スケッチ行列を用いたニュートン法による マルコフ決定過程の解法

九州大学大学院数理学府数理学専攻　新田　健人
Kento Nitta
Department of Mathematics, Graduate School of Mathematics,
Kyushu University
九州大学マス・フォア・インダストリ研究所　吉良　知文
Akifumi Kira
Institute of Mathematics for Industry,
Kyushu University

概要

多くの意思決定タスクはマルコフ決定過程 (MDP) として定式化される．価値反復法 (Value Iteration; VI) は MDP において最適な価値関数を求めるための標準的なアルゴリズムであるが, 一階法であるため, 割引率が大きい場合や大規模な問題に対しては収束が遅くなるという課題がある．そのため近年では, VI にニュートン法などの二階法を適用した手法が研究されており, 特にスケッチ行列を用いた手法は, 二階法のボトルネックとなる逆行列計算のコストを大幅に削減できるとして注目されている．本研究では, 最適化手法の一つである「Randomized Subspace Newton (RSN)」に着想を得た手法を提案する．数値実験の結果, 提案手法は従来のスケッチ化ニュートン法と比較して数値的に安定した挙動を示し, 大規模な MDP において VI や既存の二階法的手法に対する優位性が確認された．

1　研究の背景

マルコフ決定過程 (Markov Decision Processes; MDP) は, 意思決定問題を定式化するために利用される古典的な数理モデルである．強化学習 (Reinforcement Learning; RL) は, 状態遷移確率や報酬などの環境が未知の場合や、動的計画法による厳密解法の適用が困難な場合の近似手法である．多くの RL アルゴリズムは, ベルマン方程式の確率的近似として捉えることができる．例えば, Q 学習 [12] は, Q ベルマン方程式を解くための確率的な不動点反復とみなせる [7]．したがって, モデル情報が既知である MDP を解くためのより効率的なアルゴリズムを開発することは, RL アルゴリズムの性能向上にも寄与する重要な課題である．

最適な価値関数を求めるための古典的な手法として, 価値反復法 (Value Iteration; VI) と方策反復法 (Policy Iteration) がある．これらは連続最適化の理論と密接な関係があり, VI と方策反復法を一階法 (first-order method) もしくは二階法 (second-order method) と捉えることができる．包括的な知見については Grand-Clément [5] を参照されたい．VI は, ある非線形方程式の不動点を求める問題とみなせるため, アンダーソン加速やネステロフ加速といった加速法が適用されてきた

[4, 15]. また, マルコフ決定過程は線形計画問題として定式化することができ [10], これを単体法で解くことは方策反復法に対応し, 主双対内点法で解く際には内部でニュートン法が利用される. 一方, Kamanchichi ら [7] は, 最大化演算子を滑らかな関数で近似することにより, 行動価値関数のベルマン方程式に対して直接的にニュートン法を適用する方法を提案している. その後, 価値反復法にニュートン法を適用し, Newton Value Iteration (NVI) が Liu ら [9] によって提案された. NVI により, アルゴリズムの各反復あたりの計算コストは $\mathcal{O}(|\mathcal{S}|^3|\mathcal{A}|^3)$ から $\mathcal{O}(|\mathcal{S}|^3)$ へと大幅に削減された. しかし, 依然として状態数の大きな問題に対しては各反復あたりの計算コストが高く, 同文献 [9] では, 二階法の次元削減アルゴリズムである Sketched Newton-Raphson (SNR)[13] を適用した Sketched Newton Value Iteration (SNVI) が提案されている.

2 研究の目的

本研究の貢献は, 一点目は, SNVI が SNR 由来の強力な理論基盤を有している一方で, その構造からアルゴリズム中の係数行列の条件数が大きいという意味で数値的安定性に欠ける場合があることを明らかにした点である. 二点目は, 数ある二階法の次元削減手法 [3, 6, 14] の中でも, Gower ら [3] によって最適化問題に対して提案された Randomized Subspace Newton (RSN) に着想を得て類似のアルゴリズムを定式化し, VI に適用した手法を提案した点である. さらに, 提案手法が特定の MDP において SNVI と VI よりも高速に収束することを数値実験により示した.

3 導入

強化学習および動的計画法の基礎となるマルコフ決定過程（Markov Decision Processes; MDP）について定義する. MDP は, 確率的な状態遷移と報酬を伴う環境下での意思決定問題を数理的に定式化したものである.

まず, 確率過程の性質について述べる. 時間ステップ $t \in \mathbb{N} := \{0, 1, 2, \dots\}$ における状態を確率変数 S_t で表す. 確率過程 $\{S_t\}_{t\in\mathbb{N}}$ がマルコフ性を満たすとは, 将来の状態の条件付き確率分布が現在の状態のみに依存し, 過去の履歴に依存しないことを指す. すなわち, 任意の t と状態の実現値列 $s_0, \dots, s_{t+1}$ に対して, 次式が成り立つ.

$$\Pr(S_{t+1} = s_{t+1} \mid S_t = s_t, \dots, S_0 = s_0) = \Pr(S_{t+1} = s_{t+1} \mid S_t = s_t).$$

この性質を持つ確率過程をマルコフ過程（Markov Process）あるいはマルコフ連鎖（Markov Chain）と呼ぶ. MDP は, このマルコフ過程に「行動（Action）」と「報酬（Reward）」の要素を組み込んだ確率制御過程である.

n 状態 m 行動のマルコフ決定過程は, 5 つの組 $\mathcal{M} := \langle \mathcal{S}, \mathcal{A}, P, r, \gamma \rangle$ として定義される. 各要素の定義は以下の通りである.

- $\mathcal{S} := \{s_1, \dots, s_n\}$: 有限状態集合.
- $\mathcal{A} := \{a_1, \dots, a_m\}$: 有限行動集合.
- $P : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$: 状態遷移確率. ここで $\Delta(\mathcal{S})$ は $\mathcal{S}$ 上の確率単体を表す. 状態 $s \in \mathcal{S}$ で行動 $a \in \mathcal{A}$ を選択した際に, 次の状態が $s' \in \mathcal{S}$ となる確率を $P(s' \mid s, a)$ と書く. また, ベク

トルとして表したものを $P_{sa} \in \mathbb{R}^n$ と書く.

- $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$: 報酬関数. 状態 s で行動 a をとった際に得られる即時報酬を $r(s,a)$ とする. なお, 報酬関数は有界である, すなわち $|r(s,a)| \le R_{\max}$ を満たす定数 $R_{\max} > 0$ が存在すると仮定する.
- $\gamma \in [0,1)$: 割引率. 将来の報酬を現在の価値に換算するための係数である.

エージェントの行動決定ルールを方策（policy）と呼ぶ. 本研究では, 確率的方策 $\pi : \mathcal{S} \to \Delta(\mathcal{A})$ を考える. すなわち, 状態 s において行動 a を選択する確率を $\pi(a \mid s)$ と定義する. 方策 π の集合を Π と表す.

$$\Pi := \left\{ \pi : \mathcal{S} \times \mathcal{A} \to [0,1] \,\middle|\, \sum_{a \in \mathcal{A}} \pi(a \mid s) = 1,\ \forall s \in \mathcal{S} \right\}.$$

MDP における目的は, 無限期間にわたる期待割引累積報酬を最大化する方策を見つけることである. ある方策 π のもとでの状態価値関数 $V^\pi : \mathcal{S} \to \mathbb{R}$ は次のように定義される.

$$V^\pi(s) := \mathbb{E}^\pi \left[\sum_{t=0}^{\infty} \gamma^t r(S_t, A_t) \,\middle|\, S_0 = s \right].$$

ここで $\mathbb{E}^\pi$ は, 初期状態 $S_0 = s$ から開始し, 方策 π と遷移確率 P に従って生成される軌道に関する期待値を表す. 最適価値関数 V^* は, 全ての方策の中で最大の価値を与える関数として定義される.

$$V^*(s) := \max_{\pi \in \Pi} V^\pi(s).$$

最適価値関数 V^* は, 以下のベルマン最適方程式を満たすことが知られている [11].

$$V^*(s) = \max_{a \in \mathcal{A}} \left\{ r(s,a) + \gamma \sum_{s' \in \mathcal{S}} P(s' \mid s,a) V^*(s') \right\}, \quad \forall s \in \mathcal{S}. \tag{1}$$

式 (1) の解を不動点にもつ作用素をベルマン作用素という. 任意のベクトル $v \in \mathbb{R}^n$ に対し, ベルマン作用素 $T : \mathbb{R}^n \to \mathbb{R}^n$ は次のように定義される.

$$(T(v))(s) := \max_{a \in A} \left\{ r(s,a) + \gamma \sum_{s' \in S} P(s' \mid s,a) v(s') \right\}.$$

このとき, T は ∞-ノルム $\|\cdot\|_\infty$ に関して, 係数 γ の縮小写像となる. すなわち, 任意の $v, w \in \mathbb{R}^n$ に対して次式が成り立つ.

$$\|T(v) - T(w)\|_\infty \le \gamma \|v - w\|_\infty.$$

Banach の不動点定理より, T は唯一の不動点を持ち, それが最適価値関数 V^* と一致する.

$$V^* = T(V^*).$$

価値反復法（VI）は, ベルマン作用素 T を反復的に適用することで最適価値関数 V^* を求める手法である. 任意の初期ベクトル $v_0 \in \mathbb{R}^n$ から開始し, 以下の更新則に従って列 $\{v_k\}$ を生成する.

$$v_{k+1} = T(v_k), \quad k = 0, 1, \ldots$$

T が縮小写像であるため, この列は V^* に線形収束する. しかし, 割引率 γ が 1 に近い場合, 収束係数 γ に依存して収束が遅くなるという問題がある.

VI の収束速度を改善するために, ニュートン法を最適価値関数の探索問題に適用することを考える. まず関数 $F : \mathbb{R}^n \to \mathbb{R}^n$ を次のように定義する.

$$F(v) \coloneqq v - T(v).$$

このとき, v^* が T の不動点であることは, $F(v^*) = 0$ であることと同値である. したがって, MDP を解く問題は, 非線形方程式 $F(v) = 0$ の解 v^* を求める問題に帰着される. ここで, VI は $v_{k+1} = v_k - F(v_k)$ とみなすことができる. そのため, $v_{k+1} = v_k - \nabla F(v_k)^{-1} F(v_k)$ のような更新式を定義したい.

しかし, ベルマン作用素 T はその定義内に max 演算を含むため, 関数 $F(v)$ は一般に微分不可能である. したがって, ニュートン法を直接適用することは困難である. この問題を解決するために, max 演算子を滑らかな関数である LogSumExp 関数によって近似する手法が提案されている. LogSumExp 関数とは次のように定義される関数である.

定義 3.1. $x \in \mathbb{R}^n$, $\beta > 0$ に対して, $f_\beta(x)$ を次のよう定める.

$$f_\beta(x) = \frac{1}{\beta} \log \sum_{i=1}^{n} e^{\beta x_i}.$$

$f_\beta(x)$ と $f(x) \coloneqq \max_{i \in \{1,2,\dots,n\}} x_i$ の間には次の不等式が成り立つことが知られている.

$$|f(x) - f_\beta(x)| \leq \left| \frac{1}{\beta} \log \sum_{i=1}^{n} e^{\beta(x_i - f(x))} \right| \leq \left| \frac{\log n}{\beta} \right|.$$

すなわち, $f_\beta(x)$ は max 演算子を近似する関数になっており, $\beta \to \infty$ で $f(x)$ に一致する. LogSumExp 関数を用いて Smooth Bellman Operator を次のように定義する.

定義 3.2. $\beta > 0$ に対して, *Smooth Bellman Operator* $T_\beta : \mathbb{R}^n \to \mathbb{R}^n$ を次のように定義する.

$$(T_\beta(v))(s) \coloneqq \frac{1}{\beta} \log \left(\sum_{a \in \mathcal{A}} \exp \left(\beta (r(s,a) + \gamma \sum_{s' \in \mathcal{S}} P(s' \mid s, a) v(s')) \right) \right), \quad \forall s \in \mathcal{S}.$$

重要な点として, $T_\beta(v)$ は v に関して連続微分可能である. そのヤコビ行列 $\nabla T_\beta(v)$ の各成分は次のように計算される.

$$(\nabla T_\beta(v))(s, s') = \frac{\gamma \sum_{a \in A} P(s' \mid s, a) \exp\left(\beta (r(s,a) + \gamma P_{sa} v)\right)}{\sum_{a \in A} \exp\left(\beta (r(s,a) + \gamma P_{sa}^\top v)\right)}, \quad \forall s, s' \in \mathcal{S}.$$

以上の準備の下, Smooth Bellman Operator を用いてニュートン法を適用した手法が Newton Value Iteration (NVI) である. NVI では, 関数 $F_\beta(v) \coloneqq v - T_\beta^w(v)$ に対して, 以下の更新を行う.

$$v_{k+1} = v_k - (\nabla F_\beta(v_k))^{-1} F_\beta(v_k).$$

以降は $F_\beta(v)$ を単に $F(v)$ と表記する.

Liu ら [9] は, NVI が大域的に二次収束することを示している. 一方で, NVI は各反復においてヤコビ行列 $\nabla F(v_k)$ の逆行列（または線形方程式の解）を求める必要がある. この計算コストは $\mathcal{O}(n^3)$ であり, 状態数 n が大規模な問題においては計算が困難となる. この計算におけるボトルネックを解消するために, 近年研究されているのが, スケッチ（sketch）行列を用いたニュートン法である.

そのようなニュートン法の一つに Yuan ら [13] によって提案された Sketched Newton Raphson (SNR) がある. SNR を価値反復法に適用した手法が Sketched Newton Value Iteration (SNVI) である. SNVI は NVI と同様に Liu ら [9] によって提案された. 関数 $F(v) = v - T_\beta(v)$ とステップサイズ $\alpha \in \mathbb{R}$, スケッチ行列 $\mathbf{S}_k \sim \mathcal{D}_{v_k}$, $\mathbf{S}_k \in \mathbb{R}^{n\times\tau}$, $\tau \ll n$ に対して, SNVI は以下の更新則に従う.

$$v_{k+1} = v_k - \alpha\nabla F(v_k)^\top \mathbf{S}_k(\mathbf{S}_k^\top \nabla F(v_k)\nabla F(v_k)^\top \mathbf{S}_k)^\dagger \mathbf{S}_k^\top F(v_k). \tag{2}$$

ここで, $(\cdot)^\dagger$ はムーア・ペンローズ逆行列である. 更新則より逆行列の計算コストは $\mathcal{O}(n^3)$ から $\mathcal{O}(\tau^3)$ に改善される. また, スケッチ行列は仮定する分布 $\mathcal{D}_{v_k}$ によって多様に定義できる. 代表的なスケッチ行列のサンプリング方法に uniform subsampling sketch がある.

定義 3.3 (uniform subsampling sketch)**.**

$$\mathbb{P}[\mathbf{S} = \mathbf{I}_\mathcal{C}] = \frac{1}{\binom{n}{\tau}} \quad \textit{for all set } \mathcal{C} \subset \{1,\dots,n\} \textit{ s.t. } |\mathcal{C}| = \tau,$$

ここで, $\mathbf{I}_\mathcal{C}$ は単位行列 $\mathbf{I} \in \mathbb{R}^{n\times n}$ から重複を許さずにランダムに τ 個の列を抜き出し, 結合した行列である.

SNVI の収束性は SNR の理論的枠組みを適用することで直接的に示される. まず, 次のような仮定をおく.

仮定 3.1. ある $v^* \in \mathbb{R}^n$ が存在して, $F(v^*) = 0$ を満たす.

まず, 以下のように記号を定義する.

$$\mathbf{H}_\mathbf{S}(v) \coloneqq \mathbf{S}(\mathbf{S}^\top \nabla F(v)^\top \nabla F(v)\mathbf{S})^\dagger \mathbf{S}^\top, \quad f_{\mathbf{S},y}(v) \coloneqq \frac{1}{2}\|F(v)\|^2_{\mathbf{H}_\mathbf{S}(v)}, \quad f_y(v) \coloneqq \mathbb{E}[f_{\mathbf{S},y}(v)].$$

ここで, 行列 $\mathbf{M}$ が対称半正定値行列のとき, $\|x\|_\mathbf{M} \coloneqq x^\top \mathbf{M} x$ と定める. このとき, star-convexity を次のように定義する.

仮定 3.2 ($f_y(v)$ の star-convexity (Yuan et al. [13]))**.** 仮定 *3.1* が成り立つとする. 更新式 (2) によって生成される点列 $\{v_t\}$ に対して, 以下の不等式が成り立つ.

$$f_{v_t}(v^*) = 0 \geq f_{v_t}(v_t) + \langle \nabla f_{v_t}(v_t), v^* - v_t\rangle.$$

仮定 3.3 ($f_{\mathbf{S},y}(v)$ の star-convexity (Yuan et al. [13]))**.** 仮定 *3.1* が成り立つとする. 更新式 (2) によって生成される点列 $\{v_t\}$ に対して, 以下の不等式が成り立つ.

$$f_{\mathbf{S}_t,v_t}(v^*) = 0 \geq f_{\mathbf{S}_t,v_t}(v_t) + \langle \nabla f_{\mathbf{S}_t,v_t}(v_t), v^* - v_t\rangle.$$

仮定 3.3 が成り立つならば, 仮定 3.2 が従うことに注意されたい.

これらの仮定のもとで, $f_v(v)$ が期待値の意味で大域的にサブリニア収束することが知られている.

定理 3.1 (Yuan et al. [13])**.** 仮定 *3.2* が成り立つとする．$0 < \alpha < 1$ となるならば，次の不等式が成り立つ．

$$\mathbb{E}\left[\min_{t=0,\dots,k-1} f_{v_t}(v_t)\right] \leq \frac{1}{k}\sum_{t=0}^{k-1}\mathbb{E}[f_{v_t}(v_t)] \leq \frac{1}{k}\frac{\|v_0 - v^*\|^2}{2\alpha(1-\alpha)}.$$

さらに，仮定 *3.3* が成り立つならば，更新式 (2) によって生成される点列 $\{v_t\}$ は有界である．すなわち，

$$\|v_t - v^*\| \leq \|v_0 - v^*\|.$$

となる．

ここで，$\mathbf{H_S}(v)$ は対称半正定値行列であるため，$\|\cdot\|_{\mathbf{H_S}(v)}$ はセミノルムである．したがって，$f_v(v)$ が減少したとしても $\|F(v)\|$ の減少は示せない．しかし，スケッチ行列に関する仮定を課すことで，$\|F(v)\|$ が期待値の意味で大域的にサブリニア収束することが知られている．

定理 3.2 (Yuan et al. [13])**.**

$$\begin{aligned}
\rho(v) &\coloneqq \min_{u\in\mathrm{Im}(\nabla F(v)^\top)\setminus\{0\}} \frac{u^\top \nabla F(v)^\top \mathbb{E}[\mathbf{H_S}(v)]\nabla F(v)u}{\|u\|^2},\\
\rho &\coloneqq \min_{v\in\{v \mid \|v-v^*\|\leq\|v_0-v^*\|\}} \rho(v),\\
L &\coloneqq \sup_{v\in\{v \mid \|v-v^*\|\leq\|v_0-v^*\|\}} \|\nabla F(v)^\top\| > 0,
\end{aligned}$$

とする．このとき，$0 \leq \rho \leq 1$ である．もし，任意の $v \in \mathbb{R}^n$ に対して

$$F(v) \in \mathrm{Im}(\nabla F(v)) \subset \mathrm{Im}(\mathbb{E}[\mathbf{H_S}(v)]), \quad \forall v \in \mathbb{R}^n,$$

が成り立つならば任意の $v \in \mathbb{R}^n$ に対して，$\rho(v) = \lambda^+_{\min}(\nabla F(v)^\top \mathbb{E}[\mathbf{H_S}(v]]\nabla F(v)) > 0$, $\rho > 0$ となる．さらに，仮定 *3.1*，仮定 *3.2*，仮定 *3.3*，$0 < \alpha < 1$ のもとで次の不等式が成り立つ．

$$\mathbb{E}\left[\min_{t=0,\dots,n-1} \|F(v_t)\|^2_\infty\right] \leq \frac{1}{n}\frac{L^2\|v_0 - v^*\|^2}{\rho\alpha(1-\alpha)}.$$

すなわち，$\|F(v)\|_\infty$ は期待値の意味で，大域的にサブリニア収束する．

SNVI はこのように理論的な収束保証を持つ一方で，実用上は複数の MDP インスタンスにおいて収束速度が著しく低下する現象が見られる．この収束悪化の主たる原因として，擬逆行列の計算過程における数値的不安定性が挙げられる．具体的には，SNVI における係数行列の条件数が悪化し，数値誤差が増幅される可能性がある．ここで，条件数とは次のように定まる数である．

定義 3.4. 正則行列 $A \in \mathbb{R}^{n\times n}$ の条件数 $\kappa(A) \in \mathbb{R}$ を次のように定める．

$$\kappa(A) \coloneqq \|A\|\|A^{-1}\| = \frac{\sigma_{\max}(A)}{\sigma_{\min}(A)}$$

ここで，$\|\cdot\|$ は 2-ノルムであり，$\sigma_{\max}$, $\sigma_{\min}$ はそれぞれ最大特異値と最小特異値である．

数値解析の知見によれば、線形系 $Mx = b$ を解く際の誤差は条件数 $\kappa(M)$ に比例して拡大する．具体的には, 条件数が 1 桁上がるごとに有効桁数が 1 桁失われると言われる．Numpy 等で標準的な倍精度演算（約 16 桁の精度）を用いる場合、$\kappa(M) = 10^{12}$ では有効桁が 4 桁まで落ち込み, $\kappa(M) = 10^{15}$ に至っては実質的に 1 桁の精度しか確保できず, 数値的な破綻を招く恐れがある．

そこで本研究では, より数値的に安定な手法として, 最適化分野で提案されている Randomized Subspace Newton (RSN) 法 [3] に着目する．Gower ら [3] によってよって提案された RSN とは次の最適化問題を考える．

$$\min_{x\in\mathbb{R}^n} f(x),$$

ここで, $f : \mathbb{R}^n \to \mathbb{R}$ は凸で二階微分可能な関数である．このとき, RSN は以下の更新を行う．

$$x_{k+1} = x_k - \frac{1}{\hat{L}}\mathbf{S}_k(\mathbf{S}_k^\top \nabla^2 f(x_k)\mathbf{S}_k)^\dagger \nabla f(x_k).$$

ここで, $\hat{L} > 0$ は任意の $x, y \in \{x \in \mathbb{R}^n \mid f(x) \le f(x_0)\}$ に対して次の不等式を満たすような実数である．

$$f(x) \le f(y) + \langle \nabla f(y), x - y\rangle + \frac{\hat{L}}{2}\|x - y\|^2_{\nabla^2 f(y)}.$$

我々は RSN と類似した手法を不動点探索問題に対して提案し, 価値反復法に適用した手法を提案する．結果として, 提案手法において SNVI と同様に $\|F(v)\|$ が期待値の意味で大域的にサブリニア収束する．

4 提案手法

n 状態 m 行動マルコフ決定過程 $\mathcal{M} := \langle \mathcal{S}, \mathcal{A}, P, r, \gamma\rangle$ を考える．提案手法は次の更新則に従う．

$$v_{k+1} = v_k - \alpha\mathbf{S}_k \left(\mathbf{S}_k^\top \left(\nabla F(v_k) + \lambda\mathbf{I}\right)\mathbf{S}_k\right)^{-1} \mathbf{S}_k^\top F(v_k)\ ,\ \lambda > 0 \tag{3}$$

ここで, $\lambda\mathbf{I}$ は正則化項である．後述するように λ を十分大きくとれば, 提案手法は収束する．

まず, 更新式 (2), (3) 中の係数行列の条件数に関して次の定理が成り立つ．

定理 4.1. スケッチ行列 $\mathbf{S}$ を *uniform subsampling sketch* によってサンプリングする．$\lambda = 0$ での提案手法の係数行列を $A := \mathbf{S}^\top \nabla F(v)\mathbf{S}$, *SNVI* の係数行列を $B := \mathbf{S}^\top \nabla F(v)\nabla F(v)^\top \mathbf{S}$ とおく．このとき, 次の不等式が成り立つ．

$$\kappa(B) \ge \kappa(A)^2.$$

すなわち, *SNVI* の係数行列の条件数は提案手法の係数行列の条件数の 2 乗以上となる．

上記の定理より, 提案手法の更新則は SNVI と比較して数値的に安定していることがわかる．

次に提案手法の収束に関して述べる．定理の証明は SNR の証明手法を参考にしている．まず, いくつかの仮定をおく．

仮定 4.1. ある $\lambda \geq 0$ が存在し, 任意の $v \in \mathbb{R}^n$ に対して次が成り立つとする.

$$\frac{\nabla F(v) + \nabla F(v)^\top}{2} + (\lambda - 1)\mathbf{I} \succ 0.$$

仮定 4.2. 任意の $v \in \mathbb{R}^n$ に対して, $\nabla F(v)$ は正則である.

仮定 4.3. スケッチ行列 $\mathbf{S}$ は列フルランクである. すなわち, $\mathbf{S}$ の列ベクトルは線形独立である.

仮定 4.2, 仮定 4.3 は強い仮定であるが, MDP においては仮定 4.2 は満たされ, さらにスケッチ行列を uniform subsampling sketch によって生成すると, 仮定 4.3 も満たされる. 仮定 4.1 に関しては収束を理論的に保証するために必要であるが, 数値実験においては $\lambda = 0$ において最も高速に収束することが観察できた.

次のように記号を定義する.

$$\nabla F^\lambda(v) \coloneqq \nabla F(v) + \lambda \mathbf{I}, \quad \mathbf{H}_{\mathbf{S}}^\lambda(y) \coloneqq \mathbf{S}\left(\mathbf{S}^\top \nabla F^\lambda(v) \mathbf{S}\right)^\dagger \mathbf{S}^\top, \quad \mathbf{G}^\lambda(y) \coloneqq \frac{\mathbf{H}_{\mathbf{S}}^\lambda(y) + (\mathbf{H}_{\mathbf{S}}^\lambda(y))^\top}{2},$$

また, 以下の関数を考える.

$$f_{\mathbf{S},y}^\lambda(v) \coloneqq \frac{1}{2}\|F(v)\|^2_{\mathbf{G}^\lambda(y)}, \quad f_y^\lambda(v) \coloneqq \mathbb{E}\left[f_{\mathbf{S},y}^\lambda(v)\right] = \frac{1}{2}\|F(v)\|^2_{\mathbb{E}[\mathbf{G}^\lambda(y)]}.$$

SNVI と同様に $f_y^\lambda(v)$ の収束を示し, その後, $\|F(v)\|_\infty$ の収束を示す. まず, 次の定理が成り立つ.

定理 4.2. 仮定 *3.1*, 仮定 *4.1*, 仮定 *4.2*, 仮定 *4.3* を満たすとする. 更新式 (3) によって生成される点列 $\{v_k\}_{k\in\mathbb{N}}$ に対し, $f_{v_k}^\lambda(v_k)$ が次を満たすとする.

$$f_{v_k}^\lambda(v^*) = 0 \geq f_{v_k}^\lambda(v_k) + \left\langle \mathbb{E}_k\left[\mathbf{S}_k(\mathbf{S}_k^\top \nabla F^\lambda(v_k)\mathbf{S}_k)\mathbf{S}_k^\top F(v_k)\right], v^* - v_k \right\rangle. \tag{4}$$

このとき, $0 < \alpha < 1$ に対して, 大域的にサブリニア収束する. すなわち,

$$\mathbb{E}\left[\min_{t=0,\cdots k-1} f_{v_t}^\lambda(v_t)\right] \leq \frac{1}{k}\frac{1}{2\alpha(1-\alpha)}\|v_0 - v^*\|^2.$$

ここで $\mathbb{E} \coloneqq \mathbb{E}[\cdot \mid v_0]$, $\mathbb{E} \coloneqq \mathbb{E}[\cdot \mid v_k]$ である.

さらに $\|F(v)\|_\infty$ が大域的にサブリニア収束することを示すことができる.

定理 4.3. $\rho \coloneqq \inf_{v\in\mathbb{R}^n} \lambda_{\min}(\mathbf{G}^\lambda(v))$ とする. 定理 *4.2* のもとで, スケッチ行列を *uniform subsampling sketch* によって生成するとする. このとき, $\|F(v_t)\|_\infty$ は期待値の意味で大域的にサブリニア収束する. すなわち, 次の不等式が成り立つ.

$$\mathbb{E}\left[\min_{t=0,\ldots,n-1} \|F(v_t)\|_\infty^2\right] \leq \frac{1}{n}\frac{\|v_t - v^*\|^2}{2\rho\alpha(1-\alpha)}.$$

定理 4.3 より, 提案手法においても, より強い仮定のもとで SNVI と類似した性質が成り立つことがわかる. また, 仮定 3.3 に相当する不等式を提案手法においては, 定理に用いていないことに注意されたい.

5　数値実験

使用する MDP は Forest MDP, Machine Replacement MDP である．その他の MDP には例えば Healthcare MDP, Garnet MDP などがある．詳細は Goyal and Grand-Clement [4], Kamanchi, Diddigi, and Bhatnagar [7] , Delage and Mannor [2] を参照されたい． また, 特に Cordwell, Gonzalez, and Tulabandhula [1] の提供する MDP フレームワークを活用する．

条件数の比較

本節では条件数の比較を行う．$|\mathcal{S}| = 8000$, $\gamma = 0.9999$ のもとで数値実験を行う．

表 1: 各手法における条件数（Forest MDP, $S = 8000$, $\gamma = 0.9999$）

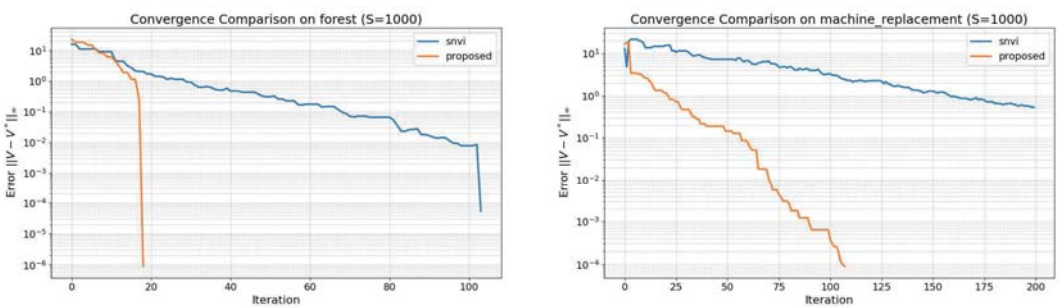
Algorithm	Min $\kappa_2(M)$	Max $\kappa_2(M)$	Mean $\kappa_2(M)$	Var $\kappa_2(M)$
提案手法	1.00	1.13×10^{7}	1.92×10^{5}	2.15×10^{12}
SNVI	1.60×10^{3}	4.43×10^{15}	2.10×10^{13}	6.69×10^{28}

$\kappa_2(M)$ は 2-ノルム条件数．最大値, 最小値, 平均, 分散を表示（有効数字 3 桁）．

実験の結果, 提案手法では条件数が 10^7 程度に収まっているのに対し, SNVI では 10^{15} 程度に達しており，有効桁数の損失は 15 桁程度に及ぶことが分かる．このような数値的不安定性は反復アルゴリズムにおいては特に顕著に収束に悪影響を及ぼすと考えられる．これは提案手法の数値的優位性を裏付ける実験的証拠となっている．

小規模実験

本節では小さい状態数・低い割引率のもとで反復回数と誤差の減少の関係を比較する．具体的には $|\mathcal{S}| = 1000$, $\gamma = 0.9$ のもとで数値実験を行う．



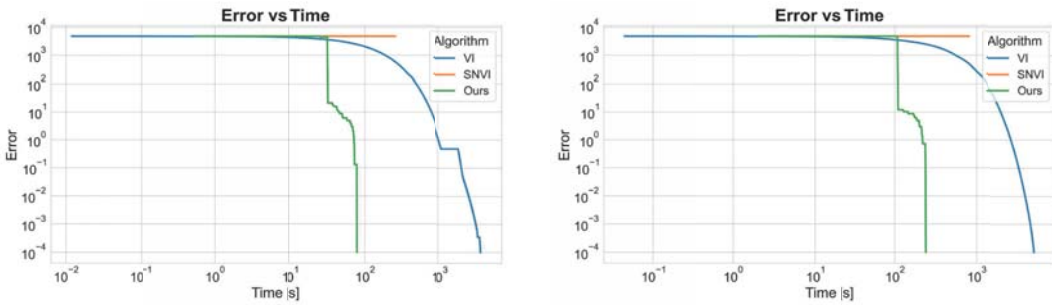
(a) Forest MDP　　(b) Machine replacement MDP

図 1: 反復回数 vs 誤差 (青線:SNVI, オレンジ線: 提案手法)

実験の結果, 提案手法は既存手法である SNVI と比較して, 反復回数あたりの誤差の減少が大きいことがわかる. 特に大規模実験においては差がより顕著になる.

大規模実験

本節では大きい状態数・高い割引率のもとで実行時間と誤差の減少の関係を比較する. 具体的には $|\mathcal{S}| = 5000$, $|\mathcal{S}| = 8000$, $|\mathcal{S}| = 10000$ の 3 種類の状態数と $\gamma = 0.9999$ のもとで Forest MDP での実験を行う. 比較する手法は VI と SNVI である. また, $\lambda = 0$ としている.



(a) S=5000　　(b) S=8000

図 2: 実行時間 vs 誤差 (青線:VI, オレンジ線:SNVI, 緑線: 提案手法)

結果的に Forest MDP において提案手法は VI と比較して 40 倍程度の高速化を達成できた. 一方で, SNVI に関しては小規模 MDP では収束するものの, 大規模 MDP においては収束が非常に遅いことが確認できた. また. Machine Replacement MDP, Healthcare MDP においては SNVI よりも高速に収束するものの VI 以上の収束は得られなかった.

6　おわりに

本研究では, RSN から着想を得て, SNVI よりも数値的に安定したアルゴリズムを提案した. また, 提案手法の収束に関する定理を与え, 数値実験によりより特定の MDP で VI, SNVI よりも高速な収束を確認した. 今後の研究課題として以下の 3 点が考えられる. 一点目は, 提案手法がどのような MDP に対して有効なのかを明らかにすること. 二点目は, 正則化パラメーター λ をノンパラメトリックに決定すること. 三点目は, 提案手法・SNVI をオンライン学習へ拡張することである.

謝辞

本研究は, 国際共同利用・共同研究拠点である京都大学数理解析研究所の支援を受けました.

参考文献

[1] S. Cordwell, Y. Gonzalez, and T. Tulabandhula: Markov decision process (MDP) toolbox for Python. *https://github.com/sawcordwell/pymdptoolbox*, Accessed 2024.

[2] E. Delage and S. Mannor: Percentile optimization for Markov decision processes with parameter uncertainty. *Operations Research*, **58** (2010), 203–213.

[3] R. Gower, D. Kovalev, F. Lieder, and P. Richtárik: RSN: Randomized subspace Newton. In *Advances in Neural Information Processing Systems*, **32** (2019), 614–623.

[4] V. Goyal and J. Grand-Clement: A first-order approach to accelerated value iteration. *Operations Research*, **71** (2023), 517–535.

[5] J. Grand-Clément: From convex optimization to MDPs: A review of first-order, second-order and quasi-Newton methods for MDPs. *arXiv preprint arXiv:2104.10677* (2021).

[6] F. Hanzely, N. Doikov, Y. Nesterov, and P. Richtárik: Stochastic subspace cubic Newton method. In *International Conference on Machine Learning* (PMLR, 2020), 4027–4038.

[7] C. Kamanchi, R. B. Diddigi, and S. Bhatnagar: Generalized second order value iteration in Markov decision processes. *Japan Journal of Industrial and Applied Mathematics*, **41** (2021), 637–657.

[8] Y. LeCun, Y. Bengio, and G. Hinton: Deep learning. *Nature*, **521** (2015), 436.

[9] J. Liu, C. Xie, Q. Deng, D. Ge, and Y. Ye: Sketched Newton value iteration for large-scale Markov decision processes. In *Proceedings of the AAAI Conference on Artificial Intelligence* (2024).

[10] A. Manne: Linear programming and sequential decisions. *Management Science*, **6**, No. 3 (1960), 259–267.

[11] M. L. Puterman: *Markov decision processes: discrete stochastic dynamic programming* (John Wiley & Sons, 2014).

[12] C. J. Watkins and P. Dayan: Q-learning. *Machine Learning*, **8** (1992), 279–292.

[13] R. Yuan, A. Lazaric, and R. M. Gower: Sketched Newton–Raphson. *SIAM Journal on Optimization*, **32** (2022), 1555–1583.

[14] C. Zhang, D. Ge, B. Jiang, and Y. Ye: DRSOM: A dimension reduced second-order method and preliminary analyses. *arXiv preprint arXiv:2208.00208* (2022).

[15] J. Zhang, B. O' Donoghue, and S. Boyd: Globally convergent type-I Anderson acceleration for nonsmooth fixed-point iterations. *SIAM Journal on Optimization*, **30** (2020), 3170–3197.