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

Jacob Adamczyk,[1] Argenis Arriojas,[1] Stas Tiomkin,[2] Rahul V. Kulkarni[1]

[1]Dept. of Physics, University of Massachusetts Boston, [2]Dept. of Computer Engineering, San Jose State University

Abstract

In RL, the ability to utilize prior knowledge from previously solved tasks can allow agents to quickly solve new problems. In some cases, these new problems may be approximately solved by composing the solutions of previously solved primitive tasks. Otherwise, prior knowledge can be used to shape the reward function in a way that leaves the optimal policy unchanged but enables quicker learning. In this work, we develop a general framework for reward shaping and task composition in **entropy-regularized** RL.

Background

Regularized RL induces stochastic optimal policies which are robust to perturbations[1] and allows for composition of basic behaviors[2]

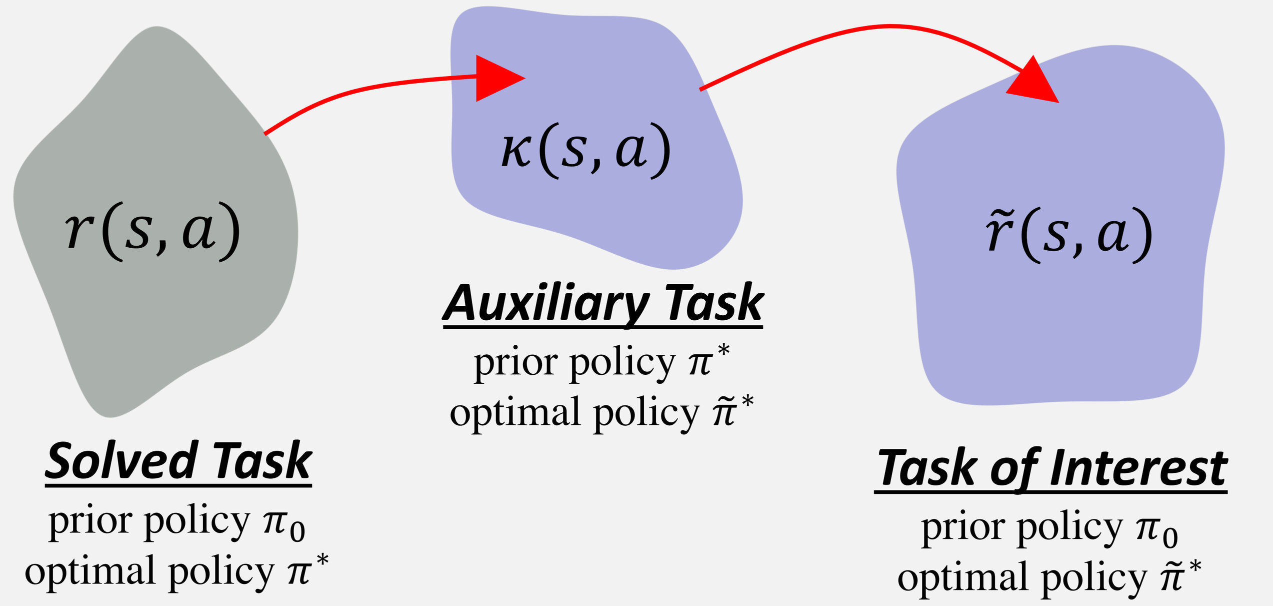
$$J(\pi) = \mathbb{E}_{\tau \sim p,\pi} \sum_{t=1}^{\infty} \gamma^t r(s_t, a_t)$$

Entropy regularization alters the objective function

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

How can prior knowledge assist the agent in solving new tasks?

Proposed Solution: Auxiliary Task



By solving a task with reward function $\kappa = \tilde{r} - r$ and a prior policy π^*, we can use prior knowledge to access the solution to the desired task.

Reward Shaping

For the solved task, set $\pi^* = \pi_0$ and $V^*(s) = \Phi(s)$. Then the corresponding reward function[4] is:

$$r = \Phi(s) - \gamma \mathbb{E} \Phi(s')$$

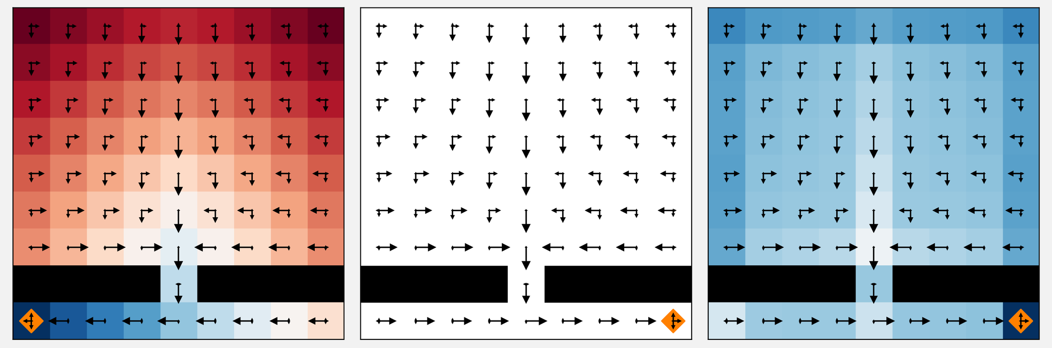
The auxiliary task's reward function is thus:

$$\kappa = \tilde{r} + \gamma \mathbb{E} \Phi(s') - \Phi(s)$$

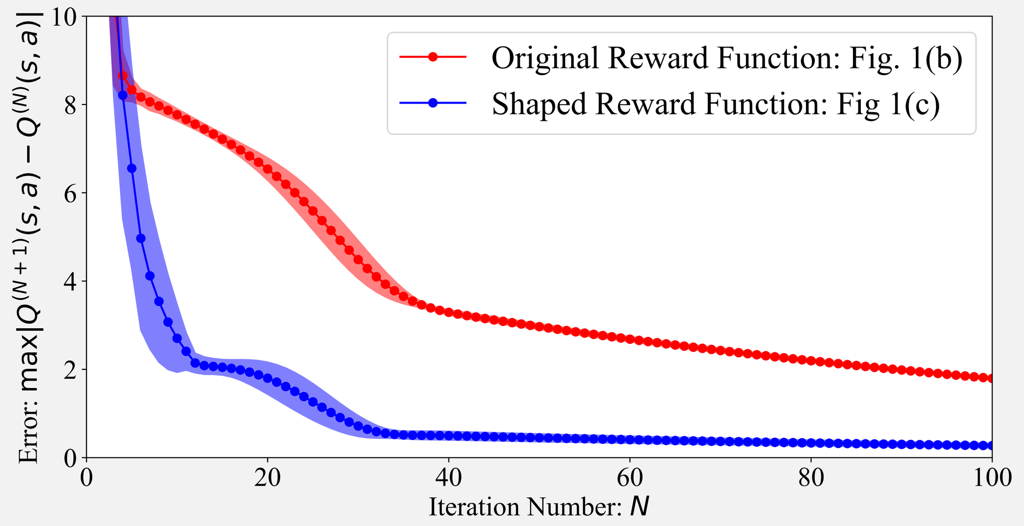
Since $\pi_K^* = \tilde{\pi}^*$, this result implies potential-based reward shaping[3] holds in entropy-regularized RL.

"Solved" Task	*Auxiliary Task*	*Task of Interest*
$\Phi(s) - \gamma \mathbb{E}\Phi(s')$	$\kappa(s,a)$	$\tilde{r}(s,a)$

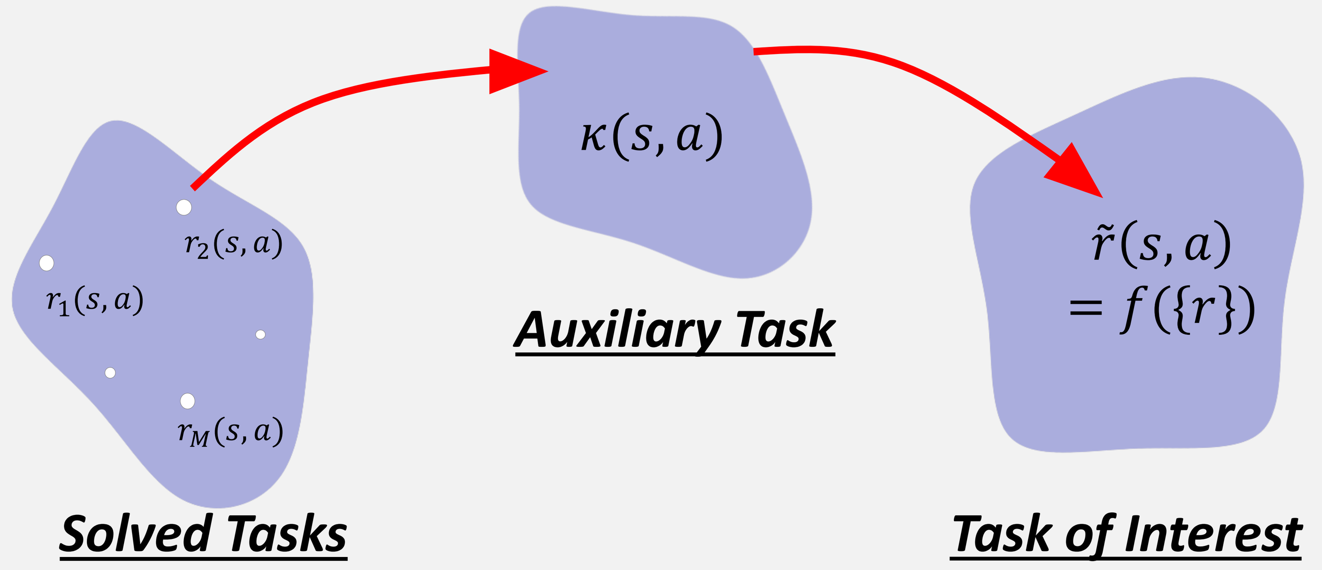
$$Q^*(s,a) + K^*(s,a) = \tilde{Q}^*(s,a)$$



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



Compositionality

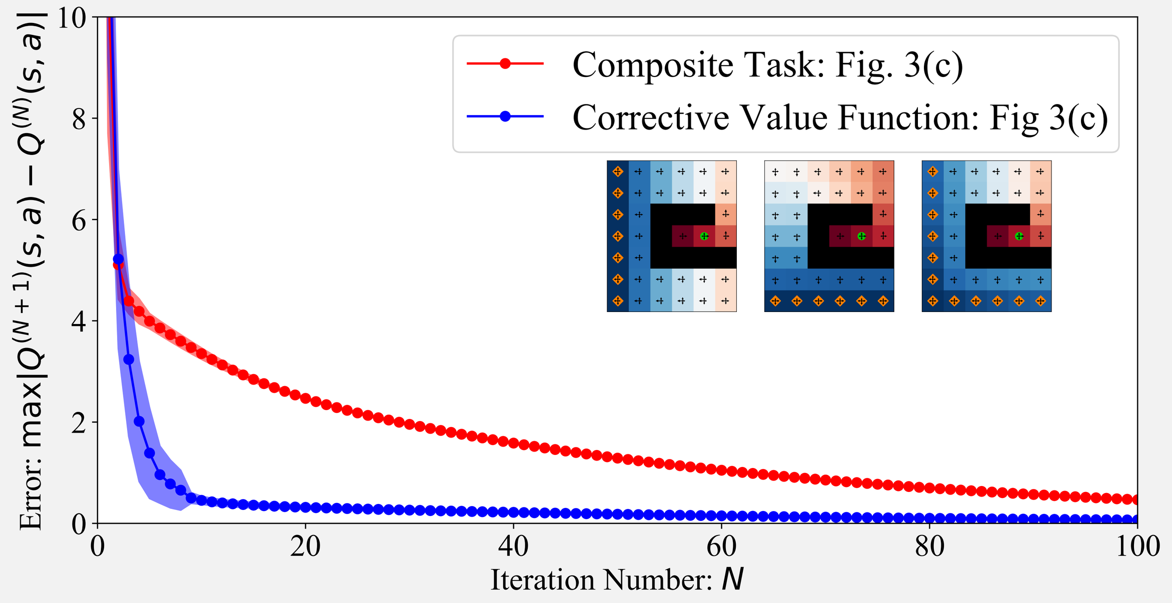


We carry out a similar analysis with *multiple* primitive tasks and a "composite" task of interest, generalizing the work of [5] and yielding:

$$f\left(\{Q_j^*(s,a)\}\right) + K^*(s,a) = \tilde{Q}^*(s,a)$$

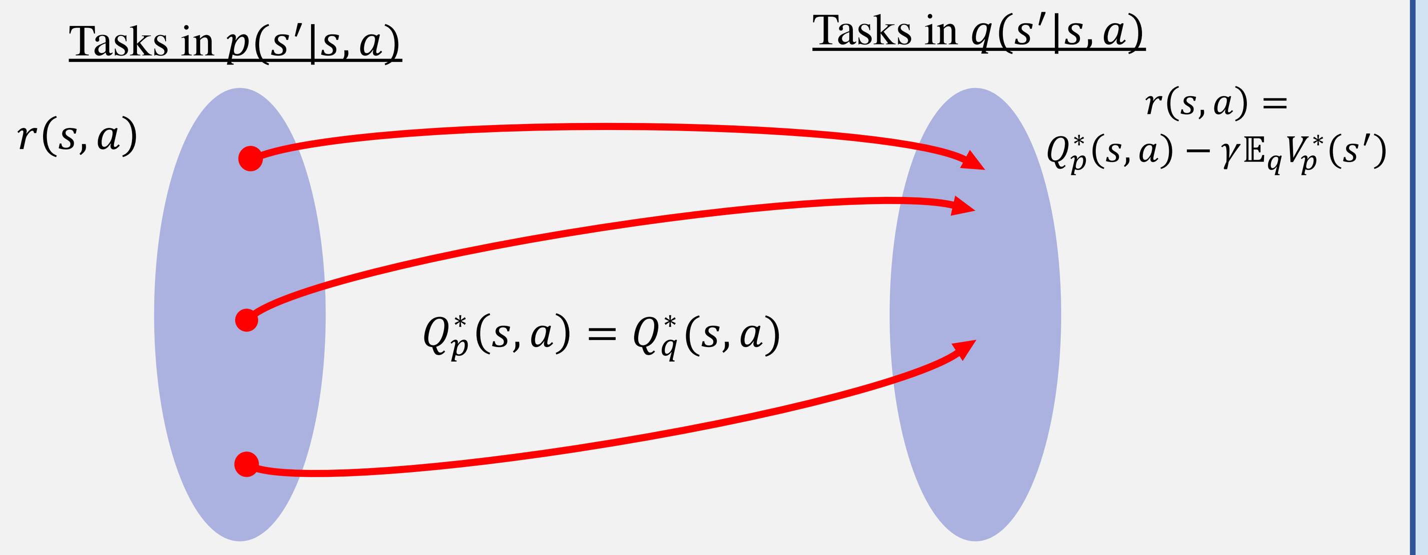
$$\kappa = f\left(\{r_j(s,a)\}\right) + \gamma \mathbb{E} V_f(s') - f\left(\{Q_j^*(s,a)\}\right)$$

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



Change in Dynamics

A similar auxiliary task can be defined for changes in dynamics. Combined with the result for a difference in rewards, we find:



Solving tasks in one setting (p) provides solutions under a different transition dynamics (q).

Future Work

In the future we would like to study the cases of continuous states and actions with function approximators, standard RL ($\beta \to \infty$), and compositionality of tasks with variable dynamics.

References

[1]: "Maximum Entropy RL (Provably) Solves Some Robust RL Problems", B. Eysenbach, S. Levine (2022); arXiv: 2103.06257

[2]: "Composable Deep Reinforcement Learning for Robotic Manipulation", T. Haarnoja, et. al. (2018); arXiv: 1803.06773

[3]: "Policy invariance under reward transformations: Theory and application to reward shaping", Ng, Harada, Russell, ICML 1999

[4]: Identifiability in inverse reinforcement learning", H. Cao, S. N. Cohen, Ł. Szpruch; arXiv:2106.03498

[5]: "Composing Entropic Policies using Divergence Correction", J.J. Hunt, et. al. (2019); arXiv:1812.02216

Acknowledgements

This work was supported by the NSF through Award DMS-1854350, the Proposal Development Grant provided by UMass Boston, the Research Foundation at SJSU, and the Alliance Innovation Lab in Silicon Valley.