Title: Preference-Based Thompson Sampling Reinforcement Learning

Authors: Ellen Novoseller, Yanan Sui, Yisong Yue, and Joel Burdick March 27, 2019

Abstract: In the preference-based reinforcement learning (RL) setting, an agent seeks to optimize its actions in an environment by receiving feedback; however, rather than receiving a numerical reward signal, as is typical in RL, the agent receives feedback in the form of preferences given by a user or expert. In particular, our work studies RL with preferences over trajectories of the agent's states and actions. Extending ideas in prior work on (1) self-sparring in the bandit setting [1] and (2) posterior sampling RL [2], we propose a preference- based RL algorithm that utilizes Thompson sampling to learn both the system dynamics and underlying utility function governing the expert's preferences. In particular, we assume that an underlying reward function over the state and action spaces generates the user's potentially-noisy preferences. Our method leverages Gaussian process modeling [3] to learn this underlying reward function, thereby tackling the credit assignment problem. This Gaussian process approach can handle large state or action spaces. Finally, we will demonstrate the algorithm's empirical performance in a simulated domain. Our ongoing work includes proving the first theoretical guarantees for bounding regret in the preference-based RL setting.

We hope to apply this work toward optimizing electrical stimulation therapy for patients with spinal cord injury. This treatment can enable patients to regain some control of their paralyzed limbs [4]; however, optimizing the parameters of the electrical stimulation remains a challenge, requiring a search over an intractably-large signal space. Dueling bandit approaches have been successfully applied toward this domain [1, 5], helping clinicians to optimize stimulation signals for aiding patients to stand independently and to grasp objects. The dueling bandit setting [6] assumes that feedback takes the form of preferences, which allows clinicians to rank patient trials rather than assign scores on an absolute scale, as human feedback is often more reliable in this form. While the bandit setting assumes that the optimal stimulation parameters are static, RL models a system in which optimal stimulation varies according to the patient's state, and in which the stimulation affects the patient's state. Given that empirical studies demonstrate that to enable a paralyzed patient to walk, different stimulation parameters should be applied in different segments of the walking cycle, we hope that our work on preference-based RL will aid spinal cord-injured patients in such relatively-complex movements as walking. Learning the function that governs the clinicians' preference-based scoring could further help create a quantifiable, interpretable model of patients' rehabilitative progress. In general, solving the credit assignment problem—modeling latent causes of the user's preferences—could yield more interpretable control policies in a variety of applications.

References

[1] Y. Sui, V. Zhuang, J. W. Burdick, and Y. Yue, "Multi-dueling bandits with dependent arms," in *Pro- ceedings of the Conference on Uncertainty in Artificial Intelligence*, 2017.

[2] I. Osband, D. Russo, and B. Van Roy, "(More) efficient reinforcement learning via posterior sampling," in *Advances in Neural Information Processing Systems*, 2013, pp. 3003–3011.

[3] C. E. Rasmussen and C. K. Williams, "Gaussian processes for machine learning," *The MIT Press*, vol. 2, no. 3, p. 4, 2006.

[4] S. Harkema, Y. Gerasimenko, J. Hodes, J. Burdick, C. Angeli, Y. Chen, C. Ferreira, A. Willhite, E. Rejc,
R. G. Grossman *et al.*, "Effect of epidural stimulation of the lumbosacral spinal cord on voluntary movement, standing, and assisted stepping after motor complete paraplegia: A case study," *The Lancet*, vol. 377, no. 9781, pp. 1938–1947, 2011.

[5] Y. Sui, Y. Yue, and J. W. Burdick, “Correlational dueling bandits with application to clinical treatment in large decision spaces,” in *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*, 2017, pp. 2793–2799.

[6] Y. Yue, J. Broder, R. Kleinberg, and T. Joachims, “The k-armed dueling bandits problem,” *Journal of Computer and System Sciences*, vol. 78, no. 5, pp. 1538–1556, 2012.