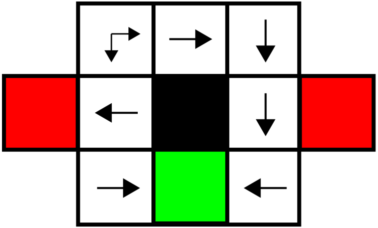
	1	0.8	0.6	0.4
1	0.8	0.6	0.4	0.2
0.8	0.6	0.4		0
0.6	0.4	0.2	0	-0.2
0.4	0.2	0	-0.2	-0.4

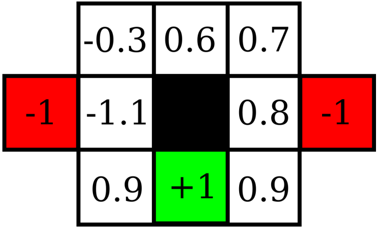
Example: Gridworld (2)

	1	0.8	0.6	0.4
1	0.8	0.6	0.4	0.2
0.8	0.6	0.4		0
0.6	0.4	0.2	0	-0.2
0.4	0.2	0	-0.2	-0.4

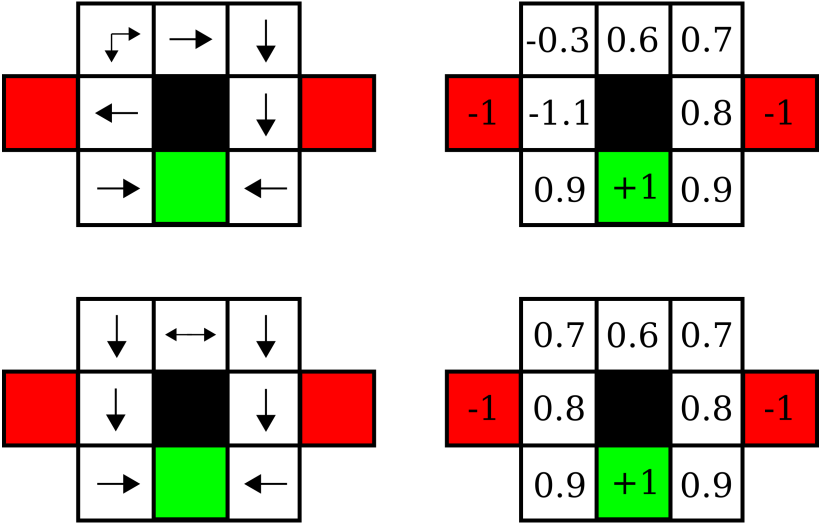
	←	←	←	←
↑	↑	↑	↑	↑
↑	↑	↑		↑
↑	↑	←	←	↑
↑	↑	←	←	←

Example: Gridworld





Example: Gridworld

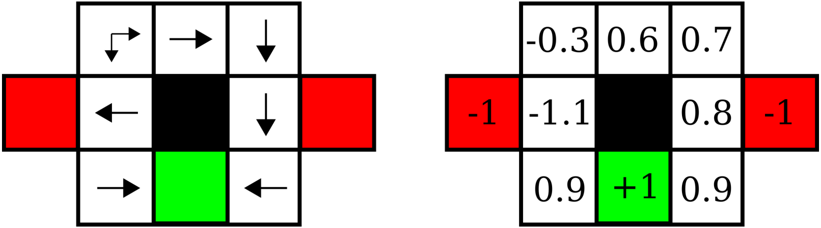


Bellman Theorem

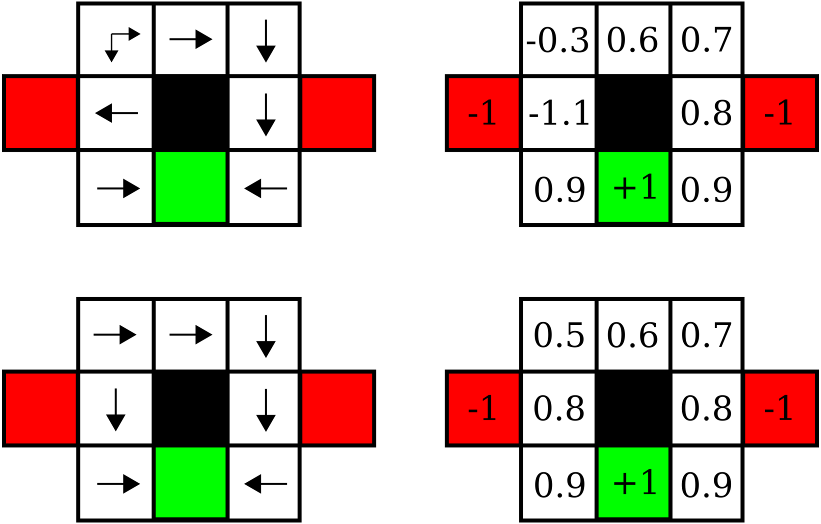
blackboard

Generalized Policy Iteration

Example: Gridworld



Example: Gridworld



Generalized Policy Iteration

function GPI
    $\pi \leftarrow$ arbitrary policy

    return π
end function

Generalized Policy Iteration

function GPI
  $\pi \leftarrow$ arbitrary policy
  repeat
  until convergence
  return π
end function

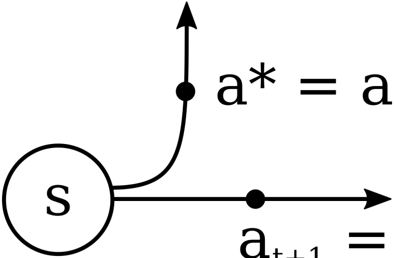
Generalized Policy Iteration

function GPI
    $\pi \leftarrow$ arbitrary policy
    repeat
        $V \leftarrow V^{\pi}$ $\qquad\qquad \triangleright$ Policy Evaluation
    until convergence
    return π
end function

Generalized Policy Iteration

```
function GPI
    π ← arbitrary policy
    repeat
        V ← V^π                 ▷ Policy Evaluation
        π ← greedy(V)           ▷ Policy Improvement
    until convergence
    return π
end function
```

Offpolicy (TD) learning

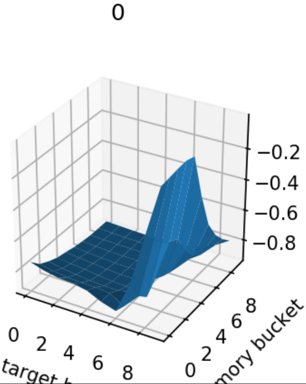


$$a^* = \arg\max_a Q(s,a)$$

$$a_{t+1} = \arg\max_a \pi(s,a)$$

Function Approximation

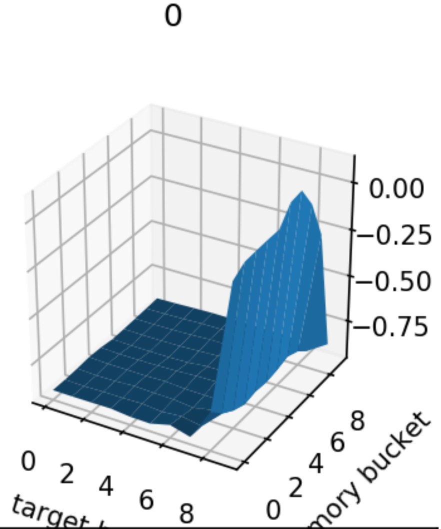
Example: Q



	10	0	0	0	0	0	0	0	0	0	0

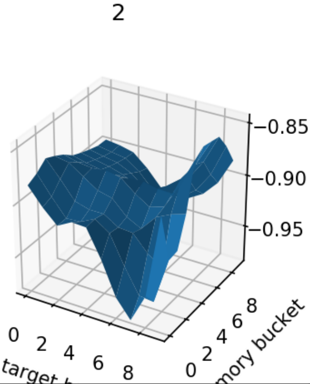
0	0	0	0	0	0	0	0	0	10

Example: Q



	10	0	0	0	0	0	0	0	0	0	0

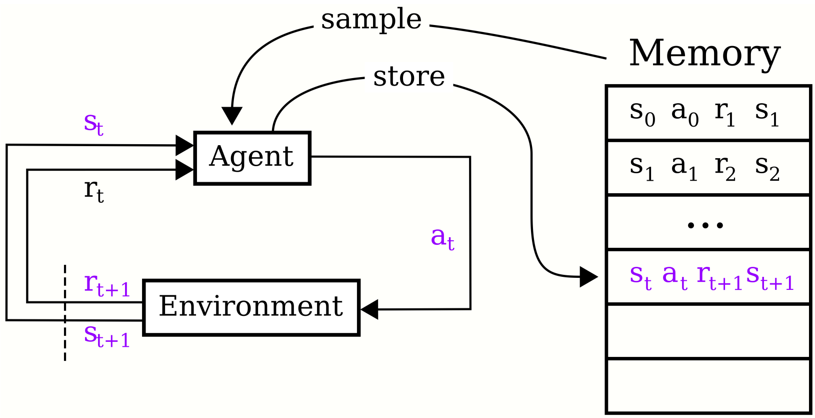
Example: Q



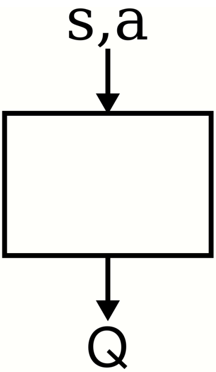
	10	0	0	0	0	0	0	0	0	0	0

0	0	0	0	0	0	0	0	0	10

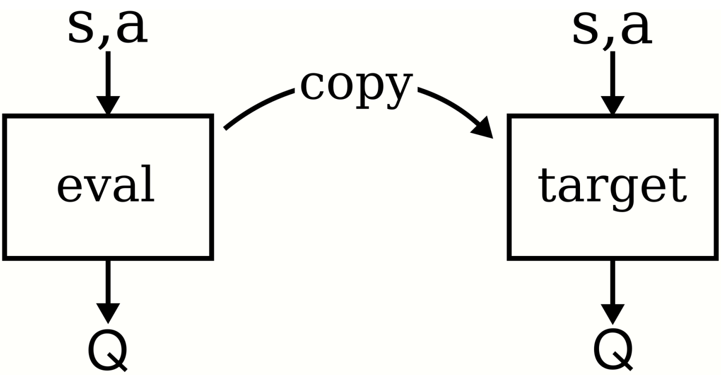
Experience Replay



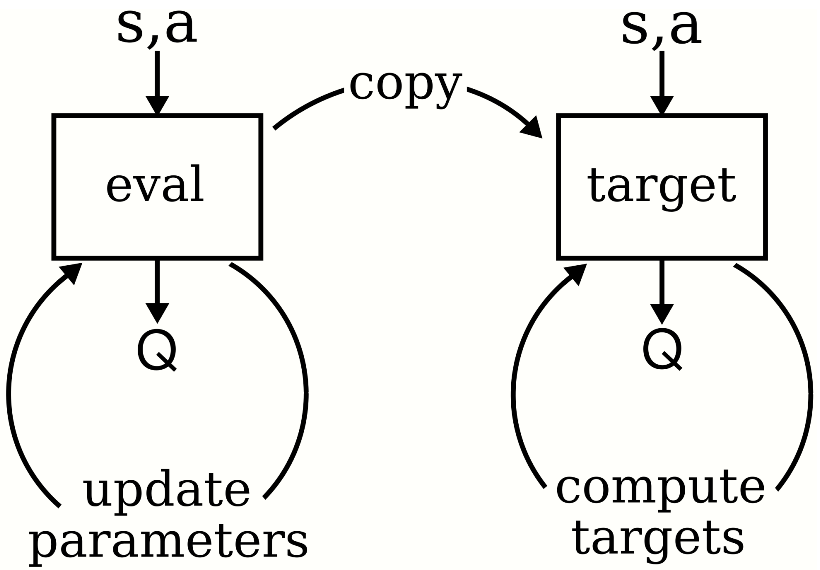
Fixed target nets



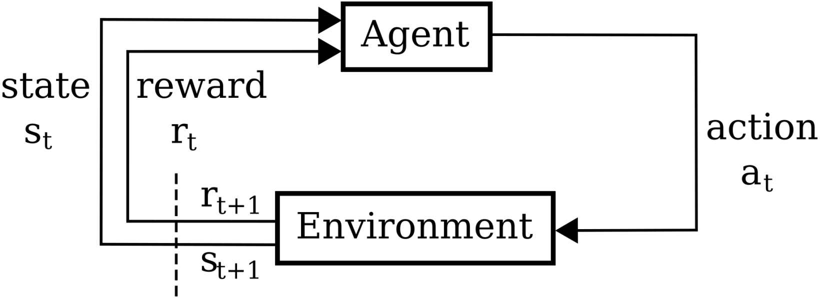
Fixed target nets



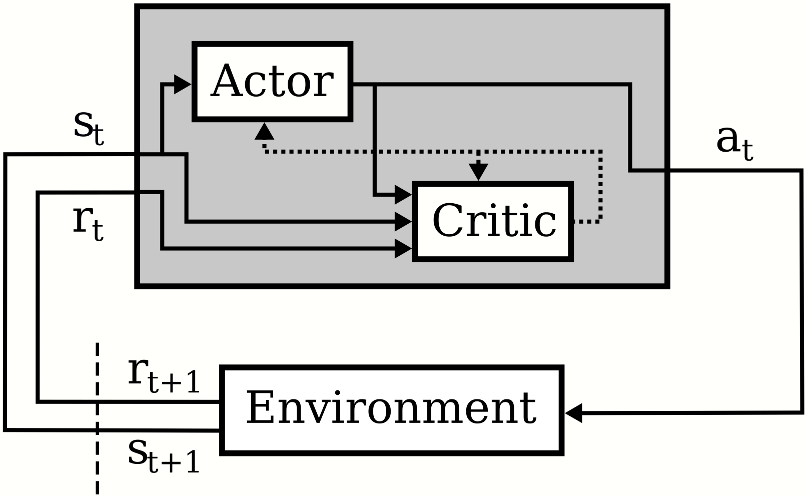
Fixed target nets



Actor Critic methods



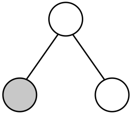
Actor Critic methods



An Example

Monte Carlo Tree Search

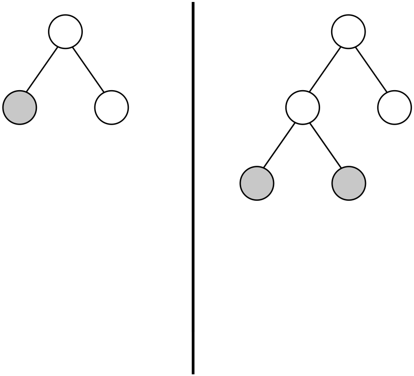
Selection



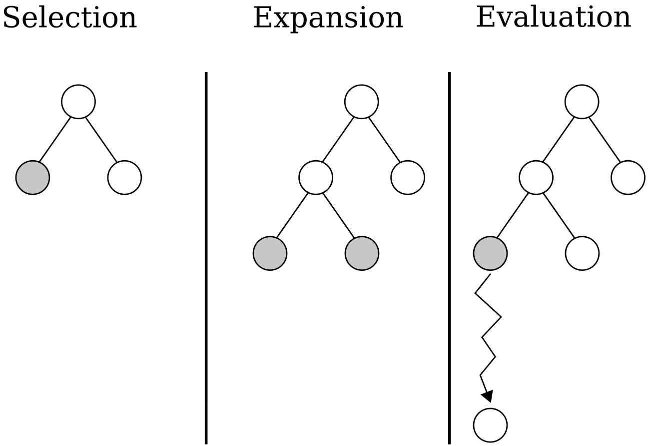
Monte Carlo Tree Search

Selection

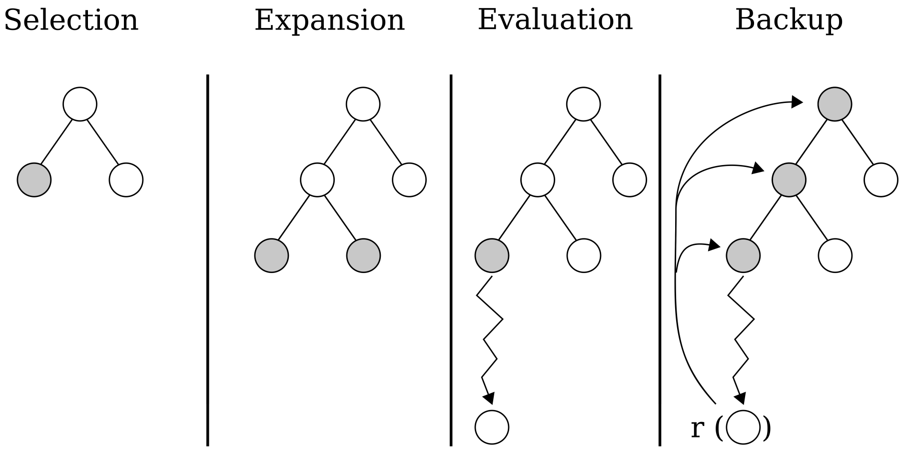
Expansion



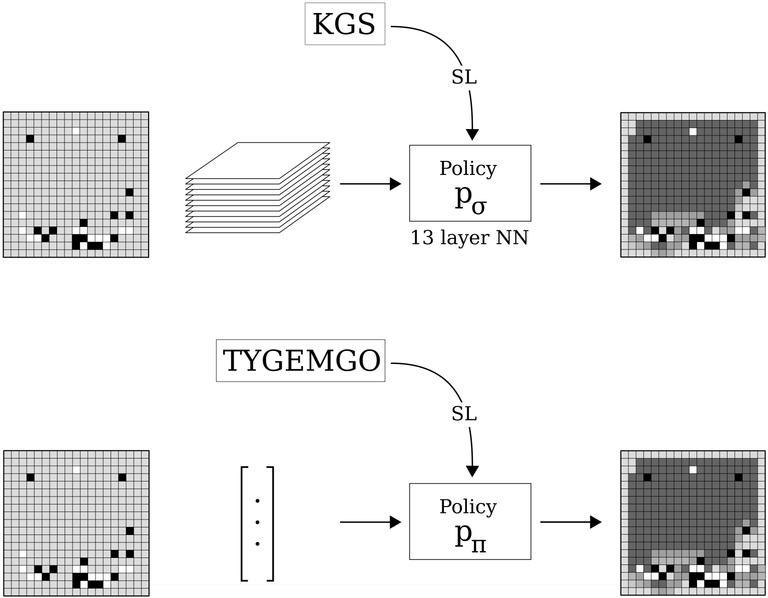
Monte Carlo Tree Search



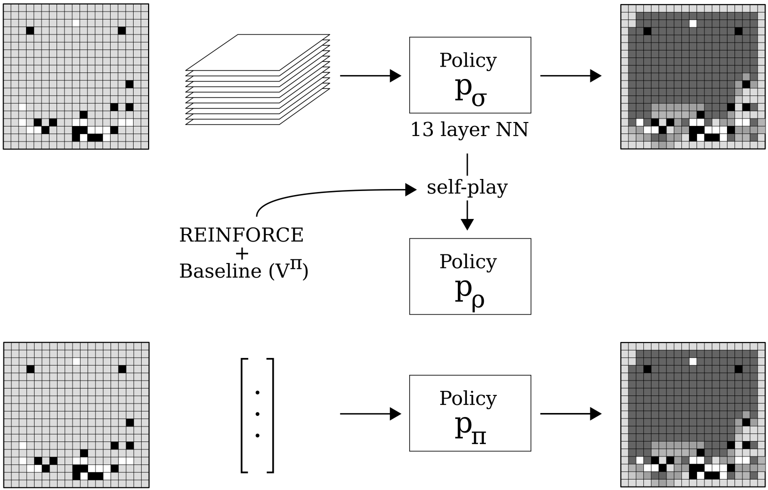
Monte Carlo Tree Search



AlphaGO

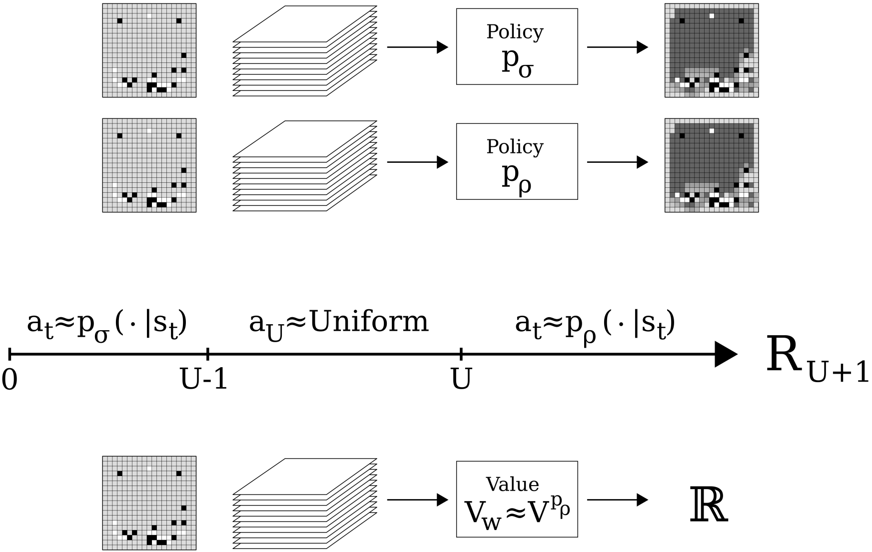


AlphaGO



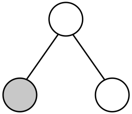
Policy
p_σ
13 layer NN
self-play
REINFORCE
+
Baseline (V^π)
Policy
p_ρ
Policy
p_π

AlphaGO


Policy
p_σ
Policy
p_ρ
$a_t \approx p_\sigma(\cdot|s_t)$
$a_U \approx \text{Uniform}$
$a_t \approx p_\rho(\cdot|s_t)$
0
U-1
U
R_{U+1}
Value
$V_w \approx V^{p_\rho}$
$\mathbb{R}$

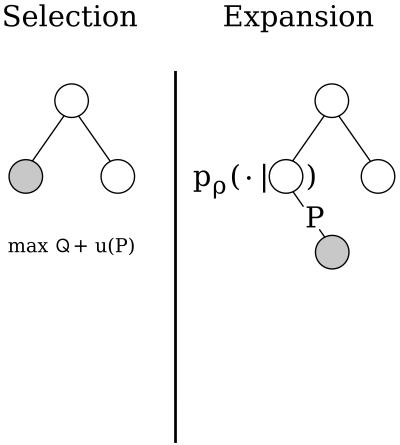
AlphaGO - MCTS

Selection

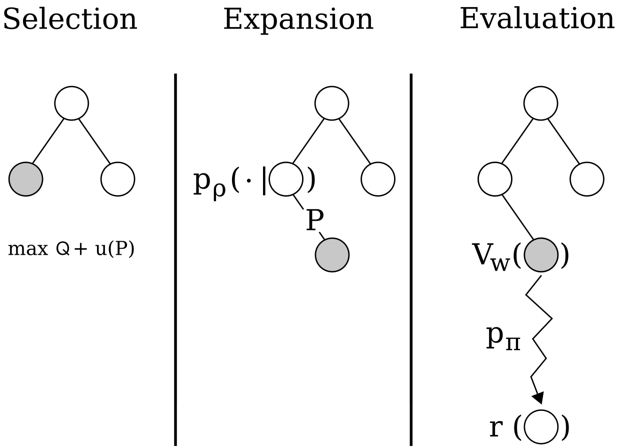


max Q + u(P)

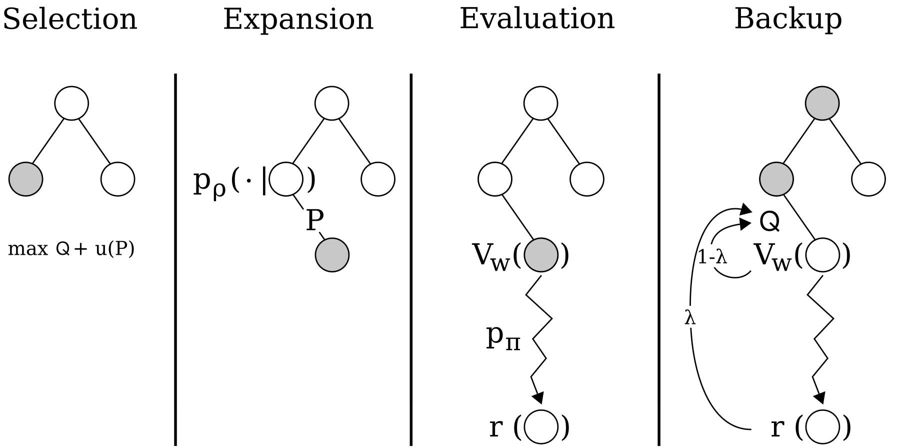
AlphaGO - MCTS



AlphaGO - MCTS



AlphaGO - MCTS



Thank you.
Any Questions?

References I

Exercises and solutions to accompany sutton's book and david silver's course. `https://github.com/dennybritz/reinforcement-learning`. Accessed: 2018-06-25.

Gym. `https://github.com/openai/gym`. Accessed: 2018-06-25.

Artwork personalization at netflix. `https://medium.com/netflix-techblog/artwork-personalization-c589f074ad76`. Accessed: 2018-06-25.

Ucl course on rl. `http://www0.cs.ucl.ac.uk/staff/d.silver/web/Teaching.html`. Accessed: 2018-06-25.

M. Andrychowicz, F. Wolski, A. Ray, J. Schneider, R. Fong, P. Welinder, B. McGrew, J. Tobin, O. P. Abbeel, and W. Zaremba. Hindsight experience replay. In **Advances in Neural Information Processing Systems**, pages 5048–5058, 2017.

S. M. Kakade. A natural policy gradient. In **Advances in neural information processing systems**, pages 1531–1538, 2002.

L. Li, W. Chu, J. Langford, and X. Wang. Unbiased offline evaluation of contextual-bandit-based news article recommendation algorithms. In **Proceedings of the fourth ACM international conference on Web search and data mining**, pages 297–306. ACM, 2011.

V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, and M. Riedmiller. Playing atari with deep reinforcement learning. **arXiv preprint arXiv:1312.5602**, 2013.

References II

V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, et al. Human-level control through deep reinforcement learning. **Nature**, 518(7540):529, 2015.

D. Silver, J. Schrittwieser, K. Simonyan, I. Antonoglou, A. Huang, A. Guez, T. Hubert, L. Baker, M. Lai, A. Bolton, et al. Mastering the game of go without human knowledge. **Nature**, 550(7676):354, 2017.

R. S. Sutton and A. G. Barto. **Reinforcement learning: An introduction**. The MIT Press, 1998. ISBN 9780262193986.

R. S. Sutton, D. A. McAllester, S. P. Singh, and Y. Mansour. Policy gradient methods for reinforcement learning with function approximation. In **Advances in neural information processing systems**, pages 1057–1063, 2000.

R. J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. In **Reinforcement Learning**, pages 5–32. Springer, 1992.

X. Zhao, L. Zhang, Z. Ding, D. Yin, Y. Zhao, and J. Tang. Deep reinforcement learning for list-wise recommendations. **arXiv preprint arXiv:1801.00209**, 2017.

Further references I

- Nuts and Bolts of Deep RL Experimentation (https://www.youtube.com/watch?v=8EcdaCk9KaQ)
- Trust Region Policy Optimization (https://arxiv.org/abs/1502.05477)
- Proximal Policy Optimization Algorithms (https://arxiv.org/abs/1707.06347) (https://blog.openai.com/openai-baselines-ppo/)
- OpenAI 5 (DOTA2) (https://blog.openai.com/openai-five-benchmark-results/)
- Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks (https://arxiv.org/abs/1703.03400)
- DQN extensions (https://medium.freecodecamp.org/improvements-in-deep-q-learning-dueling-double-dqn-prioritized-experience-replay-and-fixed-58b130cc5682)