数理解析研究所講究録
第747巻 1991年 82-92

Markov 決定過程における 政策反復法と Newton - Raphson 法について

京都大学 工学部
大西 匡光　OHNISHI Masamitsu

1 はじめに

無限計画期間上の期待総割引き費用規範のもとでの Markov 決定過程に対する政策反復法（あるいは政策改良法）が最適性方程式を各状態での最適値関数値を決定変数とする非線形方程式と見た Newton - Raphson 法と等価であることは Hordijk and Puterman [5], Puterman and Brumelle [11], Whittle [14] らにより指摘されている．本報告では無限計画期間上の長時間平均の期待費用（略して平均費用）規範のもとでの Markov 決定過程を分数計画問題と見なすことにより，実パラメータを持つ問題の族を導入し，そのパラメータを決定変数とするある非線形方程式を解くためのアルゴリズムを提案し，その政策反復法および Newton - Raphson 法との関連について考察する．

2 平均費用規範有限 Markov 決定過程

以下で定義される有限 Markov 決定過程を考える．

$S := \{0, 1, 2, \cdots, M\}$: 有限状態空間,

$A(i)$, $i \in S$: 状態 i でとり得るアクションの有限集合,

$A := \cup_{i \in S} A(i)$: 有限アクション空間,

$K := \{(i, a) : i \in S,\ a \in A(i)\}$: 許容な状態－アクションの対,

$c(i, a)$, $(i, a) \in K$: 状態 i でアクション a をとったときの期待即時費用,

$p(i, a, j)$, $(i, a) \in K$, $j \in S$: 状態 i でアクション a をとったときに状態 j に遷移する確率.

履歴および政策の形式的な定義は以下の通りである．

$H_t := K^t \times S$, $t = 0, 1, 2, \cdots$: 時刻 t での履歴の集合,

$H_\infty := K^\infty$: 標本空間.

$H_t, t = 0, 1, 2, \cdots$ の要素

$$h_t = (x_0, a_0, x_1, \cdots, x_{t-1}, a_{t-1}, x_t)$$

において $(x_s, a_s) \in K, s = 0, 1, 2, \cdots$ は時刻 s における状態とアクションの対に対応している．

政策は履歴 $h_t \in H_t$ が与えられたときの A 上の条件付き確率分布 δ_t の列 $\delta = (\delta_t; t = 0, 1, \cdots)$ として定義される，ただし

$$\delta_t(A(x_t) \mid h_t) = 1 \text{ for all } t = 0, 1, 2, \cdots \text{ and } h_t \in H_t$$

を満たすものとする．すべての政策の集合を Δ とする．各 δ_t が t および x_t のみを通して h_t に依存する政策を Markov 政策と呼び，さらに δ_t が x_t のみを通して h_t に依存する場合確率的定常政策と言う．確率的定常政策は S が与えられたときの A 上の条件付き確率分布 δ_0，ただし

$$\delta_0(A(i) \mid i) = 1 \text{ for all } i \in S,$$

により完全に規定される．条件付き確率分布 δ_0 が縮退しているとき，すなわちある S から A への写像で

$$f(i) \in A(i) \text{ for all } i \in S$$

を満たすものが存在して

$$\delta(B \mid i) = 1_B(f(i))$$

となる場合，確定的定常政策（以下では単に定常政策）と言い，簡単のため $\delta = f$ と書く，ただし $1_A(\cdot)$ は集合 A の定義関数である．定常政策の全体を F と書くことにすれば，定常政策 f は

$$(f(0), f(1), f(2), \cdots, f(M)) \in \prod_{i \in S} A(i)$$

と同一視することができることから $F = \prod_{i \in S} A(i)$ と書いてもよい．

さて我々の目的は無限計画期間上の長時間平均期待費用

$$g_\delta(i) := \limsup_{T \to \infty} \frac{1}{T+1} E_\delta \left[\sum_{t=0}^{T} c(X_t, A_t) \middle| X_0 = i \right], \ i \in S \tag{2.1}$$

をすべての初期状態 $i \in S$ に対し最小化する政策 $\delta \in \Delta$ を求めることである，ただし $E_\delta[\cdot]$ は政策 $\delta \in \Delta$ のもとでの確率測度 $P_\delta(\cdot)$ に関する期待値を表し，X_t, A_t, $t = 0, 1, 2, \cdots$ は時刻 t における状態とアクションを表す確率変数である．

以下では次の Accessibility Condition のもとで議論を行う．

仮定 2.1 ある状態 $i_0 \in S$ が存在して，任意の定常政策 $f \in F$ のもとで，i_0 は他の全ての状態から到達可能である（一般性を失うことなく，$i_0 = 0$ とする）． □

仮定 2.1のもとでは最適な定常政策が存在し，さらに最適な平均費用は初期状態に依存しないことがよく知られている．すなわち

$$\min_{\delta \in \Delta} g_\delta(i) = \min_{f \in F} g_f =: g^* \text{ for all } i \in S. \tag{2.2}$$

この問題に対する最適性方程式は次の通りである.

$$v^*(i) = \min_{a \in A(i)} \left\{ c(i,a) - g^* + \sum_{j \in S} p(i,a,j) v^*(j) \right\}, \ i \in S \tag{2.3}$$

式(2.3)における $v^*(\cdot)$ は相対値関数と呼ばれ，付加定数を除いて一意的に定まる決定すべき未知の関数である．一般性を失うことなく

$$v^*(0) = 0 \tag{2.4}$$

としてよい．最適性方程式(2.3)，(2.4)の標準的な数値解法としては政策反復法，逐次近似法（あるいは値反復法），線形計画による解法などが挙げられるが，本報告では特に次に述べる政策反復法に着目する．

アルゴリズム 2.1（政策反復法）
Step 0（初期化）：初期政策 $f_0 \in F$ を与える．
Step 1（政策評価）：現在の政策 $f_n \in F$ のもとでの平均費用 g_{f_n} と相対値関数 $v_{f_n}(i)$, $i \in S$ を次の線形方程式系を解くことで計算する．

$$v_{f_n}(i) = c(i, f_n(i)) - g_{f_n} + \sum_{j \in S} p(i, f_n(i), j) v_{f_n}(j), \ i \in S \tag{2.5}$$

$$v_{f_n}(0) = 0 \tag{2.6}$$

Step 2（政策改良）：不等式

$$v_{f_n}(i) \geq c(i, f(i)) - g_{f_n} + \sum_{j \in S} p(i, f(i), j) v_{f_n}(j) \tag{2.7}$$

をすべての状態 $i \in S$ に対し，かつ少なくとも1つの状態では狭義の不等号で成立させる政策 $f \in F$ があれば $f_{n+1} \leftarrow f$, $n \leftarrow n+1$ として Step 1 へ，さもなくば停止；f_n は最適な定常政策である． □

通常の政策反復法では Step 2（政策改良）においては各状態 $i \in S$ に対し，式(2.7)の右辺を最小化するアクションをとる政策が新しい政策 f_{n+1} として選ばれる．

3 分数計画によるアプローチ

最適な定常政策が存在することから，対象を定常政策のクラス F のみに限定してよい．任意の定常政策のもとで状態0は再帰的であるため，状態0への訪問時刻の間隔をサイクルと考えれば再生報酬過程においてよく知られた事実から

$$[\text{平均費用}] = \frac{[1\text{サイクルの期待費用}]}{[1\text{サイクルの期待時間}]}$$

が成立する．

いま　定常政策 $f \in F$ に対し

$t_f(i)$, $i \in S$: 初期状態を i としたときの状態 0 への初到達時刻（$i=0$ の場合は初帰還時刻）の期待値,

$c_f(i)$, $i \in S$: 初期状態を i としたときの状態 0 への初到達時刻（$i=0$ の場合は初帰還時刻）までの期待費用,

と定義すれば

$$g_f = \frac{c_f(0)}{t_f(0)}$$

であるから，我々は次の分数計画問題 P を解けばよい.

$$\text{P:} \qquad \min_{f \in F} \frac{c_f(0)}{t_f(0)} \tag{3.1}$$

問題 P を解くために実数パラメータ g を含む問題の族 Q(g) を導入する.

$$\text{Q}(g): \quad \min_{f \in F} \{c_f(0) - g t_f(0)\} \tag{3.2}$$

次の定理は分数計画法においてよく知られている（例えば Schaible and Ibaraki [12] ）.

定理 3.1
(1) 定常政策 $f^* \in F$ を問題 P の最適解，g^* をその最適値とすると f^* は問題 Q(g^*) の最適解でもあり，そのとき

$$c_{f^*}(0) - g^* t_{f^*}(0) = 0$$

が成立する.
(2) 逆に，ある g^+ に対して問題 Q(g^+) が定常政策 $f^+ \in F$ を最適解として持ち，その最適値が 0 ，すなわち,

$$c_{f^+}(0) - g^+ t_{f^+}(0) = 0$$

が成立すれば f^+ は問題 P の最適解でもあり，g^+ はその最適値である. □

問題 Q(g) に関連して以下の関数を定義する.

$$v_f(i;g) := c_f(i) - g t_f(i), \ f \in F, \ i \in S \tag{3.3}$$

$$v^*(i;g) := \min_{f \in F} v_f(i;g), \ i \in S \tag{3.4}$$

定義式 (3.3) , (3.4) と F が有限集合であることから，$v^*(i;g)$, $i \in S$ は g に関して単調減少，区分的に線形な凹関数である. また，ある g に対して政策 $f_g \in F$ を

$$v_{f_g}(0;g) = v^*(0;g)$$

なるものとすれば $-t_{f_g}(0)$ は関数 $v^*(0;\cdot)$ の点 g における（優）勾配となっていることが容易に示される. さらに定理 3.1 より最適平均費用 g^* は非線形方程式

$$v^*(0;g)=0 \tag{3.5}$$

の唯一解である．よって二分法（Bisection Method），はさみうち法（Regula Falsi），割線法（Secant Method）などの標準的な方法により g^* に収束する g の列を生成し，その極限として問題 $\mathrm{Q}(g^*)$ を解く数値的解法が考えられる．

本報告では g^* に上から単調に収束する g の列を生成し，その極限として問題 $\mathrm{Q}(g^*)$ を解く，次の基本アルゴリズムをまず考える．

アルゴリズム 3.1（基本アルゴリズム 1）
Step 0（初期化）：初期政策 $f_0 \in F$ を与える．
Step 1：現在の政策 $f_n \in F$ のもとでの平均費用

$$g_{f_n}=\frac{c_{f_n}(0)}{t_{f_n(0)}}$$

を計算する．
Step 2：

$$v_f(0:g_{f_n}) < v_{f_n}(0:g_{f_n})=0 \ (\Leftrightarrow g_f < g_{f_n})$$

を満たす政策 $f \in F$ があればそれを求め，$f_{n+1} \leftarrow f$，$n \leftarrow n+1$ として Step 1 へ，さもなくば停止；f_n は最適な定常政策である． □

上記アルゴリズムの Step 2 において現在の政策 f_n から平均費用が厳密に改良される政策，すなわち

$$g_f < g_{f_n} \tag{3.6}$$

を満たす政策 f を構成することは，仮定 2.1 のみのもとでは困難であり，従って式 (3.6) の厳密な不等号 $<$ を $\leq$ に緩和することを考える．

アルゴリズム 3.2（基本アルゴリズム 2）
Step 0（初期化）：初期政策 $f_0 \in F$ を与える．
Step 1：現在の政策 $f_n \in F$ のもとでの平均費用

$$g_{f_n}=\frac{c_{f_n}(0)}{t_{f_n}(0)}$$

と関数 $v_{f_n}(i;g_{f_n})$, $i \in S$ を計算する．
Step 2： 不等式

$$v_f(i;g_{f_n}) \leq v_{f_n}(i;g_{f_n})$$

をすべての状態 $i \in S$ に対し，かつ少なくとも 1 つの状態では狭義の不等号で成立させる政策 $f \in F$ があればそれを求め，$f_{n+1} \leftarrow f$，$n \leftarrow n+1$ として Step 1 へ，さもなくば停止；f_n は最適な定常政策である． □

基本アルゴリズム 2 の有限収束性は F が有限集合であることから自明であり，またその正当性も容易に示すことができる．

4 K －次政策改良 Step

基本アルゴリズム2の Step 2 の実現方法を考えよう．関数 $v^*(i;g)$, $i \in S$ は状態 $i \in S$ でアクション $a \in A(i)$ をとったときの期待即時費用を $c(i,a)-g$ とする Markov 決定過程において，初期状態を i としたときの状態 0 への初到達時刻（$i=0$ の場合は初帰還時刻）までの期待総費用の最小値を表し，次の最適性方程式を満たす．

$$v^*(i;g) = \min_{a \in A(i)} \left\{ c(i,a) - g + \sum_{j \in S-\{0\}} p(i,a,j) v^*(j;g) \right\}, \ i \in S \tag{4.1}$$

最適性方程式 (4.1) に関連して実パラメータ g を含む変換 $T_f(g)$, $f \in F$, $T^*(g)$ を以下で定義する：S 上の実数値関数 $v(\cdot)$ に対し

$$\begin{aligned} [T_f(g)v](i) &:= c(i,f(i)) - g + \sum_{j \in S-\{0\}} p(i,f(i),j) v(j), \ i \in S, \\ [T^*(g)v](i) &:= \min_{f \in F} [T_f(g)v](i) \\ &= \min_{a \in A(i)} \left\{ c(i,a) - g + \sum_{j \in S-\{0\}} p(i,a,j) v(j) \right\}, \ i \in S. \end{aligned}$$

仮定 2.1より，任意の g に対し，変換 $T_f(g)$, $f \in F$, $T^*(g)$ は $(M+1)$ －段縮小性を持ち，従って最適性方程式 (4.1) は政策反復法，逐次近似法，線形計画による解法などの標準的方法で数値的に解くことができる．

本報告では基本アルゴリズム2の Step 2 として K $(=1,2,\cdots,+\infty)$ をパラメータとして持つ次の K －次政策改良 Step を考える．

Step 2（K －次政策改良）：
Step 2.1：$i \in S$ に対し

$$u(i) \leftarrow \left[T^*(g_{f_n})^{K-1} v_{f_n}(\cdot; g_{f_n}) \right](i)$$

とする．
Step 2.2：すべての状態 $i \in S$ に対し

$$[T_{f_n}(g_{f_n})u](i) = [T^*(g_{f_n})u](i)$$

ならば停止；f_n は最適な定常政策である，さもなくば $i \in S$ に対し

$$[T_f(g_{f_n})u](i) = [T^*(g_{f_n})u](i)$$

を成立させる政策 $f \in F$ を求め，$f_{n+1} \leftarrow f$, $n \leftarrow n+1$ として Step 1 へ戻る．

上記 K －次政策改良 Step により基本アルゴリズム2の Step 2 が実現できることの証明はさほど困難でない．

さて　以下の2点を指摘しておく．

1. パラメータを $K=1$ とすれば Step 2 は通常の政策反復法 2.1の Step 2（政策改良）と一致し，基本アルゴリズム 2 は通常の政策反復法 2.1に帰着される．

2. パラメータを $K=+\infty$ とすれば Step 2.1 は最適性方程式 (4.1) を完全に解き最適値関数 $v^*(\cdot;g_{f_n})$ を求め，それを関数 u とすることに対応している（実際には無限回の反復を行うことはできないし，またその必要もない）．このとき Step 2.2 より

$$v^*(\cdot;g_{f_n})=T^*(g_{f_n})v^*(\cdot;g_{f_n})=T_{f_{n+1}}(g_{f_n})v^*(\cdot;g_{f_n})$$

が成り立ち，

$$v_{f_{n+1}}(0;g_{f_n})=v^*(0;g_{f_n})$$

を満たす．従って先に述べたように $-t_{f_{n+1}}(0)$ は関数 $v^*(0;\cdot)$ の点 g_{f_n} における（優）勾配となっており，さらに

$$g_{f_{n+1}} = \frac{c_{f_{n+1}}(0)}{t_{f_{n+1}}(0)} \tag{4.2}$$

$$= g_{f_n}-\frac{c_{f_{n+1}}(0)-g_{f_n}t_{f_{n+1}}(0)}{[-t_{f_{n+1}}(0)]} \tag{4.3}$$

$$= g_{f_n}-\frac{v_{f_{n+1}}(0;g_{f_n})}{[-t_{f_{n+1}}(0)]} \tag{4.4}$$

であるから，基本アルゴリズム 2 は非線形方程式 (3.5) に対する Newton - Raphson 法に帰着される．

パラメータ K の値が大きければ大きいほど最適性方程式 (4.1) をより正確に解くことになることから，各反復における Step 2 の実行において，より効果的な政策の改良が期待され，より少ない反復回数でのアルゴリズムの収束が予想される．一方　各反復における Step 2 の実行は計算上負担となり，よってアルゴリズム全体の計算量を問題とすれば適切なパラメータ K の選択が鍵となろう．より一般的には各反復ごとにパラメータ K の値を変化させることも可能であり，それによってもアルゴリズムの有限収束性，正当性は保持される．

任意の定常政策 $f\in F$ に対して，遷移確率行列

$$[p(i,f(i),j);\ i,\ j\in S]$$

の（状態 0 に対応する）第 1 行と第 1 列を削除した劣確率行列が上あるいは下三角行列となる場合，最適性方程式 (4.1) を完全にとくことはさほど困難でなく，この時 Newton - Raphson 法は非常に有効である．なお　この場合の Newton - Raphson 法は通常の政策反復法における Step 2（政策改良）を Gauss - Seidel 的に実行した場合に一致する．このような遷移構造を持つ問題例として Markov 的劣化システムの最適保全問題が挙げられる（例えば Morioka, Ohnishi, and Ibaraki [8] ）．

5 おわりに

本報告では 平均費用規範有限 Markov 決定過程を分数計画問題としてとらえ，その数値的解法としてパラメータ K を含む K −次政策改良 Step を組み込んだアルゴリズムを提案し，考察した．本アルゴリズムはパラメータ K の選択により，特別な場合として政策反復法と Newton - Raphson 法を含んでおり，きわめて柔軟な方法であると考えられる．様々な問題に対して本アルゴリズムの適用を試み，十分な数値実験を行うことにより，問題の構造に基づく適切なパラメータ K の選択法に関する知見を得ることが今後の課題である．

最後に常日頃から御指導頂く関西大学 三根 久教授，京都大学 茨木俊秀教授に感謝します．

参考文献

[1] Bertsekas, D. P. (1976), *Dynamic Programming and Stochastic Control*, Academic Press, New York.

[2] Denardo, E. V. (1968), "Contraction Mapping in the Theory Underlying Dynamic Programming", *SIAM Review*, Vol. 9, pp. 165 - 177.

[3] Denardo, E. V. (1982), *Dynamic Programming: Models and Applications*, Prentice - Hall, Englewood Cliffs, New Jersey.

[4] Derman, C. (1970), *Finite State Markovian Decision Processes*, Academic Press, New York.

[5] Hordijk, A. and Puterman, M. L. (1987), "On the Convergence of Policy Iteration in Finite State Undiscounted Markov Decision Processes: The Unichain Case", *Mathematics of Operations Research*, Vol. 12, pp. 163 - 176.

[6] Howard, R. A. (1960), *Dynamic Programming and Markov Processes*, The MIT Press, Cambridege, Massachusetts.

[7] McNamara, J. M. (1989), "Policy Improvement and the Newton - Raphson Algorithm for Renewal Reward Processes", *Probability in the Engineering and Information Sciences*, Vol. 3, pp. 393 - 396.

[8] Morioka, T., Ohnishi, M., and Ibaraki, T. (1987), "Optimal Inspection and Replacement Problem of Markovian Deterioration System and its Computational Algorithms", in *Reliability Theory and Applications* (Osaki, S. and Cao, J.-H. Eds.), Proceedings of the China - Japan Reliability Symposium Held in Shanghai, Xian, and Beijing, China, September 13 - 25, 1987, World Scientific Publishing, Singapore, pp. 245 - 254.

[9] Puterman, M. L. (1990), "Markov Decision Processes", in *Stochastic Models*, Handbooks in Operations Research and Management Sciences, Vol. 2 (Heyman, D. P. and Sobel, M. J.

Eds.), Elsevier Science Publishers B. V. (North - Holland), Amsterdam, Chap. 8, pp. 331 - 434.

[10] Puterman, M. L. and Brumelle, S. L. (1978), "The Analytic Theory of Policy Iteration", in *Dynamic Programming and its Applications* (Puterman, M. L. Ed.), Proceedings of the International Conference on Dynamic Programming and its Applications Held in Vancouver, Canada, April 14 - 16, 1977, Academic Press, New York, pp. 91 - 113.

[11] Puterman, M. L. and Brumelle, S. L. (1978), "On the Convergence of Policy Iteration in Stationary Dynamic Programming", *Mathematics of Operations Research*, Vol. 4, pp. 60 - 69.

[12] Schaible, S. and Ibaraki, T. (1983), "Fractional Programming", *European Journal of Operational Research*, Vol. 12, pp. 325 - 338.

[13] Tijms, H. C. (1986), *Stochastic Modelling and Analysis: A Computational Approach*, John Wiley & Sons, Chichester.

[14] Whittle, P. (1989), "A Class of Decision Processes Showing Policy - Improvement / Newton - Raphson Equivalence", *Probability in the Engineering and Information Sciences*, Vol. 3, pp. 397 - 403.

[15] Whittle, P. and Komarova, N. (1988), "Policy Improvement and the Newton - Raphson Algorithm", *Probability in the Engineering and Information Sciences*, Vol. 2, pp. 249 - 335.

A 付録：割引き費用規範問題

割引き因子を $\beta\ (\in [0,1))$ とする無限計画期間上の期待総割引き費用

$$u_{\beta,\delta}(i) := E_\delta \left[\sum_{t=0}^{\infty} \beta^t c(X_t, A_t) \middle| X_0 = i \right],\ i \in S \tag{A.1}$$

をすべての初期状態 $i \in S$ に対し最小化する政策 $\delta \in \Delta$ を求めることを考える.

最適 β −割引き費用関数を

$$u^*_\beta(i) := \min_{\delta \in \Delta} u_{\beta,\delta}(i) = \min_{f \in F} u_{\beta,f}(i),\ i \in S \tag{A.2}$$

と定義すると，この問題に対する最適性方程式は次の様になる.

$$u^*_\beta(i) = \min_{a \in A(i)} \left\{ c(i,a) + \beta \sum_{j \in S} p(i,a,j) u^*_\beta(j) \right\},\ i \in S \tag{A.3}$$

最適性方程式 (A.3) の標準的な数値解法としては，平均費用規範問題と同様，政策反復法，逐次近似法（あるいは値反復法），線形計画による解法などが挙げられる.

割引き費用規範問題にたいする政策反復法は以下の通りである.

アルゴリズム A.1（政策反復法）
Step 0（初期化）：初期政策 $f_0 \in F$ を与える．
Step 1（政策評価）：現在の政策 $f_n \in F$ のもとでの β −割引き費用関数 $u_{\beta,f_n}(i),\ i \in S$ を次の線形方程式系を解くことで計算する．

$$u_{\beta,f_n}(i) = c(i, f_n(i)) + \beta \sum_{j \in S} p(i, f_n(i), j) u_{\beta,f_n}(j),\ i \in S \tag{A.4}$$

Step 2（政策改良）：不等式

$$u_{\beta,f_n}(i) \geq c(i, f(i)) + \beta \sum_{j \in S} p(i, f(i), j) u_{\beta,f_n}(j) \tag{A.5}$$

をすべての状態 $i \in S$ に対し，かつ少なくとも 1 つの状態では狭義の不等号で成立させる政策 $f \in F$ があれば $f_{n+1} \leftarrow f$，$n \leftarrow n+1$ として Step 1 へ，さもなくば停止；f_n は β −割引き最適な定常政策である． □

通常の政策反復法では Step 2（政策改良）においては各状態 $i \in S$ に対し，式 (A.5) の右辺を最小化するアクションをとる政策が新しい政策 f_{n+1} として選ばれる．

いま S 上の実数値関数 $u(\cdot)$ に対して

$$[U^*_\beta(u)](i) := u(i) - \min_{a \in A(i)} \left\{ c(i,a) + \beta \sum_{j \in S} p(i,a,j) u(j) \right\},\ i \in S. \tag{A.6}$$

で変換（あるいは S 上の実数値関数 $u(\cdot)$ を $\mathcal{R}^{M+1}$ の要素，すなわち $M+1$ 次元実数ベクトルと見なせば $\mathcal{R}^{M+1} \to \mathcal{R}^{M+1}$ の写像） $U^*_\beta(\cdot)$ を定義すれば最適性方程式 (A.3) を解くことは非線形方程式系

$$[U^*_\beta(u)](i) = 0,\ i \in S. \tag{A.7}$$

を解くことに他ならない． いま通常の政策反復法で生成される定常政策の列を $(f_n; n = 0, 1, 2, \cdots)$ とし，それらのもとでの β −割引き費用関数の列を $\mathcal{R}^{M+1}$ の点列とみなし $(u_n; n = 0, 1, 2, \cdots)$ とすれば

$$U^*_\beta(u_n) + (I - \beta P(f_{n+1}))(u_{n+1} - u_n) = 0,\ n = 0, 1, 2, \cdots \tag{A.8}$$

あるいは

$$u_{n+1} = u_n - (I - \beta P(f_{n+1}))^{-1} U^*_\beta(u_n),\ n = 0, 1, 2, \cdots \tag{A.9}$$

を満たす，ただし I は $(M+1) \times (M+1)$ −恒等行列であり，定常政策 $f \in F$ に対して

$$P(f) := [p(i, f(i), j); i, j \in S]$$

と定義する．一方 $(M+1) \times (M+1)$ −行列 $I - \beta P(f_{n+1})$ は写像 $U^*_\beta(\cdot)$ の u_n における Support（もしすべての状態 $i \in S$ で $f_{n+1}(i) \in A(i)$ が

$$\min_{a\in A(i)}\left\{c(i,a)+\beta\sum_{j\in S}p(i,a,j)u_n(j)\right\}, \tag{A.10}$$

の min を達成する唯一のアクションならば u_n における Jacobi 行列

$$\nabla U^*_\beta(u_n) := \left[\left.\frac{\partial[U^*_\beta(u)](i)}{\partial u(j)}\right|_{u=u_n} ; i,j\in S\right]) \tag{A.11}$$

であることから，割引き費用規範問題に対する通常の政策反復法は非線形方程式系 (A.7) に対する Newton - Raphson 法と等価となる．