アルゴリズム 63－8
（1998．7．22）

シャフリングのランダムネスとそのコストの評価

岡田政則*, 岡本栄司**
* 金沢学院大学文学部, **北陸先端科学技術大学院大学情報科学研究科

実社会においては, プライバシー保護の観点からある種のランダムネスの利用の要求がある. この場合, 完全にランダムな状態でなくても低コストで十分プライバシーが守られる状態が生成できれば都合が良い. 本論文では, 最初にシャフリングを定義し, シャフリング近似の概念を導入する. そのランダムネスの尺度としては相対エントロピーを利用し, ランダムネスのコストとしてシャフリングを生成する回路の素子の個数を利用する. このためにまずその個数にコストを表す関数としての妥当性があることを検証する. 例として複数の投票所からなる投票集計のモデルを取り上げる. そこでは選挙とその投票結果のプライバシー保護をランダムネスとその生成コストの関係として考察している.

Randomness of Shuffling and Evaluation of It's Cost

Masanori Okada* and Eiji Okamoto**
*Kanazawa Gakuin University, **Japan Advanced Institute of Science and Technology

Randomness is used for a protection of privacy in the actual world. To investigate randomness, we show the definition of the shuffling S_n of n input data; and, we define the pseudo-shuffling $\hat{S}_n$ by S_2 only. The definition of randomness is represented by relative entropy, and the cost of randomness is expressed by the number of shuffling S_2 in pseudo-shuffling $\hat{S}_n$. We consider pseudo-shuffling as an approximation of shuffling with some constraints. Finally, we show a problem related to ballot boxes in the election as an example.

1 はじめに

アンケートや投票の集計を行う場合, 投票結果に対してこれら参加者のプライバシーが保護されねばならない. その保護の実現のため, 我々の社会ではシャフリング [2, 5] をしてランダムな状態を作り出すことが要求される. しかし実際の生活では必ずしも完全なシャフリングをせずにすましている.

例えば選挙投票後の開票作業において, 全ての投票をひとつの箱に入れてかき混ぜて完全なシャフリングしてから開票作業に移ることはない. 実際投票用紙が各地区の集会所や公民館毎に投票箱に集められ, その箱から直接開票作業が進められてもそれを見守る有権者は不愉快だとは感じないだろう. その理由はこのような開票処理ではその結果が変わらないことは勿論, 有権者にとっては十分プライバシーが保護されていると実感できるからである.

このように実社会においては, プライバシー保護の観点からある種のランダムネス利用の要求がある. この場合完全なシャフリングでなくとも低コストでしかも真のシャフリングに近いものが生成できると都合が良い. そこで, 本論文では, シャフリング近似の概念を導入し, ランダムネスの尺度として相対エントロピーを利用する. 例として選挙とその投票結果のプライバシー保護を取り上げる. 選挙とエントロピーに関する他の研究としては, 合田 [1] が地区割りが比較的うまくいっている政党の候補者に限っては, 相対エントロピー関数が地盤の強さの指標になりうることを示している.

2 シャフリングとは

2.1 シャフリングの定義

定義 1 (シャフリング) n 個のデータ $1,2,3,\cdots,n$ を入力した時, その任意の順列が等確率で出現するなら, その入力データはシャフルされた言う. その機能を実現する回路を S_n と表す.

定義 2 (回路 T_n の定義) 回路 T_n は, n 本の入力 $1,2,3,4,5,\cdots,n$ に対して $12345\cdots n, 21345\cdots n, 23145\cdots n, 2345\cdots n1$ の n 種類の順列が等確率に出力する回路である.

補題 1 (シャフリングの帰納的構成) 回路 S_n の存在を仮定する. 次のように S_n, T_{n+1} より S_{n+1} を構成することができる. 回路 S_n における n 個の出力の左に $n+1$ 番目のデータをおく. 合わせて $n+1$ 個のデータを回路 T_{n+1} の入力とする (Fig. 1). この回路 S_{n+1} は, 元の $n+1$ 個のデータをシャフリングしている.

2.2 S_2 を利用したシャフリング近似

シャフリング近似の方法は, 多数考え得る. ここでは, S_2 のみを組み合わせた構成で議論をすすめる. まず有限個の S_2 だけを利用して真のシャフリング S_n が構成できないことを示す. ここで Fig. 2-(1),(2) のような S_2 の接続は, 単独の S_2 と出力が同じであることに注意されたい.

定義 3 ふたつの回路 S_i の接続においてその出力が 1 つの S_i の出力と同じでない時有意な接続であると言う.

補題 2 シャフル S_n は, $n!$ 種類の順列を生成する.

proof. 定義 1 から明らかである.

定理 1 $n \neq 2$ ならば, S_n は有限個の S_2 だけを用いて構成できない. 但し全ての回路 S_2 は互いに有意な接続により結ばれているとする.

proof. n 個の入力データに対して同数の出力があり, その間には k 個の回路 S_2 のみから構成されるシャフルの機能を有する回路 S_n があるとする.

仮定により出力側に S_2 が存在しそのふたつの出力は回路全体の出力ラインの一部になっている. この S_2 は, 有意な接続がなされているはずだから生成する順列の種類を 2 倍にしている. 故に S_n からひとつの S_2 を取り外した回路を S_n^1, (Fig. 3), 引続き k 個取り外した回路を S_n^k として $g(S)$ を回路 S が生成する順列の種類の個数と考えると

$$g(S_n) = g(S_2)g(S_n^1) = 2g(S_n^1) = 2^2 g(S_n^2)$$

となる. 従って

$$g(S_n) = 2^k g(S_n^k)$$

である. また $g(S_k^n)$ は, S_n から全ての S_2 を取り去った状態での生成する順列の数を表すから 1 である.

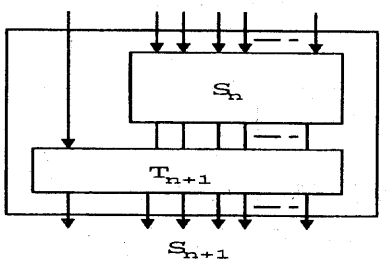

Fig. 1. The inductive construction of S_n

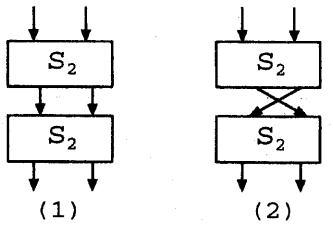

Fig. 2. The example of trivial connection

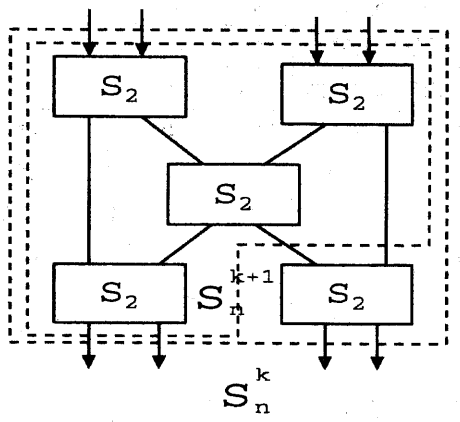

Fig. 3. S_n^k and S_n^{k+1}

補題 2により S_nは, $n!$種類の順列を生成する. しかし $g(S_n) = 2^k$ 個の順列が, $n!$種類の順列を等確率に出力することはできない. 従って有限個の S_2から構成される回路では S_nを表現できない. (証明終わり)

次に回路 S_2のみを利用して簡単な回路を作成し, それはどのような順列をどれくらいの頻度で出力できるかを観察する. ここでは入力アルファベットを $1, 2, \cdots, n$ とし, その個数 n が偶数か奇数で場合分けする.

- 最初 n が偶数の時, その例として Fig. 4(a) のように$\hat{S}_6$ を構成してみる. 最初に 3 つの S_2により, 互いに隣り合っている文字をシャフルする. 出力は $2^3 = 8$ 通りの相異なる文字列となる. このプロセスを Stage 1 と呼ぶ.

  次に Stage 1 の 8 個の出力を Stage 2 の入力とする. Fig. 4(a) のシャフリング近似$\hat{S}_6$は 4 Stage あるから, それを$\hat{S}_6^4$と表す. $\hat{S}_6^4$では, S_2が 12 個利用されている. 入力 123456 に対して, $2^{12} = 4096$ 個の出力の場合がある. その分布は Table 1 から分かるように, 136425 等 3 回ずつ現れるものが 128 個, 123546 等 4 回ずつ現れるものが 288 個, 以下 6 回ずつが 192 個, 8 回ずつが 32 個, 以下 18 回ずつが 64 個である. シャフリング S_6の定義により, 完全なシャフリングが行われるなら $6! = 720$ 種類の順列が等確率に出現する. $\hat{S}_6^4$では, 704 種の順列が前述した分布に従って出現する.

- 次に n が奇数の場合, $\hat{S}_5$を考える. Fig. 4(b) のように 4 Stages のシャリング近似$\hat{S}_5^4$において, 入力文字列の 5 文字のうち 1 つは Stage 1 を素通りせねばならない. p を各 Stage を素通りさせる入力番号とすると, その素通りする 1 本は左から数

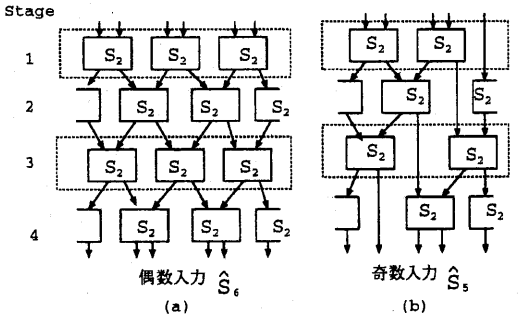

Fig. 4. Example of $\hat{S}_6, \hat{S}_5$

えて1～5番目の5通りの選択肢がある. Fig. 4(b) では, Stage p では $6-p$ 番目の入力を素通りさせている. その結果が Table 2である.

なお, Stage p までのシャフリング近似$\hat{S}_n^p$を総称して, シャフリング近似$\hat{S}_n$ と呼ぶ.

3 シャフリング近似のランダムネス

3.1 エントロピーによるランダムネスの定義

シャフルの回路S_nを$n!$種類の順列を等確率に出力する情報源と見なすことができる. 例えばシャフルS_3の出力記号は6種あり $\{123, 132, 213, 231, 312, 321\}$ となる. 一般には情報源 Sの出力記号はそれ以前に出力された記号系列の影響を受けて生起する. シャフリングやシャフリング近似では以前に出力された順列とは無関係, つまりシャフリングの出力は無記憶情報源であるとしてよい.

無記憶情報源Sのアルファベット $\{a_1, a_2, \cdots, a_M\}$ に属する任意の記号 a_iの生起確率を P_iとする. Shannon により無記憶情報源から出力さ

Table 1. Output of $\hat{S}_6$

Stage	2	3	4	5
S_2	6	9	12	15
順列の種類	64	360	704	720
未生成の順列	656	360	16	0
分布	1*64	1*256	3*128	24*16
		2*96	4*288	32*24
		8*8	6*192	33*128
			8*32	34*192
			28*64	50*64
				52*192
				68*96
				144*8
順列の数	64	512	4096	32768

Table 2. Output of $\hat{S}^5$

Stage	2	3	4	5
S_2	4	6	8	10
順列の種類	16	56	112	120
未生成の順列	104	64	8	0
分布	1*16	1*48	1*16	4*8
		2*8	2*80	6*8
			4*8	7*32
			6*8	8*24
				10*16
				11*16
				12*16
順列の数	16	64	256	1024
p	4	3	2	1

れる平均の情報量は,

$$H(S) = -\sum P_i log_2 P_i (bits)$$

となる [4]. シャフリング近似$\hat{S}_n$についてそれがどの程度真のシャフルに近いかを比べる尺度として, M元情報源Sに対する相対エントロピー

$$h(S) = \frac{H(S)}{log_2 M}$$

がある. これは

$$0 \leq h(S) \leq 1$$

を満たしている. そうすると真のシャフルに対する相対エントロピーは1となり, シャフル近似において, 等確率な出力がされない場合は, 0 に近くなる. 以下, ランダムネスの尺度としてシャフリング近似の相対エントロピーを採用する.

3.2 シャフリング近似の相対エントロピー

シャフル近似$\hat{S}_n$のStage数を変化させて, $S_n(n = 3,4,5,\cdots,12)$ に対する相対エントロピーがどのように変化するか観察する. nが奇数の時各Stageにおいて, 入力データの素通りする位置pを選択せねばならない. 今回はStage iにおいて$p = n+1-i$とした.

S_nに対する相対エントロピーを求めた結果を Table 3 に示す. Table 3においてデータが空白の欄は, メモリ不足が原因で求められなかった.

- 実験結果から$n \geq 5$の時$\hat{S}_n$のコストは次のようになる.
  - nが偶数なら$n-1$ Stageで相対エントロピーは十分1に近いと考えられる. 回路$\hat{S}_n$のS_2の個数は, $n/2$個のS_2を$n-1$ Stageつなげるので, $n(n-1)/2$となる.
  - nが奇数ならn Stageで相対エントロピーは十分1に近いと考えられる. 回路$\hat{S}_n$のS_2の個数は, $(n-1)/2$個のS_2をn Stageつなげるので, $n(n-1)/2$となる.

Table 3. The relative entropy of $\hat{S}_n$(n=3-12)

Stages	3	4	5	6	7
$\hat{S}_3$	0.97	0.99	0.99	0.99	0.99
$\hat{S}_4$	0.98	0.99	0.99	0.99	0.99
$\hat{S}_5$	0.83	0.96	0.99	0.99	0.99
$\hat{S}_6$	0.95	0.97	0.99	0.99	0.99
$\hat{S}_7$	0.69	0.85	0.94	0.97	0.99
$\hat{S}_8$	0.73	0.86	0.93	0.97	
$\hat{S}_9$	0.61	0.74	0.84		
$\hat{S}_{10}$	0.62	0.75			
$\hat{S}_{11}$	0.55	0.59			
$\hat{S}_{12}$	0.57				

つまり回路S_2を$n(n-1)/2$個利用して, シャフリング近似$\hat{S}_n$を構成すると, その相対エントロピー$h(\hat{S}_n)$は十分1に近くなりランダムな順列を生成できると考え得る.

4 シャフリングのコスト

シャフルの帰納的な構成によりシャフリングS_nは, S_{n-1}と回路T_nにより構成できる. シャフリングS_nのコストをa_n, 回路T_nのコストをb_nと書く. コストとは各回路を構成する際に必要な回路S_2の個数とする. つまり

$$a_n = a_{n-1} + b_n \quad (4.1)$$

但し$a_2 = 1$. また回路T_nは, n人から1人を等確率に選択する処理を行う操作と同値である. $n = 2^k$の時は, 次の方法により1人を公平に選択することができる.

1. n 人が2人ずつ組になってシャフル S_2 により勝者を決める．必要なシャフル S_2 の個数は 2^{k-1} 個である．

2. 次に残った 2^{k-1} 人でまた2人ずつ組になり勝者 2^{k-2} 人を決める．

3. 従って $2^{k-1}+2^{k-2}+\cdots+2+1=2^k-1$ 個のシャフル S_2 を行えば公平に $n=2^k$ 人から1人を選択することができる．

よって回路 $T_n(n=2^k)$ のコストは，2^k-1 となる．今仮に式 (4. 1) において n の値が 2^k の形でないときにも T_n のコスト b_n を $n-1$ とする．漸化式

$$a_n = a_{n-1}+n-1, a_2=1$$

を解くと，$a_n = n(n-1)/2, (n \geq 2)$ となる．従ってシャフル S_n のコストは $n(n-1)/2$ である．

従って前節の結果より自然数 n に対してシャフル S_n のコストを $n(n-1)/2$ と定義する．

5 分割による簡易シャフリング

例としてある都市における市長選挙の投票数 n と投票所の数 k の関係を取り上げる．

5.1 標本の分割とエントロピー

k 個の無記憶情報源 $V_j=\{a_{1j}, a_{2j}, \cdots, a_{n_j j}\}(j=1,2,\cdots,k)$ がある．但し $\sum_i P(a_{ij})=1$, $S=\cup_j V_j$ とする．V_j の各元 a_{ij} は，どの無記憶情報源に属しているか一意に判別できるとする．S の元 a_{ij} が出現する確率は，S から k 個の無記憶情報源 V_j のうち1つを選択し，そこから元 a_{ij} が出現すれば良いから $P(V_j)P(a_{ij})$ となる．よって S のエントロピーは

$$\begin{aligned}
H(S) &= -\sum_j\sum_i P(V_j)P(a_{ij})log_2P(V_j)P(a_{ij}) \\
&= -\sum_j\sum_i P(V_j)P(a_{ij}) \\
&\quad (log_2P(V_j)+log_2P(a_{ij})) \\
&= -\sum_j\sum_i P(V_j)P(a_{ij})log_2P(V_j) - \\
&\quad \sum_j\sum_i P(V_j)P(a_{ij})log_2P(a_{ij}) \\
&= -\sum_j P(V_j)log_2P(V_j)+\sum_j P(V_j)H(V_j)
\end{aligned}$$

となる．これは S のエントロピーは k 個の無記憶情報源 V_j の生起確率 $P(V_j)$ とエントロピー $H(V_j)$ から求めうることを表わしている．

標本の k 分割

集合 T の要素の個数が $mk = n(m,k \in N)$ であるとする．T の部分集合 $T_i = \{x_{pi}|p = 1,2,\cdots,m\}$ を 次のように定義する．

$$|T_i|=m, i\neq j \Rightarrow T_i\cap T_j=\emptyset, \cup T_i = T$$

また集合 $V_i(i=i,2,\cdots,k)$ の各元を次のように定義する．

$\forall x \in V_i, x = x_{1i}x_{2i}\cdots x_{mi}$ の並び替え $(x_{pi}\in T_i)$

つまり $|V_i|=m!$ で，$H(V_i)=log_2m!$ となる．これらの集合 V_i の各元は m 個の要素をシャフリングした時に生成される順列と考えることができる．$S=\cup V_i$ とすると，S は，$n=mk$ 個の元を k 等分した $V_i(m!$ 個の順列の集合) の和集合とみなせる．従ってエントロピー $H(S)$ を次のように計算することができて，

$$\begin{aligned}
H(S) &= -\sum_j P(V_j)log_2P(V_j) \\
&\quad +\sum_j P(V_j)H(V_j) \\
&= -\{P(V_1)log_2P(V_1)+P(V_2)log_2P(V_2)+ \\
&\quad \cdots+P(V_k)log_2P(V_k)\} \\
&\quad +P(V_1)H(V_1)+P(V_2)H(V_2)+ \\
&\quad \cdots+P(V_k)H(V_k)\} \\
&= -k\frac{1}{k}log_2\frac{1}{k}+k\frac{1}{k}H(V_i)
\end{aligned}$$

$$= log_2 k + log_2 m!$$
$$= log_2 km!$$
$$= log_2 k(\frac{n}{k})!$$

となる. ゆえに

$$h(S) = \frac{H(S)}{log_2 n!} = \frac{log_2 k(n/k)!}{log_2 n!}$$

$H(S)$ は, n 人の有権者が住む都市で k箇所の投票所を設け, その投票所に n/k人が投票すると仮定した時の票の散らばり具合をあらわしている.

5.2 分割シャフリングのコスト

$n(= mk)$ 人を k等分して, それぞれ m 個の元をシャッフルする. そのコスト $f_n(k)$ は, $m(= n/k)$ の元のシャッフルを k回, そして k個の元を含む集合のシャッフルが1回あるので

$$f_n(k) = \frac{k(k-1)}{2} + \frac{k}{2}(\frac{n}{k})(\frac{n}{k} - 1)$$

となる. 本節では n を固定し kを変化させた時,

- なるべく相対エントロピー $h(S)$ を大きくし,
- コスト $f_n(k)$ を小さくする

kを求めたい. $n = 100000$ とした時, $f_n(k), H(S)$ のグラフはFig. 5, 6の通りである. 両グラフとも横軸は分割数 k である. 縦軸はそれぞれ S_2 の個数 $f_n(k)$ とエントロピー $H(S)$ である.

コスト関数 $f_n(k)$ は下に凸で最小値を一つだけ持つ関数であり, 相対エントロピー $h(S)$ つまりエントロピー関数 $H(S)$ は単調減少の関数である. 分割数 kをパラメーターとして点 $((h(S), f_{10000}(k))$ をプロット (Fig. 7) して最適な分割数 kを求める. Fig. 7よりコスト関数 $f_n(k)$ が最小で, エントロピー $H(S)$ が最大と

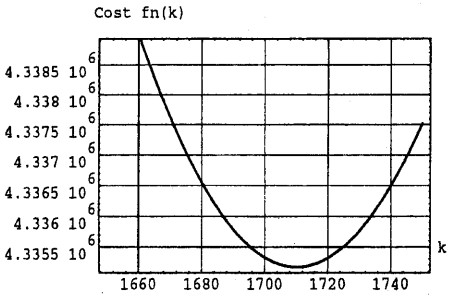

Fig. 5. The cost $f_n(k)$ such that $n = 100000$ elements are divided.

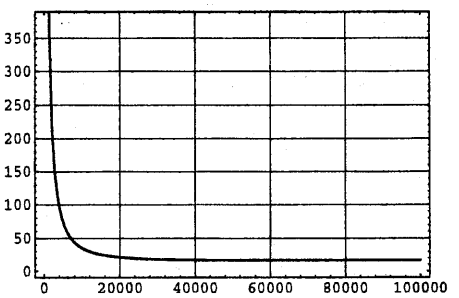

Fig. 6. The entropy such that $n = 100000$ elements are divided.

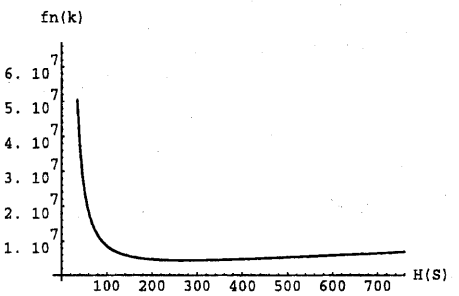

Fig. 7. The graph $(h(S), f_{10000}(k))$ such that $n = 100000$ elements are divided.

なる kを見つけることができる. いくつかの n について最適な kの値を求めると, Table. 4になる.

Table 4. The optimal entropy of cost $f_n(k)$

n	k	相対エントロピー
1000	79.53(80)	0.0043372
10000	368.57(369)	0.0008622
100000	1710.14(1710)	0.0001805
200000	2714.58(2714)	0.0001133

n	コスト	人数/投票箱
1000	8910	12.5
10000	198397	27.1
100000	4335170	58.5
200000	10950700	73.7

n=100000 で最適な分割数が1710 となった. 一つの投票箱当たりの人数は 59(58.5) 人となる. これは直観的に少なすぎると感じられるかも知れない. このような値が出た理由は, コスト関数 $f_n(k)$ の第一項と第二項の重みのかけ方による. 第一項 $k(k-1)/2$ は, 投票箱のシャッフルであり, 第二項は k個の投票箱内でのシャッフルである. 今回の実験ではその比を 1 : 1 とした. この比が適切かどうかは, 今後の課題としたい.

6 まとめ

ランダムネスの尺度として相対エントロピーを採用することにより回路 S_2のみから構成されるシャフリング近似のランダムネスを評価することができた. また, シャフリング近似の相対エントロピーをもとに回路 S_2の個数に基づくシャッフリングのコストを定義できた. その結果 $n = 2^k$の場合, シャフリング S_nの構成時に現れる回路 T_nのコストがうまく評価できた. さらに $n \neq 2^k$の場合も T_nのコストを同様に仮定することにより, $n \geq 5$ の時 S_nのコストを定義することができた.

シャフリングのコストを適用する具体例として, 複数の投票所からなる投票集計をモデルとし, コストが小さく, 大きなランダムネスを得るための投票所の数を求めることができる.

References

[1] 合田周平, 選挙のエントロピー, 別冊・数理科学, 4(1984), 34-41.

[2] J.D. ビースリー, 中村義作訳, ゲームと競技の数学, サイエンス社, 東京, 1992.

[3] Dudewicz, E.J., van der Meulen, E.C., SrimRam, M.G., Teoh, N.K.W., Entropy-based random number evaluation, American Journal of Mathematical and Management Sciences, 15(1995), 115-153.

[4] 今井秀樹, 情報理論, 昭晃堂, 東京, 1984.

[5] 中村義作, 遊びの確率論, 海鳴社, 東京, 1982.

[6] Todorov, N.S., On a general interpretation of equilibrium entropy as a measure of randomness, Annales de la Fondation Louis de Broglie, 20(1995), 169-180.

[7] Wigderson, A., Computational pseudo-randomness, Proceedings Third Israel Symposium on the Theory of Computing and Systems, (1995),218-219.