数理解析研究所講究録
1147巻 2000年 78-87

割引マルコフ決定過程におけるしきい値確率の精度保証付き数値計算について

豊永 憲治 (Kenji Toyonaga)

Graduate School of Mathematics, Kyushu University,
Fukuoka 812-8581, Japan

E-mail: toyonaga@math.kyushu-u.ac.jp

概 要

割引マルコフ決定過程における最適しきい値確率を求めるには, 理論的には種々の計算方法が提案されているが, 現実的なモデルにおける最適しきい値確率を計算する場合, コンピュータが必要不可欠である. その際, 計算を有限回で打ち切った近似解と厳密解との誤差評価が必要になる. D.J White は値反復法に対する1つの誤差評価を示しているが, 実際に計算可能な形での誤差評価は与えられていない. したがって, この論文では真の解との誤差を評価するために精度保証付きで, 最適しきい値確率を計算する方法を述べる.

1. INTRODUCTION

マルコフ決定過程における最適しきい値確率を求める1つの方法として, 値反復法がある. 実際の計算において, 単純なモデルであれば, 反復を多く行うことができるが, state 数, action 数が多い一般のモデルでは, 高々10回の反復でも膨大な時間がかかってしまう. したがって, 複雑なモデルにおいては, 反復回数を多くとることはできない. したがって, 近似解と真の解の誤差評価の重要性が増してくる. D.J White は値反復法における1つの誤差評価を示している. しかしながら, それを用いて実際の誤差を数値的に評価することは困難である.

従ってこの論文では, 複雑なモデルにおける誤差の数値評価を行えるように, 真の解を包み込むような形で近似解を求め, 精度保証的付き計算を行うことにより, 真の解との誤差を数値として評価することを試みる. また, D.J White の誤差評価で使われている値との関係も示す. 最後にアルゴリズムと一般的な数値例を示す. ただし, この論文ではしきい値確率が分布関数となるものに限って扱うことにする.

2. NOTATIONS AND FORMULATION

割引きマルコフ決定過程の定式化として, $N = \{1, 2, \ldots\}$ を離散的な time space, S を state space (finite set) とし, X_tを time $t \in N$ での state とする. A を action space (finite set), $A(s)$ を $s \in S$ でとりうる action 全体, A_tを time $t \in N$ での action とする. Y_tを $t \in N$ での random immediate reward function とし,$H > 0$ に対し, $0 \leq Y_t \leq H$ とする. Y_t は (X_t, A_t) が与えられたときの, (X_{t+1}, Y_t) に対する次の式で与えられる.

$$p^a(s', y|s) = P(X_{t+1} = s', Y_t \leq y | X_t = s, A_t = a).$$

ρ : 割引き因子, $0 < \rho < 1$ とする.

$S \times R$ を a new state space, $R = (-\infty, \infty)$. $H_1 = S \times R$, $H_{t+1} = H_t \times A \times S \times Y_t, t \in N$. このとき, H_t は時刻 t までのすべての履歴の集合とする. 時刻 t での履歴を θ_t で表す.

時刻 t での decision rule δ_t は, $h_t = (s_1, r, a_1, s_2, y_1, \ldots, a_{t-1}, s_t, y_{t-1}) \in H_t$ に対して, $\delta_t(a_t|h_t) = P(A_t = a_t|\theta_t = h_t)$ で与えられる. また, すべての $h_t = (s_1, r, a_1, \ldots, s_t, y_{t-1}) \in H_t$ について, $\delta_t(A_t \in A(s_t)|h_t) = 1$ とする. また, $\delta_t(a_t|\cdot)$ は H_t上で, Lebesgue-Stieltjes measurable function とする.

すべての decision rule の集合をΔ, policy π は decision rules の無限列 $(\delta_1, \delta_2, \ldots, \delta_t, \ldots)$ で表される. policy 全体の集合を C とする.

policy πに対する有限期間と無限期間の総割引利得は

$$Z_0 = 0, \quad Z_n^\pi = \sum_{t=1}^{n} \beta^{t-1} Y_t^\pi, \ n \geq 1, \quad Z^\pi = \sum_{t=1}^{\infty} \beta^{t-1} Y_t^\pi.$$

となる. ただし, Y_t^πは policy πに対する時刻 t での利得とする. ここで,

$$W_t = (W_1 - Z_{t-1})/\beta^{t-1}, \quad t \geq 1,$$

として, 新しい履歴 $(X_1, W_1, A_1, X_2, W_2, \ldots, A_{t-1}, X_t, W_t)$ を考え, 実現値を $h_t = (s_1, w_1, a_1, s_2, w_2, \ldots, a_{t-1}, s_t, w_t)$ として考える.decision rule, policy は新しい履歴に対しても同様に定義される. Δ_tが履歴によらず $(X_t, W_t) = (s_t, w_t)$ の関数であるとき, Markov decision rule といい, Markov decision rule 全体を Δ_Mと表す. また$\delta_t \in \Delta_M$に対し, $\delta_t(A_t = a_t|s_t, w_t) = 1$ となる $a_t \in A_t$が存在するとき, deterministic decision rule という. deterministic decision rule 全体を Δ_D とする.

policy πの各 decision rule がΔ_Mに属するとき, Markov policy といい, Markov policy 全体を C_Mで表す. また, 各 decision rule がΔ_D に属し$\delta_t = \delta_{t+1}$となるとき, deterministic Markov stationary policy といい, その policy 全体を C_Dと表す.

しきい値関数を次のように下のように書く.

$$F_n^{\pi(n)}(s, r) = P(Z_n^\pi \leq r|s), \quad F^\pi(s, r) = P(Z^\pi \leq r|s).$$

ここでは, 無限期間問題である $P(Z^\pi \leq r|s)$ の最小化問題について考察する. そのときの最適解を

$$F^*(s, r) = \inf_{\pi \in C_M} F^\pi(s, r)$$

とする.

$S \times R$ から R への関数

$$\mathcal{F} = \{F : S \times R \to R | \text{単調非減少}, \forall s \text{に対し右連続}, F(s, r) = 0 (r < 0), F(s, r) = 1 (r \geq H/1 - \rho)\}$$

を考え, 次を定義する.

$$T^a F(s, r) = \int_{S \times R} F(s', (r - y)/\rho) dp^a(s', y \mid s),$$
$$T^\delta F(s, r) = \sum_{a \in A(s)} T^a F(s, r) \delta(a \mid (s, r)),$$
$$TF(s, r) = \inf_{\delta \in \Delta} T^\delta F(s, r) = \min_{a \in A(s)} T^a F(s, r).$$

また, $(T)^{n+1} = T(T^n)$, $T^1 = T$ とする.
$F, G \in \mathcal{F}$ が, すべての (s,r) に対し $F(s,r) \leq G(s,r)$ であるとき, $F \leq G$ と書く.

3. NUMERICAL ENCLOSURE FOR AN OPTIMAL THRESHOLD PROBABILITY

LEMMA 1 $F \in \mathcal{F}$ ならば, $T^a F \in \mathcal{F}$, $TF \in \mathcal{F}$.

LEMMA 2 $F, G \in \mathcal{F}, F \leq G$ ならば $TF \leq TG$.

上の Lemma は, T の定義よりいえる.
最適解を求める 1 つの方法として値反復法がある. 以下では値反復法の近似解により, 精度保証付きで最適解を求める方法を示す.

値反復法
$L_0 \in \mathcal{F}, L_n = TL_{n-1} = (T)^n L_0$.

まず, 値反復法による最適解への収束を保証する次の定理を示す.

THEOREM 1
値反復法において $L_0 = F_0$にとると

$$\lim_{n\to\infty} (T)^n F_0 = F^*.$$

ただし,$F_0(s,r) = 0(r < 0), F_0(s,r) = 1(r \geq 0)$.
Proof. この証明は, D.J White の論文の Theorem 2 による.

THEOREM 2
値反復法において $L_0 = G_0$にとり, F^*の左連続変形を G^*とする. ここで, 左連続変形とは右連続関数をその不連続点において左連続に直した関数とする. このとき次が成り立つ.

$$\lim_{n\to\infty} (T)^n G_0 = G^*.$$

ただし, $G_0(s,r) = 0(r < H/(1-\rho)), G_0(s,r) = 1(r \geq H/(1-\rho))$.
Proof $(T)^n G_0(s,r) = F_n^*(s, r - \frac{\rho^n H}{1-\rho})$ が成り立つことを帰納法で示す. n=1 のとき

$$\begin{aligned} T[G_0(s,r)] &= \inf_{k\in K(s)} \int G_0(s', \frac{r-y}{\rho}) dp^k(s', y \mid s) = \inf_{k\in K(s)} \int_{S\times(-\infty, r-\frac{H}{1-\rho}]} dp^k(s', y \mid s) \\ &= \inf_{\pi\in C(M)} F_1^{\pi(1)}(s, r - \frac{\rho H}{1-\rho}) = F_1^*(s, \frac{\rho H}{1-\rho}). \end{aligned}$$

n のとき成り立つと仮定して, n+1 のとき

$$F_{n+1}^*(s, r - \frac{\rho^{n+1} H}{1-\rho}) = \inf_{\pi(n+1)} P(Z_{n+1}^\pi \leq r - \frac{\rho^{n+1} H}{1-\rho} \mid s)$$

$$
\begin{aligned}
&= \inf_{\pi(n+1)} P(\sum_{t=2}^{n+1} \rho^{t-2} Y_t^{\pi} \leq \frac{r - Y_1^{\pi}}{\rho} - \frac{\rho^n H}{1-\rho} \mid s) \\
&= \inf_{k \in K(s), \pi(n)} \int_I P(\sum_{t=2}^{n+1} \rho^{t-2} Y_t^{\pi} \leq \frac{r-y}{\rho} - \frac{\rho^n H}{1-\rho} \mid X_2 = s') dp^k(s', y \mid s) \\
&= \inf_{k \in K(s)} \int_I \inf_{\pi(n)} F_n^{\pi(n)}(s', \frac{r-y}{\rho} - \frac{\rho^n H}{1-\rho}) dp^k(s', y \mid s) \\
&= TF_n^*(s, r - \frac{\rho^n H}{1-\rho}).
\end{aligned}
$$

従って,

$$(T)^n G_0(s,r) = F_n^*(s, r - \frac{\rho^n H}{1-\rho}).$$

Theorem 1 から, $F_n^*(s,r) \to F^*(s,r)$ より,

$$\lim_{n \to \infty} (T)^n G_0(s,r) = G^*(s,r).$$

実際に値反復法を用いて計算する場合, state 数, action 数がふえるにしたがって膨大な計算時間がかかり, 値反復を数多く行うことができない. したがって, 途中の時点での値反復に対する誤差評価が必要となる. 誤差評価の 1 つとして D.J White[5] より次の定理が与えられている. ただし, ここで扱うしきい値確率は分布関数になっているものと仮定する.

THEOREM 3
$n = lm, \quad n, l, m \in N$(自然数).

$$\lambda(m) = \sup_{(s,r) \in I, \pi \in C_D(m)} [F_m^{\pi(m)}(s,r) - F_m^{\pi(m)}(s, r - \frac{\rho^m H}{1-\rho})]$$

とおく. ただし, $F_m^{\pi(m)}(s,r) : \pi(m)$ をとったときのしきい値確率とする.
このとき, 次が成り立つ.

$$\| L_n - F^* \| \equiv \sup_{(s,r) \in I} \mid L_n(s,r) - F^*(s,r) \mid \leq \{\lambda(m)\}^l$$

Theorem 3 の結果は, a priori な誤差評価としては意味をもつが, $\lambda(m)$ の計算において, policy πの決定ルールは (s,r) に依存するので, policy π の数は無限個存在し, 実際の計算は困難である. したがって, 計算可能な誤差の評価が必要である.

この論文では, 真の解を含むような形で近似解を計算することによって, 真の解との a posteriori な誤差評価を可能とすることを試み, 実際にそのような評価法を与える. また D.J White の論文にある$\lambda(m)$ を使っての評価との関係も示す.

THEOREM 4
$L_n = (T)^n F_0$, $L'_n = (T)^n G_0$ とする. このとき

$$L'_n \leq F^* \leq L_n$$

がなりたつ. また$\lambda(m) \equiv \sup_{(s,r)\in I, \pi\in C_D(m)} [F_m^{\pi(m)}(s,r) - F_m^{\pi(m)}(s, r - \frac{\rho^m H}{1-\rho})]$ とすると,

$$\|L_n - F^*\| \leq \|L_n - L'_n\| \leq 2\lambda(m)^l$$

が成り立つ. ただし
$F_0(s,r) = 0 \quad (r < 0), \quad 1 \quad (r \geq 0), \quad G_0(s,r) = 0 \quad (r < \frac{H}{1-\rho}), \quad 1 \quad (r \geq \frac{H}{1-\rho})$.

(proof)

$$G_0 \leq F^* \leq F_0.$$

Lemma 2 より,

$$(T)^n G_0 \leq (T)^n F^* \leq (T)^n F_0.$$

$$L'_n \leq F^* \leq L_n.$$

また

$$\|L_n - F^*\| \leq \|L_n - L'_n\|.$$

$$\|L_n - L'_n\| \leq \|L_n - F^*\| + \|F^* - L'_n\|.$$

Theorem 3 と同様な議論より, $\|F^* - L'_n\| \leq \{\lambda(m)\}^l$から

$$\|L_n - F^*\| \leq \|L_n - L'_n\| \leq 2\{\lambda(m)\}^l.$$

4. ALGORITHM

このセクションではコンピュータにより, 無限期間問題の最適解を精度保証付きで計算するアルゴリズムを与える. 一般的な問題では state 数, action 数が 1 桁の場合でも値反復法により近似解を求めるには膨大な計算量が必要なため, コンピュータが必要不可欠である. したがって, ここでは一般的なモデルを計算するときのアルゴリズムを与える. そして精度保証付きで最適解を求める例を示す.

Theorem 4 より真の解 F^*は L_nと L'_nの間に存在することが保証されるので,L_n, L'_nを計算すればよい. L_0として,F_0を選択するためTF_0の計算は不連続点の位置を決定することとなる. 従って以下では,L_n, L'_nの不連続点の位置を見出すことを基本とした L_n, L'_nの計算アルゴリズムを示す.

1. モデル推移 state の設定
   state 数 m, $(s_1, \cdots, s_m)$, action 数 n, $(a_1, \cdots, a_n)$, 推移確率 p_{ij}^k, reward w_{ij}^k,
   割引因子ρ の設定. $\quad 1 \leq i \leq m, 1 \leq j \leq m, 1 \leq k \leq n$
   ただし, p_{ij}^k, reward w_{ij}^kは $p_{ij}^k(w_{ij}) = P(X_{t+1} = s_j, Y_t = w_{ij}, | X_t = s_i, A_t = a_k)$ により決まる.

2. $L_0(s_i,r)=F_0(s_i,r)$ の設定
   $F_0(s_i,r)$ として $(x,y)=(0,1)$ を設定.

3. L_nの計算実行
   ただし, ここでは $L_n \equiv F_n$とする .

   (a) $T^{a_k}F_0(s_i,r) = \sum_j p_{ij}^{a_k} F_0(s_j,(r-w_{ij}^{a_k})/\rho)$ の計算. $\quad 1 \le i \le m, 1 \le k \le n$
   $F_0(s_i,r)$ の不連続点から, $p_{ij}^{a_k}F_0(s_j,(r-w_{ij}^{a_k})/\rho)$ の不連続点を計算する. 次に $j=1$ に対し, $p_{i1}^{a_k}F_0(s_1,(r-w_{i1}^{a_k})/\rho)$ の各不連続点 (u,v) について, u 以下で $p_{ij}^{a_k}F_0(s_j,(r-w_{ij}^{a_k})/\rho)$ $2 \le j \le m$ の x 座標が最大の不連続点の y 座標を v に加える. $j=2$ から m まで繰り返し$\sum_j$を計算する.

   (b) $F_1(s_i,r)=\min_k T^{a_k}F_0(s_i,r)$ の計算の実行. $\quad 1 \le i \le m$
   $T^{a_k}F_0(s_i,r)$ の各不連続点において, $T^{a_k}F_0(s_i,r) 1 \le k \le n$ の最小値を設定する.

   (c) $F_t(s_i,r)$ が求まったところで, (a) の $F_0(s_i,r)$ を $F_t(s_i,r)$ として, (a), (b) を繰り返し $F_{t+1}(s_i,r)$ を計算する.

4. $L_0(s_i,r)=G_0(s_i,r)$ の設定
   $G_0(s_i,r)$ として $(x,y)=(H/(1-\rho),1)$ を設定.
   ただし, H: reward $w_{ij}^k \quad 1 \le i \le m,\ 1 \le j \le m,\ 1 \le k \le n$ の最大値

5. L_n'の計算実行
   3 における F を G に置き換えて (a),(b),(c) を繰り返し計算する.

以上の計算により, L_n, L_n'が求まると, 真の解は L_n, L_n'の間に存在することが Theorem 4 より保証されるので, 真の解との誤差は. $\|L_n - L_n'\|$ 以下となる.

EXAMPLE

この例題において, 一般的なモデルにおける最適しきい値確率の精度保証付き計算を, 値反復法を用いて行う. いくつかの論文においては, state 数 2 , action 数 2 の単純なモデルを例題としているが, 一般的な問題に対してはまだ示されていない. モデルが複雑になると, 膨大な計算が必要になるため当然コンピュータが必要不可欠になる. state 数, action 数が多くなると, 値反復の回数を多くすることはできない. 実際,state 数,action 数が 1 桁でも, 反復を１０回行うには膨大な時間がかかってしまう. したがって, 近似解との誤差評価や, 精度保証付き計算が重要になるのも, モデルが複雑な場合である.

以下において, state 数 3 , action 数 3 における例を示す. 前のアルゴリズムに従えば, 任意の state 数, action 数に対するモデルの最適しきい値確率を精度保証付きで実際に計算することが可能である.

state space$\{s_1,s_2,s_3\}$, action space $\{a_1,a_2,a_3\}$, 割引因子ρ=0.05 とし,

$$p_{ij}^k(w_{ij}) = P(X_{t+1}=s_j, Y_t=w_{ij}, | X_t=s_i, A_t=a_k)$$

を次のように仮定する.

$$p_{11}^1(0)=0.5, p_{12}^1(10)=0.2, p_{13}^1(20)=0.3$$
$$p_{21}^1(40)=0.4, p_{22}^1(5)=0.1, p_{23}^1(20)=0.5$$
$$p_{31}^1(10)=0.2, p_{32}^1(5)=0.3, p_{33}^1(2)=0.5$$
$$p_{11}^2(0)=0.5, p_{12}^2(10)=0.25, p_{13}^2(30)=0.25$$
$$p_{21}^2(20)=0.6, p_{22}^2(5)=0.2, p_{23}^2(10)=0.2$$
$$p_{31}^2(5)=0.5, p_{32}^2(5)=0.3, p_{33}^2(10)=0.2$$
$$p_{11}^3(5)=0.2, p_{12}^3(10)=0.3, p_{13}^3(15)=0.5$$
$$p_{21}^3(10)=0.3, p_{22}^3(0)=0.5, p_{23}^3(3)=0.2$$
$$p_{31}^3(5)=0.2, p_{32}^3(10)=0.5, p_{33}^3(2)=0.3$$

$L_0=F_0$, ただし $F_0(s,r)=0(r<0), F(s,r)=1(r\geq H/(1-\rho))$ とおいて, 前述のアルゴリズムによって L_8, L_8'を数値的に計算すると, 不連続点の数は 10857 個生じる. このとき真の解は, L_8と L_8'の間にあることが Theorem 4 より保証され, 誤差を計算すると次のようになる.

$$\|L_8-F^*\|\leq\|L_8-L_8'\|\leq 3\times 10^{-3}$$

なお, これらの計算には浮動小数点演算は倍精度を用いているため, 実際には丸め誤差が残っている. この影響をなくすには, 厳密な区間演算用のソフトウェア (e.g. PROFIL[2]) を用いる必要がある.

以下、図1に近似解 $L_8(s_1,r)$ の概形を, 図2に $L_8(s_1,r)$ と $L_8'(s_1,r)$ のグラフの一部の拡大を示す. ただし、実線部分が $L_8(s_1,r)$, 点線部分が $L_8'(s_1,r)$ を表す。

最後に、この論文をまとめるにあたり、指導をしていただいた中尾充宏教授に深く感謝いたします。

References

1. M. Bouakiz and Y. Kebir, target-level criterion in Markov decision processes, *J. Opt. Th. Appl.* **86**(1995), 1-15.

2. Knüppel,O.,PROFIL/BIAS - A fast interval liblary, Computing **53**, (1994), 277-288.

3. M. Sobel, The variance of discounted Markov decision processes, *J. Appl. Probab.* **19** (1982), 794-802.

4. J. van der Wal, Stochastic Dynamic Programming, Mathematical Centre Tracts, **139**, Mathematisch Centrum, Amsterdam, 1981.

5. D. J. White, Markov Decision Processes, John Wiley, New York, 1993.

6. D. J. White, Minimising a threshold probability in discounted Markov decision processes, *J. Math. Anal. Appl.* **173** (1993), 634-646.

7. C. Wu and Y. Lin, Minimizing risk models in Markov decision processes with policies depending on target values, *J. Math. Anal. Appl.* **231**(1999), 47-67.

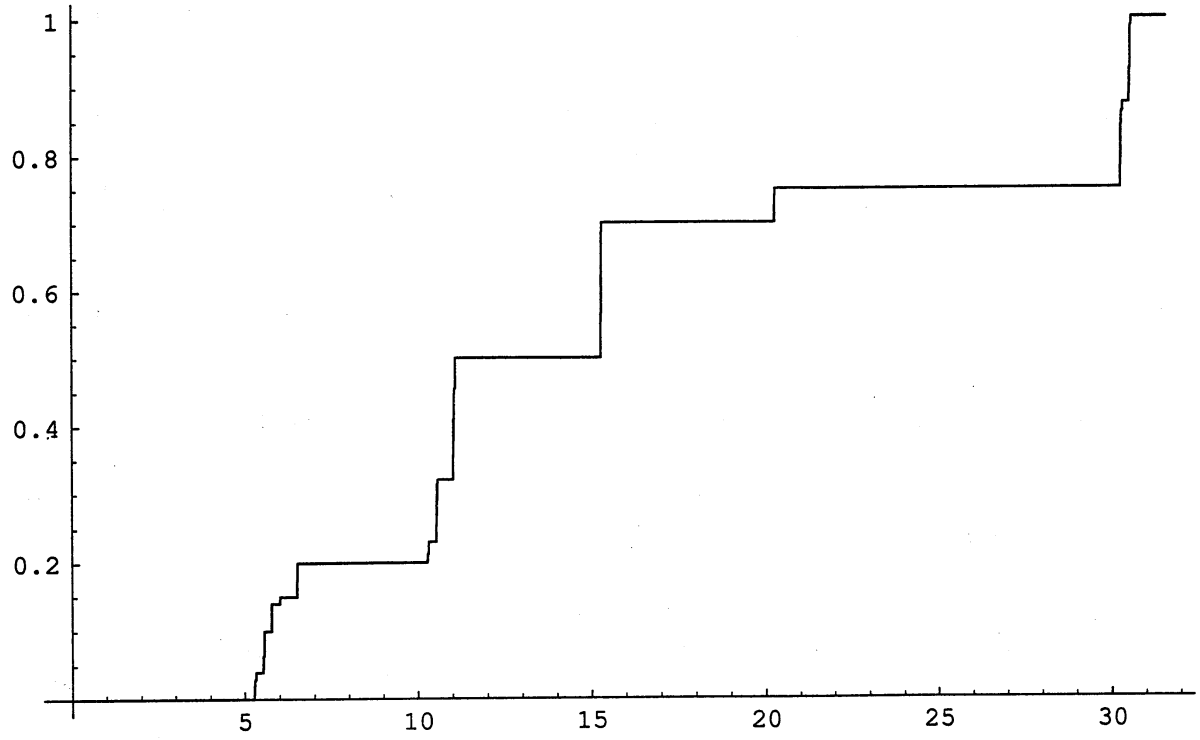

図1 近似解 $L_8(s_1, r)$ の概形

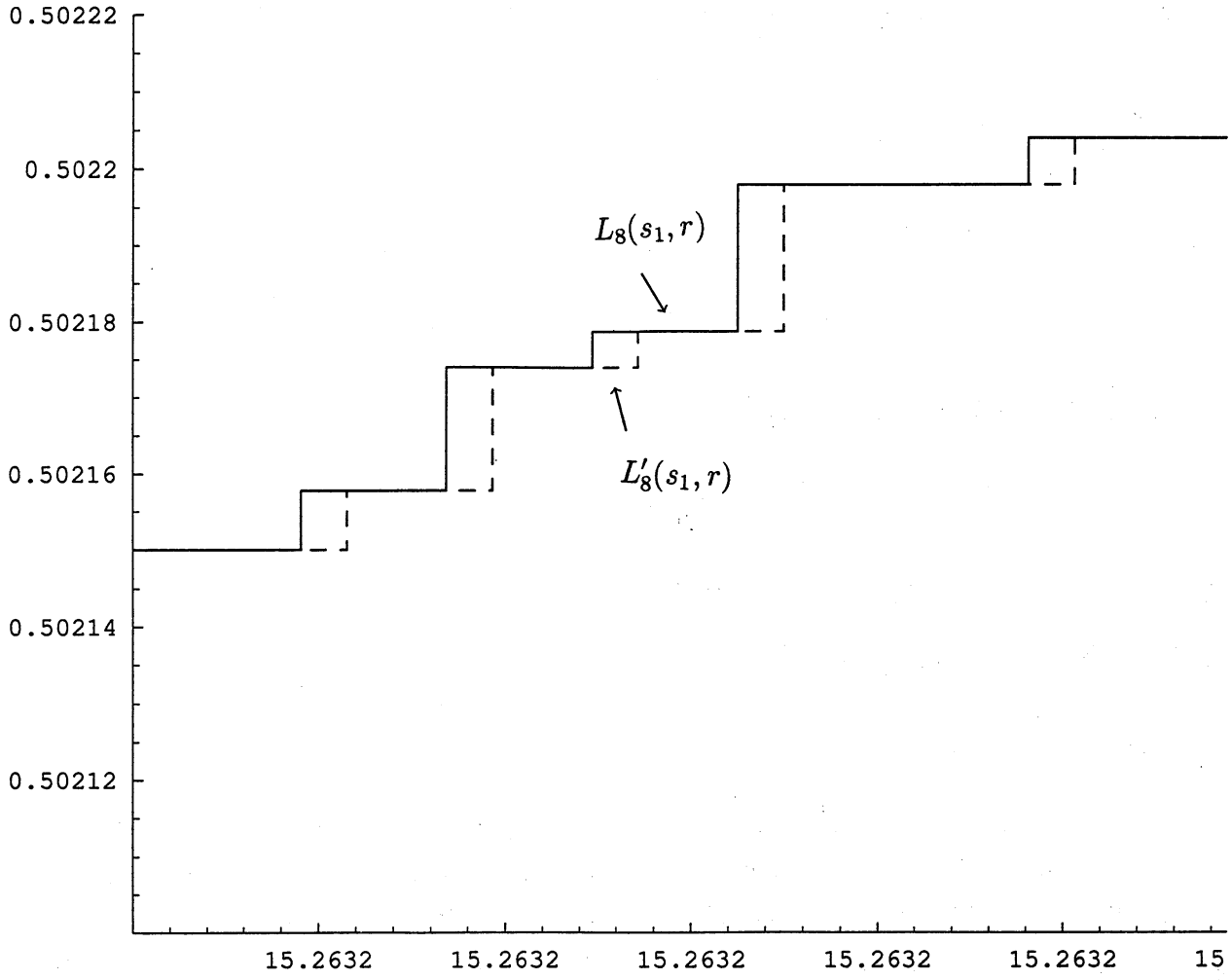

図2 最適解 $F^*(s_1, r)$ の包み込み

$(L'_8 \leq F^* \leq L_8)$