Utilizing Prior Solutions for Reward Shaping and Composition in Entropy-Regularized Reinforcement Learning

Jacob Adamczyk[1], **Argenis Arriojas**[1], **Stas Tiomkin**[2], **Rahul V. Kulkarni**[1]
[1]Department of Physics, University of Massachusetts Boston
[2]Department of Computer Engineering, San José State University
jacob.adamczyk001@umb.edu, arriojasmaldonado001@umb.edu, stas.tiomkin@sjsu.edu, rahul.kulkarni@umb.edu

Abstract

In reinforcement learning (RL), the ability to utilize prior knowledge from previously solved tasks can allow agents to quickly solve new problems. In some cases, these new problems may be approximately solved by composing the solutions of previously solved primitive tasks (task composition). Otherwise, prior knowledge can be used to adjust the reward function for a new problem, in a way that leaves the optimal policy unchanged but enables quicker learning (reward shaping). In this work, we develop a general framework for reward shaping and task composition in entropy-regularized RL. To do so, we derive an exact relation connecting the optimal soft value functions for two entropy-regularized RL problems with different reward functions and dynamics. We show how the derived relation leads to a general result for reward shaping in entropy-regularized RL. We then generalize this approach to derive an exact relation connecting optimal value functions for the composition of multiple tasks in entropy-regularized RL. We validate these theoretical contributions with experiments showing that reward shaping and task composition lead to faster learning in various settings.

Introduction

Reinforcement learning (RL) is a widely-used approach for training artificial agents to acquire complex behaviors and to engage in long-term decision making (Sutton and Barto 2018). Despite its great successes for goal-oriented tasks (e.g. board games such as chess and Go (Silver et al. 2018)), RL approaches do not fare as well when the tasks change and become more complex. The underlying problem is that RL algorithms are often incapable of effectively reusing previously-acquired knowledge; as a consequence RL agents typically start from scratch when faced with new tasks and require vast amounts of training experience to learn new solutions. Therefore, a key challenge in the field is the development of RL approaches and algorithms which are able to leverage the solutions of previous tasks for quickly solving a wide variety of new tasks. Developing approaches that enable such "transfer learning" is one of the problems we wish to address in the current work, in the context of *entropy-regularized* reinforcement learning (Ziebart 2010).

A promising approach for transfer learning is based on composing solutions for previously solved tasks to obtain solutions for new tasks. The ability to combine primitive skills to learn more complex behaviors can lead to an exponential increase in the number of new problems that an agent is able to solve (Nangue Tasse, James, and Rosman 2020; Tasse, James, and Rosman 2021). Correspondingly, there is significant interest in this idea of compositionality of tasks in RL. The observation that entropy-regularized RL provides robust solutions (Eysenbach and Levine 2022) and simple approaches for composing previous solutions (Haarnoja et al. 2018a; Van Niekerk et al. 2019; Peng et al. 2019) in specific situations has led to increased interest in this topic. Several exact results for compositionality in entropy-regularized RL have been obtained, however these are based on highly limiting assumptions on the differences between the primitive tasks. The development of more general approaches to compositionality in entropy-regularized RL is currently an important challenge in the field.

Another challenge often encountered by RL agents solving new tasks is the problem of sparse reward signals. For example, if an agent only gains a reward at the end of a long and otherwise non-rewarding trajectory, it may be difficult to learn the optimal policy in this case since the agent must be sufficiently "far-sighted". The field of reward shaping; wherein rewards are changed in a way that leaves the optimal policy invariant, has been used to address this issue (Ng, Harada, and Russell 1999). These efforts have primarily focused on the standard RL framework; to the best of our knowledge, the corresponding results for reward shaping in entropy-regularized RL have not yet been derived. Another related open question that motivates this work is understanding how we can utilize previously obtained solutions to implement reward shaping in entropy-regularized RL.

To address these issues, we focus on the core problem of deriving relations between optimal value functions for two tasks in entropy-regularized RL. Considering the first task as solved and the second as a new unsolved task, the derived relation defines a third task whose optimal value function allows us to solve the new task while leveraging prior knowledge. We show that the optimal policy for the task of interest is the same as the third task's. This observation leads to the derivation of a general result for reward shaping in entropy-regularized RL. Based on these observations connecting the optimal value functions of two tasks, we derive principled methods of approaching both task composition and reward shaping in entropy-regularized RL. In doing so, we also extend the results of (Hunt et al. 2019) for arbitrary functional transformations of rewards, and show that the theory of potential-based reward shaping of (Ng, Harada, and Russell 1999) also applies in the entropy-regularized RL formulation. By using the derived connection between optimal value functions, we also determine how a solution can remain optimal under new dynamics. Moreover, our results motivate a methodology for using previously acquired skills to learn and shape new entropy-regularized RL tasks.

Prior Work

There is a significant body of literature studying the problem of reward shaping in standard (un-regularized) reinforcement learning. This field was initiated by (Ng, Harada, and Russell 1999), whose work introduced the concept of *potential-based reward shaping* (PBRS). It was shown that PBRS functions are necessary and sufficient to describe the set of reward functions which yield the same optimal policies. In addition, the authors show that shaping is *robust* (in the sense that near-optimal policies are also invariant) and amenable to usage of prior knowledge. Therefore, solutions to previous tasks, expert knowledge, or even heuristics can be used to "shape" an agent's reward function in order to make a particular RL task easier to solve. One goal of the present work is to extend the results of (Ng, Harada, and Russell 1999) to the domain of entropy-regularized RL.

Although there exist many forms of transfer learning in RL (Taylor and Stone 2009), we shall focus on the case of concurrent skill composition by a single agent. In this work, composition refers to the combination of previous tasks through their reward functions. Task composition was introduced by (Todorov 2009) in the entropy-regularized setting and was later advanced by (Haarnoja et al. 2018a). Since composition combines previously solved tasks' reward functions in some specified functional form, it is natural to assume that those previous solutions might also be combined in the same way to obtain an approximate solution to the new task. Indeed, in standard (un-regularized) RL, it was shown by (Nangue Tasse, James, and Rosman 2020) that this holds in the case of Boolean compositions. However, for these equalities to hold, there are strong restrictions on the reward functions for the previous tasks: they may only differ on the absorbing (also known as terminal) states. In our work, we shall consider reward functions which can vary globally (over all states and actions) in entropy-regularized RL.

Previous work has shown that, in entropy-regularized RL, applying the same transformation to the solutions of previously solved tasks leads to a useful first-order approximation, depending on the specific transformation. In (Hunt et al. 2019) the authors have derived a specific correction function that can be learned and used to correct this first-order approximation to exactly solve the new task. In this work we shall extend this result (Theorem 3.2 of (Hunt et al. 2019)) to arbitrary functions (rather than only convex linear combinations) of reward functions.

Preliminaries

We consider the Markov Decision Process (MDP) model to study the entropy-regularized reinforcement learning problem. The MDP, denoted $\mathcal{M}$, consists of a state space $\mathcal{S}$, action space $\mathcal{A}$, transition dynamics $p\colon \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0,1]$, (bounded) reward function $r\colon \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$, discount factor $\gamma < 1$ and inverse temperature $\beta > 0$. We represent the MDP as a tuple $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, p, r, \gamma \rangle$. The discount factor $\gamma \in (0,1)$ discounts future rewards and assures convergence of the accumulated reward for an infinitely long trajectory ($T \to \infty$). In many instances we will also specify the particular *prior policy* $\pi_0\colon \mathcal{S} \times \mathcal{A} \to (0,1)$, a probability distribution over actions, specifying an initial exploration, data-collection, or behavior policy. We assume that any trial policy π is absolutely continuous with respect to the prior policy, which ensures that the Kullback-Liebler divergence in Equation (1) is well-defined and bounded. Although implicit in some cases, we always assume an MDP's reward function is bounded.

The entropy-regularized framework for reinforcement learning augments the standard reward-maximization objective with an entropic regularization term, relative to a reference policy π_0:

$$J(\pi) = \mathop{\mathbb{E}}_{p,\pi}\left[\sum_{t=1}^{T} \gamma^{t-1}\left(r_t - \frac{1}{\beta}\log\left(\frac{\pi(a_t|s_t)}{\pi_0(a_t|s_t)}\right)\right)\right] \quad (1)$$

This objective leads to optimal policies which remain partially exploratory (depending on the entropic term's weight, β^{-1}) and robust under perturbations to the rewards and dynamics (Eysenbach and Levine 2022). Therefore, entropy-regularized RL presents a useful method for applying reinforcement learning in real-world settings where the dynamics and reward functions may not be known with full precision.

Definition (Entropy-Regularized Task). An entropy-regularized task (or simply **task**) is an MDP $\mathcal{M}$ together with an inverse temperature β and a prior policy π_0. We denote a task by $\mathcal{T} = \mathcal{M} \cup \langle \beta, \pi_0 \rangle = \langle \mathcal{S}, \mathcal{A}, p, r, \gamma, \beta, \pi_0 \rangle$.

In the following sections, we assume the existence of a previously solved task $\mathcal{T}$ (or set of tasks $\{\mathcal{T}\}$), representing the agent's primitive knowledge. The "solution" to the task $\mathcal{T}$ refers to the optimal soft action-value function (or simply the optimal value function), Q^*, satisfying the soft Bellman optimality equation:

$$Q^*(s,a) = \mathop{\mathbb{E}}_{s'\sim p}\left[r(s,a,s') + \frac{\gamma}{\beta}\log \mathop{\mathbb{E}}_{a'\sim\pi_0} e^{\beta Q^*(s',a')}\right] \quad (2)$$

which can be solved by iterating a Bellman backup equation until convergence:

$$Q^{(N+1)}(s,a) = \mathop{\mathbb{E}}_{s'\sim p}\left[r(s,a,s') + \frac{\gamma}{\beta}\log \mathop{\mathbb{E}}_{a'\sim\pi_0} e^{\beta Q^{(N)}(s',a')}\right] \quad (3)$$

where $Q^{(0)}(s,a)$ is an arbitrary initialization function. Since the soft Bellman operator $\mathcal{B}$ is a contraction (Haarnoja et al. 2017), any bounded initialization function $Q^{(0)}$ will converge to the optimal value function: $\lim_{k\to\infty} \mathcal{B}^k Q^{(0)} = Q^*$.

When discussing the "solution" of the task $\mathcal{T}$ we also refer to the optimal policy derived from Q^*:

$$\pi^*(a|s) = \frac{\pi_0(a|s)e^{\beta Q^*(s,a)}}{\sum_{a'} \pi_0(a'|s)e^{\beta Q^*(s,a')}} \tag{4}$$

as well as the optimal state-value function:

$$V^*(s) = \frac{1}{\beta} \log \sum_a \pi_0(a|s)e^{\beta Q^*(s,a)} \tag{5}$$

In the following sections we study a solution's dependence on the underlying task's characteristics (namely, its reward function and dynamics). We show that these considerations naturally lead to the subjects of reward shaping and compositionality.

Change of Rewards

We begin by supposing that an agent has a solution to a single task, $\mathcal{T}$. The agent is then asked to solve a new problem, where only the reward function has changed. Beyond simply solving a new problem, this change in rewards may be caused by an adversary, perturbation, or general transform in the same domain. For example, suppose we have learned to reach a goal state in a maze, but now we are tasked with moving to a new goal in the same maze; as in Figure 1.

We formulate this problem by considering the two tasks $\mathcal{T} = \langle \mathcal{S}, \mathcal{A}, p, r, \gamma, \beta, \pi_0 \rangle$ and $\widetilde{\mathcal{T}} = \langle \mathcal{S}, \mathcal{A}, p, \widetilde{r}, \gamma, \beta, \pi_0 \rangle$ differing only on their reward functions. We take this opportunity to introduce the following definition.

Definition (Reward Varying Tasks). Consider a set of tasks $\{\mathcal{T}^{(k)}\}_{k=1}^N$. If the tasks only vary on their reward functions; that is, they are of the form $\mathcal{T}^{(k)} = \langle \mathcal{S}, \mathcal{A}, p, r^{(k)}, \gamma, \beta, \pi_0 \rangle$, then we say the set of tasks $\{\mathcal{T}^{(k)}\}$ is **reward varying**.

In other words, we restrict our attention to those tasks which share the same state and action spaces, transition dynamics, discount factor, temperature, and prior policy.

With these definitions in place, we first address the following question: Assuming tasks $\mathcal{T}$ and $\widetilde{\mathcal{T}}$ are reward varying, how can we utilize the solution of $\mathcal{T}$ when solving $\widetilde{\mathcal{T}}$? The answer to this question is provided by the following theorem. (The proofs for all results are provided in the Appendix of the full paper (Adamczyk et al. 2022).)

Theorem 1. *Let a task $\mathcal{T}$ with reward function r be given, with the optimal value function Q^* and corresponding optimal policy π^*. Consider a reward varying task, $\widetilde{\mathcal{T}}$ with reward function $\widetilde{r}$, with an unknown optimal action-value function, $\widetilde{Q}^*$. Define $\kappa(s,a,s') \doteq \widetilde{r}(s,a,s') - r(s,a,s')$. Denote the optimal action-value function K^* as the solution of the following Bellman optimality equation*

$$K^*(s,a) = \mathop{\mathbb{E}}_{s' \sim p}\left[\kappa(s,a,s') + \frac{\gamma}{\beta} \log \mathop{\mathbb{E}}_{a' \sim \pi^*} e^{\beta K^*(s',a')}\right] \tag{6}$$

and its corresponding state-value function

$$V_K^*(s) = \frac{1}{\beta} \log \sum_a \pi^*(a|s)e^{\beta K^*(s,a)} \tag{7}$$

Then,

$$\widetilde{Q}^*(s,a) = Q^*(s,a) + K^*(s,a) \tag{8}$$

$$\widetilde{V}^*(s) = V^*(s) + V_K^*(s) \tag{9}$$

and

$$\widetilde{\pi}^*(a|s) = \pi_K^*(a|s) \tag{10}$$

for all $s \in \mathcal{S},\ a \in \mathcal{A}$.

Therefore, by directly incorporating the solution of $\mathcal{T}$ into $\widetilde{Q}^*$ we can instead learn an auxiliary function K^* which itself happens to be an optimal action-value function. We can now use the same soft Q-learning algorithms (Haarnoja et al. 2018b) for learning this corrective value function (K^*) via Equation (6). In doing so, we learn the desired optimal value function for the new task: $\widetilde{Q}^*$.

As discussed in (Hunt et al. 2019), it is also possible to learn a corrective value function strictly offline, by using data collected for a previous task, $\mathcal{T}$. In the Appendix, we show that it is indeed possible to learn K^* using offline updates. In such a setup, the advantage is that one requires no additional samples of the environment (the previous experience can be used with appropriately re-labelled rewards). While learning the value function K^* of Theorem 1 we are implicitly solving *two* tasks simultaneously: one task corresponding to reward function κ with prior policy π^*, and another task with a reward function $\widetilde{r}$ and prior policy π_0. In the following section, we show that the former task can be mapped onto yet another task which also has a prior policy π_0.

Reward Shaping

In this section we explore the connection between the result of Theorem 1 and the field of potential-based reward shaping. Equation (10) implies that the optimal policies are the same for two distinct entropy-regularized RL problems. This is exactly the desired outcome of a shaped reward: the task's optimal policy is invariant to a change in the task's reward function.

However, these two entropy-regularized RL problems use different prior policies and therefore the result is not immediately applicable to reward shaping. To correct for this, we introduce the following lemma which describes how a change in prior policy can be accounted for in the rewards of an entropy-regularized RL task.

Lemma 2. *Suppose the task $\mathcal{T} = \langle \mathcal{S}, \mathcal{A}, p, r, \gamma, \beta, \pi_0 \rangle$ is given with associated optimal value function Q^* and optimal policy π^*. For a reward function*

$$\widetilde{r}(s,a,s') = r(s,a,s') + \frac{1}{\beta} \log \frac{\pi_0(a|s)}{\pi_1(a|s)} \tag{11}$$

the task $\widetilde{\mathcal{T}} = \langle \mathcal{S}, \mathcal{A}, p, \widetilde{r}, \gamma, \beta, \pi_1 \rangle$ has optimal value functions

$$\widetilde{Q}^*(s,a) = Q^*(s,a) + \frac{1}{\beta} \log \frac{\pi_0(a|s)}{\pi_1(a|s)} \tag{12}$$

$$\widetilde{V}^*(s) = V^*(s) \tag{13}$$

and the optimal policies of $\mathcal{T}$ and $\widetilde{\mathcal{T}}$ are equal:

$$\pi^*(a|s) = \widetilde{\pi}^*(a|s) \tag{14}$$

for all $s \in \mathcal{S}, a \in \mathcal{A}$.

Therefore, if the prior policy is changed ($\pi_0 \to \pi_1$) in an entropy-regularized RL task, we can appropriately adjust the reward function (as written in Equation (11)) in order to retain an optimal solution. Therefore, by solving a task $\mathcal{T}$ with prior policy π_0, we have also simultaneously solved all tasks with an arbitrary prior policy $\pi_1 > 0$ and corresponding reward functions $\widetilde{r}$ given in Equation (11).

Now by applying Lemma 2 to Theorem 1, we immediately have the following reward shaping result for entropy-regularized RL.

Corollary 3 (Reward Shaping)**.** *Let reward varying tasks $\mathcal{T}$ and $\widetilde{\mathcal{T}}$ be given with corresponding solutions (Q^*, V^*, π^*) and $(\widetilde{Q}^*, \widetilde{V}^*, \widetilde{\pi}^*)$. The optimal value function for another reward varying task, $\bar{\mathcal{T}}$ with the reward function*

$$\bar{r}(s, a, s') = \widetilde{r}(s, a, s') + \gamma V^*(s') - V^*(s) \tag{15}$$

is given by

$$\bar{Q}^*(s, a) = \widetilde{Q}^*(s, a) - Q^*(s, a) + \frac{1}{\beta} \log \frac{\pi^*(a|s)}{\pi_0(a|s)} \tag{16}$$

$$= \widetilde{Q}^*(s, a) - V^*(s) \tag{17}$$

and the corresponding optimal policy for $\bar{\mathcal{T}}$ is

$$\bar{\pi}^*(a|s) = \widetilde{\pi}^*(a|s). \tag{18}$$

Therefore, a shift in the reward function by $\gamma V^*(s') - V^*(s)$ does not change the optimal policy, for any reward function r with corresponding optimal soft value functions Q^* and V^*. However, Corollary 3 does not yet resemble a *general* reward shaping theorem, since it *requires* a solution to a different task ($\mathcal{T}$) to provide the shaping function, and we do not know whether such reward functions fully characterize *all* reward functions with the same optimal policy $\widetilde{\pi}^*$. To address these issues, we combine Corollary 3 with the following lemma appearing in (Cao, Cohen, and Szpruch 2021). In doing so, we arrive at a potential-based reward shaping theorem for entropy-regularized RL.

Lemma 4. *(Cao, Cohen, and Szpruch 2021) For a fixed policy $\pi(a|s) > 0$, discount factor $\gamma \in [0, 1)$, and an arbitrary choice of function $v \colon \mathcal{S} \to \mathbb{R}$, there is a unique corresponding reward function*

$$R(s, a, s') = \frac{1}{\beta} \log \frac{\pi(a|s)}{\pi_0(a|s)} + v(s) - \gamma v(s') \tag{19}$$

such that the task with reward R yields an optimal value function $V^*(s) = v(s)$ and optimal policy $\pi^* = \pi$.

Informally, this result states that one can always construct a reward function R such that an arbitrary π and v are optimal in the given environment (i.e. with fixed p, γ, β). Furthermore, for a fixed π and v in the given environment, the reward function R is uniquely defined up to a constant shift.

By applying Lemma 4 in light of Theorem 3, we see that the V^* in Equation (15) can in fact be *any* (bounded) function, $\Phi \colon \mathcal{S} \to \mathbb{R}$: it will always represent an optimal value function for a reward varying task. In fact, Cao, Cohen, and Szpruch have proven that *all* reward functions having the same optimal policy must be of the form shown in Equation (19). Therefore, the set of reward functions of this type describe *all* rewards with the same optimal policy.

This naturally leads us to the generalization of the primary result of (Ng, Harada, and Russell 1999) in the setting of entropy-regularized RL:

Theorem 5 (Potential-Based Reward Shaping)**.** *Given task $\widetilde{\mathcal{T}} = \langle \mathcal{S}, \mathcal{A}, p, \widetilde{r}, \gamma, \beta, \pi_0 \rangle$ with optimal policy $\widetilde{\pi}^*$ in entropy-regularized RL, then the reward varying task $\bar{\mathcal{T}}$ with reward function*

$$\bar{r}(s, a, s') = \widetilde{r}(s, a, s') + \gamma \Phi(s') - \Phi(s) \tag{20}$$

has the optimal policy $\bar{\pi}^* = \widetilde{\pi}^*$, and its optimal value functions satisfy

$$\bar{Q}^*(s, a) = \widetilde{Q}^*(s, a) - \Phi(s) \tag{21}$$

$$\bar{V}^*(s) = \widetilde{V}^*(s) - \Phi(s) \tag{22}$$

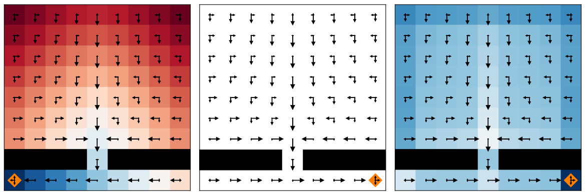
for a bounded, but otherwise arbitrary function $\Phi \colon \mathcal{S} \to \mathbb{R}$.

Furthermore, due to the aforementioned result of (Cao, Cohen, and Szpruch 2021), we note that we also have the following necessary and sufficient conditions:

- Sufficiency: Adding a *potential-based* function $F(s, a, s') = \gamma\ \Phi(s') - \Phi(s)$ to a task's reward function leaves the task's optimal policy unchanged.
- Necessity: Any shaping function which leaves a task's optimal policy invariant must be of the form F above.

Remark 6. Interestingly, the method described above for arriving at Theorem 5 provides a useful way of considering the possible "degrees of freedom" in the reward function by contrasting the work of (Ng, Harada, and Russell 1999) and (Cao, Cohen, and Szpruch 2021). Specifically, (Cao, Cohen, and Szpruch 2021) considered an arbitrary function $v(s)$ which uniquely (up to a constant shift) identifies the reward function for a given optimal policy, as noted above. To make the connection with our results, we can take this arbitrary function to be the value function $V_K^*(s)$, since we have $\pi_K^* = \widetilde{\pi}^*$. Because of Equation (9), we can equivalently take this choice of $v(s)$ to fix $V^*(s)$, which can be any arbitrary function $\Phi(s)$, as noted in Theorem 5 above. In other words, instead of considering the degree of freedom to be the value function V_K^*, we can alternatively consider it to be the value function V^*. The degree of freedom identified in (Cao, Cohen, and Szpruch 2021) corresponds to V_K^*, whereas the degree of freedom identified by (Ng, Harada, and Russell 1999) corresponds to V^*. The two choices are equivalent, given Equation (9).

The utility of Theorem 5, as compared to Corollary 3, is that we now have a way of constructing shaping functions without requiring knowledge of optimal policies in advance. Lemma 4 alone would not have allowed for this, as it requires an optimal policy π as input for the reward function R, which is more useful in the study of inverse reinforcement learning and identifiability. Of course, if the optimal policy π^* *is* known, then one can use Lemma 4 to construct the entire range of shaped rewards by iterating over all possible functions $v(s)$. Considering the space of all reward functions ($\mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$) Lemma 4 carves out a "degenerate" subspace



(a) Solved Task (b) Original Rewards (c) Shaped Rewards

Figure 1: Demonstrating the reward shaping result of Corollary 3. Tasks in question are to navigate to the left or right corner in the bottom of this simple maze. The orange diamond represents the goal state for either task (blue represents regions of higher value). (a) Task with known solution (π^* and $V^*(s)$ shown); (b) New task in the same environment (π^* and r shown); (c) Task defined in (b) with a shaped reward function (Equation (15)), having the same optimal policy as the task with an unshaped reward function (π^* and r shown). In this experiment we use the parameters $\beta = 3, \gamma = 0.99$.

or "class" of dimension $|\mathcal{S}|$, whose members are defined by tasks with the same optimal policy.

Analogous to Remark 1 of (Ng, Harada, and Russell 1999), we also have the following result in entropy-regularized reinforcement learning, which provides robustness to the shaping function F.

Remark 7. Given a potential-based reward shaping function F as described in Theorem 5, the relations

$$\bar{Q}^{\pi}(s,a) = \widetilde{Q}^{\pi}(s,a) - \Phi(s) \tag{23}$$

$$\bar{V}^{\pi}(s) = \widetilde{V}^{\pi}(s) - \Phi(s) \tag{24}$$

also hold for non-optimal policies π. In particular, if π is an ϵ-optimal policy in the reward-shaped task (i.e. $||\bar{V}^{\pi} - \bar{V}^{*}|| < \epsilon$) then, π is also ϵ-optimal in the original task (i.e. $||\widetilde{V}^{\pi} - \widetilde{V}^{*}|| < \epsilon$).

This robustness is a useful property, as it allows near-optimal policies in the reward-shaped task to be readily interpreted as near-optimal policies in the original task of interest. More generally speaking, we can say that the action of reward shaping preserves ordering in the space of policies.

Identifiability

The problem of identifiability arises in the context of inverse reinforcement learning (IRL) where one observes an optimal policy and attempts to infer the underlying reward function. In (Cao, Cohen, and Szpruch 2021), it was argued that, under certain conditions, the underlying reward function r is identifiable (up to a constant shift) when its corresponding optimal policy is observed in two sufficiently different environments with dynamics p and q. As in (Cao, Cohen, and Szpruch 2021), we suppose the environments are further diversified by different discount factors γ and $\widetilde{\gamma}$ respectively. In the following, we will see how our results provide insight into the conditions considered in (Cao, Cohen, and Szpruch 2021) for identifiability using data from these two different environments.

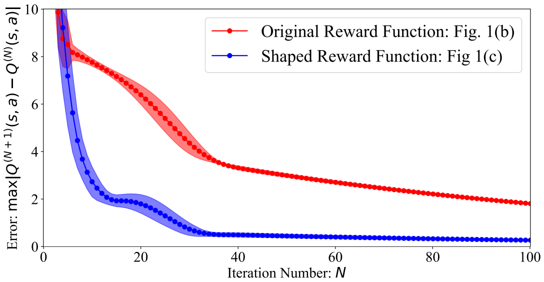


Figure 2: Convergence of the Bellman backup equation (Equation (3)) for the unshaped and shaped task considered in Fig. 1(b), 1(c) respectively. Average taken over 10 random initializations, and one standard deviation is shown in the shaded region.

Theorem 5 can be used to derive the condition which makes it impossible to determine an underlying reward function r given the optimal policies in p and q. If it is possible to shape r with a potential $\Phi(s)$ in p and shape r by another potential $\Psi(s)$ in q in a way that makes the shaped reward functions identical, then identifiability is not possible. Using Equation (20) to equate two such shaped rewards, we arrive at the following condition:

$$\gamma \mathop{\mathbb{E}}_{s' \sim p} \Phi(s) - \Phi(s') = \widetilde{\gamma} \mathop{\mathbb{E}}_{s' \sim q} \Psi(s) - \Psi(s') \tag{25}$$

If the above condition is satisfied by non-trivial shaping potentials Φ and Ψ, then there are at least two reward functions (the unshaped and correspondingly shaped rewards) in dynamics p and q which are consistent with all the observed constraints, hence the reward function will not be identifiable. The condition that there are no such non-trivial shaping potentials is imposed by Definition 1 of (Cao, Cohen, and Szpruch 2021) which then leads to their Theorem 2.

Change of Dynamics

Beyond having a new task whose sole distinction is in the reward function (reward varying tasks), we can instead consider two tasks which differ in their transition dynamics:

Definition (Dynamics Varying Tasks). Consider a set of tasks $\{\mathcal{T}^{(k)}\}_{k=1}^{N}$. If the tasks only vary on their transition dynamics; that is, they are of the form $\mathcal{T}^{(k)} = \langle \mathcal{S}, \mathcal{A}, p^{(k)}, r, \gamma, \beta, \pi_0 \rangle$ then we say the set of tasks $\{\mathcal{T}^{(k)}\}$ is **dynamics varying**.

Consider a shift in the environment's dynamics $p(s'|s,a) \rightarrow q(s'|s,a)$ as represented by two dynamics varying tasks. For example, in a discrete maze setting, the floor may become more slippery. We again derive a corrective value function that utilizes the previous solution to a dynamics varying task. We now state the corresponding result for the corrective value function in this case:

Theorem 8. *Let a task* $\mathcal{T} = \langle \mathcal{S}, \mathcal{A}, p, r, \gamma, \beta, \pi_0 \rangle$ *be given, with optimal value function* V^* *and corresponding optimal policy* π^*. *Let a dynamics varying task,* $\widetilde{\mathcal{T}} = \langle \mathcal{S}, \mathcal{A}, q, r, \gamma, \beta, \pi_0 \rangle$, *be given, with an unknown optimal action-value function* $\widetilde{Q}^*(s,a)$. *Assume that* $r = r(s,a)$.

Denote the optimal action-value function K^* as the solution of the following Bellman optimality equation

$$K^*(s,a) = \kappa(s,a) + \frac{\gamma}{\beta} \mathop{\mathbb{E}}_{s' \sim q} \log \mathop{\mathbb{E}}_{a' \sim \pi^*} e^{\beta K^*(s',a')} \tag{26}$$

where κ is the corresponding reward function:

$$\kappa(s,a) = \gamma \mathop{\mathbb{E}}_{s' \sim q} V^*(s') - \gamma \mathop{\mathbb{E}}_{s' \sim p} V^*(s'). \tag{27}$$

where $V^*(s)$ is the optimal state value function for the task defined by dynamics $p(s'|s,a)$.

Then,

$$\widetilde{Q}^*(s,a) = Q^*(s,a) + K^*(s,a) \tag{28}$$

$$\widetilde{V}^*(s) = V^*(s) + V^*_K(s) \tag{29}$$

and

$$\widetilde{\pi}^*(a|s) = \pi^*_K(a|s) \tag{30}$$

for all $s \in \mathcal{S}$, $a \in \mathcal{A}$.

Therefore, in the face of changed dynamics, an agent can adapt by learning the value function K^* instead of $\widetilde{Q}^*$. In this way, the agent uses the relevant knowledge already accumulated in a similar environment.

For simplicity, we have kept the rewards the same across the two tasks, but the results of Theorem 1 and Theorem 8 can be readily combined to accommodate those tasks which have different dynamics *and* different reward functions.

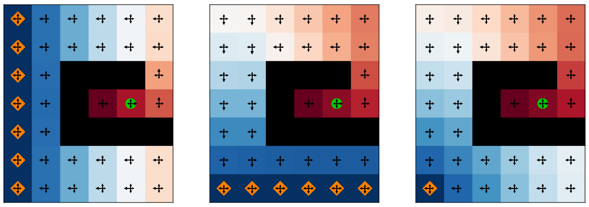
Examining Equation (28), where $\widetilde{Q}^*$ and K^* correspond to a task with dynamics q, and Q^* corresponds to dynamics p, we can instead consider $\widetilde{Q}^*$ and K^* as being related via Equation (8) of Theorem 1. With this perspective, Q^* represents the optimal value-function of a reward varying task with dynamics q and a different reward function (denoted $\bar{r}$ below). Therefore, given a solution to a task with dynamics p, we automatically have the solution to a task in dynamics q. The reward function $\bar{r}$ in a task with dynamics q to which this optimal value function corresponds, is provided by the following theorem:

Theorem 9. *Consider the task $\mathcal{T} = \langle \mathcal{S}, \mathcal{A}, p, r, \gamma, \beta, \pi_0 \rangle$ with corresponding optimal value functions $Q^*(s,a), V^*(s)$ and optimal policy π^*. Then for a reward function*

$$\bar{r}(s,a) = r(s,a) - \gamma \mathop{\mathbb{E}}_{s' \sim q} V^*(s') + \gamma \mathop{\mathbb{E}}_{s' \sim p} V^*(s') \tag{31}$$

the task given by $\bar{\mathcal{T}} = \langle \mathcal{S}, \mathcal{A}, q, \bar{r}, \gamma, \beta, \pi_0 \rangle$ has the same optimal action-value function, hence $\bar{V}^* = V^*$ and $\bar{\pi}^* = \pi^*$.

Interestingly, Theorem 9 implies that by solving problems in one environment with dynamics p, we are also simultaneously solving different problems in (arbitrary) other dynamics q. Hence, by learning in an environment that is “safer” to experiment in (p), we can obtain solutions to tasks in another environment (q), perhaps where testing is more difficult, expensive, or dangerous. By assembling this set of rewards ($\bar{r}$) for which we have the solution in dynamics q, we can either (a) attempt to solve the inverse problem of finding a task to solve in q such that that we solve the desired task in p or (b) use the forthcoming results on compositionality and previous results on reward shifts to solve the task(s) of interest.



(a) Subtask: "Left" (b) Subtask: "Down" (c) Composition: "AND"

Figure 3: Composition of two subtasks (a) and (b), with function $f = \min(\cdot)$. All subfigures illustrate π^* and $V^*(s)$. (a)-(b) Tasks with known solution (c) Composition of subtasks. In this experiment we use the parameters $\beta = 2, \gamma = 0.98$. The green circle indicates the agent's initial state and the orange diamond represents the goal state for either task.

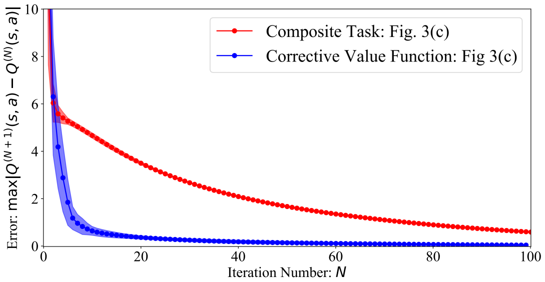


Figure 4: Convergence of the Bellman backup equation (Equation (3)) for the task considered in Fig. 3(c). Red is the composition ($\widetilde{Q}^*$) learned directly. Blue is the corrective value function (K^*). Average taken over 25 random initializations, and one standard deviation is shown in the shaded region.

Composition of Rewards

In this section we generalize the previous results by considering arbitrary compositions of M reward varying tasks. That is, we now consider a set of solved tasks $\{\mathcal{T}^{(m)}\}_{m=1}^{M}$ such that $\mathcal{T}^{(m)} = \langle \mathcal{S}, \mathcal{A}, p, r^{(m)}, \gamma, \beta, \pi_0 \rangle$. To compose these tasks, we consider applying a function f to the reward functions of $\{\mathcal{T}\}$. We also note the specific case of $M = 1$ for transformations of a single task's reward function may be of interest.

Definition (Task Composition). Consider a set of reward varying tasks, $\{\mathcal{T}^{(m)}\}_{m=1}^{M}$. The **composition** of $\{\mathcal{T}^{(m)}\}_{m=1}^{M}$ under the (bounded) function $f \colon \mathbb{R}^M \to \mathbb{R}$ is defined as the mutually reward varying task with reward function $r(s,a,s') = f(\{r^{(m)}(s,a,s')\})$.

Motivated by the results of (Hunt et al. 2019), we derive another corrective value function, which corrects the naïve guess of functionally transforming the value functions in the same way as the rewards (that is, $f(\{Q^{(m)}\})$). This naïve guess is in fact a bound in the case of convex combinations (Haarnoja et al. 2018a). Learning the corrective value function for a simple composition task is illustrated in Figures 3 and 4.

Theorem 10. *Given a set of reward varying tasks $\{\mathcal{T}^{(m)}\}$ with corresponding optimal value functions $\{Q^{(m)}\}$, denote $\widetilde{Q}^*$ as the optimal action-value function for the composition of $\{\mathcal{T}^{(m)}\}$ under f. Define the value function K^* as the solution of the following Bellman optimality equation*

$$K^*(s,a) = \mathop{\mathbb{E}}_{s'\sim p}\left[\kappa(s,a,s') + \frac{\gamma}{\beta}\log \mathop{\mathbb{E}}_{a'\sim\pi_f} e^{\beta K^*(s',a')}\right] \tag{32}$$

where κ is the corresponding reward function:

$$\kappa(s,a,s') = \left[f(\{r^{(m)}\}) + \gamma V_f(s')\right] - f(\{Q^{(m)}(s,a)\}) \tag{33}$$

with the definition

$$V_f(s) = \frac{1}{\beta}\log \mathop{\mathbb{E}}_{a\sim\pi_0} \exp \beta f(\{Q^{(m)}(s,a)\}),$$

and π_f is the policy derived from $f(\{Q^{(m)}\})$:

$$\pi_f(a|s) = \frac{\pi_0(a|s)e^{\beta f(\{Q^{(m)}(s,a)\})}}{e^{\beta V_f(s)}} \tag{34}$$

Then,

$$\widetilde{Q}^*(s,a) = f(\{Q^*(s,a)\}) + K^*(s,a) \tag{35}$$

$$\widetilde{V}^*(s) = V_f^*(s) + V_K^*(s) \tag{36}$$

and

$$\widetilde{\pi}^*(a|s) = \pi_K^*(a|s) \tag{37}$$

for all $s \in \mathcal{S},\ a \in \mathcal{A}$.

The result of this theorem can be stated as follows: calculate a policy by transforming the optimal value functions in the same way the rewards were transformed. Using this new policy as the prior policy with an appropriate reward function (κ defined in Equation (33)), we can learn the correction term, K^*, to obtain the desired optimal value function, $\widetilde{Q}^*$. We again note that, as in Theorem 1, it is possible to learn K^* in an offline manner which is described in the proof of Theorem 10 in the Appendix.

The fixed point K^* in Equation (32) generalizes the "Divergence Correction" (C^∞) introduced by (Hunt et al. 2019) for convex combinations of reward functions. Notice that the reward function κ in Theorem 10 measures the "nonlinearity" of f. For if f were linear (cf. Theorem 3.2 of (Hunt et al. 2019)), then the first term (in brackets) cancels with the total transformed Q function being subtracted, leaving the Rényi divergence between subtask policies. In addition, we have also shown that K^* is in fact the *optimal* value function for a certain task: the task with rewards and prior policy as defined in Theorem 10.

Discussion

In this work, we have studied transfer learning in entropy-regularized reinforcement learning. Specifically, we have considered reward varying tasks, dynamics varying tasks, and composition of reward varying tasks. By deriving a corrective value function in each case, we have shown that the solutions for new tasks can be informed by previous solutions. Interestingly, this study of corrective value functions also led to the derivation of potential-based reward shaping in entropy-regularized RL.

We have shown that optimal solutions under a given transition dynamics also corresponds to a set of optimal solutions under any other dynamics, by explicitly calculating the reward function in Equation (31). This change in perspective between Theorem 1 and Theorem 8 allows one to transfer a body of knowledge obtained in one dynamics p, to any other dynamics of interest. Although these are solutions to reward functions which may not be of interest *a priori*, the solutions may still prove useful when used in tandem with the results of Theorem 1 and Theorem 10.

We have also generalized the "Divergence Correction" result of (Hunt et al. 2019), allowing for general transformations and compositions over primitive tasks. All derived corrective value functions allow the agent to solve the task of interest by applying previous knowledge to the problem at hand.

Limitations and Future Work

Although the results and proofs are stated for discrete settings, it is straightforward to extend the results to continuous state and action spaces, with the usual assumptions for Bellman convergence. The tests here are demonstrated in discrete finite environments, but this work may also be extended to encompass continuous spaces.

This work is situated in the context of entropy-regularized RL, where the stochasticity of optimal policies allow for optimal solutions to be manipulated and combined for new tasks. Further work may explore the analogous problem in un-regularized RL, which can we understood as the limit $\beta \to \infty$. We also note that the case of $\gamma = 1$ has straightforward proofs in the probabilistic inference framework (Levine 2018). We intend to explore these results and their consequences in future work.

Just as we have derived corrective value functions for changes in reward function, prior policy (Lemma 2), and dynamics change, one might also consider tasks which differ in discount factor γ or temperature β^{-1}. Although not stated here, similar results can be derived for these settings as well. These generalization and their consequences will be explored in future work.

Finally, we note that there may be more general results for a definition of composition which allow the dynamics of tasks to differ as well. This topic, and possible implications for sim-to-real and general transfer learning is currently being explored and will be left to future work.

Acknowledgments

The authors would like to thank the anonymous reviewers for their helpful comments and suggestions. This work was supported by the National Science Foundation through Award DMS-1854350, the Proposal Development Grant provided by the University of Massachusetts Boston, the Research Foundation at San José State University, and the Alliance Innovation Lab in Silicon Valley.

References

Adamczyk, J.; Arriojas, A.; Tiomkin, S.; and Kulkarni, R. V. 2022. Utilizing Prior Solutions for Reward Shaping and Composition in Entropy-Regularized Reinforcement Learning. arXiv:2212.01174.

Cao, H.; Cohen, S.; and Szpruch, Ł. 2021. Identifiability in inverse reinforcement learning. *Advances in Neural Information Processing Systems*, 34: 12362–12373.

Eysenbach, B.; and Levine, S. 2022. Maximum Entropy RL (Provably) Solves Some Robust RL Problems. In *International Conference on Learning Representations*.

Haarnoja, T.; Pong, V.; Zhou, A.; Dalal, M.; Abbeel, P.; and Levine, S. 2018a. Composable Deep Reinforcement Learning for Robotic Manipulation. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE.

Haarnoja, T.; Tang, H.; Abbeel, P.; and Levine, S. 2017. Reinforcement Learning with Deep Energy-Based Policies. In Precup, D.; and Teh, Y. W., eds., *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, 1352–1361. PMLR.

Haarnoja, T.; Zhou, A.; Abbeel, P.; and Levine, S. 2018b. Soft Actor-Critic: Off-Policy Maximum Entropy Deep Reinforcement Learning with a Stochastic Actor. In Dy, J.; and Krause, A., eds., *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, 1861–1870. PMLR.

Hunt, J.; Barreto, A.; Lillicrap, T.; and Heess, N. 2019. Composing entropic policies using divergence correction. In *International Conference on Machine Learning*, 2911–2920. PMLR.

Levine, S. 2018. Reinforcement Learning and Control as Probabilistic Inference: Tutorial and Review. arXiv:1805.00909.

Nangue Tasse, G.; James, S.; and Rosman, B. 2020. A Boolean task algebra for reinforcement learning. *Advances in Neural Information Processing Systems*, 33: 9497–9507.

Ng, A. Y.; Harada, D.; and Russell, S. 1999. Policy invariance under reward transformations: Theory and application to reward shaping. In *Proceedings of the 16th International Conference on Machine Learning*, volume 99, 278–287.

Peng, X. B.; Chang, M.; Zhang, G.; Abbeel, P.; and Levine, S. 2019. MCP: Learning Composable Hierarchical Control with Multiplicative Compositional Policies. In Wallach, H.; Larochelle, H.; Beygelzimer, A.; d'Alché-Buc, F.; Fox, E.; and Garnett, R., eds., *Advances in Neural Information Processing Systems 32*, 3681–3692. Curran Associates, Inc.

Silver, D.; Hubert, T.; Schrittwieser, J.; Antonoglou, I.; Lai, M.; Guez, A.; Lanctot, M.; Sifre, L.; Kumaran, D.; Graepel, T.; et al. 2018. A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play. *Science*, 362(6419): 1140–1144.

Sutton, R. S.; and Barto, A. G. 2018. *Reinforcement learning: An introduction*. MIT press.

Tasse, G. N.; James, S.; and Rosman, B. 2021. Generalisation in Lifelong Reinforcement Learning through Logical Composition. In *Deep RL Workshop NeurIPS 2021*.

Taylor, M. E.; and Stone, P. 2009. Transfer Learning for Reinforcement Learning Domains: A Survey. *Journal of Machine Learning Research*, 10(56): 1633–1685.

Todorov, E. 2009. Compositionality of optimal control laws. In Bengio, Y.; Schuurmans, D.; Lafferty, J.; Williams, C.; and Culotta, A., eds., *Advances in Neural Information Processing Systems*, volume 22. Curran Associates, Inc.

Van Niekerk, B.; James, S.; Earle, A.; and Rosman, B. 2019. Composing value functions in reinforcement learning. In *International Conference on Machine Learning*, 6401–6409. PMLR.

Ziebart, B. D. 2010. Modeling Purposeful Adaptive Behavior with the Principle of Maximum Causal Entropy. *Carnegie Mellon University*.