ERATO セミナ 2010 - No. 36
統計学習による強化学習の考察

植野剛
京都大学情報学研究科論理生命学講座

2011/2/24

概要

強化学習は心理学、神経科学，統計，コンピュータ科学，最適制御などさまざまな研究分野に起因する機械学習法の 1 つである．近年，強化学習は未知の環境に置かれた学習エージェントが自律的に行動方策を学習する方法として, 人工知能分野でスタンダードなツールとしてとなるまで発展しており，様々な実問題に応用され成果を挙げている．

現在，強化学習アルゴリズムにおいて中心的な役割を担っているのは TD 学習に代表されるモデルフリー方策評価法を組み込んだ強化学習法である．この手法は方策の “価値”(期待累積報酬和) の推定を行う，(モデルフリー) 方策評価と, 推定した価値に基づく方策の改善, 方策改善を交互に行うことで方策の学習を行う枠組みである．この学習法の最大の特長は方策評価時にタスク環境の推定せずに現在の方策の”価値” を推定する点である，つまりタスク環境のダイナミクスを知ることなく，方策の学習を行うことができる．この望ましい性質は，多くの研究者を魅了し，多くの新しいモデルフリー方策評価法とそれを組み込んだ強化学習法がこれまで提案されている．しかしながら, 提案されたモデルフリー方策評価アルゴリズムの理論解析，特に価値推定の推定精度に関してはほとんど検証されておらず，アルゴリズム間の推定精度による比較など理論的な考察は十分に行われていない．

本発表ではモデルフリー方策評価法に着目し，統計学習の観点から統計的な性質を考察する．本研究の主たるアイディアはモデルフリー方策評価問題をより一般的なセミパラメトリック統計モデルに基づく統計推定問題として再定式化することにある．これによりこれまで統計学習分野で培われてきた統計解析手法をそのまま応用することを可能にし，モデルフリー方策評価法全体に共通する統計的な性質を明らかにすることができる．

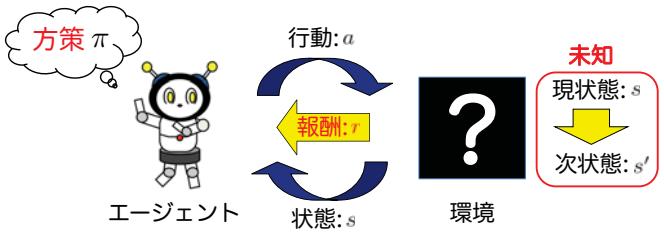
本発表は次の 3 部構成で発表する．まず第 1 部ではモデル方策評価に基づく強化学習法の簡単な導入を行い, その問題点を明らかにする．第 2 部では，我々の提案手法であるセミパラメトリック統計推論に基づく方策評価法を紹介し，その統計的な

性質を明らかにする．第３部では実応用として２足歩行ロボットの歩行学習について発表する．

統計学習による強化学習の考察

植野 剛
京都大学 情報学研究科

強化学習とは



目的

累積報酬和の期待値を最大にする行動方策を
試行錯誤を通じて自律的に獲得する

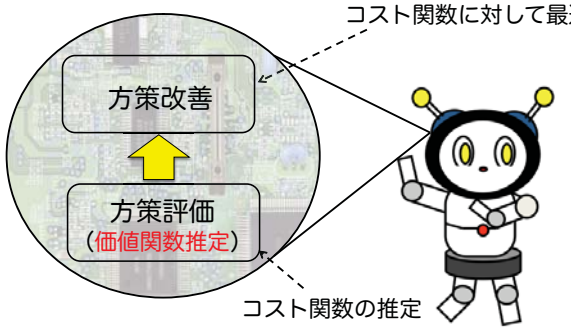
$$\pi^* = \arg\max_{\pi} \mathbb{E}\left[\sum_{t=1}^{\infty} \gamma^{t-1} r_t \middle| s_{t-1} = s, \pi\right], \forall s \in S$$

強化学習の応用

- ゲームエージェントの戦略
  （例）チェッカー, バックギャモン, ハーツ, 囲碁, 格闘ゲーム
- 最適制御, ロボット制御
  （例）ヘリコプター自動運転, 人型ロボットのモータコマンド
- オペレーションズリサーチ(OR)
  （例）エレベータ配置問題, ジョブショップ計画問題
- 神経科学・認知科学

専門家をも凌駕する性能を発揮する

方策反復 (Howard,1960)



方策評価は強化学習で重要な役割を果たす

方策評価

価値関数 (Bellman, 1957)

$$V(s_t) \equiv \lim_{T\to\infty} \mathbb{E}\left[\sum_{k=1}^{T} \gamma^{k-1} r_{t+k} \middle| s_t\right] = \mathbb{E}\left[r_{t+1}|s_t\right] + \lim_{T\to\infty} \mathbb{E}\left[\sum_{k=2}^{T} \gamma^{k-1} r_{t+k} \middle| s_t\right]$$

ベルマンの再帰方程式(ベルマン方程式)

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

- 状態遷移確率が既知
- 状態空間が離散, 低次元

$$\boldsymbol{v} = \boldsymbol{r} + \gamma \mathbf{P} \boldsymbol{v}$$

価値関数ベクトル $\boldsymbol{v}_i \equiv V(s(i))$　報酬ベクトル $\boldsymbol{r}_i \equiv \mathbb{E}[r_{t+1}|s_t = s(i)]$　状態遷移行列 $\mathbf{P}_{i,j} = p(s_{t+1} = s(j)|s_t = s(i))$

方策評価: TD学習

価値関数の関数近似: $V(s) \approx g(s, \boldsymbol{\theta})$　パラメータ: $\boldsymbol{\theta} \in \Theta, \Theta \subset \mathbb{R}^m$

$\Longrightarrow$ 近似されたベルマン方程式

$$\mathbb{E}[r_{t+1}|s_t] + \gamma \mathbb{E}[g(s_{t+1}, \boldsymbol{\theta})|s_t] - g(s_t, \boldsymbol{\theta}) = 0$$

TD学習 (Sutton,1988)

- TD誤差

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

  * 期待値を実現値で置き換え

- 学習方程式

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

- 環境の遷移ダイナミクスの推定を必要としない
- 特定の条件下で収束が保証されている

TD学習の拡張

TD学習の登場以降，多数の拡張アルゴリズムが提案されている

[オンライン法]
- TD (Sutton,1988)
- TD(λ) (Sutton,1988)
- LSPE (Nedic & Bertsekas, 2003)
- iLS
- RG
- TDC (Sutton et al., 2009)
- GTD (Sutton et al., 2009)
- GTD2 (Sutton et al., 2009) ,etc

[バッチ法]
- LSTD (Bratke, 1996)
- LSTD(λ) (Boyan, 2002) ,etc

- 環境の遷移ダイナミクスの推定を必要としない
- 特定の条件下で収束が保証されている

$\Longrightarrow$ 環境のモデルの特定を必要としない価値関数推定を
モデルフリー方策評価 と呼ぶ

未解決問題: モデルフリー方策評価

モデルフリー方策評価における価値推定の
推定量 $\hat{\theta}_T$(サンプル空間からパラメータ空間への写像)
の性質は明らかにされていない

本研究で注目する未解決問題

1. 一致推定量の設計
   - モデルが正しいとしたとき: $V(s) = g(s, \boldsymbol{\theta}),\ \forall s$
   - $\hat{\boldsymbol{\theta}}_T \to \boldsymbol{\theta},\ \text{as}\ T \to \infty$ が成り立つ $\hat{\boldsymbol{\theta}}_T$ は？
2. 異なる推定法間の推定精度比較
   - 推定量 $\hat{\boldsymbol{\theta}}'_T, \hat{\boldsymbol{\theta}}''_T$, 推定量の評価基準 $R(\hat{\boldsymbol{\theta}}_T)$ が与えられたとき
   - $R(\hat{\boldsymbol{\theta}}'_T) > R(\hat{\boldsymbol{\theta}}''_T)$ それとも $R(\hat{\boldsymbol{\theta}}'_T) < R(\hat{\boldsymbol{\theta}}''_T)$ もしくは・・・
3. モデル $g(s, \boldsymbol{\theta})$ の設計法
   - 複雑なモデル　→　オーバーフィット
   - 簡単なモデル　→　アンダーフィット

本研究のアプローチ

モデルフリー方策評価 $\Longrightarrow$ 統計推論問題
再定式化 (セミパラメトリック推論問題)

統計推論の問題として定式化することで，これまで
統計分野で研究されてきた漸近解析法を適用するを
可能にする

1. **セミパラメトリック統計推論問題としての方策評価**
2. **モデルフリー方策評価におけるリスク解析**
3. **モデルフリー方策評価におけるモデル選択**
4. **応用: 2足歩行ロボットの学習**

マルコフ報酬過程

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

初期確率 (未知)　状態遷移確率 (未知)　報酬確率 (未知)　$s_{0:T} = \{s_0, \cdots, s_T\}$, $r_{1:T} = \{r_1, \cdots, r_T\}$

状態: $s \in S$, Sは離散集合
報酬: $r \in \mathbb{R}$

・マルコフ報酬過程は唯一の定常分布 $\mu(s)$ を持つと仮定する

$$\sum_{s \in S} \mu(s)p(s'|s) = \mu(s'),\quad \forall s' \in S$$

仮定: 価値関数$V(s)$は既知の関数 $g(s, \boldsymbol{\theta})$で完全に表現できる

マルコフ報酬過程

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

初期確率 (未知)　状態遷移確率 (未知)　報酬確率 (未知)　$s_{0:T} = \{s_0, \cdots, s_T\}$, $r_{1:T} = \{r_1, \cdots, r_T\}$

状態: $s \in S$, Sは離散集合
報酬: $r \in \mathbb{R}$

・マルコフ報酬過程は唯一の定常分布 $\mu(s)$ を持つと仮定する

$$\sum_{s \in S} \mu(s)p(s'|s) = \mu(s'),\quad \forall s' \in S$$

仮定: 価値関数$V(s)$は既知の関数 $g(s, \boldsymbol{\theta})$で完全に表現できる

$\Longrightarrow$ 近似誤差は考慮しない

研究目的

モデルフリー方策評価

サンプル系列 $\{s_{0:T}, r_{1:T}\}$ $\Longrightarrow$（推定）パラメータ $\boldsymbol{\theta}$　ここで $V(s) = g(s, \boldsymbol{\theta}),\ \forall s$

ただし$p(s_{0:T}, r_{1:T})$は推定しない

望ましい推定量の性質

・サンプル数無限大の極限で $\hat{\boldsymbol{\theta}}_T \to \boldsymbol{\theta}$ が成り立つ (一致性)

・パラメータの推定誤差 $|\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}|$ が小さい

Question

・一致推定量となる関数は？

・最小のパラメータ推定誤差を実現する推定量は？

セミパラメトリックモデル

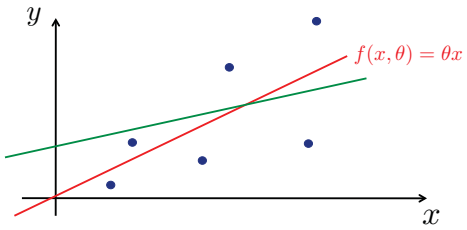
$$p(\boldsymbol{x}; \boldsymbol{\theta}, \boldsymbol{\xi})$$

知りたいパラメータ (推定パラメータ)　興味のないパラメータ (撹乱パラメータ)

セミパラメトリックモデル

Error in variable problem

$y_t = f(x_t, \theta) + \epsilon_t$　ただし x_t と ϵ_t に相関がある



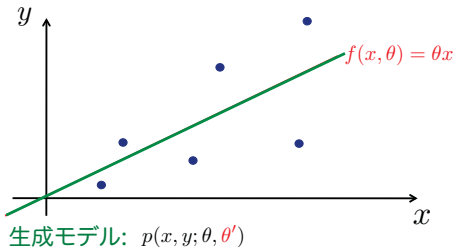
1. 最小二乗法で推定すると・・・

$\Longrightarrow$ バイアスする

セミパラメトリックモデル

Error in variable problem

$y_t = f(x_t, \theta) + \epsilon_t$　ただし x_t と ϵ_t に相関がある



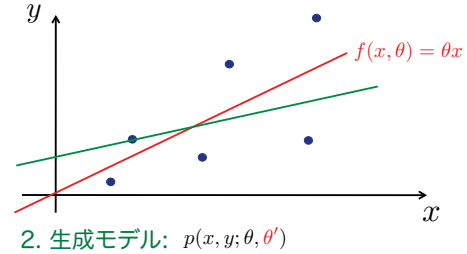
2. 生成モデル: $p(x, y; \theta, \theta')$

$\Longrightarrow$ 仮定したパラメトリックモデル $p(x, y; \theta, \theta')$ が正しければ正しい解に収束

セミパラメトリックモデル

Error in variable problem

$y_t = f(x_t, \theta) + \epsilon_t$　ただし x_t と ϵ_t に相関がある



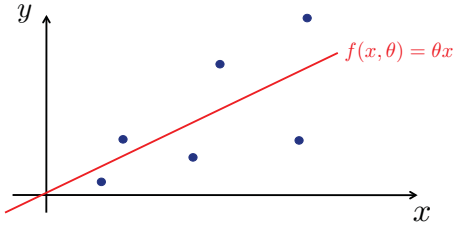
2. 生成モデル: $p(x, y; \theta, \theta')$

$\Longrightarrow$ 仮定したパラメトリックモデル $p(x, y; \theta, \theta')$ が間違っていればバイアスする

セミパラメトリックモデル

Error in variable problem

$y_t = f(x_t, \theta) + \epsilon_t$　ただし x_t と ϵ_t に相関がある



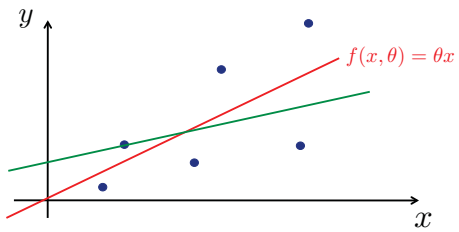
知りたいパラメータは θ

x_t と ϵ_t の相関関係は興味はない

セミパラメトリックモデル

Error in variable problem

$y_t = f(x_t, \theta) + \epsilon_t$　ただし x_t と ϵ_t に相関がある



3. セミパラメトリックモデル: $p(x, y; \theta, \xi)$

$\Longrightarrow$ 撹乱パラメータを直接推定せず，パラメータのみ推定する

セミパラメトリックモデルの推定法

$$\{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n\} \overset{\text{i.i.d.}}{\sim} p(\boldsymbol{x}; \boldsymbol{\theta}, \boldsymbol{\xi})$$

撹乱パラメータ $\boldsymbol{\xi}$ を推定せずに $\boldsymbol{\theta}$ のみを推定するには？

推定関数 (Godambe, 1991)

関数 $\boldsymbol{f}(\boldsymbol{x}, \boldsymbol{\theta})$ が任意の $\boldsymbol{\theta}, \xi$ に関して次の条件を満たすとする

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

このとき，推定方程式を解くことで
一致推定量(M-推定量)を得ることができる

$$\sum_{i=1}^{n} \boldsymbol{f}(x_i, \hat{\boldsymbol{\theta}}_n) = \boldsymbol{0}$$

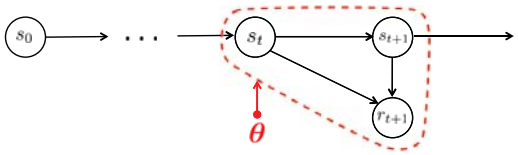
推定関数を考える利点

- 推定関数と成り得る関数の全体のクラスを特定することで，M-推定量となり得る推定量の全体のクラスを特定できる.
- 特定した推定関数に関する漸近解析を通して，全てのM-推定量のパラメータの推定誤差を計算することができる
- パラメータの推定誤差を最小とする推定関数を選択することで，漸近的に最も早く真のパラメータに収束する推定量を導出することができる.

方策評価の統計モデル

グラフィカルモデル: マルコフ報酬過程

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

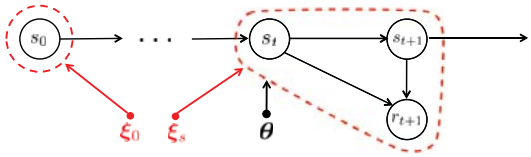


確率分布 $p(s_t, r_t|s_{t-1})$ の1次モーメント $\mathbb{E}[r_t|s_{t-1}]$ を
ベルマン方程式: $\mathbb{E}[r_t|s_{t-1}] = g(s_{t-1}, \boldsymbol{\theta}) - \gamma \mathbb{E}[g(s_t, \boldsymbol{\theta})|s_{t-1}]$
を用いて，パラメータ $\boldsymbol{\theta}$ で特徴付ける

方策評価の統計モデル

グラフィカルモデル: マルコフ報酬過程

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



確率分布 $p(s_0)$, $p(s_t, r_t|s_{t-1})$ の残りのモーメントを
撹乱パラメータ ξ_0, ξ_s で特徴付ける

方策評価の統計モデル

セミパラメトリック統計モデル

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

$$\text{s.t. } \mathbb{E}_{\boldsymbol{\theta}, \boldsymbol{\xi}_s}[r_t|s_{t-1}] = g(s_{t-1}, \boldsymbol{\theta}) - \gamma \mathbb{E}_{\boldsymbol{\theta}, \boldsymbol{\xi}_s}[g(s_t, \boldsymbol{\theta})|s_{t-1}]$$

セミパラメトリック推定問題

サンプル系列 $\{s_{0:T}, r_{1:T}\} \sim p(s_{0:T}, r_{1:T}; \boldsymbol{\theta}, \boldsymbol{\xi})$

$\Longrightarrow$ 推定　パラメータ $\boldsymbol{\theta}$　撹乱パラメータ $\boldsymbol{\xi}$

ただし，$\boldsymbol{\xi}$ を推定しない

時系列における推定関数

マーチンゲール推定関数 (Godambe, 1991)

$$\boldsymbol{f}_T(s_{0:T}, r_{1:T}, \boldsymbol{\theta}) = \sum_{t=1}^{T} \boldsymbol{\psi}_t(s_{0:t}, r_{0:t}, \boldsymbol{\theta})$$

関数 $\boldsymbol{f}_T(s_{0:T}, r_{1:T}, \boldsymbol{\theta})$ が任意の $\boldsymbol{\theta}, \boldsymbol{\xi}$ に関して
次の条件を満たすとする

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

このとき, 推定方程式を解くことで
一致推定量(M-推定量)を得ることができる

$$\sum_{t=1}^{T} \boldsymbol{\psi}_t(s_{0:t}, r_{0:t}, \hat{\boldsymbol{\theta}}_T) = \mathbf{0}$$

解析結果

定理 1.

マルコフ報酬過程における価値関数推定について
マーチンゲール推定関数と成り得る関数は以下の形に限定される

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

重み関数 × TD誤差

過去の系列 $\{s_{0:t-1}, r_{1:t-1}\}$ とパラメータ $\boldsymbol{\theta}$ の関数

$\Longrightarrow$ 重み関数を変更することで
従来提案されている推定量を全てを導くことができる

解析結果

補題 2.

$\{s_{0:T}, r_{1:T}\} \sim p(s_{0:T}, r_{1:T}; \boldsymbol{\theta}^*, \boldsymbol{\xi}^*)$ とする．このとき推定方程式

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

の解のパラメータ推定誤差は次の関係を満たす

$$\left|\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}^*\right| = O\left(\frac{\sqrt{\Omega_w}}{\sqrt{T}}\right)$$

$\sqrt{\Omega_w}$：推定精度行列
$\sqrt{T}$：サンプル数

- 全て $O(T^{-1/2})$ のオーダーで真の解 $\boldsymbol{\theta}^*$ に収束する.
- 各推定量のパラメータの推定誤差は行列 $\boldsymbol{\Omega}_w$ に依存して決まる

$\Longrightarrow$ 行列を最小にする重み関数 $\boldsymbol{w}_{t-1}$ を導出する

解析結果

定理 2

パラメータの推定誤差を最小にする推定関数は

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

ここで

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

推定方程式

$$\sum_{t=1}^{T} \boldsymbol{w}_t^*(s_{t-1}, \boldsymbol{\theta}^*) \epsilon(s_{t-1}, s_t, r_t, \hat{\boldsymbol{\theta}}_T) = \mathbf{0}$$

上記の推定方程式を解くアルゴリズム, gLSTD, gTDを提案
(本発表では省略)

この節のまとめ

- セミパラメトリック統計推論の枠組みを利用して，モデルフリー方策評価法を統計問題として定式化した.
- 推定関数理論を元に，全ての推定関数のクラスを特定した. この推定関数はこれまでに提案されてきた方策評価法を一般化する
- 特定した一般的な推定関数に漸近解析を施すことで，全ての推定量のパラメータ推定誤差を求め，パラメータ推定誤差を最小にする推定関数を導出した.
- パラメータ推定誤差を最小とするバッチアルゴリズム，オンラインアルゴリズムを提案した.

1. セミパラメトリック統計推論問題としての方策評価
2. モデルフリー方策評価におけるリスク解析
3. モデルフリー方策評価におけるモデル選択
4. 応用: 2足歩行ロボットの学習

導入

~~仮定: 価値関数 $V(s)$ は既知の関数 $g(s,\boldsymbol{\theta})$ で完全に表現できる~~

~~望ましい推定量の性質~~

~~・サンプル数無限大の極限で $\hat{\boldsymbol{\theta}}_T \to \boldsymbol{\theta}$ が成り立つ (一致性)~~
~~・パラメータの推定誤差 $|\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}|$ が小さい~~

$\Longrightarrow$ モデルが正しくない場合,
モデルによる近似誤差を考慮する必要がある

真の価値関数 $V(s)$ と価値関数の推定値 $g(s,\hat{\boldsymbol{\theta}}_T)$ との差
で推定量を評価する必要がある.

$$L(\hat{\boldsymbol{\theta}}_T) = \|V(s) - g(s,\hat{\boldsymbol{\theta}}_T)\|_D^2$$

目的

平均二乗誤差(MSE)

$$L(\hat{\boldsymbol{\theta}}_T) \equiv \mathbb{E}_\mu\left[|V(s) - g(s,\hat{\boldsymbol{\theta}}_T)|^2\right] \quad \mathbb{E}_\mu[\cdot]:\text{定常分布による期待値}$$

$$\Longrightarrow L(\hat{\boldsymbol{\theta}}_T) = L(\hat{\boldsymbol{\theta}}_T) - L(\bar{\boldsymbol{\theta}}) + L(\bar{\boldsymbol{\theta}})$$

推定誤差: パラメータの推定によって生じる誤差(サンプル依存)
近似誤差: モデルと推定量の選択によって決定する誤差(サンプル非依存)

研究目的

平均二乗誤差基準で推定量 $\hat{\boldsymbol{\theta}}_T$ の比較を1)モデルが正しい場合，2)間違っている場合それぞれで行う

$\Longrightarrow$ 推定量と推定誤差，近似誤差の関係を個別に解析する

仮定

1. モデル $g(s,\boldsymbol{\theta})$ は線形モデルに限定する: $g(s,\boldsymbol{\theta}) = \boldsymbol{\phi}(s)^\top \boldsymbol{\theta}$
2. 報酬は決定論の既知の関数とする: $r : S \times S \mapsto \mathbb{R}$

考慮する推定関数のクラス

$$\tilde{\boldsymbol{f}}_T(s_{0:T},\boldsymbol{\theta}) \equiv \sum_{t=1}^{T}\underbrace{\sum_{t'=1}^{t}\beta^{t-t'}\tilde{\boldsymbol{w}}(s_{t'-t})}_{\text{重み関数}}\underbrace{\epsilon(s_{t-1},s_t,\boldsymbol{\theta})}_{\text{TD誤差}} \quad \beta = [0,1)$$

上記の推定関数は，前節で示した推定関数のサブクラスで
あるがこれまでに提案されている全ての推定量を含む.

解析結果

$$L(\hat{\boldsymbol{\theta}}_T) = L(\hat{\boldsymbol{\theta}}_T) - L(\bar{\boldsymbol{\theta}}) + L(\bar{\boldsymbol{\theta}})$$

推定誤差

定理 6. (Liang and Jordan, 2008)

推定誤差は一般に

$$\left|L(\hat{\boldsymbol{\theta}}_T) - L(\bar{\boldsymbol{\theta}})\right| = O\left(\frac{\sqrt{\Omega'}}{\sqrt{T}}\right)$$

ただし

$\partial_{\boldsymbol{\theta}} L(\boldsymbol{\theta})|_{\boldsymbol{\theta}=\bar{\boldsymbol{\theta}}} = \mathbf{0}$ を満たすとき,

$$\left|L(\hat{\boldsymbol{\theta}}_T) - L(\bar{\boldsymbol{\theta}})\right| = O\left(\frac{\sqrt{\Omega'}}{T}\right)$$

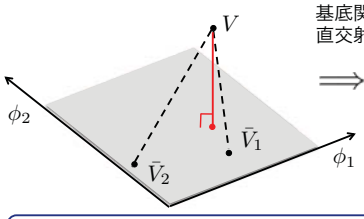
解析結果

$$L(\hat{\boldsymbol{\theta}}_T) = L(\hat{\boldsymbol{\theta}}_T) - L(\bar{\boldsymbol{\theta}}) + L(\bar{\boldsymbol{\theta}})$$

近似誤差

基底関数が張る空間への
直交射影は最小近似誤差を導く

$\Longrightarrow \partial_{\boldsymbol{\theta}} L(\boldsymbol{\theta})|_{\boldsymbol{\theta}=\bar{\boldsymbol{\theta}}} = \mathbf{0}$



$\partial_{\boldsymbol{\theta}} L(\boldsymbol{\theta})|_{\boldsymbol{\theta}=\bar{\boldsymbol{\theta}}} = \mathbf{0}$ を満たすとき，最小の近似誤差を実現する

推定量間の比較

次の代表的な３つの推定量に着目する

- TD推定量: $\hat{\boldsymbol{\theta}}_T^{\mathrm{TD}}$
  - 最も幅広く応用されている推定量
  - 実装法; TD法，LSTD法
- TD(λ)推定量: $\hat{\boldsymbol{\theta}}_T^{\mathrm{TD}(\lambda)}$
  - TD推定量の自然な拡張
  - 実装法; TD(λ)法，LSTD(λ)法
- gTD推定量: $\hat{\boldsymbol{\theta}}_T^{\mathrm{gTD}}$
  - 最小のパラメータ推定誤差を実現する
  - 実装法; gLSTD法, gTD法

推定量間の比較

1) モデルが正しい場合

全ての一致推定量で$\partial_{\boldsymbol{\theta}} L(\boldsymbol{\theta})|_{\boldsymbol{\theta}=\bar{\boldsymbol{\theta}}} = \mathbf{0}$が成り立つ

$\Longrightarrow$ 全ての一致推定量で推定誤差は$O(T^{-1})$で収束する
推定誤差の収束速度は$|\boldsymbol{\Omega}'|$で決定される

$$|\boldsymbol{\Omega}_{\mathrm{gTD}}| \preceq |\boldsymbol{\Omega}_{\mathrm{TD}}|, |\boldsymbol{\Omega}_{\mathrm{TD}(\lambda)}|$$

2) モデルが間違っている場合

$\hat{\boldsymbol{\theta}}_T^{\mathrm{TD}(\lambda=1)}$のみ $\partial_{\boldsymbol{\theta}} L(\boldsymbol{\theta})|_{\boldsymbol{\theta}=\bar{\boldsymbol{\theta}}} = \mathbf{0}$ が成り立つ

$\Longrightarrow$ $\hat{\boldsymbol{\theta}}_T^{\mathrm{TD}(\lambda=1)}$は最小の近似誤差を実現し，
また $\hat{\boldsymbol{\theta}}_T^{\mathrm{TD}(\lambda=1)}$ のみ $O(T^{-1})$ 収束する

推定量間の比較

1) モデルが正しい場合

全ての一致推定量で$\partial_{\boldsymbol{\theta}} L(\boldsymbol{\theta})|_{\boldsymbol{\theta}=\bar{\boldsymbol{\theta}}} = \mathbf{0}$が成り立つ

$\Longrightarrow$ 全ての一致推定量で推定誤差は$O(T^{-1})$で収束する
推定誤差の収束速度は$|\boldsymbol{\Omega}'|$で決定される

$$|\boldsymbol{\Omega}_{\mathrm{gTD}}| \preceq |\boldsymbol{\Omega}_{\mathrm{TD}}|, |\boldsymbol{\Omega}_{\mathrm{TD}(\lambda)}|$$

モデルが正しいとき

$$L(\hat{\boldsymbol{\theta}}_T^{\mathrm{gTD}}) \leq L(\hat{\boldsymbol{\theta}}_T^{\mathrm{TD}}), L(\hat{\boldsymbol{\theta}}_T^{\mathrm{TD}(\lambda)})$$

モデルが間違っているとき

$$L(\hat{\boldsymbol{\theta}}_T^{\mathrm{TD}(\lambda=1)}) \leq L(\hat{\boldsymbol{\theta}}_T^{\mathrm{TD}}), L(\hat{\boldsymbol{\theta}}_T^{\mathrm{gTD}})$$

$\Longrightarrow$ $\hat{\boldsymbol{\theta}}_T^{\mathrm{TD}(\lambda=1)}$は最小の近似誤差を実現し，
また $\hat{\boldsymbol{\theta}}_T^{\mathrm{TD}(\lambda=1)}$ のみ $O(T^{-1})$ 収束する

計算機実験

マルコフ過程

- 状態遷移確率行列: $P = \begin{pmatrix} 0.911 & 0.008 & 0.005 & 0.065 & 0.011 \\ 0.070 & 0.242 & 0.208 & 0.340 & 0.140 \\ 0.247 & 0.257 & 0.292 & 0.175 & 0.029 \\ 0.116 & 0.444 & 0.044 & 0.093 & 0.303 \\ 0.017 & 0.348 & 0.323 & 0.248 & 0.064 \end{pmatrix}$
- 基底関数: $\phi(s) = \left(1, s, s^2\right)^\top$

報酬関数を変更することでモデルが正しい状況
モデルが間違っている状況をそれぞれ作り出す

(a) $r^{(a)} = (-0.36, -0.98, -0.79, 0.63, -0.18)$

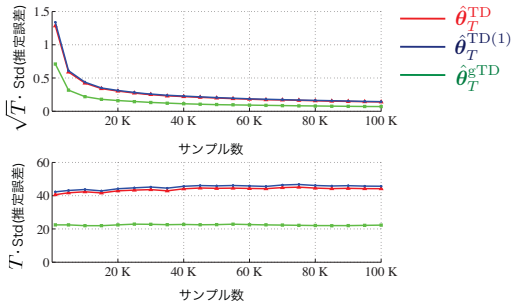
⇨ モデル $g(s, \theta)$ は真の価値関数を表現可能

(b) $r^{(b)} = (1.0, 0, 0, 0, 0)$

⇨ モデル $g(s, \theta)$ は真の価値関数を表現不可能

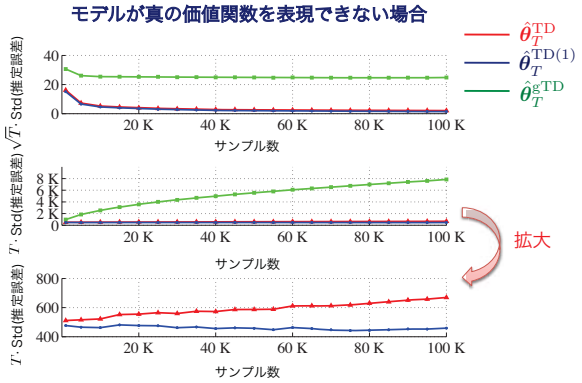
計算機実験

モデルが真の価値関数を表現できる場合



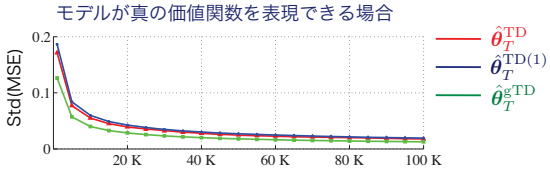
計算機実験

モデルが真の価値関数を表現できない場合

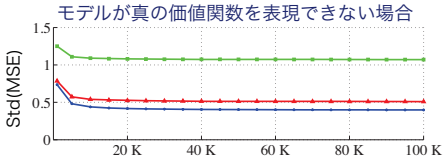


計算機実験

モデルが真の価値関数を表現できる場合



モデルが真の価値関数を表現できない場合



議論

関連研究

- MSEの有限サンプル解析に関する研究
  - 多くの研究がモデルとしてlook-up table を採用しており，近似誤差に関する考察は行われていない
  - 解析結果は基本的にバウンドである.
- MSEの漸近解析に関する研究 (Tsitsiklis and Roy,1997)
  - 近似誤差に関する議論しか行われておらず，推定誤差に関する考察はされていない

この節のまとめ

まとめ

- セミパラメトリックの枠組みを拡張し，平均二乗誤差解析を可能にした
- 代表的な3つの推定量を $\hat{\boldsymbol{\theta}}_T^{\mathrm{TD}}, \hat{\boldsymbol{\theta}}_T^{\mathrm{TD}(\lambda)}, \hat{\boldsymbol{\theta}}_T^{\mathrm{gTD}}$ 平均二乗誤差を基準に比較した．モデルが正しい場合と，正しくない場合で推定量の挙動が異なることを理論的に示した.
- 簡単な計算機実験を通して解析が正しいことを確認した.

今後の課題

- 有限サンプル下での挙動の解析
- モデル選択

導入

モデル選択の難しさ

- **単純なモデルを用いた場合**
  $\Longrightarrow$ 大きい近似誤差，小さい推定誤差 (underfitting)
- **複雑なモデルを用いた場合**
  $\Longrightarrow$ 小さい近似誤差，大きい推定誤差 (overfitting)

適切な複雑度を有するモデルを選択するためには…

推定誤差と近似誤差のトレードオフを制御する必要がある

アプローチ

統計学習分野で知られる情報量基準の考えを応用して，モデルの良さをサンプルから推定する

モデル選択の基本的な手順

評価基準: 平均二乗誤差

$$L(\hat{\boldsymbol{\theta}}_T) \equiv \mathbb{E}_\mu \left[|V(s) - g(s, \hat{\boldsymbol{\theta}}_T)|^2 \right]$$

基本的な手順

1. 事前に候補となるモデル集合 $\mathcal{G}$ を準備する
2. モデルを 1 つ選択して，推定量 $\hat{\boldsymbol{\theta}}_T$ を計算する
3. $L(\hat{\boldsymbol{\theta}}_T)$**を評価する** ← サンプルから$L(\hat{\boldsymbol{\theta}}_T)$の推定量(情報量基準)を計算する
4. 2-3 を繰り返す
5. $L(\hat{\boldsymbol{\theta}}_T)$ を最小とするモデルを採用する

もし $L(\hat{\boldsymbol{\theta}}_T)$ の評価が可能なら用意したモデル集合から適切なモデルを選択できるが・・・

$L(\hat{\boldsymbol{\theta}}_T)$ の推定

$$L(\hat{\boldsymbol{\theta}}_T) = \mathbb{E}_\mu \left[|V(s) - g(s, \hat{\boldsymbol{\theta}}_T)|^2 \right]$$

サンプル平均で置き換え　$V(s_{t-1}) \leftarrow \mathcal{R}_{t-1}(s_{t-1:T}) \equiv \sum_{t'=t}^{T} \gamma^{t'-t} r_{t'}$

単純な推定法

$$\zeta_T(s_{0:T}, \hat{\boldsymbol{\theta}}_T) \equiv \frac{1}{T} \sum_{t=1}^{T} |\mathcal{R}_{t-1}(s_{t-1:T}) - g(s_{t-1}, \hat{\boldsymbol{\theta}}_T)|^2$$

- 推定量の計算で用いたサンプルを期待値の近似に再度使用している．このような計算は推定にバイアスを生じさせてしまう.
- このバイアスの大きさは，パラメータの次元数によって異なるため，$\zeta_T(s_{0:T}, \hat{\boldsymbol{\theta}}_T)$によるモデル評価は公平なモデル選択を導くことができない

$L(\hat{\boldsymbol{\theta}}_T)$ の推定

$$L(\hat{\boldsymbol{\theta}}_T) = \mathbb{E}_\mu \left[|V(s) - g(s, \hat{\boldsymbol{\theta}}_T)|^2 \right]$$

バイアスを補正した推定法

$$\zeta_T(s_{0:T}, \hat{\boldsymbol{\theta}}_T) - \underbrace{\left(\zeta_T(s_{0:T}, \hat{\boldsymbol{\theta}}_T) - L(\hat{\boldsymbol{\theta}}_T) \right)}_{\text{バイアス}}$$

- このバイアス項を評価できれば，バイアスを補正することで公平なモデル選択を行うことができる

$\Longrightarrow$ バイアスの期待値を漸近解析を通して求める

仮定

1. モデル $g(s, \boldsymbol{\theta})$ は線形モデルに限定する: $g(s, \boldsymbol{\theta}) = \boldsymbol{\phi}(s)^\top \boldsymbol{\theta}$
2. 報酬は決定論の既知の関数とする: $r : S \times S \mapsto \mathbb{R}$

考慮する推定関数のクラス

$$\tilde{\boldsymbol{f}}_T(s_{0:T}, \boldsymbol{\theta}) \equiv \sum_{t=1}^{T} \underbrace{\sum_{t'=1}^{t} \beta^{t-t'} \tilde{\boldsymbol{w}}(s_{t'-t})}_{\text{重み関数}} \underbrace{\epsilon(s_{t-1}, s_t, \boldsymbol{\theta})}_{\text{TD誤差}} \qquad \beta = [0, 1)$$

上記の推定関数は，前節で示した推定関数のサブクラスであるがこれまでに提案されている全ての推定量を含む．

解析結果

定理 7

$$\mathbb{E}\left[\zeta_T(s_{0:T}, \hat{\boldsymbol{\theta}}_T)\right] = \mathbb{E}\left[L(\hat{\boldsymbol{\theta}}_T)\right] + b_1 + \frac{1}{T} b_2(\bar{\boldsymbol{\theta}}) + o\left(\frac{1}{T}\right)$$

$$b_1 = \text{constant}$$

$$b_2(\bar{\boldsymbol{\theta}}) \equiv -2\text{tr}\{\mathbf{A}^{-1}\mathbf{G}(\bar{\boldsymbol{\theta}})\}$$

ここで $\mathbf{A} = \mathbb{E}\left[\sum_{t'=1}^{\infty} \beta^{t'-1} \tilde{\boldsymbol{w}}(s_0)(\gamma \boldsymbol{\phi}(s_{t'}) - \boldsymbol{\phi}(s_{t'-1}))\right]$

$$\mathbf{G}(\bar{\boldsymbol{\theta}}) \equiv \mathbb{E}\left[\bar{\psi}(s_{0:\infty}, \bar{\boldsymbol{\theta}}) \bar{\boldsymbol{h}}(s_{0:\infty}, \bar{\boldsymbol{\theta}})^\top\right] + \sum_{t=1}^{\infty} \mathbb{E}\left[\bar{\psi}(s_{t:\infty}, \bar{\boldsymbol{\theta}}) \bar{\boldsymbol{h}}(s_{0:\infty}, \bar{\boldsymbol{\theta}})^\top\right] + \sum_{t=1}^{\infty} \mathbb{E}\left[\bar{\psi}(s_{0:\infty}, \bar{\boldsymbol{\theta}}) \bar{\boldsymbol{h}}(s_{t:\infty}, \bar{\boldsymbol{\theta}})^\top\right]$$

解析結果

$$b_2(\bar{\boldsymbol{\theta}}) = -2\text{tr}\{\mathbf{A}^{-1}\mathbf{G}(\bar{\boldsymbol{\theta}})\}$$

期待値はサンプルにより置き換える　推定量 $\hat{\boldsymbol{\theta}}_T$ による置き換え

バイアスの推定量

$$\hat{b}_2 = -2\text{tr}\{\hat{\mathbf{A}}_T^{-1}\hat{\mathbf{G}}_T(\hat{\boldsymbol{\theta}}_T)\}$$

$$\hat{\mathbf{A}}_T \equiv \frac{1}{T}\sum_{t=1}^{T}\sum_{t'=1}^{t} \beta^{t-t'} \tilde{\boldsymbol{w}}(s_{t-t'}) \left(\gamma \boldsymbol{\phi}(s_t) - \boldsymbol{\phi}(s_{t-1})\right)^\top$$

$$\hat{\mathbf{G}}_T(\hat{\boldsymbol{\theta}}_T) \equiv \frac{1}{T}\sum_{t=1}^{T} \tilde{\psi}_t(s_{0:t}, \hat{\boldsymbol{\theta}}_T) \boldsymbol{h}_t(s_{0:t}, \hat{\boldsymbol{\theta}}_T)^\top + \sum_{t=1}^{T-1}\left(\frac{1}{T-t}\sum_{t'=1}^{T-t} \tilde{\psi}_t(s_{0:t}, \hat{\boldsymbol{\theta}}_T) \boldsymbol{h}_{t+t'}(s_{t':t+t'}, \hat{\boldsymbol{\theta}}_T)^\top\right) + \sum_{t=1}^{T-1}\left(\frac{1}{T-t}\sum_{t'=1}^{T-t} \tilde{\psi}_{t+t'}(s_{t':t+t'}, \hat{\boldsymbol{\theta}}_T) \boldsymbol{h}_t(s_{0:t}, \hat{\boldsymbol{\theta}}_T)^\top\right).$$

補題 7

$\hat{\mathbf{A}}_T \to \mathbf{A}$,　$\hat{\mathbf{G}}_T(\hat{\boldsymbol{\theta}}_T) \to \mathbf{G}(\bar{\boldsymbol{\theta}})$

方策評価の情報量基準

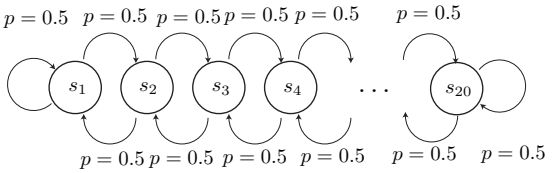
情報量基準

$$\text{IC}(s_{0:T}, \hat{\boldsymbol{\theta}}_T) = \frac{1}{T}\sum_{t=1}^{T} \left|\mathcal{R}_{t-1}(s_{t-1:T}) - g(s_{t-1}, \hat{\boldsymbol{\theta}}_T)\right|^2 + \frac{2}{T}\text{tr}\left\{\hat{\mathbf{A}}_T^{-1}\hat{\mathbf{G}}_T(\hat{\boldsymbol{\theta}}_T)\right\}$$

提案手法

1. 事前に候補となるモデル集合 $\mathcal{G}$ を準備する
2. モデルを 1 つ選択して，推定量 $\hat{\boldsymbol{\theta}}_T$ を計算する
3. **情報量基準 $\text{IC}(s_{0:T}, \hat{\boldsymbol{\theta}}_T)$ を評価する**
4. 2-3 を繰り返す
5. 情報量基準を最小とするモデルを採用する

シミュレーション実験



報酬関数: $r(s, s') = r^{(c)}(s, s')$　割引率: $\gamma = 0.95$

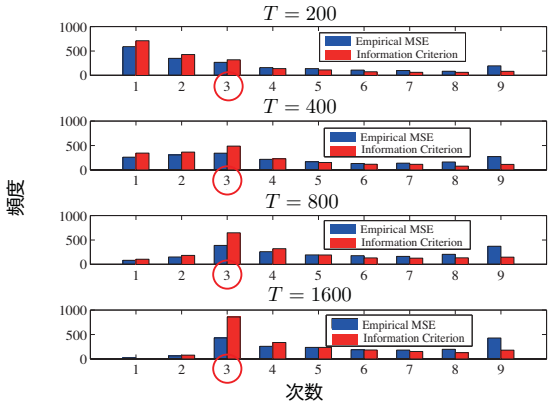
$\Longrightarrow$ このとき $V(s) = -1 + 0.9s - 0.12s^2 + 0.004s^3$

モデル: m' 次多項式 $g(s, \boldsymbol{\theta}) = \boldsymbol{\phi}(s)^\top \boldsymbol{\theta} = \theta_0 + \sum_{i=1}^{m'} s^i \theta_i$

推定量: TD 推定量

目的: 情報量基準を用いて次数を推定する

シミュレーション結果



議論

関連研究

- **基底関数の選択, 調節に関する研究**
  - Menache et al. (2005)
    - 基底関数のメタパラメータをTD誤差の二乗を最小にするように勾配法で調節する
  - Keller et al. (*ICML*, 2006) and Parr et al. (2007, 2008)
    - TD誤差を小さくするように基底関数を生成し，近似誤差が０になるまで追加し続ける

    $\Longrightarrow$ 近似誤差と推定誤差のトレードオフは考慮していない
- **ベイズ方策評価におけるモデル選択に関する研究**
  - Fard and Pineau (*NIPS*, 2010)
    - ベイズ方策評価の事前分布のパラメータの調節をPAC Boundを基準に行う

この節のまとめ

まとめ

- 情報量基準の考え方を方策評価に導入し，方策評価における新しいモデル選択法を提案した
- 提案したモデル選択法は，方策評価の幅広いモデル選択問題に適用することができる

今後の課題

- $\hat{\mathbf{A}}_T, \hat{\mathbf{G}}_T$の少数サンプル下での安定した推定

  $\Longrightarrow$ ベイズモンテカルロによる近似 (O'Hagan, 1991)

背景

準受動歩行(Wisse, Ruina, Collins, Osuka, Hosoda, Takuma, etc)

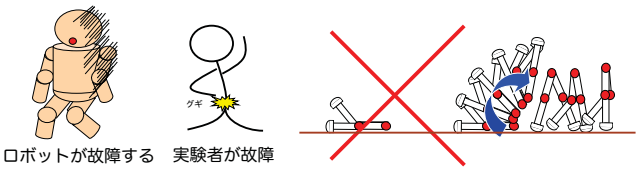
- ロボットの動特性を利用して，弾道学的な歩行をする

- 従来の2足歩行ロボット(ASHIMO (Hirai et al,2002)) の10倍以上のエネルギー効率を実現 (Collins et al, 2005)
- 制御パラメータは，環境，ロボット構造に非常に敏感に変化し，設計が困難

$\Longrightarrow$ 制御パラメータは自動的に調節されることが望ましい

2足歩行ロボットの学習の難しさ

1. 解の探索範囲が非常に広い
   - ２足歩行ロボットは他リンク構造
   - 出力，状態はいずれも連続量で様々な値を取り得る
2. サンプルが大幅に制限される
   - ロボットは転倒後，自発的に起き上がることはない
   - 実機実験特有の要因: 実験コスト



ロボットが故障する　実験者が故障

本研究の目的

不安定で転倒する実機の2足歩行ロボットに
強化学習を用いて
準受動歩行の頑健な制御器を獲得する

学習対象ロボット
大阪大学浅田研究室
QU-KAKU

解の探索範囲が非常に広い

$\Longrightarrow$ 先行研究 (Takuma et al, 2007), (Hitomi et al., 2006) を元に制御器を大きく限定する

サンプルが大幅に制限される

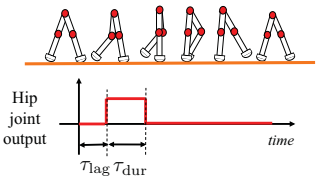
$\Longrightarrow$ 過去に得たサンプルを再利用可能なアルゴリズム 方策オフ型自然勾配法, off-NAC法 (Mori et al, 2005) を採用する

先行研究 (Hitomi, 2006)

Simulation で受動歩行の安定した制御器の獲得に成功している

- 両足着地時に着目して出力を決定すること
- 矩形波上の出力パターンを設定すること
- 膝関節にあらかじめ定められた制御を用いること

で制御器のモデルを大幅に限定している.



制御器のモデル

パラメータ行列

$$\bar{\boldsymbol{\tau}} = \tanh(\mathbf{W}\boldsymbol{x})$$

出力ベクトル $\bar{\boldsymbol{\tau}} = (\tau_{\rm lag}, \tau_{\rm dur})^\top$

入力ベクトル $\boldsymbol{x} = (1, x)^\top$

パラメータ行列 $\mathbf{W}$ をoff-NAC法で学習する

方策オフ型自然勾配法 off-NAC (Mori, 2005)

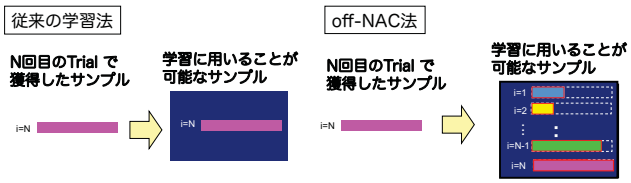
- 方策を直接パラメトライズし，確率勾配法でパラメータを学習する

方策: $p(a_t|s_t;\vartheta)$　　方策パラメータ: $\vartheta \in \Theta', \Theta' \subset \mathbb{R}^n$

- 重点サンプリグ重み(ISW)で過去の方策で獲得したサンプル系列を重み付けして勾配推定に利用する

$$\mathcal{IW} = \frac{p(x_{0:T}, a_{0:T}; \vartheta)}{p(x_{0:T}, a_{0:T}; \vartheta_b)} = \prod_{t=1}^{T} \frac{p(a_t|s_t;\vartheta)}{p(a_t|s_t;\vartheta_b)}$$

現在の方策 / 行動方策



実装

方策関数

- 方策関数は確率的な方策を使用する．実際に出力する制御信号 $\boldsymbol{\tau}$ は制御モデルの出力 $\bar{\boldsymbol{\tau}}$ にガウスノイズを付加する

$$\boldsymbol{\tau} \sim p(\boldsymbol{\tau}|x;\mathbf{W}) \equiv \mathcal{N}(\boldsymbol{\tau}|\bar{\boldsymbol{\tau}}(x,\mathbf{W}),\boldsymbol{\Sigma})$$

行動 $(a \leftarrow \boldsymbol{\tau})$　状態 $(s \leftarrow x)$　方策パラメータ $(\vartheta \leftarrow \mathbf{W})$

報酬関数

- 定常歩行時の性質に着目して報酬を以下のように設定する

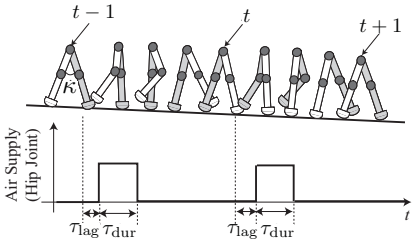
$$r_t = \exp(-\varsigma'(x_t - x_{t-1})) + r_{\rm falldown}$$

1 歩前の状態に近いほど高い報酬　転倒時のペナルティ項

シミュレーション実験

サンプル再利用の有効性を確認するため
シミュレーション実験を行う

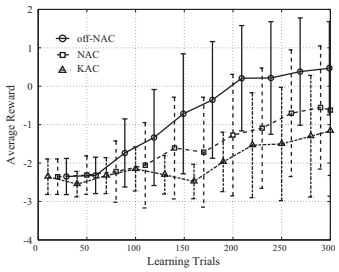
状態変数の設定: $x \leftarrow \kappa$



シミュレーション実験

サンプル再利用の有効性を確認するシミュレーション

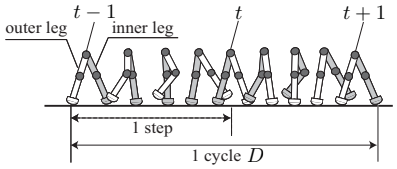
- Kimura's Actor-Critic (Kimura et al., 1999)
- Natural Actor-Critic (Peters et al., 2003)
- off-Policy Natural Actor Critic (Mori et al. 2005)



サンプル再利用型の方が
高速で安定に学習できた

実機実験

状態変数の設定: $x \leftarrow D$



(a) 学習実験

(b) ロバスト実験