一般化TD学習

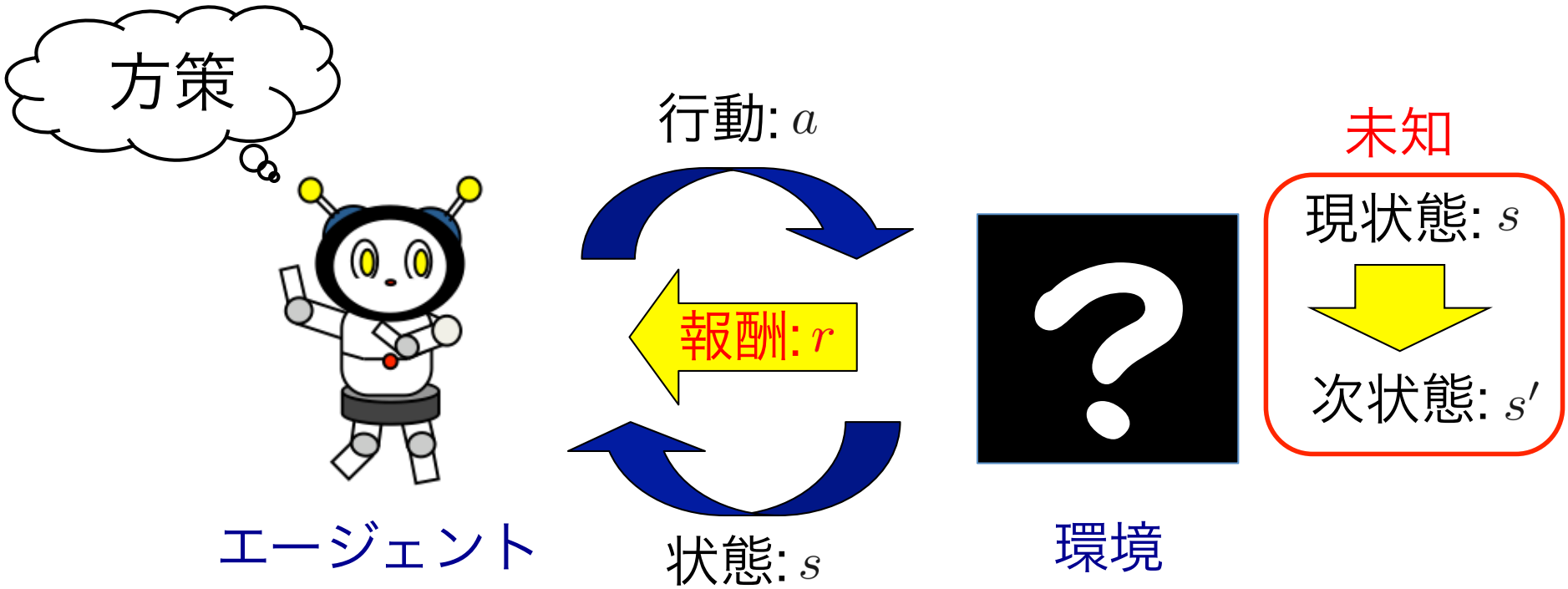
～ セミパラメトリック統計からのTD学習の一般化 ～

植野 剛[1], 前田 新一[1], 川鍋 一晃[2], 石井 信[1].

[1]京都大学 大学院情報学研究科

[2]Fraunhofer FIRST

強化学習

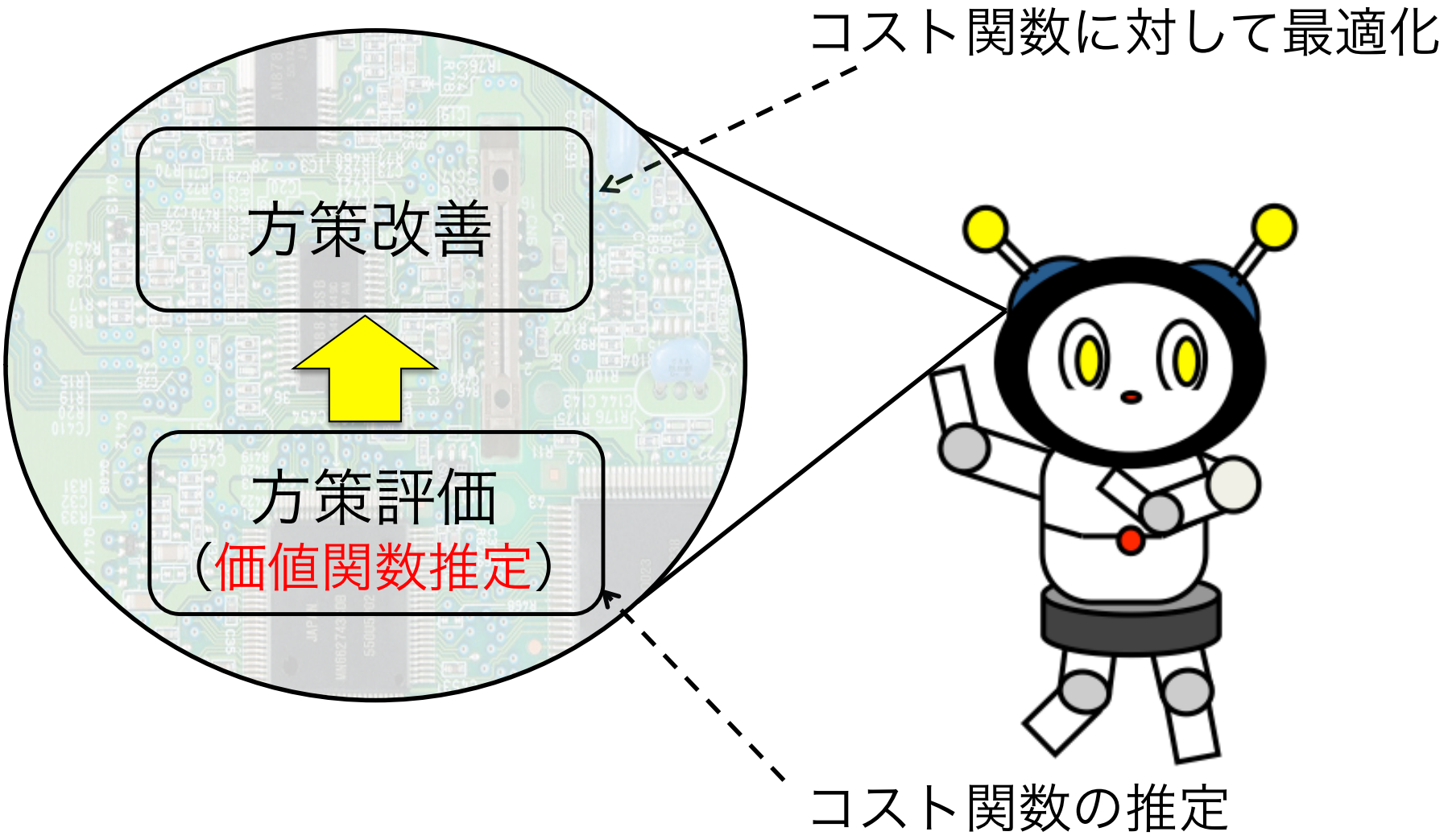


Goal

割引累積報酬和 R_t の期待値を最大とする方策を求める.

$$R_t = r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + \cdots$$

方策反復



方策評価は強化学習で重要な役割を果たす.

方策評価とベルマン方程式

価値関数:

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

$$= \mathbb{E}[r_{t+1}|s_t] + \lim_{T\to\infty} \mathbb{E}\left[\sum_{k=1}^{T} \gamma^k r_{t+k+1} \middle| s_0 = s_t\right]$$

ベルマン方程式:

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

*価値関数はベルマン方程式の根

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

関数近似: $V(s_t) \approx g(s_t, \boldsymbol{\theta})$

方策評価とTD学習

TD学習

- TD誤差(ベルマン残差)

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

  - 条件付き期待値: $\mathbb{E}\left[\epsilon(s_t, s_{t+1}, r_{t+1}, \boldsymbol{\theta}) | s_t\right] = 0$

- 確率勾配法

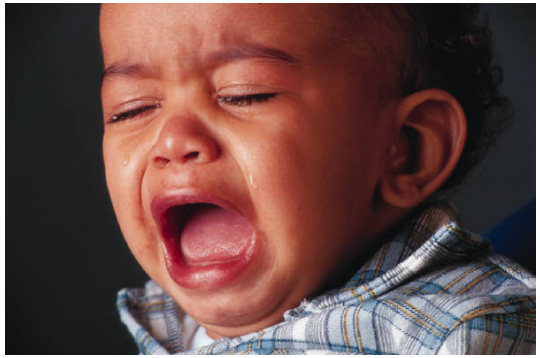
$$\boldsymbol{\theta}_{t+1} \leftarrow \boldsymbol{\theta}_t - \eta_t \partial_{\boldsymbol{\theta}} g(s_t, \boldsymbol{\theta}_t) \epsilon(s_t, s_{t+1}, r_{t+1}, \boldsymbol{\theta}_t)$$

TD学習の特徴

簡単・汎用的・遅い

TD学習の拡張

オンライン学習	バッチ学習
• TD, TD(λ) • NTD, NTD(λ) • LSPE, LSPE(λ) • RG • iLSTD, iLSTD(λ) • TDC • GTD, GTD2	• LSTD, LSTD(λ) • LSTDc • gLSTD



1. 各方策評価法の”収束速度”は？
2. 最速の方策評価法は？

本研究の目的

方策評価にセミパラメトリック統計と
推定関数の視点を導入する.

1. 各方策評価法の”収束速度”は？

TD法, その拡張手法を推定関数により
一般化する. 一般化した推定関数の漸近解析
を通して, 各手法の推定分散を評価する.

2. 最速の方策評価法は？

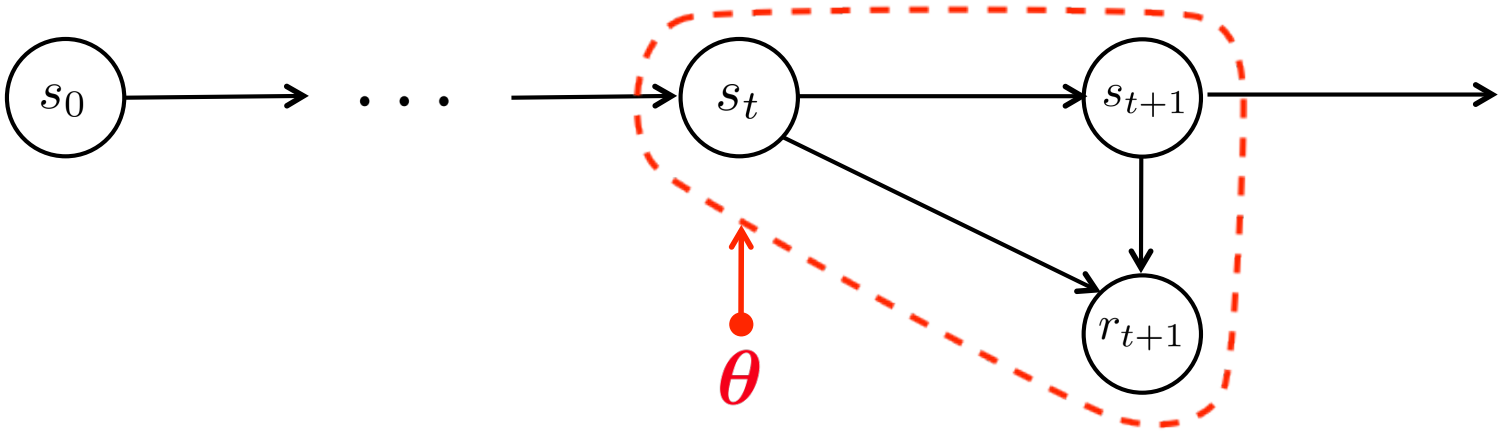
漸近推定分散を最小にする推定関数を導出する.

セミパラメトリック統計からみた方策評価

方策評価からセミパラ方策評価へ

マルコフ報酬過程:

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



- 確率分布 $p(r_{t+1}, s_{t+1}|s_t)$ の1次モーメント: $\mathbb{E}[r_{t+1}|s_t]$

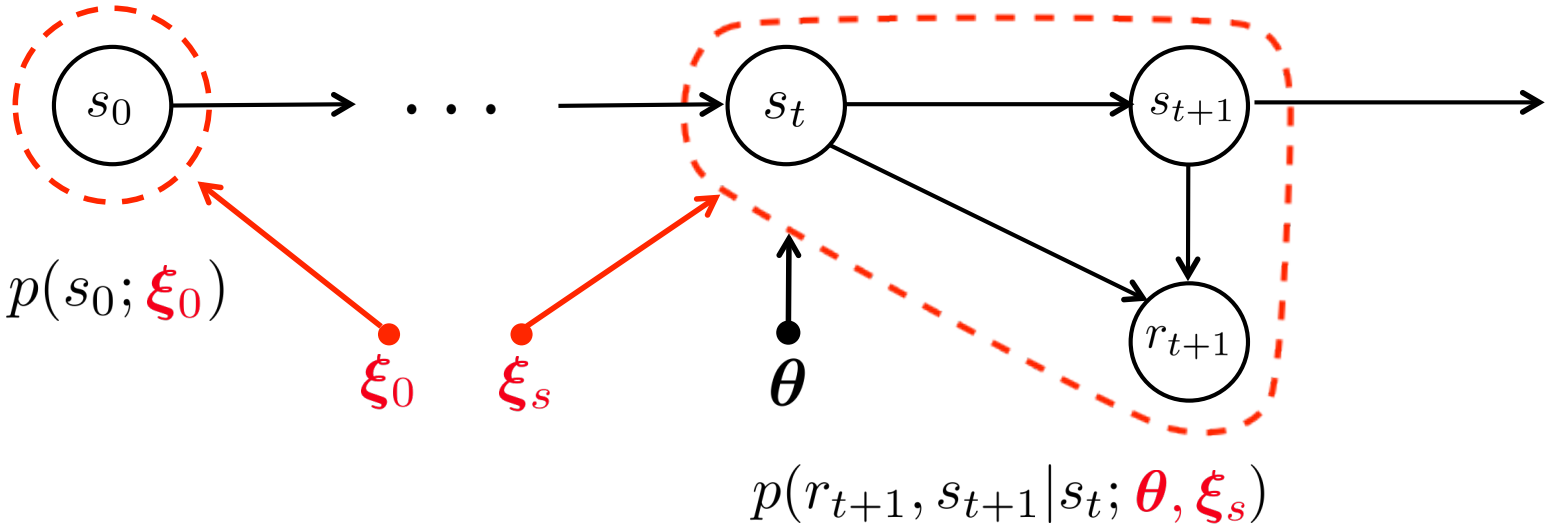
$\Longrightarrow$ ベルマン方程式により, $\boldsymbol{\theta}$ で特徴付ける.

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

方策評価からセミパラメ方策評価へ

マルコフ報酬過程:

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



- 確率分布 $p(s_0)$ と $p(r_{t+1}, s_{t+1}|s_t)$ の残りのモーメント

$\Longrightarrow$ 撹乱パラメータ $\boldsymbol{\xi}_0, \boldsymbol{\xi}_s$ によって特徴付ける.
(無限自由度をとることが可能なパラメータ)

方策評価からセミパラ方策評価へ

セミパラメトリックモデル（マルコフ報酬過程）

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

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

セミパラメトリック推定問題

$$\{s_{0:T}, r_{1:T}\} \sim p(s_{0:T}, r_{1:T}; \boldsymbol{\theta}, \boldsymbol{\xi}) \Longrightarrow \boldsymbol{\theta}$$（$\boldsymbol{\xi}$に依存せず）

セミパラメトリック推定

推定関数

次の条件を満たす $\boldsymbol{f}(s_{0:T}, r_{1:T}, \boldsymbol{\theta}) = \sum_{t=1}^{T} \boldsymbol{\psi}_t(s_{0:t}, r_{1:t}, \boldsymbol{\theta})$ を推定関数とよぶ.

$$\mathbb{E}\left[\boldsymbol{\psi}_t(s_{0:t}, r_{1:t}, \boldsymbol{\theta}) \middle| s_{t-1}\right] = \mathbf{0}, \quad \forall t, \boldsymbol{\theta}, \boldsymbol{\xi}$$

推定方程式

$$\sum_{t=1}^{T} \boldsymbol{\psi}_t(s_{0:t}, r_{1:t}, \hat{\boldsymbol{\theta}}) = 0 \qquad \{s_{0:T}, r_{1:T}\} \sim p(s_{0:T}, r_{1:T}; \boldsymbol{\theta}, \boldsymbol{\xi})$$

推定方程式の解 $\hat{\theta}$ は
撹乱パラメータに依存せず，真の解に収束する.

セミパラ方策評価における推定関数

定理 1.

すべての推定関数は以下の式で表すことができる.

$$\boldsymbol{f}(s_{0:T}, r_{1:T}, \boldsymbol{\theta}) = \sum_{t=0}^{T-1} \underbrace{\boldsymbol{w}_t(s_{0:t}, r_{1:T})}_{\text{時刻}t-1\text{までの観測にのみ依存する関数}} \underbrace{\epsilon(s_t, s_{t+1}, r_{t+1}, \boldsymbol{\theta})}_{\text{TD誤差}}$$

$\Longrightarrow$ あらゆる方策評価法は
関数$\boldsymbol{w}$の選択により説明できる.

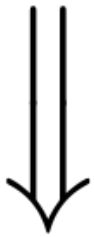
TD学習の拡張

オンライン学習	バッチ学習
• TD, TD(λ) • NTD, NTD(λ) • LSPE, LSPE(λ) • RG • iLSTD, iLSTD(λ) • TDC • GTD, GTD2	• LSTD, LSTD(λ) • LSTDc • gLSTD

$$
\begin{cases}
\boldsymbol{w}_t = \partial_\theta g(s_t, \boldsymbol{\theta}) \\
\boldsymbol{w}_t = \mathbb{E}[\partial_\theta \epsilon(s_t, s_{t+1}, r_{t+1}) | s_t] \\
\boldsymbol{w}_t = \sum_{t'=0}^{t} \lambda^{t'-t} \partial_\theta g(s_t, \boldsymbol{\theta}) \\
\boldsymbol{w}_t = \partial_\theta g(s_t, \boldsymbol{\theta}) + c \\
\boldsymbol{w}_t = \mathbb{E}[\epsilon(s_t, s_{t+1}, r_{t+1}; \theta)^2 | s_t] \mathbb{E}[\partial_\theta \epsilon(s_t, s_{t+1}, r_{t+1}) | s_t]
\end{cases}
$$

最適な推定関数

$$\boldsymbol{f}(s_{0:T}, r_{1:T}, \boldsymbol{\theta}) = \sum_{t=0}^{T-1} \underbrace{\boldsymbol{w}_t(s_{0:t}, r_{1:T})}\, \epsilon(s_t, s_{t+1}, r_{t+1}, \boldsymbol{\theta})$$



漸近推定分散 $\mathbb{E}\left[(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)^\top\right]$
を最小にするように $\boldsymbol{w}_t$ を選択

定理 2.

最小推定分散を導く推定関数は以下の式で与えられる.

$$\boldsymbol{f}_T^*(s_{0:T}, r_{1:T}, \boldsymbol{\theta}) = \sum_{t=1}^{T} \boldsymbol{w}_{t-1}^*(s_{t-1})\epsilon(s_{t-1}, s_t, r_t, \boldsymbol{\theta})$$

ここで

$$\boldsymbol{w}_{t-1}^*(s_{t-1}) := \mathbb{E}[\epsilon(s_{t-1}, s_t, r_t, \boldsymbol{\theta}^*)^2 | s_{t-1}]^{-1}\mathbb{E}[\partial_{\boldsymbol{\theta}}\epsilon(s_{t-1}, s_t, r_t, \boldsymbol{\theta}^*) | s_{t-1}]$$

まとめ & 今後の課題

まとめ

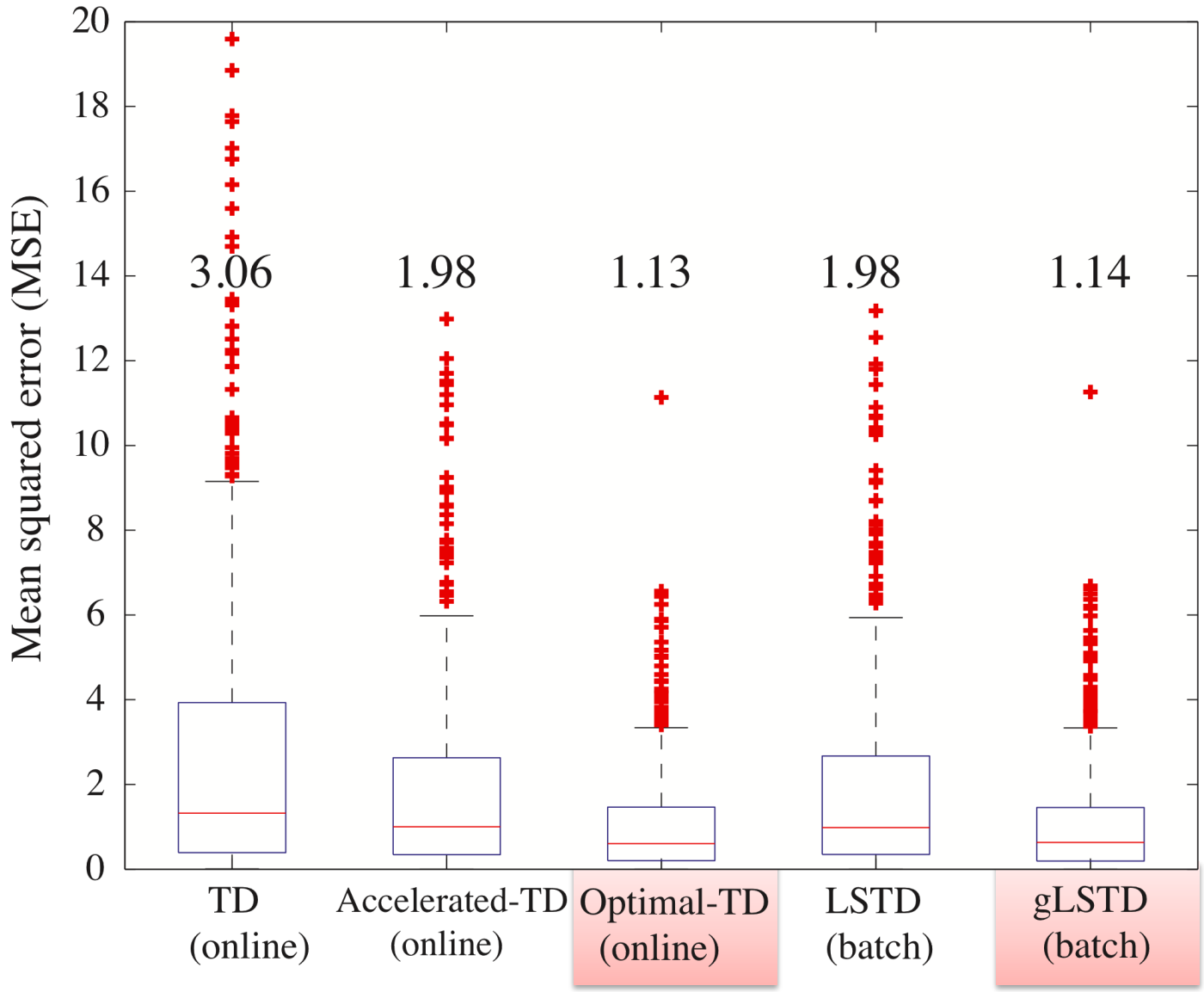
- 方策評価をセミパラメトリック推論問題として再定式化し，全ての方策評価法を一般化した.
- 一般化された推定関数の漸近解析を通して，方策評価法の推定精度を解析した
- 最小推定分散を導く推定関数を導出することで，最適な方策評価法を提案した.

今後の課題

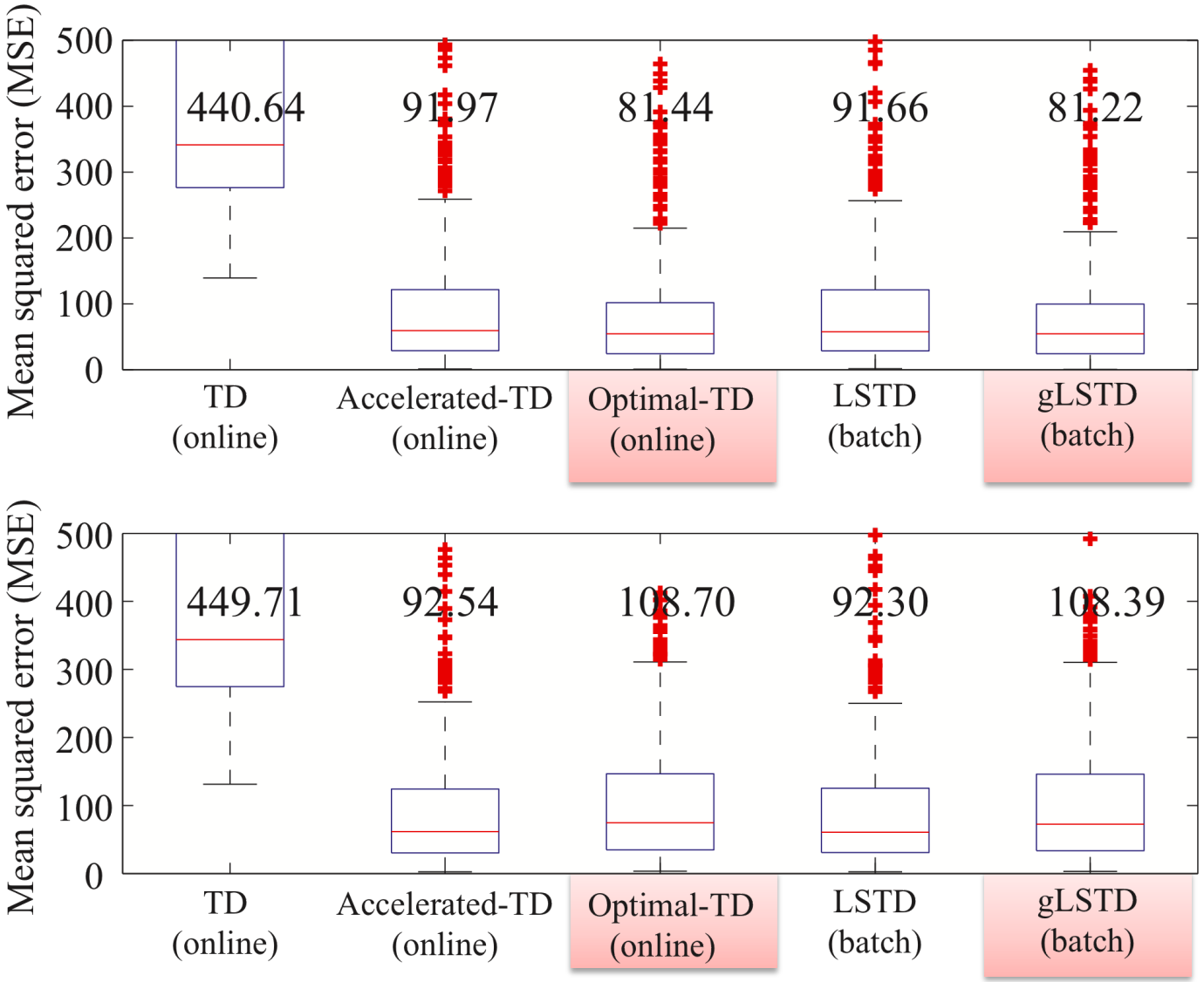
- 方策改善（マルコフ報酬過程→マルコフ決定過程）
- 関数近似誤差への対応

ご清聴ありがとうございました

トイプロブレム (5-states Markov chain)



(20-states Markov chain)



推定関数から学習アルゴリズムへ

THEOREM 4(Convergence Rate of Online Learning) *Consider the following learning process*

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

where $\hat{\mathbf{R}}_t = \{1/t \sum_{i=1}^{t} \partial_{\boldsymbol{\theta}} \boldsymbol{\psi}_i(Z_i, \hat{\boldsymbol{\theta}}_{i-1})\}$. *If the learning process converges to the true parameter almost surely, then the convergence rate is given as*

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

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