Bounding the Optimal Value Function in Compositional Reinforcement Learning (Supplementary Material)

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

1 INTRODUCTION

In the following, we discuss the results of additional experiments in the four room domain. In these experiments, we want to answer the following questions:

- How do the optimal policies and value functions compare to those calculated from the zero-shot approximations using the derived bounds?
- What are other examples of compositions and functional transformations that can be analyzed using our approach?
- Does warmstarting (using the derived bounds for initialization) in the tabular case improve the convergence?

To address these issues, we modify OpenAI's frozen lake environment Brockman et al. [2016] to allow for stochastic dynamics.

In the tabular experiments, numerical solutions for the optimal Q functions were obtained by solving the Bellman backup equations iteratively. Iterations are considered converged once the maximum difference between successive iterates is less than 10^{-10}.

Beyond the motivating example shown in the main text, we have included video files demonstrating a full range of zero-shot compositions with convex weights between the Bottom Left (BL) room and Bottom Right (BR) room subtasks, in both entropy-regularized ($\beta = 5$) and standard RL with deterministic dynamics. These videos, along with all code for the above experiments are made publicly available at a repository on `https://github.com/JacobHA/Q-Bounding-in-Compositional-RL`.

2 EXPERIMENTS

2.1 FUNCTION APPROXIMATORS

For function approximator experiments (as shown in the main text), we use the DQN implementation from Stable-Baselines3 Raffin et al. [2021]. We first fully train the subtasks (seen in Fig. 1 of the main text). Then, we perform hyperparameter sweeps for each possible clipping option. Several hyperparameters are kept fixed (Table 1), and we sweep with the range and distribution shown below in Table 2. Finally, we use the optimal hyperparameters (as measured by those which maximize the accumulated reward throughout training). These values are shown in Table 3.

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

Table 1: Hyperparameters shared by all Deep Q Networks

Hyperparameter	Value
Buffer Size	1,000,000
Discount factor, γ	0.99
$\epsilon_{\text{initial}}$	1.0
ϵ_{final}	0.05
"learning starts"	5,000
Target Update Interval	10000

Table 2: Hyperparameter Ranges Used for Finetuning

Hyperparameter	Sampling Distribution	Min. Value	Max. Value
Learning Rate	Log Uniform	10^{-4}	10^{-1}
Batch Size	Uniform	32	256
Exploration Fraction	Uniform	0.1	0.3
Polyak Update, τ	Uniform	0.5	1.0

Table 3: Hyperparameters used for different clipping methods

Hyperparameter	None	Soft	Hard	Soft-Hard
Learning Rate	7.825×10^{-4}	3.732×10^{-3}	1.457×10^{-3}	3.184×10^{-3}
Batch Size	245	247	146	138
Exploration Fraction	0.137	0.1075	0.1243	0.1207
Polyak Update, τ	0.9107	0.9898	0.5545	0.7682

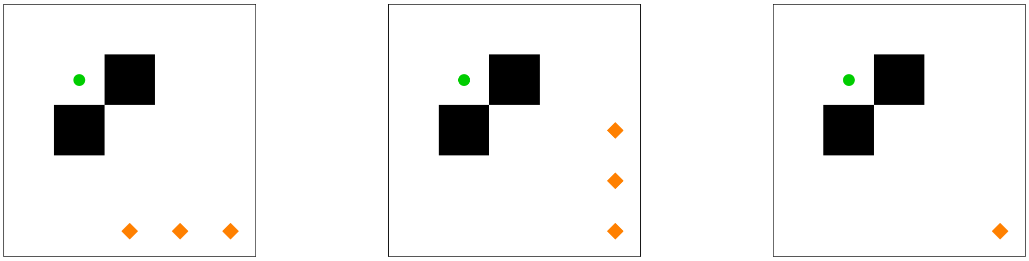
2.2 TABULAR EXPERIMENTS

In these experiments we will demonstrate on simple discrete environments the effect of increasingly stochastic dynamics and increasingly dense rewards. As a proxy for measuring the usefulness or accuracy of the bound $f(Q)$, we calculate the mean difference between $f\left(Q(s,a)\right) - \widetilde{Q}(s,a)$, as well as the mean Kullback-Liebler (KL) divergence between π (the true optimal policy) and π_f, the policy derived from the bound $f(Q)$. The proceeding experiments are situated in the entropy-regularized formalism (unless $\beta = \inf$ as shown in Fig. 5) with the uniform prior policy $\pi_0(a|s) = 1/|\mathcal{A}|$.

2.2.1 Stochasticity of Dynamics

In this experiment, we investigate the effect of stochastic dynamics on the bounds. Specifically, we vary the probability that taking an action will result in the intended action. This is equivalent to a slip probability.

We notice in the following plots that at near-deterministic dynamics the bound becomes tighter. We also remark that the Kullback-Liebler divergence is lowest in very highly-stochastic environments. This is because for any $\beta > 0$, the cost of changing the policy π away from the prior policy is not worth it: the dynamics are so stochastic that there will be no considerable difference in trajectories even if significant controls (nearly deterministic choices) are applied via π.

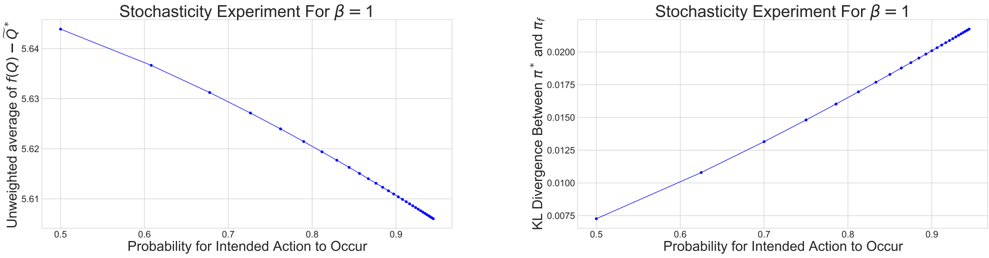


(a) Task 1

(b) Task 2

(c) Task 3: AND composition

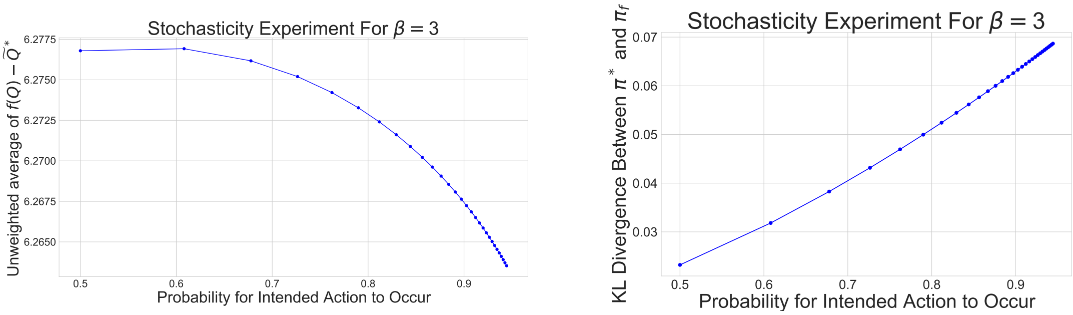
Figure 1: Reward functions for a simple maze domain; used for stochasticity experiments. We place reward (whose cost is half the default step penalty of -1) at the edges of the room, denoted by an orange diamond.



(a)

(b)

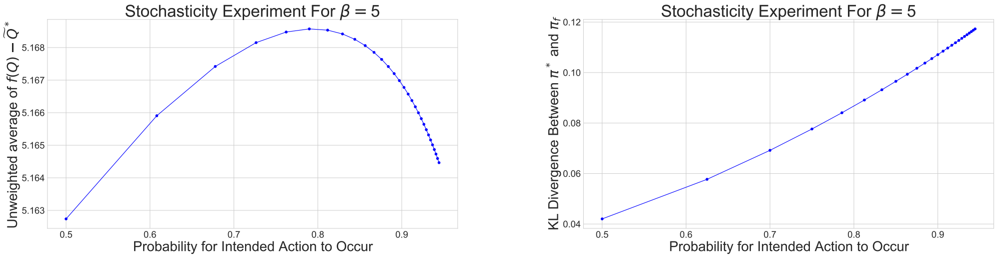
Figure 2: $\beta = 1$ KL divergence between π and π_f and average difference between optimal Q function and presented bound.



(a)

(b)

Figure 3: $\beta = 3$ KL divergence between π and π_f and average difference between optimal Q function and presented bound.



(a)

(b)

Figure 4: $\beta = 5$ KL divergence between π and π_f and average difference between optimal Q function and presented bound.

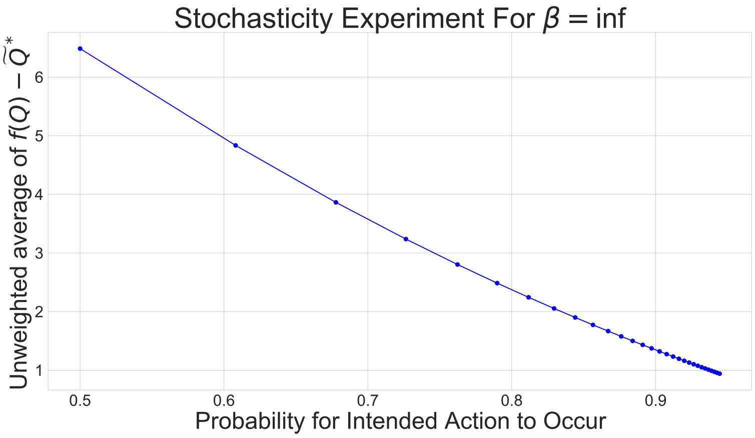
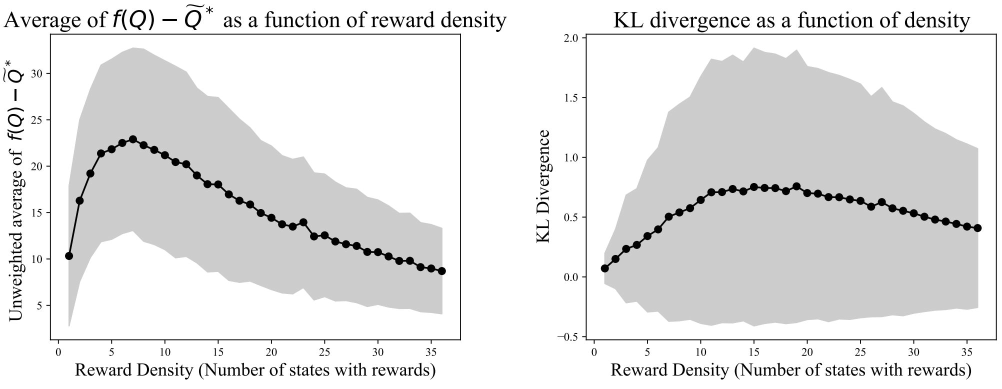


Figure 5: $\beta = \mathrm{inf}$, standard RL. Average difference between optimal Q function and presented bound. Note that we do not plot a KL divergence for this case as π is greedy and hence the divergence is always infinite.

2.3 SPARSITY OF REWARDS

In this experiment, we consider an empty environment ($|S| \times |S|$ empty square) with reward $r = 0$ everywhere and deterministic dynamics. No other rewards or obstacles are present. Then fix an integer $0 < n < |S|$. Drawing randomly (without repetition), we choose one of the states of the environment to grant a reward, drawn uniformly between $(0, 1)$. We do this again for another copy of the empty environment.

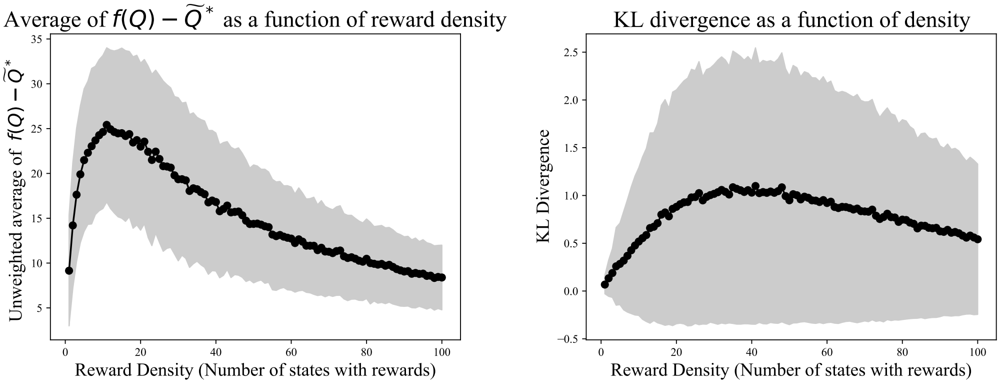
We then compose these two (randomly generated as described) subtasks by using a simple average $F(r^{(1)}, r^{(2)}) = 0.5r^{(1)} + 0.5r^{(2)}$. We have used $\beta = 5$ for all experiments in this subsection.



(a)

(b)

Figure 6: 6×6 environment. KL divergence between π and π_f and average difference between optimal Q function and presented bound, with the shaded region representing one standard deviation over 1000 runs.



(a)

(b)

Figure 7: 10×10 environment. KL divergence between π and π_f and average difference between optimal Q function and presented bound, with the shaded region representing one standard deviation over 1000 runs.

Interestingly, we find a somewhat universal behavior, in that there is a certain level of density which makes the bound a poor approximation to the true Q function. We also note that the bound is a better approximation at low densities.

3 BOOLEAN COMPOSITION DEFINITIONS

In this section, we explicitly define the action of Boolean operators on subtask reward functions. These definitions are similar to those used by Tasse et al. [2020].

Definition 3.1 (OR Composition)**.** Given subtask rewards $\{r^{(1)}, r^{(2)}, \ldots, r^{(M)}\}$, the OR composition among them is given by the *maximum* over all subtasks, at each state-action pair:

$$r^{(\text{OR})}(s,a) = \max_k r^{(k)}(s,a). \tag{1}$$

Definition 3.2 (AND Composition)**.** Given subtask rewards $\{r^{(1)}, r^{(2)}, \ldots, r^{(M)}\}$, the AND composition among them is given by the *minimum* over all subtasks, at each state-action:

$$r^{(\text{AND})}(s,a) = \min_k r^{(k)}(s,a). \tag{2}$$

Definition 3.3 (NOT Gate)**.** Given a subtask reward function r, applying the NOT gate transforms the reward function by negating all rewards (i.e. rewards → costs):

$$r^{(\text{NOT})}(s,a) = -r(s,a), \tag{3}$$

The proofs in all subsequent sections follow an inductive form based on the Bellman backup equation, whose solution converges to the optimal Q function. This is a similar approach as employed by Haarnoja et al. [2018a] and Hunt et al. [2019], but with the extension to all applicable functions; rather than (linear) convex combinations.

4 PROOFS FOR STANDARD RL

Let X be the codomain for the Q function of the primitive task ($Q : \mathcal{S} \times \mathcal{A} \to X \subseteq \mathbb{R}$).

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 (for stochastic dynamics);*
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. *$f(\max_a \mathcal{Q}(s,a)) \leq \max_a f(\mathcal{Q}(s,a))$ for all $\mathcal{Q} : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$.*

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:

$$f(Q(s,a)) \leq \widetilde{Q}(s,a) \leq f(Q(s,a)) + C(s,a) \tag{4}$$

where C is the optimal value function for a task with reward

$$r_C(s,a) = f(r(s,a)) + \gamma \mathbb{E}_{s'} V_f(s') - f(Q(s,a)). \tag{5}$$

Proof. We will prove all inequalities by induction on the number of backup steps, N. We start with the lower bound $\widetilde{Q} \geq f(Q)$. The base case, $N = 1$ is trivial since $f(r(s,a)) = f(r(s,a))$. The inductive step is the assumption $\widetilde{Q}^{(N)}(s,a) \geq f(Q^{(N)}(s,a))$ for some $N > 1$. In the case of standard RL, the Bellman backup equation for transformed rewards is given by:

$$\widetilde{Q}^{(N+1)}(s,a) = f\left(r(s,a)\right) + \gamma \mathbb{E}_{s' \sim p(s'|s,a)} \max_{a'} \widetilde{Q}^{(N)}(s',a') \tag{6}$$

Using the inductive assumption,

$$\widetilde{Q}^{(N+1)}(s,a) \geq f\left(r(s,a)\right) + \gamma \mathbb{E}_{s' \sim p(s'|s,a)} \max_{a'} f\left(Q^{(N)}(s',a')\right) \tag{7}$$

The condition $v_f(s) \geq f(v(s))$ is used on the right hand side to give:

$$\widetilde{Q}^{(N+1)}(s,a) \geq f\left(r(s,a)\right) + \gamma \mathbb{E}_{s' \sim p(s'|s,a)} f\left(\max_{a'} Q^{(N)}(s',a')\right) \tag{8}$$

Since f is convex, we use Jensen's inequality to factor it out of the expectation. Note that this condition on f is only required for stochastic dynamics. The error introduced by swapping these operators is characterized by the "Jensen's gap" for the transformation function f.

$$\widetilde{Q}^{(N+1)}(s,a) \geq f\left(r(s,a)\right) + \gamma f\left(\mathbb{E}_{s' \sim p(s'|s,a)} \max_{a'} Q^{(N)}(s',a')\right) \tag{9}$$

Finally, using both sublinearity conditions

$$\widetilde{Q}^{(N+1)}(s,a) \geq f\left(r(s,a) + \gamma \mathbb{E}_{s' \sim p(s'|s,a)} \max_{a'} Q^{(N)}(s',a')\right) \tag{10}$$

where the right-hand side is simply $f(Q^{(N+1)}(s,a))$. Since this inequality holds for all N, we take the limit $N \to \infty$ wherein $Q^{(N)}$ converges to the optimal Q-function. For the right-hand side of Eq. (10), we thus have (by continuity of f):

$$\lim_{N\to\infty} f\left(Q^{(N)}(s,a)\right) = f\left(\lim_{N\to\infty} Q^{(N)}(s,a)\right) = f(Q(s,a)) \tag{11}$$

where $Q(s,a)$ is the optimal action value function for the primitive task. Combined with the limit of the left-hand side, we arrive at the desired inequality:

$$\widetilde{Q}(s,a) \geq f\left(Q(s,a)\right). \tag{12}$$

This completes the proof of the lower bound. To prove the upper bound we again use induction on the backup equation of $\widetilde{Q}^{(N)}$. We wish to show $\widetilde{Q}^{(N)} \leq f\left(Q(s,a)\right) + C^{(N)}(s,a)$ holds for all N, with the definition of C provided in Lemma 4.1.

Let f satisfy the convex conditions. Consider the backup equation for $\widetilde{Q}$. Again, the base case ($N = 1$) is trivially satisfied with equality. Using the inductive assumption, we find

$$\begin{aligned}
\widetilde{Q}^{(N+1)}(s,a) &= f(r(s,a)) + \gamma \mathop{\mathbb{E}}_{s'} \max_{a'} \widetilde{Q}^{(N)}(s',a') \\
&\leq f(r(s,a)) + \gamma \mathop{\mathbb{E}}_{s'} \max_{a'} \left(f(Q(s',a')) + C^{(N)}(s',a')\right) \\
&\leq f(r(s,a)) + \gamma \mathop{\mathbb{E}}_{s'} \max_{a'} f(Q(s',a')) + \gamma \mathbb{E} \max_{a'} C^{(N)}(s',a') \\
&= f(Q(s,a)) + \left[f(r_i) + \gamma \mathop{\mathbb{E}}_{s'} V_f(s') - f(Q(s,a))\right] + \gamma \mathop{\mathbb{E}}_{s'} \max_{a'} C^{(N)}(s',a') \\
&= f(Q(s,a)) + C^{(N+1)}(s,a)
\end{aligned}$$

□

At this point, we verify that $C(s,a) > 0$ which ensures the double-sided bounds above are valid.

To do so, we can simply bound the reward function $r_C(s,a)$. By determining $r_C(s,a) > 0$, this will ensure $C(s,a) > \min r_C/(1-\gamma) > 0$.

$$\begin{aligned}
r_C(s,a) &= f(r(s,a)) + \gamma \mathbb{E}_{s'} V_f(s') - f(Q(s,a)) \\
&\geq f(r(s,a)) + \gamma \mathbb{E}_{s'} f(V(s')) - f(Q(s,a)) \\
&\geq f(r(s,a)) + f(\gamma \mathbb{E}_{s'} V(s')) - f(Q(s,a)) \\
&\geq f(r(s,a) + \gamma \mathbb{E}_{s'} V(s')) - f(Q(s,a)) \\
&\geq 0
\end{aligned}$$

where each line follows from the required conditions in Lemma 4.1. A similar proof holds for showing the quantities $\hat{C}$, D, $\hat{D}$ are all non-negative.

We now prove the policy evaluation bound for standard RL.

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:*

$$\widetilde{Q}(s,a) - \widetilde{Q}^{\pi_f}(s,a) \leq D(s,a) \tag{13}$$

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

$$r_D = \gamma \mathbb{E}_{s',a' \sim \pi_f}\left[\max_a \big(f(Q(s',a')) + C(s',a')\big) - f(Q(s,a))\right]$$

Proof. We will again prove this bound by induction on steps in the Bellman backup equation for the value of π_f, as given by the following fixed point equation:

$$\widetilde{Q}^{\pi_f}(s,a) = f\left(r(s,a)\right) + \gamma \mathop{\mathbb{E}}_{s',a' \sim \pi_f} \widetilde{Q}^{\pi_f}(s',a') \tag{14}$$

We consider the following initial conditions: $\widetilde{Q}^{\pi_f(0)}(s,a) = \widetilde{Q}(s,a), D(s,a) = 0$. We note that there is freedom in the choice of initial conditions, as the final statement (regarding the optimal value functions) holds regardless of initialization. As usual, the base case is trivially satisfied. We will now show that the equivalent inequality

$$\widetilde{Q}^{\pi_f(N)}(s,a) \geq \widetilde{Q}(s,a) - D^{(N)}(s,a) \tag{15}$$

holds for all N. Similar to the previous proofs, we will subsequently take the limit $N \to \infty$ to recover the desired result.

To do so, we consider the next step of the Bellman backup, and apply the inductive hypothesis:

$$\begin{aligned}
\widetilde{Q}^{\pi_f(N+1)}(s,a) &= f\left(r(s,a)\right) + \gamma \mathop{\mathbb{E}}_{s',a' \sim \pi_f} \left(\widetilde{Q}^{\pi_f(N)}(s',a')\right) && (16)\\
&\geq f\left(r(s,a)\right) + \gamma \mathop{\mathbb{E}}_{s',a' \sim \pi_f} \left(\widetilde{Q}(s',a') - D^{(N)}(s',a')\right) && (17)\\
&\geq f\left(r(s,a)\right) + \gamma \mathop{\mathbb{E}}_{s',a' \sim \pi_f} \left(f\left(Q(s',a')\right) - D^{(N)}(s',a')\right) && (18)\\
&= f\left(r(s,a)\right) + \gamma \mathop{\mathbb{E}}_{s'} \widetilde{V}(s') + \gamma \mathop{\mathbb{E}}_{s',a' \sim \pi_f} \left(f\left(Q(s',a')\right) - D^{(N)}(s',a') - \widetilde{V}(s')\right) && (19)\\
&\geq \widetilde{Q}(s,a) + \gamma \mathop{\mathbb{E}}_{s',a' \sim \pi_f} \left(f\left(Q(s',a')\right) - D^{(N)}(s',a') - \max_{a'} \left\{f\left(Q(s',a')\right) + C(s',a')\right\}\right) && (20)\\
&= \widetilde{Q}(s,a) - \gamma \mathop{\mathbb{E}}_{s',a' \sim \pi_f} \left(\max_{a'} \left\{f\left(Q(s',a')\right) + C(s',a')\right\} - f\left(Q(s',a')\right) + D^{(N)}(s',a')\right) && (21)\\
&= \widetilde{Q}(s,a) - \left(r_D(s,a) + \gamma \mathop{\mathbb{E}}_{s',a' \sim \pi_f} D^{(N)}(s',a')\right) && (22)\\
&= \widetilde{Q}(s,a) - D^{(N+1)}(s,a) && (23)
\end{aligned}$$

The third and fifth line follow from the previous bounds (Lemma 4.1). In the limit $N \to \infty$, we can thus see that the fixed point D corresponds to the policy evaluation for π_f in an environment with reward function r_D. □

Now we prove similar results, but for the "concave conditions" presented in the main text.

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 (for stochastic dynamics);*
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:

$$f(Q(s,a)) - \hat{C}(s,a) \leq \widetilde{Q}(s,a) \leq f(Q(s,a)) \tag{24}$$

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

$$\hat{r}_C(s,a) = f(Q(s,a)) - f(r(s,a)) - \gamma \mathop{\mathbb{E}}_{s'} V_f(s') \tag{25}$$

Proof. The proof of $\widetilde{Q} \leq f(Q)$ is the same as the preceding theorem's lower bound but with all inequalities reversed. To prove the upper bound involving $\hat{C}$, we use a similar approach

$$\begin{aligned}
\widetilde{Q}^{(N+1)}(s,a) &= f(r(s,a)) + \gamma \mathop{\mathbb{E}}_{s'} \max_{a'} \widetilde{Q}^{(N)}(s',a') \\
&\geq f(r(s,a)) + \gamma \mathop{\mathbb{E}}_{s'} \max_{a'} \left(f(Q(s',a')) - \hat{C}^{(N)}(s',a') \right) \\
&\geq f(r(s,a)) + \gamma \mathop{\mathbb{E}}_{s'} \left(\max_{a'} f(Q(s',a')) - \max_{a'} \hat{C}^{(N)}(s',a') \right) \\
&= f(Q(s,a)) - \left[f(Q(s,a)) - f(r(s,a)) - \gamma \mathop{\mathbb{E}}_{s'} V_f(s') + \gamma \mathop{\mathbb{E}}_{s'} \max_{a'} \hat{C}^{(N)}(s',a') \right] \\
&= f(Q(s,a)) - \hat{C}^{(N+1)}(s,a)
\end{aligned}$$

The second line follows from the inductive hypothesis. The third line follows from the $\max$ of a difference. In the penultimate line, we add and subtract $f(Q)$, and identify the definitions for V_f and the backup equation for $\hat{C}$. In the limit $N \to \infty$, we have the desired result. □

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:*

$$\widetilde{Q}(s,a) - \widetilde{Q}^{\pi_f}(s,a) \leq \hat{D}(s,a) \tag{26}$$

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

$$\hat{r}_D = \gamma \mathbb{E}_{s',a' \sim \pi_f} \left[V_f(s') - f(Q(s',a')) + \hat{C}(s',a') \right] \tag{27}$$

Proof. The proof of this result is similar to that of Lemma 4.2, except now we must employ the corresponding results of Lemma 4.3. Beginning with a substitution of the inductive hypothesis:

$$\widetilde{Q}^{\pi_f (N+1)}(s,a) = f\left(r(s,a)\right) + \gamma \mathop{\mathbb{E}}_{s',a' \sim \pi_f} \left(\widetilde{Q}^{\pi_f (N)}(s',a') \right) \tag{28}$$

$$\geq f\left(r(s,a)\right) + \gamma \mathop{\mathbb{E}}_{s',a' \sim \pi_f} \left(\widetilde{Q}(s',a') - \hat{D}^{(N)}(s',a') \right) \tag{29}$$

$$\geq f\left(r(s,a)\right) + \gamma \mathop{\mathbb{E}}_{s',a' \sim \pi_f} \left(f\left(Q(s',a')\right) - \hat{C}(s',a') - \hat{D}^{(N)}(s',a') \right) \tag{30}$$

$$= f\left(r(s,a)\right) + \gamma \mathop{\mathbb{E}}_{s'} \widetilde{V}(s') + \gamma \mathop{\mathbb{E}}_{s',a' \sim \pi_f} \left(f\left(Q(s',a')\right) - \hat{C}(s',a') - \hat{D}^{(N)}(s',a') - \widetilde{V}(s') \right) \tag{31}$$

$$\geq \widetilde{Q}(s,a) + \gamma \mathop{\mathbb{E}}_{s',a' \sim \pi_f} \left(f\left(Q(s',a')\right) - \hat{C}(s',a') - \hat{D}^{(N)}(s',a') - V_f(s') \right) \tag{32}$$

$$= \widetilde{Q}(s,a) - \gamma \mathop{\mathbb{E}}_{s',a' \sim \pi_f} \left(V_f(s') - f\left(Q(s',a')\right) + \hat{C}(s',a') + \hat{D}^{(N)}(s',a') \right) \tag{33}$$

$$= \widetilde{Q}(s,a) - \left(\hat{r}_D(s,a) + \gamma \mathop{\mathbb{E}}_{s',a' \sim \pi_f} \hat{D}^{(N)}(s',a') \right) \tag{34}$$

$$= \widetilde{Q}(s,a) - \hat{D}^{(N+1)}(s,a) \tag{35}$$

□

Now we provide further details on the technical conditions for compositions (rather than transformations) of primitive tasks to satisfy the derived bounds.

Lemma 4.5 (Convex Composition of Primitive Tasks)**.** *Suppose $F : \bigotimes_k X^{(k)} \to \mathbb{R}$ is convex on its domain and is sublinear (separately in each argument), that is:*

$$F(x^{(1)} + y^{(1)}, x^{(2)}, \ldots, x^{(M)}) \leq F(x^{(1)}, x^{(2)}, \ldots, x^{(M)}) + F(y^{(1)}, x^{(2)}, \ldots, x^{(M)}) \tag{36}$$

$$F(x^{(1)}, x^{(2)} + y^{(2)}, \ldots, x^{(M)}) \leq F(x^{(1)}, y^{(2)}, \ldots, x^{(M)}) + F(x^{(1)}, y^{(2)}, \ldots, x^{(M)}) \tag{37}$$

and similarly for the remaining arguments.

$$F(\gamma x^{(1)}, \ldots, \gamma x^{(M)}) \leq \gamma F(x^{(1)}, \ldots x^{(M)}) \tag{38}$$

and also satisfies

$$F\left(\max_a \mathcal{Q}^{(1)}(s,a), \ldots, \max_a \mathcal{Q}^{(M)}(s,a)\right) \leq \max_a F\left(\mathcal{Q}^{(1)}(s,a), \ldots, \mathcal{Q}^{(M)}(s,a)\right) \tag{39}$$

for all functions $\mathcal{Q}^{(k)} : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$. Then,

$$F(\vec{Q}(s,a)) \leq \widetilde{Q}(s,a) \leq F(\vec{Q}(s,a)) + C(s,a) \tag{40}$$

where we use a vector notation to emphasize that the function acts over the set of optimal value functions $\{Q^{(k)}\}$ corresponding to each primitive task, defined by $r^{(k)}$.

Proof. The proof of this statement is identical to the proof of Lemma 4.1, now using the fact that F is a multivariable function $F : X^N \to Y$, with each argument obeying the required conditions. C takes the analogous definition as provided for the original result. □

Lemma 4.6 (Concave Composition of Primitive Tasks)**.** *If on the other hand F is concave and superlinear in each argument, and also satisfies*

$$F\left(\max_a \mathcal{Q}^{(1)}(s,a), \ldots, \max_a \mathcal{Q}^{(M)}(s,a)\right) \leq \max_a F\left(\mathcal{Q}^{(1)}(s,a), \ldots, \mathcal{Q}^{(M)}(s,a)\right) \tag{41}$$

for all functions $\mathcal{Q}^{(k)} : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, then

$$F(\vec{Q}(s,a)) - \hat{C}(s,a) \leq \widetilde{Q}(s,a) \leq F(\vec{Q}(s,a)). \tag{42}$$

Proof. Again, the proof of this statement is identical to the proof of Lemma 4.3, now using the fact that F is a multivariable function $F : X^N \to Y$, with each argument obeying the required conditions. □

4.1 EXAMPLES OF TRANSFORMATIONS AND COMPOSITIONS

In this section, we consider the examples of transformations and compositions mentioned in the main text, and discuss the corresponding results in standard RL.

Remark 4.7. Given the convex composition of subtasks $r^{(c)} \equiv f(\{r^{(k)}\}) = \sum_k \alpha_k r^{(k)}$ considered by Haarnoja et al. [2018a] and Hunt et al. [2019], we can use the results of Lemma 4.6 to bound the optimal Q function by using the optimal Q functions for the primitive tasks:

$$Q^{(c)}(s,a) \leq \sum_k \alpha_k Q^{(k)}(s,a) \tag{43}$$

Proof. In standard RL, we need only show that $f(\max_i x_{1i}, \dots, \max_i x_{ni}) \geq \max_i f(x_i, \dots, x_n)$:

$$\sum_k \alpha_k \max_i x_i^{(k)} \geq \max_i \sum_k \alpha_k x_i^{(k)} \tag{44}$$

which holds given $\alpha_k \geq 0$ for all k. We also note that in this case the result clearly holds for general $\alpha_k \geq 0$ not necessarily with $\sum_k \alpha_k = 1$ (as assumed in Haarnoja et al. [2018a] and Hunt et al. [2019]). □

Remark 4.8. Given the AND composition defined above and considered in Tasse et al. [2020], we have the following result in standard RL:

$$Q^{\text{AND}}(s, a) \leq \min_k \left\{ Q^{(k)}(s, a) \right\} \tag{45}$$

Proof. We could proceed via induction as in the previous proofs, or simply use the above remark, and prove the necessary conditions on the function $f(\cdot) = \min(\cdot)$. The function $\min(\cdot)$ is concave in each argument. It is also straightforward to show that $\min(\cdot)$ is subadditive over all arguments. □

Remark 4.9. Result of (hard) OR composition result in standard RL:

$$Q^{\text{OR}}(s, a) \geq \max_k \left\{ Q^{(k)}(s, a) \right\} \tag{46}$$

Proof. The proof is analogous to the (hard) AND result: $\max$ is a convex, superadditive function.

□

Remark 4.10. Result for NOT operation in standard RL:

$$Q^{\text{NOT}}(s, a) \geq -Q(s, a) \tag{47}$$

Proof. Since the "NOT" gate is a unary function, and we are in the standard RL setting, we must check the conditions of Lemma 4.1 or 4.3. Moreoever, since the transformation function applied to the rewards, $f(r) = -r$ is linear, we must check the final condition: $\max_i\{-x_i\} = -\min_i\{x_i\} \geq -\max_i\{x_i\}$. This is the condition required by the concave conditions. □

5 PROOFS FOR ENTROPY-REGULARIZED RL

Let X be the codomain for the Q function of the primitive task ($Q : \mathcal{S} \times \mathcal{A} \to X \subseteq \mathbb{R}$).

Lemma 5.1 (Convex Conditions)**.** *Given a bounded, continuous transformation function $f : X \to \mathbb{R}$ which satisfies:*

1. *f is convex on its domain X (for stochastic dynamics);*
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 $\mathcal{Q} : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$.*

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:

$$f(Q(s, a)) \leq \widetilde{Q}(s, a) \leq f(Q(s, a)) + C(s, a) \tag{48}$$

Proof. We will again prove the result with induction, beginning by writing the backup equation for the optimal soft Q function in the transformed reward environment to prove the upper bound on $\widetilde{Q}$:

$$\widetilde{Q}^{(N+1)}(s, a) = f(r(s, a)) + \gamma \mathbb{E}_{s' \sim p(s'|s,a)} \frac{1}{\beta} \log \mathbb{E}_{a' \sim \pi_0(a'|s')} \exp \left(\beta Q^{(N)}(s', a') \right) \tag{49}$$

where p is the dynamics and π_0 is the prior policy. Applying the inductive assumption,

$$\widetilde{Q}^{(N+1)}(s,a) \geq f(r(s,a)) + \gamma \mathbb{E}_{s' \sim p(s'|s,a)} \frac{1}{\beta} \log \mathbb{E}_{a' \sim \pi_0(a'|s')} \exp\left(f\left(\beta Q^{(N)}(s',a')\right)\right) \tag{50}$$

Next, using the third condition on f as well as its convexity, we may factor f out of the expectations by Jensen's inequality:

$$\widetilde{Q}^{(N+1)}(s,a) \geq f(r(s,a)) + \gamma f\left(\mathbb{E}_{s' \sim p(s'|s,a)} \frac{1}{\beta} \log \mathbb{E}_{a' \sim \pi_0(a'|s')} \exp\left(\beta Q^{(N)}(s',a')\right)\right) \tag{51}$$

Finally, using the sublinearity conditions of f, we arrive at

$$\widetilde{Q}^{(N+1)}(s,a) \geq f\left(r(s,a) + \gamma \mathbb{E}_{s' \sim p(s'|s,a)} \frac{1}{\beta} \log \mathbb{E}_{a' \sim \pi_0(a'|s')} \exp\left(\beta Q^{(N)}(s',a')\right)\right) \tag{52}$$

The right hand side is $f\left(Q^{(N+1)}(s,a)\right)$. In the limit $N \to \infty,\ Q^{(N)}(s,a) \to Q(s,a)$ so the inductive proof for the upper bound is complete.

Let f satisfy the "convex conditions". Consider the backup equation for $\widetilde{Q}$. For the initialization (base case) we let $\widetilde{Q}^{(0)}(s,a) = f\left(Q(s,a)\right)$ and $C^{(0)}(s,a) = 0$. Using the inductive assumption,

$$\begin{aligned}
\widetilde{Q}^{(N+1)}(s,a) &= f(r(s,a)) + \frac{\gamma}{\beta} \mathop{\mathbb{E}}_{s' \sim p} \log \mathop{\mathbb{E}}_{a' \sim \pi_0} \exp \beta \widetilde{Q}^{(N)}(s',a') \\
&\leq f(r(s,a)) + \frac{\gamma}{\beta} \mathop{\mathbb{E}}_{s'} \log \mathop{\mathbb{E}}_{a'} \exp \beta \left(f(Q(s',a')) + C^{(N)}(s',a')\right) \\
&\leq f(r(s,a)) + \frac{\gamma}{\beta} \mathop{\mathbb{E}}_{s'} \left(\log \mathop{\mathbb{E}}_{a'} \exp \beta f(Q(s',a')) + \max_{a'} C^{(N)}(s',a')\right) \\
&= f(Q(s,a)) + f(r(s,a)) + \gamma \mathop{\mathbb{E}}_{s'} V_f(s') - f(Q(s,a)) + \gamma \mathop{\mathbb{E}}_{s'} \max_{a'} C^{(N)}(s',a') \\
&= f(Q(s,a)) + C^{(N+1)}(s,a)
\end{aligned}$$

Therefore in the limit $N \to \infty$, we have: $\widetilde{Q}(s,a) \leq f(Q(s,a)) + C(s,a)$ as desired. We note that since $f(r(s,a)) + \gamma\ \mathbb{E}_{s'}\ V_f(s') \geq f(Q(s,a))$, we immediately have $C(s,a) \geq 0$, as is required for the bound to be non-vacuous. □

Lemma 5.2. *Consider the soft value of the policy $\pi_f \propto \exp \beta f(Q)$ 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:*

$$\widetilde{Q}(s,a) - \widetilde{Q}^{\pi_f}(s,a) \leq D(s,a) \tag{53}$$

where D is the value of the policy π_f with reward

$$r_D(s,a) = \gamma \mathbb{E}_{s' \sim p}\left[\max_b \left\{f\left(Q(s',b)\right) + C(s',b)\right\} - V_f(s')\right] \tag{54}$$

Proof. To prove the (soft) policy evaluation bound, we use iterations of soft-policy evaluation Haarnoja et al. [2018b] and denote iteration N of the evaluation of π_f in the composite environment as $\widetilde{Q}^{\pi_f (N)}$. Beginning with the definitions $\widetilde{Q}^{\pi_f (0)}(s,a) = Q(s,a)$ (since the evaluation is independent of the initialization), and $D^{(0)} = 0$, the $N = 0$ step is trivially satisfied. Assuming the inductive hypothesis, we consider the next step of soft policy evaluation: As in the previous policy evaluation results, we prove an equivalent result with induction.

$$\begin{aligned}
\widetilde{Q}^{\pi_f(N+1)}(s,a) &= f(r(s,a)) + \gamma \mathop{\mathbb{E}}_{s'\sim p} \mathop{\mathbb{E}}_{a'\sim\pi_f} \left[\widetilde{Q}^{\pi_f(N)}(s',a') - \frac{1}{\beta}\log\frac{\pi_f(a'|s')}{\pi_0(a'|s')}\right] \\
&\geq f(r(s,a)) + \gamma \mathop{\mathbb{E}}_{s',a'}\left[\widetilde{Q}(s',a') - D^{(N)}(s',a') - f(Q(s',a')) + V_f(s')\right] \\
&= f(r(s,a)) + \gamma\mathop{\mathbb{E}}_{s'}\widetilde{V}(s') + \gamma\mathop{\mathbb{E}}_{s',a'}\left[\widetilde{Q}(s',a') - D^{(N)}(s',a') - f(Q(s',a')) + V_f(s') - \widetilde{V}(s')\right] \\
&\geq \widetilde{Q}(s,a) + \gamma\mathop{\mathbb{E}}_{s',a'}\left[f(Q(s',a')) - D^{(N)}(s',a') - f(Q(s',a')) + V_f(s') - \widetilde{V}(s')\right] \\
&\geq \widetilde{Q}(s,a) + \gamma \mathop{\mathbb{E}}_{s',a'}\left[-D^{(N)}(s',a') + V_f(s') - \max_b\left\{f\left(Q(s',b)\right) + C(s',b)\right\}\right] \\
&\geq \widetilde{Q}(s,a) - D^{(N+1)}(s,a)
\end{aligned}$$

where we have used $\widetilde{Q}(s,a) \geq f(Q(s,a))$ in the fourth line.

where we have used the fact that $\widetilde{V}(s) \leq \max_b \left\{f\left(Q(s,b)\right) + \max_a C(s,a)\right\}$ and $\widetilde{Q}(s,a) - f(Q(s,a)) \geq 0$ which both follow from the previously stated bounds. □

Lemma 5.3 (Concave Conditions)**.** *Given a bounded, continuous transformation function $f : X \to \mathbb{R}$ which satisfies:*

1. *f is concave on its domain X (for stochastic dynamics);*
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. *$f\left(\log \mathbb{E}\exp \mathcal{Q}(s,a)\right) \geq \log\mathbb{E}\exp f\left(\mathcal{Q}(s,a)\right)$ for all functions $\mathcal{Q} : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$.*

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

$$f\left(Q(s,a)\right) - \hat{C}(s,a) \leq \widetilde{Q}(s,a) \leq f\left(Q(s,a)\right) \tag{55}$$

Proof. The proof of the upper bound is the same as the preceding theorem's lower bound with all inequalities reversed. For the lower bound involving C,

Again consider the backup equation for $\widetilde{Q}$. Using the definitions and inductive assumption as before, we have

$$\begin{aligned}
\widetilde{Q}^{(N+1)}(s,a) &= f(r(s,a)) + \frac{\gamma}{\beta}\mathop{\mathbb{E}}_{s'\sim p}\log\mathop{\mathbb{E}}_{a'\sim\pi_0}\exp\beta\widetilde{Q}^{(N)}(s',a') \\
&\geq f(r(s,a)) + \frac{\gamma}{\beta}\mathop{\mathbb{E}}_{s'}\log\mathop{\mathbb{E}}_{a'}\exp\beta\left(f(Q(s',a')) - \hat{C}^{(N)}(s',a')\right) \\
&\geq f(r(s,a)) + \frac{\gamma}{\beta}\mathop{\mathbb{E}}_{s'}\left(\log\mathop{\mathbb{E}}_{a'}\exp\beta f(Q(s',a')) - \max_{a'}\hat{C}^{(N)}(s',a')\right) \\
&= f(Q(s,a)) - \left[f(Q(s,a)) - f(r(s,a)) - \gamma\mathop{\mathbb{E}}_{s'}V_f(s') + \gamma\mathop{\mathbb{E}}_{s'}\max_{a'}\hat{C}^{(N)}(s',a')\right] \\
&= f(Q(s,a)) - \hat{C}^{(N+1)}(s,a)
\end{aligned}$$

Therefore in the limit $N \to \infty$, we have: $\widetilde{Q}(s,a) \geq f(Q(s,a)) - \hat{C}(s,a)$ as desired. □

Lemma 5.4. *Consider the soft value of the policy $\pi_f \propto \exp\beta f(Q)$ 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:*

$$\widetilde{Q}(s,a) - \widetilde{Q}^{\pi_f}(s,a) \leq \hat{D}(s,a) \tag{56}$$

where $\hat{D}$ is the fixed point of

$$\hat{D}(s,a) \leftarrow \gamma\mathbb{E}_{s'\sim p}\mathbb{E}_{a'\sim\pi_f}\left[\hat{C}(s',a') + \hat{D}(s',a')\right] \tag{57}$$

Proof. We will show the policy evaluation result by induction, by evaluating $\pi_f \propto \exp(\beta f(Q))$ in the environment with rewards $f(r)$. We shall denote iterations of policy evaluation for π_f in the environment with rewards $f(r)$ by $\widetilde{Q}^{\pi_f(N)}(s,a)$.

$$\begin{aligned}
\widetilde{Q}^{\pi_f(N+1)}(s,a) &= f(r(s,a)) + \gamma \mathop{\mathbb{E}}_{s'\sim p} \mathop{\mathbb{E}}_{a'\sim\pi_f} \left[\widetilde{Q}^{\pi_f(N)}(s',a') - \frac{1}{\beta}\log\frac{\pi_f(a'|s')}{\pi_0(a'|s')}\right] \\
&\geq f(r(s,a)) + \gamma \mathop{\mathbb{E}}_{s',a'} \left[\widetilde{Q}(s',a') - \hat{D}^{(N)}(s',a') - (f(Q(s',a')) - V_f(s'))\right] \\
&\geq f(r(s,a)) + \gamma \mathop{\mathbb{E}}_{s',a'} \left[\widetilde{Q}(s',a') - \hat{D}^{(N)}(s',a') - \widetilde{Q}(s',a') - \hat{C}(s',a') + V_f(s')\right] \\
&\geq f(r(s,a)) + \gamma \mathop{\mathbb{E}}_{s'} \widetilde{V}(s') - \gamma \mathop{\mathbb{E}}_{s',a'} \left[\hat{D}^{(N)}(s',a') + \hat{C}(s',a')\right] \\
&= \widetilde{Q}(s,a)) - \hat{D}^{(N+1)}(s,a)
\end{aligned}$$

where we have used the inductive assumption and $V_f(s) \geq \widetilde{V}(s)$ and which follows from the previously stated bounds. Therefore in the limit $N \to \infty$, we have: $\widetilde{Q}^{\pi_f}(s,a) \geq \widetilde{Q}(s,a) - \hat{D}(s,a)$ as desired. □

Lemma 5.5 (Convex Composition of Primitive Tasks)**.** *Suppose $F : X^N \to Y$ is convex on its domain X^N and satisfies all conditions of Lemma 5.1 (Main Text) component-wise. Then,*

$$F(\vec{Q}(s,a)) \leq \widetilde{Q}(s,a) \leq F(\vec{Q}(s,a)) + C(s,a) \tag{58}$$

and

$$\widetilde{Q}^{\pi_f}(s,a) \geq \widetilde{Q}(s,a) - D(s,a) \tag{59}$$

where we use a vector notation to emphasize that the function acts over the set of optimal $\{Q_k\}$ functions corresponding to each subtask, defined by r_k.

Proof. The proof of this statement is identical to the previous proofs, now using the fact that F is a multivariable function $F : X^N \to Y$, with each argument obeying the required conditions. □

Lemma 5.6 (Concave Composition of Primitive Tasks)**.** *If on the other hand F is concave and and satisfies all conditions of Lemma 5.2 (Main Text) component-wise, then*

$$F(\vec{Q}(s,a)) - \hat{C}(s,a) \leq \widetilde{Q}(s,a) \leq F(\vec{Q}(s,a)). \tag{60}$$

and

$$\widetilde{Q}^{\pi_f}(s,a) \geq \widetilde{Q}(s,a) - \hat{D}(s,a) \tag{61}$$

Proof. Again, the proof of this statement is identical to the previous proofs, now using the fact that F is a multivariable function $F : X^N \to Y$, with each argument obeying the required conditions. □

5.1 EXAMPLES OF TRANSFORMATIONS AND COMPOSITIONS

In this section we consider several examples mentioned in the main text, and show how they are proved with our results in entropy-regularized RL.

Remark 5.7. Given the convex composition of subtasks $r^{(c)} \equiv F(\{r^{(k)}\}) = \sum_k \alpha_k r^{(k)}$ considered by Haarnoja et al. [2018a] and Hunt et al. [2019], we can use the results of Lemma 5.6 to bound the optimal Q function by using the optimal Q functions for the primitive tasks:

$$Q^{(c)}(s,a) \leq \sum_k \alpha_k Q^{(k)}(s,a) \tag{62}$$

Proof. In entropy-regularized RL we need to show that the final condition holds (in vectorized form). This is simply Hölder's inequality Hardy et al. [1952] for vector-valued functions in a probability space (with measure defined by π_0).

□

Remark 5.8. Given the AND composition defined above and considered in Tasse et al. [2020], we have the following result in standard RL:

$$Q^{\text{AND}}(s,a) \leq \min_k \left\{ Q^{(k)}(s,a) \right\} \tag{63}$$

Proof. The function $\min(\cdot)$ is concave in each argument. It is also straightforward to show that $\min(\cdot)$ is subadditive over all arguments. For the final condition, the $\min$ acts globally over all subtasks:

$$\min_k \left\{ \frac{1}{\beta} \log \mathbb{E}_{a\sim\pi_0(a|s)} \exp\left(\beta \mathcal{Q}^{(k)}(s,a)\right) \right\} \leq \frac{1}{\beta} \log \mathbb{E}_{a\sim\pi_0(a|s)} \exp\left(\beta \min_k \left\{ \mathcal{Q}^{(k)}(s,a) \right\}\right). \tag{64}$$

□

Remark 5.9. Result of (hard) OR composition result in standard RL:

$$Q^{\text{OR}}(s,a) \geq \max_k \left\{ Q^{(k)}(s,a) \right\} \tag{65}$$

Proof. The proof is analogous to the (hard) AND result: $\max$ is a convex, superadditive function. For the final condition, the $\max$ again acts globally over all subtasks:

$$\max_k \left\{ \frac{1}{\beta} \log \mathbb{E}_{a\sim\pi_0(a|s)} \exp\left(\beta \mathcal{Q}^{(k)}(s,a)\right) \right\} \geq \frac{1}{\beta} \log \mathbb{E}_{a\sim\pi_0(a|s)} \exp\left(\beta \max_k \left\{ \mathcal{Q}^{(k)}(s,a) \right\}\right). \tag{66}$$

□

Remark 5.10. Again we consider the NOT operation defined above, now in entropy-regularized RL, which yields the bound:

$$Q^{\text{NOT}}(s,a) \geq -Q(s,a) \tag{67}$$

Proof. As in the standard RL case, we need only consider the third condition of either Lemma 5.1 or 5.3. In particular, we show

$$f\left(\log \mathbb{E} \exp \mathcal{Q}(s,a)\right) \leq \log \mathbb{E} \exp f\left(\mathcal{Q}(s,a)\right) \tag{68}$$

for all functions $\mathcal{Q} : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$. This follows from

$$\frac{1}{\mathbb{E} \exp \mathcal{Q}(s,a)} \leq \mathbb{E} \frac{1}{\exp \mathcal{Q}(s,a)} \tag{69}$$

which is given by Jensen's inequality, since the function $f(x) = 1/x$ is convex.

□

Remark 5.11 (Linear Scaling). Given some $k \in (0,1)$ the function $f(x) = kx$ satisfies the results of the first theorem. Conversely, if $k \geq 1$, $f(x) = kx$ satisfies the results of the second theorem.

Proof. This result (specifically the third condition of Lemma 5.1, 5.3) follows from the monotonicity of ℓ_p norms. □

Since we have already shown the case of $k = -1$ (NOT gate), with the result of Theorem 7.2, the case for all $k \in \mathbb{R}$ has been characterized.

6 EXTENSION FOR ERROR-PRONE Q-VALUES

In this section, we provide some discussion on the case of inexact Q-values, as often occurs in practice (discussed at the end of Section 4.1 in the main text). We focus on the case of task transformation in standard RL. The corresponding statements in the settings of composition and entropy-regularized RL follow similarly.

As our starting point, we assume that an "ε-optimal estimate" $\overline{Q}(s,a)$ for a primitive task's exact value function $Q(s,a)$ is known.

Definition 6.1. An ε-optimal Q-function, $\overline{Q}$, satisfies

$$|Q(s,a) - \overline{Q}(s,a)| \leq \varepsilon \tag{70}$$

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

To allow the derived double-sided bounds on the transformed tasks' Q-values to carry over to this more general setting, we assume that the transformation function is L-Lipschitz continuous. With these assumptions, we prove the following extensions of Lemma 4.1 and 4.3:

Lemma 4.1A (Convex Conditions, Error-Prone). *Given a primitive task with discount factor γ, corresponding ε-optimal value function $\overline{Q}$, and a bounded, continuous, L-Lipschitz transformation function $f : X \to \mathbb{R}$ which satisfies:*

1. *f is convex on its domain X (for stochastic dynamics);*
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. *$f(\max_a \mathcal{Q}(s,a)) \leq \max_a f(\mathcal{Q}(s,a))$ for all $\mathcal{Q} : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$.*

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:

$$f(\overline{Q}(s,a)) - L\varepsilon \leq \widetilde{Q}(s,a) \leq f(\overline{Q}(s,a)) + \overline{C}(s,a) + \frac{2}{1-\gamma} L\varepsilon \tag{71}$$

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

$$\overline{r_C}(s,a) = f(r(s,a)) + \gamma \mathbb{E}_{s'} \overline{V_f}(s') - f(\overline{Q}(s,a)). \tag{72}$$

with $\overline{V_f}(s) = \max_a f(\overline{Q}(s,a))$.

Note that as $\varepsilon \to 0$, the exact result (Lemma 4.1) is recovered. If the function $\overline{C}$ is not known exactly, one can similarly exchange $\overline{C}$ for $\overline{\overline{C}}$, an ε-optimal estimate for $\overline{C}$. This consideration loosens the upper-bound by an addition of ε, shown at the end of the proof.

We will make use a well-known result (cf. proof of Lemma 1 in Barreto et al. [2017]) that bounds the difference in optimal Q-values for two tasks with different reward functions.

Lemma 6.2. *Let two tasks, only differing in their reward functions, be given with reward $r_1(s,a)$ and $r_2(s,a)$, respectively. Suppose $|r_1(s,a) - r_2(s,a)| \leq \delta$ Then, the optimal value functions for the tasks satisfies:*

$$|Q_1(s,a) - Q_2(s,a)| \leq \frac{\delta}{1-\gamma} \tag{73}$$

Now we are in a position to prove Lemma 4.1A:

Proof. To prove the lower bound, we begin with the original lower bound in Lemma 4.1, for the optimal primitive task Q-values:

$$\widetilde{Q}(s,a) \geq f(Q(s,a)), \tag{74}$$

or equivalently

$$-\widetilde{Q}(s,a) \leq -f(Q(s,a)) \tag{75}$$

$$-\widetilde{Q}(s,a) \leq -f(Q(s,a)) + f(\overline{Q}(s,a)) - f(\overline{Q}(s,a)) \tag{76}$$

$$-\widetilde{Q}(s,a) \leq |f(Q(s,a)) - f(\overline{Q}(s,a))| - f(\overline{Q}(s,a)) \tag{77}$$

$$\widetilde{Q}(s,a) \geq -|f(Q(s,a)) - f(\overline{Q}(s,a))| + f(\overline{Q}(s,a)) \tag{78}$$

$$\widetilde{Q}(s,a) \geq -L|Q(s,a) - \overline{Q}(s,a)| + f(\overline{Q}(s,a)) \tag{79}$$

$$\widetilde{Q}(s,a) \geq f(\overline{Q}(s,a)) - L\varepsilon \tag{80}$$

$$\tag{81}$$

Where the final steps follow from the function f being L-Lipschitz and the definition of ε-optimality of $\overline{Q}(s,a)$.

To prove the upper bound, we take a similar approach, noting that the reward function r_C in Lemma 4.1 must be updated to account for the inexact Q-values. Therefore, we must account for the following error propagations:

$$\begin{gathered} Q(s,a) \to \overline{Q}(s,a) \\ V_f(s) \to \overline{V_f}(s) \\ r_C(s,a) \to \overline{r_C}(s,a). \end{gathered}$$

We first find the difference between r_C and $\overline{r_C}$ to be bounded by $(1+\gamma)L\varepsilon$:

$$|r_C(s,a) - \overline{r_C}(s,a)| = |\gamma \mathop{\mathbb{E}}_{s'\sim p} V_f^*(s') - f(Q^*(s,a)) - \gamma \mathop{\mathbb{E}}_{s'\sim p} V_f(s') + f(Q(s,a))| \tag{82}$$

$$\leq \gamma \mathop{\mathbb{E}}_{s'} |V_f^*(s') - V_f(s')| + |f(Q^*(s,a)) - f(Q(s,a))| \tag{83}$$

$$\leq \gamma \mathop{\mathbb{E}}_{s'} \max_{a'} |f(Q^*(s',a')) - f(Q(s',a'))| + |f(Q^*(s,a)) - f(Q(s,a))| \tag{84}$$

$$\leq (1+\gamma)L\varepsilon \tag{85}$$

where in the third line we have used the bound $|\max_x f(x) - \max_x g(x)| \leq \max_x |f(x) - g(x)|$.

Now, applying Lemma 6.2 to the reward functions r_C and $\overline{r_C}$:

$$|C(s,a) - \overline{C}(s,a)| \leq \frac{(1+\gamma)}{1-\gamma} L\varepsilon \tag{86}$$

With the same technique as was used above for the lower bound, we find:

$$\widetilde{Q}(s,a) \leq f(Q(s,a)) + C(s,a) \tag{87}$$

$$\leq f(\overline{Q}(s,a)) + L\varepsilon + C(s,a) \tag{88}$$

$$= f(\overline{Q}(s,a)) + L\varepsilon + \overline{C}(s,a) - \overline{C}(s,a) + C(s,a) \tag{89}$$

$$\leq f(\overline{Q}(s,a)) + L\varepsilon + |C(s,a) - \overline{C}(s,a)| + \overline{C}(s,a) \tag{90}$$

$$\leq f(\overline{Q}(s,a)) + L\varepsilon + \overline{C}(s,a) + \frac{(1+\gamma)}{1-\gamma} L\varepsilon \tag{91}$$

$$= f(\overline{Q}(s,a)) + \overline{C}(s,a) + \frac{2}{1-\gamma} L\varepsilon \tag{92}$$

Further extending the result to the case where only an ε-optimal estimate of $\overline{C}$ is known, denoted by $\overline{\overline{C}}$, we find:

$$\widetilde{Q}(s,a) \leq f(\overline{Q}(s,a)) + \overline{C}(s,a) + \frac{2}{1-\gamma} L\varepsilon \tag{93}$$

$$\leq f(\overline{Q}(s,a)) + \overline{\overline{C}}(s,a) + |\overline{\overline{C}}(s,a) - \overline{C}(s,a)| + \frac{2}{1-\gamma} L\varepsilon \tag{94}$$

$$\leq f(\overline{Q}(s,a)) + \overline{\overline{C}}(s,a) + \varepsilon + \frac{2}{1-\gamma} L\varepsilon \tag{95}$$

$$= f(\overline{Q}(s,a)) + \overline{\overline{C}}(s,a) + \left(1 + \frac{2}{1-\gamma} L\right)\varepsilon \tag{96}$$

□

Similarly, Lemma 4.3 from the main text can be extended under the same conditions:

Lemma 4.3A (Concave Conditions, Error-Prone)**.** *Given a primitive task with discount factor γ, corresponding ε-optimal value function $\overline{Q}$, and a bounded, continuous, L-Lipschitz transformation function $f : X \to \mathbb{R}$ which satisfies:*

1. *f is concave on its domain X (for stochastic dynamics);*
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:

$$f(\overline{Q}(s,a)) - \overline{\hat{C}}(s,a) - \frac{2}{1-\gamma} L\varepsilon \leq \widetilde{Q}(s,a) \leq f(\overline{Q}(s,a)) + L\varepsilon \tag{97}$$

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

$$\overline{\hat{r}_C}(s,a) = f(\overline{Q}(s,a)) - f(r(s,a)) - \gamma \underset{s' \sim p}{\mathbb{E}} \overline{V_f}(s') \tag{98}$$

with $\overline{V_f}(s) = \max_a f(\overline{Q}(s,a))$.

The proof of Lemma 4.3A is the same as that given above for Lemma 4.1A, with all signs flipped.

Finally, we note that both extensions of Lemma 4.1A and 4.3A hold for the entropy-regularized case. The only differences required to prove the results are showing that Lemma 6.2 and $|V_f(s) - \overline{V_f}(s)| \leq L\varepsilon$ hold in entropy-regularized RL. Both statements are trivial given that the necessary soft-max operation is 1-Lipschitz. Similar results can be derived for the case of compositions, when each subtasks' Q-function is replaced by an ε-optimal estimate thereof.

7 RESULTS APPLYING TO BOTH ENTROPY-REGULARIZED AND STANDARD RL

As we have discussed in the main text; an agent with a large library of accessible functions will be able to transform and compose their primitive knowledge in a wider variety of ways. Therefore, we would like to extend $\mathcal{F}$ to encompass as many functions as possible. Below, we will show that the functions $f \in \mathcal{F}$ characterizing the Transfer MDP Library have two closure properties (additivity and function composition) which enables more accessible transfer functions.

First, let $\mathcal{F}^+$ denote the set of functions $f \in \mathcal{F}$ obeying the convex conditions, and similarly let $\mathcal{F}^-$ denote the set of functions obeying the concave conditions.

In standard RL, we have the following closure property for addition of functions.

Theorem 7.1. *Let $f, g \in \mathcal{F}^+$. Then $f + g \in \mathcal{F}^+$. Similarly, if $f, g \in \mathcal{F}^-$, then $f + g \in \mathcal{F}^-$.*

Proof. Let $f, g \in \mathcal{F}^+$.

Convexity: The sum of two convex functions is convex.

Subadditive: $(f+g)(x+y) = f(x+y) + g(x+y) \leq f(x) + g(x) + f(y) + g(y) = (f+g)(x) + (f+g)(y)$.

Submultiplicative: $(f+g)(\gamma x) = f(\gamma x) + g(\gamma x) \leq \gamma f(x) + \gamma g(x) = \gamma(f+g)(x)$.

The proof for $f, g \in \mathcal{F}^-$ is the same with all signs flipped, except for the additional final condition: $(f+g)(\max_i x_i) = f(\max_i x_i) + g(\max_i x_i) = \max f(x) + \max g(x) \geq \max f(x) + g(x)$. Although this is not equality as shown in the main text, the condition still suffices. For the case of a single function (no addition, as seen in main text), it can never be the cases that $\max_i f(x_i) > \max f(x)$ and therefore was excluded. (Just as $\max_i f(x_i) \leq \max f(x)$ is automatically satisfied for the convex conditions.) □

Theorem 7.2 (Function Composition)**.** *For any reward-mapping functions $f,\ g \in \mathcal{F}^{+}$ ($\mathcal{F}^{-}$) with f non-decreasing, the composition of functions f and g, $h(x) = f(g(x)) \in \mathcal{F}^{+}(\mathcal{F}^{-})$.*

Proof. Let $f, g \in \mathcal{F}^{+}$ assume $f : B \to C$ and $g : A \to B$, and let f be non-decreasing. This guarantees that $f(g(x))$ is convex. Additionally, $f(g(x+y)) \leq f(g(x)+g(y)) \leq f(g(x))+f(g(y))$ by the sublinearity of g, f respectively. Similarly $f(g(\gamma x)) \leq f(\gamma g(x)) \leq \gamma f(g(x))$.

For the standard RL (concave) condition, note that for all functions $\mathcal{Q} :\ \mathcal{S}\ \times\ \mathcal{A} \to X$:

$$f\left(g\left(\max_a \mathcal{Q}(s,a)\right)\right) \geq f\left(\max_a\ g\left(\mathcal{Q}(s,a)\right)\right) \geq \max_a\ f\left(g\left(\mathcal{Q}(s,a)\right)\right) \tag{99}$$

For the entropy-regularized condition, we first apply the condition to g:

$$f\left(g\left(\frac{1}{\beta}\log \mathbb{E}_{a\sim\pi_0(a|s)} \exp(\beta\mathcal{Q}(s,a))\right)\right) \leq f\left(\frac{1}{\beta}\log \mathbb{E}_{a\sim\pi_0(a|s)} \exp(\beta g(\mathcal{Q}(s,a)))\right) \tag{100}$$

Then to f:

$$f\left(g\left(\frac{1}{\beta}\log \mathbb{E}_{a\sim\pi_0(a|s)} \exp(\beta\mathcal{Q}(s,a))\right)\right) \leq \frac{1}{\beta}\log \mathbb{E}_{a\sim\pi_0(a|s)} \exp\left(\beta f\left(g\left(\mathcal{Q}(s,a)\right)\right)\right) \tag{101}$$

The reversed statement, when $f, g \in \mathcal{F}^{-}$ with f non-decreasing has a similar proof and is omitted. □

With this result established, we are now able to concatenate multiple transformations. This allows for multiple gates in Boolean logic statements, for example. As stated in the main text, this ability to compose multiple functions will greatly expand the number of tasks in the Transfer MDP Library which the agent may (approximately) solve.

References

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.

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

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.

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

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.

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.

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.