数字割り当てゲームと動的計画法

堀口　正之

本稿では，数字を割り当てるゲームを題材に，動的計画法 (Dynamic programming, DP) によって数理モデルとして定式化し最適解を導くとともに，問題のもつ不確実性に関して未知パラメータをもつ場合の DP による問題解決の方法についてみていく．

キーワード：動的計画法，多段決定モデル，sequential allocation model, bandit processes

1. はじめに

具体的に，問題を説明しよう．0 から 9 までの数字が 1 つずつ書かれているルーレットが 1 つある．これを 5 回回転させて，毎回得られる数字を使って次のルールで 5 桁の数を作る．

ルール 1：出てきた数字を順に，空いているいずれかの位（くらい）に置く．

ルール 2：一度置いた数字は，場所を変更できない．

問題：どのような順番で数字を配置したら，より大きな 5 桁の数字が作れるか．

この問題は，もともとはアメリカの子供向け数学教育番組「Square One」の 1 つのコーナーで，2 チームに分かれて，司会者がルーレットを回して毎回出てくる数字を使ってそれぞれのチームが 5 桁の数字を作り，より大きな数字を作ることを競うゲームである．Puterman 著 “Markov Decision Processes” (1994) [1] では，この問題をマルコフ決定過程の研究へと誘う身近な事例の 1 つとして，第 1 章 Introduction で紹介している．

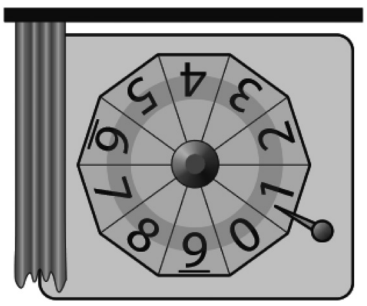

図 1　確率分布が既知のルーレット

2. 問題の定式化

数字を何桁目に置くか決めることを戦略と呼ぶことにし，どのようにすればより大きな数ができるか，具体的な例をもとに考えてみる．ルーレットを回して得られる数字は，等確率であると仮定する．すなわち，X をルーレットの数字を表す確率変数とするとき，$P(X=i)=1/10,\ i=0,1,2,\ldots,9$ であるとする（例：図 1）．簡単のため，ルーレットが 2 回だけ回るような 2 桁の数字を作る問題を考えてみる．

A 君は，1 回目に 7 以上の数字が出たら十の位にその数字を置き，それ以外の 6 以下の数字なら一の位に置く戦略を考えた．例えば，3，7 の順で出たら，この戦略でできる 2 桁の数は 73 である．また，別に，8，9 の順で出たら，できる 2 桁の数は 89 である（例：図 2）．

A 君になったつもりで，繰り返し数値実験をしてみると，もっとほかに良い戦略がありそうである．

ここでは，作られた数字（2 桁あるいはそれ以上の桁数）を確率変数とするときに，最適戦略とは，その確率変数の期待値を最大化する戦略のことをいう．

2 桁の数字を作る問題に対して，仮に 2 回目のルーレットで残りの 1 桁の置き方の状況について考えてみると，そのときの（条件付き）期待値は，残りの空い

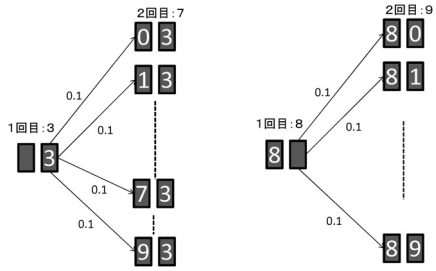

図 2　確率の木の例

ほりぐち　まさゆき
神奈川大学理学部
〒 259–1293 神奈川県平塚市土屋 2946

表 1　$N=2$ 桁の数字を作るときの第 1 回目の最適戦略

1 つ目の数字 k	1 回目のルーレットで十の位に置く	1 回目のルーレットで一の位に置く
0	4.5	45
1	14.5	46
2	24.5	47
3	34.5	48
4	44.5	49
5	54.5	50
6	64.5	51
7	74.5	52
8	84.5	53
9	94.5	54

ている桁が一の位ならば，$1/10(0+1+\cdots+9)=4.5$，十の位ならば $1/10(0+10+\cdots+90)=45$ であるから，ここで 1 つ戻って 1 回目のルーレットで出た数に対して 2 桁の数字を作ったときのそれぞれの期待値を求めると次の表のようになる（表 1）．

2 桁の数字を作る問題では，$0 \leqq k \leqq 4$ のときは初めに一の位に置く，$5 \leqq k \leqq 9$ のときは初めに十の位に置くことが最適戦略となる．表 1 のアンダーラインを引いた部分に相当する戦略が最適戦略である．

N 桁の数字を作る問題として定式化してみよう．

問題：N 桁の数字を作る戦略に対して，その戦略で作られる数字を確率変数とするとき，各戦略ごとの確率変数の期待値を最大化せよ．

多段決定モデルの構成要素として，状態空間 S，決定空間 A，利得関数 $r(k,a)$，推移法則 p_{ij} は次のようになる．状態空間 $S=\{0,1,2,\ldots,9\}$（出現する数字）．決定空間 $A=\{1,2,3,\ldots,N\}$（右から第 $a \in A$ 桁目へ数字を置く），A_n：残り n 回のルーレットを回すことが可能なときの空いている桁の場所 $(n=1,2,\ldots,N)$，ただし，$A_{i-1}=A_i-\{a_i\}$（右辺は差集合），$i=N,N-1,\ldots,2$ として A_i は定まり，$A_0=\emptyset$ であって，また，a_i は残りの i 回のルーレットを回すことが可能なときに，第 a_i 桁目に数字を置く決定を表す．利得 $r(k,a)=k\times 10^{a-1}$（数字 k が出現し，右から第 a 桁目に数字を置いたときに生じる数値）．p_{ij} を現在，数字 i が出現していて，次の回に数字 j が現れる確率とし，$\sum_j p_{ij}=1$ である．

また，k を出現した数字，B を現在空いている桁の状態を表す集合，$V(k,B)$ を k を得たときの状況 B のもとで，それ以後の最適期待利得を表す価値関数とすれば，動的計画法における最適性の原理から次のような最適方程式（Bellman 方程式）を得る：

$$V(k,A_n)=\max_{a\in A_n}\left\{r(k,a)+\sum_i p_{ki}V(i,A_n-\{a\})\right\},$$
$$k=0,1,\ldots,9,\ V(i,\emptyset)=0.$$

また，$p_i(i=0,1,\ldots,9)$，$\sum_i p_i=1$ を初期分布とするとき，すなわち，第 1 回目のルーレットで数字 k の出現する確率を p_k で表すとき，状況 A_n での最適期待利得は

$$V(A_n)=\sum_i p_i V(i,A_n)$$

と表される．ただし，$V(\emptyset)=0$ である．

具体例：

$p_{ki}=p_i=1/10$，$k,i=0,1,\ldots,9$ とする．

$N=1$ のとき：

$V(i,\{1\})=i$（1 桁の数字の問題）．

$N=2$ のとき：

$$V(i,\{2,1\})$$
$$=\max_{a\in\{2,1\}}\left\{i\times 10^{a-1}+\sum_j p_{ij}V(j,\{2,1\}-\{a\})\right\}$$
$$=\max\{10i+4.5,i+45\}.$$

最適戦略は，各 i に対して，右辺の maximizer となる位（桁の位置）を選択すればよい（先述のとおり）．

$N=3$ では，$B=\{3,2,1\}$ とするとき，

$$V(i,B)$$
$$=\max_{a\in\{3,2,1\}}\left\{i\times 10^{a-1}+\sum_j p_{ij}V(j,B-\{a\})\right\}$$
$$=\max\left\{100i+\sum_j p_{ij}V(j,\{2,1\}),\right.$$
$$10i+\sum_j p_{ij}V(j,\{3,1\}),$$
$$\left.i+\sum_j p_{ij}V(j,\{3,2\})\right\}$$

であって，$V(j,\{3,1\}),V(j,\{3,2\})$ の値は，各 j について $V(j,\{2,1\})$ での対応する maximizer となる戦略を当てはめることで得られる．

例えば，$V(j,\{3,1\})$ は表 2 の各行で大きいほうを選択する戦略が最適戦略であって，$N=3$ 桁の場合の最適戦略は，表 3 の各行で大きい値のほうを選択する．

$N=4$ 桁以上の数を作成する問題では，例えば残り 3 回の試行において空いている桁の状況が $\{4,2,1\}$ であるときの期待値の表は次のようになる（表 4）．

ちなみに，$N=5$ 桁の数字の作成における最適戦略は次のように表される（表 5）（[1], p. 15 参照）．

表 2　$N=2$ 桁の作成で $\{3,1\}$ 桁目が空いているときの期待利得

1 つ目の数字 k	1 回目のルーレットで百の位に置く	1 回目のルーレットで一の位に置く
0	4.5	450
1	104.5	451
2	204.5	452
3	304.5	453
4	404.5	454
5	504.5	455
6	604.5	456
7	704.5	457
8	804.5	458
9	904.5	459

表 3　$N=3$ 桁を作成するときの第 1 回目の最適戦略

1 つ目の数字 k	1 回目に百の位	1 回目に十の位	1 回目に一の位
0	60.75	578.25	607.5
1	160.75	588.25	608.5
2	260.75	598.25	609.5
3	360.75	608.25	610.5
4	460.75	618.25	611.5
5	560.75	628.25	612.5
6	660.75	638.25	613.5
7	760.75	648.25	614.5
8	860.75	658.25	615.5
9	960.75	668.25	616.5

表 4　$N=3$ 桁を作成で，$\{4,2,1\}$ 桁目が空いているときの期待利得

1 つ目の数字 k	1 回目に千の位	1 回目に十の位	1 回目に一の位
0	60.75	5753.25	5782.5
1	1060.75	5763.25	5783.5
2	2060.75	5773.25	5784.5
3	3060.75	5783.25	5785.5
4	4060.75	5793.25	5786.5
5	5060.75	5803.25	5787.5
6	6060.75	5813.25	5788.5
7	7060.75	5823.25	5789.5
8	8060.75	5833.25	5790.5
9	9060.75	5843.25	5791.5

表 5 の見方は，$N=5$ 桁の数字を作るとき，第 1 回目に $k=0,1,2$ の数字が出れば，一番右の位 ($\{1\}$) に置き，$k=3$ が出れば右から 2 番目の位 ($\{2\}$) に置き，$k=4,5$ が出れば右から 3 番目の位 ($\{3\}$) に置き，$k=6$ が出れば右から 4 番目の位 ($\{4\}$) に置き，$k=7,8,9$ が出れば右から 5 番目の位 ($\{5\}$) に置く．その次に，残り 4 桁分の空いている位に対して同様にルーレットを回して出てきた数字に対して，右から第何桁に置けばよいかを $N=4$ の列が示している．$N=3,2$ も同様である．また，$N=5$ の問題に対する最適戦略の期待値は 78733.8 である．

表 5　$N=5$ 桁を作成するときの最適戦略

k	$N=5$	$N=4$	$N=3$	$N=2$
0	1	1	1	1
1	1	1	1	1
2	1	1	1	1
3	2	2	1	1
4	3	2	2	1
5	3	3	2	2
6	4	3	3	2
7	5	4	3	2
8	5	4	3	2
9	5	4	3	2

表 6　確率分布が単調増加（左）と単調減少（右）の場合の最適戦略

k	$N=5$	$N=4$	$N=3$	$N=2$
0	1,1	1,1	1,1	1,1
1	1,1	1,1	1,1	1,1
2	1,2	1,2	1,1	1,1
3	1,3	1,2	1,2	1,1
4	2,3	2,3	1,2	1,2
5	3,4	2,3	2,3	1,2
6	3,5	3,4	2,3	2,2
7	4,5	3,4	3,3	2,2
8	5,5	4,4	3,3	2,2
9	5,5	4,4	3,3	2,2

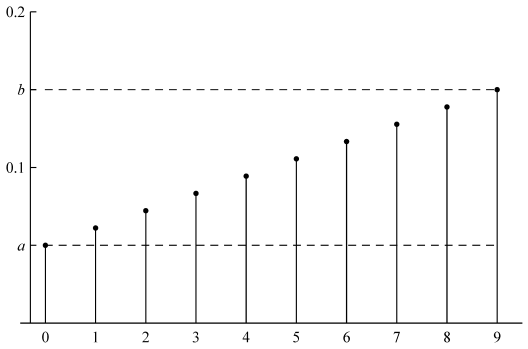

図 3　単調分布の例

次に，単調な確率分布に対して最適戦略がどのようになるか考えてみる．$p_0=a$，$p_9=b$ であって，$p_i=((9-i)a+ib)/9$，$i=0,1,\ldots,9$ であるような単調分布を考えてみる（例：図 3）．容易にわかるように，p_i が確率分布であるためには $a+b=1/5$ であればよいから，例として $a=1/20$，$b=3/20$ の場合

と，$a = 3/20$，$b = 1/20$ の場合とで，最適戦略がどのように変化しているか表 6 にまとめてみる．最適戦略の表の見方は表 5 と同様である．この表からわかることは，単調増加の例では，大きな数字の出現が予想できる分，4 や 5 などが得られた場合には，早く低い位に置いていく傾向がみられ，また，単調減少の例では，小さな数字が出現しやすいことから，4 や 5 は真ん中の位かそれ以上の高い位に置く傾向をみてとれる．単調増加の例の場合の期待値は 85826.32，単調減少の例の場合の期待値は 67187.07 であった．

3. retire のあるモデル

次に，各期での決定において，retire (stop) が含まれている場合について考えてみよう．そのときの Bellman 方程式は次のように表される．ただし M は retire を選択したときに得られる終端利得 (terminal reward) を表す．

$$V(k,B) = \max\left\{ M, k\times 10^{a-1} + \sum_i p_{ki} V(i, B-\{a\}) \right\},$$
$$a \in A, \quad k = 1,2,\ldots,9.$$

このとき，S.M. Ross (1983, p. 132) [2] と同様の議論により，各 k に対して retire するか continue するかを決める $M(k)$ が得られる．もし M が $M \geqq M(k)$ ならば，k を観測したときに retire を選択することが最適である．等確率 1/10 の場合の $M(k)$ の値を以下に示す．この表は，retire なしの問題における期待利得から導出できる（表 7）．

表 7　$M(k)$ の表

k	$M(k)$
0	74827.35
1	74828.35
2	74829.35
3	74838.585
4	74874.235
5	74974.235
6	75717.735
7	77482.735
8	87482.735
9	97482.735

4. 確率分布が未知である場合

次に，ルーレットが示す数字の確率分布が未知である場合について考えてみよう．問題を簡単にするために，2 つのルーレットを用意して，一方はすべての数字が等確率 $q = 1/10$（既知）であり，もう 1 つは，1 つのパラメータ p によって確率分布が決まるがその値が未知である場合を考える（例：図 4）．未知パラメータ p によって定まる確率分布は，第 2 節にあったような単調増加または単調減少のいずれかであると想定されて，最も大きい数字 9 の出現の得やすいほうのルーレットを選択する方法について考える．

これは，2 台の機械 X, Y がそれぞれ成功確率 p（未知）と q（既知）である two-armed bandit problem として定式化される（Bradt & Johnson & Karlin (1956)[3]，坂口実 (1970)[4]）．

X を未知の成功確率 p のルーレット，Y を既知の成功確率 q のルーレットとする．X と Y のいずれも 9 の数字が現れることを成功と呼ぶことにする．p の事前分布の分布関数を $F = F(p)$ で表し，N 期間での最適計画による期待成功回数を $W_N(F,q)$ とおくと，

$$W_N(F,q) = \begin{cases} \mu_1(F)(1 + W_N(F^s,q)) \\ \qquad + (1-\mu_1(F))W_{N-1}(F^f,q), \\ \qquad\qquad (X \text{ を選択したとき}) \\ q + W_{N-1}(F,q), \\ \qquad\qquad (Y \text{ を選択したとき}) \end{cases}$$

が成立する．ただし，$\mu_1(F)$ は，F による p の 1 次のモーメント $\int_0^1 p\,dF(p)$ であり，$F^s(F^f)$ は事前分布が F であるときに試行が成功（失敗）したときの事後分布を表す．このとき，次の定理が成り立つ．

定理 1([3])

(i) ある $\hat{p}_N(F)$ が存在して，

$$W_N(F,q) = \begin{cases} \mu_1(F)(1 + W_N(F^s,q)) \\ \qquad + (1-\mu_1(F))W_{N-1}(F^f,q), \\ \qquad\qquad (\hat{p}_N(F) \geqq q \text{ のとき}) \\ Nq + W_{N-1}(F,q), \\ \qquad\qquad (\hat{p}_N(F) < q \text{ のとき}) \end{cases}$$

(ii) 最適計画は stay with a winner の性質をもつ．すなわち，X について成功だったら再び X を選ぶ．

定理 2 ([3])

$$\hat{p}_N(F) = \max_{k} \frac{E_k(S_x)}{E_k(N_x)}$$

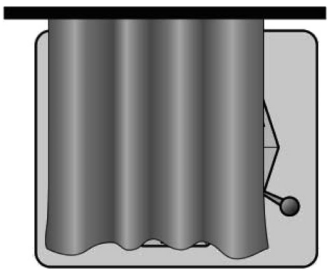

図 4　確率分布が未知のルーレット

ただし，$\boldsymbol{k}$ は，ルーレット X から試行を始める実験計画で，N 期間問題での最適計画の持つ性質を考慮したものを表し，S_x はルーレット X の成功回数，N_x はルーレット X の試行回数，$E_{\boldsymbol{k}}(\cdot)$ は $(\boldsymbol{k}, p)$ が与えられたときの条件付き期待値を表す．

計画 $\boldsymbol{k}=(r,0,\ldots,0)$ をルーレット X から始めて初めの r 回のなかで試行が失敗したならば，すぐに成功確率が既知のルーレット Y に変更し，もし，r 回のなかで失敗が一度も起こらなければ，残りの $(N-r)$ 回も全部 X を回し続けることを意味する．このとき，

$$E_{\boldsymbol{k}}(S_x)=\mu_1+\mu_2+\ldots+\mu_r+(N-r)\mu_{r+1}$$
$$E_{\boldsymbol{k}}(N_x)=1+\mu_1+\ldots+\mu_{r-1}+(N-r)\mu_r$$

を得る．ただし，μ_i は F による p の i 次のモーメント $\mu_i=\int_0^1 p^i\,dF(p)$，$i=1,2,\ldots$ である．

$N=2,3$ のときは，$\hat{p}_N(F)$ の計算は容易で，

$$\hat{p}_2(F)=\frac{\mu_1+\mu_2}{1+\mu_1},$$
$$\hat{p}_3(F)=\max\left\{\frac{\mu_1+\mu_2+\mu_3}{1+\mu_1+\mu_2},\frac{\mu_1+2\mu_2}{1+2\mu_1}\right\}$$

として得られる．

$N=3$ のとき，一様分布を事前分布とする場合は文献 [4] に示されている．その手順に沿って，同様の計算を事前分布がベータ分布 $Beta(\alpha,\beta)$ の場合について考えてみる．一般に，事前分布を $\xi(\theta)$，パラメータ θ に対する確率密度関数を $f(x|\theta)$ とすると，x を観測したときの θ の事後確率密度関数 $\xi(\theta|x)$ は

$$\xi(\theta|x)\propto\xi(\theta)f(x|\theta)$$

として得られる．今の場合，

$$f(x|\theta)=\begin{cases}\theta, & x：成功\ (=1)\\ (1-\theta), & x：失敗\ (=0)\end{cases}$$

$$\xi(\theta)=\frac{1}{B(\alpha,\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}I(\theta|(0,1))$$

であることから

$$\xi(\theta|x)\sim Beta(\alpha_1,\beta_1)$$

ただし，$\alpha_1=\alpha+x,\beta_1=\beta+1-x$ である．

これに基づいて，次のように

$$\mu_i(F)=\frac{B(\alpha+i,\beta)}{B(\alpha,\beta)},\ i=1,2,3,$$
$$\mu_i(F^f)=\frac{B(\alpha+i,\beta+1)}{B(\alpha,\beta+1)},\ i=1,2,$$
$$\mu_1(F^{sf})=\frac{B(\alpha+2,\beta+1)}{B(\alpha+1,\beta+1)},$$
$$\mu_1(F^{ff})=\frac{B(\alpha+1,\beta+2)}{B(\alpha,\beta+2)}$$

であることから，未知パラメータ p が 3/20 の値に近い形の事前分布の具体例として $\alpha=2,\beta=18$ のベータ分布を用いると，

$$\hat{p}_3(F)=\max\left(\frac{15}{143},\frac{3}{28}\right)=\frac{3}{28},$$
$$\hat{p}_2(F^f)=\frac{25}{253},$$
$$\hat{p}_1(F^{sf})=\frac{3}{22},$$
$$\hat{p}_1(F^{ff})=\frac{1}{11}$$

であって，

$$\hat{p}_1(F^{ff})<\hat{p}_2(F^f)<0.1<\hat{p}_3(F)<\hat{p}_1(F^{sf})$$

となる．したがって，1 回目に X を試行し，失敗なら Y を選び成功なら 2 回目も X を試行する．2 回目は成功でも失敗でも 3 回目に X を試行することが $N=3$ の場合の最適計画になる．

謝辞　本稿の前半の数字を割り当てるゲームについては，千葉大学名誉教授の藏野正美先生の計画数学入門に関する講義録も参考にさせていただき有益なアドバイスもいただきました．ここに感謝を表します．

参考文献

[1] M. L. Puterman, *Markov Decision Processes: Discrete Stochastic Dynamic Programming*, John Wiley & Sons Inc., 1994.
[2] S. Ross, *Introduction to Stochastic Dynamic Programming*, Academic Press, 1983.
[3] R. N. Bradt, S. M. Johnson and S. Karlin, "On sequential designs for maximizing the sum of n observations," *The Annals of Mathematical Statistics*, **27**, 1060–1074, 1956.
[4] 坂口実，経済分析と動的計画，東洋経済新報社，1970.