Bounding the Optimal Value Function in Compositional Reinforcement Learning

Jacob Adamczyk[1] **Volodymyr Makarenko**[2] **Argenis Arriojas**[1] **Stas Tiomkin**[2] **Rahul V. Kulkarni**[1]

[1]Department of Physics, University of Massachusetts Boston, Boston, MA, USA
[2]Department of Computer Engineering, San José State University, San José, CA, USA

Abstract

In the field of reinforcement learning (RL), agents are often tasked with solving a variety of problems differing only in their reward functions. In order to quickly obtain solutions to unseen problems with new reward functions, a popular approach involves functional composition of previously solved tasks. However, previous work using such functional composition has primarily focused on specific instances of composition functions whose limiting assumptions allow for exact zero-shot composition. Our work unifies these examples and provides a more general framework for compositionality in both standard and entropy-regularized RL. We find that, for a broad class of functions, the optimal solution for the composite task of interest can be related to the known primitive task solutions. Specifically, we present double-sided inequalities relating the optimal composite value function to the value functions for the primitive tasks. We also show that the regret of using a zero-shot policy can be bounded for this class of functions. The derived bounds can be used to develop clipping approaches for reducing uncertainty during training, allowing agents to quickly adapt to new tasks.

1 INTRODUCTION

Reinforcement learning has seen great success recently, but still suffers from poor sample complexity and task generalization. Generalizing and transferring domain knowledge to similar tasks remains a major challenge in the field. To combat this, different methods of transfer learning have been proposed; such as the option framework [Sutton et al., 1999, Barreto et al., 2019], successor features [Dayan, 1993, Barreto et al., 2017, Hunt et al., 2019, Nemecek and Parr, 2021], and functional composition [Todorov, 2009, Haarnoja et al., 2018a, Peng et al., 2019, Tasse et al., 2021, Van Niekerk et al., 2019]. In this work, we focus on the latter method of "compositionality" for transfer learning.

Research in compositionality has focused on the development of approaches to combine previously learned optimal behaviors to obtain solutions to new tasks. In the process, many instances of functional composition in the literature have required limiting assumptions on the dynamics and allowable class of reward functions (goal-based rewards in [Tasse et al., 2020]) in order to derive exact results. Furthermore, previous work has focused on isolated examples of particular functions in either standard or entropy-regularized RL and a framework for studying a general class of composition functions without limiting assumptions is currently lacking. One of the main contributions of our work is to provide a unifying general framework to study compositionality in reinforcement learning.

In our approach, we focus on "primitive" tasks which differ only in their associated reward functions. More specifically, we consider those downstream tasks whose reward functions can be written as a global function of the known source tasks' reward functions. To maintain generality, we do not assume that transition dynamics are deterministic. We also do not assume that reward functions are limited to the goal-based setting, in which there are a limited number of absorbing "goal" states [Todorov, 2009, Van Niekerk et al., 2019] defining the primitive task. Given the generality of this setting, we cannot expect to obtain exact solutions for compositions as in prior work. Instead, we provide a class of functions which can be used to obtain approximate solutions and bounds on the corresponding downstream tasks.

Given the solutions to a set of primitive tasks, we show that it is possible to leverage such information to obtain approximate solutions for a large class of compositely-defined tasks. To do so, we relate the solution of the downstream composite task to the solved primitive (source) tasks. Specifically, we derive relations on the optimal value function of interest. From such a relation, a "zero-shot" (i.e. not requiring further

Accepted for the 39th Conference on Uncertainty in Artificial Intelligence (UAI 2023).

training) policy can be extracted for use in the composite domain of interest. We then show that the suboptimality (regret) of this zero-shot policy is upper bounded.

Our results support the idea that RL agents can focus on obtaining domain knowledge for simpler tasks, and later use this knowledge to effectively solve more difficult tasks. The primary contributions of the present work are as follows:

Main contributions

- Establishing a general framework for analyzing reward transformations and compositions for the case of stochastic dynamics, globally varying reward structures, and continuing tasks.
- Derivation of bounds on the respective optimal value functions for transformed and composite tasks.
- Demonstration of zero-shot approximate solutions and value-based clipping of new tasks based on the known optimal solutions for primitive tasks.

2 BACKGROUND

In this work, we analyze the case of finite, discrete state and action spaces, with the Markov Decision Process (MDP) model [Sutton and Barto, 2018]. Let $\Delta(X)$ represent the set of probability distributions over X. Then the MDP is represented as a tuple $\mathcal{T} = \langle \mathcal{S}, \mathcal{A}, p, \mu, r, \gamma \rangle$ where $\mathcal{S}$ is the set of available states; $\mathcal{A}$ is the set of possible actions; $p : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ is a transition function describing the system dynamics; $\mu \in \Delta(\mathcal{S})$ is the initial state distribution; $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is a (bounded) reward function which associates a reward (or cost) with each state-action pair; and $\gamma \in (0, 1)$ is a discount factor which discounts future rewards and guarantees the convergence of total reward for infinitely long trajectories ($T \to \infty$).

In "standard" (un-regularized) RL, the agent maximizes an objective function which is the expected future reward:

$$J(\pi) = \mathop{\mathbb{E}}_{\tau \sim p, \pi} \left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \right]. \tag{1}$$

This objective has since been generalized for the setting of entropy-regularized RL [Ziebart, 2010, Levine, 2018], which augments the standard RL objective in Eq. (1) by appending an entropic regularization term for the policy:

$$J(\pi) = \mathop{\mathbb{E}}_{\tau \sim p, \pi} \left[\sum_{t=0}^{\infty} \gamma^t \left(r(s_t, a_t) - \frac{1}{\beta} \log \left(\frac{\pi(a_t|s_t)}{\pi_0(a_t|s_t)} \right) \right) \right] \tag{2}$$

where $\pi_0 : \mathcal{S} \to \Delta(\mathcal{A})$ is the fixed prior policy. The inverse temperature parameter, $\beta \in (0, \infty)$, regulates the contribution of entropic costs relative to the accumulated rewards. The additional entropic control cost discourages the agent from choosing policies that deviate too much from the prior policy. Importantly, entropy-regularized MDPs lead to stochastic optimal policies that are provably robust to perturbations of rewards and dynamics [Eysenbach and Levine, 2022]; making them a more suitable choice for real-world problems.

By "solution to the RL problem", we hereon refer to the corresponding *optimal* action-value function $Q(s, a)$ from which an *optimal* control policy can be derived: $\pi(s) \in \text{argmax}_a Q(s, a)$ for standard RL; and $\pi(a|s) \propto \exp(\beta Q(s, a))$ for entropy-regularized RL. Note that these definitions are consistent with the limit $\beta \to \infty$ in which the standard RL objective is recovered from Eq. (2). For both standard and entropy-regularized RL, the optimal Q-function can be obtained by iterating a recursive Bellman equation. For standard RL, the Bellman optimality equation is given by [Sutton and Barto, 2018]:

$$Q(s, a) = r(s, a) + \gamma \mathbb{E}_{s' \sim p(\cdot|s,a)} \max_{a'} (Q(s', a')) \tag{3}$$

The entropy term in the objective function for entropy-regularized RL modifies the previous optimality equation to [Ziebart, 2010, Haarnoja et al., 2018b]:

$$Q(s, a) = r(s, a) + \frac{\gamma}{\beta} \mathop{\mathbb{E}}_{s'} \log \mathbb{E}_{a' \sim \pi_0(\cdot|s')} e^{\beta Q(s', a')} \tag{4}$$

One of the primary goals of research in compositionality and transfer learning is deriving results for the optimal Q function for new tasks based on the known optimal Q function(s) for primitive tasks. There exist many forms of composition and transfer learning in RL, as discussed by Taylor and Stone [2009]. In this paper, we focus on the case of concurrent skill composition by a single agent as opposed to an options-based approach [Sutton et al., 1999], or other hierarchical compositions [Pateria et al., 2021, Saxe et al., 2017]. We elaborate on this point with the definitions below.

To formalize our problem setup, we adopt the relevant definitions provided by [Adamczyk et al., 2022]:

Definition 2.1. A **primitive RL task** is specified by an MDP $\mathcal{T} = \langle \mathcal{S}, \mathcal{A}, p, r, \gamma \rangle$ for which the optimal Q function is known.

In this work, we focus on primitive tasks with general reward functions, i.e. including both goal-based (sparse rewards on absorbing sets such as in the linearly solvable MDP framework of [Todorov, 2009, Tasse et al., 2020, Van Niekerk et al., 2019]) *and* arbitrary reward landscapes [Haarnoja et al., 2018a].

Definition 2.2. The **transformation of an RL task** is defined by its (bounded and continuous) transformation function: $f : \mathbb{R} \to \mathbb{R}$ and a primitive task $\mathcal{T}$. The transformed task shares the same states, actions, dynamics, and discount factor as $\mathcal{T}$ but has a transformed reward function $\widetilde{r}(s, a) = f(r(s, a))$.

Definition 2.3. The **composition of M RL tasks** is defined by a (bounded and continuous) function $F : \mathbb{R}^M \to \mathbb{R}$ and a set of primitive tasks $\{\mathcal{T}^{(k)}\}$. The **composite** RL task is defined by a new reward function $\widetilde{r}(s, a) = F(\{r^{(k)}(s, a)\})$; and shares the same states, actions, dynamics, and discount factor as all the primitive RL tasks.

Finally, we define the Transfer Library, the set of functions which obey the hypotheses of our subsequent results (see Sections 4 and 5). This definition serves to facilitate the general discussion of results obtained.

Definition 2.4. Given a set of primitive tasks $\{\mathcal{T}^{(k)}\}$, the **Transfer Library**, denoted by $\mathcal{F}$, is the set of all transformation (or composition, when $M > 1$) functions f which admit double-sided bounds (see Sections 4 and 5) on the composite task's optimal Q function ($\widetilde{Q}$).

Specifically, $\mathcal{F} = \{f \mid f \text{ satisfies Lemma 4.1 or 4.3}\}$ for standard RL and $\mathcal{F} = \{f \mid f \text{ satisfies Lemma 5.1 or 5.3}\}$ for entropy-regularized RL.

We have empirically found (cf. Fig. 1 and Supplementary Material) that by using the derived bounds for the optimal value function, the agent can learn the optimal policy more efficiently for tasks in the Transfer Library.

3 PREVIOUS WORK

There is much previous work concerning compositionality and transfer learning in reinforcement learning. In this section we will give a brief overview by highlighting the work most relevant for the current discussion.

In this work, we focus on value-based composition; rather than policy-based composition Peng et al. [2019], features-based composition [Barreto et al., 2017], or hierarchical (e.g. options-based) composition [Alver and Precup, 2021, Sutton et al., 1999, Barreto et al., 2019].

Value based methods of composition use the optimal value functions of lower-level or simpler "primitive" tasks to derive an approximation (or in some cases exact solution) for the composite task of interest. In the optimal control framework, [Todorov, 2009] has shown that optimal value functions can be composed exactly for linearly-solvable MDPs with a LogSumExp or "soft OR" composition over primitive tasks; assuming that tasks share the same absorbing set (boundary states). With a similar assumption of the shared absorbing set, [Tasse et al., 2020] show that exact optimal value functions for Boolean compositions may be recovered from primitive task solutions; thereby allowing an exponential improvement in knowledge acquisition.

In more recent work, in the context of MaxEnt RL, Haarnoja et al. [2018a] have shown that linear convex-weighted compositions in stochastic environments result in a bound on optimal value functions, and the policy extracted from this zero-shot bound is indeed useful for solving the composite task. The same premise of convex-weighted reward structures was studied by Hunt et al. [2019] where the difference between the bound of [Haarnoja et al., 2018a] and the optimal value function can itself be learned, effectively tightening the bound until convergence. This notion of a corrective function was subsequently generalized by Adamczyk et al. [2022] to allow for arbitrary functions of composition in entropy-regularized RL.

Other authors have considered the question of linear task decomposition, for instance [Barreto et al., 2017] where a convex weight vector over learned *features* can be calculated to solve the transfer problem over linearly-decomposable reward functions in standard RL. More recent developments on this line of research include [Hong et al., 2022] where a more general "bilinear value decomposition", conditioned on various goals, is learned. In [Kim et al., 2022], the authors consider the successor features (SFs) framework of [Barreto et al., 2017], and propose lower and upper bounds on the optimal value function of interest. They show that by replacing standard generalized policy improvement (GPI) with a constrained version which respects their bounds, they are able to transfer knowledge more successfully to future tasks in the successor features framework.

With our reduced assumptions (any constant dynamics, constant discount factor, any rewards) it is not generally possible to solve the transformed or composed tasks based only on primitive knowledge. Nevertheless, we are able to derive bounds on the optimal Q-functions in both standard and entropy-regularized RL, from which we can immediately derive policies which fare well in the transformed and composed problem settings. Additionally, we are able to prove that the derived policies have a bounded regret, in a similar form as Haarnoja et al. [2018a]'s Theorem 1; but in a more general setting.

4 STANDARD RL

4.1 TRANSFORMATION OF PRIMITIVE TASK

In this section, we consider **transformations of a primitive task** in the "standard" (un-regularized) RL setting. We assume a solved primitive task is given with reward function $r(s, a)$. Transforming this underlying reward function gives rise to a new reward function, $f(r(s, a))$ which specifies a new RL task to solve. All other variables defining the

MDP $(\mathcal{S}, \mathcal{A}, p, \mu, \gamma)$ are assumed to be fixed. In this new setting, we consider how to use the solution to the primitive task (that is, with rewards r) to inform the solution of the new, transfer task (that is, with rewards $f(r)$). The set of all applicable functions f for which we can derive bounds, forms the aforementioned *Transfer Library* with respect to the primitive task, for standard RL.

For a general class of transformations of reward functions (as defined below), we show that the optimal value function for the transformed task is bounded by an analogous functional transformation of the optimal value function for the primitive task. (The proofs for all theoretical results are provided in the Supplementary Material.)

We use the following definitions in the subsequent (standard RL) results: Let X be the codomain for the Q function of the primitive task ($Q : \mathcal{S} \times \mathcal{A} \to X \subseteq \mathbb{R}$). Let V_f denote the state-value function derived from the transformation function $f(Q)$: $V_f(s) = \max_a f(Q(s,a))$.

Lemma 4.1 (Convex Conditions)**.** *Given a primitive task with discount factor γ and a bounded, continuous transformation function $f : X \to \mathbb{R}$ which satisfies:*

1. *f is convex on its domain X[1];*
2. *f is sublinear:*
   (i) $f(x+y) \leq f(x) + f(y)$ for all $x, y \in X$
   (ii) $f(\gamma x) \leq \gamma f(x)$ for all $x \in X$
3. *$f\left(\max_a \mathcal{Q}(s,a)\right) \leq \max_a f\left(\mathcal{Q}(s,a)\right)$ for all functions $\mathcal{Q} : \mathcal{S} \times \mathcal{A} \to X$.*[2]

then the optimal action-value function for the transformed rewards, $\widetilde{Q}$, is now related to the optimal action-value function with respect to the original rewards by:

$$f(Q(s,a)) \leq \widetilde{Q}(s,a) \leq f(Q(s,a)) + C(s,a) \quad (5)$$

where $C(s,a)$ is the optimal value function for a task with reward

$$r_C(s,a) = f(r(s,a)) + \gamma \mathop{\mathbb{E}}_{s' \sim p} V_f(s') - f(Q(s,a)). \quad (6)$$

that is, C satisfies the following recursive equation:

$$C(s,a) = r_C(s,a) + \gamma \mathop{\mathbb{E}}_{s' \sim p} \max_{a'} C(s',a'). \quad (7)$$

With this result, we have a double-sided bound on the values of the optimal Q-function for the composite task. In particular, the lower bound ($f(Q)$) provides a zero-shot approximation for the optimal Q-function. It is thus of interest to analyze how well a policy π_f extracted from such an estimate ($f(Q)$) might perform. To this end, we provide the following result which bounds the suboptimality of π_f as compared to the optimal policy.

Lemma 4.2. *Consider the value of the policy $\pi_f(s) = \max_a f(Q(s,a))$ on the transformed task of interest, denoted by $\widetilde{Q}^{\pi_f}(s,a)$. The sub-optimality of π_f is then upper bounded by:*

$$\widetilde{Q}(s,a) - \widetilde{Q}^{\pi_f}(s,a) \leq D(s,a) \quad (8)$$

where D is the value of the policy π_f in a task with reward

$$r_D(s,a) = \gamma \mathop{\mathbb{E}}_{s' \sim p} \mathop{\mathbb{E}}_{a' \sim \pi_f} \Big[\max_b \left\{ f(Q(s',b)) + C(s',b) \right\} - f(Q(s',a')) \Big]$$

that is, D satisfies the following recursive equation:

$$D(s,a) = r_D(s,a) + \gamma \mathop{\mathbb{E}}_{s' \sim p} \mathop{\mathbb{E}}_{a' \sim \pi_f} D(s',a'). \quad (9)$$

Interestingly, the previous result shows that for functions f admitting a tight double-sided bound (that is, a relatively small value of C), the associated zero-shot policy π_f can be expected to perform near-optimally in the composite domain.

Another class of functions for which general bounds can be derived arises when f satisfies the following "reverse" conditions.

Lemma 4.3 (Concave Conditions)**.** *Given a primitive task with discount factor γ and a bounded, continuous transformation function $f : X \to \mathbb{R}$ which satisfies:*

1. *f is concave on its domain X[1];*
2. *f is superlinear:*
   (i) $f(x+y) \geq f(x) + f(y)$ for all $x, y \in X$
   (ii) $f(\gamma x) \geq \gamma f(x)$ for all $x \in X$
3. *$f\left(\max_a \mathcal{Q}(s,a)\right) \geq \max_a f\left(\mathcal{Q}(s,a)\right)$ for all functions $\mathcal{Q} : \mathcal{S} \times \mathcal{A} \to X$.*

then the optimal action-value functions are now related in the following way:

$$f(Q(s,a)) - \hat{C}(s,a) \leq \widetilde{Q}(s,a) \leq f(Q(s,a)) \quad (10)$$

where $\hat{C}$ is the optimal value function for a task with reward

$$\hat{r}_C(s,a) = f(Q(s,a)) - f(r(s,a)) - \gamma \mathop{\mathbb{E}}_{s' \sim p} V_f(s') \quad (11)$$

One obvious way to satisfy the final condition in the preceding lemma is to consider functions $f(x)$ which are monotonically increasing. Note that the definitions of C and $\hat{C}$

[1] This condition is not required for deterministic dynamics.

[2] Although this condition is automatically satisfied, it allows for a smoother connection to the analogous hypotheses in Lemmas 4.3, 5.1, 5.3 and compositional results in the Supplementary Material.

guarantee them to be positive, as is required for the bounds to be meaningful (this statement is shown explicitly in the Supplementary Material). Furthermore, by again considering the derived policy $\pi_f(a|s)$, we next provide a similar result for concave conditions, noting the difference in definitions between D and $\hat{D}$.

Lemma 4.4. *Consider the value of the policy $\pi_f(s) = \max_a f(Q(s,a))$ on the transformed task of interest, denoted by $\widetilde{Q}^{\pi_f}(s,a)$. The sub-optimality of π_f is then upper bounded by:*

$$\widetilde{Q}(s,a) - \widetilde{Q}^{\pi_f}(s,a) \leq \hat{D}(s,a) \tag{12}$$

where $\hat{D}$ is the value of the policy π_f in a task with reward

$$\hat{r}_D = \gamma \mathop{\mathbb{E}}_{s' \sim p} \mathop{\mathbb{E}}_{a' \sim \pi_f} \left[V_f(s') - f(Q(s',a')) + \hat{C}(s',a') \right]$$

Standard RL Results	
Transformation	Result
Linear Map:	$\widetilde{Q}(s,a) = kQ(s,a)$
Convex conditions:	$\widetilde{Q}(s,a) \geq f(Q(s,a))$
Concave conditions:	$\widetilde{Q}(s,a) \leq f(Q(s,a))$
OR Composition:	$\widetilde{Q}(s,a) \geq \max_k \{Q^{(k)}(s,a)\}$
AND Composition:	$\widetilde{Q}(s,a) \leq \min_k \{Q^{(k)}(s,a)\}$
NOT Gate:	$\widetilde{Q}(s,a) \geq -Q(s,a)$
Conical combination:	$\widetilde{Q}(s,a) \leq \sum_k \alpha_k Q^{(k)}(s,a)$

Table 1: **Standard Transfer Library.** Lemmas 4.1, 4.3 stated in Section 4 lead to a broad class of applicable transfer functions in standard RL. In this table we list several common examples which are demonstrated throughout the paper and in the Supplementary Materials. We show only one side of the bounds from Eq. (5), (10) which requires no additional training.

We remark that the conditions imposed on the function f are not very restrictive. For example, the Boolean functions and linear combinations considered in previous work are all included in our framework, while we also include novel transformations not considered in previous work (see Table 1). Furthermore, the conditions for f can be further relaxed if specific conditions are met. For the case of deterministic dynamics, the first condition is not required (f need not be convex nor concave).

We have shown that in the standard RL case, quite general conditions (convexity and sublinearity) lead to a wide class of applicable functions defining the Transfer Library. The conditions given in Lemmas 4.1 and 4.3 are straightforward to check for general functions. When given a primitive task defined by a reward function r, one can therefore bound the optimal Q function for a general transformation of the rewards, $f(r)$, when f obeys the conditions above. This new set of transformed tasks defines the Transfer Library from a given set of primitive tasks.

The previous (and following) results are presented for the case in which the primitive task Q-values are known *exactly*. In practice, however, this is not typically the case, even in tabular settings. In continuous environments where the use of function approximators is necessary, the error that is present in learned Q-values is further increased. To address this issue, we provide an extension of all double-sided bounds for the case where an ε-optimal estimate of the primitive task's Q-values is known, such that $|Q(s,a) - \bar{Q}(s,a)| \leq \varepsilon$ for all s, a. To derive such an extension, we further require that the composition function f is L-Lipschitz continuous (essentially a bounded first derivative), i.e. $|f(x_1) - f(x_2)| \leq L|x_1 - x_2|$ for all $x_1, x_2 \in X$, the domain of f (in the present case, the x_i are the primitive task's Q-values). To maintain the focus of the main text, we provide these results and the corresponding proofs in the Supplementary Material. We note that all functions listed in Table 1 and 2 are indeed 1-Lipschitz continuous.

4.2 GENERALIZATION TO COMPOSITION OF PRIMITIVE TASKS

The previous lemmas can be extended to the case of multi-variable transformations (see Supplementary Material for details), where $X \to \bigotimes X^{(k)}$ (the Cartesian product of primitive codomains). That is, with a function $F : \bigotimes X^{(k)} \to \mathbb{R}$ and a collection of M subtasks, $\{r^{(k)}(s,a)\}_{k=1}^{M}$, one can synthesize a new, **composition of subtasks**, with reward defined by $r^{(c)}(s,a) = F(r^{(1)}(s,a), \ldots, r^{(M)}(s,a))$.

In this vectorized format, F must obey the above conditions in each argument:

- F is convex (concave) in each argument,
- F is sublinear (superlinear) in each argument.

For the final conditions, we also require a similar vectorized inequality, which we spell out in detail in the Supplementary Material.

As an example of composition in standard RL, we consider the possible sums of reward functions, with each task having a positive weight associated to it.

In such a setup, the agent has learned to solve a set of primitive tasks, then it must solve a task with a new compositely-defined reward function, say $f\left(r^{(1)}, \ldots, r^{(M)}\right) \doteq \sum_{k=1}^{M} \alpha_k r^{(k)}$ for (possibly many) target tasks defined by the weights $\{\alpha_k\}$. To determine which bound is satisfied for such a composition function,

we look to the vectorized conditions above. This function is linear in all arguments, so we must only check the final condition. Since the inequality

$$\sum_k \alpha_k \max_a Q^{(k)}(s,a) \geq \max_a \sum_k \alpha_k Q^{(k)}(s,a) \quad (13)$$

holds for any set of $\alpha_k > 0$, this function conforms to the concave vectorized conditions, implying that $\widetilde{Q}(s,a) \leq f(Q^{(k)}(s,a)) = \sum_k \alpha_k Q^{(k)}(s,a)$. We can then use the right-hand side of this bound to calculate the associated state-value function ($V_f(s) = \max_a f(Q^{(k)}(s,a))$) and the associated greedy policy ($\pi_f(s) = \text{argmax}_a f(Q^{(k)}(s,a))$). This result agrees with an independent result by Nemecek and Parr [2021] (the upper bound in Theorem 1 therein) without accounting for approximation errors.

5 ENTROPY-REGULARIZED RL

5.1 TRANSFORMATION OF PRIMITIVE TASK

We will now extend the results obtained in the previous section to the case of entropy-regularized RL. Again we first consider the single-reward transformation $f(r)$ for some function f. Here we state the conditions that must be met by functions f, which define the Transfer Library for entropy-regularized RL.

We now use the following definitions in the subsequent (entropy-regularized RL) results. In the following results, we set $\beta = 1$ for brevity, and the expectation in the final condition is understood to be over actions, sampled from the prior policy. Full details can be found in the proofs provided in the Supplementary Material.

Lemma 5.1 (Convex Conditions)**.** *Given a primitive task with discount factor γ and a bounded, continuous transformation function $f : X \to \mathbb{R}$ which satisfies:*

1. *f is convex on its domain X[1];*
2. *f is sublinear:*
   - *(i) $f(x+y) \leq f(x) + f(y)$ for all $x, y \in X$*
   - *(ii) $f(\gamma x) \leq \gamma f(x)$ for all $x \in X$*
3. *$f(\log \mathbb{E} \exp \mathcal{Q}(s,a)) \leq \log \mathbb{E} \exp f(\mathcal{Q}(s,a))$ for all functions $\mathcal{Q} : \mathcal{S} \times \mathcal{A} \to X$.*

then the optimal action-value function for the transformed rewards, $\widetilde{Q}$, is now related to the optimal action-value function with respect to the original rewards by:

$$f(Q(s,a)) \leq \widetilde{Q}(s,a) \leq f(Q(s,a)) + C(s,a) \quad (14)$$

[3]Note that linear reward scaling can also be viewed as a linear scaling in the temperature parameter.

[4]This extends to the case $\sum_k \alpha_k \geq 1$ by composing with a linear scaling, which respects the same inequality.

Entropy-Regularized RL Results	
Transformation	Result
Linear Map, $k \in (0,1)$[3]:	$\widetilde{Q}(s,a) \geq kQ(s,a)$
Linear Map, $k > 1$:	$\widetilde{Q}(s,a) \leq kQ(s,a)$
Convex conditions:	$\widetilde{Q} \geq f(Q(s,a))$
Concave conditions:	$\widetilde{Q} \leq f(Q(s,a))$
OR Composition:	$\widetilde{Q}(s,a) \geq \max_k \{Q^{(k)}(s,a)\}$
AND Composition:	$\widetilde{Q}(s,a) \leq \min_k \{Q^{(k)}(s,a)\}$
NOT Gate:	$\widetilde{Q}(s,a) \geq -Q(s,a)$
Convex combination[4]:	$\widetilde{Q}(s,a) \leq \sum_k \alpha_k Q^{(k)}(s,a)$

Table 2: **Entropy-Regularized Transfer Library.** Lemmas 5.1, 5.3 lead to a broad class of applicable transfer functions in entropy-regularized RL. In this table we list several common examples which are demonstrated throughout the paper and in the Supplementary Materials. We show only one side of the bounds from Eq. (14), (16) which requires no additional training.

We note that C has the same definition as before, but with V_f replaced by its entropy-regularized analog: $V_f(s) \doteq \log \mathbb{E}_{a\sim\pi_0} \exp f(Q(s,a))$.

Lemma 5.2. *Consider the soft value of the policy $\pi_f(a|s) = \pi_0(a|s) \exp(f(Q(s,a)) - V_f(s))$ on the transformed task of interest, denoted by $\widetilde{Q}^{\pi_f}(s,a)$. The suboptimality of π_f is then upper bounded by:*

$$\widetilde{Q}(s,a) - \widetilde{Q}^{\pi_f}(s,a) \leq D(s,a) \quad (15)$$

where D is the soft value of the policy π_f with reward

$$r_D(s,a) = \gamma \mathop{\mathbb{E}}_{s'} \left[\max_b \{f(Q(s',b)) + C(s',b)\} - V_f(s') \right].$$

Conversely, for concave conditions we have

Lemma 5.3 (Concave Conditions)**.** *Given a primitive task with discount factor γ and a bounded, continuous transformation function $f : X \to \mathbb{R}$ which satisfies:*

1. *f is concave on its domain X[1];*
2. *f is superlinear:*
   - *(i) $f(x+y) \geq f(x) + f(y)$ for all $x, y \in X$*
   - *(ii) $f(\gamma x) \geq \gamma f(x)$ for all $x \in X$*
3. *$f(\log \mathbb{E} \exp \mathcal{Q}(s,a)) \geq \log \mathbb{E} \exp f(\mathcal{Q}(s,a))$ for all functions $\mathcal{Q} : \mathcal{S} \times \mathcal{A} \to X$.*

then the optimal action-value function for the transformed rewards obeys the following inequality:

$$f(Q(s,a)) - \hat{C}(s,a) \leq \widetilde{Q}(s,a) \leq f(Q(s,a)) \quad (16)$$

As in the preceding section, we provide a similar result for the derived policy π_f, given the concave conditions provided.

Lemma 5.4. *Consider the soft value of the policy $\pi_f(a|s)$ on the transformed task of interest, denoted by $\widetilde{Q}^{\pi_f}(s,a)$. The sub-optimality of π_f is then upper bounded by:*

$$\widetilde{Q}(s,a) - \widetilde{Q}^{\pi_f}(s,a) \leq \hat{D}(s,a) \tag{17}$$

where $\hat{D}$ satisfies the following recursive equation

$$\hat{D}(s,a) = \gamma \mathop{\mathbb{E}}_{s'\sim p} \mathop{\mathbb{E}}_{a'\sim\pi_f} \left(\hat{C}(s',a') + \hat{D}(s',a')\right). \tag{18}$$

Now, by taking $V_f(s)$ as the previously defined *soft* value function, the fixed points C and $\hat{C}$ have the same definitions as presented in Lemma 4.1 and 4.3, respectively with this new definition of V_f.

This final constraint (in Lemma 5.1 and 5.3) on f arises out of the requirements for extending the previous results to entropy-regularized RL. Although the final condition (similar to a log-convexity) appears somewhat cumbersome, we show that it is nevertheless possible to satisfy it for several non-trivial functions (Table 2). For instance, functions defining Boolean composition over subtasks ($\max(\cdot)$, $\min(\cdot)$), which have not been considered in previous entropy-regularized results [Haarnoja et al., 2018a, Van Niekerk et al., 2019] as well as new functional transformations such as the NOT gate (Table 2).

5.2 GENERALIZATION TO COMPOSITION OF PRIMITIVE TASKS

As we have done in the standard RL setting (Section 4.2), we can also extend the previous results to include compositionality: functions operating over multiple primitive tasks.

In this case, Haarnoja et al. [2018a] have demonstrated a special case of Lemma 5.3 for the composition function $f(\{r^{(k)}\}) = \sum_k \alpha_k r^{(k)}$ for convex weights α_k. This can also be shown in our framework by proving the final condition of Lemma 5.3 (since the others are automatic given that f is linear). This vectorized condition can be proven via Hölder's inequality.

Besides this previously studied composition function, we can now readily derive value function bounds for other transformations and compositions, for example Boolean compositions as defined previously. The corresponding results for entropy-regularized RL are summarized in Table 2.

6 EXPERIMENTS

To test our theoretical results using function approximators (FAs), we consider a deterministic "gridworld" MDP amenable to task composition[5]. Figure 1 shows the environments of the trained primitive tasks "6×6 L" and "6×6 D", whose reward functions are then combined to produce a composite task, "6×6 L OR D". The agent has 4 possible actions (in each of the cardinal directions) and begins at the green circle in all cases. The agent's goal is to navigate to the orange states which provide a reward. We note that these states are *not absorbing* unlike the cases considered in prior work. The red "X" indicates a penalizing state where the agent's episode is immediately terminated. Finally, a wall (black square) is added for the agent to navigate around. The primitive tasks are assumed to be solved with high accuracy (i.e. we assume the Q-values for primitive tasks to be exactly known). Although the domain is rather simple, we use such an experiment as a means of validating our theoretical results while gaining insight on the experimental effects of *clipping* (discussed below) during training.

With the primitive tasks solved, we now consider training on a target composite ("OR") task. We learn from scratch (with no prior information or bounds being applied) as our baseline (blue line, denoted "none", in Fig. 1 and Fig. 2).

To implement the derived bounds, we consider the one-sided bound, thereby not requiring further training. In this case (standard RL, "OR" composition) we have the following lower bound (see Table 1): $Q^{(\text{OR})} \geq \max\{Q^{(\text{L})}, Q^{(\text{D})}\}$. There are many ways to implement such a bound in practice. One naïve method is to simply clip the target network's new (proposed target) value to be within the allowed region for each of the $\widetilde{Q}(s,a)$ that are currently being updated. We term this method "hard clipping". Inspired by Section 3.2 of [Kim et al., 2022], we can also use an additional penalty by adding to the loss function the absolute value of bound violations that occur (the quantity "BV" defined in Eq. (19)). We term this method as "soft clipping". As mentioned by [Kim et al., 2022], this method of clipping could produce a new hyperparameter (the relative weight for this term relative to the Bellman residual). We keep this coefficient fixed (to unity) for simplicity, and we intend on exploring the possibility of a variable weight in future work.

$$\text{BV} \doteq ||\widetilde{Q}(s,a) - f(\{Q^{(k)}(s,a)\}_{k=1}^{M})|| \tag{19}$$

Similar to Eq. (21) of [Kim et al., 2022] we also considered a clipping at test-time only, with some differences in how the bounds are applied. This discrepancy is due to the difference in frameworks: [Kim et al., 2022] leverages the GPI framework, and in our setting we are learning a new policy from scratch while imposing said bounds. Our method is as follows: Whenever the agent acts greedily and samples from the policy network, it first applies (hard) clipping to the network's value; then the agent extracts an action via greedy argmax. We term this method of clipping as "test clipping".

[5] Source code available at `https://github.com/JacobHA/Q-Bounding-in-Compositional-RL`

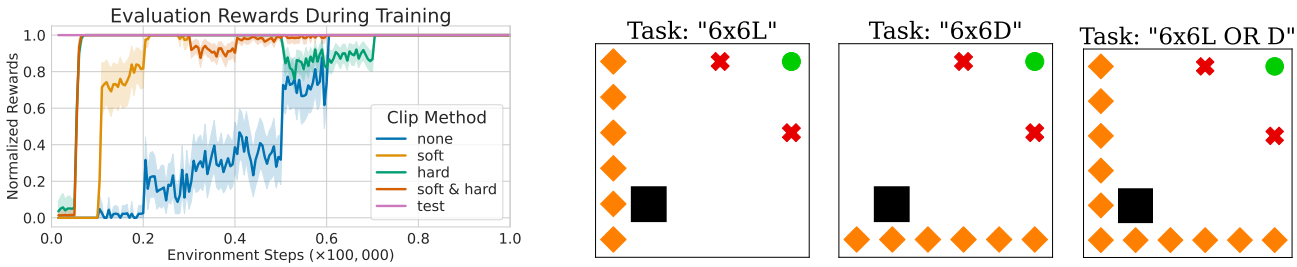

Figure 1: In the first panel, we show learning curves for each of the clipping methods proposed, averaged over 50 trials, with a 95% confidence interval shown in the shaded region. In the next two panels, we depict the primitive tasks with rewarding states (orange diamonds) on the left side and bottom side of the maze, respectively. In the rightmost panel, we show the composite task of interest, with the multivariable "OR" composition function $\max_k\{...\}$ used to define its reward function. The agent first solves the two primitive tasks with a deep Q-network (DQN, as implemented by Stable-Baselines3 [Raffin et al., 2021]) with results shown in the first panel. Training hyperparameters are given in the Supplementary Material.

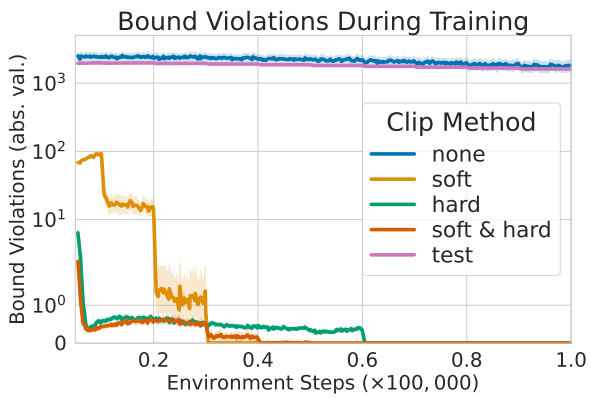

Figure 2: Mean bound violation, shown with shaded 95% confidence intervals. The bound violation measures the difference between the Q-network's estimate and the allowed bound $\widetilde{Q} \geq f(Q)$ for a given batch during training. Note that the x-axis corresponds to zero bound violation (symlog y-scale). Although "test" clipping does very well in terms of its evaluation performance, it does not respect the bounds, even long after its apparent convergence.

The results for each method (as well as a combination of both hard and soft clipping) are shown in Fig. 1 and 2.

Interestingly, we find that by directly incorporating the bound violations into the loss function (via the "soft" clipping mechanism); the bound violations most quickly become (and remain) zero (Fig. 2) as opposed to the other methods considered. We find that reduction in bound violation also generally correlates with a high evaluation reward during training. One exception to this observation (for the particular environment shown) is the case of "test" clipping.

For this particular composition, either primitive task will solve the composite task, thus yielding high evaluation rewards (Fig. 1). However, the Q-values are not accurate, which leads to a high frequency of clipping, comparable to the baseline without clipping (Fig. 2). In order to ensure the agent has learned accurate Q-values, it is therefore important to monitor the bound violations rather than only the evaluation performance which may not be representative of convergence of Q-values.

7 DISCUSSION

In summary, we have established a general theoretical treatment for functional transformation and composition of primitive tasks. This extends the scope of previous work, which has primarily focused on isolated instances of reward transformations and compositions without general structure. Additionally, we have theoretically addressed the broader setting of stochastic dynamics, with rewards varying on both terminal and non-terminal (i.e. boundary and interior) states. In this work, we have shown that it is possible to derive a general class of functions which obey transfer bounds in standard and entropy-regularized RL beyond those cases discussed in previous work. In particular, we show that by using the same functional form on the optimal Q functions as used on the reward, we can bound the transformed optimal Q function. The derived bound can then be used to calculate a zero-shot solution. We have used these functions to define a Transfer Library: a set of tasks which can immediately be addressed by our bounds. Since our approach via the optimal backup equation is general, we apply it to both standard RL and entropy-regularized RL.

The newly-defined fixed point C ($\hat{C}$) has an interesting interpretation. Rather than simply being an arbitrary function, for both the standard RL and entropy-regularized RL bounds, C represents an optimal value function for a standard RL task with reward function given by r_C ($\hat{r}_C$ for $\hat{C}$).

The function C bounds the total gap between $f(Q(s,a))$ and $\widetilde{Q}(s,a)$ at the level of state-actions. We also note the simple relationship between reward functions $r_C = -\hat{r}_C$.

The fixed point D ($\hat{D}$) is not an optimal value function, but the value of the zero-shot policy π_f in some other auxil-

iary task. The auxiliary task takes various "rewards", e.g. the function $\gamma\hat{C}$ in Lemma 5.4. Although for general functions f, the rewards do not have a simple interpretation (i.e. Rényi divergence between two policies as in [Haarnoja et al., 2018a]), we see that r_C essentially measures the non-linearities of the composition function f with respect to the given dynamics, and hence accounts for the errors made in using the bounding conditions of f. Furthermore, we can bound C (and thus the difference between the optimal value and the suggested zero-shot approximation $f(Q)$) in a simple way: by bounding the rewards corresponding to C. By simply calculating the maximum of r_C for example, one easily finds $C(s,a) \leq \frac{1}{1-\gamma}\max_{s,a} r_C(s,a)$ (and similarly for $\hat{C}$).

Interestingly, Mann and Choe [2013] have shown the provable usefulness of using an upper bound when used for "warmstarting" the training in new domains. In particular, it appears that $f(Q)$ (for the concave conditions) is related to their proposed "α-weak admissible heuristic" for $\widetilde{\mathcal{T}}$. In future work, we hope to precisely connect to such theoretical results in order to obtain provable benefits to our derived bounds. Experimentally we have observed that this warmstarting procedure does indeed improve convergence times, however a detailed study of this effect is beyond the scope of the present work and will be explored in future work. The derived results have also been used to devise protocols for clipping which improve performance and reduce variance during training based on the experiments presented.

In the future, we hope that the class of functions discussed in this work will be broadened further, allowing for a larger class of non-trivial zero-shot bounds for the Transfer Library. By adding these known transformations and compositions to the Transfer Library, the RL agent will be able to approach significantly more novel tasks without the need for further training.

The current research has also emphasized questions for transfer learning in this context, such as: *Which primitives should be prioritized for learning?* (Discussed in Tasse et al. [2021], Nemecek and Parr [2021], Alver and Precup [2021].) *What other functions can be used for transfer? How tight are these bounds? How does the Transfer Library depend on the parameters γ and β?*

In this work, we provide general bounds for the discrete MDP setting and an extension of the theory to continuous state-action spaces is deferred to future work. It would be of interest to explore if it is possible to prove general bounds for this extension, given sufficient smoothness conditions on the dynamics and the function of transformation. Other extensions can be considered as well, for instance: the applicability to other value-based or actor-critic methods, the warmstarting of function approximators, learning the C and D functions, and adjusting the "soft" clipping weight parameter.

In future work, we also aim to discover other functions satisfying the derived conditions; and will attempt to find necessary (rather than sufficient) conditions that classify the functions $f \in \mathcal{F}$. It would be of interest to explore if extensions of the current approach can further enable agents to expand and generalize their knowledge base to solve complex dynamic tasks in Deep RL.

Acknowledgements

The authors would like to thank the anonymous reviewers for their helpful comments and suggestions. JA, AA, and RVK acknowledge funding support from the NSF through Award No. DMS-1854350. VM and ST acknowledge funding support from the NSF through Award No. 2246221. JA would like to acknowledge the use of the supercomputing facilities managed by the Research Computing Department at the University of Massachusetts Boston. The work of JA and AA was supported in part by the College of Science and Mathematics Dean's Doctoral Research Fellowship through fellowship support from Oracle, project ID R20000000025727. ST and VM acknowledge support from the Alliance Innovation Lab in Silicon Valley.

References

Jacob Adamczyk, Argenis Arriojas, Stas Tiomkin, and Rahul V Kulkarni. Utilizing prior solutions for reward shaping and composition in entropy-regularized reinforcement learning. arXiv preprint arXiv:2212.01174, 2022.

Safa Alver and Doina Precup. Constructing a good behavior basis for transfer using generalized policy updates. arXiv preprint arXiv:2112.15025, 2021.

André Barreto, Will Dabney, Rémi Munos, Jonathan J Hunt, Tom Schaul, Hado P van Hasselt, and David Silver. Successor features for transfer in reinforcement learning. Advances in Neural Information Processing Systems, 30, 2017.

André Barreto, Diana Borsa, Shaobo Hou, Gheorghe Comanici, Eser Aygün, Philippe Hamel, Daniel Toyama, Shibl Mourad, David Silver, Doina Precup, et al. The option keyboard: Combining skills in reinforcement learning. Advances in Neural Information Processing Systems, 32, 2019.

Greg Brockman, Vicki Cheung, Ludwig Pettersson, Jonas Schneider, John Schulman, Jie Tang, and Wojciech Zaremba. Openai gym. arXiv preprint arXiv:1606.01540, 2016.

Peter Dayan. Improving generalization for temporal difference learning: The successor representation. Neural computation, 5(4):613–624, 1993.

Benjamin Eysenbach and Sergey Levine. Maximum entropy RL (provably) solves some robust RL problems. In International Conference on Learning Representations, 2022. URL https://openreview.net/forum?id=PtSAD3caaA2.

Tuomas Haarnoja, Vitchyr Pong, Aurick Zhou, Murtaza Dalal, Pieter Abbeel, and Sergey Levine. Composable deep reinforcement learning for robotic manipulation. In 2018 IEEE International Conference on Robotics and Automation (ICRA). IEEE, May 2018a. doi: 10.1109/icra.2018.8460756. URL https://doi.org/10.1109/icra.2018.8460756.

Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In Jennifer Dy and Andreas Krause, editors, Proceedings of the 35th International Conference on Machine Learning, volume 80 of Proceedings of Machine Learning Research, pages 1861–1870. PMLR, 10–15 Jul 2018b. URL https://proceedings.mlr.press/v80/haarnoja18b.html.

Tuomas Haarnoja, Aurick Zhou, Kristian Hartikainen, George Tucker, Sehoon Ha, Jie Tan, Vikash Kumar, Henry Zhu, Abhishek Gupta, Pieter Abbeel, and Sergey Levine. Soft actor-critic algorithms and applications, 2019.

Godfrey Harold Hardy, John Edensor Littlewood, George Pólya, György Pólya, et al. Inequalities. Cambridge university press, 1952.

Zhang-Wei Hong, Ge Yang, and Pulkit Agrawal. Bilinear value networks. arXiv preprint arXiv:2204.13695, 2022.

Jonathan Hunt, Andre Barreto, Timothy Lillicrap, and Nicolas Heess. Composing entropic policies using divergence correction. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, Proceedings of the 36th International Conference on Machine Learning, volume 97 of Proceedings of Machine Learning Research, pages 2911–2920. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/hunt19a.html.

Jaekyeom Kim, Seohong Park, and Gunhee Kim. Constrained gpi for zero-shot transfer in reinforcement learning. In Advances in Neural Information Processing Systems, 2022. URL https://openreview.net/forum?id=sWNT5lT7l9G.

Kimin Lee, Michael Laskin, Aravind Srinivas, and Pieter Abbeel. Sunrise: A simple unified framework for ensemble learning in deep reinforcement learning. In International Conference on Machine Learning, pages 6131–6141. PMLR, 2021.

Sergey Levine. Reinforcement Learning and Control as Probabilistic Inference: Tutorial and Review. arXiv, May 2018. URL https://arxiv.org/abs/1805.00909v3.

Timothy A. Mann and Yoonsuck Choe. Directed exploration in reinforcement learning with transferred knowledge. In Marc Peter Deisenroth, Csaba Szepesvári, and Jan Peters, editors, Proceedings of the Tenth European Workshop on Reinforcement Learning, volume 24 of Proceedings of Machine Learning Research, pages 59–76, Edinburgh, Scotland, 30 Jun–01 Jul 2013. PMLR. URL https://proceedings.mlr.press/v24/mann12a.html.

Mark Nemecek and Ronald Parr. Policy caches with successor features. In International Conference on Machine Learning, pages 8025–8033. PMLR, 2021.

Shubham Pateria, Budhitama Subagdja, Ah-hwee Tan, and Chai Quek. Hierarchical reinforcement learning: A comprehensive survey. ACM Computing Surveys (CSUR), 54(5):1–35, 2021.

Xue Bin Peng, Michael Chang, Grace Zhang, Pieter Abbeel, and Sergey Levine. Mcp: Learning composable hierarchical control with multiplicative compositional policies. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, Advances in Neural Information Processing Systems 32, pages 3681–3692. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/file/95192c98732387165bf8e396c0f2dad2-Paper.pdf.

Antonin Raffin, Ashley Hill, Adam Gleave, Anssi Kanervisto, Maximilian Ernestus, and Noah Dormann. Stable-baselines3: Reliable reinforcement learning implementations. Journal of Machine Learning Research, 22(268):1–8, 2021. URL http://jmlr.org/papers/v22/20-1364.html.

Andrew M. Saxe, Adam C. Earle, and Benjamin Rosman. Hierarchy through composition with multitask LMDPs. In Doina Precup and Yee Whye Teh, editors, Proceedings of the 34th International Conference on Machine Learning, volume 70 of Proceedings of Machine Learning Research, pages 3017–3026. PMLR, 06–11 Aug 2017. URL https://proceedings.mlr.press/v70/saxe17a.html.

Richard S Sutton and Andrew G Barto. Reinforcement learning: An introduction. MIT press, 2018.

Richard S. Sutton, Doina Precup, and Satinder Singh. Between mdps and semi-mdps: A framework for temporal abstraction in reinforcement learning. Artificial

Intelligence, 112(1):181–211, 1999. ISSN 0004-3702. doi: https://doi.org/10.1016/S0004-3702(99)00052-1. URL https://www.sciencedirect.com/science/article/pii/S0004370299000521.

Geraud Nangue Tasse, Steven James, and Benjamin Rosman. A boolean task algebra for reinforcement learning. Advances in Neural Information Processing Systems, 33: 9497–9507, 2020.

Geraud Nangue Tasse, Steven James, and Benjamin Rosman. Generalisation in lifelong reinforcement learning through logical composition. In Deep RL Workshop NeurIPS 2021, 2021. URL https://openreview.net/forum?id=kO-Cgmasm7.

Matthew E. Taylor and Peter Stone. Transfer learning for reinforcement learning domains: A survey. Journal of Machine Learning Research, 10(56):1633–1685, 2009. URL http://jmlr.org/papers/v10/taylor09a.html.

Emanuel Todorov. Compositionality of optimal control laws. In Y. Bengio, D. Schuurmans, J. Lafferty, C. Williams, and A. Culotta, editors, Advances in Neural Information Processing Systems, volume 22. Curran Associates, Inc., 2009. URL https://proceedings.neurips.cc/paper/2009/file/3eb71f6293a2a31f3569e10af6552658-Paper.pdf.

Benjamin Van Niekerk, Steven James, Adam Earle, and Benjamin Rosman. Composing value functions in reinforcement learning. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, Proceedings of the 36th International Conference on Machine Learning, volume 97 of Proceedings of Machine Learning Research, pages 6401–6409. PMLR, 06 2019. URL https://proceedings.mlr.press/v97/van-niekerk19a.html.

Brian D. Ziebart. Modeling Purposeful Adaptive Behavior with the Principle of Maximum Causal Entropy. Carnegie Mellon University, 12 2010. doi: 10.1184/R1/6720692.v1. URL https://kilthub.cmu.edu/articles/thesis/Modeling_Purposeful_Adaptive_Behavior_with_the_Principle_of_Maximum_Causal_Entropy/6720692.