Policy invariance under reward transformations: Theory and application to reward shaping

Andrew Y. Ng and Daishi Harada and Stuart Russell

(presented by Toryn Klassen)

November 24, 2016

Outline

MDP review

Reward shaping

To provide guidance, policies can be learned on an MDP with a modified reward function, and then used on the original MDP (with varying results).

Potential-based reward shaping

To ensure that good policies for a modified reward function are also good for the original, it suffices to base the rewards on a **potential function**.

Experiments

Some potential-based shaping functions are evaluated.

MDP review

Definition

A **Markov decision process (MDP)** is a tuple $M = \langle S, A, T, \gamma, R \rangle$ where

- S is a finite set of **states**,
- $A = \{a_1, \ldots, a_k\}$ is a set of **actions**,
- $T = \{P_{sa} : s \in S, a \in A\}$ specifies **transition probabilities**; $P_{sa}(s')$ is the probability of transitioning from s to s' with action a,
- γ is the **discount factor**, and
- $R : S \times A \times S \to \mathbb{R}$ is the **reward function**.

Definition

A **policy** over a set of states S is a function $\pi : S \to A$.

MDP review

Definition

Given a policy π and MDP $M = \langle S, A, T, \gamma, R \rangle$, the **value function** V_M^π is defined by

$$V_M^\pi(s) = \mathbb{E}[R_1 + \gamma R_2 + \gamma^2 R_3 + \dots ; \pi, s]$$

where R_i is the reward received on the ith step of following π, starting from s.

Definition

The Q-**function** is

$$Q_M^\pi(s, a) = \mathbb{E}_{s' \sim P_{sa}}[R(s, a, s') + \gamma V_M^\pi(s')]$$

MDP review

- The **optimal value function** is $V_M^*(s) = \sup_\pi V_M^\pi(s)$.
- The **optimal Q-function** is $Q_M^*(s, a) = \sup_\pi Q_M^\pi(s, a)$.
- The **optimal policy** is $\pi_M^*(s) = \operatorname{argmax}_{a \in A} Q_M^*(s, a)$.

Regularity conditions for undiscounted MDPs

When the discount γ is 1, we'll assume:

- There is an **absorbing** state s_0 s.t.
  - s_0 can never be left once entered, and
  - from s_0, no further rewards can be gained.
- The transition probabilities T are **proper**: starting from any state, following any policy will lead to s_0 with probability 1.

Modifying the reward function to provide guidance

To learn a policy for an MDP

$$M = \langle S, A, T, \gamma, R \rangle$$

we could instead run our reinforcement learning algorithm on a transformed MDP

$$M' = \langle S, A, T, \gamma, R' \rangle$$

where

$$R' = R + F$$

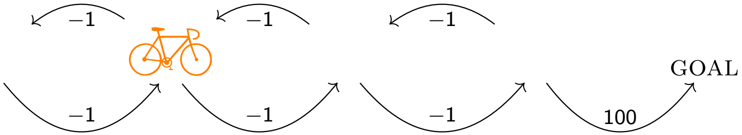
is the transformed reward function, and

$$F : S \times A \times S \to \mathbb{R}$$

is the **shaping reward function**.
When will an optimal (or good) policy for M' also be optimal (or good) for M?

Difficulties in reward shaping

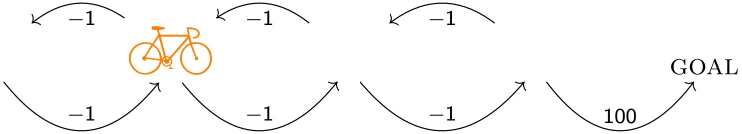
Consider this (undiscounted) problem:



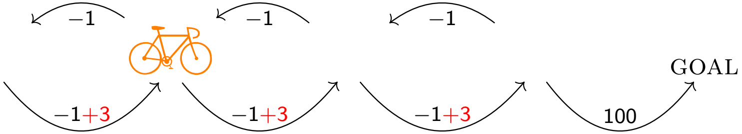
How can we modify the reward function to make the agent more quickly learn to move rightward to the goal?

Difficulties in reward shaping

Consider this (undiscounted) problem:

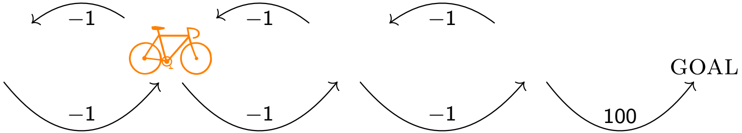


What if we give extra reward for going in the right direction?

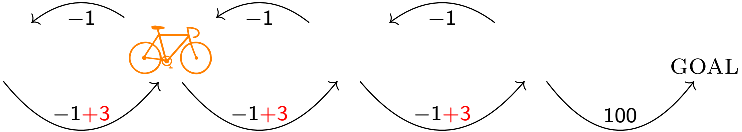


Difficulties in reward shaping

Consider this (undiscounted) problem:



What if we give extra reward for going in the right direction?



Problem: it's now better for the bicycle to try to go in a circle than to go the goal.

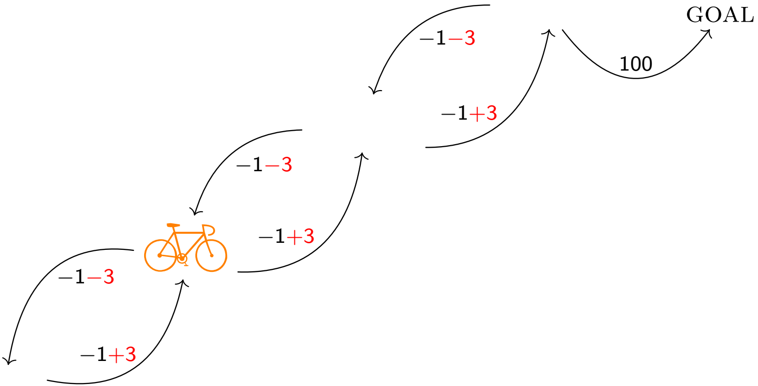
This problem isn't just a contrived artificial example.

Consider this description of work on a (more complicated) bicycle driving domain:

> *In our first experiments we rewarded the agent for driving towards the goal but did not punish it for driving away from it. Consequently the agent drove in circles with a radius of 20–50 meters around the starting point. Such behavior was actually rewarded by the reinforcement function [...]*
>
> — Randløv and Alstrøm (1998)

Idea: use a potential function

Associate a **potential** value $\Phi(s)$ to each state s, and add to the reward of a transition the difference of potentials.



$\Phi(s_1) = 0 \quad \Phi(s_2) = 3 \quad \Phi(s_3) = 6 \quad \Phi(s_4) = 9 \quad \Phi(s_0) = 9$

Definition

A shaping reward function $F : S \times A \times S \to \mathbb{R}$ is **potential-based** if there exists $\Phi : S \to \mathbb{R}$ s.t.

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

for all $s \neq s_0, a, s'$.

Theorem

If F is a potential-based shaping function, then every optimal policy in $M' = \langle S, A, T, \gamma, R + F\rangle$ will also be an optimal policy in $M = \langle S, A, T, \gamma, R\rangle$ (and vice versa).

Theorem

If F is a potential-based shaping function, then every optimal policy in $M' = \langle S, A, T, \gamma, R + F \rangle$ will also be an optimal policy in $M = \langle S, A, T, \gamma, R \rangle$ (and vice versa).

Q^*_M satisfies the Bellman equation:

$$Q^*_M(s, a) = \mathbb{E}_{s' \sim P_{sa}} \left[R(s, a, s') + \gamma \max_{a' \in A} Q^*_M(s', a') \right]$$

Let's subtract $\Phi(s)$ from both sides:

$$\begin{aligned} Q^*_M(s, a) - \Phi(s) &= \mathbb{E}_{s' \sim P_{sa}} \left[R(s, a, s') + \gamma \max_{a' \in A} Q^*_M(s', a') \right] - \Phi(s) \\ &= \mathbb{E}_{s' \sim P_{sa}} \left[R(s, a, s') + \gamma \Phi(s') + \gamma \max_{a' \in A} (Q^*_M(s', a') - \Phi(s')) \right] - \Phi(s) \\ &= \mathbb{E}_{s' \sim P_{sa}} \left[R(s, a, s') + \gamma \Phi(s') - \Phi(s) + \gamma \max_{a' \in A} \left(Q^*_M(s', a') - \Phi(s') \right) \right] \end{aligned}$$

Theorem

If F is a potential-based shaping function, then every optimal policy in $M' = \langle S, A, T, \gamma, R + F\rangle$ will also be an optimal policy in $M = \langle S, A, T, \gamma, R\rangle$ (and vice versa).

So $Q^*_M(s, a) - \Phi(s)$ is equal to

$$\mathbb{E}_{s' \sim P_{sa}} \left[R(s, a, s') + \gamma\Phi(s') - \Phi(s) + \gamma \max_{a' \in A} \left(Q^*_M(s', a') - \Phi(s') \right) \right].$$

Theorem

If F is a potential-based shaping function, then every optimal policy in $M' = \langle S, A, T, \gamma, R + F \rangle$ will also be an optimal policy in $M = \langle S, A, T, \gamma, R \rangle$ (and vice versa).

So $Q^*_M(s, a) - \Phi(s)$ is equal to

$$\mathbb{E}_{s' \sim P_{sa}} \left[R(s, a, s') + \gamma \Phi(s') - \Phi(s) + \gamma \max_{a' \in A} \left(Q^*_M(s', a') - \Phi(s') \right) \right].$$

Let

$$\hat{Q}_{M'}(s, a) \coloneqq Q^*_M(s, a) - \Phi(s).$$

and recall that

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

Therefore,

$$\begin{aligned} \hat{Q}_{M'}(s, a) &= \mathbb{E}_{s' \sim P_{sa}} \left[R(s, a, s') + F(s, a, s') + \gamma \max_{a' \in A} \left(\hat{Q}_{M'}(s', a') \right) \right] \\ &= \mathbb{E}_{s' \sim P_{sa}} \left[R'(s, a, s') + \gamma \max_{a' \in A} \left(\hat{Q}_{M'}(s', a') \right) \right] \end{aligned}$$

Theorem

If F is a potential-based shaping function, then every optimal policy in $M' = \langle S, A, T, \gamma, R + F\rangle$ will also be an optimal policy in $M = \langle S, A, T, \gamma, R\rangle$ (and vice versa).

$$\begin{aligned}\hat{Q}_{M'}(s,a) &= \mathbb{E}_{s' \sim P_{sa}}\left[R(s,a,s') + F(s,a,s') + \gamma \max_{a' \in A}\left(\hat{Q}_{M'}(s',a')\right)\right] \\ &= \mathbb{E}_{s' \sim P_{sa}}\left[R'(s,a,s') + \gamma \max_{a' \in A}\left(\hat{Q}_{M'}(s',a')\right)\right]\end{aligned}$$

This is the Bellman equation for M', so

$$\hat{Q}_{M'} = Q^*_{M'}.$$

(In the undiscounted case, $s = s_0$ has to be treated as a special case.)

Corollary

Suppose $F(s, a, s') = \gamma\Phi(s') - \Phi(s)$ (and, if $\gamma = 1$, that $\Phi(s_0) = 0$). Then, for all s, a:

$$Q^*_{M'}(s, a) = Q^*_M(s, a) - \Phi(s) \qquad V^*_{M'} = V^*_M(s) - \Phi(s)$$

Remark

The identities above actually hold for any policy π:

$$Q^\pi_{M'}(s, a) = Q^\pi_M(s, a) - \Phi(s) \qquad V^\pi_{M'} = V^\pi_M(s) - \Phi(s)$$

Therefore, potential-based shaping also preserves near-optimal policies.

- Note that setting $\Phi(s) = V^*_M(s)$ would make $V^*_{M'} \equiv 0$, which would make learning easy.
- This suggests that a way to define a good potential function might be to try to approximate $V^*_M(s)$.

MDP review

Reward shaping
: To provide guidance, policies can be learned on an MDP with a modified reward function, and then used on the original MDP (with varying results).

Potential-based reward shaping
: To ensure that good policies for a modified reward function are also good for the original, it suffices to base the rewards on a **potential function**.

Experiments
: Some potential-based shaping functions are evaluated.

A grid world

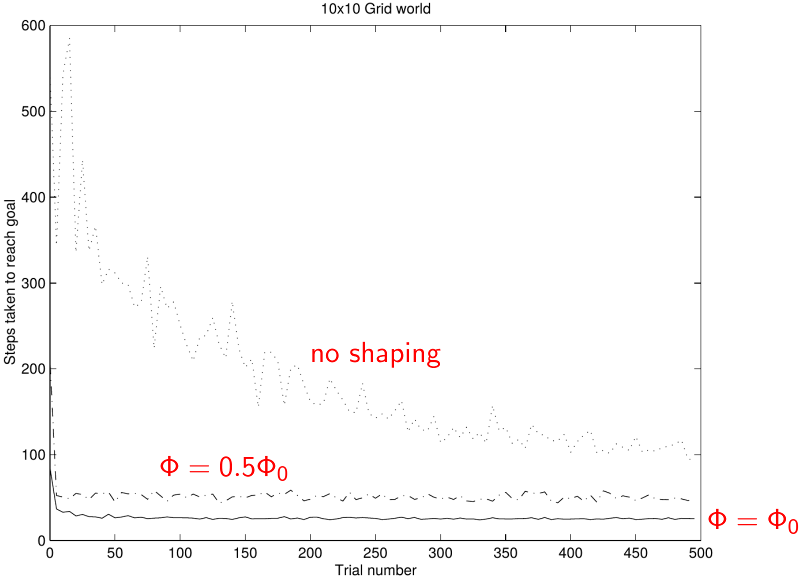
- **States**: an $n \times n$ grid, with start state and (absorbing) goal state in opposite corners.
- **Actions**: can attempt to move in any of the four cardinal directions (N, S, E, W)
- **Transition probabilities**: attempting to move in a direction succeeds with probability 0.8 and goes in a random direction otherwise
- **Discount factor**: $\gamma = 1$ (no discounting)
- **Reward function**: -1 per step

Finding a potential function to approximate V_M^*

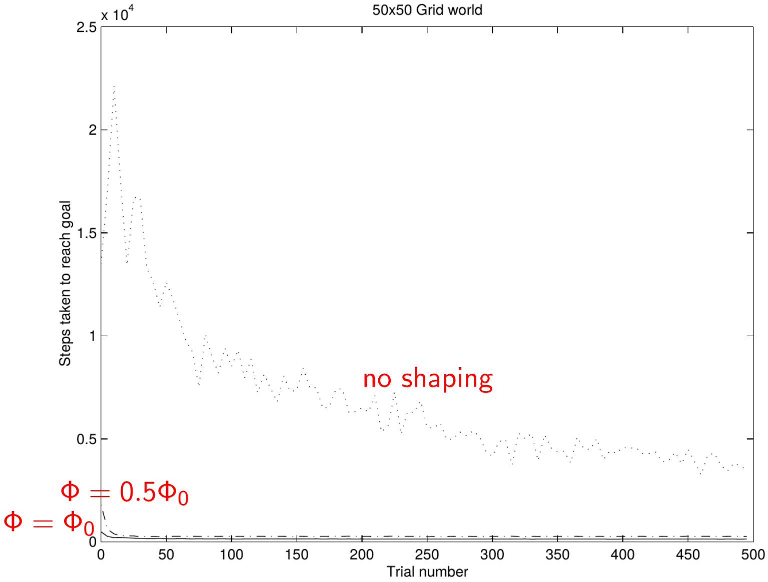
- From most states, trying to move towards the goal could be expected to make roughly 0.8 units of progress.
- Therefore, one estimate of the value function is

$$\Phi_0(s) = -\text{MANHATTAN}(s, \text{GOAL})/0.8$$

- The experiments try using Φ_0 and $0.5\Phi_0$ as potential functions.



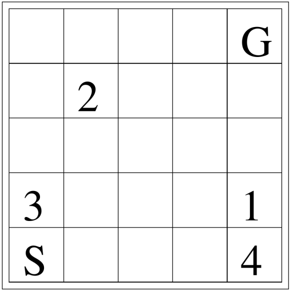
Graph from Figure 1(a) (with red labels added)



Graph from Figure 1(b) (with red labels added)

Grid world with flags

- Extend the grid world so that numbered flags have to be picked up in order.
- The state space is enlarged to keep track of the flags picked up so far.



The agent (S) needs to go to 1, 2, 3, 4, G in order.[1]

[1] Image taken from Figure 2(a)

Grid world with flags

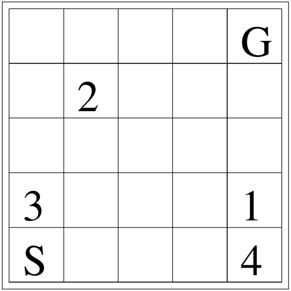
An estimate of the value function is

$$\Phi_0(s) = -\frac{(5 - n - 0.5)}{5} t$$

where

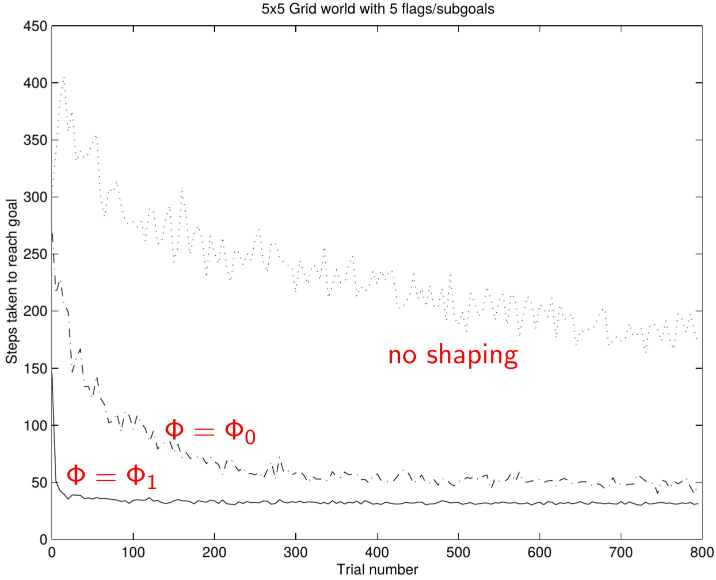
- n is the number of subgoals that have been accomplished in state s, and
- t is an estimate of the number of steps needed to reach G directly.

Experiments were done with Φ_0 and also a function Φ_1 which was a more fine-tuned estimate.



The agent (S) needs to go to 1, 2, 3, 4, G in order.[1]

[1] Image taken from Figure 2(a)



Graph from Figure 2(b) (with red labels added)

Conclusion

We've seen that

- Reward shaping can change what the optimal policy is.
- But, using potential-based shaping functions guarantees that the optimal policy will not be changed.
- The idea of potential functions can help us find useful shaping functions in practice.

References

Jette Randløv and Preben Alstrøm. Learning to drive a bicycle using reinforcement learning and shaping. In *Proceedings of the Fifteenth International Conference on Machine Learning*, ICML '98, pages 463–471, San Francisco, CA, USA, 1998. Morgan Kaufmann Publishers Inc.