情報理論における情報源符号化の統計力学的解釈

中央大学研究開発機構　只木孝太郎[*1]

1　はじめに

情報源符号化は、通信路符号化と並ぶ、情報理論の主要テーマである。例えば、コンピュータでテキストファイルを圧縮する場合など、0, 1 記号を使って情報をできる限り少ない容量で記憶する問題が、情報源符号化の問題である。本稿では、情報源符号化で最も基本的な設定である、瞬時符号による無歪み情報源符号化の枠組に対して、平衡統計力学的解釈を与える。

一般に、統計力学とは外見上隔たりのある現象や枠組に対して、統計力学的解釈を施すためには、そこにミクロカノニカルアンサンブルを同定することが重要である。ミクロカノニカルアンサンブルさえ同定できれば、後は、通常の統計力学の教科書で行われている平衡統計力学の理論の理論展開に従うことにより、ほぼ自動的に、その枠組上に統計力学的解釈を展開することが可能となる。

本稿では、最適な瞬時符号による無歪み情報源符号化の枠組の中に、ミクロカノニカルアンサンブルを同定する。そして、それを足掛かりとして、情報源符号化の枠組の上に平衡統計力学を展開する。これにより、(統計力学的) エントロピー、温度、熱平衡など、平衡統計力学における諸概念が情報源符号化の枠組に導入される。その上で、我々はこれら諸概念の情報理論的な意味について明らかにする。

我々の統計力学的解釈では、瞬時符号の符号語系列と、統計力学で取り扱われる大自由度の量子系のエネルギー固有状態とを同一視する。そして、その符号語系列の長さと、このエネルギー固有状態のエネルギーとを同一視する。符号語系列の長さの離散性は、量子系におけるエネルギー固有値の離散性に自然に対応する。ところで、統計力学でよく知られているように、量子系のエネルギースペクトルに有限の最大値がある場合、そのような系は負の温度を持ち得る。どんな瞬時符号でも符号語は有限個しかないので、無歪み情報源符号化に対する我々の統計力学的解釈でも、負の温度が現れる。

特に我々は、(統計力学的) エントロピーの概念に基づいて、この統計力学的解釈

[*1] E-mail: tadaki@kc.chuo-u.ac.jp　　WWW: http://www2.odn.ne.jp/tadaki/

では、温度 1 は瞬時符号の平均符号長に対応することを明らかにする。この事実は、温度一定の符号語系列の集合に対するフラクタル次元（ボックス次元）を用いた解析によっても更に明らかとなり、温度 1 がティピカル列に対応していることが確認される。このように、我々の統計力学的解釈において、温度 1 は特別な役割を果たす。

更に、我々は、熱平衡概念に基づいて、この統計力学解釈の情報理論的な応用について考察する。

なお、本稿では、議論の数学的厳密さには拘らない。我々の目標は、情報源符号化の枠組の中に隠された統計力学的な構造を明らかにすることなので、統計力学に倣い、本稿での議論も統計力学と同程度の数学的厳密さで進める。

2 無歪み情報源符号化と瞬時符号

本節では、以下で必要となる数学的概念、並びに情報源符号化の基本概念、特に瞬時符号の定義と性質について簡単に復習する（詳しくは [11, 2, 8, 4] などを参照されたい）。

任意の集合 S に対して、S に属する元の個数を $\#S$ で表す。有限二進列（即ち、$0,1$ 記号からなる有限列）全体の集合を $\{0,1\}^*$ で表す。任意の $s \in \{0,1\}^*$ に対して、s の長さを $|s|$ で表す。$\{0,1\}^*$ の部分集合 S が **prefix-free** であるとは、S に属するどのような 2 つの異なる有限二進列についても、一方が他方の接頭語になることはない、ということである。χ が空でない有限集合のとき、χ を**アルファベット**と呼ぶ。

X を、アルファベット χ に値をとり、確率分布 $p_X(x) = \Pr\{X = x\}$ $(x \in \chi)$ を持つ任意の確率変数とする。このとき、X の**エントロピー** $H(X)$ は次式で定義される。

$$H(X) = -\sum_{x \in \chi} p_X(x) \log p_X(x).$$

ここで、対数の底は 2 である。我々は、後程、統計力学的エントロピーの概念を導入するが、これは上記 $H(X)$ とは（少なくとも、定義に関しては）異なるものであることを注意されたい。確率変数 X に対する**瞬時符号**とは、χ から $\{0,1\}^*$ への単射 C で、その像 $C(\chi) = \{C(x) \mid x \in \chi\}$ が prefix-free であるものをいう。このとき、各 $x \in \chi$ に対し、有限二進列 $C(x)$ は、x に対応する**符号語**と呼ばれる。なお、以下で $|C(x)|$ は $l(x)$ と表す。各 i について $x_i \in \chi$ となる系列 $x_1, x_2, \ldots, x_N$ は、**情報源系列**と呼ばれる。一方、有限二進列 $C(x_1)C(x_2)\cdots C(x_N)$ は、情報源系列

$x_1, x_2, \ldots, x_N$ に対応する**符号語系列**と呼ばれる。

瞬時符号は、無歪み情報源符号化の問題において、基本的かつ重要な役割を果たす。さて、$X_1, X_2, \ldots, X_N$ を、確率分布 $p_X(x)$ に従う定常独立な確率変数列（即ち、i.i.d. 情報源）とする。瞬時符号による無歪み情報源符号化問題の目的は、N を非常に大きな数とした場合に、確率変数列 $\{X_i\}$ によって生成される情報源系列 $x_1, x_2, \ldots, x_N$ に対応する符号語系列 $C(x_1)C(x_2)\cdots C(x_N)$ の長さを、最小にすることである。この目的のためには、N とは無関係に、確率変数 X に対する瞬時符号 C の**平均符号長**

$$L_X(C) = \sum_{x \in \chi} l(x) p_X(x)$$

を考察することで十分である。この平均符号長は、X のエントロピーとの間で次の重要な関係を満たす。即ち、X に対する任意の瞬時符号 C について、

$$L_X(C) \geq H(X) \tag{1}$$

が成り立つのである。従って、エントロピー $H(X)$ は、瞬時符号による無歪み情報源符号化問題において、圧縮限界を与える。この圧縮限界を達成する瞬時符号は、情報源符号化の目的からして特に重要な存在であり、そのような瞬時符号、即ち (1) で等号が成り立つ瞬時符号 C は、確率変数 X に対して**最適**であると言われる。このとき、次の定理が成り立つ。

定理 1. 確率変数 X に対する瞬時符号 C が最適であることと、任意の $x \in \chi$ に対して $p_X(x) = 2^{-l(x)}$ が成り立つことは同値である。 □

最後に、以下では各 $x^N = (x_1, x_2, \ldots, x_N) \in \chi^N$ について、

$$p_X(x_1)p_X(x_2)\cdots p_X(x_N)$$

を $p_X(x^N)$ で表す。

3 統計力学の基本的枠組の復習

無歪み情報源符号化に対して統計力学的解釈を与えるにあたり、本節では、まず統計力学の基本的枠組について復習する。

一般に、統計力学 [9, 15, 10] では、極めて多数の同一の部分系から構成される量子系 $\mathcal{S}_{\mathrm{total}}$ を考察する。N をそのような部分系の個数としよう。例えば、$1\,\mathrm{cm}^3$ の常

温の気体では $N \sim 10^{22}$ である。ここで我々は、これら極めて多数の部分系は（原理的には）互いに区別できるものと仮定する。即ち、我々は、Bose-Einstein 統計でも Fermi-Dirac 統計でもなく、Maxwell-Boltzmann 統計に基づいて統計力学を展開する。この仮定の下では、各 $i = 1, \ldots, N$ に対して、i 番目の量子部分系 $\mathcal{S}_i$ を識別することができる。さて、一般に量子力学では、系の状態は、量子状態によって完全に記述される。そのような量子状態の中でも、特に統計力学で重要となるのは、**エネルギー固有状態**である。各部分系 $\mathcal{S}_i$ のエネルギーの固有状態は、**量子数**と呼ばれる正整数 $n = 1, 2, 3, \ldots$ によって指定される。そして、量子数 n によって指定されるエネルギー固有状態にある部分系は、確定したエネルギー E_n を持つ。このとき、全系 $\mathcal{S}_{\text{total}}$ のエネルギー固有状態は、N 対の量子数 $(n_1, n_2, \ldots, n_N)$ によって指定される。もし全系 $\mathcal{S}_{\text{total}}$ が、$(n_1, n_2, \ldots, n_N)$ で指定されるエネルギー固有状態にあるならば、各部分系 $\mathcal{S}_i$ は n_i で指定されるエネルギー固有状態にあり、全系 $\mathcal{S}_{\text{total}}$ 自身は、確定したエネルギー $E_{n_1} + E_{n_2} + \cdots + E_{n_N}$ を持つ。

以上の設定の下、統計力学の基本仮定である **The Principle of Equal Probability（等重率の原理）**は、次のように与えられる。

The Principle of Equal Probability: 系 $\mathcal{S}_{\text{total}}$ のエネルギーが、E から $E + \delta E$ の間にあることがわかっているものとする。ここで、δE は、系 $\mathcal{S}_{\text{total}}$ のエネルギーの測定における不確定の幅である。このとき、系 $\mathcal{S}_{\text{total}}$ は、$E \leq E_{n_1} + E_{n_2} + \cdots + E_{n_N} \leq E + \delta E$ を満たす $(n_1, n_2, \ldots, n_N)$ で指定される各エネルギー固有状態に、同様の確からしさで実現される。 □

$\Omega(E, N)$ を、$E \leq E_{n_1} + E_{n_2} + \cdots + E_{n_N} \leq E + \delta E$ を満たす、N 対の量子数 $(n_1, n_2, \ldots, n_N)$ の総数とする。このとき、The Principle of Equal Probability の主張は「系 $\mathcal{S}_{\text{total}}$ のエネルギー固有状態で、そのエネルギーが E と $E + \delta E$ の間にあるものは、どれも同じ確率 $1/\Omega(E, N)$ で実現される」というものである。これは、そのエネルギーが E と $E + \delta E$ との間にあるエネルギー固有状態の一様分布を、統計力学の議論の出発点にしようとするものであり、この一様分布は**ミクロカノニカルアンサンブル**と呼ばれる。

そして統計力学では、$\Omega(E, N)$ に基づいて、全系 $\mathcal{S}_{\text{total}}$ の**エントロピー** $S(E, N)$ を次式で定義する。

$$S(E, N) = k \ln \Omega(E, N). \tag{2}$$

ここで、k は **Boltzmann 定数**と呼ばれる定数であり、また、ln は自然対数を表す。

部分系 1 個当りの平均エネルギー ε は、E/N によって与えられる。一般に、ε が有限値となる通常の場合、エントロピー $S(E,N)$ は N に比例する。一方、エネルギーの不確定の幅 δE は項 $k\ln\delta E$ を通じて $S(E,N)$ に寄与するが、δE が小さ過ぎない限り、N に比べてこの項は無視される。従って、δE の大きさは、それが小さ過ぎない限り、エントロピー $S(E,N)$ の値に影響を与えない。

更に統計力学では、エントロピー $S(E,N)$ に基づいて、全系 $\mathcal{S}_{\rm total}$ の**温度** $T(E,N)$ を次式で定義する。

$$\frac{1}{T(E,N)}=\frac{\partial S}{\partial E}(E,N).$$

従って、温度は E と N の関数である。

4　最適な瞬時符号の統計力学的解釈

前節での統計力学の復習を踏まえ、本節では、最適な瞬時符号による無歪み情報源符号化に対して、統計力学的解釈を与える。

X をアルファベット χ に値をとる任意の確率変数とし、C を確率変数 X に対する最適な瞬時符号とする。$X_1,X_2,\ldots,X_N$ を、確率分布 $p_X(x)$ に従う定常独立な確率変数列とする。ここで、N は 10^{22} 程度の非常に大きな数とする。瞬時符号 C による無歪み情報源符号化の枠組は、前節で展開した統計力学の基本的枠組に、以下のように対応する。

確率変数列 $X_1,X_2,\ldots,X_N$ は、量子系 $\mathcal{S}_{\rm total}$ 対応する。その際、各確率変数 X_i は i 番目の量子部分系 $\mathcal{S}_i$ に対応する。各 $x\in\chi$ に対し、x（同じことであるが、$C(x)$）は、部分系のエネルギー固有状態に対応し、$l(x)=|C(x)|$ は、そのエネルギー E_n に対応する。このとき、情報源系列 $(x_1,\ldots,x_N)\in\chi^N$（同じことであるが、符号語系列 $C(x_1)\cdots C(x_N)$）は、$(n_1,\ldots,n_N)$ で指定される系 $\mathcal{S}_{\rm total}$ のエネルギー固有状態に対応する。そして、$l(x_1)+\cdots+l(x_N)$ は、系 $\mathcal{S}_{\rm total}$ のそのエネルギー固有状態のエネルギー $E_{n_1}+\cdots+E_{n_N}$ に対応する。

なお、この無歪み情報源符号化の統計力学的解釈においては、$l(x)=|C(x)|$ であるので、“部分系” では、“エネルギー固有状態” である有限二進列 $C(x)$ の長さが、その状態が持つ “エネルギー” に等しいという状況が成立している。しかし、$l(x_1)+\cdots+l(x_N)=|C(x_1)\cdots C(x_N)|$ が成り立つので、これは “部分系” に限らず “全系” でも成り立っていることがわかる。このように、無歪み情報源符号化に対す

る我々の統計力学的解釈においては、一般に、“エネルギー固有状態” である有限二進列の長さは、その状態が持つ “エネルギー” に等しい。

$A(L,N)$ を、符号語系列 $C(x_1)\cdots C(x_N)$ のうち、その長さが L と $L+\delta L$ の間にあるもの全てからなる集合とする。そして、$\Omega(L,N)$ を $\#A(L,N)$ で定義する。即ち、$\Omega(L,N)$ とは、N 個の符号語からなる符号語系列のうち、その長さが L と $L+\delta L$ の間にあるものの総数である。さてこのとき、C が最適な瞬時符号であることから、定理 1 より、$C(x_1)\cdots C(x_N)\in A(L,N)$ ならば $2^{-(L+\delta L)}\leq p(x^N)\leq 2^{-L}$ となることがわかる。従って、$A(L,N)$ に属する符号語系列 $C(x_1)\cdots\ C(x_N)$ は、全て同じ確率 2^{-L} で生起することがわかる。なおここで、我々は前節の統計力学での議論に倣い、δL の大きさは問題とせず、L に比べて十分に小さいものとして、無視してしまうことに注意されたい。このようにして、**The Principle of Equal Conditional Probability** と名付ける次の原理が成り立つ。

The Principle of Equal Conditional Probability: 符号語系列の長さが L に等しいという条件の下で、各符号語系列が生起する条件付確率は、どれも $1/\Omega(L,N)$ である。 □

我々は、この The Principle of Equal Conditional Probability に基づいて、無歪み情報源符号化の枠組の中に、ミクロカノニカルアンサンブルを同定する。そして、このミクロカノニカルアンサンブルから出発して、無歪み情報源符号化の枠組の上に、平衡統計力学を構築する。その方法は、前節で見た通常の平衡統計力学に対するものと同じであり、通常の平衡統計力学の理論展開を単になぞり返せばよい。

ところで、統計力学においては、The Principle of Equal Probability は、現実的な物理系では、まだその成立が完全に証明させていない仮説である。これに対し、無歪み情報源符号化に対する我々の統計力学的解釈では、The Principle of Equal Conditional Probability は、仮定なしで自動的に成り立つものである。

前節ので統計力学の展開に倣い、瞬時符号 C の**統計力学的エントロピー** $S(L,N)$ を次式で定義する。

$$S(L,N)=\log\Omega(L,N). \tag{3}$$

ここで、統計力学のエントロピーの定義 (2) に現れる Boltzmann 定数 k は、$1/\ln 2$ に選んでいることに注意されたい。我々の統計力学的解釈は、無歪み情報源符号化の形式に対する解釈であり、実際の物理系とは関係がない。従って、k の選び方は任意であるが、二進数に立脚している情報理論との整合性を考えると、このように k を選

ぶのが自然である。この選択は、次節以降の議論によって正当化される。

このように定義した統計力学的エントロピー $S(L,N)$ に基づいて、統計力学の場合と全く同様に、瞬時符号 C の**温度** $T(L,N)$ を次式で定義する。

$$\frac{1}{T(L,N)} = \frac{\partial S}{\partial L}(L,N). \tag{4}$$

従って、温度は L と N の関数である。符号語系列の 1 符号語当りの平均の長さ λ は、L/N によって与えられる。平均の長さ λ は、前節の統計力学での平均エネルギー ε に対応する。

5 統計力学的エントロピーの性質

一般に、統計力学に基づく研究においては、量子部分系 $\mathcal{S}_i$ のエネルギー値 E_n を、全ての量子数 n について知ることが重要である。それらの値が量子系 $\mathcal{S}_{\rm total}$ のエントロピー $S(E,N)$ を決定し、系の物理的性質を決定するからである。この事実に対応し、無歪み情報源符号化に対する我々の統計力学的解釈においても、全ての $x \in \chi$ に対する $l(x)$ の具体的な値が重要な役割を果たす。特に、それらの値により統計力学的エントロピー $S(L,N)$ が決定される。本節では、関数 $l(x)$ に基づいて、$S(L,N)$ と $T(L,N)$ の性質を調べる。

統計力学でよく知られている通り、量子系 $\mathcal{S}_{\rm total}$ のエネルギーに上限値がある場合には、系は負の温度を持ち得る。これと同じ状況が、無歪み情報源符号化に対する我々の統計力学的解釈でも起こる。どんな瞬時符号でも、それの符号語は有限個しかないからである。

$l_{\min}$ と $l_{\max}$ のそれぞれを、$\min\{l(x) \mid x \in \chi\}$ と $\max\{l(x) \mid x \in \chi\}$ で定義する。さて、N を固定した場合について考えよう。このとき、統計力学的エントロピー $S(L,N)$ は、L の単峰型関数であり、$Nl_{\min}$ と $Nl_{\max}$ の間でのみ $S(L,N)$ は非零となる。$S(L,N)$ を最大にする L の値を L_0 とおくと、温度の定義 (4) から、$L < L_0$ ならば $T(L,N) > 0$ であり、他方、$L > L_0$ ならば $T(L,N) < 0$ である。温度 $T(L,N)$ は、$L = L_0$ で $\pm\infty$ となる。

Boltzmann-Planck の方法 ([1] 参照) に従うことにより、

$$S(L,N) = NH(G(C,T(L,N))) \tag{5}$$

を示すことができる。ここで $G(C,T)$ は、アルファベット χ に値をとる確率変数で

あり、その確率分布 $p_{G(C,T)}(x) = \Pr\{G(C,T) = x\}$ は

$$p_{G(C,T)}(x) = \frac{2^{-l(x)/T}}{\sum_{a\in\chi} 2^{-l(a)/T}} \tag{6}$$

で定義されるものである。この $G(C,T)$ は瞬時符号 C と実数 T に依存している。式 (5) で温度 $T(L,N)$ は、式

$$\frac{L}{N} = \sum_{x\in\chi} l(x) p_{G(C,T(L,N))}(x) \tag{7}$$

を通じて、陰に、L と N の関数として決定される。

以上の $S(L,N)$ と $T(L,N)$ の性質は、集合 $A(L,N)$ の組み合わせ論的な側面のみに基づいて導かれたものであることに注意されたい。これに対し、以下では、確率変数列 $X_1, X_2, \ldots, X_N$ の導入によってもたらされる確率論的な問題を考察しよう。

瞬時符号 C は最適なので、定理 1 より、長さ L の特定の符号語系列は、確率 2^{-L} で生じる。これゆえ、長さ L の符号語系列のどれかが生じる確率は、$2^{-L}\Omega(L,N)$ で与えられる。従って、$2^{-L}\Omega(L,N)$ を L について微分し、結果を 0 とおくことで、符号語の数 N を一定とした場合における、符号語系列の長さの最も確からしい値 L^* を決定することができる。対数をとってから微分するようにすると、この L^* は次式を満たす。

$$\frac{\partial}{\partial L}\{-L + S(L,N)\}\Big|_{(L,N)=(L^*,N)} = 0.$$

この式から $T(L^*,N) = 1$ が得られる。従って、温度 1 は、(N を一定とした場合の) 符号語系列の最も確からしい長さ L^* に対応していることがわかる。他方、温度 $T(L^*,N) = 1$ では、まず式 (6) から $p_{G(C,1)}(x) = 2^{-l(x)}$ であり、これと式 (7) により $L^*/N = L_X(C) (= H(X))$ が得られる。この結果は、大数の法則に合致している。このように、温度 1 において、符号語系列の 1 符号語当りの平均の長さ λ は、平均符号長 $L_X(C)$ と一致し、温度 1 は瞬時符号 C の平均符号長 $L_X(C)$ に対応している。

6 二つの瞬時符号の間の熱平衡

本節では、二つの瞬時符号の間に生じる “熱平衡概念” について考察する。X^{I} をアルファベット χ^{I} に値をとる任意の確率変数とし、C_{I} をそれに対する最適な瞬時符号とする。そして、$X^{\mathrm{I}}_1, X^{\mathrm{I}}_2, \ldots, X^{\mathrm{I}}_{N_{\mathrm{I}}}$ を確率分布 $p_{X^{\mathrm{I}}}(x)$ に従う定常独立な確率変数

列とする。ここで、N_{I} は非常に大きな数とする。他方、X^{II} をアルファベット χ^{II} に値をとる任意の確率変数とし、C_{II} をそれに対する最適な瞬時符号とする。そして、$X_1^{\mathrm{II}}, X_2^{\mathrm{II}}, \ldots, X_{N_{\mathrm{II}}}^{\mathrm{II}}$ を確率分布 $p_{X^{\mathrm{II}}}(x)$ に従う定常独立な確率変数列とする。ここで、やはり N_{II} は非常に大きな数であるとする。なお、C_{I} と C_{II} は同じものである必要はないし、$N_{\mathrm{I}} = N_{\mathrm{II}}$ である必要もない。

さてこのとき、次の問題を考えよう: 確率変数列 $\{X_i^{\mathrm{I}}\}$ に対する瞬時符号 C_{I} による符号語系列の長さ L_{I} と、確率変数列 $\{X_i^{\mathrm{II}}\}$ に対する瞬時符号 C_{II} による符号語系列の長さ L_{II} の和 $L_{\mathrm{I}} + L_{\mathrm{II}}$ が、与えられた L に等しいという条件の下で、最も確からしい L_{I} と L_{II} の値を見つける。

この問題を解くために、統計力学における熱平衡の概念を用いることができる。初めに次の事実に注意しよう。長さが L_{I} である C_{I} の特定の符号語系列と、長さが L_{II} である C_{II} の特定の符号語系列は、確率 $2^{-L_{\mathrm{I}}}2^{-L_{\mathrm{II}}} = 2^{-L}$ で同時に現れる。これは、瞬時符号 C_{I} と C_{II} が最適だからである。これゆえ、C_{I} の符号語系列と C_{II} の符号語系列の特定の対は、対を形成するそれら二つ列の長さの和が L に等しいという条件の下では、どの特定の対も同じ確率で生起する。従って、最も確からしい $L = L_{\mathrm{I}} + L_{\mathrm{II}}$ の配分 $(L_{\mathrm{I}}^*, L_{\mathrm{II}}^*)$ は、積 $\Omega_{\mathrm{I}}(L_{\mathrm{I}}, N_{\mathrm{I}})\Omega_{\mathrm{II}}(L_{\mathrm{II}}, N_{\mathrm{II}})$ を最大にするものである。この条件は、統計力学的エントロピー及び温度の定義式 (3)、(4) から、次の等式に同値である。

$$T_{\mathrm{I}}(L_{\mathrm{I}}^*, N_{\mathrm{I}}) = T_{\mathrm{II}}(L_{\mathrm{II}}^*, N_{\mathrm{II}}).$$

ここで、関数 T_{I} と T_{II} は、それぞれ、C_{I} と C_{II} の温度である。この等式は、統計力学において、全エネルギーが一定の場合の二つの系の間の熱平衡の条件に対応するものである。

さて、L_{I}^* および L_{II}^* の具体的計算方法であるが、まず、T を未知数とする次の方程式を解くことにより、$T_{\mathrm{I}}(L_{\mathrm{I}}^*, N_{\mathrm{I}})(= T_{\mathrm{II}}(L_{\mathrm{II}}^*, N_{\mathrm{II}}))$ が求まる[*2]。

$$\frac{N_{\mathrm{I}}}{L} \sum_{x \in \chi^{\mathrm{I}}} |C_{\mathrm{I}}(x)|\, p_{G(C_{\mathrm{I}},T)}(x) + \frac{N_{\mathrm{II}}}{L} \sum_{x \in \chi^{\mathrm{II}}} |C_{\mathrm{II}}(x)|\, p_{G(C_{\mathrm{II}},T)}(x) = 1. \tag{8}$$

$T_{\mathrm{I}}(L_{\mathrm{I}}^*, N_{\mathrm{I}})$ が式 (8) を満たすことは、式 (7) と $L = L_{\mathrm{I}} + L_{\mathrm{II}}$ から確認できる。一旦 $T_{\mathrm{I}}(L_{\mathrm{I}}^*, N_{\mathrm{I}})$ の値が求まれば、再び式 (7) により、L_{I}^* および L_{II}^* が計算できる。

[*2] 実際の計算の際には、T ではなく、まず $2^{-\frac{1}{T}}$ について (8) を解く。

7　符号語系列集合のフラクタル次元

フラクタル次元の概念は、フラクタル幾何学において中心的な役割を果たしている[7]。本節では、フラクタル次元、特にボックス次元を用いて、無歪み情報源符号化に対する我々の統計力学解釈を更に調べる。

F を $\mathbb{R}$ の有界な部分集合とし、$N_n(F)$ を F と交わる 2^{-n}-mesh cubes の個数とする。ここで、2^{-n}-mesh cube とは、ある整数 m に対して $[m2^{-n}, (m+1)2^{-n}]$ の形をした $\mathbb{R}$ の部分集合のことである。このとき、F の**ボックス次元** $\dim_B F$ は次式で定義される。

$$\dim_B F = \lim_{n\to\infty} \frac{\log N_n(F)}{n}.$$

さて、

$$\{0,1\}^\infty = \{b_1 b_2 b_3 \cdots \mid b_i = 0,1 \text{ for all } i = 1,2,3,\dots\}$$

を無限二進列（即ち、$0,1$ 記号からなるの片側無限列）全体の集合とする。我々は、[12, 13] で、アルゴリズム的情報理論（algorithmic information theory）の文脈において、無限の長さを持つ符号語系列の集合のフラクタル次元（特に、ハウスドルフ次元）を研究した。ここでは、その方法に従い、瞬時符号 C の符号語系列で無限の長さを持つものを考察し、特に、それらの中でも一定の温度を持つものから成る集合を考え、その集合のボックス次元を調べる。

式 (7) により、比 L/N は、温度 T から一意に決定されることがわかる。従って、比 L/N を一定に保ちながら $L, N \to \infty$ とすることにより、集合 $A(L,N)$ は、$\{0,1\}^\infty$ の部分集合とみなすことができる。この種の極限操作は、統計力学では、**熱力学的極限**と呼ばれる。熱力学的極限をとった結果において、$A(L,N)$ を $F(T)$ と表すことにする。その際、T は式 (7) を通じて、L/N の極限値に関係付けられる。ところで、このように定義された $F(T)$ は、本来 $\{0,1\}^\infty$ の部分集合であるが、任意の無限二進列 $\alpha \in \{0,1\}^\infty$ を、二進表示の実数 $0.\alpha$ と同一視することにより、$F(T)$ を $[0,1]$ の部分集合とみなすことが可能である。このようにして、$F(T)$ に対して、そのボックス次元 $\dim_B F(T)$ を定義することができる。

以下では、$-\infty \le T \le \infty$ の範囲で、$\dim_B F(T)$ の温度 T に対する依存性を調べ

る。初めに、次が成り立つことがわかる。

$$\begin{aligned}\dim_B F(T) &= \lim_{L,N\to\infty} \frac{\log \Omega(L,N)}{L} \\ &= \lim_{L,N\to\infty} \frac{S(L,N)}{L}.\end{aligned}$$

ここで極限 lim は、各 T に対して、式 (7) を満たしながら $L, N \to \infty$ としている。この式を見ると、瞬時符号 C の統計力学的エントロピー $S(L,N)$ と、$F(T)$ のボックス次元 $\dim_B F(T)$ とは強く関連していることがわかる。この式と、式 (5)、(6)、(7) により、$\dim_B F(T)$ は具体的には次式で与えられる。

$$\dim_B F(T) = \frac{1}{T} + \frac{1}{\lambda(T)} \log \sum_{x\in\chi} 2^{-l(x)/T}. \tag{9}$$

ここで、$\lambda(T)$ は次式で定義される。

$$\lambda(T) = \sum_{x\in\chi} l(x) p_{G(C,T)}(x).$$

瞬時符号 C の “最大エネルギー固有値” と “最低エネルギー固有値” のそれぞれの縮退度 $d_{\max}$ と $d_{\min}$ を、

$$\begin{aligned} d_{\max} &= \#\{x \in \chi \mid l(x) = l_{\max}\}, \\ d_{\min} &= \#\{x \in \chi \mid l(x) = l_{\min}\} \end{aligned}$$

によって定義する。なお、C は最適なので $\sum_{x\in\chi} 2^{-l(x)} = 1$ であり、それゆえ $d_{\max}$ は偶数である。さて、比 L/N（即ち、$\lambda(T)$）に関して増加する順に、$\dim_B F(T)$ の値を見て行くと、式 (9) より、次のようになる。

$$\begin{aligned} \lim_{T\to+0} \dim_B F(T) &= \frac{\log d_{\min}}{l_{\min}}, \\ \dim_B F(1) &= 1, \\ \lim_{T\to\pm\infty} \dim_B F(T) &= \frac{n \log n}{\sum_{x\in\chi} l(x)}, \\ \lim_{T\to-0} \dim_B F(T) &= \frac{\log d_{\max}}{l_{\max}}. \end{aligned}$$

ここで、C の全ての符号語の長さが同一でない限り、$n \log n < \sum_{x\in\chi} l(x)$ であることに注意されたい。また、やはりそのような自明な場合を除けば、明らかに

$\log d_{\min}/l_{\min} < 1$ かつ $\log d_{\max}/l_{\max} < 1$ となる。従って、一般に、$\dim_B F(T)$ は温度 $T = 1$ で最大となることがわかる。

この事実はまた、式 (9) を利用する $\dim_B F(T)$ の微分計算に基づいて確認することができる。即ち、C の全ての符号語の長さが同一でない限り、次が成り立つことが、この微分計算で確認できる。

$$\text{(i)}\ \ \frac{d}{dT}\dim_B F(T)\Big|_{T=T_0} = 0 \iff T_0 = 1,$$

$$\text{(ii)}\ \ \frac{d^2}{dT^2}\dim_B F(T)\Big|_{T=1} < 0.$$

さて、$\dim_B F(1) = 1$ であり、$\dim_B F(1)$ は $\dim_B[0,1]$ に一致するから、集合 $F(1)$ は、或る意味において、集合 $[0,1]$ くらいに大きい。これは次のように説明することができる。

C は最適な瞬時符号なので、無限の長さを持つ符号語系列 $C(x_1)C(x_2)\cdots$ は、全体で集合 $\{0,1\}^\infty$ を埋め尽くし、従って $[0,1]$ を埋め尽くす。更に、C は最適なので、これらの符号語系列は、$[0,1]$ 上の一様分布、即ち、$[0,1]$ 上の Lebesgue 測度に従って生起する。一方、N が十分に大きい場合には、大数の法則により、N 個の符号語から成る符号語系列の長さは、ほぼ確実に $NL_X(C)$ に等しい。これらの観察は、N が十分に大きい場合には、長さが $NL_X(C)$ となる符号語系列が、或る意味において、$[0,1]$ を埋め尽くすことを示す。このとき、第 5 節で見たように、温度 $T = 1$ は、長さが $NL_X(C)$ となるこれらの符号語系列に対応するので、$T = 1$ の場合の $F(T)$、即ち $F(1)$ は、集合 $[0,1]$ くらいに “大きい” のである。

このように、$F(1)$ は、或る意味において、情報源系列のティピカル列に対応する符号語系列の全てを含み、これにより $\dim_B F(1) = 1$ が成り立つ。

8 まとめ

本稿では、最適な瞬時符号による無歪み情報源符号化の枠組おいて、ミクロカノニカルアンサンブルを同定し、それに基づいて、情報源符号化の枠組に対し平衡統計力学的解釈を与えた。その際、統計力学的エントロピー、温度、熱平衡など、統計力学における諸概念を情報理論に移植し、それらの情報理論的な性質を調べた。特に、この統計力学的解釈では、温度 1 は瞬時符号の平均符号長 $L_X(C)$ に対応することを発

見し、この対応の存在は、ボックス次元を用いた解析によっても確認できた。

なお、本稿では議論しなかったが、無歪み情報源符号化の枠組に対するこの統計力学的解釈では、通常の平衡統計力学の理論展開に更に従うことにより、瞬時符号のカノニカルアンサンブルの概念、並びに、瞬時符号の化学ポテンシャルや、それで特徴付けられる粒子の交換に関する平衡や、グランドカノニカルアンサンブルの概念を導入することが可能である。そして、それらの情報理論的な意義について考察を行うことが可能である。この事実は、本稿で我々が行った、無歪み情報源符号化の枠組におけるミクロカノニカルアンサンブルを同定が、自然かつ普遍的なものであることを示している。

参考文献

[1] 有村卓, 確率・情報・エントロピー. 森北出版, 1980.

[2] R. B. Ash, *Information Theory.* Dover Publications, Inc., New York, 1990.

[3] C. S. Calude and M. A. Stay, "Natural halting probabilities, partial randomness, and zeta functions," *Inform. and Comput.*, vol. 204, pp. 1718–1739, 2006.

[4] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. John Wiley & Sons, Inc., New York, 2006.

[5] A. Dembo and O. Zeitouni, *Large Deviations Techniques and Applications*, 2nd ed. Springer, New York, 1998.

[6] P. A. M. Dirac, *The Principles of Quantum Mechanics*, 4th ed. Oxford University Press, London, 1958.

[7] K. Falconer, *Fractal Geometry*, *Mathematical Foundations and Applications.* John Wiley & Sons, Inc., Chichester, 1990.

[8] 韓太瞬, 小林欣吾, 情報と符号化の数理. 培風館, 1999.

[9] F. Reif, *Fundamentals of Statistical and Thermal Physics.* McGraw-Hill, Inc., Singapore, 1965.

[10] D. Ruelle, *Statistical Mechanics*, *Rigorous Results*, 3rd ed. Imperial College Press and World Scientific Publishing Co. Pte. Ltd., Singapore, 1999.

[11] C. E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, pt. I, pp. 379–423, 1948; pt. II, pp. 623–656, 1948.

[12] 只木孝太郎, アルゴリズム的情報理論とフラクタル集合. 1999 年情報論的学習理論ワークショップ (IBIS'99) 予稿集, pp. 105–110, 1999 年 8 月 26, 27 日, 静岡県 田方郡 修善寺.

[13] K. Tadaki, "A generalization of Chaitin's halting probability Ω and halting self-similar sets," *Hokkaido Math. J.*, vol. 31, pp. 219–253, 2002.

[14] K. Tadaki, A statistical mechanical interpretation of instantaneous codes. Proceedings of 2007 IEEE International Symposium on Information Theory (ISIT2007), pp. 1906–1910, June 24-29, 2007, Nice, France.

[15] 戸田盛和, 久保亮五編, 統計物理学, 岩波講座 現代物理学の基礎 [第 2 版] 5. 岩波書店, 1978.