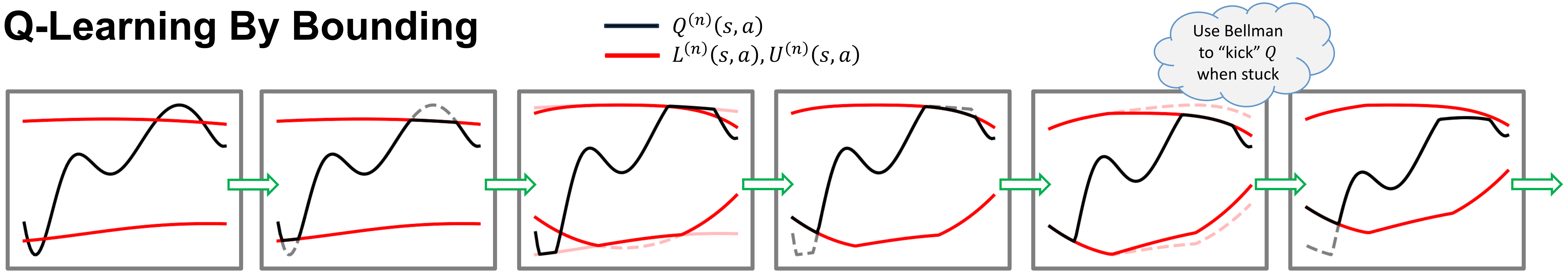
Boosting Soft Q-Learning By Bounding

Jacob Adamczyk,[1] Volodymyr Makarenko,[2] 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

An agent's ability to leverage past experience is critical for efficiently solving new tasks. Prior work has focused on using value function estimates to obtain zero-shot approximations for solutions to a new task. In soft Q-learning, we show how any value function estimate can also be used to derive double-sided bounds on the optimal value function. The derived bounds lead to new approaches for boosting training performance which we validate experimentally. Notably, we find that the proposed framework suggests an alternative method for updating the Q-function, leading to improved performance.

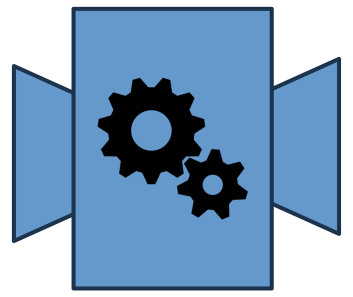
Q-Learning By Bounding



Initialize Q-values and calculate bounds → Clip the Q-values based on current bounds → Calculate new (tightest) upper and lower bounds → Clip the Q-values based on current bounds → Calculate new (tightest) upper and lower bounds → Clipping did not occur: Use TD update and continue

Main Result

Double-sided Bound on $Q^*(s,a)$ From Any Estimate $\hat{Q}(s,a)$

$Q^{(n)}(s,a)$  $L^{(n)} \leq Q^* \leq U^{(n)}$

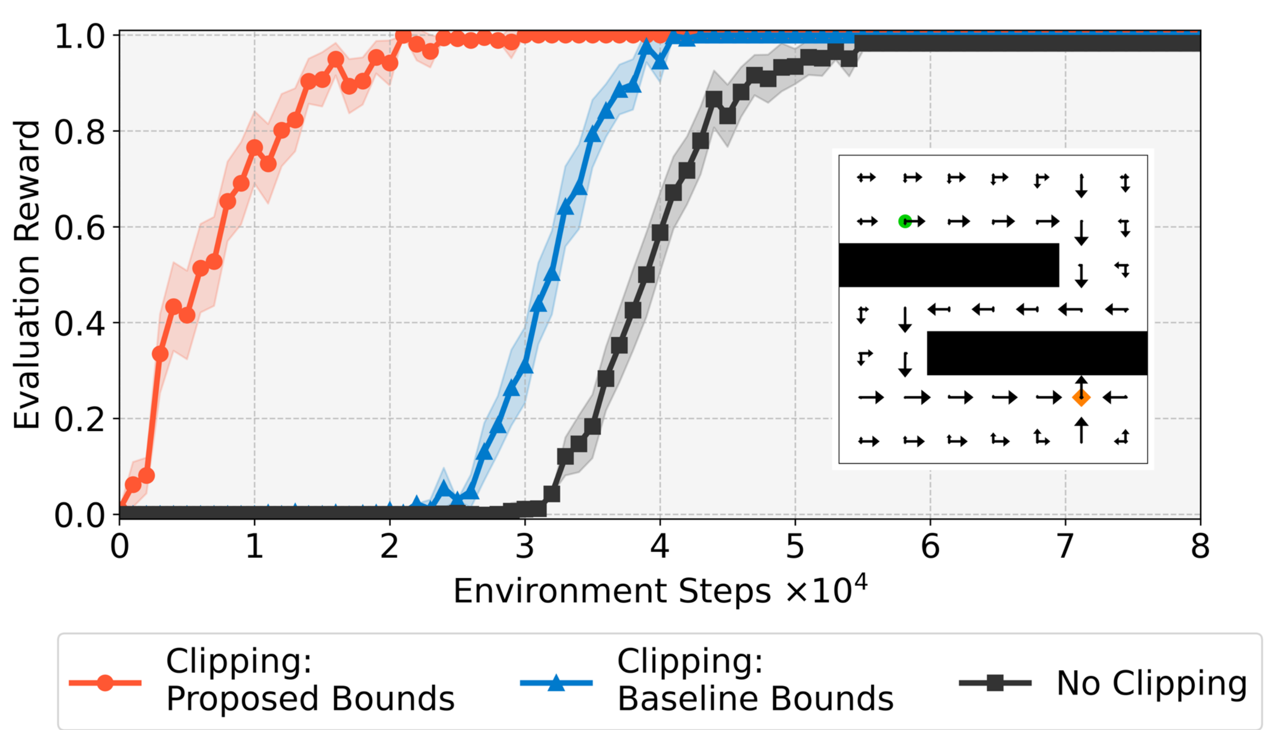
$$L^{(n)}(s,a) = r + \gamma V(s') + \frac{\min \Delta}{1-\gamma}$$

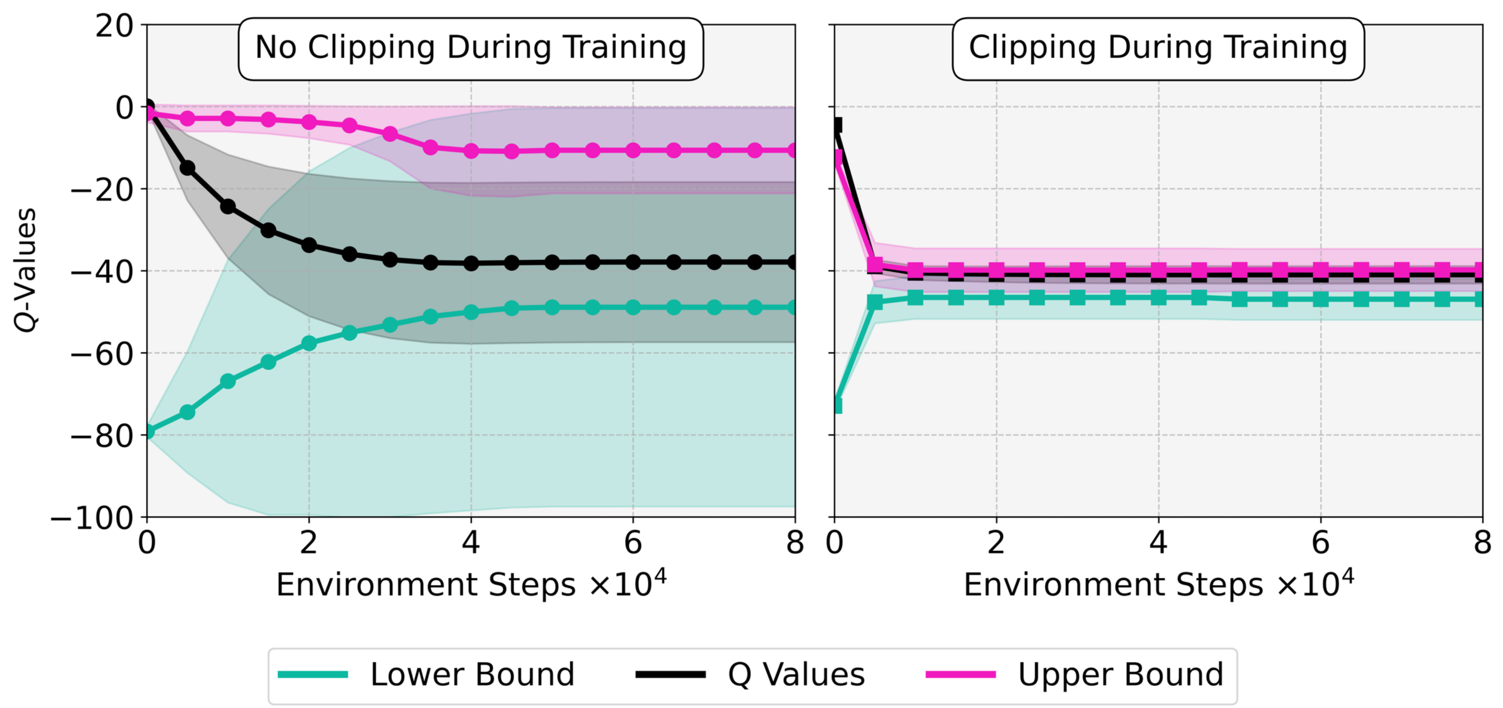
$$U^{(n)}(s,a) = r + \gamma V(s') + \frac{\max \Delta}{1-\gamma}$$

$$\Delta(s,a) = r(s,a) + \gamma V(s') - Q(s,a)$$

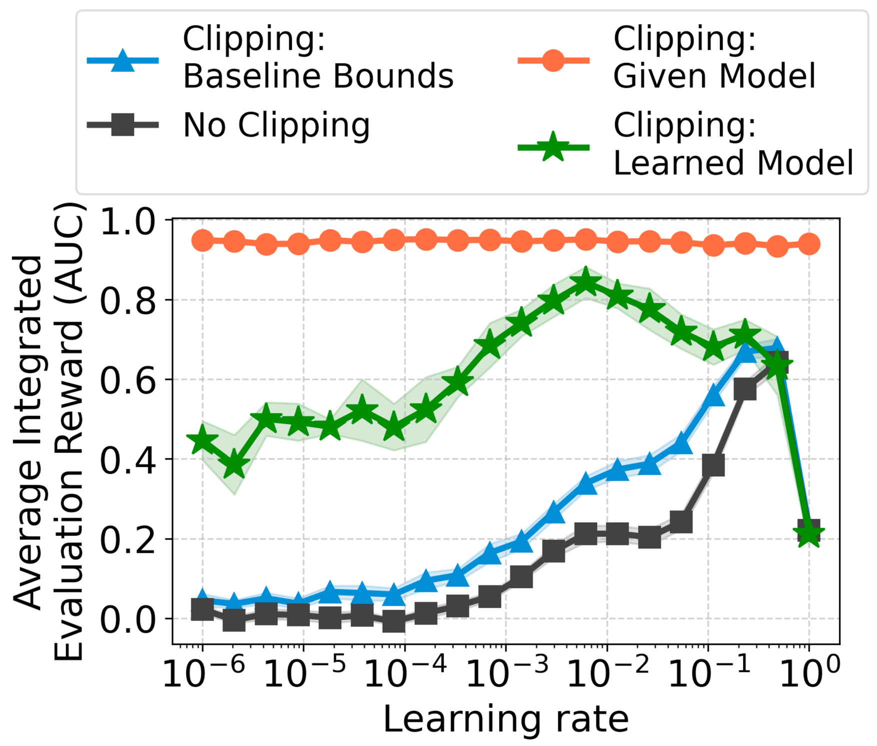
As $n \to \infty$: $L^{(n)}, U^{(n)} \to Q^*$

Boosted Training – Tight Bounds





Clipping is All You Need*



Conclusions

- Method for generating progressively tighter bounds without prior knowledge
- Our clipping method pushes *away* from invalid Q-values, whereas TD pulls *toward* valid Q-values
- We find the former (clipping) to be significantly faster as it quickly reduces potential solution space
- We derive theoretical results and run initial validation experiments in the deep RL setting

References & Acknowledgements

[1]: B. Eysenbach, S. Levine. "Maximum Entropy RL (Provably) Solves Some Robust RL Problems" (2022);
[2]: **JA**, Volodymyr Makarenko, Argenis Arriojas, Stas Tiomkin, and Rahul V. Kulkarni. "Bounding the optimal value function in compositional reinforcement learning", UAI 2023;
[3]: Cédric Malherbe and Nicolas Vayatis. "Global optimization of Lipschitz functions", ICML 2017;
[4]: Emmanuel Rachelson and Michail G. Lagoudakis. "On the locality of action domination in sequential decision making," ISIAM 2010.
[5]: Jaekyeom Kim, Seohong Park, and Gunhee Kim. "Constrained GPI for zero-shot transfer in reinforcement learning", NeurIPS 2022

This work was supported by the National Science Foundation through Award DMS-1854350, the Proposal Development Grant provided by the UMB the CSM Dean's Doctoral Research Fellowship through fellowship support from Oracle, project ID R20000000025727, the Research Foundation at San Jose State University, and the Alliance Innovation Lab in Silicon Valley.