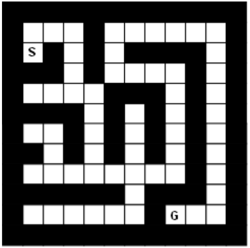
Potential Based Reward Shaping Tutorial

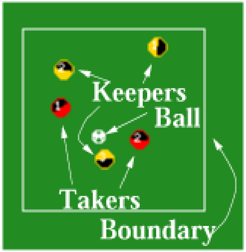
ALA 2014
Sam Devlin

THE UNIVERSITY of York

Knowledge-Based Reinforcement Learning

- Commonly, RL algorithms assume no prior knowledge
- Including domain knowledge can simplify learning





Reward Shaping

Q-Learning

- A popular RL algorithm
- $Q(s,a) \leftarrow Q(s,a) + \alpha[r + \gamma \max_{a'} Q_i(s',a') - Q(s,a)]$

Reward Shaping

- Provide heuristic knowledge by an additional reward
- $Q(s,a) \leftarrow Q(s,a) + \alpha[r + F(s,s') + \gamma \max_{a'} Q_i(s',a') - Q(s,a)]$

Potential-Based Reward Shaping

$$F(s, s') = \gamma\Phi(s') - \Phi(s)$$

- $F(s, s')$: Additional reward when moving from state s to s'
- γ : Discount factor
- $\Phi(s)$: Potential of state s

Potential-Based Reward Shaping

Formal Definition

- $F(s, s') = \gamma\Phi(s') - \Phi(s)$

Guarantees

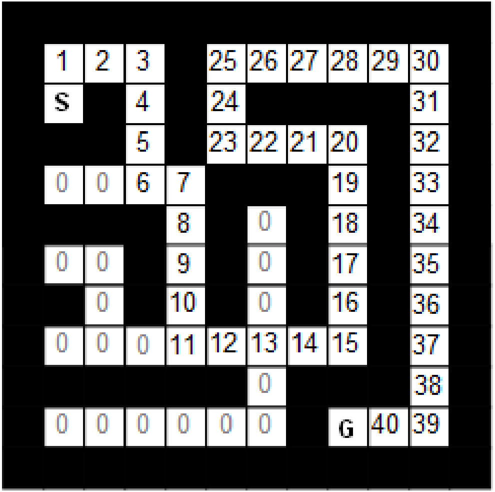
- Policy invariance (optimal policy unchanged) in single agent[1]

Can

- Increase/Decrease time taken to learn optimal policy

[1] Ng, Russell and Harada. "Policy Invariance Under Reward Transformations: Theory And Application To Reward Shaping." ICML, 1999.

An Example Potential Function



Potential-Based Reward Shaping

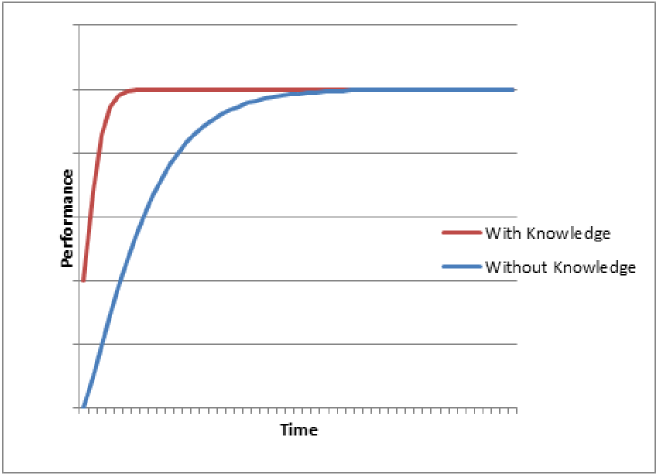


Figure : A Typical Single Agent Result

Proof of Policy Invariance [2]

$$
\begin{aligned}
U_\Phi(\bar{s}) &= \sum_{j=0}^{\infty} \gamma^j (r_j + \gamma \Phi(s_{j+1}) - \Phi(s_j)) \\
&= \sum_{j=0}^{\infty} \gamma^j r_j + \sum_{j=0}^{\infty} \gamma^{j+1} \Phi(s_{j+1}) - \sum_{j=0}^{\infty} \gamma^j \Phi(s_j) \\
&= U(\bar{s}) + \sum_{j=1}^{\infty} \gamma^j \Phi(s_j) - \Phi(s_0) - \sum_{j=1}^{\infty} \gamma^j \Phi(s_j) \\
&= U(\bar{s}) - \Phi(s_0)
\end{aligned}
$$

[2] Asmuth, Littman and Zinkov. "Potential-based shaping in model-based reinforcement learning." AAAI, 2008.

Proof of Policy Invariance

$$U_{\Phi}(\bar{s}) = \sum_{j=0}^{\infty} \gamma^{j}(r_j + \gamma\Phi(s_{j+1}) - \Phi(s_j))$$

Proof of Policy Invariance

$$
\begin{aligned}
U_\Phi(\bar{s}) &= \sum_{j=0}^{\infty} \gamma^j (r_j + \gamma\Phi(s_{j+1}) - \Phi(s_j)) \\
&= \sum_{j=0}^{\infty} \gamma^j r_j + \sum_{j=0}^{\infty} \gamma^{j+1}\Phi(s_{j+1}) - \sum_{j=0}^{\infty} \gamma^j \Phi(s_j)
\end{aligned}
$$

Proof of Policy Invariance

$$
\begin{aligned}
U_\Phi(\bar{s}) &= \sum_{j=0}^{\infty} \gamma^j (r_j + \gamma\Phi(s_{j+1}) - \Phi(s_j)) \\
&= \sum_{j=0}^{\infty} \gamma^j r_j + \sum_{j=0}^{\infty} \gamma^{j+1}\Phi(s_{j+1}) - \sum_{j=0}^{\infty} \gamma^j \Phi(s_j) \\
&= U(\bar{s}) + \sum_{j=1}^{\infty} \gamma^j \Phi(s_j) - \Phi(s_0) - \sum_{j=1}^{\infty} \gamma^j \Phi(s_j)
\end{aligned}
$$

Proof of Policy Invariance

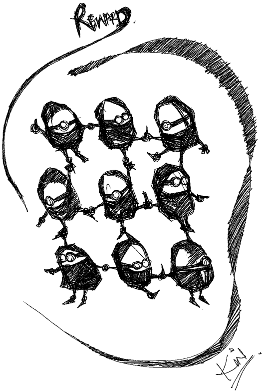
$$
\begin{aligned}
U_\Phi(\bar{s}) &= \sum_{j=0}^{\infty} \gamma^j (r_j + \gamma\Phi(s_{j+1}) - \Phi(s_j)) \\
&= \sum_{j=0}^{\infty} \gamma^j r_j + \sum_{j=0}^{\infty} \gamma^{j+1}\Phi(s_{j+1}) - \sum_{j=0}^{\infty} \gamma^j \Phi(s_j) \\
&= U(\bar{s}) + \sum_{j=1}^{\infty} \gamma^j \Phi(s_j) - \Phi(s_0) - \sum_{j=1}^{\infty} \gamma^j \Phi(s_j) \\
&= U(\bar{s}) - \Phi(s_0)
\end{aligned}
$$

Q-Table Initialization

- Wiewiora: "Potential-based shaping and Q-value initialization are equivalent." (JAIR, 2003)
- ...If the potential function is **static**.

Multi-Agent Reinforcement Learning

- Multiple agents learning concurrently in the same environment
- Typically learn a Nash equilibrium
- No clear notion of an optimal policy



Multi-Agent Potential-Based Reward Shaping

Guarantees

- Nash Equilibria not altered [3]

Can

- Increase/Decrease time taken to reach a stable joint policy
- Change final joint policy

[3] Devlin and Kudenko, "Theoretical Considerations Of Potential-Based Reward Shaping For Multi-Agent Systems", AAMAS, 2011.

Potential-Based Reward Shaping

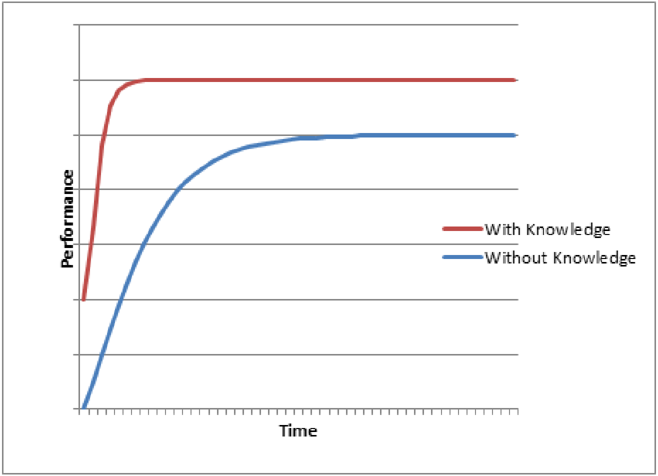
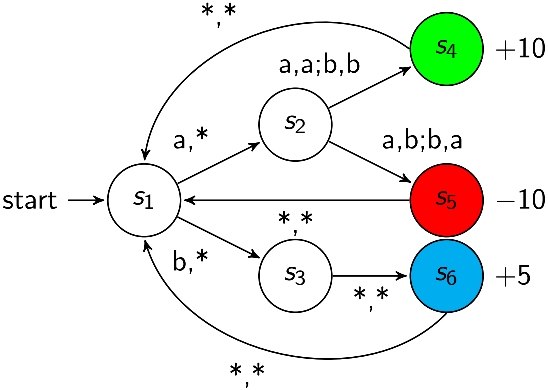


Figure : A Typical Multi-Agent Result

Exploration Altered

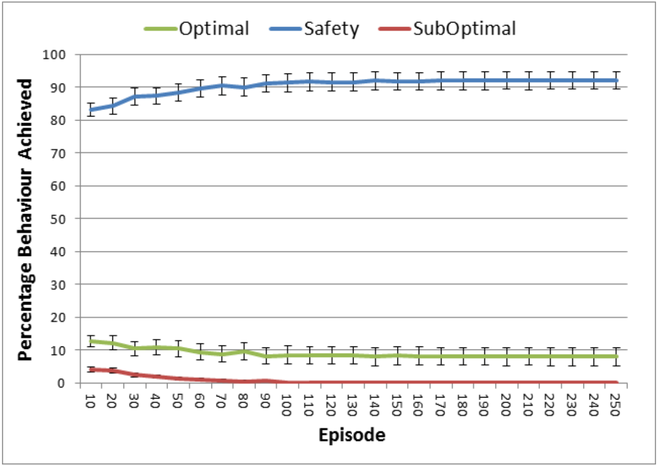
- Reward shaping alters exploration / which actions are chosen
- In single-agent, this affects time to convergence
- In multi-agent, this may cause the agents to reach a different point of equilibrium
  - Wellman and Hu (1998) showed the joint policy converged upon in a learning MAS is highly sensitive to initial belief

Multi-Agent Example [4]



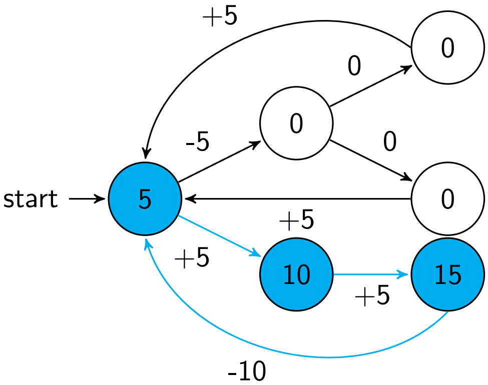
[4] Boutillier, "Sequential Optimality And Coordination In Multiagent Systems", IJCAI, 1999.

Results

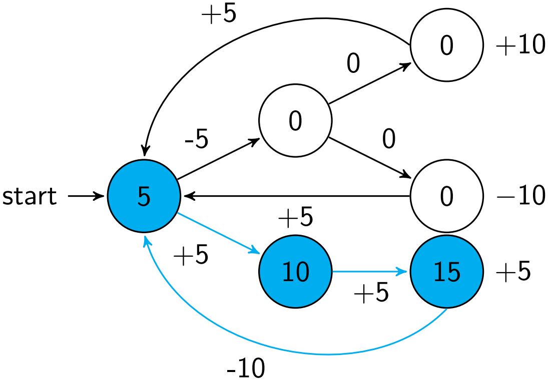


(a) Without Reward Shaping

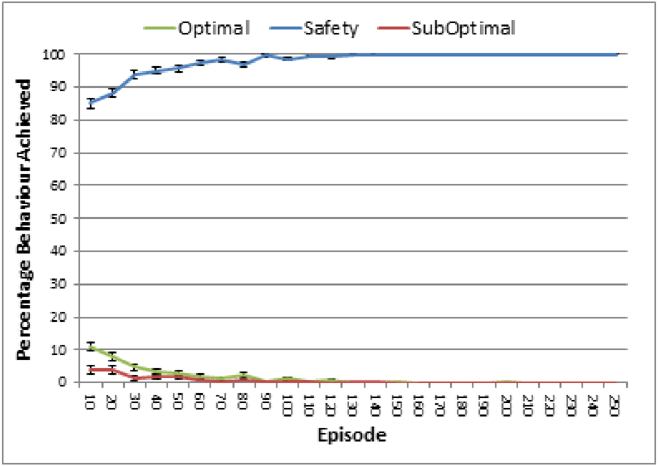
Safe Reward Shaping



Safe Reward Shaping

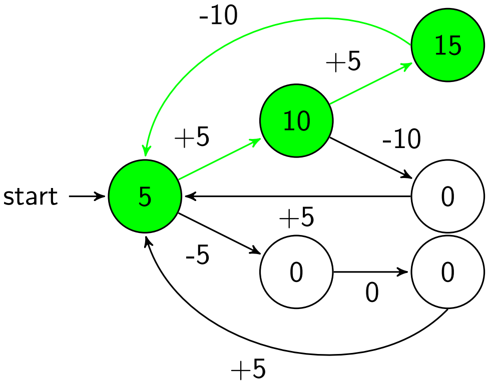


Results

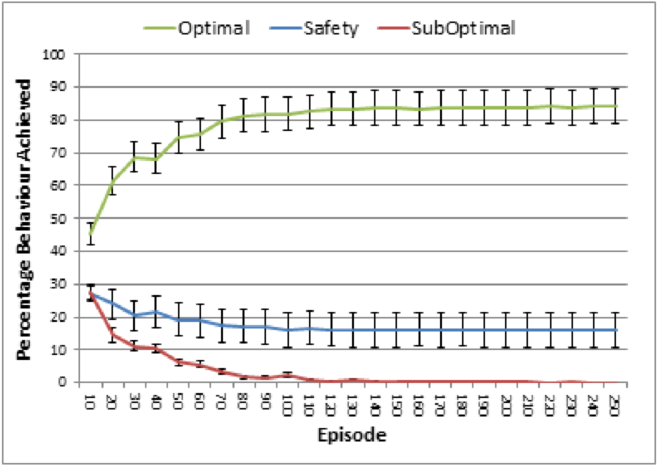


(b) With Safe Reward Shaping

Coordinated Reward Shaping

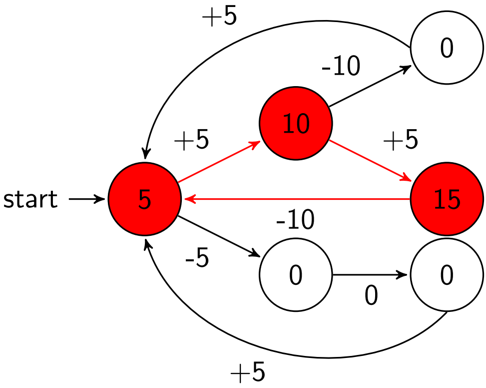


Results

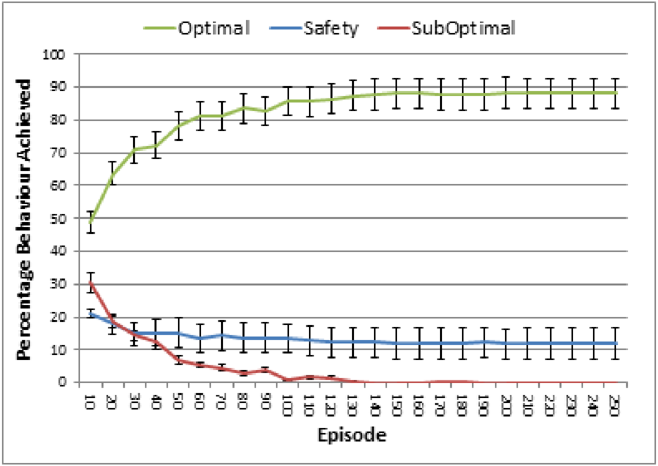


(c) With Coordinated Reward Shaping

Miscoordinated Reward Shaping

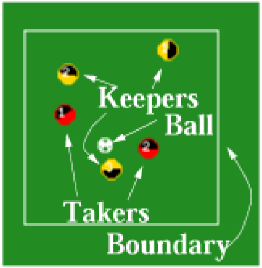


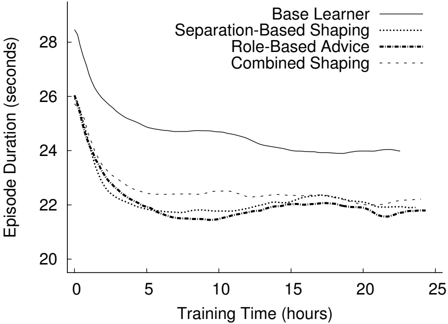
Results



(d) With Miscoordinated Reward Shaping

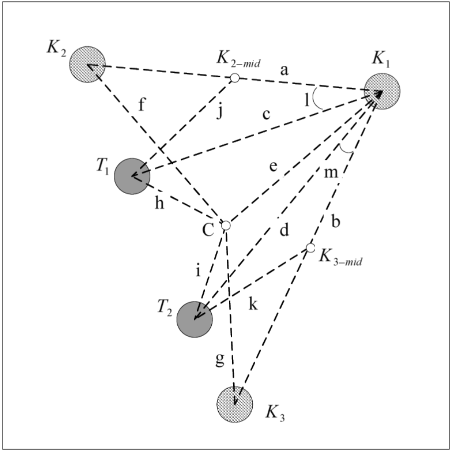
Multiagent Example 2: RoboCup KeepAway [5]





[5] Devlin, Grzes and Kudenko. "An Empirical Study Of Potential-Based Reward Shaping And Advice In Complex, Multi-Agent Systems." ACS, 2011.

Practical Exercise: Design a Potential Function



- a: $dist(K_1, K_2)$
- b: $dist(K_1, K_3)$
- c: $dist(K_1, T_1)$
- d: $dist(K_1, T_2)$
- e: $dist(K_1, C)$
- f: $dist(K_2, C)$
- g: $dist(K_3, C)$
- h: $dist(T_1, C)$
- i: $dist(T_2, C)$
- j: $\min\_dist(K_{2-mid}, T_j)$
- k: $\min\_dist(K_{3-mid}, T_j)$
- l: $\min\_ang(K_2, K_1, T_j)$
- m: $\min\_ang(K_3, K_1, T_j)$

Figure : State Representation

Past Assumptions

- Previous theoretical guarantees assume a **static** potential function
- Some claim the potential function must converge before the agent can [6]

[6] Laud, "Theory And Application Of Reward Shaping In Reinforcement Learning", PhD Thesis, 2004.

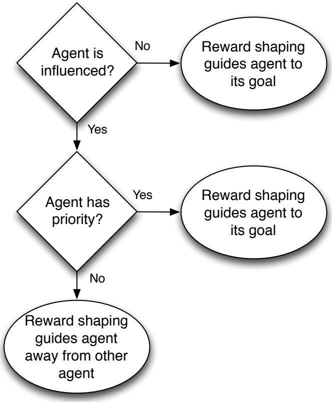
Dynamic Potential Based Reward Shaping [7]

- Guarantees policy invariance or consistant Nash equilibria, provided:

$$F(s, t, s', t') = \gamma\Phi(s', t') - \Phi(s, t)$$

[7] Devlin and Kudenko. "Dynamic Potential-Based Reward Shaping." AAMAS, 2012.

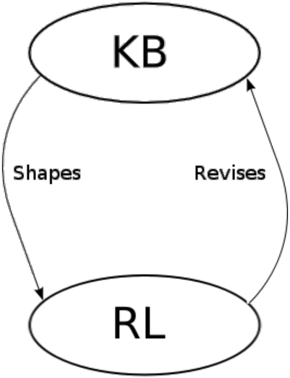
Context Sensitive Reward Shaping [8]

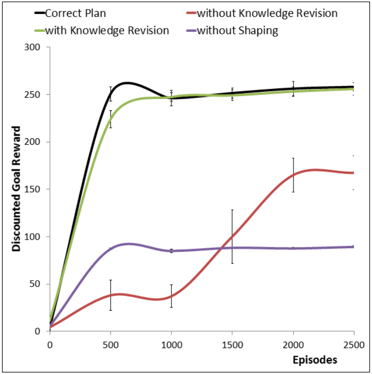


- In different contexts we often reccomend different behaviours

[8] De Hauwere, Devlin, Kudenko and Nowe. "Context Sensitive Reward Shaping for Sparse Interaction Multi-Agent Systems." BNAIC, 2013.

Belief Revision [9]





[9] Efthymiadis, Devlin, and Kudenko. "Overcoming Erroneous Domain Knowledge in Plan-Based Reward Shaping." AAMAS, 2013.

Q-Table Initialization

- Wiewiora: "Potential-based shaping and Q-value initialization are equivalent." (JAIR, 2003)
- ...If the potential function is **static**.

State and Action Shaping [10]

Look-Ahead Advice

- $F(s, a, s', a') = \gamma\Phi(s', a') - \Phi(s, a)$
- $\pi(s) = argmax_a\{Q(s, a) + \Phi(s, a)\}$
- Maintains all previous guarantees

Look-Back Advice

- $F(s, a, s', a') = \Phi(s', a') - \gamma^{-1}\Phi(s, a)$
- No guarantees proven

[10] Wiewiora, Cottrell and Elkan. "Principled methods for advising reinforcement learning agents." ICML, 2003.

Partial Observability

Formal Definition

- $F(o, o') = \gamma \Phi(o') - \Phi(o)$

Guarantees [11]

- Equivalence to Q-table initialisation
- Policy invariance (optimal policy unchanged) in single agent
- Consistant Nash equilibria in multi-agent systems

[11] Eck, Soh, Devlin and Kudenko. "Potential-Based Reward Shaping for Partially Observable Markov Decision Processes." AAMAS, 2013.

Closing Remarks

Implementation Advice

- γ must be equal to update rule
- Use an absorbing state
- Store current potential for next iteration
- Avoid negative potentials [12]

[12] Grzes and Kudenko. "Theoretical and empirical analysis of reward shaping in reinforcement learning." ICMLA, 2009.

General Effect

- Does not modify any property of the underlying MDP or SG invariant to changes in absolute value of expected return.

- Provided a property is only reliant on the **relative difference** or **order** of expected returns, potential-based reward shaping will not affect it.

Neccessity

- For every reward shaping function that is not potential-based, there is an MDP where the optimal policy differs with and without reward shaping. [13]

[13] Ng, Russell and Harada. "Policy Invariance Under Reward Transformations: Theory And Application To Reward Shaping." ICML, 1999.

References

- Ng, Russell and Harada. "Policy Invariance Under Reward Transformations: Theory And Application To Reward Shaping." ICML, 1999.
- Wiewiora. "Potential-based shaping and Q-value initialization are equivalent." JAIR, 2003
- Wiewiora, Cottrell and Elkan. "Principled methods for advising reinforcement learning agents." ICML, 2003.
- Asmuth, Littman and Zinkov. "Potential-based shaping in model-based reinforcement learning." AAAI 2008
- Devlin and Kudenko, "Theoretical Considerations Of Potential-Based Reward Shaping For Multi-Agent Systems", AAMAS, 2011.
- Devlin and Kudenko. "Dynamic Potential-Based Reward Shaping." AAMAS, 2012.

AAMAS 2014

Potential-Based Difference Rewards for Multiagent Reinforcement Learning.

Sam Devlin, Logan Yliniemi, Daniel Kudenko and Kagan Tumer

Learning I
Miles Davis A & B
Wednesday 09:00