A Semiparametric Statistical Approach to Model-Free Policy Evaluation

Tsuyoshi Ueno[†] TSUYOS-U@SYS.I.KYOTO-U.AC.JP
Motoaki Kawanabe[‡] MOTOAKI.KAWANABE@FIRST.FRAUNHOFER.DE
Takeshi Mori[†] TAK-MORI@SYS.I.KYOTO-U.AC.JP
Shin-ichi Maeda[†] ICHI@SYS.I.KYOTO-U.AC.JP
Shin Ishii[†] ISHII@I.KYOTO-U.AC.JP

[†]Graduate School of Informatics, Kyoto University, Gokasho, Uji, Kyoto 611-0011, Japan
[‡]Fraunhofer FIRST, IDA, Kekuléstr. 7, 12489 Berlin, Germany

Abstract

Reinforcement learning (RL) methods based on least-squares temporal difference (LSTD) have been developed recently and have shown good practical performance. However, the quality of their estimation has not been well elucidated. In this article, we discuss LSTD-based policy evaluation from the new viewpoint of semiparametric statistical inference. In fact, the estimator can be obtained from a particular estimating function which guarantees its convergence to the true value asymptotically, without specifying a model of the environment. Based on these observations, we 1) analyze the asymptotic variance of an LSTD-based estimator, 2) derive the optimal estimating function with the minimum asymptotic estimation variance, and 3) derive a suboptimal estimator to reduce the computational burden in obtaining the optimal estimating function.

1. Introduction

Reinforcement learning (RL) is a machine learning framework based on reward-related interactions with environments (Sutton & Barto, 1998). In many RL methods, policy evaluation, in which a value function is estimated from sample trajectories, is an important step for improving a current policy. Since RL problems often involve high-dimensional state spaces, the value functions are often approximated by low-dimensional parametric models. Linear function approximation has mostly been used due to their simplicity and computational convenience.

To estimate the value function with a linear model, an online procedure called temporal difference (TD) learning (Sutton & Barto, 1998) and a batch procedure called least-squares temporal difference (LSTD) learning are widely used (Bradtke & Barto, 1996). LSTD can achieve fast learning, because it uses entire sample trajectories simultaneously. Recently, efficient procedures for policy improvement combined with policy evaluation by LSTD have been developed, and have shown good performance in realistic problems. For example, the least squares policy iteration (LSPI) method maximizes the Q-function estimated by LSTD (Lagoudakis & Parr, 2003), and the natural actor-critic (NAC) algorithm uses the natural policy gradient obtained by LSTD (Peters et al., 2005). Although variance reduction techniques have been proposed for other RL algorithms (Greensmith et al., 2004; Mannor et al., 2007), the important issue of how to evaluate and reduce the estimation variance of LSTD learning remains unresolved.

In this article, we discuss LSTD-based policy evaluation in the framework of semiparmetric statistical inference, which is new to the RL field. Estimation of linearly-represented value functions can be formulated as a semiparametric inference problem, where the statistical model includes not only the parameters of interest but also additional nuisance parameters with innumerable degrees of freedom (Godambe, 1991; Amari & Kawanabe, 1997; Bickel et al., 1998). We approach this problem by using estimating functions, which provide a well-established method for semiparametric estimation (Godambe, 1991). We then show that the instrumental variable method, a technique used in LSTD

Appearing in *Proceedings of the 25th International Conference on Machine Learning*, Helsinki, Finland, 2008.

learning, can be constructed from an estimating function which guarantees its consistency (asymptotic lack of bias) by definition.

As the main results, we show the asymptotic estimation variance in a general instrumental variable method (Lemma 2) and the optimal estimating function that yields the minimum asymptotic variance of the estimation (Theorem 1). We also derive a suboptimal instrumental variable, based on the idea of the c-estimator (Amari & Kawanabe, 1997), to reduce the computational difficulty of estimating the optimal instrumental variable (Theorem 2). As a proof of concept, we compare the mean squared error (MSE) of our new estimators with that of LSTD on a simple example of the Markov decision process (MDP).

2. Background

2.1. MDPs and Policy Evaluation

RL is an approach to finding an optimal policy for sequential decision-making in an unknown environment. We consider a finite MDP, which is defined as a quadruple $(\mathcal{S}, \mathcal{A}, p, r)$: $\mathcal{S}$ is a finite set of states; $\mathcal{A}$ is a finite set of actions; $p(s_{t+1}|s_t, a_t)$ is the transition probability to a next state s_{t+1} when taking an action a_t at state s_t; and $r(s_t, a_t, s_{t+1})$ is a reward received with the state transition. Let $\pi(s_t, a_t) = p(a_t|s_t)$ be a stochastic policy that the agent follows. We introduce the following assumption concerning the MDP.

Assumption 1. *An MDP has a stationary state distribution $d^\pi(s) = p(s)$ under the policy $\pi(s_t, a_t)$.*

There are two major choices in definition of the state value function: discounted reward accumulation and average reward (Bertsekas & Tsitsiklis, 1996). With the former choice, the value function is defined as

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

where $\mathbb{E}^\pi[\cdot|s_0 = s]$ is the expectation with respect to the sample trajectory conditioned on $s_0 = s$ and $r_{t+1} := r(s_t, a_t, s_{t+1})$. $\gamma \in [0, 1)$ is the discount factor. With the latter choice, on the other hand, the value function is defined as

$$V^\pi(s) := \sum_{t=0}^{\infty} \mathbb{E}^\pi \left[r_{t+1} - \bar{r} | s_0 = s\right], \tag{2}$$

where $\bar{r} := \sum_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}} \sum_{s'\in\mathcal{S}} d^\pi(s)\pi(s, a)p(s'|s, a)r(s, a, s')$ denotes the average reward over the stationary distribution.

According to the Bellman equation, eq. (2) can be rewritten as

$$V^\pi(s_t) = \sum_{s_{t+1}\in\mathcal{S}} p(s_{t+1}|s_t)\bar{r}(s_t, s_{t+1}) - \bar{r} + \sum_{s_{t+1}\in\mathcal{S}} p(s_{t+1}|s_t)V^\pi(s_{t+1}), \tag{3}$$

where $p(s_{t+1}|s_t) := \sum_{a_t\in\mathcal{A}} \pi(s_t, a_t)p(s_{t+1}|s_t, a_t)$ and $\bar{r}(s_t, s_{t+1}) := \frac{\sum_{a_t\in\mathcal{A}} \pi(s_t,a_t)p(s_{t+1}|s_t,a_t)r(s_t,a_t,s_{t+1})}{p(s_{t+1}|s_t)}$. Throughout this article, we assume that the linear function approximation is faithful, and discuss only asymptotic estimation variance. (In general cases, bias becomes non-negligible and selection of basis functions is more important.)

Assumption 2. *The value function can be represented as a linear function of some features:*

$$V^\pi(s_t) = \boldsymbol{\phi}(s_t)^\top \boldsymbol{\theta} = \boldsymbol{\phi}_t^\top \boldsymbol{\theta}, \tag{4}$$

where $\boldsymbol{\phi}(s) : \mathcal{S} \to \mathcal{R}^m$ is a feature vector and $\boldsymbol{\theta} \in \mathcal{R}^m$ is a parameter vector.

Here, the symbol $\top$ denotes a transpose and the dimensionality of the feature vector m is smaller than the number of states $|\mathcal{S}|$. Substituting eq. (4) for eq. (3), we obtain the following equation

$$\left\{\boldsymbol{\phi}_t - \sum_{s_{t+1}\in\mathcal{S}} p(s_{t+1}|s_t)\boldsymbol{\phi}_{t+1}\right\}^\top \boldsymbol{\theta} = \sum_{s_{t+1}\in\mathcal{S}} p(s_{t+1}|s_t)\bar{r}(s_t, s_{t+1}) - \bar{r}. \tag{5}$$

When the matrix $\mathbb{E}^\pi\left[\left(\boldsymbol{\phi}_t - \sum_{s_{t+1}\in\mathcal{S}} p(s_{t+1}|s_t)\boldsymbol{\phi}_{t+1}\right)\left(\boldsymbol{\phi}_t - \sum_{s_{t+1}\in\mathcal{S}} p(s_{t+1}|s_t)\boldsymbol{\phi}_{t+1}\right)^\top\right]$ is non-singular and $p(s_{t+1}|s_t)$ is known, we can easily obtain the parameter $\boldsymbol{\theta}$. However, since $p(s_{t+1}|s_t)$ is unknown in normal RL settings, we have to estimate this parameter from the sample trajectory $\{s_0, a_0, r_1, \cdots, s_{N-1}, a_{N-1}, r_N\}$ alone, instead of using it directly.

Eq. (5) can be rewritten as

$$y_t = \boldsymbol{x}_t^\top \boldsymbol{\theta} + \epsilon_t, \tag{6}$$

where y_t, $\boldsymbol{x}_t$ and ϵ_t are defined as

$$y_t := r_{t+1} - \bar{r}, \quad \boldsymbol{x}_t := \boldsymbol{\phi}_t - \boldsymbol{\phi}_{t+1}$$
$$\epsilon_t := \left\{\boldsymbol{\phi}_{t+1} - \sum_{s_{t+1}\in\mathcal{S}} p(s_{t+1}|s_t)\boldsymbol{\phi}_{t+1}\right\}^\top \boldsymbol{\theta} + r_{t+1} - \sum_{s_{t+1}\in\mathcal{S}} p(s_{t+1}|s_t)\bar{r}(s_t, s_{t+1}). \tag{7}$$

When we use the discounted reward accumulation for the value function, eq. (6) also holds with

$$y_t := r_{t+1}, \quad \boldsymbol{x}_t := \boldsymbol{\phi}_t - \gamma \boldsymbol{\phi}_{t+1}$$
$$\epsilon_t := \gamma \left\{ \boldsymbol{\phi}_{t+1} - \sum_{s_{t+1} \in \mathcal{S}} p(s_{t+1}|s_t) \boldsymbol{\phi}_{t+1} \right\}^\top \boldsymbol{\theta} + r_{t+1} - \sum_{s_{t+1} \in \mathcal{S}} p(s_{t+1}|s_t) \bar{r}(s_t, s_{t+1}). \quad (8)$$

Because $\mathbb{E}^\pi[\epsilon_t] = 0$, eq. (6) can be seen as a linear regression problem, where $\boldsymbol{x}$, y and ϵ are an input, an output and observation noise, respectively (Bradtke & Barto, 1996). Note that

$$\mathbb{E}^\pi[\epsilon_t g(s_t, s_{t-1}, \cdots, s_0)] = 0 \quad (9)$$

holds for any function $g(s_t, s_{t-1}, \cdots, s_0)$ because of the Markov property. The regression problem (6) has an undesirable property, however, which is known as an "error-in-variable problem" (Young, 1984): the input $\boldsymbol{x}_t$ and observation noise variables ϵ_t are mutually dependent.

It is not easy to solve such an error-in-variable problem in a rigorous manner; the simple least-squares method lacks consistency. Therefore, LSTD learning has used the instrumental variable method (Bradtke & Barto, 1996), a standard method to solve the error-in-variable problem that employs an "instrumental variable" to remove the effects of correlation between the input and the observation noise. When $\boldsymbol{X} = [\boldsymbol{x}_0, \boldsymbol{x}_1, \cdots, \boldsymbol{x}_{N-1}]$ and $\boldsymbol{y} = [y_0, y_1, \cdots, y_{N-1}]^\top$, the estimator of the instrumental variable method is given by

$$\hat{\boldsymbol{\theta}} = [\boldsymbol{Z}\boldsymbol{X}^\top]^{-1}[\boldsymbol{Z}\boldsymbol{y}], \quad (10)$$

where $\boldsymbol{Z} = [\boldsymbol{z}_0, \boldsymbol{z}_1, \cdots, \boldsymbol{z}_{N-1}]$, and $\boldsymbol{z}_t$ is an instrumental variable that is assumed to be correlated with the input $\boldsymbol{x}_t$ but uncorrelated with the observation noise ϵ_t.

2.2. Semiparametric Model and Estimating Functions

In the error-in-variable problem, if it is possible to assume a reasonable model with a small number of parameters on the joint input-output probability $p(\boldsymbol{x}, y)$, a proper estimator with consistency can be obtained by the maximum likelihood method. Since the transition probability $p(s_{t+1}|s_t)$ is unknown and usually difficult to estimate, it is practically impossible to construct such a parametric model. Let $\boldsymbol{k}_x$ and $\boldsymbol{k}_\epsilon$ be parameters which characterize the input distribution $p(\boldsymbol{x})$ and the conditional distribution $p(y|\boldsymbol{x})$ of output y given $\boldsymbol{x}$, respectively. Then, the joint distribution becomes

$$p(\boldsymbol{x}, y; \boldsymbol{\theta}, \boldsymbol{k}_x, \boldsymbol{k}_\epsilon) = p(y|\boldsymbol{x}; \boldsymbol{\theta}, \boldsymbol{k}_\epsilon) p(\boldsymbol{x}; \boldsymbol{k}_x). \quad (11)$$

We would like to estimate the parameter $\boldsymbol{\theta}$ representing the value function in the presence of the extra unknowns $\boldsymbol{k}_x$ and $\boldsymbol{k}_\epsilon$, which can have innumerable degrees of freedom. Statistical models which contain such (possibly infinite-dimensional) nuisance parameters in addition to parameters of interest are called semiparametric (Bickel et al., 1998). In semiparametric inference, one established way of estimating parameters is to employ an estimating function (Godambe, 1991), which can give a consistent estimator of $\boldsymbol{\theta}$ without estimation of the nuisance parameters $\boldsymbol{k}_x$ and $\boldsymbol{k}_\epsilon$. Now we begin with a short overview of the estimating function in the simple i.i.d. case, and then discuss the Markov chain case.

We consider a general semiparametric model $p(\boldsymbol{x}|\boldsymbol{\theta}, \boldsymbol{\kappa})$, where $\boldsymbol{\theta}$ is an m-dimensional parameter and $\boldsymbol{\kappa}$ is a nuisance parameter. An m-dimensional vector function $\boldsymbol{f}(\boldsymbol{x}; \boldsymbol{\theta})$ is called an estimating function when it satisfies the following conditions for any $\boldsymbol{\theta}$, $\boldsymbol{\kappa}$;

$$\mathbb{E}[\boldsymbol{f}(\boldsymbol{x}; \boldsymbol{\theta})|\boldsymbol{\theta}, \boldsymbol{\kappa}] = \boldsymbol{0} \quad (12)$$
$$\det \left| \mathbb{E}\left[\frac{\partial}{\partial \boldsymbol{\theta}} \boldsymbol{f}(\boldsymbol{x}; \boldsymbol{\theta}) \middle| \boldsymbol{\theta}, \boldsymbol{\kappa} \right] \right| \neq 0 \quad (13)$$
$$\mathbb{E}\left[||\boldsymbol{f}(\boldsymbol{x}; \boldsymbol{\theta})||^2 | \boldsymbol{\theta}, \boldsymbol{\kappa} \right] < \infty, \quad (14)$$

where $\mathbb{E}[\cdot|\boldsymbol{\theta}, \boldsymbol{\kappa}]$ denotes the expectation with respect to $\boldsymbol{x}$, which obeys the distribution $p(\boldsymbol{x}; \boldsymbol{\theta}, \boldsymbol{\kappa})$. The notations $\det|\cdot|$ and $||\cdot||$ denote the determinant and the Euclidean norm, respectively. Consider that i.i.d. samples $\{\boldsymbol{x}_0, \boldsymbol{x}_1, \cdots, \boldsymbol{x}_{N-1}\}$ are obtained from the true model $p(\boldsymbol{x}; \boldsymbol{\theta} = \boldsymbol{\theta}^*, \boldsymbol{\kappa} = \boldsymbol{\kappa}^*) = p(\boldsymbol{x}; \boldsymbol{\theta}^*, \boldsymbol{\kappa}^*)$ for the observed trajectory. If there is an estimating function $\boldsymbol{f}$, by solving the estimating equation

$$\sum_{t=0}^{N-1} \boldsymbol{f}(\boldsymbol{x}_t; \hat{\boldsymbol{\theta}}) = \boldsymbol{0}, \quad (15)$$

we can obtain an estimator $\hat{\boldsymbol{\theta}}$ with good asymptotic properties. A solution of eq. (15) is called an "M-estimator" in statistics; the M-estimator is consistent, i.e., converges to the true parameter $\boldsymbol{\theta}^*$ regardless of the true nuisance parameter $\boldsymbol{\kappa}^*$ when the sample size N reaches infinity. In addition, the asymptotic variance $\mathrm{AV}[\hat{\boldsymbol{\theta}}]$ is given by

$$\mathrm{AV}[\hat{\boldsymbol{\theta}}] = \mathbb{E}[(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)^\top] = \frac{1}{N} \boldsymbol{A}^{-1} \boldsymbol{M} \boldsymbol{A}^{-\top}, \quad (16)$$

where $\boldsymbol{A} = \mathbb{E}\left[\frac{\partial}{\partial \boldsymbol{\theta}} f(\boldsymbol{x};\boldsymbol{\theta})|\boldsymbol{\theta}^*, \boldsymbol{\kappa}^*\right]$ and $\boldsymbol{M} = \mathbb{E}\left[\boldsymbol{f}(\boldsymbol{x};\boldsymbol{\theta})\boldsymbol{f}^\top(\boldsymbol{x};\boldsymbol{\theta})|\boldsymbol{\theta}^*, \boldsymbol{\kappa}^*\right]$. The symbol $-\top$ denotes transpose of the inverse matrix. We omit the time index t, unless it is necessary to clarify. Note that the asymptotic variance AV depends on the true parameters, $\boldsymbol{\theta}^*$ and $\boldsymbol{\kappa}^*$, not on the samples $\{\boldsymbol{x}_0, \boldsymbol{x}_1, \cdots, \boldsymbol{x}_{N-1}\}$.

The notion of the estimating function can be extended to cases in which samples are given by a certain stochastic process (Godambe, 1985). In the semiparametric model for policy evaluation, under Assumption 1, there exist sufficient conditions of estimating functions which are almost the same as eqs. (12) - (14). The instrumental variable method is a type of estimating function method for semiparametric problems where the unknown distribution is given by eq. (11).

Lemma 1. *Suppose $\{\boldsymbol{x}_t, y_t\}$ is given by eq. (7) or (8), and $\boldsymbol{z}_t$ is given by a function of $\{s_t, \cdots, s_{t-T}\}$. If $\mathbb{E}^\pi[\boldsymbol{z}_t\boldsymbol{x}_t^\top]$ is nonsingular and $\mathbb{E}^\pi\left[||\boldsymbol{z}_t(\boldsymbol{x}_t^\top\boldsymbol{\theta} - y_t)||^2\right]$ is finite, then*

$$\boldsymbol{z}_t(\boldsymbol{x}_t^\top\boldsymbol{\theta} - y_t) \tag{17}$$

is an estimating function for the parameter $\boldsymbol{\theta}$. Therefore, the estimating equation is given by

$$\sum_{t=0}^{N-1} \boldsymbol{z}_t(\boldsymbol{x}_t^\top\boldsymbol{\theta} - y_t) = \boldsymbol{0}. \tag{18}$$

Proof For all t, the conditions corresponding to (13) and (14) are satisfied by the assumptions, and the condition (12) is satisfied as $\mathbb{E}^\pi[\boldsymbol{z}_t(\boldsymbol{x}_t^\top\boldsymbol{\theta} - y_t)] = \mathbb{E}^\pi[\boldsymbol{z}_t\epsilon_t] = \boldsymbol{0}$ from the property in eq. (9). (Q.E.D.)

LSTD is specifically an instrumental variable method in which the feature vector $\boldsymbol{z}_t = \boldsymbol{\phi}(s_t) = \boldsymbol{\phi}_t$ is used as an instrumental variable:

$$\boldsymbol{f}_{\mathrm{LSTD}} = \boldsymbol{\phi}_t(\boldsymbol{x}_t^\top\boldsymbol{\theta} - y_t). \tag{19}$$

The solution of the estimating equation is an M-estimator, and its asymptotic variance is given as follows.

Lemma 2. *Let $\boldsymbol{z}_t$ be a function of $\{s_t, ..., s_{t-T}\}$ satisfying the two conditions in Lemma 1 and $\epsilon_t^* = \boldsymbol{x}_t^\top\boldsymbol{\theta}^* - y_t$ be the residual for the true parameter $\boldsymbol{\theta}^*$. Then, the solution $\hat{\boldsymbol{\theta}}$ of the estimating equation (18) has the asymptotic variance*

$$\mathrm{AV}[\hat{\boldsymbol{\theta}}] = \frac{1}{N}\boldsymbol{A}_{\mathrm{IV}}^{-1}\boldsymbol{M}_{\mathrm{IV}}\boldsymbol{A}_{\mathrm{IV}}^{-\top}, \tag{20}$$

where $\boldsymbol{A}_{\mathrm{IV}} = \mathbb{E}_d[\boldsymbol{z}_t\boldsymbol{x}_t^\top]$, $\boldsymbol{M}_{\mathrm{IV}} = \mathbb{E}_d[(\epsilon_t^*)^2\boldsymbol{z}_t\boldsymbol{z}_t^\top]$. $E_d[\cdot]$ denotes the expectation when the sample trajectory starts from the stationary distribution $d^\pi(s_0)$. [1]

Proof The estimating equation (18) can be expressed as

$$\boldsymbol{Z}\boldsymbol{y} = \boldsymbol{Z}\boldsymbol{X}^\top\hat{\boldsymbol{\theta}} \tag{21}$$

where $\boldsymbol{Z} = [\boldsymbol{z}_0, \cdots, \boldsymbol{z}_{N-1}]$, $\boldsymbol{X} = [\boldsymbol{x}_0, \cdots, \boldsymbol{x}_{N-1}]$, $\boldsymbol{y} = [y_0, \cdots, y_{N-1}]^\top$. On the other hand, from eq. (6), the left hand side of eq. (21) is equal to $\boldsymbol{Z}\boldsymbol{X}^\top\boldsymbol{\theta}^* + \boldsymbol{Z}\boldsymbol{X}^\top\boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} = [\epsilon_0^*, \cdots, \epsilon_{N-1}^*]^\top$. Thus the asymptotic variance of the estimator $\hat{\boldsymbol{\theta}}$ is obtained as

$$\mathbb{E}^\pi[(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)^\top] = \mathbb{E}^\pi[(\boldsymbol{Z}\boldsymbol{X}^\top)^{-1}\boldsymbol{Z}\boldsymbol{\epsilon}\boldsymbol{\epsilon}^\top\boldsymbol{Z}^\top(\boldsymbol{X}\boldsymbol{Z}^\top)^{-1}]$$
$$\overset{N\to\infty}{\longrightarrow} \boldsymbol{A}_{\mathrm{IV}}^{-1}\frac{1}{N^2}\mathbb{E}^\pi[\boldsymbol{Z}\boldsymbol{\epsilon}\boldsymbol{\epsilon}^\top\boldsymbol{Z}^\top]\boldsymbol{A}_{\mathrm{IV}}^{-\top}$$

where we used the fact that the matrix $\boldsymbol{Z}\boldsymbol{X}^\top$ has the limit

$$\frac{1}{N}(\boldsymbol{Z}\boldsymbol{X}^\top) = \frac{1}{N}\sum_{t=0}^{N-1}\boldsymbol{z}_t\boldsymbol{x}_t^\top \overset{N\to\infty}{\longrightarrow} \boldsymbol{A}_{\mathrm{IV}}.$$

Also, the matrix $\mathbb{E}^\pi[\boldsymbol{Z}\boldsymbol{\epsilon}\boldsymbol{\epsilon}^\top\boldsymbol{Z}^\top]$ has the following limit:

$$\frac{1}{N}\mathbb{E}^\pi[\boldsymbol{Z}\boldsymbol{\epsilon}\boldsymbol{\epsilon}^\top\boldsymbol{Z}^\top] = \frac{1}{N}\sum_{t=0}^{N-1}\mathbb{E}^\pi[(\epsilon_t^*)^2\boldsymbol{z}_t\boldsymbol{z}_t^\top] \overset{N\to\infty}{\longrightarrow} \boldsymbol{M}_{\mathrm{IV}},$$

where we used the property in eq. (9). Therefore, we have

$$\mathbb{E}^\pi[(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)^\top] \overset{N\to\infty}{\longrightarrow} \frac{1}{N}\boldsymbol{A}_{\mathrm{IV}}^{-1}\boldsymbol{M}_{\mathrm{IV}}\boldsymbol{A}_{\mathrm{IV}}^{-\top}.$$

(Q.E.D.)

To summarize, if we have an instrumental variable which satisfies the assumptions in Lemmas 1 and 2, we can obtain an M-estimator from the estimating equation (18) with the asymptotic variance eq. (20). When more than one instrumental variable exists, it is appropriate to choose the one whose estimator has the minimum asymptotic variance.

3. Main Results

In this section, we show that estimating functions for the semiparametric model of policy evaluation are limited to the type of equation used in the instrumental variable method. Furthermore, we derive the optimal

[1] We remark that the definitions of $\boldsymbol{A}_{\mathrm{IV}}$ and $\boldsymbol{M}_{\mathrm{IV}}$ do not depend on the t. Nevertheless, we keep the time index t for clarification.

instrumental variable having the minimum asymptotic variance of the estimation.

We first remark on the invariance property of the instrumental variable method.

Lemma 3. *The value function estimation* $\hat{V}(s_t) = \boldsymbol{\phi}_t^\top \hat{\boldsymbol{\theta}} = \boldsymbol{\phi}_t^\top [\boldsymbol{Z}\boldsymbol{X}^\top]^{-1}[\boldsymbol{Z}\boldsymbol{y}]$ *is invariant with respect to the application of any regular linear transformation to either the instrumental variable* $\boldsymbol{z}_t$ *or the basis functions* $\boldsymbol{\phi}_t$.

Proof Assume that the instrumental variable and the basis functions are both transformed by any regular matrices $\boldsymbol{W}_z$ and $\boldsymbol{W}_\phi$ as $\boldsymbol{z}'_t = \boldsymbol{W}_z \boldsymbol{z}_t$ and $\boldsymbol{\phi}'_t = \boldsymbol{W}_\phi \boldsymbol{\phi}_t$. Noting that the linear transformation of $\boldsymbol{\phi}_t$ yields the linear transformation of the input $\boldsymbol{x}'_t = \boldsymbol{W}_\phi \boldsymbol{x}_t$, the estimator of the instrumental variable method given by eq. (10) becomes $\hat{\boldsymbol{\theta}}' = [\boldsymbol{Z}'(\boldsymbol{X}')^\top]^{-1}[\boldsymbol{Z}'\boldsymbol{y}] = \boldsymbol{W}_\phi^{-1}\hat{\boldsymbol{\theta}}$. This means that the estimated value function is invariant as $(\boldsymbol{\phi}'_t)^\top \hat{\boldsymbol{\theta}}' = \boldsymbol{\phi}_t^\top \hat{\boldsymbol{\theta}}$. (Q.E.D.)

When the basis functions span over the whole space of functions of the state, any set of basis functions can be represented by applying a linear transformation to another set of basis functions. This observation leads to the following Corollary.

Corollary 1. *When the basis functions* $\boldsymbol{\phi}_t$ *span the whole space of functions of the state, the value function estimation is invariant with respect to the choice of basis functions and of the instrumental variable.*

An instrumental variable may depend not only on the current state s_t, but also on the previous states $\{s_{t-1}, \cdots, s_{t-T}\}$, because such an instrumental variable does not violate the condition, $\mathrm{cov}[\boldsymbol{z}_t, \epsilon_t] = \boldsymbol{0}$. However, we do not need to consider such instrumental variables, as the following Lemma shows.

Lemma 4. *Let* $\boldsymbol{z}_t(s_t, \cdots, s_{t-T})$ *be any instrumental variable depending on the current and previous states which satisfies the conditions in Lemmas 1 and 2. Then, there is necessarily an instrumental variable depending only on the current state whose corresponding estimator has equal or minimum asymptotic variance.*

Proof We show that the conditional expectation $\tilde{\boldsymbol{z}}_t = \mathbb{E}^\pi[\boldsymbol{z}_t | s_t]$ which depends only on the current state s_t, gives an equally good or better estimator. The matrices in the asymptotic variance, eq. (20), can be calculated as

$$\begin{aligned}\boldsymbol{A}_{\boldsymbol{z}} &= \mathbb{E}_d\left[\tilde{\boldsymbol{z}}_t \boldsymbol{x}_t^\top\right] + \mathbb{E}_d\left[(\tilde{\boldsymbol{z}}_t - \boldsymbol{z}_t)\boldsymbol{x}_t^\top\right] = \boldsymbol{A}_{\tilde{\boldsymbol{z}}} \\ \boldsymbol{M}_{\boldsymbol{z}} &= \mathbb{E}_d[(\epsilon_t^*)^2(\tilde{\boldsymbol{z}}_t + \boldsymbol{z}_t - \tilde{\boldsymbol{z}}_t)(\tilde{\boldsymbol{z}}_t + \boldsymbol{z}_t - \tilde{\boldsymbol{z}}_t)^\top] \\ &= \boldsymbol{M}_{\tilde{\boldsymbol{z}}} + \mathbb{E}_d[(\epsilon_t^*)^2(\boldsymbol{z}_t - \tilde{\boldsymbol{z}}_t)(\boldsymbol{z}_t - \tilde{\boldsymbol{z}}_t)^\top],\end{aligned}$$

where we have used eq. (9). This implies that

$$\mathrm{AV}[\hat{\boldsymbol{\theta}}_{\boldsymbol{z}}] = \frac{1}{N}\boldsymbol{A}_{\boldsymbol{z}}^{-1}\boldsymbol{M}_{\boldsymbol{z}}\boldsymbol{A}_{\boldsymbol{z}}^{-\top} \succeq \frac{1}{N}\boldsymbol{A}_{\tilde{\boldsymbol{z}}}^{-1}\boldsymbol{M}_{\tilde{\boldsymbol{z}}}\boldsymbol{A}_{\tilde{\boldsymbol{z}}}^{-\top} = \mathrm{AV}[\hat{\boldsymbol{\theta}}_{\tilde{\boldsymbol{z}}}].$$

(Q.E.D.)

Here, the inequality $\succeq$ denotes the semipositive definiteness of the subtraction. Now, we consider the general form of estimating functions for inference of the value function. In the following, we consider only 'admissible' estimating functions. More precisely, we discard 'inadmissible' estimating functions whose estimators are always inferior to those of other estimating functions in the sense of asymptotic variance. To simplify analysis, we only consider the limited set of estimating functions which are defined on a one-step sample $\{s, a, s'\}$.

Proposition 1. *For the semiparametric model of eqs. (6), (7) or eqs. (6), (8), all admissible estimating functions of only the one-step sample* $\{s, a, s'\}$ *must have the form of* $\boldsymbol{f} = \boldsymbol{z}(y - \boldsymbol{x}^\top\boldsymbol{\theta})$, *where* $\boldsymbol{z}$ *is any function which does not depend on* s' *and satisfies the assumption in Lemma 1.*

Proof Due to space limitation, we will just outline the proof. To be an estimating function, the function $\boldsymbol{f}$ must satisfy $\mathbb{E}_d[\boldsymbol{f}] = \sum_{s\in\mathcal{S}} d^\pi(s) \sum_{s'\in\mathcal{S}} \sum_{a\in\mathcal{A}} p(s'|s,a)\pi(s,a)\boldsymbol{f}(s,a,s') = \boldsymbol{0}$. Because we can prove that the stationary distribution $d^\pi(s)$ takes any probability vector, $\sum_{s\in\mathcal{S}} d^\pi(s)\boldsymbol{v}(s) = \boldsymbol{0}$ implies that $\boldsymbol{v}(s) = \boldsymbol{0}$ for any state s, where $\boldsymbol{v}(s) := \sum_{s'\in\mathcal{S}} \sum_{a\in\mathcal{A}} p(s'|s,a)\pi(s,a)\boldsymbol{f}(s,a,s')$. Furthermore, the Bellman equation (6) holds, whatever the $p(s'|s,a)$ is. To fulfil $\boldsymbol{v} = \boldsymbol{0}$, $\boldsymbol{f}$ must have the form of $\boldsymbol{f} = \boldsymbol{z}(y - \boldsymbol{x}^\top\boldsymbol{\theta}) + \boldsymbol{h}$, where $\boldsymbol{z}$ does not depend on s' or a, and $\boldsymbol{h}$ is any function that satisfies $\sum_{a\in\mathcal{A}} \pi(s,a)\boldsymbol{h}(s,a,s') = \boldsymbol{0}$. However, the addition of such a function $\boldsymbol{h}$ necessarily enlarges the asymptotic variance of the estimation. Therefore, the admissible estimating function is restricted to the form of $\boldsymbol{f} = \boldsymbol{z}(y - \boldsymbol{x}^\top\boldsymbol{\theta})$. (Q.E.D)

We are currently working on the conjecture that whether Proposition 1 can be extended to general estimating functions depend on all previous states and actions. If this is true, from Lemma 4, it is sufficient to consider the instrumental variable method with $\boldsymbol{z}_t$ depending only on the current state s_t for the semiparametric inference problem. Therefore, we next discuss the optimal instrument variable of this type in terms of asymptotic variance, which corresponds to the optimal estimating function.

Algorithm 1 The pseudo code of gLSTD.

gLSTD($\mathcal{D}$, ϕ)
// $D = \{s_0, r_1, \cdots, s_{N-1}, r_N\}$: Sample sequence
// ϕ: Basis functions
// Calculate the initial parameter and its residual
$\hat{\boldsymbol{\theta}}_0 \leftarrow \left[\sum_{t=0}^{N-1} \boldsymbol{\phi}_t \boldsymbol{x}_t^\top\right]^{-1} \left[\sum_{t=0}^{N-1} \boldsymbol{\phi}_t y_t\right]$
$\hat{\epsilon}_t \leftarrow \boldsymbol{x}_t^\top \hat{\boldsymbol{\theta}}_0 - y_t$
// Calculate the estimator $\widetilde{\mathbb{E}^\pi[(\hat{\epsilon}_t)^2|s_t]}, \widetilde{\mathbb{E}^\pi[\boldsymbol{x}_t|s_t]}$
// of the conditional expectations
// and construct the instrumental variable
$\hat{\boldsymbol{z}}_t \leftarrow \widetilde{\mathbb{E}^\pi[(\hat{\epsilon}_t)^2|s_t]}^{-1} \widetilde{\mathbb{E}^\pi[\boldsymbol{x}_t|s_t]}$
// Calculate the parameter
$\hat{\boldsymbol{\theta}}_g \leftarrow \left[\sum_{t=0}^{N-1} \hat{\boldsymbol{z}}_t \boldsymbol{x}_t^\top\right]^{-1} \left[\sum_{t=0}^{N-1} \hat{\boldsymbol{z}}_t y_t\right]$
Return $\hat{\boldsymbol{\theta}}_g$

Theorem 1. *The optimal instrumental variable gives the minimum asymptotic variance*

$$\boldsymbol{z}_t^* = \begin{cases} \mathbb{E}^\pi\left[(\epsilon_t^*)^2|s_t\right]^{-1}\left(\boldsymbol{\phi}_t - \gamma\mathbb{E}^\pi[\boldsymbol{\phi}_{t+1}|s_t]\right) \\ \quad (\textit{discounted reward accumulation}) \\ \mathbb{E}^\pi\left[(\epsilon_t^*)^2|s_t\right]^{-1}\left(\boldsymbol{\phi}_t - \mathbb{E}^\pi[\boldsymbol{\phi}_{t+1}|s_t]\right) \\ \quad (\textit{average reward}). \end{cases} \tag{22}$$

The proof is given in Appendix A. Note that the definition of the optimal instrumental variable includes both the residual ϵ_t^* and the conditional expectations $\mathbb{E}^\pi[\boldsymbol{\phi}_{t+1}|s_t]$ and $\mathbb{E}^\pi[(\epsilon_t^*)^2|s_t]$. To make this estimator practical, we replace the residual ϵ_t^* with that of the LSTD estimator, and approximate the expectation, $\mathbb{E}^\pi[\boldsymbol{\phi}_{t+1}|s_t]$ and $\mathbb{E}^\pi[(\epsilon_t^*)^2|s_t]$, by using function approximation. We call this procedure "gLSTD learning" (see Algorithm 1 for its pseudo code).

To avoid estimating the functions depending on the current state, $\mathbb{E}^\pi[\boldsymbol{\phi}_{t+1}|s_t]$ and $\mathbb{E}^\pi[(\epsilon_t^*)^2|s_t]$, which appear in the instrumental variable, we simply replace them by constants. When $\boldsymbol{z}$ is an instrumental variable, addition of any constant value to $\boldsymbol{z}$, $\boldsymbol{z}' = \boldsymbol{z} + \boldsymbol{c}$, leads to another valid instrumental variable; because of Lemma 1, it is easily confirmed that $\boldsymbol{f}_{\mathrm{c}} = (\boldsymbol{z}_t + \boldsymbol{c})(\boldsymbol{x}_t^\top\boldsymbol{\theta} - y_t)$ is an estimating function. Therefore, obtaining the optimal constant $\boldsymbol{c}$ yields a suboptimal instrumental variable within instrumental variables produced by constant shifts.

Theorem 2. *The optimal shift is given by*

$$\boldsymbol{c}^* := -\frac{\mathbb{E}_d[(\epsilon_t^*)^2\boldsymbol{z}_t] - \mathbb{E}_d[(\epsilon_t^*)^2\boldsymbol{z}_t\boldsymbol{z}_t^\top]\mathbb{E}_d[\boldsymbol{x}_t\boldsymbol{z}_t^\top]^{-1}\mathbb{E}_d[\boldsymbol{x}_t]}{\mathbb{E}_d[(\epsilon_t^*)^2] - \mathbb{E}_d[(\epsilon_t^*)^2\boldsymbol{z}_t^\top]\mathbb{E}_d[\boldsymbol{x}_t\boldsymbol{z}_t^\top]^{-1}\mathbb{E}_d[\boldsymbol{x}_t]}. \tag{23}$$

Algorithm 2 The pseudo code of LSTDc

LSTDc($\mathcal{D}$, ϕ)
// $D = \{s_0, r_1, \cdots, s_{N-1}, r_N\}$: Sample sequence
// ϕ: Basis functions
// Calculate the initial parameter and its residual
$\hat{\boldsymbol{\theta}}_0 \leftarrow \left[\sum_{t=0}^{N-1} \boldsymbol{\phi}_t \boldsymbol{x}_t^\top\right]^{-1} \left[\sum_{t=0}^{N-1} \boldsymbol{\phi}_t y_t\right]$
$\hat{\epsilon}_t \leftarrow \boldsymbol{x}_t^\top \hat{\boldsymbol{\theta}}_0 - y_t$
// Construct the suboptimal
// instrumental variable with optimal shift
$$\hat{\boldsymbol{c}} \leftarrow -\frac{\left[\sum_{t=0}^{N-1} \hat{\epsilon}_t^2\boldsymbol{\phi}_t\right] - \left[\sum_{t=0}^{N-1} \hat{\epsilon}_t^2\boldsymbol{\phi}_t\boldsymbol{\phi}_t^\top\right]\left[\sum_{t=0}^{N-1} \boldsymbol{x}_t\boldsymbol{\phi}_t^\top\right]^{-1}\left[\sum_{t=0}^{N-1} \boldsymbol{x}_t\right]}{\left[\sum_{t=0}^{N-1} \hat{\epsilon}_t^2\right] - \left[\sum_{t=0}^{N-1} \hat{\epsilon}_t^2\boldsymbol{\phi}_t^\top\right]\left[\sum_{t=0}^{N-1} \boldsymbol{x}_t\boldsymbol{\phi}_t^\top\right]^{-1}\left[\sum_{t=0}^{N-1} \boldsymbol{x}_t\right]}$$
$\hat{\boldsymbol{z}}_t = \boldsymbol{\phi}_t + \hat{\boldsymbol{c}}$
//Calculate the parameter
$\hat{\boldsymbol{\theta}}_c \leftarrow \left[\sum_{t=0}^{N-1} \hat{\boldsymbol{z}}_t \boldsymbol{x}_t^\top\right]^{-1} \left[\sum_{t=0}^{N-1} \hat{\boldsymbol{z}}_t y_t\right]$
Return $\hat{\boldsymbol{\theta}}_c$

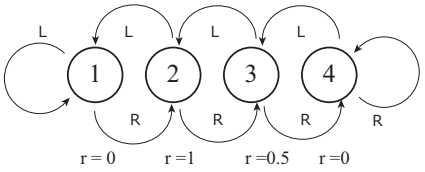

Figure 1. A four-state MDP.

The proof is given in Appendix B. In eq. (23), however, the residual ϵ_t^* is again unknown; hence, we need to approximate this, too, as in the gLSTD learning. We call this procedure "LSTDc learning" (see Algorithm 2 for its pseudo code).

4. Simulation Experiments

So far, we have discussed the asymptotic variance under the assumption that we have an infinite number of samples. In this section, we evaluate the performance of the proposed estimator in a practical situation with a finite number of samples. We use an MDP defined on a simple Markov random walk, which was also used in a previous study (Lagoudakis & Parr, 2003). This MDP incorporates a one-dimensional chain walk with four states (Figure 1). Two actions, "left"(L) and "right"(R), are available at every state. Rewards 1 and 0.5 are given when states '2' and '3' are visited, respectively.

We adopt the simplest direct representation of states; the state variable took $s = 1$, $s = 2$, $s = 3$ or $s = 4$,

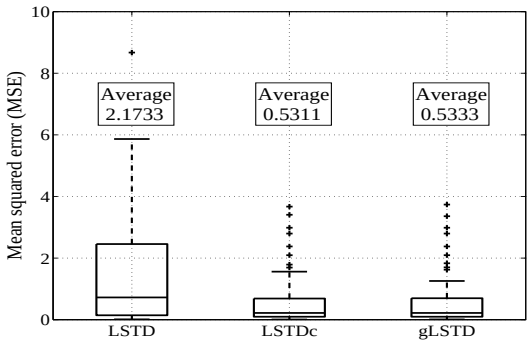

Figure 2. Simulation result.

when the corresponding state was visited. The value function was defined as the average reward, eq. (2), and was approximated by a linear function with a three-dimensional basis function: $\boldsymbol{\phi}(s) = [s, s^2, s^3]^\top$. The policy was set at random, and at the beginning of each episode an initial state was randomly selected according to the stationary distribution of this Markov chain.

Under these conditions, we performed 100 episodes each of which consisted of 100 random walk steps. We evaluated the "mean squared error" (MSE) of the value function, i.e., $\sum_{i\in\{1,2,3,4\}} d^\pi(i)|\hat{V}(i) - V^*(i)|^2$; where $\hat{V}$ and V^* denote $\hat{V}(i) = \boldsymbol{\phi}(s=i)^\top\hat{\boldsymbol{\theta}}$ and $V^*(i) = \boldsymbol{\phi}(s=i)^\top\boldsymbol{\theta}^*$, respectively.

Figure 2 shows box-plots of the MSEs of LSTD, LSTDc, and gLSTD. For this example, estimators of the conditional expectations in gLSTD can be calculated by sample average in each state, because there were only four discrete states. In continuous state problems, however, estimation of such conditional expectations would become much harder.

In Figure 2, the y-axis denotes the MSE of the value function. The center line and the upper and lower sides of each box denote the median of MSEs and the upper and lower quartiles, respectively. The number above each box represents the average MSE. There is significant difference between the MSE of LSTD and those of LSTDc and gLSTD. The estimators for LSTDc and for gLSTD both achieved a much smaller MSE than that for the ordinary LSTD.

5. Conclusion

In this study, we have discussed LSTD-based policy evaluation in the framework of semiparametric statistical inference. We showed that the standard LSTD algorithm is indeed an estimating function method which is guaranteed to be consistent regardless of the stochastic properties of the environments. Based on the optimal estimating functions in the two classes of estimating functions, we constructed two new policy evaluation methods called gLSTD and LSTDc. We also evaluated the asymptotic variance of the general instrumental variable methods for MDP. Moreover, we showed that the form of possible estimating functions for the value function estimation is restricted to be the same as those used in the instrumental variable methods. We then demonstrated, through an experiment using a simple MDP problem, that the gLSTD and LSTDc estimators reduce substantially the asymptotic variance of the LSTD estimator.

Further work is necessary to construct procedures for policy updating based on evaluation by gLSTD and LSTDc. It should be possible to incorporate our proposed ideas into the least-squares policy iteration (Lagoudakis & Parr, 2003) and the natural actor-critic method (Peters et al., 2005).

A. Proof of Theorem 1: The Optimal Instrumental Variable

As shown in eq. (20), the asymptotic variance of the estimator $\hat{\boldsymbol{\theta}}_{\boldsymbol{z}}$ is given by

$$\mathrm{AV}[\hat{\boldsymbol{\theta}}_{\boldsymbol{z}}] = \frac{1}{N}\boldsymbol{A}_{\boldsymbol{z}}^{-1}\boldsymbol{M}_{\boldsymbol{z}}\boldsymbol{A}_{\boldsymbol{z}}^{-\top},$$

where $\boldsymbol{A}_{\boldsymbol{z}} := \mathbb{E}_d[\boldsymbol{z}_t\boldsymbol{x}_t^\top]$ and $\boldsymbol{M}_{\boldsymbol{z}} := \mathbb{E}_d[(\epsilon_t^*)^2\boldsymbol{z}_t\boldsymbol{z}_t^\top]$. If we add a small change $\boldsymbol{\delta}_t(s_t, \cdots, s_{t-T})$ to the instrumental variable $\boldsymbol{z}_t$, the matrices become

$$\begin{aligned}\boldsymbol{A}_{\boldsymbol{z}+\boldsymbol{\delta}} &= \boldsymbol{A}_{\boldsymbol{z}} + \mathbb{E}_d[\boldsymbol{\delta}_t\boldsymbol{x}_t^\top],\\ \boldsymbol{M}_{\boldsymbol{z}+\boldsymbol{\delta}} &= \boldsymbol{M}_{\boldsymbol{z}} + \mathbb{E}_d[(\epsilon_t^*)^2(\boldsymbol{\delta}_t\boldsymbol{z}_t^\top + \boldsymbol{z}_t\boldsymbol{\delta}_t^\top)].\end{aligned}$$

Therefore, the deviation of the trace of asymptotic variance can be calculated as

$$\begin{aligned}&\mathrm{Tr}\left[\boldsymbol{A}_{\boldsymbol{z}+\boldsymbol{\delta}}^{-1}\boldsymbol{M}_{\boldsymbol{z}+\boldsymbol{\delta}}\boldsymbol{A}_{\boldsymbol{z}+\boldsymbol{\delta}}^{-\top}\right] - \mathrm{Tr}\left[\boldsymbol{A}_{\boldsymbol{z}}^{-1}\boldsymbol{M}_{\boldsymbol{z}}\boldsymbol{A}_{\boldsymbol{z}}^{-\top}\right]\\ =&-\mathrm{Tr}\left\{\boldsymbol{A}_{\boldsymbol{z}}^{-1}\mathbb{E}_d\left[\boldsymbol{\delta}_t\boldsymbol{x}_t^\top\right]\boldsymbol{A}_{\boldsymbol{z}}^{-1}\boldsymbol{M}_{\boldsymbol{z}}\boldsymbol{A}_{\boldsymbol{z}}^{-\top}\right\}\\ &-\mathrm{Tr}\left\{\boldsymbol{A}_{\boldsymbol{z}}^{-1}\mathbb{E}_d\left[\boldsymbol{x}_t\boldsymbol{\delta}_t^\top\right]\boldsymbol{A}_{\boldsymbol{z}}^{-1}\boldsymbol{M}_{\boldsymbol{z}}\boldsymbol{A}_{\boldsymbol{z}}^{-\top}\right\}\\ &+\mathrm{Tr}\left\{\boldsymbol{A}_{\boldsymbol{z}}^{-1}\mathbb{E}_d\left[(\epsilon_t^*)^2\left(\boldsymbol{z}_t\boldsymbol{\delta}_t^\top + \boldsymbol{\delta}\boldsymbol{z}_t^\top\right)\right]\boldsymbol{A}_{\boldsymbol{z}}^{-\top}\right\}\\ =&2\mathbb{E}_d\left[\boldsymbol{\delta}_t^\top\boldsymbol{A}_{\boldsymbol{z}}^{-\top}\boldsymbol{A}_{\boldsymbol{z}}^{-1}\mathbb{E}^\pi\left[(\epsilon_t^*)^2|s_t, \cdots, s_{t-T}\right]\boldsymbol{z}_t\right]\\ &-2\mathbb{E}_d\left[\boldsymbol{\delta}_t^\top\boldsymbol{A}_{\boldsymbol{z}}^{-\top}\boldsymbol{A}_{\boldsymbol{z}}^{-1}\boldsymbol{M}_{\boldsymbol{z}}\boldsymbol{A}_{\boldsymbol{z}}^{-\top}\mathbb{E}^\pi[\boldsymbol{x}_t|s_t, \cdots, s_{t-T}]\right].\end{aligned}$$

By using the condition that the deviation becomes $\mathbf{0}$ for any small change $\boldsymbol{\delta}_t(s_t, \cdots, s_{t-T})$, the optimal instrumental variable can be obtained as

$$\boldsymbol{z}_t^* = \mathbb{E}^\pi\left[(\epsilon_t^*)^2|s_t\right]^{-1}\boldsymbol{M}_{\boldsymbol{z}}\boldsymbol{A}_{\boldsymbol{z}}^{-\top}\mathbb{E}^\pi[\boldsymbol{x}_t|s_t].$$

Considering Lemma 3, the optimal instrumental variable is restricted as

$$\boldsymbol{z}_t^* = \mathbb{E}^\pi\left[(\epsilon_t^*)^2|s_t\right]^{-1}\mathbb{E}^\pi[\boldsymbol{x}_t|s_t],$$

or as its transformation by any regular matrix.

Now, we show that eq. (22) also satisfies the global optimality. Substituting $\boldsymbol{z}_t^*$ to the matrix $\boldsymbol{A}_{\boldsymbol{z}^*}$, we obtain

$$\boldsymbol{A}_{\boldsymbol{z}^*} = \boldsymbol{M}_{\boldsymbol{z}^*}\boldsymbol{A}_{\boldsymbol{z}^*}^{-\top}\boldsymbol{F},$$

where

$$\boldsymbol{F} := \mathbb{E}_d\left[\mathbb{E}^\pi[(\epsilon_t^*)^2|s_t]^{-1}\mathbb{E}^\pi[\boldsymbol{x}_t|s_t]\mathbb{E}^\pi[\boldsymbol{x}_t^\top|s_t]\right].$$

Furthermore, the matrices at $\boldsymbol{z}_t^* + \boldsymbol{\delta}_t$ become

$$\begin{aligned}\boldsymbol{A}_{\boldsymbol{z}^*+\boldsymbol{\delta}} &= \boldsymbol{M}_{\boldsymbol{z}^*}\boldsymbol{A}_{\boldsymbol{z}^*}^{-\top}\boldsymbol{F} + \mathbb{E}_d[\boldsymbol{x}_t\boldsymbol{\delta}_t^\top],\\ \boldsymbol{M}_{\boldsymbol{z}^*+\boldsymbol{\delta}} &= \boldsymbol{M}_{\boldsymbol{z}^*}\boldsymbol{A}_{\boldsymbol{z}^*}^{-\top}\boldsymbol{F}\boldsymbol{A}_{\boldsymbol{z}^*}^{-1}\boldsymbol{M}_{\boldsymbol{z}^*} + \boldsymbol{M}_{\boldsymbol{z}^*}\boldsymbol{A}_{\boldsymbol{z}^*}^{-\top}\mathbb{E}_d[\boldsymbol{x}_t\boldsymbol{\delta}_t^\top]\\ &\quad + \mathbb{E}_d[\boldsymbol{\delta}_t\boldsymbol{x}_t^\top]\boldsymbol{A}_{\boldsymbol{z}^*}^{-1}\boldsymbol{M}_{\boldsymbol{z}^*} + \mathbb{E}_d[(\epsilon_t^*)^2\boldsymbol{\delta}_t\boldsymbol{\delta}_t^\top].\end{aligned}$$

Therefore,

$$\begin{aligned}&\boldsymbol{A}_{\boldsymbol{z}^*+\boldsymbol{\delta}}^{-1}\boldsymbol{M}_{\boldsymbol{z}^*+\boldsymbol{\delta}}\boldsymbol{A}_{\boldsymbol{z}^*+\boldsymbol{\delta}}^{-\top} - \boldsymbol{A}_{\boldsymbol{z}^*}^{-1}\boldsymbol{M}_{\boldsymbol{z}^*}\boldsymbol{A}_{\boldsymbol{z}^*}^{-\top}\\ &= \boldsymbol{A}_{\boldsymbol{z}^*+\boldsymbol{\delta}}^{-1}(\boldsymbol{M}_{\boldsymbol{z}^*+\boldsymbol{\delta}} - \boldsymbol{A}_{\boldsymbol{z}^*+\boldsymbol{\delta}}\boldsymbol{A}_{\boldsymbol{z}^*}^{-1}\boldsymbol{M}_{\boldsymbol{z}^*}\boldsymbol{A}_{\boldsymbol{z}^*}^{-\top}\boldsymbol{A}_{\boldsymbol{z}^*+\boldsymbol{\delta}}^{\top})\boldsymbol{A}_{\boldsymbol{z}^*+\boldsymbol{\delta}}^{-\top}\\ &= \boldsymbol{A}_{\boldsymbol{z}^*+\boldsymbol{\delta}}^{-1}(\mathbb{E}_d[(\epsilon_t^*)^2\boldsymbol{\delta}_t\boldsymbol{\delta}_t^\top] - \mathbb{E}_d[\boldsymbol{\delta}_t\boldsymbol{x}_t^\top]\boldsymbol{F}^{-1}\mathbb{E}_d[\boldsymbol{x}_t\boldsymbol{\delta}_t^\top])\boldsymbol{A}_{\boldsymbol{z}^*+\boldsymbol{\delta}}^{-\top}\\ &= \boldsymbol{A}_{\boldsymbol{z}^*+\boldsymbol{\delta}}^{-1}\Big\{\mathbb{E}_d\Big[\mathbb{E}^\pi[(\epsilon_t^*)^2|s_t]\\ &\quad\times\left(\boldsymbol{\delta}_t - \mathbb{E}^\pi[(\epsilon_t^*)^2|s_t]^{-1}\mathbb{E}_d[\boldsymbol{\delta}_t\boldsymbol{x}_t^\top]\boldsymbol{F}^{-1}\mathbb{E}^\pi[\boldsymbol{x}_t|s_t]\right)\\ &\quad\times\left(\boldsymbol{\delta}_t - \mathbb{E}^\pi[(\epsilon_t^*)^2|s_t]^{-1}\mathbb{E}_d[\boldsymbol{\delta}_t\boldsymbol{x}_t^\top]\boldsymbol{F}^{-1}\mathbb{E}^\pi[\boldsymbol{x}_t|s_t]\right)^\top\Big]\\ &\quad\Big\}\boldsymbol{A}_{\boldsymbol{z}^*+\boldsymbol{\delta}}^{-\top} \succeq \boldsymbol{0}.\end{aligned}$$

The equality holds only when $\boldsymbol{\delta}_t(s_t) \propto \boldsymbol{z}_t^*$. (Q.E.D.)

B. Proof of Theorem 2: The Optimal $\boldsymbol{c}$

By differentiating the trace of eq. (20), the optimal constant $\boldsymbol{c}$ must satisfy

$$\mathbb{E}_d[(\epsilon_t^*)^2]\boldsymbol{c} + \mathbb{E}_d[(\epsilon_t^*)^2\boldsymbol{\phi}_t] = \boldsymbol{M}_{\boldsymbol{c}}\boldsymbol{A}_{\boldsymbol{c}}^{-\top}\mathbb{E}_d[\boldsymbol{x}_t].$$

Using the well-known matrix inversion lemma (Horn & Johnson, 1985), the solution can be obtained as eq. (23).

In addition, the global optimality among those applied by constant shifts can be proved using a similar argument to that in Appendix A. (Q.E.D.)

Acknowledgments

The authors wish to thank the anonymous reviewers for their helpful comments and suggestions.

References

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

Bertsekas, D., & Tsitsiklis, J. (1996). *Neuro-Dynamic Programming*. Athena Scientific.

Bickel, D., Ritov, D., Klaassen, C., & Wellner, J. (1998). *Efficient and Adaptive Estimation for Semiparametric Models*. Springer.

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

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

Godambe, V. (Ed.). (1991). *Estimating Functions*. Oxford Science.

Greensmith, E., Bartlett, P., & Baxter, J. (2004). Variance reduction techniques for gradient estimates in reinforcement learning. *Journal of Machine Learning Research*, *5*, 1471–1530.

Horn, R., & Johnson, C. (1985). *Matrix analysis*. Cambridge University Press.

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

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

Peters, J., Vijayakumar, S., & Schaal, S. (2005). Natural actor-critic. *Proceedings of the 16th European Conference on Machine Learning* (pp. 280–291).

Sutton, R., & Barto, A. (1998). *Reinforcement Learning: An Introduction*. MIT Press.

Young, P. (1984). *Recursive Estimation and Time-series Analysis*. Springer-Verlag.