Bounding the Optimal Value Function in Compositional RL

Jacob Adamczyk[1], Volodymyr Makarenko[2], Argenis Arriojas[1], Stas Tiomkin[2], Rahul V Kulkarni[1]
[1]Physics Department, University of Massachusetts Boston, [2]Department of Computer Engineering, San José State University

Abstract

RL agents often solve a variety of tasks differing only in reward function. One popular approach for obtaining new solutions in this setting involves functional composition of previously solved Q-values. Our work unifies previous examples, providing a general framework for composition in both standard and entropy-regularized RL. For many functions, we show the composite task's solution is related to the known task solutions via double-sided bounds on the optimal Q-value. We find the suboptimality of using the zero-shot greedy policy is bounded for this class of functions. We present clipping approaches for reducing uncertainty during training, thereby allowing agents to quickly adapt to new tasks.

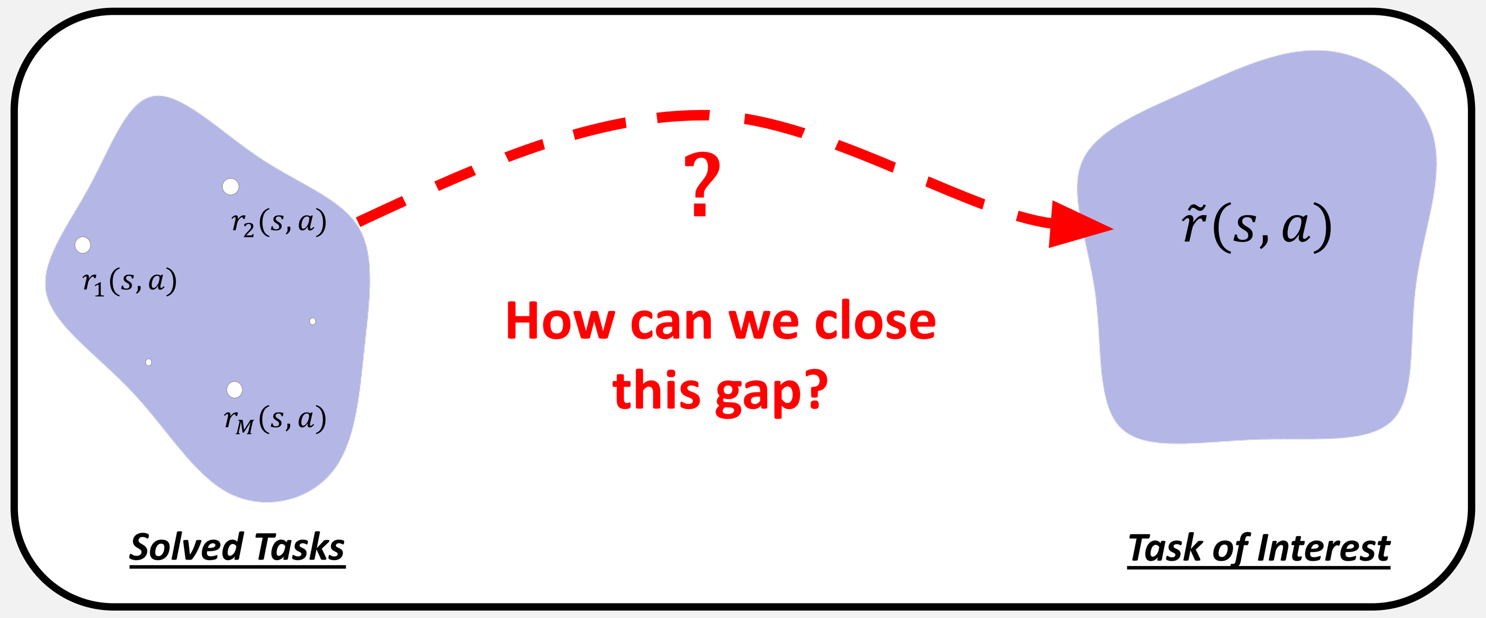
Motivation

Reinforcement Learning (RL) objective[1]:

$$Q^{\pi}(s,a) = \mathbb{E}_{\tau\sim p,\pi}\left[\sum_{t=0}^{\infty}\gamma^t r(s_t,a_t)\right]$$

Compositional RL

We take primitive tasks as given and wish to transfer to a target task by functionally combining previous solutions:



Setting

$$\tilde{r}(s,a) \doteq f(\{r_1(s,a), r_2(s,a), \dots, r_M(s,a)\})$$

Ansatz

$$\tilde{Q}(s,a) \approx f(\{Q_1(s,a), Q_2(s,a), \dots, Q_M(s,a)\})$$

Prior Work

This setup has been previously considered with specific functions[3-5] and assumptions on MDP structure.

Proposed Solution

Based on this ansatz, we present a set of functions whose corresponding MDPs enjoy double-sided Q-value bounds and bounded suboptimality.

Also holds for entropy-regularized RL[2,3] and error-prone Q-values!

Convex conditions

Given a convex function additionally satisfying:

$$f(x+y) \le f(x) + f(y)$$
$$f(\gamma x) \le \gamma f(x)$$

The composite task's Q-values are upper and lower bounded:

$$f(Q(s,a)) \le \tilde{Q}(s,a) \le f(Q(s,a)) + C(s,a)$$

Where C satisfies a Bellman backup equation:

$$C(s,a) = r_C(s,a) + \gamma \mathbb{E} \max_{a'} C(s',a')$$
$$r_C(s,a) = f(r(s,a)) + \gamma \mathbb{E} V_f(s') - f(Q(s,a))$$

Similar results for "concave conditions" and multiple primitive Q functions are provided.

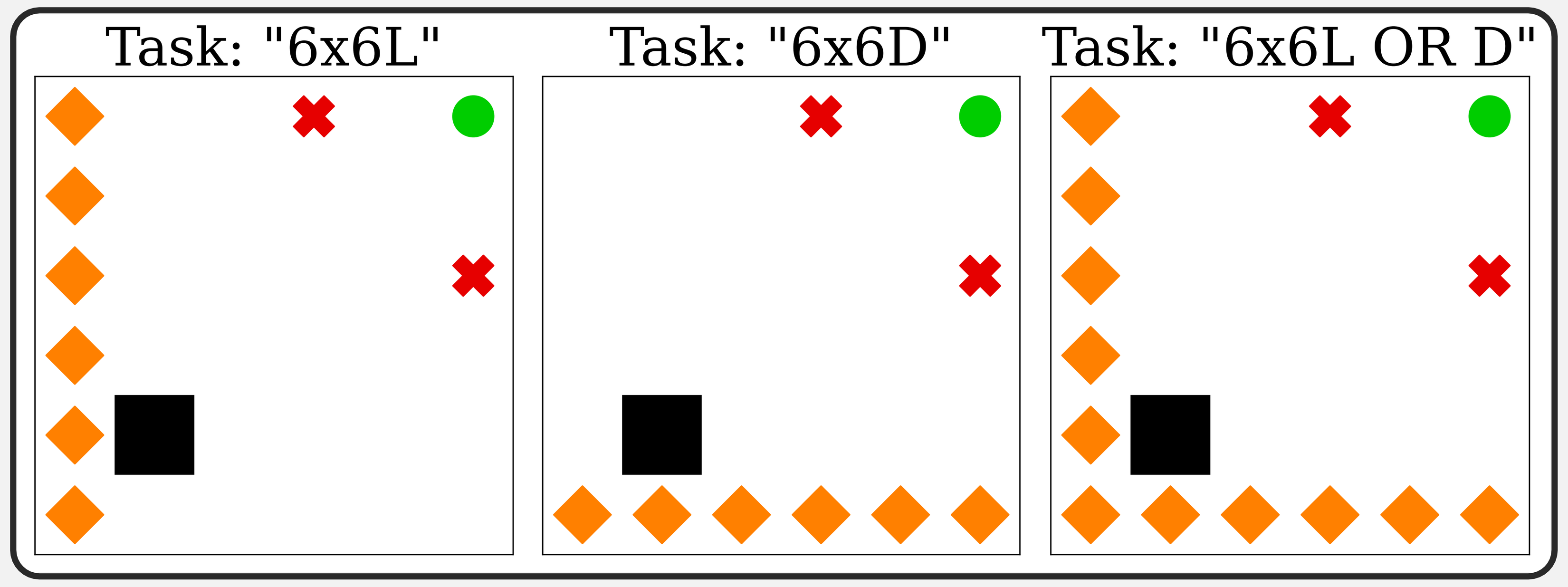
Bounded suboptimality

The greedy composition policy $\pi_f(a|s) = \text{argmax}_a f(Q(s,a))$ has a bounded suboptimality:

$$\tilde{Q}(s,a) - \tilde{Q}^{\pi_f}(s,a) \le D(s,a)$$

$$D(s,a) = r_D(s,a) + \gamma \mathbb{E} D(s',a')$$
$$r_D(s,a) = \gamma \mathbb{E}\left[\max_A\{f(Q(s',A)) + C(s',A)\} - f(Q(s',a'))\right]$$

Example



Pre-trained "primitive" tasks → Target "composed" task

$$\{Q^{(L)}(s,a), Q^{(D)}(s,a)\} \rightarrow \tilde{Q}(s,a) \le \max\{Q^{(L)}(s,a), Q^{(D)}(s,a)\}$$

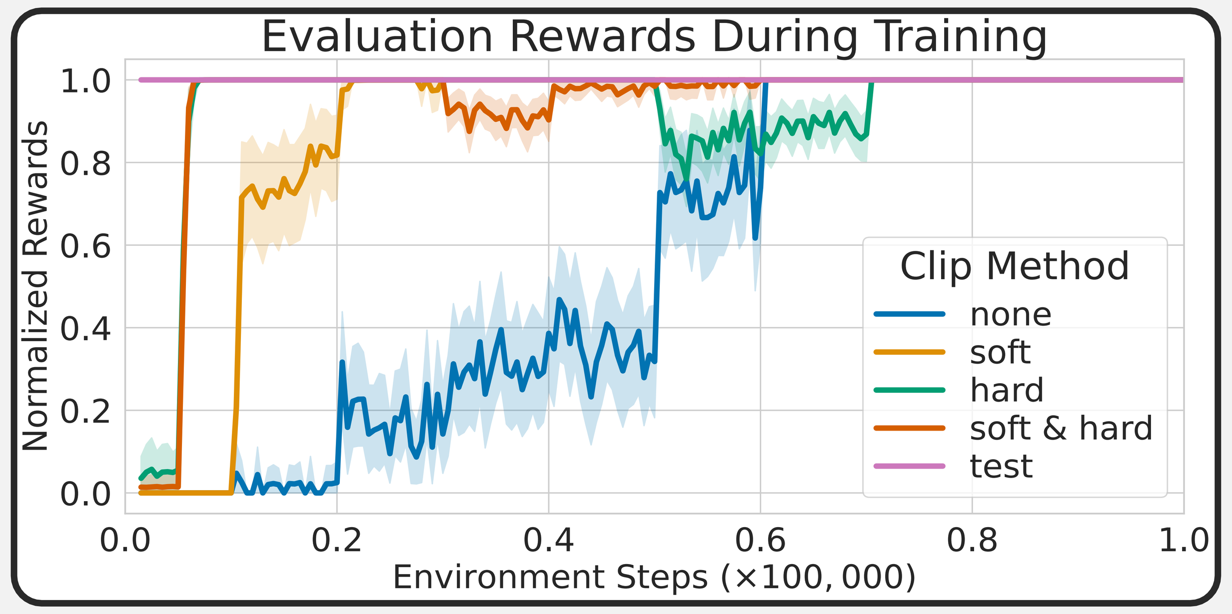
Primitive tasks provide a bound on the optimal value function for the composite task

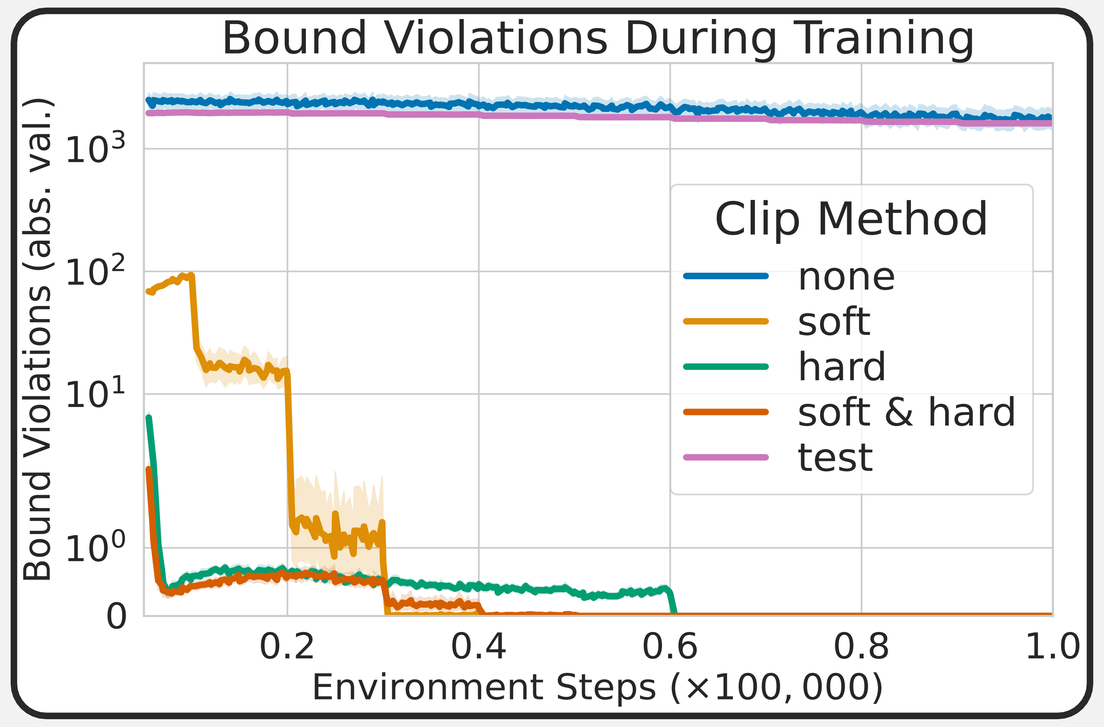
Applications

To restrict **bound violations** (BVs) that occur during training, the standard Bellman loss function is amended. Inspired by [6], we use the following clipping methods:

- **Soft clipping**: Bound violations appended to loss function
- **Hard clipping**: Clip the proposed Q-values to respect bounds
- **Test clipping**: Hard clipping at evaluation time only

These clipping methods improve training while reducing BVs.
Note: True Q-values are not being learned when BVs are nonzero.





References

[1]: R. Sutton and A. Barto "Reinforcement Learning: An introduction." MIT Press 2018
[2]: B. Ziebart "Modeling Purposeful Adaptive Behavior with the Principle of Maximum Causal Entropy" CMU PhD Thesis, 2010
[3]: T. Haarnoja, et. al. "Composable deep reinforcement learning for robotic manipulation." ICRA 2018
[4]: G.N. Tasse, et. al. "A Boolean task algebra for reinforcement learning." NeurIPS 2020
[5]: E. Todorov "Compositionality of optimal control laws" NeurIPS 2009
[6]: J. Kim, et. al. "Constrained GPI for zero-shot transfer in reinforcement learning" NeurIPS 2022

Link to Paper:

Acknowledgements

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. JA and RVK would like to acknowledge support from the Proposal Development Grant provided by the University of Massachusetts Boston. ST and VM acknowledge support from the Alliance Innovation Lab in Silicon Valley