Distributional Reward Shaping: Point Estimates Are All You Need

Michael Gimelfarb **Scott Sanner** **Chi-Guhn Lee**
Department of Mechanical and Industrial Engineering
University of Toronto
Toronto, Canada, M5S 3G8
{mike.gimelfarb,ssanner,cglee}@{mail,mie,mie}.utoronto.ca

Abstract

Potential-based reward shaping is a powerful approach for incorporating value-based advice in order to accelerate the convergence of reinforcement learning algorithms on problems with sparse reward. We propose the idea of distributional reward shaping, in which the shaping signal is a probability distribution over hypothetical returns in each state-action pair. A natural setting in which such advice could be useful is the distributional reinforcement learning (DRL) that has recently provided state-of-the-art results on a number of benchmark problems. However, it is largely unclear how to incorporate distributional advice while maintaining policy invariance guarantees as in standard RL. To this end, our first contribution is to show that distributional reward shaping maintains policy invariance if the policy is derived by maximization of the expected return. By drawing on several examples from the literature, our second contribution is to illustrate that such results do not hold generally in the risk-sensitive RL setting, in which the agent optimizes a non-linear utility function of the return. However, we show that the utility of the distributional reward shape provides an ideal deterministic reward signal, that does not require making independence assumptions nor limiting the class of utility functions that can be used.

Keywords: reward shaping, transfer learning, distributional reinforcement learning, risk-sensitive, risk-aware, safety

1 Introduction

Potential-based reward shaping (PBRS) is a powerful technique for transforming a reinforcement learning problem with a sparse reward into one with a dense reward without changing the optimal policies [Ng et al., 1999]. Recent literature has demonstrated its ability and flexibility in transferring knowledge from a variety of sources [Brys et al., 2015a,b, Suay et al., 2016]. However, none of the existing work has studied the problem of transferring knowledge that occurs naturally in the form of a probability distribution over hypothetical return outcomes. Such "distributional" advice could be particularly useful as a means of communicating notions of danger or uncertainty between learning agents, or from humans to agents, and represents a richer vocabulary for transferring value-based advice than point estimates alone. One practical setting is the training of an RL agent using the *distributional RL* (DRL) framework [Bellemare et al., 2017], and using the learned value function distribution as the input to another agent who later solves a different (but somewhat related) task, perhaps measuring risk in a different way. The goal of this abstract is two-fold. First, we prove that policy invariance holds for DRL agents that optimize expected returns (Theorem 1). Second, we demonstrate that such results do not generalize directly to the risk-sensitive setting (Example 1, 2). Instead, the utility of the distributional advice provides a PBRS signal that preserves invariance, as long as the utility function is cash-invariant, and also has favourable mathematical properties (Lemma 1).

2 Preliminaries

2.1 Distributional Reinforcement Learning

Given a *Markov decision process* (MDP) with state space $\mathcal{S}$, actions $\mathcal{A}$, reward function r, transition function P, and discount factor $\gamma \in (0, 1)$, the goal of DRL is to learn the probability distribution $Z^*(s, a) \in \mathcal{Z}$ in a set of bounded random variables $\mathcal{Z}$ by solving the distributional Bellman equation (DBE):

$$Z^*(s, a) = r(s, a, S') + \gamma Z^*(S', \pi^*(S')), \quad \pi^*(s') \in \operatorname{argmax}_{a'} \mathbb{E}_{Z^*}[Z^*(s', a')].$$

Bellemare et al. [2017] approximated $Z^*(s, a)$ by a histogram with equidistant points, and learned a parametric representation $Z_\theta(s, a)$ by minimizing the KL-divergence between $Z_\theta(s, a)$ and the distribution of the single-step Bellman update. Furthermore, Bellemare et al. [2017] showed that the DBE operator is a contraction mapping under the p-Wasserstein metric. More recently, Dabney et al. [2018b] proposed QR-DQN to estimate $Z(s, a)$ by applying the discretization to the quantiles of the distribution rather than the outcomes, and used quantile regression to learn the approximate quantiles of $Z^*(s, a)$.

Dabney et al. [2018a] extended the framework of QR-DQN to learn risk-sensitive policies. Given a utility function $U : \mathbb{R} \to \mathbb{R}$, the goal of *risk-aware RL* is to learn a policy that optimizes the expected utility of the return associated with the Bellman equation:

$$Z^*(s, a) = r(s, a, S') + \gamma Z^*(S', \pi^*(S')), \quad \pi^*(s') \in \arg\max_{a'} \mathbb{E}_{Z^*}[U(Z^*(s', a'))].$$

IQN restricted U to the class of *distortion risk measures*, that intuitively can be seen as computing the expected value of the return under a distorted distribution of $Z(s, a)$. More formally, U is associated with a distortion function $\beta : [0, 1] \to [0, 1]$, such that $\mathbb{E}_Z[U(Z(s, a))] = \mathbb{E}_{\tau \sim U([0,1])}[Z_{\beta(\tau)}(s, a)]$ where $Z_\tau(s, a) = F^{-1}_{Z(s,a)}(\tau)$ denotes the quantile function of $Z(s, a)$.

2.2 Utility Function

In our developments, we further abstract the idea of expected utility by considering the general class of *concave utility* functions $\mathcal{U} : \mathcal{Z} \to \mathbb{R}$ that satisfy [Föllmer and Schied, 2002]:

A1. $\mathcal{U}[0] = 0$

A2. if $Z_1, Z_2 \in \mathcal{Z}$ such that $\mathbb{P}(Z_1 \geq Z_2) = 1$, then $\mathcal{U}[Z_1] \geq \mathcal{U}[Z_2]$

A3. if $c \in \mathbb{R}$ and $Z \in \mathcal{Z}$, then $\mathcal{U}[Z + c] = \mathcal{U}[Z] + c$

A4. if $Z_1, Z_2 \in \mathcal{Z}$ then $\mathcal{U}[Z_1 + Z_2] \geq \mathcal{U}[Z_1] + \mathcal{U}[Z_2]$,

which can be seen as necessary criteria for rational and risk-averse decision-making. We also assume:

A5. if $Z_t \in \mathcal{Z}$ converges in distribution to $Z \in \mathcal{Z}$ with $\min Z_t \to \min Z$ and $\max Z_t \to \max Z$, then $\mathcal{U}[Z_t] \to \mathcal{U}[Z]$,

which ensures that the optimization criterion $\arg\max_a \mathcal{U}[Z_t(s, a)]$ remains a meaningful policy as Z_t converges to the true return distribution.

2.3 Reward Shaping

In general, the successful application of a reinforcement learning algorithm – and by extension DRL – depends largely on the quality of the chosen reward function. Given a reward function $r : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$ and a shaping function $F : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$, reward shaping produces an augmented reward function $r'(s, a, s') = r(s, a, s') + F(s, a, s')$, that in many cases can allow an agent to converge faster towards the optimal policy than r. However, care is necessary to ensure that the set of optimal policies remains *invariant* with respect to the above change in the reward function [Randløv and Alstrøm, 1998]. Ng et al. [1999] has shown that, when the shaping function is potential-based, e.g. $F(s, a, s') = \gamma\Phi(s') - \Phi(s)$ for some function $\Phi : \mathcal{S} \to \mathbb{R}$, then policy-invariance is guaranteed.

3 Distributional Reward Shaping

In this section, we discuss the distributional setting in which the advice and the agent are both return distributions in $\mathcal{Z}$.

3.1 Distributional Policy Invariance in DRL

To establish a characterization of policy invariance in DRL, we first study the risk-neutral setting $\mathcal{U} = \mathbb{E}$, where advice is given as a probability distribution over hypothetical return.

Theorem 1. *Let $\Phi : \mathcal{S} \to \mathcal{Z}$ be independent at each evaluation. If $\mathcal{U} = \mathbb{E}$, then PBRS leaves the optimal policies unchanged.*

Proof. We begin with the Bellman equation for an MDP $\mathcal{M}$ with dynamics P and reward function r,

$$Z_t(s, a) = r(s, a, S') + \gamma Z_{t+1}(S', \arg\max_{a'} \mathcal{U}[Z_{t+1}(S', a')]) := \mathcal{T} Z_{t+1}(s, a),$$

and subtract $\Phi(s)$ from both sides to obtain

$$Z_t(s, a) - \Phi(s) = r(s, a, S') + \gamma\Phi(S') - \Phi(s) + \gamma \left(Z_{t+1}(S', \arg\max_{a'} \mathcal{U}[Z_{t+1}(S', a')]) - \Phi(S') \right). \tag{1}$$

Here and throughout, the equality is to be understood as equivalence in distribution. Next, define the random variable $Z'_t(s, a)$ and $r'_t(s, a, s')$ such that:

$$Z'_t(s, a) := Z_t(s, a) - \Phi(s) \tag{2}$$

$$r'(s, a, S') := r(s, a, S') + \gamma\Phi(S') - \Phi(s). \tag{3}$$

Substituting the last two equations into (1), we obtain

$$Z'_t(s, a) = r'(s, a, S') + \gamma Z'_{t+1}(S', \arg\max_{a'} \mathcal{U}[Z_{t+1}(S', a')]).$$

Next, given **A3**, (2) and linearity of expectation:

$$\begin{aligned}
\pi_{t+1}(s') &\in \arg\max_{a'} \mathcal{U}[Z_{t+1}(s', a')] \\
&= \arg\max_{a'} \left\{ \mathcal{U}[Z'_{t+1}(s', a')] + \mathcal{U}[\Phi(s')] \right\} \\
&= \arg\max_{a'} \mathcal{U}[Z'_{t+1}(s', a')] := \pi'_{t+1}(s'),
\end{aligned}$$

and hence:

$$\begin{aligned}
Z'_t(s, a) &= r'(s, a, S') + \gamma Z'_{t+1}(S', \arg\max_{a'} \mathcal{U}[Z_{t+1}(S', a')]) \\
&= r'(s, a, S') + \gamma Z'_{t+1}(S', \arg\max_{a'} \mathcal{U}[Z'_{t+1}(S', a')]) := \mathcal{T}' Z'_{t+1}(s, a),
\end{aligned}$$

is the distributional Bellman operator of the MDP with immediate reward function r'. Finally, we observe that the roles of the two MDPs with reward r and r' can be interchanged, and so the optimal policies are preserved, as claimed. □

Observe that (2) foreshadows a result in Ng et al. [1999], namely that PBRS shifts the return in each state-action pair down by the value of the potential $\Phi(s)$ in each state of the MDP. However, a critical distinction in the DRL setting is that PBRS leads to a *convolution* of $Z_t(s, a)$ and $\Phi(s)$. Also, (3) suggests that immediate rewards r' can be stochastic even after observing $s' \sim S'$. This does not complicate our analysis, however, since the stochasticity can be absorbed directly into the estimate $Z_t(s, a)$.

3.2 The Failure of Policy Invariance in Risk-Sensitive DRL

A natural question to ask is whether policy invariance still holds for arbitrary concave utility functions. Unfortunately, this turns out to be false in general, since the analysis above shows that $\mathcal{U}$ must be strictly additive, e.g.

$$\mathcal{U}[Z_t(s,a) + \Phi(s)] = \mathcal{U}[Z_t(s,a)] + \mathcal{U}[\Phi(s)] \tag{4}$$

for all choices of Z_t and Φ. In more intuitive terms, *policy invariance holds if the potential function does not permit the learning agent to improve the overall risk through diversification among two lotteries: (1) the agent's current knowledge of the return $Z_t(s,a)$, and (2) the expert's current knowledge of the return $\Phi(s)$*. Without further knowledge about the dependence between $Z_t(s,a)$ and $\Phi(s)$, it is therefore not possible to establish policy invariance except when $\mathcal{U} = \mathbb{E}$. To make this point clearer, we illustrate how this would require making careful choices for $\mathcal{U}$ that depend on how $Z_t(s,a)$ and $\Phi(s)$ are jointly distributed.

Example 1. Suppose that $Z_t(s,a)$ and $\Phi(s)$ are independent for each (s,a), and consider the *entropic utility* function $\mathcal{U}_\beta$

$$\mathcal{U}_\beta[Z] = \frac{1}{\beta} \log \mathbb{E}[e^{\beta Z}],$$

where $\beta \in \mathbb{R}$ is an arbitrary control parameter, and which satisfies **A1-A4**. Furthermore, we also define the *weighted entropic utility* $\mathcal{U}_w$ as

$$\mathcal{U}_w[Z] = \int \mathcal{U}_\beta[Z]\, w(\beta)\, \mathrm{d}\beta$$

where $w : \mathbb{R} \to [0, \infty)$, and which satisfies **A1-A4** provided that $\int w(\beta)\, \mathrm{d}\beta = 1$. It is also easy to verify that:

$$\mathcal{U}_w[Z_t(s,a) + \Phi(s)] = \mathcal{U}_w[Z_t(s,a)] + \mathcal{U}_w[\Phi(s)].$$

Furthermore, it can be shown that $\mathcal{U}_w$ is the *only* class of utility functions that satisfies **A1-A4** and **A5** under the independence assumption [Goovaerts et al., 2004].

While the form of $\mathcal{U}$ in the aforementioned example appears restrictive, we note that entropic utility essentially guarantees policy invariance as long as $Z_t(s,a)$ and $\Phi(s)$ are simulated from uncoupled stochastic processes. This setting is quite natural, for instance, when $\Phi(s)$ and $Z_t(s,a)$ are learned on different tasks using IQN in an end-to-end framework, since samples $\Phi(s) \sim \max_a Z_{\beta(\tau)}(s,a)$ and $G_t \sim Z_{\beta(\tau')}(s,a)$ are typically generated according to i.i.d. samples $\tau, \tau' \sim U([0,1])$. In many instances however, such as when the immediate rewards or the return realizations from different tasks are directly correlated [Zhang et al., 2021], the entropic utility is no longer a viable quantifier of risk.

Example 2. A converse example to independence is perfect dependence or *commonoticity*, in which $Z_t(s,a) = F^{-1}_{Z_t(s,a)}(\tau)$ and $\Phi(s) = F^{-1}_{\Phi(s)}(\tau)$ for the same realization of $\tau \sim U([0,1])$, where equality holds in distribution. We define $Q_X(p)$ to denote the p-quantile function of X, and the *tail value-at-risk*

$$\mathrm{TVaR}_p[X] = \frac{1}{1-p} \int_p^1 Q_X(q)\, \mathrm{d}q.$$

This utility function is additive for commonotone random variables [Dhaene et al., 2006], so (4) holds.

Taken together, the previous two examples have thus shown that *policy invariance holds only when the dependence structure between the model returns of the agent and the expert are precisely known, and the utility function $\mathcal{U}$ is chosen carefully*. In other words, ignoring the hidden interaction between the agents can lead to unintended behaviours in end-to-end or multi-task approaches. Another shortcoming of the distributional approach to shaping is that it requires computing the convolution of several probability distributions for each sample and can be computationally demanding. Finally, the choice of utility should be dictated by the external environment (e.g. regulators), rather than solely from the specificity of each task.

3.3 Point Estimates are All You Need

If arbitrary distributional advice cannot be accounted for directly without knowing the structure of Z_t and Φ, we ask whether it can be incorporated into PBRS in other ways. Given an arbitrary distribution-valued potential $\Phi_t \in \mathcal{Z}$, one such candidate that is consistent with the agent's decision-making criteria is $\phi(s) = \mathcal{U}[\Phi(s)]$. With this simplification, (4) holds immediately due to cash invariance, and thus $\phi(s)$ provides a policy-invariant reward signal.

The next natural question to ask is: what kind of advice allows DRL to learn *efficiently*? We can answer this question briefly by showing that the nature of the ideal *deterministic* potential Φ takes the form of the utility of the return distribution Z^*, with similar results having been shown only for standard RL [Zou et al., 2021].

Lemma 1. *For an MDP with return distribution Z^*, the PBRS signal $\phi(s) = \max_a \mathcal{U}[Z^*(s,a)]$ defines an MDP $\mathcal{M}'$ with a dense reward whose optimal utility is zero along the optimal trajectory $\pi^*(s)$ and negative everywhere else.*

Proof. The consequences of Theorem 1 hold for the deterministic signal $\phi(s)$, and so the optimal policy remains unchanged. Furthermore, according to (2) and **A3**:

$$\mathcal{U}[Z'(s,a)] = \mathcal{U}[Z^*(s,a) + (-\phi(s))] = \mathcal{U}[Z^*(s,a)] - \phi(s) = \mathcal{U}[Z^*(s,a)] - \max_a \mathcal{U}[Z^*(s,a)] \leq 0,$$

where equality holds only if $a \in \arg\max_a \mathcal{U}[Z^*(s,a)] = \pi^*(s)$, as claimed. □

3.4 Generalization to State-Action Potentials

As a final observation, we note that it is easy to extend our previous analysis to the setting in which $\Phi_t(s,a)$ is time and action-dependent, by considering the look-ahead reward shaping function

$$F_t(s,a,s',a') = \gamma\Phi_{t+1}(s',a') - \Phi_t(s,a),$$

where a is the action chosen in state s, a' is the corresponding action chosen in state s', and so forth. By extending the derivations of Wiewiora et al. [2003], Devlin and Kudenko [2012] to our setting, it is easy to check that

$$\mathcal{U}[Z'_t(s,a)] = \mathcal{U}[Z_t(s,a) - \Phi_t(s,a)].$$

Thus, under assumptions **A1-A5**, policy invariance holds once again as long as the actions in the new MDP are chosen according to

$$\pi'_t(s) \in \operatorname{argmax}_a \left\{\mathcal{U}[Z_t(s,a)] + \mathcal{U}[\Phi_t(s,a)]\right\},$$

where $\phi(s,a) = \mathcal{U}[Z^*(s,a)]$ satisfies the usual guarantees for PBRS.

4 Conclusion

We demonstrated that PBRS can be generalized to a distribution over returns, and that it preserves policy invariance in standard DRL. A negative result is illustrated, namely that policy invariance in the risk-sensitive setting cannot be guaranteed for arbitrary non-linear utility functions of the return. Instead, to accommodate many rational choices of utility functions such as CVaR, we propose to map the distributional advice to a point estimate by using the given utility function, in which case the resulting PBRS signal is risk-sensitive, policy-invariant, and is optimal in certain circumstances.

References

Marc G Bellemare, Will Dabney, and Rémi Munos. A distributional perspective on reinforcement learning. In *ICML*, pages 449–458. PMLR, 2017.

Tim Brys, Anna Harutyunyan, Halit Bener Suay, Sonia Chernova, Matthew E Taylor, and Ann Nowé. Reinforcement learning from demonstration through shaping. In *AAAI*, 2015a.

Tim Brys, Anna Harutyunyan, Matthew E Taylor, and Ann Nowé. Policy transfer using reward shaping. In *AAMAS*, pages 181–188, 2015b.

Will Dabney, Georg Ostrovski, David Silver, and Rémi Munos. Implicit quantile networks for distributional reinforcement learning. In *ICML*, pages 1096–1105. PMLR, 2018a.

Will Dabney, Mark Rowland, Marc G Bellemare, and Rémi Munos. Distributional reinforcement learning with quantile regression. In *AAAI*, 2018b.

Sam Michael Devlin and Daniel Kudenko. Dynamic potential-based reward shaping. In *AAMAS*, pages 433–440. IFAAMAS, 2012.

Jan Dhaene, Steven Vanduffel, Marc Goovaerts, Rob Kaas, Qihe Tang, and David Vyncke. Risk measures and comonotonicity: A review. *Stochastic Models*, 22:573–606, 12 2006. doi: 10.1080/15326340600878016.

Hans Föllmer and Alexander Schied. Convex measures of risk and trading constraints. *Finance and stochastics*, 6(4):429–447, 2002.

Marc J Goovaerts, Rob Kaas, Roger JA Laeven, and Qihe Tang. A comonotonic image of independence for additive risk measures. *Insurance: Mathematics and Economics*, 35(3):581–594, 2004.

Andrew Y Ng, Daishi Harada, and Stuart Russell. Policy invariance under reward transformations: Theory and application to reward shaping. In *ICML*, volume 99, pages 278–287, 1999.

Jette Randløv and Preben Alstrøm. Learning to drive a bicycle using reinforcement learning and shaping. In *ICML*, volume 98, pages 463–471. Citeseer, 1998.

Halit Bener Suay, Tim Brys, Matthew E Taylor, and Sonia Chernova. Learning from demonstration for shaping through inverse reinforcement learning. In *AAMAS*, pages 429–437, 2016.

Eric Wiewiora, Garrison W Cottrell, and Charles Elkan. Principled methods for advising reinforcement learning agents. In *ICML*, pages 792–799, 2003.

Pushi Zhang, Xiaoyu Chen, Li Zhao, Wei Xiong, Tao Qin, and Tie-Yan Liu. Distributional reinforcement learning for multi-dimensional reward functions. *NeurIPS*, 34, 2021.

Haosheng Zou, Tongzheng Ren, Dong Yan, Hang Su, and Jun Zhu. Learning task-distribution reward shaping with meta-learning. In *AAAI*, volume 35, pages 11210–11218, 2021.