# Optimal Online Learning Procedures for Model-Free Policy Evaluation

Tsuyoshi Ueno[1], Shin-ichi Maeda[1], Motoaki Kawanabe[2], and Shin Ishii[1]

[1] Graduate School of Informatics, Kyoto University
{tsuyos-u,ichi,ishii}@sys.i.kyoto-u.ac.jp
[2] Fraunhofer FIRST and Berlin Institute of Technology, Germany
motoaki.kawanabe@first.fraunhofer.de

**Abstract.** In this study, we extend the framework of semiparametric statistical inference introduced recently to reinforcement learning [1] to online learning procedures for policy evaluation. This generalization enables us to investigate statistical properties of value function estimators both by batch and online procedures in a unified way in terms of estimating functions. Furthermore, we propose a novel online learning algorithm with optimal estimating functions which achieve the minimum estimation error. Our theoretical developments are confirmed using a simple chain walk problem.

## 1 Introduction

Reinforcement learning is a class of machine learning based on reward-related interactions with environments, and has successfully been applied to various control problems [2]. In order to find out optimal strategies, it is important, in particular in model-free approaches, to estimate the value function which denotes goodness of the current policy, from a given sample trajectory. There are two major ways in value function estimation. The temporal difference (TD) learning [2] updates the current estimator step-by-step whose step uses a relatively small number of samples (online procedure). On the other hand, the least squares temporal difference (LSTD) learning [3,4] obtains an estimator in one shot by using all samples in the given trajectory (batch procedure). Other algorithms proposed so far are also categorized into one of these two groups.

Recently, [1] introduced a novel 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 may have infinite degrees of freedom [5,6,7]. For estimating the parameters of interest in such models, estimating functions provide a well-established toolbox: they give consistent estimators (M-estimators) without knowing the nuisance parameters [5,8]. Applying this technique to Markov decision processes (MDP), they discussed asymptotic properties of LSTD-like learning procedures and proposed the generalized LSTD (gLSTD) based on the optimal estimating function that achieved the minimum error. Although the framework by [1] has potential to bring new insights to reinforcement learning, their theory could only

W. Buntine et al. (Eds.): ECML PKDD 2009, Part II, LNAI 5782, pp. 473–488, 2009.

deal with batch procedures and a bunch of online algorithms such as TD were excluded.

In this article, we extend their semiparametric statistical techniques to be applicable to online learning procedures as to follow the existing analysis of online learning [9]. This extension leads to a general class of online learning procedures for model-free policy evaluation derived from estimating functions, which includes many popular algorithms [2,10] such as TD learning [2] and least squares policy evaluation (LSPE) [11]. This generalization also allows us to examine the convergence of statistical error and hence to see that online algorithms can achieve the same asymptotic performance as their batch counterparts if a matrix factor is properly tuned (Theorem 4). Based on this fact, we can accelerate TD learning (Section 5.4). Furthermore, we can derive the optimal choice of the estimating function and construct a novel online learning algorithm which achieves the *minimum estimation error* asymptotically (Algorithm 1).

This article is organized as follows. In Section 2, a semiparametric setting of Markov reward processes (MRPs) is presented. We explain the concept of estimating functions in Section 3, before going into those for MRPs in Section 4. Then, in Section 5, we discuss online learning procedures derived from estimating functions. Convergence theorems for such algorithms will be presented, followed by a novel algorithm with the optimal estimating function. In Section 6, the performance of the proposed algorithms are compared to a couple of well-established algorithms using a simple chain walk problem.

## 2 Markov Reward Process

Following the literature of policy evaluation [12], we consider Markov Reward Processes (MRPs) in this study. However, extension to Markov Decision Processes (MDPs) is straightforward as long as focusing on policy evaluation (hence the policy is fixed).

An MRP is defined by the initial state probability $p(s_0)$, the state transition probability $p(s_{t+1}|s_t)$ and the reward probability $p(r_{t+1}|s_t, s_{t+1})$. The state variable $s$ is an element of a finite set $S$ and the reward variable $r \in R$ can be either discrete or continuous, but a finite value.

The joint distribution of a sample trajectory $Z_T := \{s_0, s_1, r_1 \cdots, s_T, r_T\}$ of the MRP is described as

$$p(Z_T) = p(s_0) \prod_{t=0}^{T-1} p(r_{t+1}|s_t, s_{t+1}) p(s_{t+1}|s_t). \tag{1}$$

We also impose the following assumptions on MRPs.

**Assumption 1.** *Under $p(s_{t+1}|s_t)$, MRP has a unique invariant stationary distribution $\mu(s)$.*

**Assumption 2.** *For any time $t$, the state $s_t$ and the reward $r_t$ are uniformly bounded.*

Here, we introduce a statistical framework by confirming that the value function estimation can be interpreted as the estimation of certain statistics of MRP (1).

**Proposition 1.** *[10] Consider a conditional probability of* $\{r_{t+1}, s_{t+1}\}$ *given* $s_t$*,*

$$p(r_{t+1}, s_{t+1}|s_t) = p(r_{t+1}|s_t, s_{t+1})p(s_{t+1}|s_t).$$

*Then, there is such a function* $V$ *that*

$$\mathbb{E}\left[r_{t+1}|s_t\right] = V(s_t) - \gamma\mathbb{E}[V(s_{t+1})|s_t] \tag{2}$$

*holds for any state* $s_t$*. Here,* $\mathbb{E}\left[\cdot|s\right]$ *denotes the conditional expectation for a given state* $s$*. The function* $V$ *that satisfies eq. (2) is unique and found to be a value function;*

$$V(s) := \lim_{T\to\infty} \mathbb{E}\left[\sum_{t=0}^{T} \gamma^t r_{t+1} \middle| s_0 = s\right], \tag{3}$$

*where* $\gamma \in [0, 1)$ *is a constant called the discount factor.*

We assume throughout this article that the value function can be represented by a certain parametric function, including a nonlinear function with respect to the parameter.

**Assumption 3.** *The value function given by eq. (3) is represented by a parametric function* $g(s, \boldsymbol{\theta})$*;*

$$V(s) = g(s, \boldsymbol{\theta}),$$

*where* $g : S \to \mathbb{R}$*,* $\boldsymbol{\theta} \in \mathbb{R}^m$ *is a parameter. Moreover,* $g(s, \boldsymbol{\theta})$ *is assumed to be twice-differentiable with respect to* $\boldsymbol{\theta}$*, and* $g(s, \boldsymbol{\theta}) < \infty$ *for any* $s \in S$ *and* $\boldsymbol{\theta}$*.*

Under Assumption 3, $p(r_{t+1}|s_t)$ is partially parameterized by $\boldsymbol{\theta}$, through its conditional mean

$$\mathbb{E}[r_{t+1}|s_t] = g(s_t, \boldsymbol{\theta}) - \gamma\mathbb{E}[g(s_{t+1}, \boldsymbol{\theta})|s_t]. \tag{4}$$

Our goal is to find out such a value of the parameter $\boldsymbol{\theta}$ that the function $g(s, \boldsymbol{\theta})$ satisfies eq. (4), that is, it coincides with the true value function.

In order to specify the probabilistic model (4) completely, we need usually extra parameters other than $\boldsymbol{\theta}$. Let $\boldsymbol{\xi}_0$ and $\boldsymbol{\xi}_s$ be such extra parameters that initial distribution $p(s_0, \boldsymbol{\xi}_0)$ and transition distribution $p(r, s|s; \boldsymbol{\theta}, \boldsymbol{\xi}_s)$ are completely identified, respectively. In such a case, the joint distribution of the trajectory $Z_T$ is expressed as

$$p(Z_T; \boldsymbol{\theta}, \boldsymbol{\xi}) = p(s_0; \boldsymbol{\xi}_0) \prod_{t=0}^{T-1} p(r_{t+1}, s_{t+1}|s_t; \boldsymbol{\theta}, \boldsymbol{\xi}_s), \tag{5}$$

where $\boldsymbol{\xi} = (\boldsymbol{\xi}_0, \boldsymbol{\xi}_s)$. Since there is no way to know the complexity of the target system, we attempt to estimate the parameter $\boldsymbol{\theta}$ without estimating the extra $\boldsymbol{\xi}$, which may have innumerable degrees of freedom. Statistical models which contain such (possibly infinite-dimensional) nuisance parameters ($\boldsymbol{\xi}$) in addition to the parameter of interest ($\boldsymbol{\theta}$) are said semiparametric [6]. We emphasize that the nuisance parameters are necessary only for theoretical discussions. In actual estimation of the parameters, same as in other model-free policy evaluation algorithms, we neither define them concretely, nor estimate them. This can be achieved by usage of estimating functions which is a well-established technique to obtain a consistent estimator of the parameter without estimating the nuisance parameter [5,7]. The advantages of considering such semiparametric models behind model-free approaches are:

(a) we can characterize all possible model-free algorithms,
(b) we can discuss asymptotic properties of the estimators in a unified way and obtain the optimal one with the asymptotically *minimum estimation error*.

We will summarize the estimating function method in the next section.

## 3 Estimating Functions in Semiparametric Models

We begin with a short overview of the estimating function theory in the i.i.d. case and then discuss the MRP case in the next section. We consider a general semiparametric model $p(\boldsymbol{x}; \boldsymbol{\theta}, \boldsymbol{\xi})$, where $\boldsymbol{\theta}$ is an $m$-dimensional parameter of interest and $\boldsymbol{\xi}$ is a nuisance parameter which can have infinite degrees of freedom. An $m$-dimensional vector function $\boldsymbol{f}$ is called an *estimating function* when it satisfies the following conditions for any $\boldsymbol{\theta}$ and $\boldsymbol{\xi}$;

$$\mathbb{E}_{\boldsymbol{\theta},\boldsymbol{\xi}}[\boldsymbol{f}(\boldsymbol{x}, \boldsymbol{\theta})] = \mathbf{0} \tag{6}$$

$$\det|\mathbf{A}| \neq 0, \quad \text{where } \mathbf{A} = \mathbb{E}_{\boldsymbol{\theta},\boldsymbol{\xi}}\left[\partial_{\boldsymbol{\theta}} \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{\theta})\right] \tag{7}$$

$$\mathbb{E}_{\boldsymbol{\theta},\boldsymbol{\xi}}\left[||\boldsymbol{f}(\boldsymbol{x}, \boldsymbol{\theta})||^2\right] < \infty, \tag{8}$$

where $\partial_{\boldsymbol{\theta}} = \partial/\partial\boldsymbol{\theta}$ is the partial derivative with respect to $\boldsymbol{\theta}$, and $\det|\cdot|$ and $||\cdot||$ denote the determinant and the Euclidean norm, respectively. Here $\mathbb{E}_{\boldsymbol{\theta},\boldsymbol{\xi}}[\cdot]$ means the expectation over $\boldsymbol{x}$ with $p(\boldsymbol{x}; \boldsymbol{\theta}, \boldsymbol{\xi})$ and we further remark that the parameter $\boldsymbol{\theta}$ in $\boldsymbol{f}(\boldsymbol{x}, \boldsymbol{\theta})$ and $\mathbb{E}_{\boldsymbol{\theta},\boldsymbol{\xi}}[\cdot]$ must be the same.

Suppose i.i.d. samples $\{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N\}$ are generated from the model $p(\boldsymbol{x}; \boldsymbol{\theta}^*, \boldsymbol{\xi}^*)$. If there is an estimating function $\boldsymbol{f}(\boldsymbol{x}, \boldsymbol{\theta})$, we can obtain an estimator $\hat{\boldsymbol{\theta}}$ which has good asymptotic properties, by solving the following estimating equation;

$$\sum_{i=1}^{N} \boldsymbol{f}(\boldsymbol{x}_i, \hat{\boldsymbol{\theta}}) = \mathbf{0}. \tag{9}$$

A solution of the estimating equation (9) is called an *M-estimator* in statistics [5]. The M-estimator is consistent, that is, it converges to the true value *regardless*

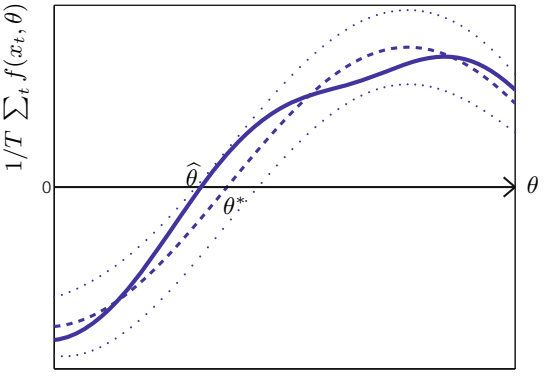


**Fig. 1.** An illustrative plot of $1/T \sum_t f(x_t, \theta)$ as a function of $\theta$ (the solid line). Due to the effect of finite samples, the function is slightly apart from its expectation $\mathbb{E}_{\theta^*,\xi^*}[f(x,\theta)]$ (the dashed line) which takes 0 at $\theta = \theta^*$ because of the condition (6). The condition (8) means that the expectation (the dashed line) has a non-zero slope around $\theta^*$, which ensures the local uniqueness of the zero crossing point. On the other hand, the condition (7) guarantees that its standard deviation shown by the two dotted lines shrinks in the order of $1/\sqrt{T}$, thus we can expect to find asymptotically at least one solution $\hat{\theta}$ of the estimating equation (9) near the true value $\theta^*$. This situation holds regardless of that the true nuisance parameter $\xi^*$ takes any possible value.

*of the nuisance parameter* $\boldsymbol{\xi}^*$. Moreover, it is normally distributed, that is, $\hat{\boldsymbol{\theta}} \sim \mathcal{N}(\boldsymbol{\theta}^*, \mathrm{Av})$ when the sample size $N$ approaches infinity. The matrix Av, which is called the asymptotic variance, can be calculated by

$$\mathrm{Av} := \mathrm{Av}(\hat{\boldsymbol{\theta}}) = \frac{1}{N}\mathbf{A}^{-1}\mathbb{E}_{\boldsymbol{\theta}^*,\boldsymbol{\xi}^*}\left[\boldsymbol{f}(\boldsymbol{x},\boldsymbol{\theta}^*)\boldsymbol{f}(\boldsymbol{x},\boldsymbol{\theta}^*)^\top\right](\mathbf{A}^\top)^{-1},$$

where $\mathbf{A} = \mathbb{E}_{\boldsymbol{\theta}^*,\boldsymbol{\xi}^*}\left[\partial_{\boldsymbol{\theta}}\boldsymbol{f}(\boldsymbol{x},\boldsymbol{\theta}^*)\right]$, and the symbol $\top$ denotes the matrix transpose. Note that Av depends on $(\boldsymbol{\theta}^*, \boldsymbol{\xi}^*)$, but not on the samples $\{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N\}$. We illustrate in Fig. 3 the left hand side of the estimating equation (9) in order to explain the reason why an M-estimator has nice properties and the meaning of conditions (6)-(8).

## 4 Estimating Functions in the MRP Model

The notion of estimating function has been extended to be applicable to Markov time-series [13,14]. To make it applicable to MRPs, we need similar extension. For convenience, we write the triplet at time $t$ as $z_t := \{s_{t-1}, s_t, r_t\}$. and the trajectory up to time $t$ as $Z_t := \{s_0, s_1, r_1, \ldots, s_t, r_t\}$.

We consider an $m$-dimensional vector valued function of the form:

$$\boldsymbol{f}_T(Z_T, \boldsymbol{\theta}) = \sum_{i=1}^{T} \boldsymbol{\psi}_t(Z_t, \boldsymbol{\theta}),$$

we attempt to estimate the parameter $\boldsymbol{\theta} \in \mathbb{R}^m$ for the given trajectory $Z_T$. This is similar to the left hand side of (9) in the i.i.d. case, but now each term $\boldsymbol{\psi}_t$ depends also on the previous observations, that is, a function of the sequence up to time $t$. If the sequence of the functions $\{\boldsymbol{\psi}_t\}$ satisfies the following properties for any $\boldsymbol{\theta}$ and $\boldsymbol{\xi}$, function $\boldsymbol{f}_T$ becomes an estimating function.

$$\mathbb{E}_{\boldsymbol{\theta},\boldsymbol{\xi}_s}\left[\boldsymbol{\psi}_t(Z_t,\boldsymbol{\theta})|Z_{t-1}\right] = \mathbf{0}, \quad \forall t \tag{10}$$

$$\det|\mathbf{A}| \neq 0, \quad \text{where} \quad \mathbf{A} := \mathbb{E}_{\boldsymbol{\theta},\boldsymbol{\xi}}\left[\partial_{\boldsymbol{\theta}} \boldsymbol{f}_T(Z_T,\boldsymbol{\theta})\right] \tag{11}$$

$$\mathbb{E}_{\boldsymbol{\theta},\boldsymbol{\xi}}\left[\left\|\boldsymbol{\psi}_t(Z_t,\boldsymbol{\theta})\right\|^2\right] < \infty, \quad \forall t. \tag{12}$$

Note that the estimating function $\boldsymbol{f}_T(Z_T,\boldsymbol{\theta})$ satisfies the martingale properties because of the condition (10). Therefore, it is called a *martingale estimating function* in literature [5]. Although time-series estimating functions can be defined in a more general form, the above definition is enough for our theoretical consideration.

### 4.1 Characterizing the Class of Estimating Functions

In this section, we characterize possible estimating functions in MRPs. Let $\epsilon_{t+1}$ be the TD error, that is,

$$\epsilon_{t+1} := \epsilon(z_{t+1},\boldsymbol{\theta}) := g(s_t,\boldsymbol{\theta}) - \gamma g(s_{t+1},\boldsymbol{\theta}) - r_{t+1}.$$

From (4), its conditional expectation $\mathbb{E}_{\boldsymbol{\theta},\boldsymbol{\xi}_s}\left[\epsilon_{t+1}|s_t\right]$ is equal to 0 for any state $s_t$. Furthermore, this zero-mean property holds even when multiplied by any weight function $\boldsymbol{w}_t := \boldsymbol{w}_t(Z_t)$ which depends only on the past observations, that is,

$$\mathbb{E}_{\boldsymbol{\theta},\boldsymbol{\xi}_s}\left[\boldsymbol{w}_t(Z_t)\epsilon_{t+1}|s_t\right] = \boldsymbol{w}_t(Z_t)\mathbb{E}_{\boldsymbol{\theta},\boldsymbol{\xi}_s}\left[\epsilon_{t+1}|s_t\right] = \mathbf{0},$$

for any $s_t$. From this observation, we can obtain a class of estimating functions $\boldsymbol{f}_T(Z_T,\boldsymbol{\theta})$ in MRPs.

**Lemma 1.** *Suppose that the random sequence $Z_T$ is generated from the distribution of the semiparametric model $\{p(Z_T;\boldsymbol{\theta},\boldsymbol{\xi}) \,|\, \boldsymbol{\theta},\boldsymbol{\xi}\}$ defined by (5). If the matrix $\mathbb{E}_{\boldsymbol{\theta},\boldsymbol{\xi}}\left[\sum_{t=1}^T \boldsymbol{w}_{t-1}(Z_{t-1})\left\{\partial_{\boldsymbol{\theta}}\epsilon(z_t,\boldsymbol{\theta})\right\}^\top\right]$ is nonsingular for any $\boldsymbol{\theta}$ and $\boldsymbol{\xi}$, then*

$$\boldsymbol{f}_T(Z_T,\boldsymbol{\theta}) = \sum_{t=1}^T \boldsymbol{\psi}_t(Z_t,\boldsymbol{\theta}) := \sum_{t=1}^T \boldsymbol{w}_{t-1}(Z_{t-1})\epsilon(z_t,\boldsymbol{\theta}) \tag{13}$$

*becomes an estimating function.*

From Lemma 1, we can obtain an M-estimator $\hat{\boldsymbol{\theta}}$ by solving the estimating equation

$$\sum_{t=1}^T \boldsymbol{\psi}_t(Z_t,\hat{\boldsymbol{\theta}}) = \mathbf{0}. \tag{14}$$

In general, estimating equations can be nonlinear with respect to the parameter $\boldsymbol{\theta}$. Therefore, in order to obtain a solution we need to employ iterative procedures, for example, online learning procedures as will be discussed in Section 5. The estimator derived from the estimating equation (14) has such an asymptotic variance that described by the following lemma.

**Lemma 2.** *Suppose that the random sequence* $\{Z_T\}$ *is generated from the distribution* $p(Z_T; \boldsymbol{\theta}^*, \boldsymbol{\xi}^*)$ *and* $\boldsymbol{w}_t$ *is a function of* $\{s_{0:t}, r_{1:t}\}$ *satisfying the condition of Lemma 1. Then, the M-estimator derived from eq. (14) has the asymptotic variance*

$$\mathrm{Av} = \mathrm{Av}(\hat{\boldsymbol{\theta}}) = \frac{1}{T}\mathbf{A}^{-1}\boldsymbol{\Sigma}\left(\mathbf{A}^\top\right)^{-1},$$

*where* $\mathbf{A} = \mathbf{A}(\boldsymbol{\theta}^*, \boldsymbol{\xi}^*) = \lim_{t\to\infty} \mathbb{E}_{\boldsymbol{\theta}^*,\boldsymbol{\xi}^*}\left[\boldsymbol{w}_{t-1}\left\{\partial_{\boldsymbol{\theta}}\epsilon(z_t, \boldsymbol{\theta}^*)\right\}^\top\right]$,
$\boldsymbol{\Sigma} = \boldsymbol{\Sigma}(\boldsymbol{\theta}^*, \boldsymbol{\xi}^*) = \lim_{t\to\infty} \mathbb{E}_{\boldsymbol{\theta}^*,\boldsymbol{\xi}^*}\left[(\epsilon_t^*)^2 \boldsymbol{w}_{t-1}\boldsymbol{w}_{t-1}^\top\right]$ *and* $\epsilon_t^* := \epsilon(z_t, \boldsymbol{\theta}^*)$ *denotes the TD error with the optimal parameter* $\boldsymbol{\theta}^*$.

Interestingly, the converse of Lemma 1 can also be shown; any martingale estimating functions for MRP take the form (13).

**Theorem 1.** *Any martingale estimating functions in the semiparametric model* $\{p(Z_T; \boldsymbol{\theta}, \boldsymbol{\xi}) \,|\, \boldsymbol{\theta}, \boldsymbol{\xi}\}$ *of MRP can be expressed as*

$$\boldsymbol{f}_T(Z_T, \boldsymbol{\theta}) = \sum_{t=1}^{T} \boldsymbol{\psi}_t(Z_t, \boldsymbol{\theta}) = \sum_{t=1}^{T} \boldsymbol{w}_{t-1}(Z_{t-1})\epsilon(z_t, \boldsymbol{\theta}). \tag{15}$$

*Proof.* Due to space limitation, we just sketch the proof here. From the martingale property, for any $t$, we have

$$\mathbb{E}_{\boldsymbol{\theta},\boldsymbol{\xi}_s}\left[\boldsymbol{f}_{t+1}(Z_{t+1}, \boldsymbol{\theta}) - \boldsymbol{f}_t(Z_t, \boldsymbol{\theta})|s_t\right] = \mathbf{0},$$

which should hold for any nuisance parameter $\boldsymbol{\xi}$. It can be shown that the TD error $\epsilon_{t+1}$ is the unique one that satisfies $\mathbb{E}_{\boldsymbol{\theta},\boldsymbol{\xi}}\left[\epsilon_{t+1}|s_t\right] = 0$ for any $s_t$ and $\boldsymbol{\xi}$. This implies $\boldsymbol{f}_{t+1}(Z_{t+1}, \boldsymbol{\theta}) - \boldsymbol{f}_t(Z_t, \boldsymbol{\theta}) = \boldsymbol{w}_t(Z_t)\epsilon(z_{t+1}, \boldsymbol{\theta})$. By induction, we see that $\boldsymbol{f}_T(Z_T, \boldsymbol{\theta})$ must have the form (15). □

### 4.2 Optimal Estimating Function

Since Theorem 1 has specified the set of all martingale estimating functions, we can now discuss the optimal estimating function among them which gives an M-estimator with *minimum asymptotic variance*. Because of the same reason as described in [1], it is suffice to consider the estimating function (15) with the weight $\boldsymbol{w}_t = \boldsymbol{w}_t(s_t)$ which depends only on the current state $s_t$. Furthermore, by the calculus of variations, we can obtain the optimal estimating function as stated by the following theorem.

**Theorem 2.** *When the random sequence $Z_T$ is generated from the distribution $p(Z_T; \boldsymbol{\theta}^*, \boldsymbol{\xi}^*)$, the optimal estimating function is given by*

$$\boldsymbol{f}_T^*(Z_T, \boldsymbol{\theta}) = \sum_{t=1}^{T} \boldsymbol{\psi}^*(z_t, \boldsymbol{\theta}) := \sum_{t=1}^{T} \boldsymbol{w}_{t-1}^*(s_{t-1})\epsilon(z_t, \boldsymbol{\theta}), \tag{16}$$

$$\text{where} \quad \boldsymbol{w}_t^*(s_t) := \mathbb{E}_{\boldsymbol{\theta}^*, \boldsymbol{\xi}_s^*}[\epsilon(z_{t+1}, \boldsymbol{\theta}^*)^2 | s_t]^{-1} \mathbb{E}_{\boldsymbol{\theta}^*, \boldsymbol{\xi}_s^*}[\partial_{\boldsymbol{\theta}} \epsilon(z_{t+1}, \boldsymbol{\theta}^*) | s_t].$$

Note that the optimal weighting function $\boldsymbol{w}_t^*$ depends on the true parameter $\boldsymbol{\theta}^*$ (but unknown) and needs the expectation with respect to $p(r_{t+1}, s_{t+1} | s_t; \boldsymbol{\theta}^*, \boldsymbol{\xi}_s^*)$, which is also unknown. Therefore, we need to substitute initial estimators for them as we will explain later. It is noted, however, that there is no need to estimate the nuisance parameter $\boldsymbol{\xi}$ itself and that consistency is always guaranteed, even if the initial estimators are based on rough approximation.

The minimum asymptotic variance can be obtained from Lemma 2 and Theorem 2.

**Corollary 1.** *The minimum asymptotic variance is given by*

$$\mathrm{Av}[\hat{\boldsymbol{\theta}}] = \frac{1}{T} \mathbf{Q}^{-1},$$

*where* $\mathbf{Q} = \lim_{t \to \infty} \mathbb{E}_{\boldsymbol{\theta}^*, \boldsymbol{\xi}^*} \left[ \partial_{\boldsymbol{\theta}} \boldsymbol{\psi}^*(z_t, \boldsymbol{\theta}^*) \right] = \lim_{t \to \infty} \mathbb{E}_{\boldsymbol{\theta}^*, \boldsymbol{\xi}^*} \left[ \boldsymbol{\psi}^*(z_t, \boldsymbol{\theta}^*) \boldsymbol{\psi}^*(z_t, \boldsymbol{\theta}^*)^\top \right]$.

We remark that the positive definite matrix $\mathbf{Q}$ measures information of the optimal estimating function. In general, the information associated with this matrix $\mathbf{Q}$ is smaller than Fisher information, since we trade efficiency for robustness against the nuisance parameter [7].

## 5 Learning Algorithms

This section describes the learning algorithm of the parameter $\boldsymbol{\theta}$. In reinforcement learning, online learning is often preferred to batch learning because of its computational efficiency and adaptability to even time-variant situations. Estimating functions provide not only batch algorithms via estimating equations, but also online ones as follows. An online estimator of $\boldsymbol{\theta}$ at time $t$ is denoted as $\hat{\boldsymbol{\theta}}_t$. Suppose that the sequence $\{\boldsymbol{\psi}_1(Z_1, \boldsymbol{\theta}), \ldots, \boldsymbol{\psi}_T(Z_T, \boldsymbol{\theta})\}$ forms a martingale estimating function for MRP. Then, an online update rule can be given by

$$\hat{\boldsymbol{\theta}}_t = \hat{\boldsymbol{\theta}}_{t-1} - \eta_t \boldsymbol{\psi}_t(Z_t, \hat{\boldsymbol{\theta}}_{t-1}), \tag{17}$$

where $\eta_t$ denotes a nonnegative scalar stepsize. In fact, there exist other online update rules derived from the same estimating function
$\boldsymbol{f}_t(Z_t, \boldsymbol{\theta}) = \sum_{i=1}^{t} \boldsymbol{\psi}_i(Z_i, \boldsymbol{\theta})$ as,

$$\hat{\boldsymbol{\theta}}_t = \hat{\boldsymbol{\theta}}_{t-1} - \eta_t \mathbf{R}(\hat{\boldsymbol{\theta}}_{t-1}) \boldsymbol{\psi}_t(Z_t, \hat{\boldsymbol{\theta}}_{t-1}), \tag{18}$$

where $\mathbf{R}(\boldsymbol{\theta})$ denotes an $m \times m$ nonsingular matrix depending only on $\boldsymbol{\theta}$ [15]. These variations come from the fact that $\mathbf{R}(\boldsymbol{\theta}) \sum_{i=1}^{t} \boldsymbol{\psi}_i(Z_i, \boldsymbol{\theta})$ gives the same roots as its original for any $\mathbf{R}(\boldsymbol{\theta})$. This equivalence guarantees that both learning procedures, (17) and (18), have the same stable point, while their dynamics may be different; that is, even if the plain algorithm (17) is unstable, it can be stabilized by introducing an appropriate $\mathbf{R}(\boldsymbol{\theta})$ as (18).

In the next two sections, we will discuss convergence of the online learning algorithm (18).

### 5.1 Convergence to the True Value

Here, we give sufficient conditions to guarantee the convergence of the online learning (18) to the true parameter $\boldsymbol{\theta}^*$. For the sake of simplicity, we focus on the final convergence phase: $\hat{\boldsymbol{\theta}}_t$ are confined in a neighborhood of $\boldsymbol{\theta}^*$. Now we introduce the following conditions for the convergence.

**Condition 1**
*(a) For any t,* $(\hat{\boldsymbol{\theta}}_t - \boldsymbol{\theta}^*)^\top \mathbf{R}(\hat{\boldsymbol{\theta}}_t) \mathbb{E}_{\boldsymbol{\theta}^*, \boldsymbol{\xi}_s^*} \left[ \boldsymbol{\psi}_{t+1}(Z_{t+1}, \hat{\boldsymbol{\theta}}_t) \middle| s_t \right]$ *is nonnegative .*
*(b) For any t, there exist such nonnegative constants* $c_1$ *and* $c_2$ *that* $\|\mathbf{R}(\hat{\boldsymbol{\theta}}_t) \mathbb{E}_{\boldsymbol{\theta}^*, \boldsymbol{\xi}_s^*}[\boldsymbol{\psi}_{t+1}(Z_{t+1}, \hat{\boldsymbol{\theta}}_t) | s_t]\|^2 \leq c_1 + c_2 \|\hat{\boldsymbol{\theta}}_t - \boldsymbol{\theta}^*\|$.

Then, the following theorem guarantees the (local) convergence of $\hat{\boldsymbol{\theta}}_t$ to $\boldsymbol{\theta}^*$.

**Theorem 3.** *Suppose that Condition 1 holds. If the stepsizes* $\{\eta_t\}$ *are all positive and satisfy* $\sum_{t=1}^{\infty} \eta_t = \infty$ *and* $\sum_{t=1}^{\infty} \eta_t^2 < \infty$*, then the online algorithm (18) converges to the true parameter* $\boldsymbol{\theta}^*$ *almost surely.*

*Proof.* The proof is given in Appendix A.

Theorem 3 ensures that an online algorithm of the form (18) is consistent, if we can find such a matrix $\mathbf{R}(\boldsymbol{\theta})$ that satisfies Condition 1.

### 5.2 Convergence Rate

In general, the convergence rate of an online algorithm is slow when compared to a batch algorithm that tries to obtain the solution of the estimating equation using all available samples. However, if we choose an appropriate matrix $\mathbf{R}(\boldsymbol{\theta})$ and adjust the stepsizes $\{\eta_t\}$ appropriately, then it is possible to achieve the same convergence rate with the batch algorithm [9]. First, we characterize the learning process of the batch algorithm.

**Lemma 3.** *Let* $\tilde{\boldsymbol{\theta}}_t$ *and* $\tilde{\boldsymbol{\theta}}_{t-1}$ *be solutions of the estimating equations* $1/t \sum_{i=1}^{t} \boldsymbol{\psi}_i(Z_i, \tilde{\boldsymbol{\theta}}_t) = \mathbf{0}$ *and* $1/(t-1) \sum_{i=1}^{t-1} \boldsymbol{\psi}_i(Z_i, \tilde{\boldsymbol{\theta}}_{t-1}) = \mathbf{0}$*, respectively. Then, we have*

$$\tilde{\boldsymbol{\theta}}_t = \tilde{\boldsymbol{\theta}}_{t-1} - \frac{1}{t}\hat{\mathbf{R}}_t^{-1}(\tilde{\boldsymbol{\theta}}_{t-1})\boldsymbol{\psi}_t(Z_t, \tilde{\boldsymbol{\theta}}_{t-1}) + \mathcal{O}\left(\frac{1}{t^2}\right), \tag{19}$$

*where* $\hat{\mathbf{R}}_t^{-1}(\tilde{\boldsymbol{\theta}}_{t-1}) = \{1/t\sum_{i=1}^{t} \partial_{\boldsymbol{\theta}}\boldsymbol{\psi}_i(Z_i, \tilde{\boldsymbol{\theta}}_{t-1})\}^{-1}$.

Note that (19) defines the sequence of $\tilde{\boldsymbol{\theta}}_t$ as a recursive stochastic process that is essentially same as the online learning (18) for the same $\mathbf{R}$. In other words, Lemma 3 implies that online algorithms can converge with the same convergence rate as batch counterparts by an appropriate choice of the matrix $\mathbf{R}$. Finally, the following theorem addresses the convergence rate of the (stochastic) learning process such as (19).

**Theorem 4.** *Consider the following learning process*

$$\hat{\boldsymbol{\theta}}_t = \hat{\boldsymbol{\theta}}_{t-1} - \frac{1}{t}\hat{\mathbf{R}}_t^{-1}\boldsymbol{\psi}_t(Z_t, \hat{\boldsymbol{\theta}}_{t-1}) + \mathcal{O}\left(\frac{1}{t^2}\right), \tag{20}$$

*where* $\hat{\mathbf{R}}_t = \{1/t\sum_{i=1}^{t} \partial_{\boldsymbol{\theta}}\boldsymbol{\psi}_i(Z_i, \hat{\boldsymbol{\theta}}_{i-1})\}$.
*Assume that:*

*(a)* $\hat{\mathbf{R}}_t^{-1}$ *can be written as* $\hat{\mathbf{R}}_t^{-1} = \mathbb{E}_{\boldsymbol{\theta}^*,\boldsymbol{\xi}_s^*}[\hat{\mathbf{R}}_t^{-1}|s_{t-1}] + o\left(t^{-1}\right)$.
*(b) For any* $t$, $\hat{\mathbf{R}}_t$ *is a nonsingular matrix.*

*If the learning process (20) converges to the true parameter almost surely, then the convergence rate is given as*

$$\mathbb{E}_{\boldsymbol{\theta}^*,\boldsymbol{\xi}^*}\left[\|\hat{\boldsymbol{\theta}}_t - \boldsymbol{\theta}^*\|^2\right] = \frac{1}{t}\mathrm{Tr}\left[\mathbf{A}^{-1}\boldsymbol{\Sigma}(\mathbf{A}^{-1})^\top\right] + o\left(\frac{1}{t}\right), \tag{21}$$

*where* $\mathbf{A} = \lim_{t\to\infty} \mathbb{E}_{\boldsymbol{\theta}^*,\boldsymbol{\xi}^*}[\boldsymbol{w}_{t-1}\{\partial_{\boldsymbol{\theta}}\epsilon(z_t, \boldsymbol{\theta}^*)\}^\top]$ *and*
$\boldsymbol{\Sigma} = \lim_{t\to\infty} \mathbb{E}_{\boldsymbol{\theta}^*,\boldsymbol{\xi}^*}\left[\epsilon(z_t, \boldsymbol{\theta}^*)^2\boldsymbol{w}_{t-1}\boldsymbol{w}_{t-1}^\top\right]$.

Theorem 4 applies to both the online and batch sequences. Note that this convergence rate (21) is neither affected by the third term of (20) nor by small variations on the matrix $\hat{\mathbf{R}}_t^{-1}$.

### 5.3 Implementation of Online Algorithm with Optimal Estimating Function

We now construct an optimal online learning which yields the minimum estimation error. Roughly speaking, this is given by the optimal estimating function in Theorem 2 with the best (i.e., with the fastest convergence) choice of the nonsingular matrix in Theorem 4;

$$\hat{\boldsymbol{\theta}}_t = \hat{\boldsymbol{\theta}}_{t-1} - \frac{1}{t}\hat{\mathbf{Q}}_t^{-1}\boldsymbol{\psi}^*(z_t, \hat{\boldsymbol{\theta}}_{t-1}), \tag{22}$$

where $\hat{\mathbf{Q}}_t^{-1} = \{1/t \sum_{i=1}^{t} \partial_{\boldsymbol{\theta}} \boldsymbol{\psi}^*(z_i, \hat{\boldsymbol{\theta}}_{i-1})\}^{-1}$ and $\boldsymbol{\psi}^*(z_t, \boldsymbol{\theta})$ is defined by eq. (16). If the learning equation (22) satisfies Condition 1 and Theorem 4, then it converges to the true parameter with the minimum estimation error, $(1/t)\mathbf{Q}^{-1}$.

However, the learning rule (22) still contains unknown parameters and quantities, so is impractical. For practical implementation, it is necessary to evaluate $\mathbb{E}_{\boldsymbol{\theta}^*, \boldsymbol{\xi}_s^*}[\epsilon(z_{t+1}, \boldsymbol{\theta}^*)^2 | s_t]$ and $\mathbb{E}_{\boldsymbol{\theta}^*, \boldsymbol{\xi}_s^*}[\partial_{\boldsymbol{\theta}} \epsilon(z_{t+1}, \boldsymbol{\theta}^*) | s_t]$ appearing in the optimal estimating function. Therefore, we apply the online function approximation for them. Let $\zeta(s_t, \boldsymbol{\alpha}_t)$ and $\boldsymbol{\varphi}(s_t, \boldsymbol{\beta}_t)$ be the approximations of $\mathbb{E}_{\boldsymbol{\theta}^*, \boldsymbol{\xi}_s^*}[\epsilon(z_{t+1}, \boldsymbol{\theta}_t)^2 | s_t]$ and $\mathbb{E}_{\boldsymbol{\theta}^*, \boldsymbol{\xi}_s^*}[\partial_{\boldsymbol{\theta}} \epsilon(z_{t+1}, \boldsymbol{\theta}_t) | s_t]$, respectively:

$$\begin{aligned} \zeta(s_t, \boldsymbol{\alpha}_t) &\approx \mathbb{E}_{\boldsymbol{\theta}^*, \boldsymbol{\xi}_s^*}[\epsilon(z_{t+1}, \hat{\boldsymbol{\theta}}_t)^2 | s_t] \\ \boldsymbol{\varphi}(s_t, \boldsymbol{\beta}_t) &\approx \mathbb{E}_{\boldsymbol{\theta}^*, \boldsymbol{\xi}_s^*}[\partial_{\boldsymbol{\theta}} \epsilon(z_{t+1}, \hat{\boldsymbol{\theta}}_t) | s_t], \end{aligned}$$

where $\boldsymbol{\alpha}_t$ and $\boldsymbol{\beta}_t$ are adjustable parameters. $\boldsymbol{\alpha}_t$ and $\boldsymbol{\beta}_t$ are adjusted in an online manner;

$$\begin{aligned} \hat{\boldsymbol{\alpha}}_t &= \hat{\boldsymbol{\alpha}}_{t-1} - \eta_t^{\boldsymbol{\alpha}} \partial_{\boldsymbol{\alpha}} \zeta(s_{t-1}, \hat{\boldsymbol{\alpha}}_{t-1}) \left( \zeta(s_{t-1}, \hat{\boldsymbol{\alpha}}_{t-1}) - \epsilon(z_t, \hat{\boldsymbol{\theta}}_{t-1})^2 \right) \\ \hat{\boldsymbol{\beta}}_t &= \hat{\boldsymbol{\beta}}_{t-1} - \eta_t^{\boldsymbol{\beta}} \partial_{\boldsymbol{\beta}} \boldsymbol{\varphi}(s_{t-1}, \hat{\boldsymbol{\beta}}_{t-1}) \left( \boldsymbol{\varphi}(s_{t-1}, \hat{\boldsymbol{\beta}}_{t-1}) - \partial_{\boldsymbol{\theta}} \epsilon(z_t, \hat{\boldsymbol{\theta}}_{t-1}) \right), \end{aligned}$$

where $\eta_t^{\boldsymbol{\alpha}}$ and $\eta_t^{\boldsymbol{\beta}}$ are stepsizes. By using these parameterized functions, we can replace $\boldsymbol{\psi}_t^*(z_t, \hat{\boldsymbol{\theta}}_{t-1})$ and $\hat{\mathbf{Q}}_t^{-1}$ by

$$\begin{aligned} \boldsymbol{\psi}_t^*(z_t, \hat{\boldsymbol{\theta}}_{t-1}) &= \zeta(s_{t-1}, \hat{\boldsymbol{\alpha}}_{t-1})^{-1} \boldsymbol{\varphi}(s_{t-1}, \hat{\boldsymbol{\beta}}_{t-1}) \epsilon(z_t, \hat{\boldsymbol{\theta}}_t) \\ \hat{\mathbf{Q}}_t^{-1} &= \left( \frac{1}{t} \sum_{i=1}^{t} \zeta(s_{i-1}, \hat{\boldsymbol{\alpha}}_{i-1})^{-1} \boldsymbol{\varphi}(s_{i-1}, \hat{\boldsymbol{\beta}}_{i-1}) \partial_{\boldsymbol{\theta}} \epsilon(z_i, \hat{\boldsymbol{\theta}}_{i-1})^\top \right)^{-1}. \end{aligned} \quad (23)$$

Note that the update (23) can be done in an online manner by applying the well-known matrix inversion lemma [16]. We summarize our implementation of the optimal online learning algorithm in Algorithm 1. The empirical results of this algorithm will be shown in Section 6.

### 5.4 Acceleration of TD Learning

TD learning is a traditional online approach to model-free policy evaluation and has been as one of the most important algorithms in reinforcement learning. Although the TD learning is widely used due to its simplicity, it is known to converge rather slowly. In this section, we discuss the TD learning from the viewpoint of the estimating function method and propose a new online algorithm which can achieve faster convergence than the usual TD learning.

To simplify the following discussions, let $g(s, \boldsymbol{\theta})$ be a linear function of features:

$$V(s_t) := \boldsymbol{\phi}(s_t)^\top \boldsymbol{\theta} := \boldsymbol{\phi}_t^\top \boldsymbol{\theta},$$

**Algorithm 1.** The proposed online learning algorithm

Initialize $\hat{\boldsymbol{\alpha}}_0$, $\hat{\boldsymbol{\beta}}_0$, $\hat{\boldsymbol{\theta}}_0$, $\hat{\mathbf{Q}}_0^{-1} = \epsilon\mathbf{I}$, $a_1$, $a_2$
{$\epsilon$ and $\mathbf{I}$ denote a small constant and an $m \times m$ identical matrix, respectively. }

**for** $t = 1, 2, \cdots$ **do**
Obtain a new sample $z_t = \{s_{t-1}, s_t, r_t\}$

Compute the optimal weight function $\boldsymbol{w}_{t-1}^*$
$\hat{\boldsymbol{\alpha}}_t \leftarrow \hat{\boldsymbol{\alpha}}_{t-1} - \eta_t^{\boldsymbol{\alpha}} \partial_{\boldsymbol{\alpha}} \zeta(s_{t-1}, \hat{\boldsymbol{\alpha}}_{t-1}) \{\zeta(s_{t-1}, \hat{\boldsymbol{\alpha}}_{t-1}) - \epsilon(z_t, \hat{\boldsymbol{\theta}}_{t-1})^2\}$
$\hat{\boldsymbol{\beta}}_t \leftarrow \hat{\boldsymbol{\beta}}_{t-1} - \eta_t^{\boldsymbol{\beta}} \partial_{\boldsymbol{\beta}} \boldsymbol{\varphi}(s_{t-1}, \hat{\boldsymbol{\beta}}_{t-1}) \left(\boldsymbol{\varphi}(s_{t-1}, \hat{\boldsymbol{\beta}}_{t-1}) - \partial_{\boldsymbol{\theta}} \epsilon(z_t, \hat{\boldsymbol{\theta}}_{t-1})\right)$
$\boldsymbol{w}_{t-1}^* \leftarrow \zeta(s_{t-1}, \hat{\boldsymbol{\alpha}}_{t-1})^{-1} \boldsymbol{\varphi}(s_{t-1}, \hat{\boldsymbol{\beta}}_{t-1})$

Update $\hat{\mathbf{Q}}_t^{-1}$ using matrix inversion lemma
$\hat{\mathbf{Q}}_t^{-1} \leftarrow \frac{1}{t-1}\hat{\mathbf{Q}}_{t-1}^{-1} - \frac{1}{t} \frac{\hat{\mathbf{Q}}_{t-1}^{-1} \boldsymbol{w}_{t-1}^* \partial_{\boldsymbol{\theta}} \epsilon(z_t, \hat{\boldsymbol{\theta}}_{t-1})^\top \hat{\mathbf{Q}}_{t-1}^{-1}}{1 + \partial_{\boldsymbol{\theta}} \epsilon(z_t, \hat{\boldsymbol{\theta}}_{t-1})^\top \hat{\mathbf{Q}}_{t-1}^{-1} \boldsymbol{w}_{t-1}^*}$

Update the parameter
$\tau \leftarrow \max(a_1, t - a_2)$
$\hat{\boldsymbol{\theta}}_t \leftarrow \hat{\boldsymbol{\theta}}_{t-1} - \frac{1}{\tau}\hat{\mathbf{Q}}_t^{-1} \boldsymbol{w}_{t-1}^* \epsilon(z_t, \hat{\boldsymbol{\theta}}_{t-1})$
**end for**

where $\boldsymbol{\phi}(s) : S \to \mathbb{R}^m$ is a feature vector and $\boldsymbol{\theta} \in \mathbb{R}^m$ is a parameter vector. In this case, we have two ways to solve the linear estimating equation; one is a batch procedure:

$$\hat{\boldsymbol{\theta}} = \{\sum_{t=1}^{T} \boldsymbol{w}_{t-1}(\boldsymbol{\phi}_{t-1} - \gamma\boldsymbol{\phi}_t)^\top\}^{-1} \{\sum_{t=1}^{T} \boldsymbol{w}_{t-1} r_t\}$$

and the other is an online procedure:

$$\hat{\boldsymbol{\theta}}_t = \hat{\boldsymbol{\theta}}_{t-1} - \eta_t \boldsymbol{w}_{t-1} \epsilon(z_t, \hat{\boldsymbol{\theta}}_{t-1}).$$

When the weight function $\boldsymbol{w}_t$ is set to $\boldsymbol{\phi}_t$, the online procedure and batch procedure correspond to the TD learning and LSTD algorithm, respectively. Note that both TD and LSTD share the same estimating function. Therefore, from Lemma 3 and Theorem 4, we can in principle construct an accelerated TD learning which converges at the same speed as the LSTD algorithm.

Here, we consider the following learning equation;

$$\hat{\boldsymbol{\theta}}_t = \hat{\boldsymbol{\theta}}_{t-1} - \frac{1}{t}\hat{\mathbf{R}}_t^{-1} \boldsymbol{\phi}_{t-1} \epsilon(z_t, \hat{\boldsymbol{\theta}}_{t-1}), \tag{24}$$

where $\hat{\mathbf{R}}_t^{-1} = \{1/t \sum_{i=1}^{t} \boldsymbol{\phi}_{i-1}(\boldsymbol{\phi}_{i-1} - \gamma\boldsymbol{\phi}_i)^\top\}^{-1}$. Since $\hat{\mathbf{R}}_t^{-1}$ converges to $\mathbf{A}^{-1} = \lim_{t\to\infty} \mathbb{E}_{\boldsymbol{\theta}^*, \boldsymbol{\xi}^*}[\boldsymbol{\phi}_{t-1}(\boldsymbol{\phi}_{t-1} - \gamma\boldsymbol{\phi}_t)^\top]^{-1}$ and $\mathbf{A}^{-1}$ must be a positive definite matrix (see Lemma 6.4 in [10]), the online algorithm (24) also converges to the true parameter almost surely. Then, if $\hat{\mathbf{R}}_t$ satisfies the condition in Theorem 4, it can achieve same convergence rate as LSTD. We call this procedure the *accelerated-TD learning*.

In both the optimal online learning and the accelerated-TD learning, it is necessary to maintain the inverse of the scaling matrix $\hat{\mathbf{R}}_t$. Since this matrix inversion operation costs $\mathcal{O}(m^2)$ in each step, maintaining the inverse matrix becomes expensive when the dimensionality of parameters increases. The computational cost can be dramatically reduced by maintaining a coarse approximation of the scaling matrix (e.g. diagonal, block diagonal, etc.). An appropriate setting ensures the convergence rate remains $\mathcal{O}(1/t)$ without spoiling computational efficiency.

## 6 Simulation Experiments

In order to validate our theoretical developments, we compared the performance (statistical error) of the proposed online algorithms (accelerated-TD algorithm and the optimal online learning algorithm) with those of the baselines: TD algorithm [2] (online), LSTD algorithm [3] (batch), and gLSTD algorithm [1] (batch) in a toy problem. An MRP trajectory was generated from a simple Markov random walk on a chain with ten states ($s = 1, \cdots, 10$) as depicted in Fig. 2. At each time $t$, the state changes to either of its left ($-1$) or right ($+1$) with equal probability of 0.5. A reward was given by the deterministic function $r = \exp(-0.5(s-5)^2/3^2)$, and the discount factor was set to 0.95. The value function was approximated by a linear function with three-dimensional basis functions, that is, $V(s) \approx \sum_{n=1}^{3} \theta_n \phi_n(s)$. The basis functions $\phi_n(s)$ were generated according to a diffusion model [17]. This approximation was not faithful; i.e. there remained tiny bias.

We generated $M = 200$ trajectories (episodes) each of which consisted of $T = 200$ random walk steps. The value function was estimated for each episode. We evaluated the "mean squared error" (MSE) of the value function, that is, $\frac{1}{M}\frac{1}{10}\sum_{k=1}^{M}\sum_{i\in\{1,\cdots,10\}} \|\boldsymbol{\phi}_i^\top \hat{\boldsymbol{\theta}}_k - V^*(i)\|^2$ where $V^*$ denotes the true value function.

As is often done in online procedures, we utilized some batch procedures to obtain initial estimates of the parameter. More specifically, the first 20 steps in each episode were used to obtain an initial estimator in a batch manner and the online algorithm started after 20 steps. In this random walk problem, owing to the linear approximation, the parameter by the batch algorithm can be obtained analytically. In general situations, on the other hand, an online algorithm has

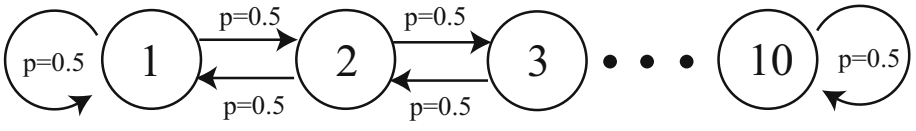


**Fig. 2.** A ten-states MRP

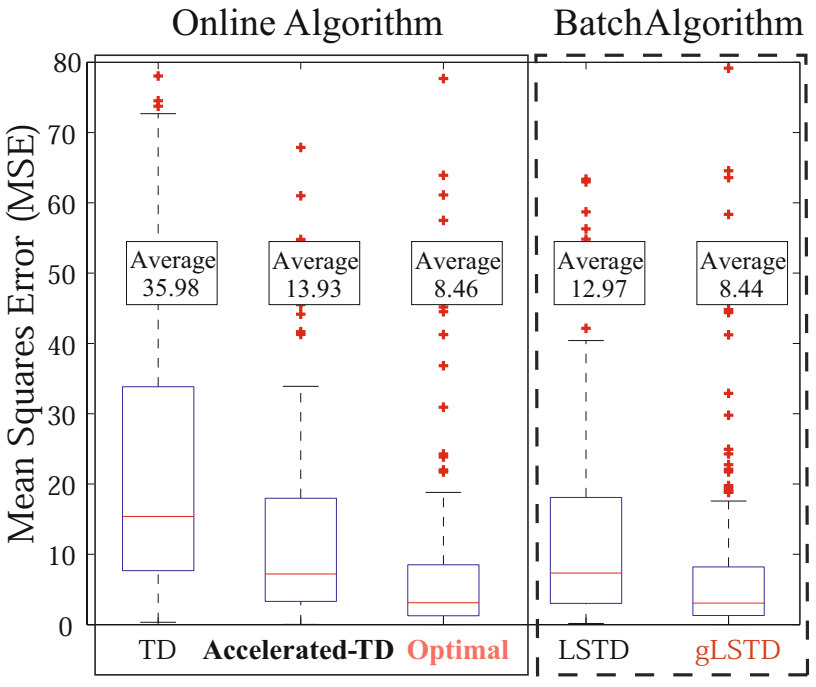


**Fig. 3.** Simulation results

a merit, because online procedures require less memory and are computationally more efficient. They perform only a single update at each time, while the batch algorithms must keep all trajectories and need to iterate computation until convergence which is serious when employing nonlinear estimating equations.

In the proposed online algorithms, the stepsizes were decreased as simple as $1/t$. On the other hand, the convergence of TD learning was too slow in simple $1/t$ setting due to fast decay of the stepsizes but also in certain well-chosen constant stepsize. Therefore, we adopt an ad-hoc adjustment for the stepsizes as $1/\tau$, where $\tau = \max(10, t - 100)$.

Fig. 3 shows the MSEs of the value functions estimated by our proposed algorithms and the existing algorithms, in which the MSEs of all 200 episodes are shown by box-plots; the center line, and the upper and lower sides of each box denote the median of MSE, and the upper and lower quartiles, respectively. The number above each box is the average MSE. As is shown in Fig. 3, the optimal online learning algorithm (Optimal) and the optimal batch learning algorithm (gLSTD) achieve the minimum MSE among the online and batch algorithms, respectively, and these two MSEs are very close. It should be noted that the accelerated-TD algorithm (accelerated-TD) performs significantly better than the ordinary TD algorithm showing the matrix $\mathbf{R}$ was effective for accelerating the convergence as expected by our theoretical analysis.

## 7 Conclusion

In this study, we extended the framework of semiparametric statistics inference for value function estimation to be applicable to online learning procedures. Based on this extension, we derived the general form of estimating functions for the model-free value function estimation in MRPs, which provides the statistical basis to many existing batch and online learning algorithms. Moreover, we found the optimal estimating function, which yields the minimum asymptotic estimation variance amongst the general class, and presented a new online learning algorithm (optimal algorithm) based on it. Using a simple MRP problem, we confirmed the validity of our analysis, that is, the optimal algorithm achieves the minimum MSE of the value function estimation and converges with almost the same speed with the batch algorithm gLSTD.

Throughout this article, we assumed that the function approximation is faithful, that is, there is no model misspecification for the value function, and analyzed only its asymptotic variance. Even in misspecified cases, the asymptotic variance can be correctly evaluated [14]. Therefore, if we can reduce the bias term to be much smaller than the variance, our optimal and accelerated-TD procedures could also improve significantly the existing algorithms. Moreover, it is an important future issue to find out a good parametric function $g$ or set of basis functions $\phi_n$ for linearly approximating the value function.

## References

1. Ueno, T., Kawanabe, M., Mori, T., Maeda, S., Ishii, S.: A semiparametric statistical apparoach to model-free policy evaluation. In: Proceedings of the 25th International Conference on Machine Learning, pp. 1072–1079 (2008)
2. Sutton, R., Barto, A.: Reinforcement Learning: An Introduction. MIT Press, Cambridge (1998)
3. Bradtke, S., Barto, A.: Linear least-squares algorithms for temporal difference learning. Machine Learning 22(1), 33–57 (1996)
4. Boyan, J.: Technical update: Least-squares temporal difference learning. Machine Learning 49(2), 233–246 (2002)
5. Godambe, V. (ed.): Estimating Functions. Oxford Science, Oxford (1991)
6. Bickel, D., Ritov, D., Klaassen, C., Wellner, J.: Efficient and Adaptive Estimation for Semiparametric Models. Springer, Heidelberg (1998)
7. Amari, S., Kawanabe, M.: Information geometry of estimating functions in semiparametric statistical models. Bernoulli 3(1), 29–54 (1997)
8. van der Vaart, A.: Asymptotic Statistics. Cambridge University Press, Cambridge (1998)
9. Bottou, L., LeCun, Y.: On-line learning for very large datasets. Applied Stochastic Models in Business and Industry 21(2), 137–151 (2005)
10. Bertsekas, D., Tsitsiklis, J.: Neuro-Dynamic Programming. Athena Scientific, Belmont (1996)
11. Nedić, A., Bertsekas, D.: Least squares policy evaluation algorithms with linear function approximation. Discrete Event Dynamic Systems 13(1), 79–110 (2003)

12. Mannor, S., Simester, D., Sun, P., Tsitsiklis, J.: Bias and variance in value function estimation. In: Proceedings of the twenty-first international conference on Machine learning. ACM, New York (2004)
13. Godambe, V.: The foundations of finite sample estimation in stochastic processes. Biometrika 72(2), 419–428 (1985)
14. Sørensen, M.: On asymptotics of estimating functions. Brazilian Journal of Probability and Statistics 13(2), 419–428 (1999)
15. Amari, S.: Natural gradient works efficiently in learning. Neural Computation 10(2), 251–276 (1998)
16. Horn, R., Johnson, C.: Matrix analysis. Cambridge University Press, Cambridge (1985)
17. Mahadevan, S., Maggioni, M.: Proto-value Functions: A Laplacian Framework for Learning Representation and Control in Markov Decision Processes. The Journal of Machine Learning Research 8, 2169–2231 (2007)

## A Proof: Theorem 3

To simplify the following proof, we assume the true parameter is located on the origin without loss of generality: $\boldsymbol{\theta}^* = \mathbf{0}$. Let $h_t$ be $\|\hat{\boldsymbol{\theta}}_t\|^2$. The conditional expectation of variation of $h_t$ can be derived as

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\theta}^*,\boldsymbol{\xi}_s^*}\left[h_{t+1} - h_t | s_t\right] = &-2\eta_{t+1}\hat{\boldsymbol{\theta}}_t^\top \mathbf{R}(\hat{\boldsymbol{\theta}}_t)\mathbb{E}_{\boldsymbol{\theta}^*,\boldsymbol{\xi}_s^*}\left[\boldsymbol{\psi}_{t+1}(Z_{t+1},\hat{\boldsymbol{\theta}}_t)|s_t\right] \\ &+ \eta_{t+1}^2 \mathbb{E}_{\boldsymbol{\theta}^*,\boldsymbol{\xi}_s^*}\left[\|\mathbf{R}(\hat{\boldsymbol{\theta}}_t)\boldsymbol{\psi}_{t+1}(Z_{t+1},\hat{\boldsymbol{\theta}}_t)\|^2 | s_t\right].\end{aligned}$$

From Condition 1, the second term of this equation is bounded by the second moment, thus we obtain

$$\begin{aligned}&\mathbb{E}_{\boldsymbol{\theta}^*,\boldsymbol{\xi}_s^*}\left[h_{t+1} - (1+\eta_{t+1}^2 c_2)h_t | s_t\right] \\ &\leq -2\eta_{t+1}\hat{\boldsymbol{\theta}}_t^\top \mathbf{R}(\hat{\boldsymbol{\theta}}_t)\mathbb{E}_{\boldsymbol{\theta}^*,\boldsymbol{\xi}_s^*}\left[\boldsymbol{\psi}_{t+1}(Z_{t+1},\hat{\boldsymbol{\theta}}_t)|s_t\right] + \eta_{t+1}^2 c_1. \qquad (25)\end{aligned}$$

Now, let $\chi_t = \prod_{k=0}^{t-1} \frac{1}{1+\eta_{k+1}^2 c_2}$ and $h'_t = \chi_t h_t$. From the assumption $\sum_{t=1}^{\infty} \eta_t^2 < \infty$, we easily verify that $0 < \chi_t < 1$. Multiplying both sides of eq. (25) by $\chi_{t+1}$, we obtain

$$\begin{aligned}&\mathbb{E}_{\boldsymbol{\theta}^*,\boldsymbol{\xi}^*}\left[h'_{t+1} - h'_t | \mathcal{P}_t\right] \\ &\leq -2\eta_{t+1}\chi_{t+1}\hat{\boldsymbol{\theta}}_t^\top \mathbf{R}(\hat{\boldsymbol{\theta}}_t)\mathbb{E}_{\boldsymbol{\theta}^*,\boldsymbol{\xi}_s^*}\left[\boldsymbol{\psi}_{t+1}(Z_{t+1},\hat{\boldsymbol{\theta}}_t)|s_t\right] + \eta_{t+1}^2\chi_{t+1}c_1.\end{aligned}$$

The first term of this upper bound is negative because of Condition 1, and the second term is nonnegative because $\eta_t$, $\chi_{t+1}$, and $c_1$ are nonnegative, and the sum of the second terms $\sum_{t=1}^{\infty} \eta_t^2 \chi_{t+1} c_1$ is finite. Then, the supermartingale convergence theorem [10] guarantees that $h'_t$ converges to a nonnegative random variable almost surely, and $\sum_{t=1}^{\infty} \eta_{t+1}\chi_{t+1}\hat{\boldsymbol{\theta}}_t^\top \mathbf{R}_t(\hat{\boldsymbol{\theta}}_t)\mathbb{E}_{\boldsymbol{\theta}^*,\boldsymbol{\xi}_s^*}\left[\boldsymbol{\psi}_{t+1}(Z_{t+1},\hat{\boldsymbol{\theta}}_t)|s_t\right] < \infty$. Since $\sum_{t=1}^{\infty}\eta_t = \infty$ and $\lim_{t\to\infty}\chi_t = \chi_\infty > 0$, we have $\hat{\boldsymbol{\theta}}_t^\top \mathbf{R}(\hat{\boldsymbol{\theta}}_t)\mathbb{E}_{\boldsymbol{\theta}^*,\boldsymbol{\xi}_s^*}\left[\boldsymbol{\psi}_{t+1}(Z_{t+1},\hat{\boldsymbol{\theta}}_t)|s_t\right] \xrightarrow{a.s.} \mathbf{0}$, where $\xrightarrow{a.s.}$ denotes the almost sure convergence. This result suggests the conclusion that the online learning algorithm converges almost surely: $\hat{\boldsymbol{\theta}}_t \xrightarrow{a.s.} \boldsymbol{\theta}^* = \mathbf{0}$.