Semiparametric Statistical Approach to Reinforcement Learning

Tsuyoshi Ueno

Japan Science and Technology
Minato Discrete Strucure Manipulation System Project

Summary of This Talk

Background

- Reinforcement learning (RL) = sampling-based stochastic optimal control
- An open issue in RL is to rigorously evaluate statistical properties of RL algorithms and compare their performance.

Contributions

1. Reformulate model-free policy evaluation, which is a key of RL, as a general **semiparametric statistical inference problem**
2. Derive a general class of consistent estimators which leads to almost all of model-free policy evaluation algorithms proposed so far
3. Propose a new estimator which minimizes the estimation variance in asymptotics among the general class

Outline

1. What is Reinforcement Learning ?
2. Introduction to RL Algorithms
3. Semiparametric Statistical Inference Approach to RL
4. Summary & Future Works

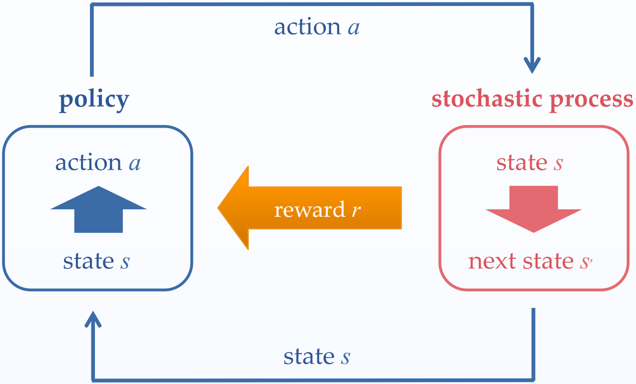
What is RL ?

What is RL ?

RL is a solution for optimal control for Markovian stochastic processes
by iterating between sampling and inference

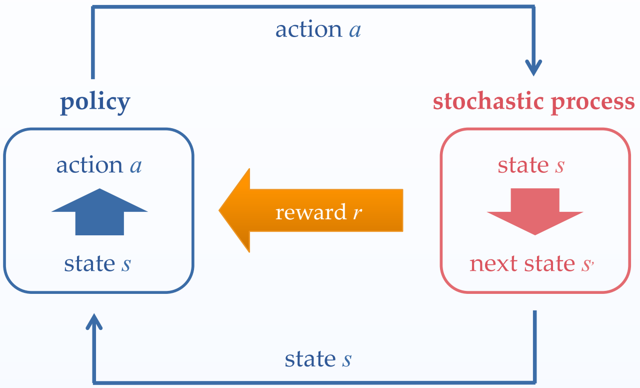
What is RL ?

RL is a solution for optimal control for Markovian stochastic processes
by iterating between sampling and inference



What is RL ?

RL is a solution for optimal control for Markovian stochastic processes
by iterating between sampling and inference



Infer the best policy that maximizes some measure incorporating long-term future rewards from sequences of states, actions and rewards

Optimal Control vs RL

Classical Optimal Control Scheme

1. **System Identification**
   Identify the stochastic process using a statistical model
2. **Dynamic Programming** [Bellman, 1957]
   Optimize the policy based on the identified model

Optimal Control vs RL

Classical Optimal Control Scheme

1. **System Identification**
   Identify the stochastic process using a statistical model
2. **Dynamic Programming** [Bellman, 1957]
   Optimize the policy based on the identified model

RL Scheme

Iterate the following two steps until the convergence:

1. **Sampling**
   Generate the sequence under the current policy
2. **Inference for Policy**
   Infer the better policy than the current one from the sequence directly

Optimal Control vs RL

Classical Optimal Control Scheme

1. **System Identification**
   Identify the stochastic process using a statistical model
2. **Dynamic Programming** [Bellman, 1957]
   Optimize the policy based on the identified model

RL Scheme

Iterate the following two steps until the convergence:

1. **Sampling**
   Generate the sequence under the current policy
2. **Inference for Policy**
   Infer the better policy than the current one from the sequence directly

RL can find the optimal policy without the system identification.

Acrobot Swing-up Task [Yoshimoto et al., 2005]

- Acrobot: 2 link 1 actuator
- Goal: make the acrobot stand upside-down at the top
- Reward: take the higher value when the acrobot is close to standing up

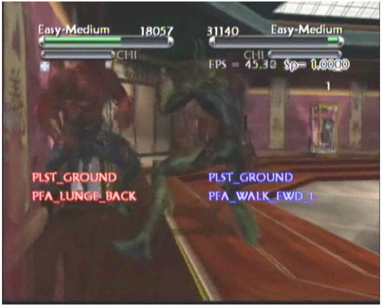
(befor)

(after)

Fighting Game
[Graepel et al., 2004]

- Video game for Xbox: Tao feng published by Microsoft
- Learn the non player character's motions by RL

Reward: damage of enemy
(before) (after)



Reward: damage of enemy with the remained life of learning agent
(after)

Outline

Optimal Control Problem

Markov Decision Processes (MDPs)

- state $s \in S$
- action $a \in A$
- state transition distribution $p(s_t | s_{t-1}, a_{t-1})$
- reward function $r_{t+1} := r(s_t, a_t, s_{t+1})$
- policy $\pi(a_t | s_t) := p(a_t | s_t)$

Optimal Control Problem

Markov Decision Processes (MDPs)

- state $s \in S$
- action $a \in A$
- state transition distribution $p(s_t | s_{t-1}, a_{t-1})$
- reward function $r_{t+1} := r(s_t, a_t, s_{t+1})$
- policy $\pi(a_t | s_t) := p(a_t | s_t)$

Goal of Optimal Control

Find the optimal policy that maximizes the value function $V_\pi(s)$ for any $s \in S$

$$V_\pi(s) := \lim_{T \to \infty} \mathbb{E}_\pi\left(\sum_{t'=t}^{T} \beta^{t-t'} r_{t+1} \,\middle|\, s_t = s \right) \quad \text{where } \beta \in [0, 1)$$

How to Solve Optimal Control Problems

Policy Iteration [Howard, 1960]

How to Solve Optimal Control Problems

Policy Iteration [Howard, 1960]

- Is the common mathmatical basis for both optimal control and RL

How to Solve Optimal Control Problems

Policy Iteration [Howard, 1960]

- Is the common mathmatical basis for both optimal control and RL
- Iterate the following two procedures:

How to Solve Optimal Control Problems

Policy Iteration [Howard, 1960]

- Is the common mathmatical basis for both optimal control and RL
- Iterate the following two procedures:
  1. **Policy Evaluation**
     Evaluate the value function under the current policy based on the identified model $\widehat{p}(s'|s,a)$

How to Solve Optimal Control Problems

Policy Iteration [Howard, 1960]

- Is the common mathmatical basis for both optimal control and RL
- Iterate the following two procedures:
  1. **Policy Evaluation**
     Evaluate the value function under the current policy based on the identified model $\widehat{p}(s'|s, a)$
  2. **Policy Improvement**
     Update the policy so as to maximize the value function

How to Solve Optimal Control Problems

Policy Iteration [Howard, 1960]

- Is the common mathmatical basis for both optimal control and RL
- Iterate the following two procedures:
  1. **Policy Evaluation**
     Evaluate the value function under the current policy based on the identified model $\widehat{p}(s'|s,a)$
  2. **Policy Improvement**
     Update the policy so as to maximize the value function
- Converge the optimal policy
  as long as the value function can be exactly evaluated

How to Solve Optimal Control Problems

Policy Iteration based RL

- Iterate the following two procedures:
  1. **Model-free Policy Evaluation**
     Estimate the value function from the sequence of observations directly
  2. **Policy Improvement**
     Update the policy so as to maximize the value function
- Converge the optimal policy
  as long as the value function can be exactly evaluated

How to Solve Optimal Control Problems

Policy Iteration based RL

- Iterate the following two procedures:
  1. **Model-free Policy Evaluation**
     Estimate the value function from the sequence of observations directly
  2. **Policy Improvement**
     Update the policy so as to maximize the value function
- Converge the optimal policy
  as long as the value function can be exactly evaluated

All of current model-free policy evaluation algorithms were constructed based on TD learning

Model-free Policy Evaluation: TD Learning [Sutton, 1984]

Model for Value Function

Assume that $V_\pi(s)$ can be represented by a parametric model $g(s, \theta)$:

$$V_\pi(s) \approx g(s, \theta) \quad \text{for any } s \in S.$$

Model-free Policy Evaluation: TD Learning [Sutton, 1984]

Model for Value Function

Assume that $V_\pi(s)$ can be represented by a parametric model $g(s, \theta)$:

$$V_\pi(s) \approx g(s, \theta) \quad \text{for any } s \in S.$$

Bellman Equation

- Express the value function as

$$\begin{aligned} V_\pi(s_t) &= \lim_{T \to \infty} \mathbb{E}_\pi\left(\sum_{t'=t}^{T} \beta^{t'-t} r_{t'+1} \middle| s_t \right) \\ &= \mathbb{E}_\pi\left(r_{t+1} \middle| s_t \right) + \beta \cdot \lim_{T \to \infty} \mathbb{E}_\pi\left(\sum_{t'=t+1}^{T} \beta^{t'-t-1} r_{t'+1} \middle| s_t \right) \end{aligned}$$

Model-free Policy Evaluation: TD Learning [Sutton, 1984]

Model for Value Function

Assume that $V_\pi(s)$ can be represented by a parametric model $g(s, \theta)$:

$$V_\pi(s) \approx g(s, \theta) \quad \text{for any } s \in S.$$

Bellman Equation

- Express the value function as

$$\begin{aligned} V_\pi(s_t) &= \lim_{T \to \infty} \mathbb{E}_\pi \left(\sum_{t'=t}^{T} \beta^{t'-t} r_{t'+1} \middle| s_t \right) \\ &= \mathbb{E}_\pi \left(r_{t+1} | s_t \right) + \beta \cdot \lim_{T \to \infty} \mathbb{E}_\pi \left(\sum_{t'=t+1}^{T} \beta^{t'-t-1} r_{t'+1} \middle| s_t \right) \end{aligned}$$

- Derive the Bellman equation by

$$V_\pi(s_t) = \mathbb{E}_\pi \left(r_{t+1} | s_t \right) + \beta \mathbb{E}_\pi \left(V_\pi(s_{t+1}) | s_t \right)$$

$$g(s_t, \theta) = \mathbb{E}_\pi(r_{t+1} | s_t) + \beta \mathbb{E}_\pi(g(s_{t+1}, \theta) | s_t)$$

Model-free Policy Evaluation: TD Learning [Sutton, 1984]

Model-free Policy Evaluation: TD Learning [Sutton, 1984]

- Define the **temporal difference (TD) error** as

$$\varepsilon(z_t, \theta) := r_{t+1} + \beta g(s_{t+1}, \theta) - g(s_t, \theta), \quad \text{where } z_t := (s_{t-1}, s_t, r_t).$$

Model-free Policy Evaluation: TD Learning [Sutton, 1984]

- Define the **temporal difference (TD) error** as

$$\varepsilon(z_t, \theta) := r_{t+1} + \beta g(s_{t+1}, \theta) - g(s_t, \theta), \quad \text{where } z_t := (s_{t-1}, s_t, r_t).$$

- Satisfy $\mathbb{E}_\pi(\varepsilon(z_t, \theta)|s_t) = 0$ for any $s_t \in S$ because

$$\mathbb{E}_\pi(\varepsilon(z_t, \theta)|s_t) = \mathbb{E}_\pi(r_{t+1} + \beta g(s_{t+1}, \theta)|s_t) - g(s_t, \theta) = g(s_t, \theta) - g(s_t, \theta) = 0,$$

where we used the Bellman equation $g(s_t, \theta) = \mathbb{E}_\pi(r_{t+1} + \beta g(s_{t+1})|s_t)$

Model-free Policy Evaluation: TD Learning [Sutton, 1984]

- Define the **temporal difference (TD) error** as

$$\varepsilon(z_t, \theta) := r_{t+1} + \beta g(s_{t+1}, \theta) - g(s_t, \theta), \quad \text{where } z_t := (s_{t-1}, s_t, r_t).$$

- Satisfy $\mathbb{E}_\pi(\varepsilon(z_t, \theta)|s_t) = 0$ for any $s_t \in S$ because

$$\mathbb{E}_\pi(\varepsilon(z_t, \theta)|s_t) = \mathbb{E}_\pi(r_{t+1} + \beta g(s_{t+1}, \theta)|s_t) - g(s_t, \theta) = g(s_t, \theta) - g(s_t, \theta) = 0,$$

  where we used the Bellman equation $g(s_t, \theta) = \mathbb{E}_\pi(r_{t+1} + \beta g(s_{t+1})|s_t)$

- Update the parameter incrementally by stochastic gradient descent:

$$\widehat{\theta}_{t+1} = \widehat{\theta}_t - \alpha_t \partial_\theta g(s_t, \theta)|_{\theta = \widehat{\theta}_t} \cdot \varepsilon(z_t, \widehat{\theta}_t), \text{ where } \alpha_t \text{ is the stepsize parameter}$$

Model-free Policy Evaluation: TD Learning [Sutton, 1984]

- Define the **temporal difference (TD) error** as

$$\varepsilon(z_t, \theta) := r_{t+1} + \beta g(s_{t+1}, \theta) - g(s_t, \theta), \quad \text{where } z_t := (s_{t-1}, s_t, r_t).$$

- Satisfy $\mathbb{E}_\pi(\varepsilon(z_t, \theta) | s_t) = 0$ for any $s_t \in S$ because

$$\mathbb{E}_\pi(\varepsilon(z_t, \theta)|s_t) = \mathbb{E}_\pi(r_{t+1} + \beta g(s_{t+1}, \theta)|s_t) - g(s_t, \theta) = g(s_t, \theta) - g(s_t, \theta) = 0,$$

  where we used the Bellman equation $g(s_t, \theta) = \mathbb{E}_\pi(r_{t+1} + \beta g(s_{t+1})|s_t)$

- Update the parameter incrementally by stochastic gradient descent:

$$\widehat{\theta}_{t+1} = \widehat{\theta}_t - \alpha_t \partial_\theta g(s_t, \theta)\big|_{\theta=\widehat{\theta}_t} \cdot \varepsilon(z_t, \widehat{\theta}_t), \text{ where } \alpha_t \text{ is the stepsize parameter}$$

- Converges to the true parameter
  that can represent the value function under some conditions

Model-free Policy Evaluation: TD Learning [Sutton, 1984]

- Define the **temporal difference (TD) error** as

$$\varepsilon(z_t, \theta) := r_{t+1} + \beta g(s_{t+1}, \theta) - g(s_t, \theta), \quad \text{where } z_t := (s_{t-1}, s_t, r_t).$$

- Satisfy $\mathbb{E}_\pi(\varepsilon(z_t, \theta) | s_t) = 0$ for any $s_t \in S$ because

$$\mathbb{E}_\pi(\varepsilon(z_t, \theta) | s_t) = \mathbb{E}_\pi(r_{t+1} + \beta g(s_{t+1}, \theta) | s_t) - g(s_t, \theta) = g(s_t, \theta) - g(s_t, \theta) = 0,$$

  where we used the Bellman equation $g(s_t, \theta) = \mathbb{E}_\pi(r_{t+1} + \beta g(s_{t+1}) | s_t)$

- Update the parameter incrementally by stochastic gradient descent:

$$\widehat{\theta}_{t+1} = \widehat{\theta}_t - \alpha_t \partial_\theta g(s_t, \theta)|_{\theta = \widehat{\theta}_t} \cdot \varepsilon(z_t, \widehat{\theta}_t), \text{ where } \alpha_t \text{ is the stepsize parameter}$$

- Converges to the true parameter
  that can represent the value function under some conditions

TD learning does not require any knowledge of the stochastic process, but can estimate the value function exactly.

Extensions of TD Learning

Online algorithms

- TD [Sutton, 1984]
- TD(λ) [Sutton and Barto, 1998]
- LSPE [Nedić and Bertsekas, 2003]
- iLSTD [Geramifard et al., 2006]
- RG [Baird, 1995]
- TDC [Sutton et al., 2009a]
- GTD [Sutton et al., 2009b]
- GTD2 [Sutton et al., 2009a]

Batch algorithms

- LSTD [Bradtke and Barto, 1996]
- LSTD(λ) [Boyan, 2002]
- LSTDc [Ueno et al., 2008]

Extensions of TD Learning

Online algorithms

- TD [Sutton, 1984]
- TD(λ) [Sutton and Barto, 1998]
- LSPE [Nedić and Bertsekas, 2003]
- iLSTD [Geramifard et al., 2006]
- RG [Baird, 1995]
- TDC [Sutton et al., 2009a]
- GTD [Sutton et al., 2009b]
- GTD2 [Sutton et al., 2009a]

Batch algorithms

- LSTD [Bradtke and Barto, 1996]
- LSTD(λ) [Boyan, 2002]
- LSTDc [Ueno et al., 2008]

- The validation of the performance of their proposed algorithms has been performed only in numerical experiments.
- The methodology for evaluating and comparing the performance of various policy evaluation algorithms has not been established yet.

Challenge

Challenge

Analyze the statistical properties of the variously-presented model-free policy evaluation algorithms in a unified way and derive the optimal model-free policy evaluation algorithm

Challenge

Analyze the statistical properties of the variously-presented model-free policy evaluation algorithms in a unified way and derive the optimal model-free policy evaluation algorithm

Main Idea

Challenge

Analyze the statistical properties of the variously-presented model-free policy evaluation algorithms in a unified way and derive the optimal model-free policy evaluation algorithm

Main Idea

- Reformulate the model-free policy evaluation as a general semiparametric statistical inference problem

Challenge

Analyze the statistical properties of the variously-presented model-free policy evaluation algorithms in a unified way and derive the optimal model-free policy evaluation algorithm

Main Idea

- Reformulate the model-free policy evaluation as a general semiparametric statistical inference problem
- Enable to apply various analysis techniques to the problem of estimating the value function, and to discuss theoretical properties which are common over model-free policy evaluation problems

Outline

Preliminary

Discrete-time Markov Reward Process (MRP)

- state $s \in S$ (S is a discrete state space)
- reward $r \in \mathbb{R}$
- state transition probability $p(s_t|s_{t-1})$
- reward probability $p(r_t|s_{t-1}, s_t)$

Preliminary

Discrete-time Markov Reward Process (MRP)

- state $s \in S$ (S is a discrete state space)
- reward $r \in \mathbb{R}$
- state transition probability $p(s_t | s_{t-1})$
- reward probability $p(r_t | s_{t-1}, s_t)$

Value Function

Define the value function as

$$V(s) := \mathbb{E}\left(\sum_{t'=t}^{\infty} \beta^{t'-t} r_{t+1} \middle| s_t = s\right), \quad \text{where } \beta \in [0,1)$$

Preliminary

Discrete-time Markov Reward Process (MRP)

- state $s \in S$ (S is a discrete state space)
- reward $r \in \mathbb{R}$
- state transition probability $p(s_t | s_{t-1})$
- reward probability $p(r_t | s_{t-1}, s_t)$

Value Function

Define the value function as

$$V(s) := \mathbb{E}\left(\sum_{t'=t}^{\infty} \beta^{t'-t} r_{t+1} \middle| s_t = s \right), \quad \text{where } \beta \in [0, 1)$$

Model for Value Function

Characterize the value function by a parametric model $g(s, \theta)$: $V(s) \approx g(s, \theta)$

Assumptions

1. The MRP satisfies ergodicity.
2. The model $g(s, \theta)$ can completely represent the value function: $V(s) = g(s, \theta)$ for any $s \in S$.

Assumptions

1. The MRP satisfies ergodicity.
2. The model $g(s,\theta)$ can completely represent the value function: $V(s) = g(s,\theta)$ for any $s \in S$.

We do not consider the model error here, and focus solely on the estimation error of the parameter.

Assumptions

1. The MRP satisfies ergodicity.
2. The model $g(s, \theta)$ can completely represent the value function: $V(s) = g(s, \theta)$ for any $s \in S$.

We do not consider the model error here, and focus solely on the estimation error of the parameter.

Model-Free Policy Evaluation Problem on MRPs

Given an initial state s_0,
the sequence of states and rewards $Z_T := \left((s_t)_{t=0}^{T}, (r_t)_{t=1}^{T}\right)$ is obtained by

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

Assumptions

1. The MRP satisfies ergodicity.
2. The model $g(s, \theta)$ can completely represent the value function: $V(s) = g(s, \theta)$ for any $s \in S$.

**We do not consider the model error here,
and focus solely on the estimation error of the parameter.**

Model-Free Policy Evaluation Problem on MRPs

Given an initial state s_0,
the sequence of states and rewards $Z_T := \left((s_t)_{t=0}^T, (r_t)_{t=1}^T\right)$ is obtained by

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

Then, we estimate the parameter θ from the sample sequence Z_T
without identifying $p(r_t, s_t | s_{t-1})$.

Reformulation of Model-free Policy Evaluation as Semiparametric Inference

Reformulation of Model-free Policy Evaluation as Semiparametric Inference

Bellman Equation

Recall that

$$V(s_{t-1}) = \mathbb{E}(r_t|s_{t-1}) + \beta\mathbb{E}(V(s_t)|s_{t-1}) \quad \forall s_{t-1} \in S$$

$$\Rightarrow \mathbb{E}(r_t|s_{t-1}) = g(s_{t-1}, \theta) - \beta\mathbb{E}(g(s_t, \theta)|s_{t-1}) \tag{1}$$

Reformulation of Model-free Policy Evaluation as Semiparametric Inference

Bellman Equation

Recall that

$$\begin{aligned} V(s_{t-1}) &= \mathbb{E}(r_t|s_{t-1}) + \beta\mathbb{E}(V(s_t)|s_{t-1}) \quad \forall s_{t-1} \in S \\ \Rightarrow \mathbb{E}(r_t|s_{t-1}) &= g(s_{t-1}, \theta) - \beta\mathbb{E}(g(s_t, \theta)|s_{t-1}) \end{aligned} \tag{1}$$

Semi-parameterization of MRPs

Reformulation of Model-free Policy Evaluation as Semiparametric Inference

Bellman Equation

Recall that

$$\begin{aligned} V(s_{t-1}) &= \mathbb{E}(r_t|s_{t-1}) + \beta\mathbb{E}(V(s_t)|s_{t-1}) \quad \forall s_{t-1} \in S \\ \Rightarrow \mathbb{E}(r_t|s_{t-1}) &= g(s_{t-1}, \theta) - \beta\mathbb{E}(g(s_t, \theta)|s_{t-1}) \end{aligned} \tag{1}$$

Semi-parameterization of MRPs

1. Specify the first-order moment $\mathbb{E}(r_t|s_{t-1})$ of $p(r_t, s_t|s_{t-1})$ by the parameter θ through **Bellman equation** (1)

Reformulation of Model-free Policy Evaluation as Semiparametric Inference

Bellman Equation

Recall that

$$\begin{aligned} V(s_{t-1}) &= \mathbb{E}(r_t|s_{t-1}) + \beta\mathbb{E}(V(s_t)|s_{t-1}) \quad \forall s_{t-1} \in S \\ \Rightarrow \mathbb{E}(r_t|s_{t-1}) &= g(s_{t-1}, \theta) - \beta\mathbb{E}(g(s_t, \theta)|s_{t-1}) \end{aligned} \tag{1}$$

Semi-parameterization of MRPs

1. Specify the first-order moment $\mathbb{E}(r_t|s_{t-1})$ of $p(r_t, s_t|s_{t-1})$ by the parameter θ through **Bellman equation (1)**
2. Specify the other moments by **nuisance parameters** η

Reformulation of Model-free Policy Evaluation as Semiparametric Inference

Semiparametric Model

The semiparametric model of MRP is given by

$$p_{\theta,\eta}(Z_T) = \prod_{t=1}^{T} p_{\theta,\eta}(s_t, r_t | s_{t-1})$$

$$\text{s.t. } \mathbb{E}_{\theta,\eta}\left(r_t | s_{t-1}\right) = g(s_{t-1}, \theta) - \beta \mathbb{E}_{\theta,\eta}\left(g(s_t, \theta) | s_{t-1}\right),$$

where $\mathbb{E}_{\theta,\eta}(\cdot | s_{t-1})$ is the expectation with respect to $p_{\theta,\eta}(r_t, s_t | s_{t-1})$.

Reformulation of Model-free Policy Evaluation as Semiparametric Inference

Semiparametric Model

The semiparametric model of MRP is given by

$$p_{\theta,\eta}(Z_T) = \prod_{t=1}^{T} p_{\theta,\eta}(s_t, r_t | s_{t-1})$$

$$\text{s.t. } \mathbb{E}_{\theta,\eta}\left(r_t | s_{t-1}\right) = g(s_{t-1}, \theta) - \beta \mathbb{E}_{\theta,\eta}\left(g(s_t, \theta) | s_{t-1}\right),$$

where $\mathbb{E}_{\theta,\eta}(\cdot | s_{t-1})$ is the expectation with respect to $p_{\theta,\eta}(r_t, s_t | s_{t-1})$.

Semiparametric Inference Problem

Given an initial state s_0, the sequence of states and rewards $Z_T := \left((s_t)_{t=0}^{T}, (r_t)_{t=1}^{T}\right)$ is obtained by

$$Z_T \sim p_{\theta,\eta}(Z_T) := \prod_{t=1}^{T-1} p_{\theta,\eta}(r_t, s_t | s_{t-1}).$$

Then, we estimate the parameter θ from the sample sequence Z_T without knowing η.

How to Solve ?

How to Solve ?

Martingale Estimating Function [Godambe, 1991]

The function $f_T(Z_T, \theta) = \sum_{t=1}^{T} \psi_t(Z_t, \theta)$ is called **martingale estimating function** when $\psi_t(Z_t, \theta)$ satisfies

$$\mathbb{E}_{\theta,\eta}\left(\psi_t(Z_t, \theta) | Z_{t-1}\right) = 0, \quad \text{for any } \theta, \eta \text{ and } t.$$

How to Solve ?

Martingale Estimating Function [Godambe, 1991]

The function $f_T(Z_T, \theta) = \sum_{t=1}^{T} \psi_t(Z_t, \theta)$ is called **martingale estimating function** when $\psi_t(Z_t, \theta)$ satisfies

$$\mathbb{E}_{\theta,\eta}\left(\psi_t(Z_t, \theta)|Z_{t-1}\right) = 0, \quad \text{for any } \theta, \eta \text{ and } t.$$

M-estimators

If there is a martingale estimating function, we can obtain a consistent estiamtor $\widehat{\theta}_T := \widehat{\theta}_T(Z_T)$, so-called **M-estimator**, by solving the following estimating equation:

$$\sum_{t=1}^{T} \psi_t(Z_t, \widehat{\theta}_T) = 0.$$

Design of Estimating Functions

Temporal Difference (TD) Error

$$\varepsilon(z_t, \theta) := r_t + \beta g(s_t, \theta) - g(s_{t-1}, \theta), \text{ where } z_t := (s_{t-1}, s_t, r_t).$$

Design of Estimating Functions

Temporal Difference (TD) Error

$$\varepsilon(z_t, \theta) := r_t + \beta g(s_t, \theta) - g(s_{t-1}, \theta), \quad \text{where } z_t := (s_{t-1}, s_t, r_t).$$

- TD error satisfies $\mathbb{E}_{\theta,\eta}(\varepsilon(z_t, \theta) | Z_{t-1}) = 0$ for any θ, η and t.

Design of Estimating Functions

Temporal Difference (TD) Error

$$\varepsilon(z_t, \theta) := r_t + \beta g(s_t, \theta) - g(s_{t-1}, \theta), \;\text{ where } z_t := (s_{t-1}, s_t, r_t).$$

- TD error satisfies $\mathbb{E}_{\theta,\eta}(\varepsilon(z_t, \theta) | Z_{t-1}) = 0$ for any θ, η and t.
- Zero mean property holds
  even when multiplied by any weight function $w_{t-1} := w_{t-1}(Z_{t-1}, \theta)$:

$$\mathbb{E}_{\theta,\eta}(w_{t-1} \cdot \varepsilon(z_t, \theta) | Z_{t-1}) = w_{t-1} \cdot \mathbb{E}_{\theta,\eta}(\varepsilon(z_t, \theta) | Z_{t-1}) = 0, \;\text{ for any } \theta, \eta \text{ and } t.$$

Design of Estimating Functions

Temporal Difference (TD) Error

$$\varepsilon(z_t, \theta) := r_t + \beta g(s_t, \theta) - g(s_{t-1}, \theta), \;\; \text{where } z_t := (s_{t-1}, s_t, r_t).$$

- TD error satisfies $\mathbb{E}_{\theta,\eta}(\varepsilon(z_t, \theta)|Z_{t-1}) = 0$ for any θ, η and t.
- Zero mean property holds even when multiplied by any weight function $w_{t-1} := w_{t-1}(Z_{t-1}, \theta)$:

$$\mathbb{E}_{\theta,\eta}(w_{t-1} \cdot \varepsilon(z_t, \theta)|Z_{t-1}) = w_{t-1} \cdot \mathbb{E}_{\theta,\eta}(\varepsilon(z_t, \theta)|Z_{t-1}) = 0, \;\; \text{for any } \theta, \eta \text{ and } t.$$

$f(Z_t, \theta) = \sum_{t=1}^{T} w_{t-1}(Z_{t-1}, \theta) \cdot \varepsilon(z_t, \theta)$ **is a candidate of martingale estimating functions.**

Main Results

Theorem 1

Any martingale estimating functions in semiparametric model $\{p_{\theta,\eta}(Z_T)|\theta,\eta\}$ can be expressed as

$$f_T(Z_T,\theta) = \sum_{t=1}^{T} \underbrace{w_{t-1}(Z_{t-1},\theta)}_{\text{weight}} \cdot \underbrace{\varepsilon(z_t,\theta)}_{\text{TD error}}.$$

Main Results

Theorem 1

Any martingale estimating functions in semiparametric model $\{p_{\theta,\eta}(Z_T)|\theta,\eta\}$ can be expressed as

$$f_T(Z_T,\theta) = \sum_{t=1}^{T} \underbrace{w_{t-1}(Z_{t-1},\theta)}_{\text{weight}} \cdot \underbrace{\varepsilon(z_t,\theta)}_{\text{TD error}}.$$

This estimating function generalizes almost all of the conventional model-free policy evaluation algorithms.

Extensions of TD Learning

Online algorithms

- TD [Sutton, 1984]
- TD(λ) [Sutton and Barto, 1998]
- LSPE [Nedić and Bertsekas, 2003]
- iLSTD [Geramifard et al., 2006]
- RG [Baird, 1995]
- TDC [Sutton et al., 2009a]
- GTD [Sutton et al., 2009b]
- GTD2 [Sutton et al., 2009a]

Batch algorithms

- LSTD [Bradtke and Barto, 1996]
- LSTD(λ) [Boyan, 2002]
- LSTDc [Ueno et al., 2008]

Extensions of TD Learning

Online algorithms

- TD [Sutton, 1984]
- TD(λ) [Sutton and Barto, 1998]
- LSPE [Nedić and Bertsekas, 2003]
- iLSTD [Geramifard et al., 2006]
- RG [Baird, 1995]
- TDC [Sutton et al., 2009a]
- GTD [Sutton et al., 2009b]
- GTD2 [Sutton et al., 2009a]

Batch algorithms

- LSTD [Bradtke and Barto, 1996]
- LSTD(λ) [Boyan, 2002]
- LSTDc [Ueno et al., 2008]

$$w_t = \partial g(s_t, \theta) \qquad w_t = \sum_{t'=1}^{t} \lambda^{t-t'} \partial g(s_t, \theta)$$

$$w_t = \mathbb{E}_{\theta^\star, \eta^\star}\left[\partial \varepsilon(z_t, \theta) | s_{t-1}\right] \qquad w_t = g(s_t, \theta) + c$$

The variation of the weight functions lead to many major model-free policy evaluation algorithms

Main Results

Lemma 2

Suppose that sample sequence Z_T is generated by $p_{\theta^\star,\eta^\star}(\mathrm{Z}_T)$.
Also suppose that the estimator $\widehat{\theta}_T$ is obtained from

$$\sum_{t=1}^{T} w_{t-1}(\mathrm{Z}_{t-1},\widehat{\theta}_T)\cdot\varepsilon(\mathrm{z}_t,\widehat{\theta}_T)=0. \tag{2}$$

Then, under reasonable assumptions, we have

$$\sqrt{T}\left(\widehat{\theta}_T-\theta^\star\right)\sim\mathcal{N}\left(0,\mathrm{Av}(\widehat{\theta}_T)\right),$$

where $\mathrm{Av}(\widehat{\theta}_T):=\mathbb{E}_{\theta^\star,\eta^\star}((\widehat{\theta}_T-\theta^\star)(\widehat{\theta}_T-\theta^\star)^\top)=A^{-1}MA^{-\top}$
is the estimation variance.

Main Results

Lemma 2

Suppose that sample sequence Z_T is generated by $p_{\theta^\star,\eta^\star}(\mathrm{Z}_T)$.
Also suppose that the estimator $\widehat{\theta}_T$ is obtained from

$$\sum_{t=1}^{T} w_{t-1}(\mathrm{Z}_{t-1},\widehat{\theta}_T)\cdot\varepsilon(\mathrm{z}_t,\widehat{\theta}_T)=0. \tag{2}$$

Then, under reasonable assumptions, we have

$$\sqrt{T}\left(\widehat{\theta}_T-\theta^\star\right)\sim\mathcal{N}\left(0,\mathrm{Av}(\widehat{\theta}_T)\right),$$

where $\mathrm{Av}(\widehat{\theta}_T):=\mathbb{E}_{\theta^\star,\eta^\star}((\widehat{\theta}_T-\theta^\star)(\widehat{\theta}_T-\theta^\star)^\top)=A^{-1}MA^{-\top}$
is the estimation variance.

The optimal estimator among the class of estimators given by Eq. (2) can be derived by minimizing $\mathrm{Av}(\widehat{\theta}_T)$.

Main Results

Theorem 3

The martingale estimating function with the minimum estimation variance is given by

$$f_T^\star(Z_T, \theta) := \sum_{t=1}^T w_t^\star(s_{t-1}, \theta^\star) \cdot \varepsilon(z_t, \theta),$$

where

$$w_t^\star(s_{t-1}, \theta^\star) := \frac{\mathbb{E}_{\theta^\star, \eta^\star}(\partial_\theta \varepsilon(z_t, \theta)|_{\theta=\theta^\star} | s_{t-1})}{\mathbb{E}_{\theta^\star, \eta^\star}(\varepsilon(z_t, \theta^\star)^2 | s_{t-1})}.$$

Main Results

Theorem 3

The martingale estimating function with the minimum estimation variance is given by

$$f_T^\star(Z_T, \theta) := \sum_{t=1}^{T} w_t^\star(s_{t-1}, \theta^\star) \cdot \varepsilon(z_t, \theta),$$

where

$$w_t^\star(s_{t-1}, \theta^\star) := \frac{\mathbb{E}_{\theta^\star,\eta^\star}(\partial_\theta \varepsilon(z_t, \theta)|_{\theta=\theta^\star} | s_{t-1})}{\mathbb{E}_{\theta^\star,\eta^\star}(\varepsilon(z_t, \theta^\star)^2 | s_{t-1})}.$$

- The true parameter $\theta^\star$ and the conditional expectation $\mathbb{E}_{\theta^\star,\eta^\star}(\cdot|s)$ are unknown.

Main Results

Theorem 3

The martingale estimating function with the minimum estimation variance is given by

$$f_T^\star(Z_T,\theta) := \sum_{t=1}^{T} w_t^\star(s_{t-1},\theta^\star)\cdot\varepsilon(z_t,\theta),$$

where

$$w_t^\star(s_{t-1},\theta^\star) := \frac{\mathbb{E}_{\theta^\star,\eta^\star}(\partial_\theta\varepsilon(z_t,\theta)|_{\theta=\theta^\star}|s_{t-1})}{\mathbb{E}_{\theta^\star,\eta^\star}(\varepsilon(z_t,\theta^\star)^2|s_{t-1})}.$$

- The true parameter $\theta^\star$ and the conditional expectation $\mathbb{E}_{\theta^\star,\eta^\star}(\cdot|s)$ are unknown.

We have proposed online and batch approximation methods

See the details in [Ueno et al., 2011].

Outline

Summary

- Introduced a framework of semiparametric statistical inference for policy evaluation which can be applied to analyzing statistical properties for policy evaluation
- Derived the general form of estimating function for policy evaluation in MRPs, which provides a statistical basis to many model-free policy evaluation algorithms
- Found an estimating function which yields the minimum asymptotic estimation variance among the general class

Future Directions

- **Robustness**
  Propose estimators for the value function which provide robustness against unpredictable outliers
- **Model Selection**
  Construct the scheme for selecting an appropriate model for the value function from observations
- **Asymptotic Behavior of Policy Improvement**
  Analyze statistical properties not only for estimating the value function, but also for estimating the policy

Collaborators

- Shin Ishii (Kyoto University)
- Shin-ichi Maeda (Kyoto University)
- Motoaki Kawanabe (ATR)
- Mori Takeshi

Reference I

[Baird, 1995] Baird, L. (1995). Residual algorithms: Reinforcement learning with function approximation. In International Conference on Machine Learning, pages 30--37.

[Bellman, 1957] Bellman, R. E. (1957). Dynamic Programming. Princeton University Press.

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

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

[Geramifard et al., 2006] Geramifard, A., Bowling, M., and Sutton, R. S. (2006). Incremental least-squares temporal difference learning. In Proceegings of National Conference on Artificial Intelligence, pages 356--361. AAAI Press.

[Godambe, 1991] Godambe, V. P., editor (1991). Estimating Functions. Oxford University Press.

[Graepel et al., 2004] Graepel, T., Herbrich, R., and Gold, J. (2004). Learning to fight. In Proceedings of the International Conference on Computer Games: Artificial Intelligence, Design and Education, pages 193--200.

Reference II

[Howard, 1960] Howard, R. A. (1960). Dynamic programming and markov processes..

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

[Sutton, 1984] Sutton, R. S. (1984). Temporal credit assignment in reinforcement learning. PhD thesis.

[Sutton and Barto, 1998] Sutton, R. S. and Barto, A. G. (1998). Reinforcement Learning: An Introduction. MIT Press.

[Sutton et al., 2009a] Sutton, R. S., Maei, H. R., Precup, D., Bhatnagar, S., Silver, D., Szepesvári, C., and Wiewiora, E. (2009a). Fast gradient-descent methods for temporal-difference learning with linear function approximation. In International Conference on Machine Learning, pages 993--1000.

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

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

Reference III

[Ueno et al., 2011] Ueno, T., Maeda, S., Kawanabe, M., and Ishii, S. (2011). Generalized TD learning. Journal of Machine Learning Research, 12:1977--2020.

[Yoshimoto et al., 2005] Yoshimoto, J., Nishimura, M., Tokita, Y., and Ishii, S. (2005). Acrobot control by learning the switching of multiple controllers. Artificial Life and Robotics, 9(2):67--71.