Efficient Sample Reuse by Covariate Shift Adaptation in Value Function Approximation

Department of Computer Science, Tokyo Institute of Technology,
2-12-1, O-okayama, Meguro-ku, Tokyo, 152-8552, Japan

Hirotaka Hachiya hachiya@sg.cs.titech.ac.jp
Takayuki Akiyama akiyama@sg.cs.titech.ac.jp
Masashi Sugiyama sugi@cs.titech.ac.jp

Policy iteration is a general framework to obtain the optimal policy by iteratively performing value function approximation and policy improvement [6]. A traditional practice of policy iteration is, when policies are updated, new data samples are gathered following the new policy and are used for value function approximation. However, this approach is inefficient particularly when the sampling cost is high since previously gathered data samples are simply discarded; it would be more efficient if we could reuse the data collected in the past. A situation where the behavior policy (a policy used for gathering data samples) and the current target policy are different is called *off-policy* reinforcement learning [6].

In the off-policy situation, simply employing a standard policy iteration method (such as *least-squares* policy iteration [2]) does not lead to the optimal policy due to the bias caused by the difference between behavior and target policies. This policy mismatch problem could be eased by the use of *importance sampling* techniques [1]—the bias caused by the policy mismatch is asymptotically canceled. However, the approximation error is not necessarily small when the bias is reduced by importance sampling; the variance of estimators should also be taken into account since the approximation error is the sum of squared bias and variance. Due to large variance, existing importance sampling techniques tend to be unstable [6], [3].

To overcome the instability problem, we propose using an *adaptive importance sampling* technique used in statistics [4]. The proposed adaptive method, which smoothly bridges the ordinary estimator and importance-weighted estimator, allows us to control the trade-off between bias and variance. Thus, given that the adaptive parameter is chosen carefully, the optimal performance can be acheived in terms of both bias and variance. However, the optimal value of the adaptive parameter is heavily dependent on problems, and therefore using a prefixed adaptive parameter may not be always effective in practice.

For optimally choosing the value of the trade-off parameter, we reformulate the value function approximation problem as a supervised regression problem and propose using an automatic model selection method based on a statistical machine learning theory [5]. The method called *importance-weighted cross-validation* enables us to estimate the approximation error of value functions in an unbiased manner even under off-policy situations. Thus we can actively determine the adaptive parameter based on data samples at hand. We demonstrate the usefulness of the proposed approach in standard chain-walk and mountain-car benchmark problems.

REFERENCES

[1] G. S. Fishman. *Monte Carlo: Concepts, Algorithms, and Applications*. Springer-Verlag, Berlin, 1996.
[2] M. G. Lagoudakis and R. Parr. Least-squares policy iteration. *Journal of Machine Learning Research*, (4):1107–1149, 2003.
[3] D. Precup, R. S. Sutton, and S. Singh. Eligibility traces for off-policy policy evaluation. Morgan Kaufmann, 2000.
[4] H. Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference*, 90(2):227–244, 2000.
[5] M. Sugiyama, M. Krauledat, and K.-R. Müller. Covariate shift adaptation by importance weighted cross validation. *Journal of Machine Learning Research*, 8:985–1005, May 2007.
[6] R. S. Sutton and A. G. Barto. *Reinforcement Learning: An Introduction*. The MIT Press, 1998.