Elastic Net を用いた状態価値関数の推定における有限標本解析

Finite sample analysis of value function estimation via the Elastic Net

数学専攻　小林 知陽

KOBAYASHI, Kazuaki

1　はじめに

近年, デジタル化や人工知能の発展に伴い, 多種多様で膨大なデータを扱うことが可能となり, 最適な意思決定のルールを扱う強化学習が注目されるようになった. そして, 強化学習と深層学習を組み合わせる手法が次々と提案され, 多くの領域で性能を発揮している. 強化学習は, その他にも医療統計分野において治療を行動として強化学習の枠組みを用いる動的治療計画の研究 [2] や通信ネットワーク経路の最適化におけるネットワークモデルなど, 多岐にわたる領域での応用が期待されている.

強化学習では, 与えられた環境の下で報酬関数や状態遷移確率を近似しながら状態価値関数 (目的関数) を最大にするような行動をエージェントに学習させる. しかし, 状態数が膨大, あるいは状態変数が連続値をとる場合には状態価値関数を求めることが困難になる. この問題に対し, 状態価値関数を線形関数で近似する方法が提案されている. 線形関数近似を用いた強化学習では, 推定したい状態価値関数を特徴ベクトルとパラメータを用いて線形関数で近似し, そのパラメータを推定する. 推定法としては, 一致性をもつパラメータを推定する Least-Squares Temporal Difference (LSTD) 法 [1] や, 状態価値関数に対して影響が強い特徴ベクトルに重きを置いてパラメータを推定する Lasso-TD [5] などがある. これらの推定法は履歴データに基づいているので, 報酬関数や状態遷移確率を近似する必要がなく, 汎用性が高い.

Lasso-TD は高々サンプルサイズ分までの特徴ベクトルしか選択することができない. 本研究では, サンプルサイズに依存せず, 状態価値関数に大きく影響を与える特徴ベクトルを適切に選択できる方法を提案する.

2　強化学習

強化学習は, ある環境の下においてエージェントが現在の状態を観測し, その情報から取るべき行動を学習する機械学習の一種である. 強化学習では, エージェントと環境の相互作用によって主体的にデータを収集し, 得られた環境の情報をもとに最適な行動を決定することを目的とする.

強化学習では, エージェントが環境に対して行動を入力することで, 環境の次の状態と報酬を観測する. この操作を十分なデータが得られるまで繰り返す. このような状況をモデル化するために, 状態の確率変数 $\mathcal{S}$ と行動の確率変数 $\mathcal{A}$ の確率制御過程 $\{S_t, A_t | t \in \mathbb{N}_0\}$ を考える. 確率制御過程には一般にマルコフ性を仮定する. マルコフ連鎖に行動と報酬を組み入れた確率制御過程をマルコフ決定過程と呼び, 以下の 5 つの組 $\mathrm{M} = \{\mathcal{S}, \mathcal{A}, P_{s_0}, P, R\}$ で構成される.

有限状態集合　$\mathcal{S} = \{s_1, \ldots, s_{|\mathcal{S}|}\} \ni s$

有限行動集合　$\mathcal{A} = \{a_1, \ldots, a_{|\mathcal{A}|}\} \ni a$

初期状態確率関数　$p_{s_0} : \mathcal{S} \to [0, 1],\ p_{s_0} = \Pr(S_0 = s)$

状態遷移確率関数　$P : \mathcal{S} \times \mathcal{S} \times \mathcal{A} \to [0, 1], \quad P(s'|s, a) = \Pr(S_{t+1} = s' | S_t = s, A_t = a)$

報酬関数　$R : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$

状態遷移確率関数 P はマルコフ性に従い, 現在の状態と行動のみに依存して次の状態が確率的に定まる. 強化学習では多くの場合, 環境をマルコフ決定過程でモデル化する.

マルコフ決定過程では, エージェントが現在の状態 s あるいは過去の履歴から次のとるべき行動 a を選択する. この行動選択のルールを決定する関数を方策という. 一般的には, 現在の状態のみに依存して確率的に行動を選択する確率的方策

$$\pi(a|s) = \Pr(A = a|S = s) \tag{2.1}$$

が用いられる. 方策 $\pi(a|s)$ を環境に入力することで, データ $\{s_0, a_0, r_0, \ldots, s_t, a_t, r_t, \ldots\}$ が得られる.

強化学習の目的関数としては割引累積報酬の期待値である状態価値関数

$$V^\pi(s) = \mathrm{E}\left[\sum_{t=0}^{\infty} \gamma^t R(s_t, a_t) \,\middle|\, s_0 = s\right]$$

を用いる. ここで $\gamma \in [0, 1)$ は割引率と呼ばれ, 長期的な報酬和をどの程度考慮するのかを調整するハイパーパラメータである. 割引累積報酬を考慮することにより, エージェントが直近で得られる報酬に重きを置くようになる. 任意の初期状態 $s \in \mathcal{S}$ に対する状態価値関数を推定し, 状態価値関数を最大にする方策

$$\pi^* = \arg\max_\pi V^\pi(s) \tag{2.2}$$

を求める. 強化学習では収集した次の状態や報酬のデータを用いて, 各状態の状態価値関数を推定して, それをもとに方策 π^* を求める.

3 スパースモデルにおける線形関数近似を用いた状態価値関数の推定

状態数が膨大な場合や状態変数が連続値をとる場合, 状態価値関数を推定することが困難になる. このような状況で用いられる方法として, 状態価値関数を k 個の基底関数 $\{\phi_i | i = 1, \ldots, k\}$ からなる特徴ベクトル $\boldsymbol{\phi} \in \mathbb{R}^k$, パラメータ $\boldsymbol{w} \in \mathbb{R}^k$ を用いて線形関数として近似する方法がある. このとき, 状態価値関数を

$$V^\pi(s) \approx \sum_{i=1}^{k} w_i \phi_i(s) = \boldsymbol{w}^\top \boldsymbol{\phi}(s) \tag{3.1}$$

と表す. 線形関数近似を用いた強化学習では, パラメータ $\boldsymbol{w}$ を推定することが目的となる.

本節では, 線形関数近似を用いた強化学習の代表的な手法として Least-Squares Temporal Difference (LSTD) 法 [1] と Lasso-TD [5] を説明する. さらに, Lasso の問題点 [4] を解決する方法として, LSTD 法の目的関数に Elasitic Net [3] を用いた方法を提案する.

3.1 線形関数近似を用いた状態価値関数のパラメータ推定

線形関数近似を用いた強化学習のパラメータの推定法として Least-Squares Temporal Difference (LSTD) 法がある. LSTD 法は, 得られる状態価値関数がベルマン方程式を近似的に満たすようにパラメータを推定する方法である. 強化学習では報酬関数と状態遷移確率は未知なので, 代わりにマルコフ決定過程から得られたサンプルデータ

$$\boldsymbol{\phi}(s) = \begin{pmatrix} \phi_1(s) \\ \vdots \\ \phi_k(s) \end{pmatrix}, \boldsymbol{\Phi} = \begin{pmatrix} \boldsymbol{\phi}(s_1)^\top \\ \vdots \\ \boldsymbol{\phi}(s_n)^\top \end{pmatrix}, \boldsymbol{\Phi}' = \begin{pmatrix} \boldsymbol{\phi}(s_1')^\top \\ \vdots \\ \boldsymbol{\phi}(s_n')^\top \end{pmatrix}, \boldsymbol{R} = \begin{pmatrix} r_1 \\ \vdots \\ r_n \end{pmatrix} \tag{3.2}$$

を用いる. ただし $\boldsymbol{\Phi}$ は状態の特徴行列, $\boldsymbol{\Phi}'$ は次状態の特徴行列, $\boldsymbol{R}$ は報酬ベクトルである. このとき, LSTD 法ではパラメータを以下のように推定する:

$$\boldsymbol{w} = \arg\min_{\boldsymbol{u} \in \mathbb{R}^k} \frac{1}{2} \left\| \boldsymbol{\Phi}\boldsymbol{u} - (\boldsymbol{R} + \gamma \boldsymbol{\Phi}' \boldsymbol{w}) \right\|_2^2 . \tag{3.3}$$

ここで $\boldsymbol{u}$ は線形関数空間上に射影させたときの状態価値関数の不動点のパラメータを示す. 式 (3.3) の解は簡単な計算により

$$\boldsymbol{w} = \left(\boldsymbol{\Phi}^\top(\boldsymbol{\Phi} - \gamma\boldsymbol{\Phi}')\right)^{-1}\boldsymbol{\Phi}^\top\boldsymbol{R} \tag{3.4}$$

であることがわかる.

Lasso の考え方を援用して, LSTD 法の目的関数に L_1 ノルムを追加し,

$$\boldsymbol{w} = \underset{\boldsymbol{u}\in\mathbb{R}^k}{\arg\min}\,\frac{1}{2}\left\|\boldsymbol{\Phi}\boldsymbol{u} - (\boldsymbol{R} + \gamma\boldsymbol{\Phi}'\boldsymbol{w})\right\|_2^2 + \lambda\|\boldsymbol{u}\|_1 \tag{3.5}$$

の最適化によりパラメータを推定する方法を Lasso-TD [5] と呼ぶ. ただし $\lambda > 0$ は正則化パラメータである. 特徴ベクトルの次元数 k がサンプルサイズ n よりも多いとき, Lasso-TD は過学習を起こさないので, 精度のよい推定値が得られる. さらに, Lasso-TD による推定はパラメータの推定値を 0 とすることができるので, 推定と同時に状態価値関数に影響のある特徴量ベクトルの選択を行うことができる.

3.2 Elastic Net を用いた状態価値関数

Lasso は高々サンプルサイズ分までの特徴量しか選択できない. また, 観測した状態の特徴量に相関がある場合, どちらか一方しか選択されない [4]. この問題点をもつ Lasso では, 目的変数に影響を与える説明変数を適切に抽出することが難しい. そこで, LSTD 法の目的関数に L_1 ノルムと L_2 ノルムの重み付き和を追加した式

$$\boldsymbol{w} = \underset{\boldsymbol{u}\in\mathbb{R}^k}{\arg\min}\,\frac{1}{2}\|\boldsymbol{\Phi}\boldsymbol{u} - (\boldsymbol{R} + \gamma\boldsymbol{\Phi}'\boldsymbol{w})\|_2^2 + \alpha\lambda\|\boldsymbol{u}\|_1 + \frac{(1-\alpha)}{2}\lambda\|\boldsymbol{u}\|_2^2 \tag{3.6}$$

を用いてパラメータを推定する. ただし $\boldsymbol{w}$ はベルマン作用素によって更新したパラメータ, $0 \le \alpha \le 1$ は調整パラメータである. Elastic Net を用いることで, サンプルサイズに依存せず, 推定と同時に状態価値関数に大きく影響を与える特徴量を適切に選択できる.

3.3 Elastic Net を用いた状態価値関数のパラメータ推定

式 (3.6) のパラメータ推定法として交互方向乗数法 (alternating determination method of multipliers) がある. 交互方向乗数法では, ラグランジュの未定乗数法とペナルティ法を組み合わせた拡張ラグランジュ関数

$$L_\rho(\boldsymbol{u},\boldsymbol{z},\boldsymbol{h}) = \frac{1}{2}\|\boldsymbol{\Phi}\boldsymbol{u} - (\boldsymbol{R} + \gamma\boldsymbol{\Phi}'\boldsymbol{w})\|_2^2 + \alpha\lambda\|\boldsymbol{z}\|_1 + \frac{1-\alpha}{2}\lambda\|\boldsymbol{z}\|_2^2 + \frac{\rho}{2}\|\boldsymbol{w} - \boldsymbol{z} + \boldsymbol{u}\|_2^2 + \frac{\rho}{2}\|\boldsymbol{h}\|_2^2 \tag{3.7}$$

をもとに更新式を構築する. ただし $\boldsymbol{h}$ はラグランジュ乗数, ρ は正の調整パラメータである. 推定したいパラメータ $\boldsymbol{u},\boldsymbol{z}$ に対して拡張ラグランジュ関数の最小化を行い, その最小値を用いてラグランジュ乗数を勾配法で更新していく. これを収束するまで繰り返す. LSTD 法と同様に, 更新したパラメータ $\boldsymbol{u}$ を線形関数空間に射影させるために, $\boldsymbol{u} = \boldsymbol{w}$ と置換することで更新式を得る:

$$\begin{cases} \boldsymbol{w}^{k+1} = (\boldsymbol{A} + \rho\boldsymbol{I})^{-1}\left(\boldsymbol{b} + \rho(\boldsymbol{z}^k - \boldsymbol{u}^k)\right), & (3.8)\\ \boldsymbol{z}^{k+1} = S_{\frac{\alpha\lambda}{(1-\alpha)\lambda+\rho}}\left(\dfrac{\rho}{(1-\alpha)\lambda+\rho}(\boldsymbol{w}^{k+1} + \boldsymbol{u}^k)\right), & (3.9)\\ \boldsymbol{h}^{k+1} = \boldsymbol{h}^k + \boldsymbol{w}^{k+1} - \boldsymbol{z}^{k+1}. & (3.10) \end{cases}$$

ここで sign は符号関数であり, パラメータ $\boldsymbol{A},\boldsymbol{b}$ や軟閾値作用素 S は

$$\boldsymbol{A} = \left(\boldsymbol{\Phi}^\top(\boldsymbol{\Phi} - \gamma\boldsymbol{\Phi}')\right)^{-1}, \quad \boldsymbol{b} = \boldsymbol{\Phi}^\top\boldsymbol{R},$$

$$S_{\frac{\alpha\lambda}{(1-\alpha)\lambda+\rho}}\left(\frac{\rho}{(1-\alpha)\lambda+\rho}(\boldsymbol{w}^{k+1} + \boldsymbol{u}^k)\right) = \operatorname{sign}\left(\frac{\rho}{(1-\alpha)\lambda+\rho}(\boldsymbol{w}^{k+1} + \boldsymbol{u}^k)\right) \cdot \max\left\{\left|\frac{\rho}{(1-\alpha)\lambda+\rho}(\boldsymbol{w}^{k+1} + \boldsymbol{u}^k)\right| - \frac{\alpha\lambda}{(1-\alpha)\lambda}\right\}$$

である.

3.4 Elastic Net を用いた状態価値関数の有限標本解析

Elastic Net では, 状態価値関数に対して影響が強い特徴ベクトルに重きを置いてパラメータを推定するので, 状態価値関数のパラメータの真値と一致しない. そのため, Elastic Net で推定したパラメータによる状態価値関数と真の状態価値関数には誤差が生じる. この誤差を近似誤差と推定誤差で評価すると, 以下の定理の結果が得られる.

定理 3.1 $\{s_t\}_{t=1}^n$ をマルコフ報酬過程における状態とする. また, $\boldsymbol{v}, \boldsymbol{f}_{\hat{\boldsymbol{w}}}, \boldsymbol{f}_{\boldsymbol{w}}$ をそれぞれ $\{s_t\}_{t=1}^n$ の真の状態価値関数, Elastic Net により推定したパラメータを用いた状態価値関数, 真の状態価値関数に対して線形関数近似したものとする. このとき, 任意の $\delta > 0$ に対して確率 $1-\delta$ で

$$\|\boldsymbol{v} - \boldsymbol{f}_{\hat{\boldsymbol{w}}}\|_n \leq \frac{1}{1-\gamma}\left\{\|\boldsymbol{v} - \boldsymbol{f}_{\boldsymbol{w}}\|_n + \sqrt{4\gamma V_{\max} L}\left(\left(\frac{2\log(2k/\delta)}{n}\right)^{1/4} + \frac{1}{\sqrt{2n}}\right)\right.$$
$$\left.\cdot\left(\sqrt{(2\alpha+1)\|\boldsymbol{w}\|_1 + (1-\alpha)\|\boldsymbol{w}\|_2^2}\right)\right\} \tag{3.11}$$

が成り立つ.

定理 3.1 より, 式 (3.11) の推定誤差はサンプルサイズ n が大きなるほど減少し, 特徴ベクトルの次元数が大きくなると増加することがわかる.

4 まとめと今後の展望

本研究では, 線形関数近似を用いた強化学習のパラメータ推定に焦点を当て, LSTD 法の目的関数に Elastic Net を取り入れた推定法を提案した. さらに, Elastic Net で推定したパラメータによる状態価値関数の評価を与えた. 今後の課題として, 正則化パラメータ λ の制御が挙げられる. 一般的なスパースモデルでは交差検証や拡張 BIC などを用いて決めるが, 現時点で実装を行うときは λ を決め打ちで設定することしかできない. そのため, 適切な λ を設定することができず, 状態価値関数に影響を与える特徴ベクトルを適切に選択できない可能性がある. そこで, 目的関数の自由度に基づいた評価基準を作成することが今後の課題である.

参考文献

[1] Bradtke, S., Barto, A. (1996). Linear least-squares algorithms for temporal difference. In *Machine Learning*, pp.22,33-57.

[2] Chakraborty, B., Moodie, E. E. (2013). Statistical methods for dynamic treatment regimes, In *Springer*

[3] Hui, Z., Trevor, H. (2005). Regularization and Variable Selection via the Elastic Net In *Journal of the Royal Statistical Society. Series B*, pp.301-320.

[4] Ryan, J, T. (2013). The lasso problem and uniqueness, In *Electronic Journal of statistics 7*, pp1456-1490.

[5] Tsipinakis, N., Nelson, J. D. B. (2015). Sparse Temporal Difference Learning via Alternating Direction Method of Multipliers In *Proceedings of the 20th international joint conference on Artifical intelligence*, pp.2586-2591.