Asymptotic Performance Guarantee for Online Reinforcement Learning with the Least-Squares Regression

Mohammad Gheshlaghi Azar
Hilbert J. Kappen

February 10, 2011

We introduce a new online reinforcement learning (RL) method called least-squares action preference learning (LS-APL) for learning the near-optimal policy in Markovian decision processes (MDPs) (Bertsekas, 2007). Online RL aims at learning a policy to control a system in an incremental fashion, such that some measure of long-term performance is maximized by that (optimal) policy. A typical setting where online RL operates is as follows: Given the state x and the behavior policy $\bar{\pi}$ the controller calculates a control action a which is sent back to the system. The system then makes a transition to the new state x' and issues a control feedback (reward) and the cycle is repeated. The learning problem is to gradually improve the estimate of optimal control based on a history of observations (state-action-reward).[1] Although many online RL algorithms with various levels of success have been proposed during last 20 years (Maei et al., 2010; Melo et al., 2008; Szepesvari and Smart, 2004), we know of no theoretical guarantee in terms of performance loss for general function approximation. This paper provides an asymptotic performance guarantee for online RL, relying on a new variant of dynamic programming (DP) for iterating the control policy.

Dating all the way back to Boyan (1999) the least-squares regression (LS) has been widely used to scale up reinforcement learning algorithms to large state-action problems (Szepesvari, 2009; Munos and Szepesvári, 2008; Lagoudakis and Parr, 2003). Among several least-squares RL, least-squares fitted Q-iteration (LSFQI) and the least-squares policy iteration (LSPI), have gained attraction of many theoreticians. The asymptotic as well as the finite-time behavior of these algorithms have been throughly analyzed for both parametric and non-parametric type of function approximation (massoud Farahmand et al., 2008; Antos et al., 2008, 2007). For both LSFQI and LSPI one can show that the asymptotic loss compared to the optimal policy converges to zero provided that the function space grows by the number of samples in a controlled fashion. However, the fact

[1] Note that, in general, the behavior policy is not required to follow the latest estimate of the optimal policy. However, to maximize the long-term performance online RL methods often rely on this estimate for decision making.

that the quality of these performance loss bounds depends on the number of samples per iteration and not the total number of samples, see Munos and Szepesvári (2008), makes it difficult, if not impossible, to apply these results for online applications, since, in online problems, we typically have limited sampling and computational budget per iteration.

One approach to dealing with this issue may involve looking for a kind of performance-loss bound that relaxes dependency on estimation error. A potential solution may be found in very recent work by massoud Farahmand et al. (2010) that establishes a new fine-time performance loss bounds for approximate policy iteration (API) and approximate value iteration (AVI)(Bertsekas, 2007). This new result shows that the contribution of approximation (estimation) error to performance loss is more prominent in latter iterations of AVI/API algorithm and the effect of an error term in early iterations decays exponentially fast. In other words, it is better to put more effort on having lower approximation error at later iterations of API/AVI. In sampling-based algorithms like LSFQI and LSPI, this can be done by gradually increasing the number of samples throughout iterations.

We take a different approach to reduce the sample complexity based on a new variant of dynamic programming (DP) algorithm called dynamic policy programming (DPP)(Azar and Kappen, 2010). LS-APL can be characterized as a sampling-based version of DPP with least-squares regression, for which we establish asymptotic performance loss compare to the optimal policy. Given the existence of an upper-bound for the approximation error of DPP operator, we prove that the asymptotic performance loss of LS-APL is bounded with probability (w.p.) 1 by some finite value. The bound is similar to those of LSFQI/LSPI (Antos et al., 2008, 2007) in many respects. The key difference is that, to achieve non-trivial performance loss bound w.p.1, LS-APL requires a finite number of samples per iteration, whereas LSFQI/LSPI require infinitely many samples throughout iterations. This result is applied to online settings, where the learning algorithms can only make use of few samples at each step of learning.

Our analysis is based on the observation that the performance loss of DPP depends on some measure of average of approximation (estimation) error instead of max-norm error in the case of AVI and API. The idea is to show that for LS-APL variant of approximate DPP this average normed-error concentrate around a value which is bounded by some approximation error of DPP operator which depends on the capacity of function space. For measure concentration, we rely on the recent results from the Martingale difference processes literature concerning the concentration measure of the function of strongly dependent random processes (Kontorovich and Ramanan, 2007).

References

Antos, A., Munos, R., and Szepesvári, C. (2007). Fitted q-iteration in continuous action-space mdps. In *Proceedings of the 21st Annual Conference on Neural Information Processing Systems*.

Antos, A., Szepesvári, C., and Munos, R. (2008). Learning near-optimal policies with bellman-residual minimization based fitted policy iteration and a single sample path. *Machine Learning*, 71(1):89–129.

Azar, M. G. and Kappen, H. J. (2010). Dynamic policy programming. *CoRR*, abs/1004.2027.

Bertsekas, D. P. (2007). *Dynamic Programming and Optimal Control*, volume II. Athena Scientific, Belmount, Massachusetts, third edition.

Boyan, J. A. (1999). Least-squares temporal difference learning. In *ICML*, pages 49–56.

Kontorovich, L. and Ramanan, K. (2007). Concentration inequalities for dependent random variables via the martingale method. *The Annals of Probability*, 36(6):2126–2158.

Lagoudakis, M. G. and Parr, R. (2003). Least-squares policy iteration. *Journal of Machine Learning Research*, 4:1107–1149.

Maei, H. R., Szepesvári, C., Bhatnagar, S., and Sutton, R. S. (2010). Toward off-policy learning control with function approximation. In *ICML*, pages 719–726.

massoud Farahmand, A., Ghavamzadeh, M., Szepesvári, C., and Mannor, S. (2008). Regularized fitted q-iteration: Application to planning. In *EWRL*, pages 55–68.

massoud Farahmand, A., Munos, R., and Szepesvari, C. (2010). Error propagation for approximate policy and value iteration. In Lafferty, J., Williams, C. K. I., Shawe-Taylor, J., Zemel, R., and Culotta, A., editors, *Advances in Neural Information Processing Systems 23*, pages 568–576.

Melo, F., Meyn, S., and Ribeiro, I. (2008). An analysis of reinforcement learning with function approximation. In *Proceedings of 25st International Conference on Machine Learning*.

Munos, R. and Szepesvári, C. (2008). Finite-time bounds for fitted value iteration. *Journal of Machine Learning Research*, 9:815–857.

Szepesvari, C. (2009). Reinforcement learning algorithms for mdps – a survey. Technical Report TR09–13, Department of Computing Science, University of Alberta, Edmonton, Alberta, Canada.

Szepesvari, C. and Smart, W. (2004). Interpolation-based q-learning. In *Proceedings of 21st International Conference on Machine Learning*, volume 69 of *ACM International Conference Proceeding Series*, page 100, Banff, Alberta, Canada.