Submitted 10/09; Revised 5/09; Published 6/11

Generalized TD Learning

Tsuyoshi Ueno TSUYOS-U@SYS.I.KYOTO-U.AC.JP
Shin-ichi Maeda ICHI@SYS.I.KYOTO-U.AC.JP
Graduate School of Informatics
Kyoto University
Gokasho, Uji, Kyoto, 611-0011 Japan

Motoaki Kawanabe[*] MOTOAKI.KAWANABE@FIRST.FRAUNHOFER.DE
Fraunhofer FIRST.IDA
Kekulestrasee 7
12489, Berlin, Germany

Shin Ishii ISHII@I.KYOTO-U.AC.JP
Graduate School of Informatics
Kyoto University
Gokasho, Uji, Kyoto, 611-0011 Japan

Editor: Shie Mannor

Abstract

Since the invention of temporal difference (TD) learning (Sutton, 1988), many new algorithms for model-free policy evaluation have been proposed. Although they have brought much progress in practical applications of reinforcement learning (RL), there still remain fundamental problems concerning statistical properties of the value function estimation. To solve these problems, we introduce a new framework, *semiparametric statistical inference*, to model-free policy evaluation. This framework generalizes TD learning and its extensions, and allows us to investigate statistical properties of both of batch and online learning procedures for the value function estimation in a unified way in terms of *estimating functions*. Furthermore, based on this framework, we derive an optimal estimating function with the *minimum asymptotic variance* and propose batch and online learning algorithms which achieve the optimality.

Keywords: reinforcement learning, model-free policy evaluation, TD learning, semiparametirc model, estimating function

1. Introduction

Studies in reinforcement learning (RL) have provided a methodology for optimal control and decision making in various practical applications, for example, job scheduling (Zhang and Dietterich, 1995), backgammon (Tesauro, 1995), elevator dispatching (Crites and Barto, 1996), and dynamic channel allocation (Singh and Bertsekas, 1997). Although the tasks in these studies are large-scale and complicated, RL has achieved good performance which exceeds that of human experts. These successes were attributed to model-free policy evaluation, that is, the value function which evaluates the expected cumulative reward is estimated from a given sample trajectory without specifying the task environment. Since the policy is updated based on the estimated value function, the quality

[*] Also at Berlin Institute of Technology, Computer Science, Machine Learning Group, Franklinstr. 28/29, 10587, Berlin, Germany.

of its estimation directly affects policy improvement. Hence, it is important for research in RL to develop efficient model-free policy evaluation techniques.

This article introduces a novel framework, *semiparametric statistical inference*, to model-free policy evaluation. This framework generalizes previously developed model-free algorithms, which include temporal difference learning and its extensions, and moreover, enables us to investigate the statistical properties of these algorithms, which have not been yet elucidated.

The overall framework can be summarized as follows. We focus on the policy evaluation like in previous studies (Singh and Dayan, 1998; Mannor et al., 2004; Grunëwälder and Obermayer, 2006; Mannor et al., 2007); then we deal with the Markov Reward Process (MRP), in which the initial, transition, and the reward probabilities are assumed to be unknown. From a sample trajectory given by MRP, the value function is estimated without directly identifying those probabilities. Central to our proposed framework is the notion of *semiparametric statistical models* which include not only parameters of interest but also additional nuisance parameters with possibly infinite degrees of freedom. We specify the MRP as a semiparametric model, where only the value function is modeled parametrically with a smaller number of parameters than necessary, while the other unspecified part of MRP corresponds to the nuisance parameters. 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 (Godambe, 1960, 1991; Huber and Ronchetti, 2009; van der Vaart, 2000). In this sense, the semiparametric inference is a promising approach to model-free policy evaluation.

Our contributions are summarized as follows:

(a) A set of all estimating functions is shown explicitly: the set constitutes a general class of consistent estimators (Theorem 4). Furthermore, by applying the asymptotic analysis, we derive the asymptotic estimation variance of general estimating functions (Lemma 3) and the optimal estimating function that yields the *minimum asymptotic variance* of estimation (Theorem 6).

(b) We discuss two types of learning algorithms based on estimating functions. One is the class of batch algorithms which obtain estimators in one shot by using all samples in the given trajectory such as least squares temporal difference (LSTD) learning (Bradtke and Barto, 1996). The other is the class of online algorithms which update the estimators step-by-step such as temporal difference (TD) learning (Sutton, 1988). In the batch algorithm, we assume that the value function is represented as a parametrically linear function and derive a new least squares-type algorithm, *gLSTD* learning, which achieves the minimum asymptotic variance (Algorithm 1).

(c) Following previous work (Amari, 1998; Murata and Amari, 1999; Bottou and LeCun, 2004, 2005), we examine the convergence of statistical deviations of the online algorithms. We then show that the online algorithms can achieve the same asymptotic performance as their batch counterparts if the parameters controlling learning processes are appropriately tuned (Lemma 9 and Theorem 10). We derive the optimal choice of the estimating function and construct the online learning algorithm that achieves the minimum estimation error asymptotically (Algorithm 2). We also propose an acceleration of TD learning, which is called *accelerated TD learning* (Algorithm 3).

(d) We then show that our proposed framework generalizes almost all of the conventional model-free policy evaluation algorithms, such as TD learning, TD(λ) learning (Sutton, 1988; Sutton

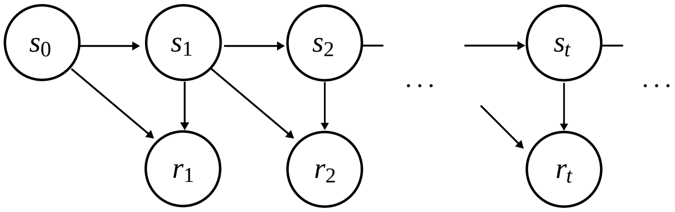

Figure 1: Graphical model for infinite horizon MRP. s and r denote state variable and reward, respectively.

and Barto, 1998), Bellman residual (RG) learning (Baird, 1995), LSTD learning (Bradtke and Barto, 1996), LSTD(λ) learning (Boyan, 2002), least squares policy evaluation (LSPE) learning (Nedić and Bertsekas, 2003), and incremental LSTD (iLSTD) learning (Geramifard et al., 2006, 2007) (Table 1).

We compare the performance of the proposed online algorithms with a couple of well-established algorithms in simple numerical experiments and show that the results support our theoretical findings.

The rest of this article is organized as follows. First, we give background of MRP and define the semiparametric statistical model for estimating the value function (Section 2). After providing a short overview of estimating functions (Section 3), we present the main contribution, fundamental statistical analysis based on the estimating function theory (Section 4). Then, we explain the construction of practical learning algorithms, derived from estimating functions, as both batch and online algorithms (Section 5). Furthermore, relations of our proposed methods to current algorithms in RL are discussed (Section 6). Finally, we report our experimental results (Section 7), and discuss open questions and future direction of this study (Section 8).

2. Markov Reward Processes

Figure 1 shows a graphical model for an infinite horizon MRP[1] which is defined by the initial state probability $p(s_0)$, the state transition probability $p(s_t|s_{t-1})$ and the reward probability $p(r_t|s_t,s_{t-1})$. State variable s is an element of a finite set S and reward variable $r \in R$ can be either discrete or continuous. The joint distribution of a sample trajectory $Z_T \equiv \{s_0, s_1, r_1 \cdots, s_T, r_T\}$ of the MRP is described as

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

We further impose the following assumptions on MRPs.

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

Assumption 2 *For any time t, reward r_t is uniformly bounded.*

[1] In this study, we only consider MRPs; however, extension to Markov Decision Processes (MDPs) is straightforward as long as considering the policy evaluation problem (hence the policy is fixed).

Before introducing the statistical framework, we begin by confirming that the value function estimation can be interpreted as estimation of certain statistics of MRP (1).

Proposition 1 *(Bertsekas and Tsitsiklis, 1996) Consider a conditional probability of* $\{r_t, s_t\}$ *given* s_{t-1}*,*

$$p(r_t, s_t|s_{t-1}) = p(r_t|s_t, s_{t-1})p(s_t|s_{t-1}).$$

Then, there is such a function V *that*

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

holds for any state $s_{t-1} \in S$*, where* $\gamma \in [0,1)$ *is a constant called discount factor. Here,* $\mathbb{E}\left[\cdot|s\right]$ *denotes the conditional expectation for the given state* s*. The function* V *that satisfies Equation (2) is unique and found to be the value function:*

$$V(s) \equiv \lim_{T\to\infty} \mathbb{E}\left[\sum_{t=1}^{T} \gamma^{t-1} r_t \middle| s_0 = s\right]. \tag{3}$$

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

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

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

Here, $g : S \times \Theta \mapsto \mathbb{R}$ *and* $\boldsymbol{\theta} \in \Theta$ *is a certain parameter in a parameter space* $\Theta \subseteq \mathbb{R}^m$*. Also, the dimension of the parameter* $\boldsymbol{\theta}$ *is smaller than that of the state space:* $m < |S|$*. Moreover,* $g(s, \boldsymbol{\theta})$ *is assumed to be twice continuously differentiable with respect to* $\boldsymbol{\theta}$*.*

Under Assumption 3, $p(r_t|s_{t-1})$ is partially parametrized by $\boldsymbol{\theta}$, through its conditional mean

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

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

To specify the probabilistic model (1) altogether, we usually need extra parameters other than $\boldsymbol{\theta}$. Let $\boldsymbol{\xi}_0$ and $\boldsymbol{\xi}_s$ be such additional parameters that $p(s_0, \boldsymbol{\xi}_0)$ and $p(r, s|s; \boldsymbol{\theta}, \boldsymbol{\xi}_s)$ can completely represent the initial and transition distributions, respectively. In such a case, the joint distribution of the trajectory Z_T is expressed as

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

where $\boldsymbol{\xi} \equiv (\boldsymbol{\xi}_0, \boldsymbol{\xi}_s)$.

Since it is in general quite difficult to know the complexity of the target system, we attempt to estimate the parameter $\boldsymbol{\theta}$ representing the value function beside the presence of 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 called semiparametric (Bickel et al., 1998; Amari and Kawanabe, 1997; van der Vaart, 2000). We emphasize that the nuisance parameters are necessary only for developing theoretical frameworks. In actual estimation procedures of the parameter $\boldsymbol{\theta}$, same as in other model-free policy evaluation algorithms, we neither define them concretely, nor estimate them. This can be achieved by using estimating functions which is a well-established technique to obtain a consistent estimator of the parameter without estimating the nuisance parameters (Godambe, 1960, 1991; Amari and Kawanabe, 1997; Huber and Ronchetti, 2009). The advantages of considering such semiparametric models behind the model-free policy evaluation 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 *minimum estimation error*.

We review 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 independent and identically distributed (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 $\boldsymbol{f}$ of $\boldsymbol{x}$ and $\boldsymbol{\theta}$ is called an *estimating function* (Godambe, 1960, 1991) when it satisfies the following conditions for any $\boldsymbol{\theta}$ and $\boldsymbol{\xi}$ for sufficiently large values of T;

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

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

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

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 respect to the distribution $p(\boldsymbol{x};\boldsymbol{\theta},\boldsymbol{\xi})$ 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 $\{\boldsymbol{x}_1,\cdots,\boldsymbol{x}_T\}$ 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 $\hat{\boldsymbol{\theta}}_T$ which has good asymptotic properties, by solving the following estimating equation:

$$\sum_{t=1}^{T}\boldsymbol{f}(\boldsymbol{x}_t,\hat{\boldsymbol{\theta}}_T) = \mathbf{0}. \tag{9}$$

A solution of the estimating Equation (9) is called an *M-estimator* in statistics (Huber and Ronchetti, 2009; van der Vaart, 2000). The M-estimator is consistent, that is, it converges to the true value *regardless of the nuisance parameter* $\boldsymbol{\xi}^*$.[2] Moreover, it is normally distributed, that is,

[2] In this study, ‘consistency’ means ‘local consistency’ as well as in the previous works (Amari and Kawanabe, 1997; Amari and Cardoso, 2002; Kawanabe and Müller, 2005).

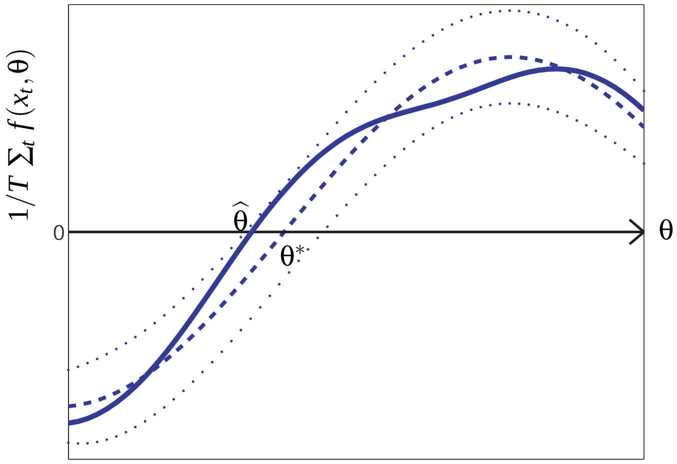

Figure 2: An illustrative plot of $1/T\sum_t f(x_t,\theta)$ as function of θ (solid line). Due to the effect of finite samples, the function is slightly apart from its expectation $\mathbb{E}_{\theta^*,\xi^*}[f(x,\theta)]$ (dashed line) which takes 0 at $\theta=\theta^*$ because of condition (6). Condition (7) means that the expectation (dashed line) has a non-zero slope around θ^*, which ensures the local uniqueness of the zero crossing point. On the other hand, condition (8) 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 $\hat{\theta}_T$ of estimating Equation (9) near the true value θ^*. This situation holds regardless of the value of the true nuisance parameter ξ^*.

$\hat{\boldsymbol{\theta}}_T \sim \mathcal{N}(\boldsymbol{\theta}^*, \mathrm{Av})$, when the sample size T approaches infinity. The matrix Av, which is called the asymptotic variance, can be calculated by

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

where $\mathbf{A}=\mathbb{E}_{\boldsymbol{\theta}^*,\boldsymbol{\xi}^*}[\partial_{\boldsymbol{\theta}}\boldsymbol{f}(\boldsymbol{x},\boldsymbol{\theta}^*)]$, and the symbol $\top$ denotes the matrix transpose. Note that the matrix Av depends on $(\boldsymbol{\theta}^*,\boldsymbol{\xi}^*)$, but not on the samples $\{\boldsymbol{x}_1,\cdots,\boldsymbol{x}_T\}$. We illustrate in Figure 2 the left side of the estimating Equation (9) normalized by the sample size T to explain why an M-estimator has good properties and to show the meaning of conditions (6)-(8).

4. Estimating Functions in MRP Model

The notion of estimating functions has been extended to be applicable to Markov time-series (Godambe, 1985, 1991; Wefelmeyer, 1996; Sørensen, 1999). We need a similar extension to enable it to be applied to MRPs. For convenience, we write the triplet at time t as $z_t \equiv \{s_{t-1}, s_t, r_t\} \in S^2\times R$ and the trajectory up to time t as $Z_t \equiv \{s_0,s_1,r_1,\ldots,s_t,r_t\} \in S^{t+1}\times R^t$.

Let us consider an m-dimensional vector-valued function of the form $\boldsymbol{f}_T : S^{T+1}\times R^T\times\Theta \mapsto \mathbb{R}^m$:

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

This is similar to the left side of (9) in the i.i.d. case, but now each term $\psi_t : S^{t+1} \times R^t \times \Theta \mapsto \mathbb{R}^m$ depends also on previous observations, that is, a function of the sequence up to time t. If the sequence of the functions $\{\psi_t\}$ satisfies the following properties for any $\boldsymbol{\theta}$ and $\boldsymbol{\xi}$, the function $\boldsymbol{f}_T$ becomes an estimating function for T sufficiently large (Godambe, 1985, 1991).

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

$$\det|\mathbf{A}| \neq 0, \ \text{ where } \ \mathbf{A} \equiv \lim_{t\to\infty} \mathbb{E}_{\boldsymbol{\theta},\boldsymbol{\xi}}[\partial_{\boldsymbol{\theta}}\psi_t(Z_t,\boldsymbol{\theta})], \tag{11}$$

$$\lim_{t\to\infty} \mathbb{E}_{\boldsymbol{\theta},\boldsymbol{\xi}}\left[\|\psi_t(Z_t,\boldsymbol{\theta})\|^2\right] < \infty. \tag{12}$$

Note that the estimating function $\boldsymbol{f}_T(Z_T,\boldsymbol{\theta})$ satisfies the martingale properties because of condition (10). Therefore, it is called a *martingale estimating function* in the literature (Godambe, 1985, 1991; Wefelmeyer, 1996; Sørensen, 1999).[3] Although time-series estimating functions can be defined in a more general form, the above definition is sufficient for our theoretical consideration.

4.1 Characterizing Class of Estimating Functions

In this section, we characterize possible estimating functions in MRPs. Let $\varepsilon : S^2 \times R \times \Theta \mapsto \mathbb{R}^1$ be the so-called temporal difference (TD) error, that is,

$$\varepsilon_t \equiv \varepsilon(z_t,\boldsymbol{\theta}) \equiv g(s_{t-1},\boldsymbol{\theta}) - \gamma g(s_t,\boldsymbol{\theta}) - r_t.$$

From Equation (4), its conditional expectation $\mathbb{E}_{\boldsymbol{\theta},\boldsymbol{\xi}_s}[\varepsilon_t|Z_{t-1}] = \mathbb{E}_{\boldsymbol{\theta},\boldsymbol{\xi}_s}[\varepsilon_t|s_{t-1}]$ is equal to 0 for any t. Furthermore, this zero-mean property holds even when multiplied by any weight function $\boldsymbol{w}_{t-1}(Z_{t-1},\boldsymbol{\theta})$, which depends on past observations and the parameter, that is,

$$\mathbb{E}_{\boldsymbol{\theta},\boldsymbol{\xi}_s}[\boldsymbol{w}_{t-1}(Z_{t-1},\boldsymbol{\theta})\varepsilon(z_t,\boldsymbol{\theta})|Z_{t-1}] = \boldsymbol{w}_{t-1}(Z_{t-1},\boldsymbol{\theta})\mathbb{E}_{\boldsymbol{\theta},\boldsymbol{\xi}_s}[\varepsilon(z_t,\boldsymbol{\theta})|Z_{t-1}] = \mathbf{0}, \tag{13}$$

for any t. We can obtain a class of possible estimating functions $\boldsymbol{f}_T(Z_T,\boldsymbol{\theta})$ in MRPs from this observation if we impose some regularity conditions summarized in Assumption 4.

Assumption 4

(a) *Function $\boldsymbol{w}_t : S^{t+1} \times R^t \times \Theta \mapsto \mathbb{R}^m$ can be twice continuously differentiable with respect to parameter $\boldsymbol{\theta}$ for any t, and $\lim_{t\to\infty} \mathbb{E}_{\boldsymbol{\theta},\boldsymbol{\xi}}[|\partial_{\boldsymbol{\theta}}\boldsymbol{w}_t(Z_t,\boldsymbol{\theta})|] < \infty$ for any $\boldsymbol{\theta}$.*

(b) *There exists a limit of matrix $\mathbb{E}_{\boldsymbol{\theta},\boldsymbol{\xi}}[\boldsymbol{w}_{t-1}(Z_{t-1},\boldsymbol{\theta})\{\partial_{\boldsymbol{\theta}}\varepsilon(z_t,\boldsymbol{\theta})\}^\top]$, and the matrix $\lim_{t\to\infty} \mathbb{E}_{\boldsymbol{\theta},\boldsymbol{\xi}}[\boldsymbol{w}_{t-1}(Z_{t-1},\boldsymbol{\theta})\{\partial_{\boldsymbol{\theta}}\varepsilon(z_t,\boldsymbol{\theta})\}^\top]$ is nonsingular for any $\boldsymbol{\theta}$ and $\boldsymbol{\xi}$.*

(c) *$\mathbb{E}_{\boldsymbol{\theta},\boldsymbol{\xi}}[\|\boldsymbol{w}_{t-1}(Z_{t-1},\boldsymbol{\theta})\varepsilon(z_t,\boldsymbol{\theta})\|^2]$ is finite for any t, $\boldsymbol{\theta}$ and $\boldsymbol{\xi}$.*

[3] Strictly speaking, strict consistency of M-estimator given by function $\boldsymbol{f}(Z_T,\boldsymbol{\theta})$ requires some additional conditions. To show consistency rigorously, we have to impose further conditions for exchange between limit and expectation operators in the neighborhood of the true parameter (more detailed discussion is shown in Theorem 3.6 in Sørensen 1999). In this article, for the sake of readability, we do not show such strict discussion.

Lemma 2 *Suppose that random sequence Z_T is generated from a distribution of semiparametric model $\{p(Z_T;\boldsymbol{\theta},\boldsymbol{\xi})\,|\,\boldsymbol{\theta},\boldsymbol{\xi}\}$ defined by Equation (5). If the conditions in Assumptions 1-4 are satisfied, then*

$$\boldsymbol{f}_T(Z_T,\boldsymbol{\theta}) = \sum_{t=1}^{T}\boldsymbol{\psi}_t(Z_t,\boldsymbol{\theta}) \equiv \sum_{t=1}^{T}\boldsymbol{w}_{t-1}(Z_{t-1},\boldsymbol{\theta})\varepsilon(z_t,\boldsymbol{\theta}) \tag{14}$$

becomes an estimating function.

The proof is given in Appendix C. From Lemma 2, we can obtain an M-estimator $\hat{\boldsymbol{\theta}}_T : S^{T+1}\times R^T \mapsto \mathbb{R}^m$ by solving the estimating equation

$$\sum_{t=1}^{T}\boldsymbol{\psi}_t(Z_t,\hat{\boldsymbol{\theta}}_T) = \mathbf{0}. \tag{15}$$

Practical procedures for finding the solution of the estimating Equation (15) will be discussed in Section 5. The estimator derived from the estimating Equation (15) has an asymptotic variance summarized in the following lemma.

Lemma 3 *Suppose that random sequence Z_T is generated from distribution $p(Z_T;\boldsymbol{\theta}^*,\boldsymbol{\xi}^*)$. If the conditions in Assumptions 1-4 are satisfied, then the M-estimator derived from Equation (15) has asymptotic estimation variance*

$$\mathrm{Av} = \mathrm{Av}(\hat{\boldsymbol{\theta}}_T) = \frac{1}{T}\mathbf{A}^{-1}\boldsymbol{\Sigma}\left(\mathbf{A}^\top\right)^{-1}, \tag{16}$$

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

The proof is given in Appendix D. Interestingly, for the MRP model, we can specify all possible estimating functions. More specifically, the converse of Lemma 2 also holds; any martingale estimating functions for MRP must take the form (14).

Theorem 4 *Suppose that the conditions in Assumptions 1-4 are satisfied. Then, any martingale estimating function $\boldsymbol{f}_T(Z_T,\boldsymbol{\theta}) = \sum_{t=1}^{T}\boldsymbol{\psi}_t(Z_t,\boldsymbol{\theta})$ in the semiparametric model $\{p(Z_T;\boldsymbol{\theta},\boldsymbol{\xi})\,|\,\boldsymbol{\theta},\boldsymbol{\xi}\}$ of MRP can be expressed as*

$$\boldsymbol{f}_T(Z_T,\boldsymbol{\theta}) = \sum_{t=1}^{T}\boldsymbol{\psi}_t(Z_t,\boldsymbol{\theta}) = \sum_{t=1}^{T}\boldsymbol{w}_{t-1}(Z_{t-1},\boldsymbol{\theta})\varepsilon(z_t,\boldsymbol{\theta}). \tag{17}$$

The proof is given in Appendix E.

4.2 Optimal Estimating Function

Since Theorem 4 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 the *minimum asymptotic variance*. The weight function $\boldsymbol{w}_t(Z_{t-1},\boldsymbol{\theta})$ may depend not only on the current state s_t and the parameter $\boldsymbol{\theta}$, but also on the previous states and rewards. However, we do not need to consider such weight functions, as Lemma 5 shows.

Lemma 5 *Let $\boldsymbol{w}_t(Z_t, \boldsymbol{\theta})$ be any weight function that depends on the current and previous observations and the parameter, and satisfies the conditions in Assumption 4. Then, there is necessarily a weight function depending only on the current state and the parameter whose corresponding estimator has the minimum asymptotic variance among all possible weight functions.*

The proof is given in Appendix F.

We next discuss the optimal weight function of Equation (14) in terms of asymptotic variance, which corresponds to the optimal estimating function.

Theorem 6 *Suppose that random sequence Z_T is generated from distribution $p(Z_T; \boldsymbol{\theta}^*, \boldsymbol{\xi}^*)$. If the conditions in Assumptions 1-4 are satisfied, an optimal estimating function with minimum asymptotic estimation variance is given by*

$$\boldsymbol{f}_T^*(Z_T, \boldsymbol{\theta}) = \sum_{t=1}^{T} \boldsymbol{\psi}_t^*(z_t, \boldsymbol{\theta}) \equiv \sum_{t=1}^{T} \boldsymbol{w}_{t-1}^*(s_{t-1}, \boldsymbol{\theta}^*) \varepsilon(z_t, \boldsymbol{\theta}), \tag{18}$$

where

$$\boldsymbol{w}_{t-1}^*(s_{t-1}, \boldsymbol{\theta}^*) \equiv \mathbb{E}_{\boldsymbol{\theta}^*, \boldsymbol{\xi}_s^*}[\varepsilon(z_t, \boldsymbol{\theta}^*)^2 | s_{t-1}]^{-1} \mathbb{E}_{\boldsymbol{\theta}^*, \boldsymbol{\xi}_s^*}[\partial_{\boldsymbol{\theta}} \varepsilon(z_t, \boldsymbol{\theta}^*) | s_{t-1}].$$

The proof is given in Appendix G. Note that weight function $\boldsymbol{w}_{t-1}^*(Z_{t-1}, \boldsymbol{\theta}^*)$ depends on true parameter $\boldsymbol{\theta}^*$ (unknown) and requires the expectation with respect to $p(r_t, s_t | s_{t-1}; \boldsymbol{\theta}^*, \boldsymbol{\xi}_s^*)$, which is also unknown. Therefore, we need to approximate the true parameter and the expectation, which will be explained in a later section.

The asymptotic variance of the optimal estimating function can be calculated from Lemma 3 and Theorem 6.

Lemma 7 *The minimum asymptotic variance is given by*

$$\mathrm{Av} = \mathrm{Av}(\hat{\boldsymbol{\theta}}_T) = \frac{1}{T} \mathbf{Q}^{-1},$$

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

The proof is given in Appendix H. We here note that positive definite matrix $\mathbf{Q}$ is similar to the *Fisher information matrix*, which is well-known in asymptotic estimation theory. However, the information associated with this matrix $\mathbf{Q}$ is generally smaller than the Fisher information because we sacrifice statistical efficiency for robustness against the nuisance parameter (Amari and Kawanabe, 1997; Amari and Cardoso, 2002). In other words, the estimator derived from the estimating function (18) does not achieve the statistical lower bound, that is, the Cramèr-Rao lower bound.[4]

5. Learning Algorithms

In this section, we present two kinds of practical algorithms to obtain the solution of the estimating Equation (15): one is the batch learning procedure and the other is the online learning procedure. In Section 5.1, we discuss batch learning and derive new least squares-type algorithms like LSTD and

[4] If one wants more efficient estimators, it is necessary to identify the target MRP, including the nuisance parameters.

LSTD(λ) to determine the parameter $\boldsymbol{\theta}$ under the assumption that the value function is represented as a parametrically linear function. In Section 5.2, we then study convergence issues of online learning. We first analyze the sufficient condition of the convergence of the estimation and the convergence rate of various online procedures without the constraint of linear parametrization. This theoretical consideration allows us to obtain a new online learning algorithm that asymptotically converges faster than current online algorithms.

5.1 Batch Learning

Let $g(s,\boldsymbol{\theta})$ be a linear parametric function of features:

$$V(s_t) \equiv \boldsymbol{\phi}(s_t)^\top \boldsymbol{\theta}, \tag{19}$$

where $\boldsymbol{\phi}: S \mapsto \mathbb{R}^m$ is a feature vector and $\boldsymbol{\theta} \in \Theta$ is a parameter vector. Then, estimating Equation (14) is given as

$$\sum_{t=1}^{T} \boldsymbol{w}_{t-1}(Z_{t-1},\boldsymbol{\theta}) \left\{ (\boldsymbol{\phi}(s_{t-1}) - \gamma\boldsymbol{\phi}(s_t))^\top \hat{\boldsymbol{\theta}}_T - r_t \right\} = \mathbf{0}.$$

If the weight function does not depend on parameter $\boldsymbol{\theta}$, the estimator $\hat{\boldsymbol{\theta}}_T$ can be analytically obtained as

$$\hat{\boldsymbol{\theta}}_T = \left\{ \sum_{t=1}^{T} \bar{\boldsymbol{w}}_{t-1}(Z_{t-1})(\boldsymbol{\phi}(s_{t-1}) - \gamma\boldsymbol{\phi}(s_t))^\top \right\}^{-1} \left\{ \sum_{t=1}^{T} \bar{\boldsymbol{w}}_{t-1}(Z_{t-1}) r_t \right\},$$

where $\bar{\boldsymbol{w}}_t : S^{t+1} \times R^t \mapsto \mathbb{R}^m$ is a function which depends only on the previous observations. Note that when the weight function $\bar{\boldsymbol{w}}(Z_t)$ is set to $\boldsymbol{\phi}(s_t)$, this estimator is equivalent to that of the LSTD learning.

We now derive a new least-squares learning algorithm, *generalized least squares temporal difference (gLSTD)*, which achieves minimum estimation of asymptotic variance in linear estimations of value functions. If weight function $\boldsymbol{w}_t^*(Z_t,\boldsymbol{\theta}^*)$ defined in Theorem 6 is known, an estimator of the estimating function (18) can be obtained as

$$\hat{\boldsymbol{\theta}}_T = \left\{ \sum_{t=1}^{T} \boldsymbol{w}_{t-1}^*(Z_{t-1},\boldsymbol{\theta}^*)(\boldsymbol{\phi}(s_{t-1}) - \gamma\boldsymbol{\phi}(s_t))^\top \right\}^{-1} \left\{ \sum_{t=1}^{T} \boldsymbol{w}_{t-1}^*(Z_{t-1},\boldsymbol{\theta}^*) r_t \right\},$$

by recalling that $\boldsymbol{w}_{t-1}^*(Z_{t-1},\boldsymbol{\theta}^*) = \mathbb{E}_{\boldsymbol{\theta}^*,\boldsymbol{\xi}_s^*}[\varepsilon(z_t,\boldsymbol{\theta}^*)^2|s_{t-1}]^{-1}\mathbb{E}_{\boldsymbol{\theta}^*,\boldsymbol{\xi}_s^*}[\boldsymbol{\phi}(s_{t-1}) - \gamma\boldsymbol{\phi}(s_t)|s_{t-1}]$. Obviously, we do not know $\boldsymbol{w}_{t-1}^*(Z_{t-1},\boldsymbol{\theta}^*)$ because the definition of $\boldsymbol{w}_{t-1}^*(Z_{t-1},\boldsymbol{\theta}^*)$ contains the residual at the true parameter, $\varepsilon(z_t,\boldsymbol{\theta}^*)$, and unknown conditional expectations, $\mathbb{E}_{\boldsymbol{\theta}^*,\boldsymbol{\xi}_s^*}[\varepsilon(z_t,\boldsymbol{\theta}^*)^2|s_{t-1}]$ and $\mathbb{E}_{\boldsymbol{\theta}^*,\boldsymbol{\xi}_s^*}[\boldsymbol{\phi}(s_{t-1}) - \gamma\boldsymbol{\phi}(s_t)|s_{t-1}]$. Therefore, we replace the true residual $\varepsilon(z_t,\boldsymbol{\theta}^*)$ with that of the LSTD estimator and approximate the expectations $\mathbb{E}_{\boldsymbol{\theta}^*,\boldsymbol{\xi}_s^*}[\varepsilon(z_t,\boldsymbol{\theta}^*)^2|s_{t-1}]^{-1}$ and $\mathbb{E}_{\boldsymbol{\theta}^*,\boldsymbol{\xi}_s^*}[\boldsymbol{\phi}(s_{t-1}) - \gamma\boldsymbol{\phi}(s_t)|s_{t-1}]$ by using function approximations

$$\mathbb{E}_{\boldsymbol{\theta}^*,\boldsymbol{\xi}_s^*}[\varepsilon(z_t,\boldsymbol{\theta}^*)^2|s_{t-1}]^{-1} \approx v(s_{t-1},\boldsymbol{\alpha}),$$
$$\mathbb{E}_{\boldsymbol{\theta}^*,\boldsymbol{\xi}_s^*}[\boldsymbol{\phi}(s_{t-1}) - \gamma\boldsymbol{\phi}(s_t)|s_{t-1}] \approx \boldsymbol{\zeta}(s_{t-1},\boldsymbol{\beta}),$$

Algorithm 1 gLSTD learning

```
for $t = 1, 2, \cdots$ do
    Obtain sample $z_t = \{s_{t-1}, s_t, r_t\}$
end for

Set constant $k$ to a sufficiently large value
for $t = 1, 2, \cdots$ do
    Calculate LSTD estimator $\hat{\boldsymbol{\theta}}^{\mathrm{LSTD}}$ based on sample
    $\{z_1, \cdots, z_{t-1}\} \cup \{z_{t+k}, \cdots, z_T\}$
    Calculate its residual $\hat{\varepsilon}_t$
        $\hat{\varepsilon}_t \leftarrow (\boldsymbol{\phi}(s_{t-1}) - \gamma\boldsymbol{\phi}(s_t))^\top \hat{\boldsymbol{\theta}}^{\mathrm{LSTD}} - r_t$
    Calculate conditional expectations $v(s_{t-1}, \boldsymbol{\alpha})$, $\boldsymbol{\zeta}(s_{t-1}, \boldsymbol{\beta})$ by means of function approximations
    based on sample $\{z_1, \cdots, z_{t-1}\} \cup \{z_{t+k}, \cdots, z_T\}$
    Obtain the weight function
        $\hat{\boldsymbol{w}}^*_{t-1} \leftarrow v(s_{t-1}, \boldsymbol{\alpha})^{-1} \boldsymbol{\zeta}(s_{t-1}, \boldsymbol{\beta})$
end for

Obtain the gLSTD estimator
    $\hat{\boldsymbol{\theta}}^{\mathrm{gLSTD}}_T \leftarrow [\sum_{t=1}^T \hat{\boldsymbol{w}}^*_{t-1} (\boldsymbol{\phi}(s_{t-1}) - \gamma\boldsymbol{\phi}(s_t))^\top]^{-1} [\sum_{t=1}^T \hat{\boldsymbol{w}}^*_{t-1} r_t]$
```

where $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are adjustable parameters for function approximators $v(s_{t-1}, \boldsymbol{\alpha})$ and $\boldsymbol{\zeta}(s_{t-1}, \boldsymbol{\beta})$, respectively. The estimation of conditional expectations is a simpler problem than that of the conditional probability itself. Also note that if the weight function is approximated by using past observations Z_{t-1}, condition (13) still holds regardless of the approximation accuracy of the weight function, implying the consistency of gLSTD. This is because any function that only depends on past observations can be employed as a weight function. This favorable characteristic is consistent regardless of the accuracy of approximation and allows us to use any approximation techniques (e.g., sparse regression, kernel regression, or neural networks) without particular constraints. Algorithm 1 demonstrates the pseudo-code of gLSTD learning. We introduce constant non-negative integer k to Algorithm 1 to enhance the efficient use of samples. LSTD estimator $\hat{\boldsymbol{\theta}}^{\mathrm{LSTD}}$ can be obtained in an unbiased manner by using future trajectory $\{z_{t+k}, \cdots, z_T\}$ for sufficiently large positive integer k, because the MRPs defined in Equation (1) satisfy geometrically uniform mixing, implying the exponential decay of the correlation between the statistics of s_t and s_{t+k}. Although k must be infinite to guarantee consistency in a strict sense, it could be a certain moderate integer when we consider the trade-off between the accuracy of function approximations and consistency. There are also some computational difficulties in Algorithm 1 with a large k value, because we must store the sample trajectory in memory to estimate weight function $\hat{\boldsymbol{w}}^*_t$ at each time t. Thus, in the simulations in Sections 7 and 8, we set k to zero; both the LSTD estimator and conditional expectations are calculated by using whole samples. Although this simplified implementation in fact violates the condition of consistency, it works well in practice.

5.2 Online Learning

Online learning procedures in the field of RL are often preferred to batch learning ones because they require less memory and can be adapted even to time-variant situations. Here, an online estimator of $\boldsymbol{\theta}$ at time t is denoted as $\hat{\boldsymbol{\theta}}_t$. Suppose that sequence $\{\psi_1(Z_1,\boldsymbol{\theta}),\cdots,\psi_T(Z_T,\boldsymbol{\theta})\}$ forms a martingale estimating function for MRP. Then, an online update rule can simply be given by

$$\hat{\boldsymbol{\theta}}_t = \hat{\boldsymbol{\theta}}_{t-1} - \eta_t \psi_t(Z_t,\hat{\boldsymbol{\theta}}_{t-1}), \tag{20}$$

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

$$\hat{\boldsymbol{\theta}}_t = \hat{\boldsymbol{\theta}}_{t-1} - \eta_t \mathbf{R}(\hat{\boldsymbol{\theta}}_{t-1})\psi_t(Z_t,\hat{\boldsymbol{\theta}}_{t-1}), \tag{21}$$

where $\mathbf{R}(\boldsymbol{\theta})$ denotes an $m\times m$ nonsingular matrix only depending on $\boldsymbol{\theta}$ (Amari, 1998). This variation results from the fact that function $\mathbf{R}(\boldsymbol{\theta})\sum_{i=1}^t \psi_i(Z_i,\boldsymbol{\theta})$ yields the same roots as its original for any $\mathbf{R}(\boldsymbol{\theta})$. This equivalence guarantees that both learning procedures, (20) and (21), have the same equilibrium, while their dynamics may be different, that is, even if the original algorithm (20) is unstable around the required solution, it can be stabilized by introducing appropriate $\mathbf{R}(\boldsymbol{\theta})$ into (21).

We will discuss the convergence of the online learning algorithm (21) in the next two subsections.

5.2.1 CONVERGENCE TO TRUE VALUE

We will now discuss the convergence of online learning (21) to the true parameter $\boldsymbol{\theta}^*$. For the sake of simplicity, we will focus on local convergence, that is, initial estimator $\hat{\boldsymbol{\theta}}_0$ is confined in the neighborhood of the true parameter, which is assumed to be a unique solution in the neighborhood. Now let us introduce sufficient conditions for convergence.

Assumption 5

(a) For any t, $(\hat{\boldsymbol{\theta}}_t - \boldsymbol{\theta}^*)^\top \mathbf{R}(\hat{\boldsymbol{\theta}}_t)\mathbb{E}_{\boldsymbol{\theta}^*,\boldsymbol{\xi}_s^*}[\psi_{t+1}(Z_{t+1},\hat{\boldsymbol{\theta}}_t)|s_t]$ is nonnegative.

(b) For any t, there exists such nonnegative constants c_1 and c_2 that

$$\mathbb{E}_{\boldsymbol{\theta}^*,\boldsymbol{\xi}_s^*}\left[\left\|\mathbf{R}(\hat{\boldsymbol{\theta}}_t)\psi_{t+1}(Z_{t+1},\hat{\boldsymbol{\theta}}_t)\right\|^2 \middle| s_t\right] \leq c_1 + c_2\left\|\hat{\boldsymbol{\theta}}_t - \boldsymbol{\theta}^*\right\|^2.$$

Condition (a) assumes that the opposite of gradient $\mathbf{R}(\hat{\boldsymbol{\theta}}_t)\mathbb{E}_{\boldsymbol{\theta}^*,\boldsymbol{\xi}_s^*}[\psi_{t+1}(Z_{t+1},\hat{\boldsymbol{\theta}}_t)|s_t]$ must point toward the true parameter $\boldsymbol{\theta}^*$ at each time t. Then, the following theorem guarantees the convergence of $\hat{\boldsymbol{\theta}}_t$ to $\boldsymbol{\theta}^*$.

Theorem 8 *Suppose that random sequence Z_T is generated from distribution $p(Z_T;\boldsymbol{\theta}^*,\boldsymbol{\xi}^*)$. Also, suppose that the conditions in Assumptions 1-5 hold. If 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 (21) almost surely converges to true parameter $\boldsymbol{\theta}^*$.*

The proof is given in Appendix I. Theorem 8 ensures that even if the original online learning algorithm (20) does not converge to the true parameter, we can construct an online learning algorithm with local consistency by appropriately choosing matrix $\mathbf{R}(\boldsymbol{\theta})$.

5.2.2 CONVERGENCE RATE

The convergence speed of an online algorithm could generally be slower than that of its batch counterpart that tries to solve the estimating equation using all available samples. However, if we set matrix $\mathbf{R}(\boldsymbol{\theta})$ and stepsizes $\{\eta_t\}$ appropriately, then it is possible to achieve the same convergence speed as that of the batch algorithm (Amari, 1998; Murata and Amari, 1999; Bottou and LeCun, 2004, 2005). Following the discussion on the previous work (Bottou and LeCun, 2004, 2005), we elucidate the convergence speed of online learning for estimating the value function in this section. Throughout the following discussion, the notion of *stochastic orders* plays a central role. Appendix A briefly describes the definition of stochastic orders and their properties. Then, we characterize the learning process for the batch algorithm.

Lemma 9 *Let $\tilde{\boldsymbol{\theta}}_t$ and $\tilde{\boldsymbol{\theta}}_{t-1}$ be solutions to estimating equations $(1/t)\sum_{i=1}^{t}\psi_i(Z_i,\tilde{\boldsymbol{\theta}}_t) = \mathbf{0}$ and $(1/(t-1))\sum_{i=1}^{t-1}\psi_i(Z_i,\tilde{\boldsymbol{\theta}}_{t-1}) = \mathbf{0}$, respectively. We assume that the conditions in Assumptions 2-4 are satisfied. Also we assume that $\tilde{\boldsymbol{\theta}}_t$ is uniformly bounded for any t, and matrix $\tilde{\mathbf{R}}_t(\tilde{\boldsymbol{\theta}}_{t-1}) \equiv (1/t)\sum_{i=1}^{t}\partial_{\boldsymbol{\theta}}\psi_i(Z_i,\tilde{\boldsymbol{\theta}}_{t-1})$ is nonsingular for any t. Then, we have*

$$\tilde{\boldsymbol{\theta}}_t = \tilde{\boldsymbol{\theta}}_{t-1} - \frac{1}{t}\tilde{\mathbf{R}}_t^{-1}(\tilde{\boldsymbol{\theta}}_{t-1})\psi_t(Z_t,\tilde{\boldsymbol{\theta}}_{t-1}) + O_p\left(\frac{1}{t^2}\right), \tag{22}$$

where the definition of $O_p(\cdot)$ is given in Appendix A.

The proof is given in Appendix J. Note that Equation (22) defines the sequence of $\tilde{\boldsymbol{\theta}}_t$ as a recursive stochastic process that is essentially the same as online learning (21) for the same $\mathbf{R}$. In other words, Lemma 9 indicates that online algorithms can converge with the same convergence speed as their batch counterparts through an appropriate choice of matrix $\mathbf{R}$. Finally, the following theorem addresses the convergence speed of the (stochastic) learning process such as that in Equation (22).

Theorem 10 *Suppose that random sequence Z_T is generated from distribution $p(Z_T;\boldsymbol{\theta}^*,\boldsymbol{\xi}^*)$, and then consider the following learning process*

$$\hat{\boldsymbol{\theta}}_t = \hat{\boldsymbol{\theta}}_{t-1} - \frac{1}{t}\hat{\mathbf{R}}_t^{-1}\psi_t(Z_t,\hat{\boldsymbol{\theta}}_{t-1}) + O_p\left(\frac{1}{t^2}\right), \tag{23}$$

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

- *(a) For any t, $\hat{\boldsymbol{\theta}}_t$ is uniformly bounded.*
- *(b) $\hat{\mathbf{R}}_t^{-1}$ can be written as $\hat{\mathbf{R}}_t^{-1} = \mathbb{E}_{\boldsymbol{\theta}^*,\boldsymbol{\xi}_s^*}[\hat{\mathbf{R}}_t^{-1}|Z_{t-1}] + o_p(1/t)$.*
- *(c) $\mathbb{E}_{\boldsymbol{\theta}^*,\boldsymbol{\xi}^*}[\partial_{\boldsymbol{\theta}}\psi_t(Z_t,\boldsymbol{\theta}^*)\hat{\boldsymbol{\theta}}_{t-1}\hat{\boldsymbol{\theta}}_{t-1}^\top]$ can be written as $\mathbb{E}_{\boldsymbol{\theta}^*,\boldsymbol{\xi}^*}[\partial_{\boldsymbol{\theta}}\psi_t(Z_t,\boldsymbol{\theta}^*)\hat{\boldsymbol{\theta}}_{t-1}\hat{\boldsymbol{\theta}}_{t-1}^\top] = \mathbb{E}_{\boldsymbol{\theta}^*,\boldsymbol{\xi}^*}[\partial_{\boldsymbol{\theta}}\psi_t(Z_t,\boldsymbol{\theta}^*)]\mathbb{E}_{\boldsymbol{\theta}^*,\boldsymbol{\xi}^*}[\hat{\boldsymbol{\theta}}_{t-1}\hat{\boldsymbol{\theta}}_{t-1}^\top] + o(1/t)$.*
- *(d) For any t, $\hat{\mathbf{R}}_t$ is a nonsingular matrix.*

Also assume that the conditions in Assumptions 1-4 are satisfied. If learning process (23) almost surely converges to the true parameter, then the convergence rate is given as

$$\mathbb{E}_{\boldsymbol{\theta}^*,\boldsymbol{\xi}^*}\left[\|\hat{\boldsymbol{\theta}}_t - \boldsymbol{\theta}^*\|^2\right] = \frac{1}{t}\mathrm{tr}\left\{\mathbf{A}^{-1}\boldsymbol{\Sigma}(\mathbf{A}^{-1})^\top\right\} + o\left(\frac{1}{t}\right), \tag{24}$$

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

The proof is given in Appendix K. Note that this convergence rate (24) is neither affected by the third term of (23) nor by the $o_p(1/t)$ term in matrix $\hat{\mathbf{R}}_t^{-1}$.

5.2.3 GENERALIZED TD LEARNING

We now present the online learning procedure that yields the minimum estimation error. Roughly speaking, this is given by estimating function $\boldsymbol{f}_T^*(Z_T,\boldsymbol{\theta})$ in Theorem 6 with the best (i.e., with the fastest convergence) choice of the nonsingular matrix in Theorem 10:

$$\hat{\boldsymbol{\theta}}_t = \hat{\boldsymbol{\theta}}_{t-1} - \frac{1}{t}\hat{\mathbf{Q}}_t^{-1}\boldsymbol{\psi}^*(z_t,\hat{\boldsymbol{\theta}}_{t-1}), \tag{25}$$

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 $\boldsymbol{\psi}^*(z_t,\boldsymbol{\theta})$ have been defined by Equation (18). If learning equation (25) satisfies conditions in Assumptions 1-5 and Theorem 10, then it converges to the true parameter with the minimum estimation error, $(1/t)\mathbf{Q}^{-1}$. However, this is impractical as learning rule (25) contains unknown parameters and quantities. For practical implementation, we need to evaluate $\mathbb{E}_{\boldsymbol{\theta}^*,\boldsymbol{\xi}_s^*}[\varepsilon(z_t,\boldsymbol{\theta}^*)^2|s_{t-1}]$ and $\mathbb{E}_{\boldsymbol{\theta}^*,\boldsymbol{\xi}_s^*}[\partial_{\boldsymbol{\theta}}\varepsilon(z_t,\boldsymbol{\theta}^*)|s_{t-1}]$ by using function approximations, whereas standard online learning procedures do not maintain statistics as a time series to avoid increasing the amount of memory. Therefore, we apply online functional approximations to these.

Let $v(s_t,\boldsymbol{\alpha}_t)$and $\boldsymbol{\zeta}(s_t,\boldsymbol{\beta}_t)$ be the approximations of $\mathbb{E}_{\boldsymbol{\theta}^*,\boldsymbol{\xi}_s^*}[\varepsilon(z_{t+1},\boldsymbol{\theta}_t)^2|s_t]$ and $\mathbb{E}_{\boldsymbol{\theta}^*,\boldsymbol{\xi}_s^*}[\partial_{\boldsymbol{\theta}}\varepsilon(z_{t+1},\boldsymbol{\theta}_t)|s_t]$, respectively. Here, $\boldsymbol{\alpha}_t$ and $\boldsymbol{\beta}_t$ are adjustable parameters, and they are adjusted in an online manner;

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

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

$$\begin{aligned}\boldsymbol{\psi}_t^*(z_t,\hat{\boldsymbol{\theta}}_{t-1}) &= v(s_{t-1},\hat{\boldsymbol{\alpha}}_{t-1})^{-1}\boldsymbol{\zeta}(s_{t-1},\hat{\boldsymbol{\beta}}_{t-1})\varepsilon(z_t,\hat{\boldsymbol{\theta}}_{t-1})\\ \hat{\mathbf{Q}}_t^{-1} &= \left\{\frac{1}{t}\sum_{i=1}^t v(s_{i-1},\hat{\boldsymbol{\alpha}}_{i-1})^{-1}\boldsymbol{\zeta}(s_{i-1},\hat{\boldsymbol{\beta}}_{i-1})\partial_{\boldsymbol{\theta}}\varepsilon(z_i,\hat{\boldsymbol{\theta}}_{i-1})^\top\right\}^{-1}.\end{aligned} \tag{26}$$

Note that update (26) can be done in an online manner by applying the well-known matrix inversion lemma (Horn and Johnson, 1985);

$$\hat{\mathbf{Q}}_t^{-1} = \frac{1}{(1-\varepsilon_t)}\hat{\mathbf{Q}}_{t-1}^{-1} - \frac{\varepsilon_t}{1-\varepsilon_t}\frac{\hat{\mathbf{Q}}_{t-1}^{-1}\hat{\boldsymbol{w}}_{t-1}^*\partial_{\boldsymbol{\theta}}\varepsilon(z_t,\hat{\boldsymbol{\theta}}_{t-1})^\top\hat{\mathbf{Q}}_{t-1}^{-1}}{(1-\varepsilon_t)+\varepsilon_t\partial_{\boldsymbol{\theta}}\varepsilon(z_t,\hat{\boldsymbol{\theta}}_{t-1})^\top\hat{\mathbf{Q}}_{t-1}^{-1}\hat{\boldsymbol{w}}_{t-1}^*}, \tag{27}$$

where $\varepsilon_t \equiv 1/t$ and $\hat{\boldsymbol{w}}_{t-1}^* \equiv v(s_{t-1},\hat{\boldsymbol{\alpha}}_{t-1})^{-1}\boldsymbol{\zeta}(s_{t-1},\hat{\boldsymbol{\beta}}_{t-1})$. Following Amari et al. (2000), we additionally simplify update equation (27) as

$$\hat{\mathbf{Q}}_t^{-1} = (1+\varepsilon_t)\hat{\mathbf{Q}}_{t-1}^{-1} - \varepsilon_t\hat{\mathbf{Q}}_{t-1}^{-1}\hat{\boldsymbol{w}}_{t-1}^*\partial_{\boldsymbol{\theta}}\varepsilon(z_t,\hat{\boldsymbol{\theta}}_{t-1})^\top\hat{\mathbf{Q}}_{t-1}^{-1}, \tag{28}$$

Algorithm 2 Optimal TD Learning

```
Initialize $\hat{\boldsymbol{\alpha}}_0$, $\hat{\boldsymbol{\beta}}_0$, $\hat{\boldsymbol{\theta}}_0$, $\hat{\mathbf{Q}}_0^{-1} = \varepsilon\mathbf{I}_m$, $a_1$, $a_2$
{$\varepsilon$ and $\mathbf{I}_m$ denote a small constant and an $m \times m$ identical matrix.}
for $t = 1, 2, \cdots$ do
    Obtain a new sample, $z_t = \{s_{t-1}, s_t, r_t\}$
    Calculate the weight function, $\hat{\boldsymbol{w}}^*_{t-1}$
      $\hat{\boldsymbol{\alpha}}_t \leftarrow \hat{\boldsymbol{\alpha}}_{t-1} - \eta_t^{\alpha}\partial_{\boldsymbol{\alpha}} v(s_{t-1}, \hat{\boldsymbol{\alpha}}_{t-1})\{v(s_{t-1}, \hat{\boldsymbol{\alpha}}_{t-1}) - \varepsilon(z_t, \hat{\boldsymbol{\theta}}_{t-1})^2\}$
      $\hat{\boldsymbol{\beta}}_t \leftarrow \hat{\boldsymbol{\beta}}_{t-1} - \eta_t^{\beta}\partial_{\boldsymbol{\beta}} \boldsymbol{\zeta}(s_{t-1}, \hat{\boldsymbol{\beta}}_{t-1})\{\boldsymbol{\zeta}(s_{t-1}, \hat{\boldsymbol{\beta}}_{t-1}) - \partial_{\boldsymbol{\theta}}\varepsilon(z_t, \hat{\boldsymbol{\theta}}_{t-1})\}$
      $\hat{\boldsymbol{w}}^*_{t-1} \leftarrow v(s_{t-1}, \hat{\boldsymbol{\alpha}}_{t-1})^{-1}\boldsymbol{\zeta}(s_{t-1}, \hat{\boldsymbol{\beta}}_{t-1})$
    Update $\hat{\mathbf{Q}}_t^{-1}$ by using Equation (28)
      $\hat{\mathbf{Q}}_t^{-1} \leftarrow (1 + (1/t))\hat{\mathbf{Q}}_{t-1}^{-1} - (1/t)\hat{\mathbf{Q}}_{t-1}^{-1}\hat{\boldsymbol{w}}^*_{t-1}\partial_{\boldsymbol{\theta}}\varepsilon(z_t, \hat{\boldsymbol{\theta}}_{t-1})^\top\hat{\mathbf{Q}}_{t-1}^{-1}$
    Update the parameter,
      $\tau \leftarrow \min(a_1, a_2/t)$
      $\hat{\boldsymbol{\theta}}_t \leftarrow \hat{\boldsymbol{\theta}}_{t-1} - (1/\tau)\hat{\mathbf{Q}}_t^{-1}\hat{\boldsymbol{w}}^*_{t-1}\varepsilon(z_t, \hat{\boldsymbol{\theta}}_{t-1})$
end for
```

which can be obtained because ε_t is small. We call this procedure *optimal TD learning* and its pseudo-code is summarized in Algorithm 2.[5]

5.2.4 ACCELERATED TD LEARNING

TD learning is a traditional online approach to model-free policy evaluation and has been one of the most important algorithms in the RL field. Although TD learning is widely used because of its simplicity, it is known that it converges rather slowly. This section discusses TD learning from the viewpoint of the method of estimating functions and proposes a new online algorithm that can achieve faster convergence than standard TD learning.

To simplify the following discussion, we have assumed that $g(s, \boldsymbol{\theta})$ is a linear function as in Equation (19) with which we can solve the linear estimating equation using both batch and online procedures. When weight function $\boldsymbol{w}_t(Z_t, \boldsymbol{\theta})$ in Equation (13) is set to $\boldsymbol{\phi}(s_t)$, the online and batch procedures correspond to the TD and LSTD algorithms, respectively. Note that both TD and LSTD share the same estimating function. Therefore, from Lemma 9 and Theorem 10, we can theoretically construct accelerated TD learning, which converges at the same speed as LSTD learning.

Here, we consider the following learning equation:

$$\hat{\boldsymbol{\theta}}_t = \hat{\boldsymbol{\theta}}_{t-1} - \frac{1}{t}\hat{\mathbf{R}}_t^{-1}\boldsymbol{\phi}(s_{t-1})\varepsilon(z_t, \hat{\boldsymbol{\theta}}_{t-1}), \tag{29}$$

where $\hat{\mathbf{R}}_t^{-1} = \{(1/t)\sum_{i=1}^t \boldsymbol{\phi}(s_{i-1})(\boldsymbol{\phi}(s_{i-1}) - \gamma\boldsymbol{\phi}(s_i))^\top\}^{-1}$. Since $\hat{\mathbf{R}}_t^{-1}$ converges to $\mathbf{A}^{-1} = \lim_{t\to\infty} \mathbb{E}_{\boldsymbol{\theta}^*, \boldsymbol{\xi}^*}[\boldsymbol{\phi}(s_{t-1})(\boldsymbol{\phi}(s_{t-1}) - \gamma\boldsymbol{\phi}(s_t))^\top]^{-1}$ and $\mathbf{A}^{-1}$ must be a positive definite matrix (see Lemma 6.4 in Bertsekas and Tsitsiklis 1996), online algorithm (29) also almost surely converges to the true parameter. Then, if $\hat{\mathbf{R}}_t$ satisfies the conditions in Theorem 10, it can achieve the *same*

[5] Since the online approximation of the weight function only depends on past observations, optimal TD learning is necessarily consistent even when the online approximation of the weight function is inaccurate.

Algorithm 3 Accelerated-TD Learning

```
Initialize $\hat{\boldsymbol{\theta}}_0$, $\hat{\mathbf{R}}_0^{-1} = \varepsilon \mathbf{I}_m$, $a_1$, $a_2$
{$\varepsilon$ and $\mathbf{I}_m$ denote a small constant and an $m \times m$ identical matrix.}

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

    Update $\hat{\mathbf{R}}_t^{-1}$
      $\hat{\mathbf{R}}_t^{-1} \leftarrow (1 + (1/t))\hat{\mathbf{R}}_{t-1}^{-1} - (1/t)\hat{\mathbf{R}}_{t-1}^{-1}\partial_{\boldsymbol{\theta}} g(s_{t-1}, \hat{\boldsymbol{\theta}}_{t-1})\partial_{\boldsymbol{\theta}}\varepsilon(z_t, \hat{\boldsymbol{\theta}}_{t-1})\hat{\mathbf{R}}_{t-1}^{-1}$

    Update the parameter,
      $\tau \leftarrow \min(a_1, a_2/t)$
      $\hat{\boldsymbol{\theta}}_t \leftarrow \hat{\boldsymbol{\theta}}_{t-1} - (1/\tau)\hat{\mathbf{R}}_t^{-1}\partial_{\boldsymbol{\theta}} g(s_{t-1}, \hat{\boldsymbol{\theta}}_{t-1})\varepsilon(z_t, \hat{\boldsymbol{\theta}}_{t-1})$
end for
```

convergence rate as LSTD. We call this procedure *Accelerated-TD learning*. We present an implementation of Accelerated-TD learning in Algorithm 3.

6. Related Work

This section discusses the relation between current major RL algorithms and the proposed ones from the viewpoint of estimating functions. Theorem 4 describes the broadest class of estimating functions that lead to unbiased estimators. Therefore, almost all the current value-based RL methods, in which consistency is assured, can be viewed as instances of the method of estimating functions.

For simplicity, let $g(s, \boldsymbol{\theta})$ be a linear function, that is, the value function can be represented as in Equation (19). We have two ways of solving such a linear estimating equation. The first is a batch procedure:

$$\hat{\boldsymbol{\theta}}_T = \left[\sum_{t=1}^{T} \boldsymbol{w}_{t-1} \left(\boldsymbol{\phi}(s_{t-1}) - \gamma \boldsymbol{\phi}(s_t)\right)^{\top}\right]^{-1} \left[\sum_{t=1}^{T} \boldsymbol{w}_{t-1} r_t\right].$$

and the second is an online procedure:

$$\hat{\boldsymbol{\theta}}_t = \hat{\boldsymbol{\theta}}_{t-1} - \eta_t \hat{\mathbf{R}}_t \boldsymbol{w}_{t-1} \varepsilon(z_t, \hat{\boldsymbol{\theta}}_{t-1}),$$

where $\boldsymbol{w}_t$ is a weight function at time t. By choosing both weight function $\boldsymbol{w}_{t-1}$ and the learning procedure, we can derive various RL algorithms. Let $\boldsymbol{f}_T^{\mathrm{TD}}$, $\boldsymbol{f}_T^{\mathrm{TD}(\lambda)}$, $\boldsymbol{f}_T^{\mathrm{RG}}$and $\boldsymbol{f}_T^{*}$ be the estimating functions that are defined as

$$\boldsymbol{f}_T^{\mathrm{TD}} \equiv \boldsymbol{f}_T^{\mathrm{TD}}(Z_T, \boldsymbol{\theta}) = \sum_{t=1}^{T} \boldsymbol{\phi}(s_{t-1})\varepsilon(z_t, \boldsymbol{\theta}),$$

$$\boldsymbol{f}_T^{\mathrm{TD}(\lambda)} \equiv \boldsymbol{f}_T^{\mathrm{TD}(\lambda)}(Z_T, \boldsymbol{\theta}) = \sum_{t=1}^{T}\sum_{i=1}^{t} (\gamma\lambda)^{t-i} \boldsymbol{\phi}(s_{i-1})\varepsilon(z_t, \boldsymbol{\theta}),$$

$$\boldsymbol{f}_T^{\mathrm{RG}} \equiv \boldsymbol{f}_T^{\mathrm{RG}}(Z_T, \boldsymbol{\theta}) = \sum_{t=1}^{T} \mathbb{E}_{\boldsymbol{\theta}^*, \boldsymbol{\xi}_s^*}[\boldsymbol{\phi}(s_{t-1}) - \gamma\boldsymbol{\phi}(s_t) | s_{t-1}]\varepsilon(z_t, \boldsymbol{\theta}),$$

$$\boldsymbol{f}_T^{*} \equiv \boldsymbol{f}_T^{*}(Z_T, \boldsymbol{\theta}) = \sum_{t=1}^{T} \mathbb{E}_{\boldsymbol{\theta}^*, \boldsymbol{\xi}_s^*}[(\varepsilon(z_t, \boldsymbol{\theta}^*)^2 | s_{t-1}]^{-1} \mathbb{E}_{\boldsymbol{\theta}^*, \boldsymbol{\xi}_s^*}[\boldsymbol{\phi}(s_{t-1}) - \gamma\boldsymbol{\phi}(s_t) | s_{t-1}]\varepsilon(z_t, \boldsymbol{\theta}).$$

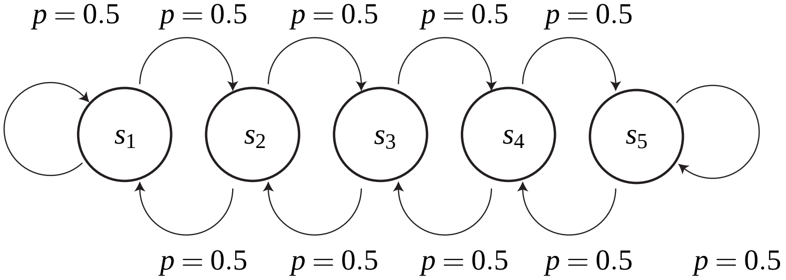

Figure 3: A five-states MRP.

Here, we remark that TD-based algorithms (TD Sutton and Barto, 1998, NTD Bradtke and Barto, 1996, LSTD Bradtke and Barto, 1996, LSPE Nedić and Bertsekas, 2003, GTD Sutton et al., 2009b, GTD2, TDC Sutton et al., 2009a and iLSTD Geramifard et al., 2006), TD (λ)-based algorithms (TD (λ) Sutton and Barto, 1998, NTD (λ) Bradtke and Barto, 1996, LSTD (λ) Boyan, 2002, LSPE (λ) Nedić and Bertsekas, 2003, and iLSTD (λ) Geramifard et al., 2007) and RG (Baird, 1995) originated from the estimating functions $\boldsymbol{f}_T^{\mathrm{TD}}$, $\boldsymbol{f}_T^{\mathrm{TD}(\lambda)}$ and $\boldsymbol{f}_T^{\mathrm{RG}}$, respectively. It should be noted that GTD, GTD2, GTDc, iLSTD, and iLSTD (λ) are specific online implementations for solving corresponding estimating equations; however, these algorithms can also be interpreted as instances of the method of estimating functions we propose. We have briefly summarized the relation between the current learning algorithms and the proposed algorithms in Table 1.

The asymptotic behavior of model-free policy evaluation has been analyzed within special contexts; Konda (2002) derived the asymptotic variance of LSTD (λ) and revealed that the convergence rate of TD (λ) was worse than that of LSTD (λ). Yu and Bertsekas (2006) derived the convergence rate of LSPE (λ) and found that it had the same convergence rate as LSTD (λ). Because these results can be seen in Lemma 3 and Theorem 8, our proposed framework generalizes previous asymptotic analyses to provide us with a methodology that can be more widely applied to carry out asymptotic analyses.

7. Simulation Experiment

In order to validate our theoretical developments, we compared the performance (statistical error) of the proposed algorithms (gLSTD, Accelerated-TD and Optimal-TD algorithms) with those of the online and batch baselines: TD algorithm (Sutton and Barto, 1998) and LSTD algorithm (Bradtke and Barto, 1996), respectively, in a very simple problem. An MRP trajectory was generated from a simple Markov random walk on a chain with five states $(s = 1, \cdots, 5)$ as depicted in Figure 3. At each time t, the state changes to either of its left (-1) or right $(+1)$ with equal probability of 0.5. A reward function was set as a deterministic function of the state:
$r = [0.6594, -0.3870, -0.9742, -0.9142, 0.9714]$[6] 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 (Mahadevan and Maggioni, 2007); basis functions were given based on the minor eigenvectors of the

[6] This reward function was prepared as follows. We first set the true value function by choosing the basis function and generating the parameter randomly, then the reward function was set so that it satisfied the Bellman equation.

Online Learning: $\hat{\boldsymbol{\theta}}_t = \hat{\boldsymbol{\theta}}_{t-1} - \eta_t \hat{\mathbf{R}}_t \boldsymbol{w}_{t-1}(Z_{t-1})\varepsilon(z_t, \hat{\boldsymbol{\theta}}_t)$

• $\boldsymbol{f}_T^{\mathrm{TD}}(Z_T, \boldsymbol{\theta}) = \sum_{t=1}^T \phi(s_{t-1})\varepsilon(z_t, \boldsymbol{\theta})$	
• TD (Sutton, 1988)	$\hat{\mathbf{R}}_t = \mathbf{R} = \mathbf{I}$
• NTD (Bradtke and Barto, 1996)	$\hat{\mathbf{R}}_t = \{(1/t)\sum_{i=1}^t \phi(s_i)^\top \phi(s_i)\}^{-1}\mathbf{I}$
• LSPE (Nedić and Bertsekas, 2003)	$\hat{\mathbf{R}}_t = \{(1/t)\sum_{i=1}^t \phi(s_i)\phi(s_i)^\top\}^{-1}$
• GTD (Sutton et al., 2009b)	See Equations (9) and (10) in the literature
• GTD2 (Sutton et al., 2009a)	See Equations (8) and (9) in the literature
• TDC (Sutton et al., 2009a)	See Equations (9) and (10) in the literature
• iLSTD (Geramifard et al., 2006)	See Algorithm 3 in the literature
• *Accelerated-TD Learning*	$\hat{\mathbf{R}}_t = \{(1/t)\sum_{i=1}^t \phi(s_{i-1})(\phi(s_{i-1}) - \gamma\phi(s_i))^\top\}^{-1}, \quad \eta_t = 1/t$
• $\boldsymbol{f}_T^{\mathrm{TD}(\lambda)}(Z_T, \boldsymbol{\theta}) = \sum_{t=1}^T \sum_{i=1}^t (\gamma\lambda)^{t-i}\phi(s_{i-1})\varepsilon(z_t, \boldsymbol{\theta})$	
• TD(λ) (Sutton, 1988)	$\hat{\mathbf{R}}_t = \mathbf{R} = \mathbf{I}$
• NTD(λ) (Bradtke and Barto, 1996)	$\hat{\mathbf{R}}_t = \{(1/t)\sum_{i=1}^t \phi^\top(s_i)\phi(s_i)\}^{-1}\mathbf{I}$
• LSPE(λ) (Nedić and Bertsekas, 2003)	$\hat{\mathbf{R}}_t = \{(1/t)\sum_{i=1}^t \phi(s_i)\phi(s_i)^\top\}^{-1}$
• iLSTD(λ) (Geramifard et al., 2007)	See Algorithm 2 in the literature
• $\boldsymbol{f}_T^{\mathrm{RG}}(Z_T, \boldsymbol{\theta}) = \sum_{t=1}^T \left(\phi(s_{t-1}) - \gamma\mathbb{E}_{\boldsymbol{\theta}^*, \boldsymbol{\xi}_s^*}[\phi(s_t) \mid s_{t-1}]\right)\varepsilon(z_t, \boldsymbol{\theta})$	
• RG (Baird, 1995)	$\hat{\mathbf{R}} = \mathbf{R} = \mathbf{I}$
• $\boldsymbol{f}_T^*(Z_T, \boldsymbol{\theta})$ given by Equation (18)	
• *Optimal-TD Learning*	$\hat{\mathbf{R}}_t = \hat{\mathbf{Q}}_t^{-1}$, $\eta_t = 1/t$

Batch Learning: $\hat{\boldsymbol{\theta}}_T = \left[\sum_{t=1}^T \boldsymbol{w}_{t-1}(Z_{t-1})(\phi(s_{t-1}) - \gamma\phi(s_t))^\top\right]^{-1}\left[\sum_{t=1}^T \boldsymbol{w}(Z_{t-1})r_t\right]$

- $\boldsymbol{f}_T^{\mathrm{TD}}(Z_T, \boldsymbol{\theta}) = \sum_{t=1}^T \phi(s_{t-1})\varepsilon(z_t, \boldsymbol{\theta})$
  - LSTD (Bradtke and Barto, 1996)
- $\boldsymbol{f}_T^{\mathrm{TD}(\lambda)}(Z_T, \boldsymbol{\theta}) = \sum_{t=1}^T \sum_{i=1}^t (\gamma\lambda)^{t-i}\phi(s_{i-1})\varepsilon(z_t, \boldsymbol{\theta})$
  - LSTD(λ) (Boyan, 2002)
- $\boldsymbol{f}_T^*(Z_T, \boldsymbol{\theta})$ given by Equation (18)
  - *gLSTD*

Table 1: Relation between the current learning and the proposed algorithms. **I** denotes an $m \times m$ identity matrix.

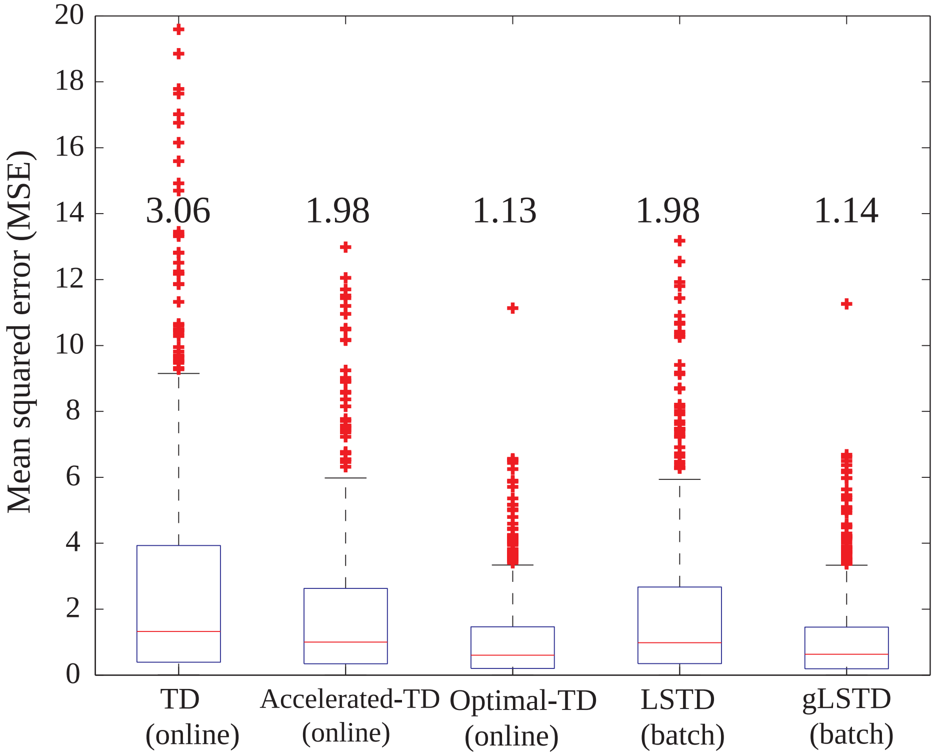

Figure 4: Boxplots of MSE both of the online (TD, Accelerated-TD and Optimal-TD) and batch (LSTD and gLSTD) algorithms. The center line, and the upper and lower sides of each box denote the median of MSE, and the upper and lower quartiles, respectively. The number above each box is the average MSE.

graph Laplacian on an undirected graph constructed by the state transition. The basis functions actually used in this simulation were $\phi(s_1) = [1, -0.6015, 0.5117]^\top$, $\phi(s_2) = [1, -0.3717, -0.1954]^\top$, $\phi(s_3) = [1, 0, -0.6325]^\top$, $\phi(s_4) = [1, 0.3717, -0.1954]^\top$, and $\phi(s_5) = [1, 0.6015, 0.5117]^\top$. In general, there is no guarantee that the true value function is included in the space spanned by the generated basis functions. In our example, however, the true value function can be represented faithfully by the basis vectors above.

We first generated $M = 500$ trajectories (episodes) each of which consisted of $T = 500$ random walk steps. The value function was estimated for each episode. We evaluated the mean squared error (MSE) between the true value function and the estimated value function, evaluated over the five states.

Figure 4 shows the boxplots of the MSE of the value functions estimated by the proposed (Accelerated-TD, Optimal-TD and gLSTD) and baseline (TD and LSTD) algorithms, in which the MSEs of all 500 episodes are shown by box-plots. For this example, the conditional expectations both in Optimal-TD and gLSTD can be calculated by sample average in each state, because there were only five states. In the online algorithms (TD, Accelerated-TD, and Optimal-TD), we used some batch procedures to obtain initial estimates of the parameters, as is often done in many online procedures. More specifically, the first 10 steps in each episode were used to obtain initial estimators in a batch manner and the online algorithm started after the 10 steps.

In the proposed online algorithms (Accelerated-TD and Optimal-TD), the stepsizes were decreased as simple as $1/t$. On the other hand, the convergence of TD learning was too slow in the simple $1/t$ setting due to fast decay of the stepsizes; this slow convergence was also observed when employing a certain well-chosen constant stepsize. Therefore, we adopt an ad-hoc adjustment for the stepsizes as $1/\tau$, where $\tau = \alpha_0(n_0+1)/(n_0+t)$. The best α_0 and n_0 have been selected by searching the sets of $\alpha_0 \in \{0.05, 0.1, 0.2, 0.3, 0.4\}$ and $n_0 \in \{10, 50, 100, 150, 200, 250, 300, 400, 500, 1000\}$, so that α_0 and n_0 are selected as 0.3 and 200, respectively.

As shown in Figure 4, the Optimal-TD and gLSTD algorithms achieved the minimum MSE among the online and batch algorithms, respectively. The MSEs by these two methods were comparable.[7] It should be noted that the Accelerated-TD algorithm performed significantly better than the ordinary TD algorithm, showing the matrix **R** was effective for accelerating the convergence of the online procedure as expected by our theoretical analysis.

Figure 5 shows how the estimation error of the estimator ($\hat{\theta}_T$) behaves as the learning proceeds, both for online (upper panel) and batch (lower panel) learning algorithms. X-axis and y-axis denote the number of learning steps and the estimation error, that is, the MSE between the true parameter and estimated parameter, average over 500 runs, respectively. The theoretical results, dotted and solid lines, exhibit good accordance with the simulation results, crosses and circles, respectively, as expected. Although our theoretical methods were mostly based on asymptotic analysis, they were supported by simulation results even in the cases of relatively small number of samples.

8. Discussion and Future Work

The contributions of this study are to present a new semiparametric approach to the model-free policy evaluation, which generalizes most of the current policy evaluation methods, and to clarify statistical properties of the policy evaluation problem. On the other hand, our framework to evaluate the policy evaluation has been restricted to situations in which the function approximation is faithful, that is, there is no model misspecification for the value function; we have not referred to statistical behaviors of our proposed algorithms in misspecified cases. In fact, the proposed algorithms may not better than current algorithms when the choice of parametric function g or the preparation of basis functions for approximating the value function introduces bias. Also, it is unsure whether our proposed online algorithms converge or not in misspecified cases. Figure 6 shows an example where the proposed algorithms (Optimal-TD and gLSTD) fail to obtain the best estimation accuracy. Here, an MRP trajectory was generated from an Markov random chain on the same dynamics as in Section 7. Rewards $+1$ and -1 were given when arriving at states '1' and '20', respectively, and the discounted factor was set at 0.98. Under this setting, we generated $M = 500$ trajectories (episodes) each of which consisted of $T = 1000$ random walk steps. We tested two linear function approximations with eight-dimensional and four-dimensional basis functions, respectively, which were also generated by the diffusion model. The former basis functions cause a tiny bias which can be ignored, whereas the latter ones make a significant bias. The upper and lower panels in Figure 6 show the

[7] In a particular implementation of the gLSTD algorithm (Algorithm 1) here, we used the whole sample trajectory to approximate the weight function $\boldsymbol{w}_t^*$, that is, $k = 0$, implying gLSTD does not necessarily hold consistency. Based on good agreement of the results between gLSTD and Optimal-TD, however, we can speculate that the approximation of the weight function using the whole sample trajectory did not affect the estimation accuracy so much.

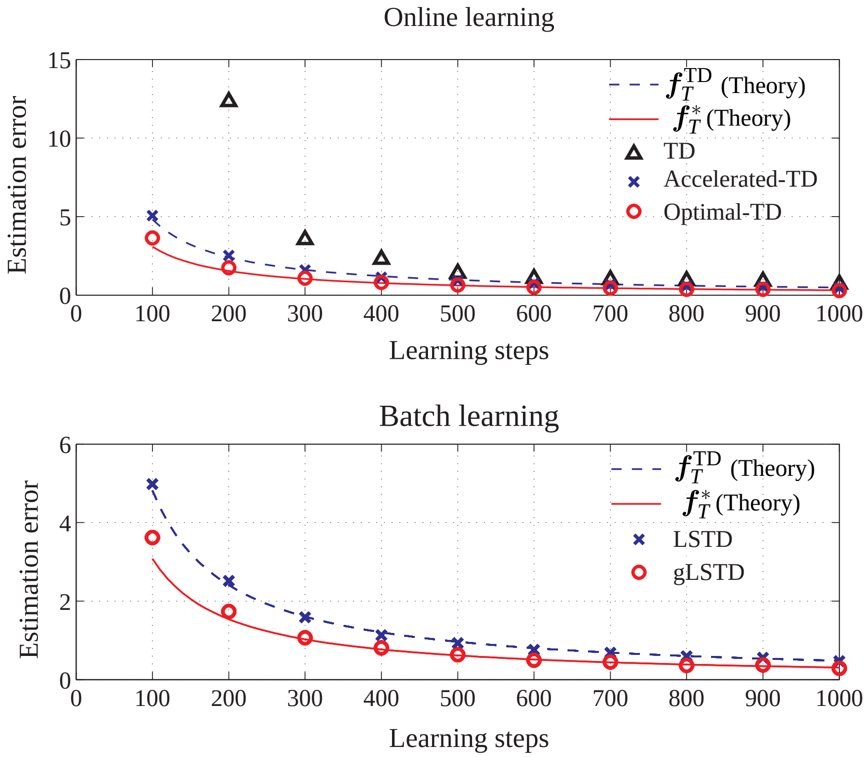

Figure 5: 500 learning runs by varying the initial conditions were performed. (Upper panel) Triangles ($\triangle$), crosses ($\times$) and circles ($\circ$) denote the simulation results for TD, Accelerated-TD and Optimal-TD, respectively. They were averaged over the 500 runs. The dotted and solid lines show the theoretical results discussed in Lemma 3 for estimating functions $\boldsymbol{f}_T^{\mathrm{TD}}$ and $\boldsymbol{f}_T^*$ described in Section 6. (Lower panel) Crosses ($\times$) and circles ($\circ$) denote the simulation results for LSTD and gLSTD, respectively.

MSEs of the value functions estimated by the proposed (Accelerated-TD, Optimal-TD and gLSTD) and baseline (TD and LSTD) algorithms employing eight-dimensional and four-dimensional basis functions, respectively. For scheduling of stepsizes in the online algorithms, we followed the same procedures as in Section 7. In the well-specified case (upper panel), the proposed algorithms achieved the smaller MSEs than the baseline algorithms as expected by our analysis, while in the misspecified case (lower panel), our proposed algorithms were inferior to the baseline algorithms. These results indicate the limitation of our analysis. When the bias-variance trade-off cannot be ignored, it is not sufficient to consider solely the asymptotic variance. Therefore, we need to analyze a *risk* $\mathcal{R}(\hat{\boldsymbol{\theta}}_T)$ which represents the deviation between the estimated value function and the true value function. Also, it is an important future work to construct good parametric representations (e.g., basis functions in linear cases) which attain small modeling biases. Furthermore, it is necessary to extend our convergence analysis for online learning algorithms to applicable to misspecified cases.

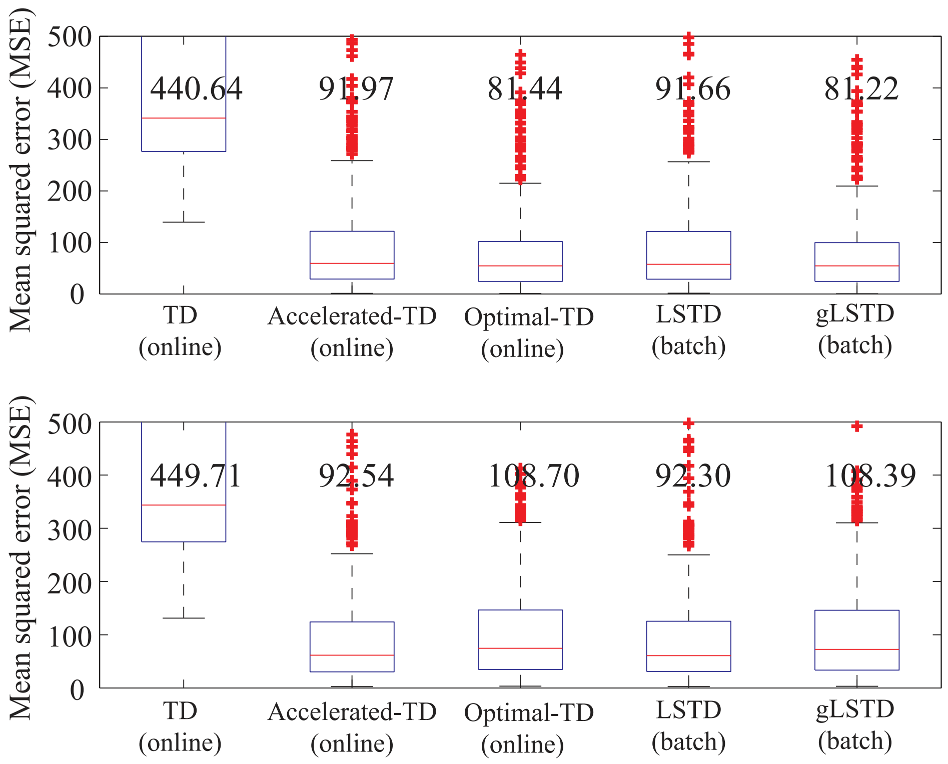

Figure 6: Boxplots of MSE for both of the online (TD, Accelerated-TD and Optimal-TD) and batch (LSTD and gLSTD) algorithms on a twenty states Markov random walk problem. (Upper panel) Simulation results on the function approximation with eight-dimensional diffusion basis functions. (Lower panel) Simulation results on the function approximation with four-dimensional diffusion basis functions.

8.1 Asymptotic Analysis in Misspecified Situations

First, let us revisit the asymptotic variance of the estimating function (15). In misspecified cases, estimating function (14) does not necessarily satisfy the martingale property, then its asymptotic variance can no longer be calculated by Equation (16). However, by introducing a notion of *uniform mixing*, the asymptotic variance can be correctly evaluated, even in misspecified cases.

To clarify the following discussion, we only consider the class of estimators given by the following estimating function $\bar{\boldsymbol{f}}_T : S^{T+1} \times R^T \times \Theta \mapsto \mathbb{R}^m$:

$$\bar{\boldsymbol{f}}_T(Z_T,\boldsymbol{\theta}) = \sum_{t=1}^{T} \bar{\boldsymbol{\psi}}_t(Z_t,\boldsymbol{\theta}) \equiv \sum_{t=1}^{T} \bar{\boldsymbol{w}}_{t-1}(Z_{t-1})\varepsilon(z_t,\boldsymbol{\theta}). \qquad (30)$$

Note that the class of estimators characterized by the above estimating function (30) is general enough for our theoretical consideration because it leads to almost all of the major algorithms for model-free policy evaluation that have been proposed so far (see Table 1). Now we demonstrate with Lemma 11 that the asymptotic variance of the estimators $\hat{\boldsymbol{\theta}}_T$ given by the estimating equation

$$\sum_{t=1}^{T} \bar{\psi}_t(Z_t, \hat{\boldsymbol{\theta}}) = \mathbf{0}. \tag{31}$$

Lemma 11 *Suppose that the random sequence Z_T is generated from the distribution $p(Z_T)$ defined by Equation (1). Assume that:*

(a) There exists such a parameter value $\bar{\boldsymbol{\theta}} \in \Theta$ that

$$\lim_{t\to\infty} \mathbb{E}\left[\bar{\psi}_t(Z_t, \bar{\boldsymbol{\theta}})\right] = \mathbf{0},$$

where that $\mathbb{E}[\cdot]$ denotes the expectation with respect to $p(Z_T)$, and $\hat{\boldsymbol{\theta}}_T$ converges to the parameter $\bar{\boldsymbol{\theta}}$ in probability.[8]

(b) There exists a limit of matrix $\mathbb{E}\left[\bar{\boldsymbol{w}}_{t-1}(Z_{t-1})\{\partial_{\boldsymbol{\theta}}\varepsilon(z_t, \bar{\boldsymbol{\theta}})\}^\top\right]$ and $\lim\limits_{t\to\infty} \mathbb{E}\left[\bar{\boldsymbol{w}}_{t-1}(Z_{t-1})\{\partial_{\boldsymbol{\theta}}\varepsilon(z_t, \bar{\boldsymbol{\theta}})\}^\top\right]$ is nonsingular.

(c) $\mathbb{E}\left[\|\bar{\boldsymbol{w}}_{t-1}(Z_{t-1})\varepsilon(z_t, \bar{\boldsymbol{\theta}})\|^2\right]$ is finite for any t.

Then, the estimator derived from estimating Equation (31) has the asymptotic variance

$$\widetilde{\mathrm{Av}} \equiv \widetilde{\mathrm{Av}}(\hat{\boldsymbol{\theta}}_T) \equiv \mathbb{E}\left[(\hat{\boldsymbol{\theta}}_T - \bar{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}}_T - \bar{\boldsymbol{\theta}})^\top\right] = \frac{1}{T}\bar{\mathbf{A}}^{-1}\bar{\boldsymbol{\Sigma}}\left(\bar{\mathbf{A}}^\top\right)^{-1}, \tag{32}$$

where

$$\begin{aligned}
\bar{\mathbf{A}} &\equiv \bar{\mathbf{A}}(\bar{\boldsymbol{\theta}}) \equiv \lim_{t\to\infty} \mathbb{E}\left[\bar{\boldsymbol{w}}_{t-1}\left\{\partial_{\boldsymbol{\theta}}\varepsilon(z_t, \bar{\boldsymbol{\theta}})\right\}^\top\right], \\
\bar{\boldsymbol{\Sigma}} &\equiv \bar{\boldsymbol{\Sigma}}(\bar{\boldsymbol{\theta}}) \equiv \lim_{t\to\infty} \mathbb{E}\left[\varepsilon(z_t, \bar{\boldsymbol{\theta}})^2 \bar{\boldsymbol{w}}_{t-1}\bar{\boldsymbol{w}}_{t-1}^\top\right] + \lim_{t\to\infty} 2\sum_{t'=1}^{\infty} \mathrm{cov}\left[\varepsilon(z_t, \bar{\boldsymbol{\theta}})\bar{\boldsymbol{w}}_{t-1}, \varepsilon(z_{t+t'}, \bar{\boldsymbol{\theta}})\bar{\boldsymbol{w}}_{t+t'-1}\right].
\end{aligned}$$

Here, $\bar{\boldsymbol{w}}_t$ and $\mathrm{cov}[\cdot,\cdot]$ denote the abbreviation of $\bar{\boldsymbol{w}}_t(Z_t)$ and the covariance function, respectively.

The proof is given in Appendix L. Since this proof required the central limit theorem under uniform mixing condition, we briefly review the notion and properties of uniform mixing in Appendix B. We note that the infinite sum of covariance in Equation (32) becomes zero when the parametric representation of the value function is faithful. This implies that Lemma 11 generalizes the result of Lemma 3.

Furthermore, we can derive the upper bound of the asymptotic variance (32).

Lemma 12 *There exists such a positive constant Υ that*

$$\frac{1}{T}\bar{\mathbf{A}}^{-1}\bar{\boldsymbol{\Sigma}}\left(\bar{\mathbf{A}}^\top\right)^{-1} \preceq \frac{\Upsilon}{T}\bar{\mathbf{A}}^{-1}\bar{\boldsymbol{\Sigma}}_0\left(\bar{\mathbf{A}}^\top\right)^{-1}$$

holds, where $\bar{\boldsymbol{\Sigma}}_0 \equiv \lim\limits_{t\to\infty} \mathbb{E}\left[\varepsilon(z_t, \bar{\boldsymbol{\theta}})^2 \bar{\boldsymbol{w}}_{t-1}\bar{\boldsymbol{w}}_{t-1}^\top\right]$.

[8] We can show the stochastic convergence of the estimator to the parameter $\bar{\boldsymbol{\theta}}$ by imposing further mild conditions to $\bar{\boldsymbol{f}}_T$. The proof can be obtained by following the procedure used in Theorem 3.6 in Sørensen (1999).

The proof is given in Appendix M. This lemma addresses that the estimators, which we have proposed so far, minimize the upper bound of the asymptotic variance in misspecified cases.

Lemma 11 allows us to see the asymptotic behavior of the risk, like done by the previous work in a different context; Liang and Jordan (2008) evaluated the quality of probabilistic model-based predictions in a structured prediction task. They analyzed the expected log-loss (risk) of composite likelihood estimators and compared it with those of generative, discriminative and pseudo-likelihood estimators, both when the probabilistic models are well-specified and misspecified. Since composite likelihood estimators are in the class of M-estimator, we will be able to evaluate the risk of various estimators by performing a similar analysis to Liang and Jordan (2008). Konishi and Kitagawa (1996) introduced generalized information criterion (GIC) which could be applied to evaluate statistical models constructed by various types of estimation procedures. GIC is the generalization of the well-known Akaike information criterion (AIC) (Akaike, 1974) and provided an unbiased estimator for the expected log-loss (risk) of statistical models obtained by M-estimators. Therefore, it may be possible to select a good model from a set of potential models by constructing an information criterion for model-free policy evaluation based on the analysis in Konishi and Kitagawa (1996).

8.2 Online Learning Procedures in Large Scale Situations

In both Optimal-TD and Accelerated-TD learning, it is necessary to maintain the inverse of the scaling matrix $\hat{\mathbf{R}}_t$. Since this matrix inversion operation costs $O(m^2)$ in each step, maintaining the inverse matrix becomes expensive when the dimensionality of the parameters increases. An efficient implementation in such a large-scale setting is to use a coarsely-represented scale matrix, for example, a diagonal or a block diagonal matrix. An appropriate setting still ensures the convergence rate of $O(1/t)$ without losing the computational efficiency. Le Roux et al. (2008) presented an interesting implementation of natural gradient learning (Amari, 1998) for large-scale settings, which was called “TONGA”. TONGA uses a low-rank approximation of the scaling matrix and casted both problems of finding the low-rank approximation and computing the gradient onto a lower-dimensional space, thereby attaining a lower computational complexity. Therefore, by applying such an idea to our proposed algorithms, we can improve the computational complexity without sacrificing the fast convergence.

9. Conclusions

We introduced a framework of semiparametric statistical inference for value function estimation which can be applied to analyzing both batch learning and online learning procedures. Based on this framework, we derived the general form of estimating functions for model-free value function estimation in MRPs, which provides a statistical basis to many batch and online learning algorithms available currently for policy evaluation. Moreover, we found an optimal estimating function, which yields the minimum asymptotic estimation variance amongst the general class, and presented new learning algorithms based on it as both batch and the online procedures. Using a simple MRP problem, we confirmed the validity of our analysis; actually, our proposed algorithms showed reasonably good performance.

Acknowledgments

This study was partly supported by Grant-in-Aid for Scientific Research (B) (No.21300113) from Japan Society for the Promotion of Science (JSPS) and the Federal Ministry of Economics and Technology of Germany (BMWi) under the project THESEUS, grant 01MQ07018. Tsuyoshi Ueno was also supported by JSPS Fellowship for Young Scientist.

Appendix A. Stochastic Order Symbols

The stochastic order symbols O_p and o_p are useful when evaluating the rate of convergence by means of asymptotic theory. Let n denote the number of observations. The stochastic order symbols are defined as follows.

Definition 13 *Let $\{X_n\}$ and $\{R_n\}$ denote a sequence of random variables and a sequence of real numbers, respectively. Then $X_n = o_p(R_n)$ if and only if X_n/R_n converges in probability to 0 when $n \to \infty$.*

Definition 14 *Let $\{X_n\}$ and $\{R_n\}$ denote a sequence of random variables and a sequence of real numbers, respectively. Then $X_n = O_p(R_n)$ if and only if X_n/R_n is bounded in probability when $n \to \infty$. "Bounded in probability" means that there exist a constant C_ε and a natural number $n_0(\varepsilon)$ such that for any $\varepsilon > 0$ and $n > n_0(\varepsilon)$,*

$$\mathbf{P}\left\{|X_n| \leq C_\varepsilon\right\} \geq 1 - \varepsilon$$

holds.

Most properties of the usual orders also apply to stochastic orders. For instance,

$$\begin{aligned} o_p(1) + o_p(1) &= o_p(1), \\ o_p(1) + O_p(1) &= O_p(1), \\ O_p(1)o_p(1) &= o_p(1), \\ (1 + o_p(1))^{-1} &= O_p(1), \\ o_p(R_n) &= R_n o_p(1), \\ O_p(R_n) &= R_n O_p(1), \\ o_p(O_p(1)) &= o_p(1). \end{aligned}$$

Moreover, by taking the expectation, the stochastic order symbol $o_p(\cdot)$ reduces to the usual order symbol $o(\cdot)$.

Remark 15 *Let $\{X_n\}$ and $\{R_n\}$ denote a sequence of random variables which satisfies $X_n = o_p(1)$ and a sequence of real numbers, respectively. Let $Y_n = X_n R_n$ denote a random variable which satisfies $Y_n = o_p(R_n)$. If the sequence of random variable Y_n is asymptotically uniformly integrable, then, the expectation of the random variables Y_n has the same normal order, $\mathbb{E}[Y_n] = o(R_n)$.*

This remark can be shown from Theorem 2.20 in van der Vaart (2000). Note that the sequence of real numbers $\{R_n\}$ which appears in Definition 13 and 14 corresponds to the *convergence rate*; then $Y_n = o_p(R_n)$ and $Y_n = O_p(R_n)$ mean that the sequence Y_n converges in probability to zero and is bounded in probability, respectively, at the rate of R_n.

Appendix B. Uniform Mixing and Central Limit Theorem

The notion of *mixing* is important when analyzing the rate of convergence in the stochastic processes which do not satisfy the martingale condition. There are several different definitions for mixing. In this section, we will especially focus on *uniform mixing* which is defined as follows.

Definition 16 *Let* $\mathrm{Y} \equiv \{Y_t : t = 1,2,\dots\}$ *be a strictly stationary process*[9] *on a probabilistic space* $(\Omega,\mathcal{F},P)$ *and* $\mathcal{F}_k^m$ *be* σ*-algebra generated by* $\{Y_k,\cdots,Y_m\}$*. Then, the process* Y *is said to be uniform mixing (*ϕ*-mixing) if* $\varphi(t) \to 0$ *as* $t \to \infty$ *where*

$$\varphi(t) \equiv \sup_{A\in\mathcal{F}_1^k, B\in\mathcal{F}_{k+t}^\infty} |P(B|A) - P(B)|, \;\; P(A) \neq 0.$$

The function $\varphi(t)$ is called *mixing coefficient*. If the mixing coefficient $\varphi(t)$ converges to zero as fast as exponential, then Y is called *geometrically uniform mixing*.

Definition 17 *Suppose that* Y *is a strictly stationary process. If there exist some constants* $C > 0$ *and* $\rho \in [0,1)$ *such that*

$$\varphi(t) < C\rho^t,$$

then Y *is said to be geometrically uniform mixing.*

Let f be a Borel function on the state space and define $\bar{f}_T = 1/T\sum_{t=1}^T f(Y_t)$. We now consider the conditions under which the central limit theorem holds for $\bar{f}_T$.

Lemma 18 *(Ibragimov and Linnik, 1971, Theorem 18.5.2.) Suppose that* $\{\mathrm{Y}_T\}$ *is a strictly stationary process with geometrically uniform mixing. If* $\lim_{t\to\infty}\mathbb{E}[\|f(Y_t)\|^2]$ *is finite, then the central limit theorem holds for* $\bar{f}$*, that is,*

$$\sqrt{T}\left(\bar{f}_T - \lim_{t\to\infty}\mathbb{E}[f(Y_t)]\right) \xrightarrow{d} \mathcal{N}(0,\sigma^2),$$

as $T \to \infty$ *where* $\sigma^2 \equiv \lim_{t\to\infty}\mathbb{E}[f(Y_t)^2] + 2\lim_{t\to\infty}\sum_{t'=1}^\infty \mathrm{cov}\left[f(Y_t), f(Y_{t+t'})\right]$.

Note that, unlike the i.i.d. or the martingale case, the variance of the asymptotic distribution involves the correlation between different times. Generally, such time dependency makes finding an exact relationship difficult; however, it may be easy to evaluate the upper bound of the time-dependent covariance.

Lemma 19 *(Ibragimov and Linnik, 1971, Theorem 17.2.3.) Suppose that* Y *is a strictly stationary process with uniform mixing. Let* f *and* g *be measurable functions with respect to* $\mathcal{F}_1^k$ *and* $\mathcal{F}_{k+t}^\infty$*, respectively. If* f *and* g *satisfy*

$$\mathbb{E}\left[|f|^p\right] < \infty, \;\; \mathbb{E}\left[|g|^q\right] < \infty,$$

where $p,q > 1$*,* $p+q = 1$*, then*

$$|\mathbb{E}\left[fg\right] - \mathbb{E}\left[f\right]\mathbb{E}\left[g\right]| \leq 2\varphi(t)^{1/p}\mathbb{E}\left[|f|^p\right]^{1/p}\mathbb{E}\left[|g|^q\right]^{1/q}.$$

Finally in this section, we consider the conditions that Markov processes satisfy the uniform mixing condition.

[9] In a strictly stationary stochastic process, joint probability distribution is consistent when shifted in time.

Lemma 20 *(Bradley, 2005, Theorem 3.1) Suppose that* Y *is a strictly stationary, finite state Markov process. Then the following statements are equivalent:*

(a) Y *is irreducible and aperiodic.*

(b) Y *is ergodic.*

(c) Y *is geometrically uniform mixing.*

Note that if a finite state Markov process has an unique and invariant stationary distribution, it implies ergodicity. Then Lemma 20 addresses that such Markov process is uniform mixing.

Appendix C. Proof of Lemma 2

Proof Condition corresponding to (12) is satisfied by condition (c) in Assumption 4. Also, condition (10) is satisfied by Equation (13). From condition (a) in Assumption 4, the expectation of the derivative of the function $\boldsymbol{w}_{t-1}(Z_{t-1},\boldsymbol{\theta})\varepsilon(z_t,\boldsymbol{\theta})$ can be expressed as

$$\begin{aligned}
&\lim_{t\to\infty}\mathbb{E}_{\boldsymbol{\theta},\boldsymbol{\xi}}\left[\partial_{\boldsymbol{\theta}}\left\{\boldsymbol{w}_{t-1}(Z_{t-1},\boldsymbol{\theta})\varepsilon(z_t,\boldsymbol{\theta})\right\}\right]\\
&=\lim_{t\to\infty}\mathbb{E}_{\boldsymbol{\theta},\boldsymbol{\xi}}\left[\partial_{\boldsymbol{\theta}}\boldsymbol{w}_{t-1}(Z_{t-1},\boldsymbol{\theta})\varepsilon(z_t,\boldsymbol{\theta})\right]+\lim_{t\to\infty}\mathbb{E}_{\boldsymbol{\theta},\boldsymbol{\xi}}\left[\boldsymbol{w}_{t-1}(Z_{t-1},\boldsymbol{\theta})\partial_{\boldsymbol{\theta}}\varepsilon(z_t,\boldsymbol{\theta})\right]\\
&=\lim_{t\to\infty}\mathbb{E}_{\boldsymbol{\theta},\boldsymbol{\xi}}\left[\partial_{\boldsymbol{\theta}}\boldsymbol{w}_{t-1}(Z_{t-1},\boldsymbol{\theta})\underbrace{\mathbb{E}_{\boldsymbol{\theta},\boldsymbol{\xi}_s}\left[\varepsilon(z_t,\boldsymbol{\theta})|Z_{t-1}\right]}_{=0}\right]+\lim_{t\to\infty}\mathbb{E}_{\boldsymbol{\theta},\boldsymbol{\xi}}\left[\boldsymbol{w}_{t-1}(Z_{t-1},\boldsymbol{\theta})\partial_{\boldsymbol{\theta}}\varepsilon(z_t,\boldsymbol{\theta})\right]\\
&=\lim_{t\to\infty}\mathbb{E}_{\boldsymbol{\theta},\boldsymbol{\xi}}\left[\boldsymbol{w}_{t-1}(Z_{t-1},\boldsymbol{\theta})\partial_{\boldsymbol{\theta}}\varepsilon(z_t,\boldsymbol{\theta})\right],
\end{aligned}$$

where we have used the fact in Equation (13). Therefore, using condition (b) in Assumption 4, we can show that condition (11) is satisfied. ■

Appendix D. Proof of Lemma 3

Proof By performing a Taylor series expansion of estimating Equation (15) around the true parameter $\boldsymbol{\theta}^*$, we obtain

$$\mathbf{0}=\sum_{t=1}^{T}\psi_t(Z_t,\boldsymbol{\theta}^*)+\sum_{t=1}^{T}\partial_{\boldsymbol{\theta}}\psi_t(Z_t,\boldsymbol{\theta}^*)(\hat{\boldsymbol{\theta}}_T-\boldsymbol{\theta}^*)+O_p\left(\left\|\hat{\boldsymbol{\theta}}_T-\boldsymbol{\theta}^*\right\|^2\right).$$

Here, high order terms of the above equation are in total represented as
$O_p(\|\hat{\boldsymbol{\theta}}_T-\boldsymbol{\theta}^*\|^2)$ because of Assumption 3, that is, the twice differentiable condition for the function $g(s,\boldsymbol{\theta})$. By applying the law of large numbers (ergodic pointwise theorem) (Billingsley, 1995, Theorem 24.1) to $(1/T)\sum_{t=1}^{T}\partial_{\boldsymbol{\theta}}\psi_t(Z_t,\boldsymbol{\theta}^*)$ and the martingale central limit theorem (Billingsley,

1961) to $(1/\sqrt{T})\sum_{t=1}^{T}\partial_{\boldsymbol{\theta}}\psi_t(Z_t,\boldsymbol{\theta}^*)$, we have

$$\begin{aligned}
\frac{1}{T}\sum_{t=1}^{T}\partial_{\boldsymbol{\theta}}\psi_t(Z_t,\boldsymbol{\theta}^*) &= \frac{1}{T}\sum_{t=1}^{T}\partial_{\boldsymbol{\theta}}\boldsymbol{w}_{t-1}(Z_{t-1},\boldsymbol{\theta}^*)\varepsilon(z_t,\boldsymbol{\theta}^*)+\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{w}_{t-1}(Z_{t-1},\boldsymbol{\theta}^*)\partial_{\boldsymbol{\theta}}\varepsilon(z_t,\boldsymbol{\theta}^*) \\
&\xrightarrow{a.s.} \underbrace{\lim_{t\to\infty}\mathbb{E}_{\boldsymbol{\theta}^*,\boldsymbol{\xi}^*}\left[\partial_{\boldsymbol{\theta}}\boldsymbol{w}_{t-1}(Z_{t-1},\boldsymbol{\theta}^*)\varepsilon(z_t,\boldsymbol{\theta}^*)\right]}_{=\mathbf{0}} \\
&\quad + \underbrace{\lim_{t\to\infty}\mathbb{E}_{\boldsymbol{\theta}^*,\boldsymbol{\xi}^*}\left[\boldsymbol{w}_{t-1}(Z_{t-1},\boldsymbol{\theta}^*)\{\partial_{\boldsymbol{\theta}}\varepsilon(z_t,\boldsymbol{\theta}^*)\}^\top\right]}_{=\mathbf{A}} \\
\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\psi_t(Z_t,\boldsymbol{\theta}^*) &= \frac{1}{\sqrt{T}}\sum_{t=1}^{T}\boldsymbol{w}_{t-1}(Z_{t-1},\boldsymbol{\theta}^*)\varepsilon(z_t,\boldsymbol{\theta}^*) \\
&\xrightarrow{d} \mathcal{N}\left(\mathbf{0},\underbrace{\lim_{t\to\infty}\mathbb{E}_{\boldsymbol{\theta}^*,\boldsymbol{\xi}^*}\left[\varepsilon(z_t,\boldsymbol{\theta}^*)^2\boldsymbol{w}_{t-1}(Z_{t-1},\boldsymbol{\theta}^*)\boldsymbol{w}_{t-1}(Z_{t-1},\boldsymbol{\theta}^*)^\top\right]}_{=\boldsymbol{\Sigma}}\right).
\end{aligned}$$

By neglecting higher order terms, we obtain

$$\sqrt{T}(\hat{\boldsymbol{\theta}}_T-\boldsymbol{\theta}^*)\sim\mathcal{N}\left(\mathbf{0},\mathbf{A}^{-1}\boldsymbol{\Sigma}(\mathbf{A}^\top)^{-1}\right).$$

Then, $\hat{\boldsymbol{\theta}}_T$ is Gaussian distributed: $\hat{\boldsymbol{\theta}}_T\sim\mathcal{N}(\boldsymbol{\theta}^*,\mathrm{Av})$, where the asymptotic variance Av is given by Equation (16). ■

Appendix E. Proof of Theorem 4

Proof From Equation (2), for any t, the value function $V(s_t)=g(s_t,\boldsymbol{\theta})$ must satisfy

$$\mathbb{E}_{\boldsymbol{\theta},\boldsymbol{\xi}_s}[r_{t+1}|s_t]=g(s_t,\boldsymbol{\theta})-\mathbb{E}_{\boldsymbol{\theta},\boldsymbol{\xi}_s}[g(s_{t+1},\boldsymbol{\theta})|s_t],$$

regardless of the nuisance parameter $\boldsymbol{\xi}$. Then, the TD error $\varepsilon(z_{t+1},\boldsymbol{\theta})=g(s_t,\boldsymbol{\theta})-\gamma g(s_{t+1},\boldsymbol{\theta})-r_{t+1}$ must satisfy $\mathbb{E}_{\boldsymbol{\theta},\boldsymbol{\xi}_s}[\varepsilon(z_{t+1},\boldsymbol{\theta})|Z_t]=0$ for any t and $\boldsymbol{\xi}$. Also, from the condition of martingale estimating functions, for any time t, the estimating function must satisfy

$$\mathbb{E}_{\boldsymbol{\theta},\boldsymbol{\xi}_s}[\boldsymbol{f}_{t+1}(Z_{t+1},\boldsymbol{\theta})-\boldsymbol{f}_t(Z_t,\boldsymbol{\theta})|Z_t]=\mathbf{0}, \tag{33}$$

regardless of the nuisance parameter $\boldsymbol{\xi}$. If we can show from Equation (33) that $\boldsymbol{f}_{t+1}(Z_{t+1},\boldsymbol{\theta})-\boldsymbol{f}_t(Z_t,\boldsymbol{\theta})=\boldsymbol{w}_t(Z_t,\boldsymbol{\theta})\varepsilon(z_{t+1},\boldsymbol{\theta})$ holds, $\boldsymbol{f}_T(Z_T,\boldsymbol{\theta})$ must have the form (17) by induction. Since this statement can be considered component-wise, we will prove the similar claim for scalar functions, that is,

$$\mathbb{E}_{\boldsymbol{\theta},\boldsymbol{\xi}_s}[h(Z_{t+1},\boldsymbol{\theta})|Z_t]=0,\ \forall\boldsymbol{\xi}_s \quad\Longrightarrow\quad h(Z_{t+1},\boldsymbol{\theta})=w(Z_t,\boldsymbol{\theta})\varepsilon(z_{t+1},\boldsymbol{\theta}), \tag{34}$$

in the following two steps.

1. We first prove in a *constructive* manner that any simple function $h(z_{t+1},\boldsymbol{\theta})$ which depends only on z_{t+1} and satisfy $\mathbb{E}_{\boldsymbol{\theta},\boldsymbol{\xi}_s}[h(z_{t+1},\boldsymbol{\theta})|s_t]=0$ and $\mathbb{E}_{\boldsymbol{\theta},\boldsymbol{\xi}_s}[\{h(z_{t+1},\boldsymbol{\theta})\}^2]<\infty$ for any t, $\boldsymbol{\theta}$ and $\boldsymbol{\xi}_s$ can be expressed as $h(z_{t+1},\boldsymbol{\theta})=w(s_t,\boldsymbol{\theta})\varepsilon(z_{t+1},\boldsymbol{\theta})$, where $w(s_t,\boldsymbol{\theta})$ is a function of s_t.

2. Our claim (34) for general function $h(Z_{t+1},\boldsymbol{\theta})$ is derived from the fact shown in the previous step, because for each fixed Z_t this problem boils down to the simple case above.

To prove the simple case first, for arbitrary fixed s_t and $\boldsymbol{\theta}$, we consider the set $\mathcal{M}(s_t,\boldsymbol{\theta})$ of all probability distributions of r_{t+1} and s_{t+1} with each of which the expectation of the TD error $\varepsilon(z_{t+1},\boldsymbol{\theta})$ vanishes. In the following discussion, s_t is treated as a *fixed constant*. In our semiparametric case, this set can be expressed as the set of all conditional distributions of r_{t+1} and s_{t+1} for given s_t which has value function $g(s_t,\boldsymbol{\theta})$ with the fixed $\boldsymbol{\theta}$, that is,

$$\begin{aligned}\mathcal{M}(s_t,\boldsymbol{\theta}) \;\equiv\; \{\, & p(r_{t+1},s_{t+1}|s_t;\boldsymbol{\theta},\boldsymbol{\xi}_s)\,|\,s_t,\,\boldsymbol{\theta}\text{: fixed},\\ & \mathbb{E}_{\boldsymbol{\theta},\boldsymbol{\xi}_s}[\varepsilon(z_{t+1},\boldsymbol{\theta})|s_t]=g(s_t,\boldsymbol{\theta})-\gamma\mathbb{E}_{\boldsymbol{\theta},\boldsymbol{\xi}_s}[g(s_{t+1},\boldsymbol{\theta})|s_t]-\mathbb{E}_{\boldsymbol{\theta},\boldsymbol{\xi}_s}[r_{t+1}|s_t]=0\,\},\end{aligned}$$

where the nuisance parameter $\boldsymbol{\xi}_s$ is designed so that it becomes bijective with the distributions in $\mathcal{M}(s_t,\boldsymbol{\theta})$. We remark that the set $\mathcal{M}(s_t,\boldsymbol{\theta})$ and the domain of the nuisance parameter $\boldsymbol{\xi}_s$ depend on s_t and $\boldsymbol{\theta}$.

Suppose that there exists a function of $h(z_{t+1})$ which satisfies $\mathbb{E}_p[h(z_{t+1})]=0$ and $\mathbb{E}_p[\{h(z_{t+1})\}^2]<\infty$ for any $p(r_{t+1},s_{t+1})\in\mathcal{M}(s_t,\boldsymbol{\theta})$. Then, because of the linearity and continuity of the integral operator, the unbiasedness condition can be extended to any function $q(r_{t+1},s_{t+1})$ which belongs to the closed span $\overline{\mathcal{M}}(s_t,\boldsymbol{\theta})$ of $\mathcal{M}(s_t,\boldsymbol{\theta})$:

$$\sum_{s_{t+1}}\int h(z_{t+1})\,q(r_{t+1},s_{t+1})\,dr_{t+1}=0. \tag{35}$$

It is also easy to show that $\overline{\mathcal{M}}(s_t,\boldsymbol{\theta})$ contains any functions (i.e., even without satisfying the non-negativity constraint of probabilities) which satisfy the condition

$$\sum_{s_{t+1}}\int \varepsilon(z_{t+1})\,q(r_{t+1},s_{t+1})\,dr_{t+1}=0. \tag{36}$$

Indeed, we can always construct a linear representation of such a function $q(r_{t+1},s_{t+1})$ with four probability distributions in $\mathcal{M}(s_t,\boldsymbol{\theta})$ which take positive values only in two regions out of $\{(r_{t+1},s_{t+1})\,|\,\varepsilon\geq 0,\ q\geq 0\}$, $\{(r_{t+1},s_{t+1})\,|\,\varepsilon\geq 0,\ q<0\}$, $\{(r_{t+1},s_{t+1})\,|\,\varepsilon<0,\ q\geq 0\}$ and $\{(r_{t+1},s_{t+1})\,|\,\varepsilon<0,\ q<0\}$.

Now, we take a distribution $p(r_{t+1},s_{t+1})$ in $\mathcal{M}(s_t,\boldsymbol{\theta})$ which is positive over its domain[10] and consider its perturbation

$$\widetilde{q}(r_{t+1},s_{t+1})\equiv p(r_{t+1},s_{t+1})\left\{1+\delta h(z_{t+1})-\delta\frac{\mathbb{E}_p[h(z_{t+1})\varepsilon(z_{t+1})]}{\mathbb{E}_p[\varepsilon(z_{t+1})^2]}\varepsilon(z_{t+1})\right\},$$

where $\delta>0$ is a small constant and $\mathbb{E}_p$ denotes the expectation over r_{t+1} and s_{t+1} with respect to $p(r_{t+1},s_{t+1})$. This function $\widetilde{q}(r_{t+1},s_{t+1})$ does not necessarily belong to the model $\mathcal{M}(s_t,\boldsymbol{\theta})$, but

[10] If there exists a region where all distributions in $\mathcal{M}(s_t,\boldsymbol{\theta})$ take 0, it is impossible to characterize the functional form of $h(z_{t+1})$ in that region. For simplicity, however, we do not consider such a pathological case in this proof.

is an element of its closed span $\overline{\mathcal{M}}(s_t,\boldsymbol{\theta})$, because it also satisfies the condition (36). Therefore, Equation (35) must hold for this perturbed function $\widetilde{q}$, leading to

$$\sum_{s_{t+1}}\int h(z_{t+1})\,\widetilde{q}(r_{t+1},s_{t+1})\,dr_{t+1} = \delta\left\{\mathbb{E}_p[h(z_{t+1})^2] - \frac{(\mathbb{E}_p[h(z_{t+1})\varepsilon(z_{t+1})])^2}{\mathbb{E}_p[\varepsilon(z_{t+1})^2]}\right\} = 0. \tag{37}$$

From Cauchy-Schwarz's inequality, this equation holds if and only if $h(z_{t+1}) \propto \varepsilon(z_{t+1})$ and otherwise $\mathbb{E}_{\widetilde{q}}[h(z_{t+1})]$, the left-hand-side of Equation (37), becomes strictly positive, which contradicts the fact (35). Since the whole argument holds for any s_t and $\boldsymbol{\theta}$, the first claim is proved.

In the general case for the function of Z_{t+1}, we just show that any function $h(Z_{t+1},\boldsymbol{\theta})$ which satisfies $\mathbb{E}_{\boldsymbol{\theta},\boldsymbol{\xi}_s}[h(Z_{t+1},\boldsymbol{\theta})|Z_t] = 0$ and $\mathbb{E}_{\boldsymbol{\theta},\boldsymbol{\xi}_s}[\{h(Z_{t+1},\boldsymbol{\theta})\}^2] < \infty$ for any s_t, $\boldsymbol{\theta}$ and $\boldsymbol{\xi}_s$ can be expressed as $h(Z_{t+1},\boldsymbol{\theta}) = w_t(Z_t,\boldsymbol{\theta})\varepsilon(z_{t+1},\boldsymbol{\theta})$, where $w_t(Z_t,\boldsymbol{\theta})$ is a function of Z_t and $\boldsymbol{\theta}$.

For arbitrary fixed Z_t, $h(Z_{t+1},\boldsymbol{\theta})$ can be regarded as a function of r_{t+1} and s_{t+1}. Therefore, the problem reduces to the case that the function only depends on z_{t+1}, so that we can say that $h(r_{t+1},s_{t+1},Z_t,\boldsymbol{\theta}) \propto \varepsilon(r_{t+1},s_{t+1},s_t)$. Since this relationship holds for any Z_t, we conclude that the function $h(Z_{t+1},\boldsymbol{\theta})$ must have the form $w_t(Z_t,\boldsymbol{\theta})\varepsilon(z_{t+1},\boldsymbol{\theta})$. ■

Appendix F. Proof of Lemma 5

Proof We show that the conditional expectation $\tilde{\boldsymbol{w}}_t(s_t,\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{\xi}_s}[\boldsymbol{w}_t(Z_t,\boldsymbol{\theta})|s_t]$, which depends only on the current state s_t and the parameter $\boldsymbol{\theta}$, gives an equally good estimator or better estimator than those by the original weight function $\boldsymbol{w}_t(Z_t,\boldsymbol{\theta})$. As shown in Equation (16), the asymptotic variance of the estimator $\hat{\boldsymbol{\theta}}_w$ with $\boldsymbol{w}_t(Z_t,\boldsymbol{\theta})$ is given by

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

where $\mathbf{A}_w = \lim_{t\to\infty}\mathbb{E}_{\boldsymbol{\theta}^*,\boldsymbol{\xi}^*}\left[\boldsymbol{w}_{t-1}\{\partial_{\boldsymbol{\theta}}\varepsilon(z_t,\boldsymbol{\theta}^*)\}^\top\right] \equiv \lim_{t\to\infty}\mathbb{E}_{\boldsymbol{\theta}^*,\boldsymbol{\xi}^*}\left[\boldsymbol{w}_{t-1}(Z_{t-1},\boldsymbol{\theta}^*)\{\partial_{\boldsymbol{\theta}}\varepsilon(z_t,\boldsymbol{\theta}^*)\}^\top\right]$ and $\boldsymbol{\Sigma}_w = \lim_{t\to\infty}\mathbb{E}_{\boldsymbol{\theta}^*,\boldsymbol{\xi}^*}\left[(\varepsilon_t^*)^2\boldsymbol{w}_{t-1}\boldsymbol{w}_{t-1}^\top\right] \equiv \lim_{t\to\infty}\mathbb{E}_{\boldsymbol{\theta}^*,\boldsymbol{\xi}^*}\left[(\varepsilon_t^*)^2\boldsymbol{w}_{t-1}(Z_{t-1},\boldsymbol{\theta}^*)\boldsymbol{w}_{t-1}(Z_{t-1},\boldsymbol{\theta}^*)^\top\right]$. Here, $\boldsymbol{w}_t$ is an abbreviation of $\boldsymbol{w}_t(Z_t,\boldsymbol{\theta}^*)$. Similarly, the asymptotic variance of the estimator $\hat{\boldsymbol{\theta}}_{\tilde{w}}$ with $\tilde{\boldsymbol{w}}_{t-1}(s_{t-1},\boldsymbol{\theta})$ is given by

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

where $\mathbf{A}_{\tilde{w}} \equiv \lim_{t\to\infty}\mathbb{E}_{\boldsymbol{\theta}^*,\boldsymbol{\xi}^*}\left[\tilde{\boldsymbol{w}}_{t-1}\{\partial_{\boldsymbol{\theta}}\varepsilon(z_t,\boldsymbol{\theta}^*)\}^\top\right] \equiv \lim_{t\to\infty}\mathbb{E}_{\boldsymbol{\theta}^*,\boldsymbol{\xi}^*}\left[\tilde{\boldsymbol{w}}_{t-1}(s_{t-1},\boldsymbol{\theta}^*)\{\partial_{\boldsymbol{\theta}}\varepsilon(z_t,\boldsymbol{\theta}^*)\}^\top\right]$ and $\boldsymbol{\Sigma}_{\tilde{w}} \equiv \lim_{t\to\infty}\mathbb{E}_{\boldsymbol{\theta}^*,\boldsymbol{\xi}^*}\left[(\varepsilon_t^*)^2\tilde{\boldsymbol{w}}_{t-1}\tilde{\boldsymbol{w}}_{t-1}^\top\right] \equiv \lim_{t\to\infty}\mathbb{E}_{\boldsymbol{\theta}^*,\boldsymbol{\xi}^*}\left[(\varepsilon_t^*)^2\tilde{\boldsymbol{w}}_{t-1}(s_{t-1},\boldsymbol{\theta}^*)\tilde{\boldsymbol{w}}_{t-1}(s_{t-1},\boldsymbol{\theta}^*)^\top\right]$. Here, $\tilde{\boldsymbol{w}}_t$ is an abbreviation of $\tilde{\boldsymbol{w}}_t(s_t,\boldsymbol{\theta})$. The matrices $\mathbf{A}_w$ and $\boldsymbol{\Sigma}_w$ can be calculated as

$$
\begin{aligned}
\mathbf{A}_w &= \lim_{t\to\infty} \mathbb{E}_{\boldsymbol{\theta}^*,\boldsymbol{\xi}^*}\left[\boldsymbol{w}_{t-1}\left\{\partial_{\boldsymbol{\theta}}\varepsilon_t(z_t,\boldsymbol{\theta}^*)\right\}^\top\right]\\
&= \lim_{t\to\infty} \mathbb{E}_{\boldsymbol{\theta}^*,\boldsymbol{\xi}^*}\left[\mathbb{E}_{\boldsymbol{\theta}^*,\boldsymbol{\xi}_s^*}\left[\boldsymbol{w}_{t-1}|s_{t-1}\right]\left\{\partial_{\boldsymbol{\theta}}\varepsilon_t(z_t,\boldsymbol{\theta}^*)\right\}^\top\right] = \mathbf{A}_{\tilde{w}},\\
\boldsymbol{\Sigma}_w &= \lim_{t\to\infty} \mathbb{E}_{\boldsymbol{\theta}^*,\boldsymbol{\xi}^*}\Big[(\varepsilon_t^*)^2\left(\mathbb{E}_{\boldsymbol{\theta}^*,\boldsymbol{\xi}_s^*}\left[\boldsymbol{w}_{t-1}|s_{t-1}\right] + \boldsymbol{w}_{t-1} - \mathbb{E}_{\boldsymbol{\theta}^*,\boldsymbol{\xi}_s^*}\left[\boldsymbol{w}_{t-1}|s_{t-1}\right]\right)\\
&\qquad\qquad\qquad \left(\mathbb{E}_{\boldsymbol{\theta}^*,\boldsymbol{\xi}_s^*}\left[\boldsymbol{w}_{t-1}|s_{t-1}\right] + \boldsymbol{w}_{t-1} - \mathbb{E}_{\boldsymbol{\theta}^*,\boldsymbol{\xi}_s^*}\left[\boldsymbol{w}_{t-1}|s_{t-1}\right]\right)^\top\Big]\\
&= \lim_{t\to\infty} \mathbb{E}_{\boldsymbol{\theta}^*,\boldsymbol{\xi}^*}\left[(\varepsilon_t^*)^2\mathbb{E}_{\boldsymbol{\theta}^*,\boldsymbol{\xi}_s^*}\left[\boldsymbol{w}_{t-1}|s_{t-1}\right]\mathbb{E}_{\boldsymbol{\theta}^*,\boldsymbol{\xi}_s^*}\left[\boldsymbol{w}_{t-1}|s_{t-1}\right]^\top\right]\\
&\quad + \lim_{t\to\infty} \mathbb{E}_{\boldsymbol{\theta}^*,\boldsymbol{\xi}^*}\left[(\varepsilon_t^*)^2\mathbb{E}_{\boldsymbol{\theta}^*,\boldsymbol{\xi}_s^*}\left[\boldsymbol{w}_{t-1}|s_{t-1}\right]\left(\boldsymbol{w}_{t-1} - \mathbb{E}_{\boldsymbol{\theta}^*,\boldsymbol{\xi}_s^*}\left[\boldsymbol{w}_{t-1}|s_{t-1}\right]\right)^\top\right]\\
&\quad + \lim_{t\to\infty} \mathbb{E}_{\boldsymbol{\theta}^*,\boldsymbol{\xi}^*}\left[(\varepsilon_t^*)^2\left(\boldsymbol{w}_{t-1} - \mathbb{E}_{\boldsymbol{\theta}^*,\boldsymbol{\xi}_s^*}\left[\boldsymbol{w}_{t-1}|s_{t-1}\right]\right)\left(\mathbb{E}_{\boldsymbol{\theta}^*,\boldsymbol{\xi}_s^*}\left[\boldsymbol{w}_{t-1}|s_{t-1}\right]\right)^\top\right]\\
&\quad + \lim_{t\to\infty} \mathbb{E}_{\boldsymbol{\theta}^*,\boldsymbol{\xi}^*}\left[(\varepsilon_t^*)^2\left(\boldsymbol{w}_{t-1} - \mathbb{E}_{\boldsymbol{\theta}^*,\boldsymbol{\xi}_s^*}\left[\boldsymbol{w}_{t-1}|s_{t-1}\right]\right)\left(\boldsymbol{w}_{t-1} - \mathbb{E}_{\boldsymbol{\theta}^*,\boldsymbol{\xi}_s^*}\left[\boldsymbol{w}_{t-1}|s_{t-1}\right]\right)^\top\right]\\
&= \boldsymbol{\Sigma}_{\tilde{\boldsymbol{w}}} + \lim_{t\to\infty} \mathbb{E}_{\boldsymbol{\theta}^*,\boldsymbol{\xi}^*}\left[(\varepsilon_t^*)^2\left(\boldsymbol{w}_{t-1} - \mathbb{E}_{\boldsymbol{\theta}^*,\boldsymbol{\xi}_s^*}\left[\boldsymbol{w}_{t-1}|s_{t-1}\right]\right)\left(\boldsymbol{w}_{t-1} - \mathbb{E}_{\boldsymbol{\theta}^*,\boldsymbol{\xi}_s^*}\left[\boldsymbol{w}_{t-1}|s_{t-1}\right]\right)^\top\right]\\
&= \boldsymbol{\Sigma}_{\tilde{\boldsymbol{w}}} + \lim_{t\to\infty} \mathbb{E}_{\boldsymbol{\theta}^*,\boldsymbol{\xi}^*}\left[(\varepsilon_t^*)^2\left(\boldsymbol{w}_{t-1} - \tilde{\boldsymbol{w}}_{t-1}\right)\left(\boldsymbol{w}_{t-1} - \tilde{\boldsymbol{w}}_{t-1}\right)^\top\right],
\end{aligned}
$$

where we have used $\mathbb{E}_{\boldsymbol{\theta}^*,\boldsymbol{\xi}_s^*}\left[\boldsymbol{w}_{t-1}|z_{t-1}\right] = \mathbb{E}_{\boldsymbol{\theta}^*,\boldsymbol{\xi}_s^*}\left[\boldsymbol{w}_{t-1}|s_{t-1}\right] = \tilde{\boldsymbol{w}}_{t-1}$ and

$$
\begin{aligned}
&\lim_{t\to\infty} \mathbb{E}_{\boldsymbol{\theta}^*,\boldsymbol{\xi}^*}\left[(\varepsilon_t^*)^2\left(\boldsymbol{w}_{t-1} - \mathbb{E}_{\boldsymbol{\theta}^*,\boldsymbol{\xi}_s^*}\left[\boldsymbol{w}_{t-1}|s_{t-1}\right]\right)\mathbb{E}_{\boldsymbol{\theta}^*,\boldsymbol{\xi}_s^*}\left[\boldsymbol{w}_{t-1}|s_{t-1}\right]^\top\right]\\
&= \lim_{t\to\infty} \mathbb{E}_{\boldsymbol{\theta}^*,\boldsymbol{\xi}^*}\left[(\varepsilon_t^*)^2\left(\mathbb{E}_{\boldsymbol{\theta}^*,\boldsymbol{\xi}_s^*}\left[\boldsymbol{w}_{t-1}|s_{t-1}\right] - \mathbb{E}_{\boldsymbol{\theta}^*,\boldsymbol{\xi}_s^*}\left[\boldsymbol{w}_{t-1}|s_{t-1}\right]\right)\left(\mathbb{E}_{\boldsymbol{\theta}^*,\boldsymbol{\xi}_s^*}\left[\boldsymbol{w}_{t-1}|s_{t-1}\right]\right)^\top\right] = \mathbf{0}.
\end{aligned}
$$

This implies that

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

where $\succeq$ denotes the semipositive definiteness of the subtraction. ■

Appendix G. Proof of Theorem 6

Proof As shown in Equation (16), the asymptotic variance of the estimator $\hat{\boldsymbol{\theta}}_w$ is given by

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

where

$$
\begin{aligned}
\mathbf{A}_w &\equiv \lim_{t\to\infty} \mathbb{E}_{\boldsymbol{\theta}^*,\boldsymbol{\xi}^*}[\boldsymbol{w}_{t-1}(Z_{t-1},\boldsymbol{\theta}^*)\{\partial_{\boldsymbol{\theta}}\varepsilon(z_t,\boldsymbol{\theta}^*)\}^\top],\\
\boldsymbol{\Sigma}_w &\equiv \lim_{t\to\infty} \mathbb{E}_{\boldsymbol{\theta}^*,\boldsymbol{\xi}^*}[\varepsilon(z_t,\boldsymbol{\theta}^*)^2\boldsymbol{w}_{t-1}(Z_{t-1},\boldsymbol{\theta}^*)\boldsymbol{w}_{t-1}(Z_{t-1},\boldsymbol{\theta}^*)^\top].
\end{aligned}
$$

For the sake of expression simplicity, the weight function $\boldsymbol{w}_t(Z_t,\boldsymbol{\theta}^*)$ the TD error $\varepsilon(z_t,\boldsymbol{\theta}^*)$ are abbreviated as $\boldsymbol{w}_t$ and ε_t, respectively; we rewrite $\mathbf{A}_w$ and Σ as

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

We first derive the weight function that minimizes the trace of the asymptotic variance, that is,

$$\boldsymbol{w}^*_{t-1} = \underset{\boldsymbol{w}_{t-1}}{\operatorname{argmin}}\, F\left(\boldsymbol{w}_{t-1}\right)$$
$$\text{where}\;\; F\left(\boldsymbol{w}_{t-1}\right) = \operatorname{tr}\{\operatorname{Av}\left(\boldsymbol{w}_{t-1}\right)\}.$$

Let $\boldsymbol{\delta}_{t-1}\equiv\boldsymbol{\delta}_{t-1}(Z_{t-1},\boldsymbol{\theta}^*)$ be an arbitrary function of Z_{t-1} and $\boldsymbol{\theta}^*$. We consider how much a functional $F(\boldsymbol{w}_{t-1})$ changes when we make a small change $h\boldsymbol{\delta}_{t-1}$ to the weight function $\boldsymbol{w}_{t-1}$. For notational convenience, we define $G(h;\boldsymbol{w}_{t-1},\boldsymbol{\delta}_{t-1})\equiv F\left(\boldsymbol{w}_{t-1}+h\boldsymbol{\delta}_{t-1}\right)$.[11] If the function $G(h;\boldsymbol{w}_{t-1},\boldsymbol{\delta}_{t-1})$ is twice differentiable with respect to h, then we have

$$G(h;\boldsymbol{w}_{t-1},\boldsymbol{\delta}_{t-1}) = G(0;\boldsymbol{w}_{t-1},\boldsymbol{\delta}_{t-1}) + h\,\partial_h G(h;\boldsymbol{w}_{t-1},\boldsymbol{\delta}_{t-1})|_{h=0} + O(h^2),$$

where ∂_h denotes the partial derivative with respect to h. Since the functional $F(\boldsymbol{w}_{t-1})$ is stationary for tiny variation in the function $\boldsymbol{w}_{t-1}$, the weight function $\boldsymbol{w}^*_{t-1}$ which minimizes the asymptotic estimation variance must satisfy

$$\partial_h G(h;\boldsymbol{w}^*_{t-1},\boldsymbol{\delta}_{t-1})\big|_{h=0} = 0,$$

for arbitrary choice of $\boldsymbol{\delta}_{t-1}$.

The definition of derivative says

$$\partial_h G(h;\boldsymbol{w}_{t-1},\boldsymbol{\delta}_{t-1})|_{h=0} = \lim_{\lambda\to0}\frac{G(\lambda;\boldsymbol{w}_{t-1},\boldsymbol{\delta}_{t-1}) - G(0;\boldsymbol{w}_{t-1},\boldsymbol{\delta}_{t-1})}{\lambda}.$$

The numerator of the above equation is written as

$$\begin{aligned} &G(\lambda;\boldsymbol{w}_{t-1},\boldsymbol{\delta}_{t-1}) - G(0;\boldsymbol{w}_{t-1},\boldsymbol{\delta}_{t-1}) \\ &= \operatorname{tr}\left[\mathbf{A}^{-1}_{w+\lambda\delta}\boldsymbol{\Sigma}_{w+\lambda\delta}(\mathbf{A}^{-1}_{w+\lambda\delta})^\top\right] - \operatorname{tr}\left[\mathbf{A}^{-1}_{w}\boldsymbol{\Sigma}_{w}(\mathbf{A}^{-1}_{w})^\top\right], \end{aligned} \tag{38}$$

where

$$\begin{aligned} \mathbf{A}_{w+\lambda\delta} &\equiv \lim_{t\to\infty}\mathbb{E}_{\boldsymbol{\theta}^*,\boldsymbol{\xi}^*}[(\boldsymbol{w}_{t-1}+\lambda\boldsymbol{\delta}_{t-1})\{\partial_{\boldsymbol{\theta}}\varepsilon(z_t,\boldsymbol{\theta}^*)\}^\top] \\ &= \mathbf{A}_w + \lambda\lim_{t\to\infty}\mathbb{E}_{\boldsymbol{\theta}^*,\boldsymbol{\xi}^*}[\boldsymbol{\delta}_{t-1}\{\partial_{\boldsymbol{\theta}}\varepsilon(z_t,\boldsymbol{\theta}^*)\}^\top] \end{aligned} \tag{39}$$

$$\begin{aligned} \boldsymbol{\Sigma}_{w+\lambda\delta} &\equiv \lim_{t\to\infty}\mathbb{E}_{\boldsymbol{\theta}^*,\boldsymbol{\xi}^*}[\varepsilon_t^2\left(\boldsymbol{w}_{t-1}+\lambda\boldsymbol{\delta}_{t-1}\right)\left(\boldsymbol{w}_{t-1}+\lambda\boldsymbol{\delta}_{t-1}\right)^\top] \\ &= \boldsymbol{\Sigma}_w + \lambda\lim_{t\to\infty}\mathbb{E}_{\boldsymbol{\theta}^*,\boldsymbol{\xi}^*}[\varepsilon_t^2(\boldsymbol{\delta}_{t-1}\boldsymbol{w}_{t-1}^\top + \boldsymbol{w}_{t-1}\boldsymbol{\delta}_{t-1}^\top)] + O(\lambda^2). \end{aligned} \tag{40}$$

[11] We used this notation to emphasize that $G(h;\boldsymbol{w}_{t-1},\boldsymbol{\delta}_{t-1})$ is a function of h, while $\boldsymbol{w}_{t-1}$ and $\boldsymbol{\delta}_{t-1}$ are regarded as auxiliary variables.

By using the matrix inversion lemma (Horn and Johnson, 1985), $\mathbf{A}_{w+\lambda\delta}^{-1}$ can be written as

$$\begin{aligned}
\mathbf{A}_{w+\lambda\delta}^{-1} &= \left(\mathbf{A}_w + \lambda \lim_{t\to\infty} \mathbb{E}_{\boldsymbol{\theta}^*,\boldsymbol{\xi}^*}[\boldsymbol{\delta}_{t-1}\{\partial_{\boldsymbol{\theta}}\varepsilon(z_t,\boldsymbol{\theta}^*)\}^\top]\right)^{-1} \\
&= \mathbf{A}_w^{-1} - \lim_{t\to\infty} \mathbf{A}_w^{-1}\left(\mathbf{I}+\lambda\mathbb{E}_{\boldsymbol{\theta}^*,\boldsymbol{\xi}^*}[\boldsymbol{\delta}_{t-1}\{\partial_{\boldsymbol{\theta}}\varepsilon(z_t,\boldsymbol{\theta}^*)\}^\top]\mathbf{A}_w^{-1}\right)^{-1}\lambda\mathbb{E}_{\boldsymbol{\theta}^*,\boldsymbol{\xi}^*}[\boldsymbol{\delta}_{t-1}\{\partial_{\boldsymbol{\theta}}\varepsilon(z_t,\boldsymbol{\theta}^*)\}^\top]\mathbf{A}_w^{-1}.
\end{aligned}$$

The matrix $\left(\mathbf{I}+\lambda\mathbb{E}_{\boldsymbol{\theta}^*,\boldsymbol{\xi}^*}[\boldsymbol{\delta}_{t-1}\{\partial_{\boldsymbol{\theta}}\varepsilon(z_t,\boldsymbol{\theta}^*)\}^\top]\mathbf{A}_w^{-1}\right)^{-1}$ can be calculated as

$$\left(\mathbf{I}+\lambda\mathbb{E}_{\boldsymbol{\theta}^*,\boldsymbol{\xi}^*}[\boldsymbol{\delta}_{t-1}\{\partial_{\boldsymbol{\theta}}\varepsilon(z_t,\boldsymbol{\theta}^*)\}^\top]\mathbf{A}_w^{-1}\right)^{-1} = \left(\mathbf{I}-\lambda\mathbb{E}_{\boldsymbol{\theta}^*,\boldsymbol{\xi}^*}[\boldsymbol{\delta}_{t-1}\{\partial_{\boldsymbol{\theta}}\varepsilon(z_t,\boldsymbol{\theta}^*)\}^\top]\mathbf{A}_w^{-1}\right) + O(\lambda^2),$$

because of

$$\left(\mathbf{I}+\lambda\mathbb{E}_{\boldsymbol{\theta}^*,\boldsymbol{\xi}^*}[\boldsymbol{\delta}_{t-1}\{\partial_{\boldsymbol{\theta}}\varepsilon(z_t,\boldsymbol{\theta}^*)\}^\top]\mathbf{A}_w^{-1}\right)\left(\mathbf{I}-\lambda\mathbb{E}_{\boldsymbol{\theta}^*,\boldsymbol{\xi}^*}[\boldsymbol{\delta}_{t-1}\{\partial_{\boldsymbol{\theta}}\varepsilon(z_t,\boldsymbol{\theta}^*)\}^\top]\mathbf{A}_w^{-1}\right) = \mathbf{I} + O\left(\lambda^2\right).$$

Thus we obtain

$$\mathbf{A}_{w+\lambda\delta}^{-1} = \mathbf{A}_w^{-1} - \lambda\lim_{t\to\infty}\mathbf{A}_w^{-1}\mathbb{E}_{\boldsymbol{\theta}^*,\boldsymbol{\xi}^*}\left[\boldsymbol{\delta}_{t-1}\{\partial_{\boldsymbol{\theta}}\varepsilon(z_t,\boldsymbol{\theta}^*)\}^\top\right]\mathbf{A}_w^{-1} + O(\lambda^2), \tag{41}$$

where high order terms are summarized as $O(\lambda^2)$.

Substituting Equations (39)-(41) to Equation (38), we have

$$\begin{aligned}
&\mathrm{tr}\left[\mathbf{A}_{w+\lambda\delta}^{-1}\boldsymbol{\Sigma}_{w+\lambda\delta}(\mathbf{A}_{w+\lambda\delta}^{-1})^\top\right] - \mathrm{tr}\left[\mathbf{A}_w^{-1}\boldsymbol{\Sigma}_w(\mathbf{A}_w^{-1})^\top\right] \\
&= -\lambda\lim_{t\to\infty}\mathrm{tr}\left[\mathbf{A}_w^{-1}\mathbb{E}_{\boldsymbol{\theta}^*,\boldsymbol{\xi}^*}\left[\boldsymbol{\delta}_{t-1}\{\partial_{\boldsymbol{\theta}}\varepsilon(z_t,\boldsymbol{\theta}^*)\}^\top\right]\mathbf{A}_w^{-1}\boldsymbol{\Sigma}_w\left(\mathbf{A}_w^{-1}\right)^\top\right] \\
&\quad -\lambda\lim_{t\to\infty}\mathrm{tr}\left[\mathbf{A}_w^{-1}\mathbb{E}_{\boldsymbol{\theta}^*,\boldsymbol{\xi}^*}\left[\partial_{\boldsymbol{\theta}}\varepsilon(z_t,\boldsymbol{\theta}^*)\boldsymbol{\delta}_{t-1}^\top\right]\mathbf{A}_w^{-1}\boldsymbol{\Sigma}_w\left(\mathbf{A}_w^{-1}\right)^\top\right] \\
&\quad +\lambda\lim_{t\to\infty}\mathrm{tr}\left[\mathbf{A}_w^{-1}\lim_{t\to\infty}\mathbb{E}_{\boldsymbol{\theta}^*,\boldsymbol{\xi}^*}[\varepsilon_t^2(\boldsymbol{\delta}_{t-1}\boldsymbol{w}_{t-1}^\top+\boldsymbol{w}_{t-1}\boldsymbol{\delta}_{t-1}^\top)]\left(\mathbf{A}_w^{-1}\right)^\top\right] + O(\lambda^2) \\
&= -2\lambda\lim_{t\to\infty}\mathrm{tr}\left[\mathbf{A}_w^{-1}\boldsymbol{\Sigma}_w(\mathbf{A}_w^{-1})^\top\mathbb{E}_{\boldsymbol{\theta}^*,\boldsymbol{\xi}^*}\left[\partial_{\boldsymbol{\theta}}\varepsilon(z_t,\boldsymbol{\theta}^*)\boldsymbol{\delta}_{t-1}^\top\right](\mathbf{A}_w^{-1})^\top\right] \\
&\quad +2\lambda\lim_{t\to\infty}\mathrm{tr}\left[\mathbf{A}_w^{-1}\mathbb{E}_{\boldsymbol{\theta}^*,\boldsymbol{\xi}^*}[\varepsilon_t^2\boldsymbol{w}_{t-1}\boldsymbol{\delta}_{t-1}^\top](\mathbf{A}_w^{-1})^\top\right] + O(\lambda^2).
\end{aligned}$$

This gives the partial derivative $\partial_h G(h;\boldsymbol{w}_{t-1},\boldsymbol{\delta}_{t-1})$ as

$$\begin{aligned}
&\partial_h G(h;\boldsymbol{w}_{t-1},\boldsymbol{\delta}_{t-1})|_{h=0} \\
&= -2\lim_{t\to\infty}\mathrm{tr}\left[\mathbf{A}_w^{-1}\boldsymbol{\Sigma}_w(\mathbf{A}_w^{-1})^\top\mathbb{E}_{\boldsymbol{\theta}^*,\boldsymbol{\xi}^*}\left[\partial_{\boldsymbol{\theta}}\varepsilon(z_t,\boldsymbol{\theta}^*)\boldsymbol{\delta}_{t-1}^\top\right](\mathbf{A}_w^{-1})^\top\right] \\
&\quad +2\lim_{t\to\infty}\mathrm{tr}\left[\mathbf{A}_w^{-1}\mathbb{E}_{\boldsymbol{\theta}^*,\boldsymbol{\xi}^*}[\varepsilon_t^2\boldsymbol{w}_{t-1}\boldsymbol{\delta}_{t-1}^\top](\mathbf{A}_w^{-1})^\top\right] \\
&= -2\lim_{t\to\infty}\mathbb{E}_{\boldsymbol{\theta}^*,\boldsymbol{\xi}^*}\left[\boldsymbol{\delta}_{t-1}^\top(\mathbf{A}_w^{-1})^\top\mathbf{A}_w^{-1}\boldsymbol{\Sigma}_w(\mathbf{A}_w^{-1})^\top\mathbb{E}_{\boldsymbol{\theta}^*,\boldsymbol{\xi}_s^*}[\partial_{\boldsymbol{\theta}}\varepsilon(z_t,\boldsymbol{\theta}^*)|Z_{t-1}]\right] \\
&\quad +2\lim_{t\to\infty}\mathbb{E}_{\boldsymbol{\theta}^*,\boldsymbol{\xi}^*}\left[\boldsymbol{\delta}_{t-1}^\top(\mathbf{A}_w^{-1})^\top\mathbf{A}_w^{-1}\mathbb{E}_{\boldsymbol{\theta}^*,\boldsymbol{\xi}_s^*}[\varepsilon_t^2|Z_{t-1}]\boldsymbol{w}_{t-1}\right] \\
&=2\lim_{t\to\infty}\mathbb{E}_{\boldsymbol{\theta}^*,\boldsymbol{\xi}^*}\left[\boldsymbol{\delta}_{t-1}^\top(\mathbf{A}_w^{-1})^\top\mathbf{A}_w^{-1}\left\{\mathbb{E}_{\boldsymbol{\theta}^*,\boldsymbol{\xi}_s^*}[\varepsilon_t^2|s_{t-1}]\boldsymbol{w}_{t-1}-\boldsymbol{\Sigma}_w(\mathbf{A}_w^{-1})^\top\mathbb{E}_{\boldsymbol{\theta}^*,\boldsymbol{\xi}_s^*}[\partial_{\boldsymbol{\theta}}\varepsilon(z_t,\boldsymbol{\theta}^*)|s_{t-1}]\right\}\right].
\end{aligned}$$

By applying the condition that the deviation becomes $\mathbf{0}$ for any function $\boldsymbol{\delta}_{t-1}(Z_{t-1},\boldsymbol{\theta}^*)$, the optimal weight function is obtained as

$$\boldsymbol{w}^*_{t-1} = \mathbb{E}_{\boldsymbol{\theta}^*,\boldsymbol{\xi}^*_s}[(\varepsilon(z_t,\boldsymbol{\theta}^*)^2|s_{t-1}]^{-1}\boldsymbol{\Sigma}_w(\mathbf{A}_w^{-1})^\top\mathbb{E}_{\boldsymbol{\theta}^*,\boldsymbol{\xi}^*_s}[\partial_{\boldsymbol{\theta}}\varepsilon(z_t,\boldsymbol{\theta}^*)|s_{t-1}].$$

Because any estimating function is invariant to transformation applied by any regular matrix, the optimal estimating function is restricted as

$$\boldsymbol{w}^*_{t-1} = \boldsymbol{w}^*_{t-1}(Z_{t-1},\boldsymbol{\theta}^*) = \mathbb{E}_{\boldsymbol{\theta}^*,\boldsymbol{\xi}^*_s}\left[\varepsilon(z_t,\boldsymbol{\theta}^*)^2|s_{t-1}\right]^{-1}\mathbb{E}_{\boldsymbol{\theta}^*,\boldsymbol{\xi}^*_s}[\partial_{\boldsymbol{\theta}}\varepsilon(z_t,\boldsymbol{\theta}^*)|s_{t-1}],$$

or its transformation applied by any regular matrix.

Now, we confirm that the estimator obtained by Equation (18) yields the minimum asymptotic variance. Substituting $\boldsymbol{w}^*_{t-1}$ to the matrix $\mathbf{A}_w$, some calculations in Appendix H lead us to

$$\mathbf{A}_{w^*} = \boldsymbol{\Sigma}_{w^*} = \mathbf{Q},$$

where

$$\mathbf{Q} \equiv \lim_{t\to\infty}\mathbb{E}_{\boldsymbol{\theta}^*,\boldsymbol{\xi}^*}\left[\mathbb{E}_{\boldsymbol{\theta}^*,\boldsymbol{\xi}^*_s}[\varepsilon_t^2|s_t]^{-1}\mathbb{E}_{\boldsymbol{\theta}^*,\boldsymbol{\xi}^*_s}[\partial_{\boldsymbol{\theta}}\varepsilon(z_t,\boldsymbol{\theta}^*)|s_{t-1}]\mathbb{E}_{\boldsymbol{\theta}^*,\boldsymbol{\xi}^*_s}[\partial_{\boldsymbol{\theta}}\varepsilon(z_t,\boldsymbol{\theta}^*)|s_{t-1}]^\top\right].$$

We consider how much the asymptotic variance Av changes when we make a small change $\bar{\boldsymbol{\delta}}_{t-1} \equiv h\boldsymbol{\delta}_{t-1}$ on $\boldsymbol{w}^*_{t-1}$. The matrices at $\boldsymbol{w}^*_{t-1}+\bar{\boldsymbol{\delta}}_{t-1}$ become

$$\begin{aligned}\mathbf{A}_{w^*+\bar{\delta}} &= \mathbf{Q} + \lim_{t\to\infty}\mathbb{E}_{\boldsymbol{\theta}^*,\boldsymbol{\xi}^*}[\bar{\boldsymbol{\delta}}_{t-1}\partial_{\boldsymbol{\theta}}\varepsilon(z_t,\boldsymbol{\theta}^*)^\top],\\ \boldsymbol{\Sigma}_{w^*+\bar{\delta}} &= \mathbf{Q} + \lim_{t\to\infty}\mathbb{E}_{\boldsymbol{\theta}^*,\boldsymbol{\xi}^*}\left[\partial_{\boldsymbol{\theta}}\varepsilon(z_t,\boldsymbol{\theta}^*)\bar{\boldsymbol{\delta}}^\top_{t-1}\right]\\ &\quad + \lim_{t\to\infty}\mathbb{E}_{\boldsymbol{\theta}^*,\boldsymbol{\xi}^*}\left[\bar{\boldsymbol{\delta}}_{t-1}\{\partial_{\boldsymbol{\theta}}\varepsilon(z_t,\boldsymbol{\theta}^*)\}^\top\right] + \lim_{t\to\infty}\mathbb{E}_{\boldsymbol{\theta}^*,\boldsymbol{\xi}^*}\left[\varepsilon_t^2\bar{\boldsymbol{\delta}}_{t-1}\bar{\boldsymbol{\delta}}^\top_{t-1}\right].\end{aligned}$$

Therefore,

$$\begin{aligned}&\mathbf{A}^{-1}_{w^*+\bar{\delta}}\boldsymbol{\Sigma}_{w^*+\bar{\delta}}\left(\mathbf{A}^{-1}_{w^*+\bar{\delta}}\right)^\top - \mathbf{A}^{-1}_{w^*}\boldsymbol{\Sigma}_{w^*}\left(\mathbf{A}^{-1}_{w^*}\right)^\top\\ &= \mathbf{A}^{-1}_{w^*+\bar{\delta}}\underbrace{\left(\boldsymbol{\Sigma}_{w^*+\bar{\delta}} - \mathbf{A}_{w^*+\bar{\delta}}\mathbf{A}^{-1}_{w^*}\boldsymbol{\Sigma}_{w^*}(\mathbf{A}^{-1}_{w^*})^\top\mathbf{A}^\top_{w^*+\bar{\delta}}\right)}_{\mathbf{C}_1}\left(\mathbf{A}^{-1}_{w^*+\bar{\delta}}\right)^\top.\end{aligned}$$

The matrix $\mathbf{C}_1$ is a semipositive definite matrix, because

$$\begin{aligned}\mathbf{C}_1 =&\boldsymbol{\Sigma}_{w^*+\bar{\delta}} - \mathbf{A}_{w^*+\bar{\delta}}\mathbf{A}^{-1}_{w^*}\boldsymbol{\Sigma}_{w^*}(\mathbf{A}^{-1}_{w^*})^\top\mathbf{A}^\top_{w^*+\bar{\delta}}\\ =&\lim_{t\to\infty}\mathbb{E}_{\boldsymbol{\theta}^*,\boldsymbol{\xi}^*}\left[\varepsilon_t^2\bar{\boldsymbol{\delta}}_{t-1}\bar{\boldsymbol{\delta}}^\top_{t-1}\right] - \lim_{t\to\infty}\mathbb{E}_{\boldsymbol{\theta}^*,\boldsymbol{\xi}^*}\left[\bar{\boldsymbol{\delta}}_{t-1}\{\partial_{\boldsymbol{\theta}}\varepsilon(z_t,\boldsymbol{\theta}^*)\}^\top\right]\mathbf{Q}^{-1}\mathbb{E}_{\boldsymbol{\theta}^*,\boldsymbol{\xi}^*}\left[\partial_{\boldsymbol{\theta}}\varepsilon(z_t,\boldsymbol{\theta}^*)\bar{\boldsymbol{\delta}}^\top_{t-1}\right]\\ =&\lim_{t\to\infty}\mathbb{E}_{\boldsymbol{\theta}^*,\boldsymbol{\xi}^*}\left[\left(\bar{\boldsymbol{\delta}}_{t-1}-\boldsymbol{\nu}_{t-1}\right)\left(\bar{\boldsymbol{\delta}}_{t-1}-\boldsymbol{\nu}_{t-1}\right)^\top\right]\succeq\mathbf{0},\end{aligned}$$

where

$$\boldsymbol{\nu}_{t-1} \equiv \mathbb{E}_{\boldsymbol{\theta}^*,\boldsymbol{\xi}^*_s}[\varepsilon_t^2|s_{t-1}]^{-1}\mathbb{E}_{\boldsymbol{\theta}^*,\boldsymbol{\xi}^*}\left[\bar{\boldsymbol{\delta}}_{t-1}\{\partial_{\boldsymbol{\theta}}\varepsilon(z_t,\boldsymbol{\theta}^*)\}^\top\right]\mathbf{Q}^{-1}\mathbb{E}_{\boldsymbol{\theta}^*,\boldsymbol{\xi}^*_s}\left[\partial_{\boldsymbol{\theta}}\varepsilon(z_t,\boldsymbol{\theta}^*)|s_{t-1}\right].$$

Thus we have

$$\mathbf{A}^{-1}_{w^*+\bar{\delta}}\boldsymbol{\Sigma}_{w^*+\bar{\delta}}\left(\mathbf{A}^{-1}_{w^*+\bar{\delta}}\right)^\top - \mathbf{A}^{-1}_{w^*}\boldsymbol{\Sigma}_{w^*}\left(\mathbf{A}^{-1}_{w^*}\right)^\top\succeq\mathbf{0},$$

where $\succeq$ denotes the semipositive definiteness of the subtraction. The equality in the above equation holds only when $\bar{\boldsymbol{\delta}}_{t-1}\propto\boldsymbol{w}^*_{t-1}$. ■

Appendix H. Proof of Lemma 7

Proof The matrix $\mathbf{A}$ in the asymptotic variance given by Equation (16) can be calculated as

$$\begin{aligned}
\mathbf{A} &= \lim_{t\to\infty} \mathbb{E}_{\boldsymbol{\theta}^*,\boldsymbol{\xi}^*}\left[\partial_{\boldsymbol{\theta}} \boldsymbol{\psi}_t^*(z_t,\boldsymbol{\theta}^*)\right] \\
&= \lim_{t\to\infty} \mathbb{E}_{\boldsymbol{\theta}^*,\boldsymbol{\xi}^*}\left[\boldsymbol{w}_{t-1}^*(Z_{t-1},\boldsymbol{\theta}^*)\left\{\partial_{\boldsymbol{\theta}}\varepsilon(z_t,\boldsymbol{\theta})\right\}^\top\right] \\
&= \lim_{t\to\infty} \mathbb{E}_{\boldsymbol{\theta}^*,\boldsymbol{\xi}^*}\Big[\mathbb{E}_{\boldsymbol{\theta}^*,\boldsymbol{\xi}_s^*}[\varepsilon(z_t,\boldsymbol{\theta}^*)^2|s_{t-1}]^{-1} \\
&\qquad\qquad \mathbb{E}_{\boldsymbol{\theta}^*,\boldsymbol{\xi}_s^*}[\partial_{\boldsymbol{\theta}}\varepsilon(z_t,\boldsymbol{\theta}^*)|s_{t-1}]\mathbb{E}_{\boldsymbol{\theta}^*,\boldsymbol{\xi}_s^*}[\partial_{\boldsymbol{\theta}}\varepsilon(z_t,\boldsymbol{\theta}^*)|s_{t-1}]^\top\Big].
\end{aligned}$$

Also the matrix $\boldsymbol{\Sigma}$ can be calculated as

$$\begin{aligned}
\boldsymbol{\Sigma} &= \lim_{t\to\infty} \mathbb{E}_{\boldsymbol{\theta}^*,\boldsymbol{\xi}^*}\left[\boldsymbol{\psi}_t^*(z_t,\boldsymbol{\theta}^*)\boldsymbol{\psi}_t^*(z_t,\boldsymbol{\theta}^*)^\top\right] \\
&= \lim_{t\to\infty} \mathbb{E}_{\boldsymbol{\theta}^*,\boldsymbol{\xi}^*}\left[\mathbb{E}_{\boldsymbol{\theta}^*,\boldsymbol{\xi}_s^*}\left[\varepsilon(z_t,\boldsymbol{\theta}^*)^2|s_{t-1}\right]\boldsymbol{w}_{t-1}^*(Z_{t-1},\boldsymbol{\theta}^*)\{\boldsymbol{w}_{t-1}^*(Z_{t-1},\boldsymbol{\theta}^*)\}^\top\right] \\
&= \lim_{t\to\infty} \mathbb{E}_{\boldsymbol{\theta}^*,\boldsymbol{\xi}^*}\Bigg[\mathbb{E}_{\boldsymbol{\theta}^*,\boldsymbol{\xi}_s^*}\left[\varepsilon(z_t,\boldsymbol{\theta}^*)^2|s_{t-1}\right] \\
&\qquad \left\{\mathbb{E}_{\boldsymbol{\theta}^*,\boldsymbol{\xi}_s^*}\left[\varepsilon(z_t,\boldsymbol{\theta}^*)^2|s_{t-1}\right]^{-1}\mathbb{E}_{\boldsymbol{\theta}^*,\boldsymbol{\xi}_s^*}[\partial_{\boldsymbol{\theta}}\varepsilon(z_t,\boldsymbol{\theta}^*)|s_{t-1}]\right\} \\
&\qquad \left\{\mathbb{E}_{\boldsymbol{\theta}^*,\boldsymbol{\xi}_s^*}\left[\varepsilon(z_t,\boldsymbol{\theta}^*)^2|s_{t-1}\right]^{-1}\mathbb{E}_{\boldsymbol{\theta}^*,\boldsymbol{\xi}_s^*}[\partial_{\boldsymbol{\theta}}\varepsilon(z_t,\boldsymbol{\theta}^*)|s_{t-1}]\right\}^\top\Bigg] \\
&= \lim_{t\to\infty} \mathbb{E}_{\boldsymbol{\theta}^*,\boldsymbol{\xi}^*}\Big[\mathbb{E}_{\boldsymbol{\theta}^*,\boldsymbol{\xi}_s^*}[\varepsilon(z_t,\boldsymbol{\theta}^*)^2|s_{t-1}]^{-1} \\
&\qquad \mathbb{E}_{\boldsymbol{\theta}^*,\boldsymbol{\xi}_s^*}[\partial_{\boldsymbol{\theta}}\varepsilon(z_t,\boldsymbol{\theta}^*)|s_{t-1}]\mathbb{E}_{\boldsymbol{\theta}^*,\boldsymbol{\xi}_s^*}[\partial_{\boldsymbol{\theta}}\varepsilon(z_t,\boldsymbol{\theta}^*)|s_{t-1}]^\top\Big] = \mathbf{A} = \mathbf{Q}.
\end{aligned}$$

These observations yield

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

■

Appendix I. Proof of Theorem 8

Proof 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

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

From Assumption 5, the second term of this equation is bounded by the second moment, thus we obtain

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{\theta}^*,\boldsymbol{\xi}_s^*}\left[h_{t+1}-(1+\eta_{t+1}^2c_2)h_t|s_t\right] \\
&\leq -2\eta_{t+1}\hat{\boldsymbol{\theta}}_t^\top\mathbf{R}(\hat{\boldsymbol{\theta}}_t)\mathbb{E}_{\boldsymbol{\theta}^*,\boldsymbol{\xi}_s^*}\left[\boldsymbol{\psi}_{t+1}(Z_{t+1},\hat{\boldsymbol{\theta}}_t)\middle| s_t\right]+\eta_{t+1}^2c_1. \qquad (42)
\end{aligned}$$

Now, let $\chi_t = \prod_{k=1}^{t} 1/(1+\eta_k^2 c_2)$ 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 the both sides of Equation (42) by χ_{t+1}, we obtain

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

The first term of this upper bound is negative because of Assumption 5, the second term is nonnegative because η_t, χ_{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 (Neveu, 1975; Bertsekas and Tsitsiklis, 1996, Proposition 4.2) guarantees that h'_t converges to a nonnegative random variable almost surely, and $\sum_{t=1}^{\infty} \eta_{t+1}\chi_{t+1}\hat{\boldsymbol{\theta}}_t^\top \mathbf{R}_t(\hat{\boldsymbol{\theta}}_t)\mathbb{E}_{\boldsymbol{\theta}^*,\boldsymbol{\xi}_s^*}\left[\boldsymbol{\psi}_{t+1}(Z_{t+1},\hat{\boldsymbol{\theta}}_t)\middle| s_t\right] < \infty$. Since $\sum_{t=1}^{\infty} \eta_t = \infty$ and $\lim_{t\to\infty} \chi_t = \chi_\infty > 0$, we have $\hat{\boldsymbol{\theta}}_t^\top \mathbf{R}(\hat{\boldsymbol{\theta}}_t)\mathbb{E}_{\boldsymbol{\theta}^*,\boldsymbol{\xi}_s^*}\left[\boldsymbol{\psi}_{t+1}(Z_{t+1},\hat{\boldsymbol{\theta}}_t)\middle| s_t\right] \xrightarrow{a.s.} \mathbf{0}$. This result suggests the conclusion that the online learning algorithm converges to the true parameter almost surely: $\hat{\boldsymbol{\theta}}_t \xrightarrow{a.s.} \boldsymbol{\theta}^* = \mathbf{0}$. ■

Appendix J. Proof of Lemma 9

Proof Using Taylor series expansion of the estimating equation $(1/t)\sum_{i=1}^{t} \boldsymbol{\psi}_i(Z_i, \tilde{\boldsymbol{\theta}}_t)$ around $\tilde{\boldsymbol{\theta}}_{t-1}$, we obtain

$$\begin{aligned}\frac{1}{t}\sum_{i=1}^{t} \boldsymbol{\psi}_i(Z_i, \tilde{\boldsymbol{\theta}}_t) =& \frac{1}{t}\sum_{i=1}^{t} \boldsymbol{\psi}_i(Z_i, \tilde{\boldsymbol{\theta}}_{t-1})\\ &+ \frac{1}{t}\sum_{i=1}^{t} \partial_{\boldsymbol{\theta}} \boldsymbol{\psi}_i(Z_i, \tilde{\boldsymbol{\theta}}_{t-1})(\tilde{\boldsymbol{\theta}}_t - \tilde{\boldsymbol{\theta}}_{t-1}) + O_p\left(\|\tilde{\boldsymbol{\theta}}_t - \tilde{\boldsymbol{\theta}}_{t-1}\|^2\right).\end{aligned}$$

Since $\sum_{i=1}^{t} \boldsymbol{\psi}_i(Z_i, \tilde{\boldsymbol{\theta}}_t) = \sum_{i=1}^{t-1} \boldsymbol{\psi}_i(Z_i, \tilde{\boldsymbol{\theta}}_{t-1}) = \mathbf{0}$, we obtain the following equation:

$$-\frac{1}{t}\boldsymbol{\psi}_t(Z_t, \tilde{\boldsymbol{\theta}}_{t-1}) = \tilde{\mathbf{R}}_t(\tilde{\boldsymbol{\theta}}_{t-1})(\tilde{\boldsymbol{\theta}}_t - \tilde{\boldsymbol{\theta}}_{t-1}) + O_p\left(\|\tilde{\boldsymbol{\theta}}_t - \tilde{\boldsymbol{\theta}}_{t-1}\|^2\right).$$

We can then rewrite the right hand side as

$$-\frac{1}{t}\boldsymbol{\psi}_t(Z_t, \tilde{\boldsymbol{\theta}}_{t-1}) = \{\tilde{\mathbf{R}}_t(\tilde{\boldsymbol{\theta}}_{t-1}) + O_p(\|\tilde{\boldsymbol{\theta}}_t - \tilde{\boldsymbol{\theta}}_{t-1}\|)\}(\tilde{\boldsymbol{\theta}}_t - \tilde{\boldsymbol{\theta}}_{t-1}),$$

and

$$(\tilde{\boldsymbol{\theta}}_t - \tilde{\boldsymbol{\theta}}_{t-1}) = -\frac{1}{t}\{\tilde{\mathbf{R}}_t^{-1}(\tilde{\boldsymbol{\theta}}_{t-1}) + O_p(\|\tilde{\boldsymbol{\theta}}_t - \tilde{\boldsymbol{\theta}}_{t-1}\|)\}\boldsymbol{\psi}_t(Z_t, \tilde{\boldsymbol{\theta}}_{t-1}).$$

Note that $\tilde{\mathbf{R}}_t^{-1}(\tilde{\boldsymbol{\theta}}_{t-1})$ is uniformly bounded because of the nonsingular condition in Lemma 9. Also $\tilde{\boldsymbol{\theta}}_t$ is uniformly bounded for any t. Furthermore, $\boldsymbol{\psi}_t(Z_t, \tilde{\boldsymbol{\theta}}_{t-1})$ is uniformly bounded for any t since the conditions in Assumptions 2-4 imply that $\boldsymbol{\psi}_t(Z_t, \tilde{\boldsymbol{\theta}}_{t-1})$ is a continuous function of uniformly bounded variables. Hence, the above equation implies that $\tilde{\boldsymbol{\theta}}_t - \tilde{\boldsymbol{\theta}}_{t-1} = O_p(1/t)$. Therefore, we can obtain the following equation

$$-\frac{1}{t}\boldsymbol{\psi}_t(Z_t, \tilde{\boldsymbol{\theta}}_{t-1}) = \tilde{\mathbf{R}}_t(\tilde{\boldsymbol{\theta}}_{t-1})(\tilde{\boldsymbol{\theta}}_t - \tilde{\boldsymbol{\theta}}_{t-1}) + O_p\left(\frac{1}{t^2}\right).$$

By using the matrix inversion operation, we derive

$$\tilde{\boldsymbol{\theta}}_t = \tilde{\boldsymbol{\theta}}_{t-1} - \frac{1}{t}\tilde{\mathbf{R}}_t^{-1}\psi_t(Z_t,\tilde{\boldsymbol{\theta}}_{t-1}) + O_p\left(\frac{1}{t^2}\right).$$

■

Appendix K. Proof of Theorem 10

Proof Similar to the proof in Appendix I, we assume the true parameter is located at the origin: $\boldsymbol{\theta}^* = \mathbf{0}$. From the assumption in Theorem 10, the online learning converges to the true parameter almost surely; this implies that $\hat{\boldsymbol{\theta}}_t = \boldsymbol{\theta}^* + o_p(1) = o_p(1)$. Note also that $\mathbf{R}_t$ converges to $\mathbf{A}$ almost surely; this implies that $\mathbf{R}_t = \mathbf{A} + o_p(1)$. Furthermore, from condition (d) in Theorem 10, the matrix $\mathbf{R}_t$ is invertible for any t; this implies that $\mathbf{R}_t^{-1} = \mathbf{A}^{-1} + o_p(1)$.

Using Equation (23), $(\hat{\boldsymbol{\theta}}_t - \boldsymbol{\theta}^*)(\hat{\boldsymbol{\theta}}_t - \boldsymbol{\theta}^*)^\top = \hat{\boldsymbol{\theta}}_t\hat{\boldsymbol{\theta}}_t^\top$ can be expressed as

$$\begin{aligned}
\hat{\boldsymbol{\theta}}_t\hat{\boldsymbol{\theta}}_t^\top = &\left(\hat{\boldsymbol{\theta}}_{t-1} - \frac{1}{t}\hat{\mathbf{R}}_t^{-1}\psi_t(Z_t,\hat{\boldsymbol{\theta}}_{t-1}) + O_p\left(\frac{1}{t^2}\right)\right)\\
&\left(\hat{\boldsymbol{\theta}}_{t-1} - \frac{1}{t}\hat{\mathbf{R}}_t^{-1}\psi_t(Z_t,\hat{\boldsymbol{\theta}}_{t-1}) + O_p\left(\frac{1}{t^2}\right)\right)^\top\\
=&\hat{\boldsymbol{\theta}}_{t-1}\hat{\boldsymbol{\theta}}_{t-1}^\top - \frac{1}{t}\hat{\boldsymbol{\theta}}_{t-1}\psi_t(Z_t,\hat{\boldsymbol{\theta}}_{t-1})^\top(\hat{\mathbf{R}}_t^{-1})^\top - \frac{1}{t}\hat{\mathbf{R}}_t^{-1}\psi_t(Z_t,\hat{\boldsymbol{\theta}}_{t-1})\hat{\boldsymbol{\theta}}_{t-1}^\top\\
&+ \frac{1}{t^2}\hat{\mathbf{R}}_t^{-1}\psi_t(Z_t,\hat{\boldsymbol{\theta}}_{t-1})\psi_t(Z_t,\hat{\boldsymbol{\theta}}_{t-1})^\top(\hat{\mathbf{R}}_t^{-1})^\top + o_p\left(\frac{1}{t^2}\right),
\end{aligned}$$

where high order terms are in total represented as $o_p\left(1/t^2\right)$ because of
$\hat{\boldsymbol{\theta}}_{t-1}O_p(1/t^2) = o_p(1)O_p(1/t^2) = o_p(1/t^2)$. Taking the conditional expectation of $\hat{\boldsymbol{\theta}}_t\hat{\boldsymbol{\theta}}_t^\top$ given Z_{t-1}, we obtain

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{\theta}^*,\boldsymbol{\xi}_s^*}\left[\hat{\boldsymbol{\theta}}_t\hat{\boldsymbol{\theta}}_t^\top|Z_{t-1}\right] =&\hat{\boldsymbol{\theta}}_{t-1}\hat{\boldsymbol{\theta}}_{t-1}^\top - \frac{1}{t}\underbrace{\hat{\boldsymbol{\theta}}_{t-1}\mathbb{E}_{\boldsymbol{\theta}^*,\boldsymbol{\xi}_s^*}\left[\psi_t(Z_t,\hat{\boldsymbol{\theta}}_{t-1})^\top(\hat{\mathbf{R}}_t^{-1})^\top\middle|Z_{t-1}\right]}_{\mathbf{C}_1^\top}\\
&- \frac{1}{t}\underbrace{\mathbb{E}_{\boldsymbol{\theta}^*,\boldsymbol{\xi}_s^*}\left[\hat{\mathbf{R}}_t^{-1}\psi_t(Z_t,\hat{\boldsymbol{\theta}}_{t-1})\middle|Z_{t-1}\right]\hat{\boldsymbol{\theta}}_{t-1}^\top}_{\mathbf{C}_1}\\
&+ \frac{1}{t^2}\underbrace{\mathbb{E}_{\boldsymbol{\theta}^*,\boldsymbol{\xi}_s^*}\left[\hat{\mathbf{R}}_t^{-1}\psi_t(Z_t,\hat{\boldsymbol{\theta}}_{t-1})\psi_t(Z_t,\hat{\boldsymbol{\theta}}_{t-1})^\top(\hat{\mathbf{R}}_t^{-1})^\top\middle|Z_{t-1}\right]}_{\mathbf{C}_2} + o_p\left(\frac{1}{t^2}\right).
\end{aligned}$$

We now express each of the terms in the above equation.

In order to express $\mathbf{C}_1$ and $\mathbf{C}_2$, we introduce the following lemma.

Lemma 21 *(Bottou and LeCun, 2005, Theorem 4) Let X_t be a uniformly bounded random variable depending on Z_t. Then we have*

$$\mathbb{E}_{\boldsymbol{\theta}^*,\boldsymbol{\xi}_s^*}\left[\hat{\mathbf{R}}_t^{-1}X_t\middle|Z_{t-1}\right] = \mathbb{E}_{\boldsymbol{\theta}^*,\boldsymbol{\xi}_s^*}\left[\hat{\mathbf{R}}_t^{-1}\middle|Z_{t-1}\right]\mathbb{E}_{\boldsymbol{\theta}^*,\boldsymbol{\xi}_s^*}\left[X_t\middle|Z_{t-1}\right] + o_p\left(\frac{1}{t}\right).$$

Proof By using assumption (b) in Theorem 10, $\mathbb{E}_{\boldsymbol{\theta}^*,\boldsymbol{\xi}_s^*}[\hat{\mathbf{R}}_t^{-1}X_t|Z_{t-1}]$ can be calculated as

$$\mathbb{E}_{\boldsymbol{\theta}^*,\boldsymbol{\xi}_s^*}\left[\hat{\mathbf{R}}_t^{-1}X_t|Z_{t-1}\right] = \mathbb{E}_{\boldsymbol{\theta}^*,\boldsymbol{\xi}_s^*}[\hat{\mathbf{R}}_t^{-1}|Z_{t-1}]\mathbb{E}_{\boldsymbol{\theta}^*,\boldsymbol{\xi}_s^*}\left[X_t|Z_{t-1}\right] + \mathbb{E}_{\boldsymbol{\theta}^*,\boldsymbol{\xi}_s^*}\left[\varepsilon_t(Z_t)X_t|Z_{t-1}\right],$$

where $\varepsilon_t(Z_t) = o_p(1/t)$. $\mathbb{E}_{\boldsymbol{\theta}^*,\boldsymbol{\xi}_s^*}[\varepsilon_t(Z_t)X_t|Z_{t-1}]$ is summarized as $o_p(1/t)$ because of the Cauchy-Schwartz's inequality:

$$\mathbb{E}_{\boldsymbol{\theta}^*,\boldsymbol{\xi}_s^*}[\varepsilon_t(Z_t)X_t|Z_{t-1}] \leq \sqrt{\mathbb{E}_{\boldsymbol{\theta}^*,\boldsymbol{\xi}_s^*}[\,\|\varepsilon_t(Z_t)\|^2|\,Z_{t-1}]}\sqrt{\mathbb{E}_{\boldsymbol{\theta}^*,\boldsymbol{\xi}_s^*}[\,\|X_t\|^2|\,Z_{t-1}]}.$$

■

Since condition (a) in Theorem 10 and Assumptions 2-4 lead $\psi_t(Z_t,\hat{\boldsymbol{\theta}}_{t-1})$ and $\hat{\mathbf{R}}_t$ to be continuous functions of uniformly bounded variables, $\psi_t(Z_t,\hat{\boldsymbol{\theta}}_{t-1})$ and $\hat{\mathbf{R}}_t$ are uniformly bounded for any t. Then, using Lemma 21, $\mathbf{C}_2$ can be expressed as

$$\begin{aligned}\mathbf{C}_2 = \Bigg\{ & \mathbb{E}_{\boldsymbol{\theta}^*,\boldsymbol{\xi}_s^*}\left[\hat{\mathbf{R}}_t^{-1}\middle|\,Z_{t-1}\right] \\ & \mathbb{E}_{\boldsymbol{\theta}^*,\boldsymbol{\xi}_s^*}\left[\psi_t(Z_t,\hat{\boldsymbol{\theta}}_{t-1})\psi_t(Z_t,\hat{\boldsymbol{\theta}}_{t-1})^\top\middle|\,Z_{t-1}\right]\mathbb{E}_{\boldsymbol{\theta}^*,\boldsymbol{\xi}_s^*}\left[(\hat{\mathbf{R}}_t^{-1})^\top\middle|\,Z_{t-1}\right]\Bigg\} + o_p\left(\frac{1}{t}\right).\end{aligned}$$

We note that $\mathbb{E}_{\boldsymbol{\theta}^*,\boldsymbol{\xi}_s^*}[\psi_t(Z_t,\hat{\boldsymbol{\theta}}_{t-1})\psi_t(Z_t,\hat{\boldsymbol{\theta}}_{t-1})^\top|Z_{t-1}]$ can be calculated as

$$\mathbb{E}_{\boldsymbol{\theta}^*,\boldsymbol{\xi}_s^*}\left[\psi_t(Z_t,\hat{\boldsymbol{\theta}}_{t-1})\psi_t(Z_t,\hat{\boldsymbol{\theta}}_{t-1})^\top\middle|\,Z_{t-1}\right] = \mathbb{E}_{\boldsymbol{\theta}^*,\boldsymbol{\xi}_s^*}\left[\psi_t(Z_t,\boldsymbol{\theta}^*)\psi_t(Z_t,\boldsymbol{\theta}^*)^\top\middle|\,Z_{t-1}\right] + o_p(1),$$

because $\hat{\boldsymbol{\theta}}_{t-1}$ converges to the true parameter and $\psi_t(Z_t,\boldsymbol{\theta})$ is uniformly bounded. Since $\hat{\mathbf{R}}_t^{-1} = \mathbf{A}^{-1} + o_p(1)$ is satisfied, $\mathbf{C}_2$ can be rewritten as

$$\mathbf{C}_2 = \mathbf{A}^{-1}\mathbb{E}_{\boldsymbol{\theta}^*,\boldsymbol{\xi}_s^*}\left[\psi_t(Z_t,\boldsymbol{\theta}^*)\psi_t(Z_t,\boldsymbol{\theta}^*)^\top\middle|\,Z_{t-1}\right](\mathbf{A}^{-1})^\top + o_p(1). \tag{43}$$

Using similar arguments, $\mathbf{C}_1$ can be expressed as

$$\begin{aligned}\mathbf{C}_1 &= \mathbb{E}_{\boldsymbol{\theta}^*,\boldsymbol{\xi}_s^*}\left[\hat{\mathbf{R}}_t^{-1}\psi_t(Z_t,\hat{\boldsymbol{\theta}}_{t-1})\middle|\,Z_{t-1}\right]\hat{\boldsymbol{\theta}}_{t-1} \\ &= \mathbb{E}_{\boldsymbol{\theta}^*,\boldsymbol{\xi}_s^*}\left[\hat{\mathbf{R}}_t^{-1}\middle|\,Z_{t-1}\right]\mathbb{E}_{\boldsymbol{\theta}^*,\boldsymbol{\xi}_s^*}\left[\psi_t(Z_t,\hat{\boldsymbol{\theta}}_{t-1})|Z_{t-1}\right]\hat{\boldsymbol{\theta}}_{t-1}^\top + o_p\left(\frac{1}{t}\right).\end{aligned}$$

We now consider $\mathbb{E}_{\boldsymbol{\theta}^*,\boldsymbol{\xi}_s^*}[\psi_t(Z_t,\hat{\boldsymbol{\theta}}_{t-1})|Z_{t-1}]$. Applying a Taylor series expansion to $\mathbb{E}_{\boldsymbol{\theta}^*,\boldsymbol{\xi}_s^*}[\psi_t(Z_t,\hat{\boldsymbol{\theta}}_{t-1})|Z_{t-1}]$ around the true parameter $\boldsymbol{\theta}^* = \mathbf{0}$, we have

$$\begin{aligned}&\mathbb{E}_{\boldsymbol{\theta}^*,\boldsymbol{\xi}_s^*}\left[\psi_t(Z_t,\hat{\boldsymbol{\theta}}_{t-1})\middle|\,Z_{t-1}\right] \\ &= \mathbb{E}_{\boldsymbol{\theta}^*,\boldsymbol{\xi}_s^*}\left[\psi_t(Z_t,\boldsymbol{\theta}^*)|\,Z_{t-1}\right] + \mathbb{E}_{\boldsymbol{\theta}^*,\boldsymbol{\xi}_s^*}\left[\partial_{\boldsymbol{\theta}}\psi_t(Z_t,\boldsymbol{\theta}^*)|\,Z_{t-1}\right]\hat{\boldsymbol{\theta}}_{t-1} + o_p\left(|\hat{\boldsymbol{\theta}}_{t-1}|\right) \\ &= \mathbb{E}_{\boldsymbol{\theta}^*,\boldsymbol{\xi}_s^*}\left[\partial_{\boldsymbol{\theta}}\psi_t(Z_t,\boldsymbol{\theta}^*)|\,Z_{t-1}\right]\hat{\boldsymbol{\theta}}_{t-1} + o_p\left(|\hat{\boldsymbol{\theta}}_{t-1}|\right),\end{aligned}$$

where we have used the fact that $\mathbb{E}_{\boldsymbol{\theta}^*,\boldsymbol{\xi}_s^*}[\psi_t(Z_t,\boldsymbol{\theta}^*)|Z_{t-1}]$ is zero. Since $\hat{\mathbf{R}}_t = \mathbf{A} + o_p(1)$ is satisfied, $\mathbf{C}_1$ can be rewritten as

$$\begin{aligned}\mathbf{C}_1 &= \left(\mathbf{A}^{-1} + o_p(1)\right)\left(\mathbb{E}_{\boldsymbol{\theta}^*,\boldsymbol{\xi}_s^*}\left[\partial_{\boldsymbol{\theta}}\psi_t(Z_t,\boldsymbol{\theta}^*)|Z_{t-1}\right]\hat{\boldsymbol{\theta}}_{t-1} + o_p\left(|\hat{\boldsymbol{\theta}}_{t-1}|\right)\right)\hat{\boldsymbol{\theta}}_{t-1}^\top + o_p\left(\frac{1}{t}\right) \\ &= \mathbf{A}^{-1}\mathbb{E}_{\boldsymbol{\theta}^*,\boldsymbol{\xi}_s^*}\left[\partial_{\boldsymbol{\theta}}\psi_t(Z_t,\boldsymbol{\theta}^*)|Z_{t-1}\right]\hat{\boldsymbol{\theta}}_{t-1}\hat{\boldsymbol{\theta}}_{t-1}^\top + o_p\left(\|\hat{\boldsymbol{\theta}}_{t-1}\|^2\right) + o_p\left(\frac{1}{t}\right).\end{aligned} \tag{44}$$

We now use Equations (43) and (44), leading to

$$
\begin{aligned}
\mathbb{E}_{\boldsymbol{\theta}^*,\boldsymbol{\xi}_s^*}\left[\hat{\boldsymbol{\theta}}_t\hat{\boldsymbol{\theta}}_t^\top|Z_{t-1}\right] = {} & \left(1+o_p\left(\frac{1}{t}\right)\right)\hat{\boldsymbol{\theta}}_{t-1}\hat{\boldsymbol{\theta}}_{t-1}^\top - \frac{1}{t}\mathbf{A}^{-1}\mathbb{E}_{\boldsymbol{\theta}^*,\boldsymbol{\xi}_s^*}\left[\partial_{\boldsymbol{\theta}}\boldsymbol{\psi}_t(Z_t,\boldsymbol{\theta}^*)|Z_{t-1}\right]\hat{\boldsymbol{\theta}}_{t-1}\hat{\boldsymbol{\theta}}_{t-1}^\top \\
& - \frac{1}{t}\left(\mathbf{A}^{-1}\mathbb{E}_{\boldsymbol{\theta}^*,\boldsymbol{\xi}_s^*}\left[\partial_{\boldsymbol{\theta}}\boldsymbol{\psi}_t(Z_t,\boldsymbol{\theta}^*)|Z_{t-1}\right]\hat{\boldsymbol{\theta}}_{t-1}\hat{\boldsymbol{\theta}}_{t-1}^\top\right)^\top \\
& + \frac{1}{t^2}\mathbf{A}^{-1}\mathbb{E}_{\boldsymbol{\theta}^*,\boldsymbol{\xi}_s^*}\left[\boldsymbol{\psi}_t(Z_t,\boldsymbol{\theta}^*)\boldsymbol{\psi}_t(Z_t,\boldsymbol{\theta}^*)^\top|Z_{t-1}\right](\mathbf{A}^{-1})^\top + o_p\left(\frac{1}{t^2}\right).
\end{aligned}
$$

Taking the expectation over the sequence, we obtain

$$
\begin{aligned}
\mathbb{E}_{\boldsymbol{\theta}^*,\boldsymbol{\xi}^*}\left[\hat{\boldsymbol{\theta}}_t\hat{\boldsymbol{\theta}}_t^\top\right] = {} & \left(1+o\left(\frac{1}{t}\right)\right)\mathbb{E}_{\boldsymbol{\theta}^*,\boldsymbol{\xi}^*}\left[\hat{\boldsymbol{\theta}}_{t-1}\hat{\boldsymbol{\theta}}_{t-1}^\top\right] - \frac{1}{t}\mathbf{A}^{-1}\mathbb{E}_{\boldsymbol{\theta}^*,\boldsymbol{\xi}^*}\left[\partial_{\boldsymbol{\theta}}\boldsymbol{\psi}_t(Z_t,\boldsymbol{\theta}^*)\hat{\boldsymbol{\theta}}_{t-1}\hat{\boldsymbol{\theta}}_{t-1}^\top\right] \\
& - \frac{1}{t}\left(\mathbf{A}^{-1}\mathbb{E}_{\boldsymbol{\theta}^*,\boldsymbol{\xi}^*}\left[\partial_{\boldsymbol{\theta}}\boldsymbol{\psi}_t(Z_t,\boldsymbol{\theta}^*)\hat{\boldsymbol{\theta}}_{t-1}\hat{\boldsymbol{\theta}}_{t-1}^\top\right]\right)^\top + \frac{1}{t^2}\mathbf{A}^{-1}\boldsymbol{\Sigma}(\mathbf{A}^{-1})^\top + o\left(\frac{1}{t^2}\right),
\end{aligned}
$$

where we have used the fact that $\mathbb{E}_{\boldsymbol{\theta}^*,\boldsymbol{\xi}^*}\left[\boldsymbol{\psi}_t(Z_t,\boldsymbol{\theta}^*)\boldsymbol{\psi}_t(Z_t,\boldsymbol{\theta}^*)^\top\right]$ converges to $\boldsymbol{\Sigma}$: $\mathbb{E}_{\boldsymbol{\theta}^*,\boldsymbol{\xi}^*}\left[\boldsymbol{\psi}_t(Z_t,\boldsymbol{\theta}^*)\boldsymbol{\psi}_t(Z_t,\boldsymbol{\theta}^*)^\top\right] = \boldsymbol{\Sigma} + o(1)$. Using assumption (c) in Theorem 10 and applying the trace operator, we obtain

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

We now introduce the following lemma.

Lemma 22 *(Bottou and LeCun, 2005, Lemma 1) Let $\{u_t\}$ be a positive sequence defined as*

$$
u_t = \left(1-\frac{\bar{\alpha}}{t}+o\left(\frac{1}{t}\right)\right)u_{t-1} + \frac{\bar{\beta}}{t^2} + o\left(\frac{1}{t^2}\right).
$$

If $\bar{\alpha} > 1$ and $\bar{\beta} > 0$ hold, then

$$
tu_t \to \frac{\bar{\beta}}{\bar{\alpha}-1}.
$$

The proof is given in Lemma 1 in Bottou and LeCun (2005). Referring the result of Lemma 22, we have

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

■

Appendix L. Proof of Lemma 11

Proof Since the MRPs defined in Section 2 are ergodic, the MRPs satisfy geometrically uniform mixing. By performing a Taylor series expansion to estimating Equation (15) around the parameter $\bar{\boldsymbol{\theta}}$, we obtain

$$
\mathbf{0} = \sum_{t=1}^{T}\bar{\boldsymbol{\psi}}_t(Z_t,\bar{\boldsymbol{\theta}}) + \sum_{t=1}^{T}\partial_{\boldsymbol{\theta}}\bar{\boldsymbol{\psi}}_t(Z_t,\bar{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}}_T-\bar{\boldsymbol{\theta}}) + O\left(\left\|\hat{\boldsymbol{\theta}}_T-\bar{\boldsymbol{\theta}}\right\|^2\right).
$$

Here, high order terms are in total represented as $O(\|\hat{\boldsymbol\theta}_T - \bar{\boldsymbol\theta}\|^2)$ because of the twice differentiable condition for the function $g(s,\boldsymbol\theta)$ as Assumption 3. By applying the law of large numbers (ergodic pointwise theorem) (Billingsley, 1995, Theorem 24.1) to $(1/T)\sum_{t=1}^T \partial_{\boldsymbol\theta}\bar\psi_t(Z_t,\bar{\boldsymbol\theta})$, we have

$$\frac{1}{T}\sum_{t=1}^{T}\partial_{\boldsymbol\theta}\bar\psi_t(Z_t,\bar{\boldsymbol\theta}) = \frac{1}{T}\sum_{t=1}^{T}\bar{\boldsymbol w}_{t-1}(Z_{t-1})\{\partial_{\boldsymbol\theta}\varepsilon(z_t,\bar{\boldsymbol\theta})\}^\top \xrightarrow{a.s.} \underbrace{\lim_{t\to\infty}\mathbb{E}\left[\bar{\boldsymbol w}_{t-1}(Z_{t-1})\partial_{\boldsymbol\theta}\{\varepsilon(z_t,\bar{\boldsymbol\theta})\}^\top\right]}_{=\bar{\mathbf A}}.$$

Let $\boldsymbol k \in \mathbb{R}^m$ be any nonzero vector. By applying the central limit theorem in Lemma 18 to $(1/\sqrt{T})\sum_{t=1}^T \boldsymbol k^\top \bar\psi_t(Z_t,\bar{\boldsymbol\theta})$, we have

$$\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\boldsymbol k^\top\bar\psi_t(Z_t,\bar{\boldsymbol\theta}) = \frac{1}{\sqrt{T}}\sum_{t=1}^{T}\boldsymbol k^\top\bar{\boldsymbol w}_{t-1}(Z_{t-1})\varepsilon_t(z_t,\bar{\boldsymbol\theta})$$

$$\xrightarrow{d} \mathcal{N}\left(0, \boldsymbol k\underbrace{\left(\lim_{t\to\infty}\mathbb{E}\left[\varepsilon_t(z_t,\bar{\boldsymbol\theta})^2\bar{\boldsymbol w}_{t-1}\bar{\boldsymbol w}_{t-1}^\top\right] + \lim_{t\to\infty}2\sum_{t'=1}^{\infty}\mathrm{cov}\left[\varepsilon(z_t,\bar{\boldsymbol\theta})\bar{\boldsymbol w}_{t-1}, \varepsilon(z_{t+t'},\bar{\boldsymbol\theta})\bar{\boldsymbol w}_{t+t'-1}\right]\right)}_{\bar{\boldsymbol\Sigma}}\boldsymbol k\right).$$

Therefore, from the Cramér-Wold theorem (van der Vaart, 2000), $(1/\sqrt{T})\sum_{t=1}^T\bar\psi_t(Z_t,\bar{\boldsymbol\theta})$ converges to a Gaussian distribution as follows;

$$\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\bar\psi_t(Z_t,\bar{\boldsymbol\theta}) \xrightarrow{d} \mathcal{N}\left(\mathbf{0},\bar{\boldsymbol\Sigma}\right).$$

By neglecting higher order terms, we obtain

$$\sqrt{T}(\hat{\boldsymbol\theta}_T - \bar{\boldsymbol\theta}) \sim \mathcal{N}\left(\mathbf{0}, \bar{\mathbf A}^{-1}\bar{\boldsymbol\Sigma}(\bar{\mathbf A}^\top)^{-1}\right).$$

Then, $\hat{\boldsymbol\theta}_T$ is Gaussian distributed: $\hat{\boldsymbol\theta}_T \sim \mathcal{N}(\bar{\boldsymbol\theta},\widetilde{\mathrm{Av}})$, where the asymptotic variance $\widetilde{\mathrm{Av}}$ is given by Equation (32). ■

Appendix M. Proof of Lemma 12

Proof Let $\boldsymbol k \in \mathbb{R}^m$ be any nonzero vector. By applying the central limit theorem to $(1/\sqrt{T})\sum_{t=1}^T \boldsymbol k^\top\bar\psi_t(Z_t,\bar{\boldsymbol\theta})$ in Lemma 18, we have

$$\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\boldsymbol k^\top\bar\psi_t(Z_t,\bar{\boldsymbol\theta}) \xrightarrow{d} \mathcal{N}\left(0,\boldsymbol k^\top\bar{\boldsymbol\Sigma}\boldsymbol k\right),$$

where

$$\boldsymbol k^\top\bar{\boldsymbol\Sigma}\boldsymbol k = \lim_{t\to\infty}\boldsymbol k^\top\mathbb{E}\left[\varepsilon(z_t,\bar{\boldsymbol\theta})^2\bar{\boldsymbol w}_{t-1}\bar{\boldsymbol w}_{t-1}^\top\right]\boldsymbol k$$
$$+\lim_{t\to\infty}2\sum_{t'=1}^{\infty}\mathrm{cov}\left[\varepsilon(z_t,\bar{\boldsymbol\theta})\boldsymbol k^\top\bar{\boldsymbol w}_{t-1}, \varepsilon(z_{t+t'},\bar{\boldsymbol\theta})\boldsymbol k^\top\bar{\boldsymbol w}_{t+t'-1}\right].$$

Since the target process is a geometrically uniform mixing, there exist some positive constants C and $\rho \in [0,1)$ such that $\varphi(t) \leq C\rho^t$. Then, by using the covariance bound in Lemma 19, we obtain

$$\begin{aligned}\left|\mathrm{cov}\left[\varepsilon(z_t,\bar{\boldsymbol{\theta}})\boldsymbol{k}^\top\bar{\boldsymbol{w}}_{t-1},\varepsilon(z_{t+t'},\bar{\boldsymbol{\theta}})\boldsymbol{k}^\top\bar{\boldsymbol{w}}_{t+t'-1}\right]\right| &\leq 2\sqrt{\varphi(t')}\lim_{t\to\infty}\boldsymbol{k}^\top\mathbb{E}\left[\varepsilon(z_t,\bar{\boldsymbol{\theta}})^2\bar{\boldsymbol{w}}_{t-1}\bar{\boldsymbol{w}}_{t-1}^\top\right]\boldsymbol{k}\\ &\leq 2\sqrt{C}\rho^{t'/2}\lim_{t\to\infty}\boldsymbol{k}^\top\mathbb{E}\left[\varepsilon(z_t,\bar{\boldsymbol{\theta}})^2\bar{\boldsymbol{w}}_{t-1}\bar{\boldsymbol{w}}_{t-1}^\top\right]\boldsymbol{k}.\end{aligned}$$

Therefore, $\boldsymbol{k}^\top\bar{\boldsymbol{\Sigma}}\boldsymbol{k}$ is bounded as

$$\begin{aligned}&\left|\boldsymbol{k}^\top\bar{\boldsymbol{\Sigma}}\boldsymbol{k}\right|\\ &=\left|\lim_{t\to\infty}\boldsymbol{k}^\top\mathbb{E}\left[\varepsilon(z_t,\bar{\boldsymbol{\theta}})^2\bar{\boldsymbol{w}}_{t-1}\bar{\boldsymbol{w}}_{t-1}^\top\right]\boldsymbol{k}+2\lim_{t\to\infty}\sum_{t'=1}^{\infty}\mathrm{cov}\left[\varepsilon(z_t,\bar{\boldsymbol{\theta}})\boldsymbol{k}^\top\bar{\boldsymbol{w}}_{t-1},\varepsilon(z_{t+t'},\bar{\boldsymbol{\theta}})\boldsymbol{k}^\top\bar{\boldsymbol{w}}_{t+t'-1}\right]\right|\\ &\leq\lim_{t\to\infty}\left|\boldsymbol{k}^\top\mathbb{E}\left[\varepsilon(z_t,\bar{\boldsymbol{\theta}})^2\bar{\boldsymbol{w}}_{t-1}\bar{\boldsymbol{w}}_{t-1}^\top\right]\boldsymbol{k}\right|+2\lim_{t\to\infty}\sum_{t'=1}^{\infty}\left|\mathrm{cov}\left[\varepsilon(z_t,\bar{\boldsymbol{\theta}})\boldsymbol{k}^\top\bar{\boldsymbol{w}}_{t-1},\varepsilon(z_{t+t'},\bar{\boldsymbol{\theta}})\boldsymbol{k}^\top\bar{\boldsymbol{w}}_{t+t'-1}\right]\right|\\ &\leq\lim_{t\to\infty}\left|\boldsymbol{k}^\top\mathbb{E}\left[\varepsilon(z_t,\bar{\boldsymbol{\theta}})^2\bar{\boldsymbol{w}}_{t-1}\bar{\boldsymbol{w}}_{t-1}^\top\right]\boldsymbol{k}\right|+4\sqrt{C}\sum_{t'=1}^{\infty}\rho^{t'/2}\lim_{t\to\infty}\left|\boldsymbol{k}^\top\mathbb{E}\left[\varepsilon(z_t,\bar{\boldsymbol{\theta}})^2\bar{\boldsymbol{w}}_{t-1}\bar{\boldsymbol{w}}_{t-1}^\top\right]\boldsymbol{k}\right|\\ &=\lim_{t\to\infty}\left|\boldsymbol{k}^\top\mathbb{E}\left[\varepsilon(z_t,\bar{\boldsymbol{\theta}})^2\bar{\boldsymbol{w}}_{t-1}\bar{\boldsymbol{w}}_{t-1}^\top\right]\boldsymbol{k}\right|\left(1+4\sqrt{C}\sum_{t'=1}^{\infty}\rho^{t'/2}\right)\\ &=\lim_{t\to\infty}\left|\boldsymbol{k}^\top\mathbb{E}\left[\varepsilon(z_t,\bar{\boldsymbol{\theta}})^2\bar{\boldsymbol{w}}_{t-1}\bar{\boldsymbol{w}}_{t-1}^\top\right]\boldsymbol{k}\right|\left(1+4\sqrt{C}\frac{\rho^{1/2}}{(1-\rho^{1/2})}\right)\\ &=\Upsilon\lim_{t\to\infty}\left|\boldsymbol{k}^\top\mathbb{E}\left[\varepsilon(z_t,\bar{\boldsymbol{\theta}})^2\bar{\boldsymbol{w}}_{t-1}\bar{\boldsymbol{w}}_{t-1}^\top\right]\boldsymbol{k}\right|=\Upsilon\left|\boldsymbol{k}^\top\bar{\boldsymbol{\Sigma}}_0\boldsymbol{k}\right|,\end{aligned}$$

where $\Upsilon = 1+4\sqrt{C}\rho^{1/2}/(1-\rho^{1/2})$. Thus, we can obtain the following relation;

$$\boldsymbol{k}^\top\left(\Upsilon\bar{\boldsymbol{\Sigma}}_0-\bar{\boldsymbol{\Sigma}}\right)\boldsymbol{k}\geq 0.$$

This implies that $\Upsilon\bar{\boldsymbol{\Sigma}}_0-\bar{\boldsymbol{\Sigma}}$ is a semipositive definite matrix, hence we derive

$$\frac{1}{T}\bar{\mathbf{A}}^{-1}\bar{\boldsymbol{\Sigma}}\left(\bar{\mathbf{A}}^{-1}\right)^\top\preceq\frac{\Upsilon}{T}\bar{\mathbf{A}}^{-1}\bar{\boldsymbol{\Sigma}}_0\left(\bar{\mathbf{A}}^{-1}\right)^\top.$$

■

References

H. Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723, 1974.

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

S. Amari and J. F. Cardoso. Blind source separation-semiparametric statistical approach. *IEEE Transactions on Signal Processing*, 45(11):2692–2700, 2002.

S. Amari and M. Kawanabe. Information geometry of estimating functions in semi-parametric statistical models. *Bernoulli*, 3(1):29–54, 1997.

S. Amari, H. Park, and K. Fukumizu. Adaptive method of realizing natural gradient learning for multilayer perceptrons. *Neural Computation*, 12(6):1399–1409, 2000.

L. Baird. Residual algorithms: Reinforcement learning with function approximation. In *Proceedings of the 12th International Conference on Machine Learning*, pages 30–37, 1995.

D. P. Bertsekas and J. N. Tsitsiklis. *Neuro-Dynamic Programming*. Athena Scientific, 1996.

P. J. Bickel, C. A. Klaassen, Y. Ritov, and J. A. Wellner. *Efficient and Adaptive Estimation for Semiparametric Models*. Springer, 1998.

P. Billingsley. The Lindeberg-Levy theorem for martingales. *Proceedings of the American Mathematical Society*, 12(5):788–792, 1961.

P. Billingsley. *Probability and Measure*. John Wiley and Sons, 1995.

L. Bottou and Y. LeCun. Large scale online learning. In *Advances in Neural Information Processing Systems 16*, 2004.

L. Bottou and Y. LeCun. On-line learning for very large datasets. *Applied Stochastic Models in Business and Industry*, 21(4):137–151, 2005.

J. A. Boyan. Technical update: Least-squares temporal difference learning. *Machine Learning*, 49 (2):233–246, 2002.

R. C. Bradley. Basic properties of strong mixing conditions. A survey and some open questions. *Probability Surveys*, 2:107–144, 2005.

S. J. Bradtke and A. G. Barto. Linear least-squares algorithms for temporal difference learning. *Machine Learning*, 22(1):33–57, 1996.

R. H. Crites and A. G. Barto. Improving elevator performance using reinforcement learning. In *Advances in Neural Information Processing Systems 8*, pages 1017–1023, 1996.

A. Geramifard, M. Bowling, and R. S. Sutton. Incremental least-squares temporal difference learning. In *Proceedings of the 21st National Conference on Artificial Intelligence*, pages 356–361. AAAI Press, 2006.

A. Geramifard, M. Bowling, M. Zinkevich, and R. S. Sutton. iLSTD: Eligibility traces and convergence analysis. In *Advances in Neural Information Processing Systems 19*, pages 441–448, 2007.

V. P. Godambe. An optimum property of regular maximum likelihood estimation. *The Annals of Mathematical Statistics*, 31(4):1208–1211, 1960.

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

V. P. Godambe, editor. *Estimating Functions*. Oxford University Press, 1991.

S. Grunëwälder and K. Obermayer. Optimality of LSTD and its relation to TD and MC. Technical report, Berlin University of Technology, 2006.

R. A. Horn and C. R. Johnson. *Matrix Analysis*. Cambridge University Press, 1985.

P. J. Huber and E. M. Ronchetti. *Robust Statistics*. John Wiley and Sons, 2009.

I. A. Ibragimov and I. U. V. Linnik. *Independent and Stationary Sequences of Random Variables*. Wolters-Noordhoff, 1971.

M. Kawanabe and K. Müller. Estimating functions for blind separation when sources have variance dependencies. *Journal of Machine Learning Research*, 6(1):453—482, 2005.

V. R. Konda. *Actor-Critic Algorithm*. PhD thesis, Massachusetts Institute of Technology, 2002.

S. Konishi and G. Kitagawa. Generalised information criteria in model selection. *Biometrika*, 83 (4):875–890, 1996.

N. Le Roux, P. A. Manzagol, and Y. Bengio. Topmoumoute online natural gradient algorithm. In J.C. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems 20*. MIT Press, 2008.

P. Liang and M. I. Jordan. An asymptotic analysis of generative, discriminative, and pseudolikelihood estimators. In *Proceedings of the 25th International Conference on Machine Learning*, pages 584–591, 2008.

S. Mahadevan and M. Maggioni. Proto-value functions: A Laplacian framework for learning representation and control in Markov decision processes. *Journal of Machine Learning Research*, 8: 2169–2231, 2007.

S. Mannor, D. Simester, P. Sun, and J. N. Tsitsiklis. Bias and variance in value function estimation. In *Proceedings of the 21st International Conference on Machine Learning*, pages 308–322, 2004.

S. Mannor, D. Simester, P. Sun, and J. N. Tsitsiklis. Bias and variance approximation in value function estimates. *Management Science*, 53(2):308–322, 2007.

N. Murata and S. Amari. Statistical analysis of learning dynamics. *Signal Processing*, 74(1):3–28, 1999.

A. Nedić and D. P. Bertsekas. Least squares policy evaluation algorithms with linear function approximation. *Discrete Event Dynamic Systems*, 13(1):79–110, 2003.

J. Neveu. *Discrete-parameter Martingales*. Elsevier, 1975.

S. Singh and D. P. Bertsekas. Reinforcement learning for dynamic channel allocation in cellular telephone systems. In *Advances in Neural Information Processing Systems 9*, pages 974–980, 1997.

S. Singh and P. Dayan. Analytical mean squared error curves for temporal difference learning. *Machine Learning*, 32(1):5–40, 1998.

M. Sørensen. On asymptotics of estimating functions. *Brazilian Journal of Probability and Statistics*, 13(2):419–428, 1999.

R. S. Sutton. Learning to predict by the methods of temporal differences. *Machine learning*, 3(1): 9–44, 1988.

R. S. Sutton and A. G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, 1998.

R. S. Sutton, H. R. Maei, D. Precup, S. Bhatnagar, D. Silver, C. Szepesvári, and E. Wiewiora. Fast gradient-descent methods for temporal-difference learning with linear function approximation. In *Proceedings of the 26thth International Conference on Machine Learning*, pages 993–1000, 2009a.

R. S. Sutton, C. Szepesvári, and R. H. Maei. A convergent $O(n)$ temporal-difference algorithm for off-policy learning with linear function approximation. In *Advances in Neural Information Processing Systems 21*, 2009b.

G. Tesauro. Temporal difference learning and TD-Gammon. *Communications of the ACM*, 38(3): 58 – 68, 1995.

A. W. van der Vaart. *Asymptotic Statistics*. Cambridge University Press, 2000.

W. Wefelmeyer. Quasi-likelihood models and optimal inference. *The Annals of Statistics*, 24(1): 405–422, 1996.

H. Yu and D. P. Bertsekas. Convergence results for some temporal difference methods based on least squares. Technical report, LIDS REPORT 2697, 2006.

W. Zhang and T. G. Dietterich. A reinforcement learning approach to job-shop scheduling. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, pages 1114–1120, 1995.