Supplement: Sufficiency and Necessity Theorem for a Potential-Based Shaping Function's Policy Invariance

Justin Burzachiello, Prabal Chhatkuli, Zeru Zhu, Touhid Hossain

The following document provides supplementary information about necessary and sufficient conditions of optimal policy invariance from an MDP to its modified MDP with reward shaping. The text is largely sourced from papers written by Ng et al. and Randløv & Preben Alstrøm. It mainly serves as a quick reference to their work, which can be found in our references.

Shaping Rewards

Shaping rewards are a powerful tool for improving reinforcement learning efficiency by modifying the reward structure of an MDP. The goal is to guide learning algorithms toward optimal policies faster by incorporating an additional **shaping reward function** F**.** For an MDP $M = (S, A, T, \gamma, R)$, instead of using the original reward R, the transformed MDP $M' = (S, A, T, \gamma, R')$ uses a new reward function $R' = R + F$. Here, F is a bounded real-valued function called the **shaping reward function**, which adds extra incentives for certain transitions to influence the agent's behavior. Imagine teaching an agent how to ride a bike to some destination. Without shaping, the agent will take a random walk until it accidentally arrives at the destination where it is rewarded. One can shape the reward to guide the bike towards the target based on its angle. In particular, one can reward motion towards the target and punish motion away from the target. However, improperly designed shaping functions can lead to undesired behaviors, such as exploitation of cycles where the agent gets distracted by repeatedly visiting certain states, earning unintended rewards without progressing toward the goal. "*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.*"[2] To prevent this and ensure consistency with the original optimal policy, F must be potential based and designed to ensure net rewards for cycles are zero, avoiding distractions.

Potential-Based Reward Shaping (PBRS)

The potential-based shaping function F is defined as $F(s, a, s') = \gamma\Phi(s') - \Phi(s)$, where $\Phi(s)$ is a potential function over states. This formulation guarantees that the agent does not exploit cycles because the total shaping reward for any cycle is zero. Ng *et al.* proved the sufficiency and necessity theorem for a potential-based shaping function's policy invariance [1].

Theorem 1

Let any S, A, γ, and any shaping reward function F be given. We say F is a potential-based shaping function if there exists a function $\Phi : S \to \mathbb{R}$ such that for all $s \in S - \{s_0\}, a \in A, s' \in S$,

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

Sufficiency: If F is a potential-based shaping function, then every optimal policy in M' will also be an optimal policy in M (and vice versa).

Necessity: If F is not a potential-based shaping function, then there exist T and R such that no optimal policy in M' is optimal in M.

Proof of Sufficiency Theorem: We need to show that any policy that optimizes $Q^*_{M'}(s, a)$ also optimizes $Q^*_M(s, a)$.
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:

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

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

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

Let

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

and recall that

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

Therefore,

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

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

This is the Bellman equation for M', so we have,

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

Thus $Q^*_{M'}(s, a) = \hat{Q}_{M'}(s, a) = Q^*_M(s, a) - \Phi(s)$, and the optimal policy for M' therefore satisfies

$$\begin{aligned} \pi^*_{M'}(s) &\in \arg\max_{a \in A} Q^*_{M'}(s, a) \\ &= \arg\max_{a \in A} (Q^*_M(s, a) - \Phi(s)) \\ &= \arg\max_{a \in A} Q^*_M(s, a). \end{aligned}$$

and is therefore also optimal in M.

Now we will show that any policy that optimizes $Q^*_M(s,a)$ also optimizes $Q^*_{M'}(s,a)$. $Q^*_{M'}$ 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 add $\Phi(s)$ to both sides:

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

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

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

Let

$$\hat{Q}_M(s,a) := Q^*_{M'}(s,a) + \Phi(s)$$

and recall that

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

Therefore,

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

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

This is the Bellman equation for M, so we have,

$$\hat{Q}_M = Q^*_M.$$

Therefore, any policy that optimizes $Q^*_M(s,a)$ also optimizes $Q^*_M(s,a) = Q^*_{M'}(s,a) + \Phi(s)$. Since $\Phi(s)$ does not depend on the action chosen in state s, this policy also maximizes $Q^*_{M'}(s,a)$. That is, an optimal policy for M is also an optimal policy for M'. This completes the proof.

Proof of Necessity Theorem We can split proof into two cases:

- **Case 1:** F depends on the action, i.e., there exist actions a, a' such that

$$\Delta = F(s,a,s') - F(s,a',s') > 0.$$

- **Case 2:** F does not depend on the action, i.e., $F(s,a,s') = F(s,s')$.

Proof-Case 1: Construct M such that $P_{sa}(s') = P_{sa'}(s') = 1.0$, $R(s,a,s') = 0$, and $R(s,a',s') = \Delta/2$. Clearly, $\pi^*_M(s) = a'$.
On the other hand, since $R' = R + F$, we have $R'(s,a,s') = F(s,a,s')$ and

$$\begin{aligned} R'(s,a',s') &= \Delta/2 + F(s,a',s') \\ &= \Delta + F(s,a',s') - \Delta/2 \\ &= F(s,a,s') - F(s,a',s') + F(s,a',s') - \Delta/2 \\ &= F(s,a,s') - \Delta/2 \\ &< R'(s,a,s') \end{aligned}$$

Therefore, $\pi^*_{M'}(s) = a$, which is not the same as $\pi^*_M(s) = a'$.
Proof-Case 2: We can assume, without loss of generality, that $F(\hat{s}_o, \hat{s}_o) = 0$, where $\hat{s}_o$ is the absorbing state if $\gamma = 1$, and some fixed state otherwise. Since F is not potential-based, for any potential function Φ, there exist states s_1, s_2 such that $\gamma\Phi(s_2) - \Phi(s_1) \neq F(s_1, s_2)$, where $s_1, s_2, \hat{s}_o$ are distinct.

Construct M as follows:

- $P_{s_1 a}(\hat{s}_0) = P_{s_1 a'}(s_2) = P_{s_2 a}(\hat{s}_0) = P_{\hat{s}_0 a}(\hat{s}_0) = 1.0$
- $\Delta = F(s_1, s_2) + \gamma F(s_2, \hat{s}_0) - F(s_1, \hat{s}_0)$
- $R(s_1, a, \hat{s}_0) = \Delta/2$, otherwise $R(\cdot,\cdot,\cdot) = 0$

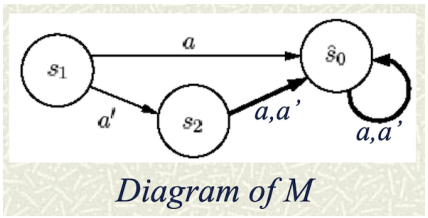

Diagram of M

Then we have

- $Q^*_M(s_1, a) = \frac{\Delta}{2} + \gamma V^*_M(\hat{s}_0) = \frac{\Delta}{2}$
- $Q^*_M(s_1, a') = 0 + \gamma V^*_M(s_2) = 0 + \gamma \cdot 0 + V^*_M(\hat{s}_0) = 0$
- $Q^*_{M'}(s_1, a) = \frac{\Delta}{2} + F(s_1, \hat{s}_0) = \Delta + F(s_1, \hat{s}_0) - \frac{\Delta}{2}$

$$= F(s_1, s_2) + \gamma F(s_2, \hat{s}_0) - F(s_1, \hat{s}_0) + F(s_1, \hat{s}_0) - \frac{\Delta}{2}$$

$$= F(s_1, s_2) + \gamma F(s_2, \hat{s}_0) - \frac{\Delta}{2}.$$

- $Q^*_{M'}(s_1, a') = 0 + F(s_1, s_2) + \gamma V^*_M(\hat{s}_0) = F(s_1, s_2) + \gamma F(s_2, \hat{s}_0)$.

Thus, we have

$$\pi^*_M(s_1) = \begin{cases} a & \text{if } \Delta > 0, \\ a' & \text{otherwise} \end{cases}$$

$$\pi^*_{M'}(s_1) = \begin{cases} a' & \text{if } \Delta > 0, \\ a & \text{otherwise} \end{cases}$$

Therefore, $\pi^*_M(s) = a$, which is not the same as $\pi^*_{M'}(s) = a'$. This completes the proof.

References

[1] Andrew Y Ng, Daishi Harada, and Stuart Russell. "Policy invariance under reward transformations: Theory and application to reward shaping". In: *Icml*. Vol. 99. 1999, pp. 278–287.

[2] Jette Randløv and Preben Alstrøm. "Learning to Drive a Bicycle Using Reinforcement Learning and Shaping." In: *ICML*. Vol. 98. 1998, pp. 463–471.