Optimal Online Learning Procedures for Model-Free Policy Evaluation

Tsuyoshi Ueno[1], Shin-ichi Maeda[1], Motoaki Kawanabe[2] and Shin Ishii[1]
[1] Kyoto University, [2] Fraunhofer FIRST

Theoretical advances in reinforcement learning have been contributed to optimal control and decision making in various practical applications. In order to find optimal strategies, it is important, in particular in model-free approaches, to estimate the value function, which indicates the goodness of the current policy, from a given sample trajectory. There are two major ways in value function estimation. The temporal difference (TD) learning updates the current estimator step-by-step manner, and each step uses a relatively small number of samples (online procedure). On the other hand, the least squares temporal difference (LSTD) learning obtains an estimator in one shot computation by using all the samples in the given trajectory (batch procedure). All the other algorithms proposed so far are also categorized into one of two classes: online procedure and batch procedure. The convergence of an online learning algorithm is slow, in general, compared to a batch learning algorithm. However, in reinforcement learning, online learning is often preferred to batch learning because of the computational efficiency and adaptability.

Recently, [1] introduced introduced a framework of semiparametric statistical inference to model-free policy evaluation. The semiparametric statistical models include not only parameters of interest but also additional nuisance parameters which can have infinite degrees of freedom. For estimating the parameters of interest in such models, estimating functions provide a well-established toolbox: they give consistent estimators (M-estimators) regardless of the nuisance parameters [2]. Applying this technique to the estimation of the value functions, they discussed asymptotic statistical property of LSTD-like learning procedures and proposed the generalized LSTD (gLSTD) based on the optimal estimating function that achieved the minimum estimation error. Although the framework by [1] has potential to bring new insights to reinforcement learning, their theory can only deal with batch procedures and a bunch of online algorithms such as TD were excluded.

In this study, we extend their semiparametric statistical framework to be applicable to online learning procedures [3, 4]. More specifically, we characterize Markov reward process (MRP) as a semiparametric model, where only the value function $V(s)$ is modelled parametrically with small number of parameters, while the other unspecified parts of MRP are regarded as the nuisance

parameters. In this semiparametric model, we investigate martingale estimating functions [5] which bring us consistent estimators of the parameters of interest *without estimating the nuisance parameters*. Furthermore, we obtain explicitly the general form of possible estimating functions and the optimal one which gives the consistent estimator with *minimum asymptotic variance*. Then, we discuss the online learning procedures derived from estimating functions in general. Our main results are:

(a) We show a general class of online procedures for model-free policy evaluation derived from estimating functions which includes many famous model-free policy evaluation methods such as TD learning and least squares policy evaluation (LSPE) learning.

(b) We examine convergence of statistical deviations of such estimators and show that online algorithms can achieve the same asymptotic performance as their batch counterparts if a matrix factor is properly tuned.

(c) Based on the above results, we derive the optimal choice of the estimating function and construct a novel online learning algorithm that achieves the minimum estimation error asymptotically. We also propose acceleration of TD learning.

We compare the performance of proposed online algorithms to a couple of well-established algorithms in a simple numerical experiment and show that the result matches well with our theoretical findings.

References

[1] Ueno, T., Kawanabe, M., Mori, T., Maeda, S., Ishii, S.: A semiparametric statistical apparoach to model-free policy evaluation. In: In Proceedings of the 25th International Conference on Machine Learning. (2008) 1072–1079

[2] Godambe, V., ed.: Estimating Functions. Oxford Science (1991)

[3] Bottou, L., LeCun, Y.: On-line learning for very large datasets. Applied Stochastic Models in Business and Industry **21**(2) (2005) 137–151

[4] Amari, S.: Natural gradient works efficiently in learning. Neural Computation **10**(2) (1998) 251–276

[5] Godambe, V.: The foundations of finite sample estimation in stochastic processes. Biometrika **72**(2) (1985) 419–428