統計数理(2009)
第 57 巻 第 2 号 221–232

特集「ネットワーク情報理論」
[研究詳解]

センシングと符号化の統計力学

村山 立人†

(受付 2009 年 1 月 7 日；改訂 2 月 26 日；採択 2 月 26 日)

要　　旨

現在，あらゆるセンサーの小型化と量産化が加速している．そして，これらは現在のコンピューター網に接続されていくと予想される．すると，ネットワークに諸センサーが統合されたシステムが情報基盤として確立する可能性は高い．これは，計測データを効率的に伝送するための通信技術の需要が拡大することを意味する．同時に，センシングのために行う符号化の理論的限界も重要になる．本稿では，情報理論と統計力学を背景にした学術的知見に基づく，センシングと符号化についての新しいアプローチを解説する．

キーワード： センサーネットワーク，符号化，統計力学．

1. はじめに

近い将来，デバイスとセンサーのネットワークは社会のあらゆる場面で活躍するようになると予想されている．この新しいタイプの次世代ネットワークは「センサーネットワーク」と標語的に呼ばれることもあり，農場管理，工場制御，犯罪監視，そして軍事利用に至るまで幅広い応用が期待されている．実際，センサーネットワークに対する半導体大手や軍事部門の注目度は極めて高く，潜在的な未来市場を開拓するための最有力技術として認識されている．しかし，そのような注目度にもかかわらず，センサーネットワークを全体としてうまく統合するための方法はあまり知られていない．つまり，デバイス，ソフトウェア，省電力化の方法などの個別テーマでは技術革新が進行しつつあるのだが，この新しい技術をシステム・レベルで理解しようとする動向は意外に弱いのである．このようなシステム的視点の確立には，新しい切り口で技術を再考する必要がある．そして，システムに誘発される協同現象とその結果としてのトレードオフを数学的に記述することができれば，今後の実用的研究にも役立つだろう．本稿ではこのような野心的な動機を持ちつつ，以下の枠組みを精密科学の立場から分析していく．

今，データセンターがあるデータ系列 $\{X(t)\}_{t=1}^{\infty}$ に興味を持っているが，これを直接計測できないものとする．そこで，データセンターは L 個のセンサーを周囲に配置したとしよう．各センサーはノイズのある環境で計測した系列 $\{Y_i(t)\}_{t=1}^{\infty}$ をそれぞれ独立に符号化する．つまり，各センサーは互いに通信することができず，したがって，事前にデータセンターに伝送する内容についていっさい協調できないものとする．データセンターは，L 個の符号語を通信回線を利用して回収し，元の系列 $\{X(t)\}_{t=1}^{\infty}$ をできるだけ復元したいと考えている．しかし，このデータ系列だけがデータセンターにとっての重要な事柄ではないので，各センサーが利用できるデータ伝送率(通信速度)の合計 R は厳密に制限されている．つまり，データセンターは一定の回線速度でしか符号語を回収できない．このような推定操作を伴った分散型通信のモデル

† NTT コミュニケーション科学基礎研究所：〒619–0237 京都府相楽郡精華町光台 2–4

は Berger-Zhang-Viswanathan によって定式化され(Berger et al., 1996), 情報理論の立場からセンサーネットワークの理論的枠組みを提供していると解釈されている．彼らの仕事によって，大規模観測系におけるいくつかの興味深い性質が明らかにされた．もしセンサーが互いに通信できるなら，センサー数 L が無限の極限において，独立に発生している計測ノイズを完全に除去することが可能となる．したがって，$D(\cdot)$ を $\{X(t)\}$ の歪み・レート関数(レート・歪み関数の逆関数)として，データセンターは任意の忠実度で「歪み」$D(R)$ を達成することが保障されている．逆に，センサーが互いに通信できないなら，有限の合計伝送率 R で歪み D を無限に小さくすることはできない．たとえ，無限個のセンサーが利用可能であったとしても，それは実現できないと証明できるのである(Berger et al., 1996)．

本稿では，有限の合計伝送率 R での分散化の極限 $L\to\infty$ の効率を見通しよく議論するために，簡単なシステムモデルを導入する．より詳細にいうと，センサーはレート・歪み符号として低密度符号を利用し，データセンターは L 個の復号系列にビットごとの「多数決」を行うことによってベイズ最適な推定操作を実現するものとする(MacKay, 2003)．このとき，合計伝送率 R を既与として，どの程度のセンサー数 L が最適であるかを議論する分散観測問題を本稿では提案する．本稿の漸近的議論によって，全システムの効率を $L\to\infty$ の極限で評価することが可能となるが，これは個々のセンサーが送信に利用できる伝送率がゼロに収束することを意味する．ここで，統計力学の計算技術である「レプリカ法」と確率論の有名定理である「中心極限定理」を組み合わせることにより，理論上の取り扱いが困難な発散項の精密な評価を行ったのが議論の特色である．

次章より，本稿は以下のように構成される．まず，第 2 章では，解析的に分析が容易なシステムのモデルを導入する．次に，第 3 章でこの方法による結果を要約し，続く第 4 章で導出の概要を情報理論と統計力学の両方向からスケッチする．そして，最終章において簡単なまとめを行う．

2. システムモデル

本稿では，現実のシステムの詳細に依存しない普遍的な性質を議論する．そこで，システムが分散符号化によって享受する情報利得を単純な形式で抽出する目的で，生成されるデータ系列は冗長性を持たないように下記のように設定する．いま，データ系列 $\{X(t)\}\in\mathcal{X}$ に共通の確率分布を $P(x)$ とする．また，$\mathcal{Y}$ を計測系列 $\{Y_i(t)\}$ に共通のアルファベットとし，$\mathcal{X}\times\mathcal{Y}$ 上で定義される確率行列を $W(y|x)$ とする $(i=1,\ldots,L,t\geq 1)$．まず，無記憶情報源 $\{X(t)\}_{t=1}^{\infty}$ に対し，同時確率分布を次のように仮定する．

$$\Pr[x,y_1,\ldots,y_L]=P(x)\prod_{i=1}^{L}W(y_i|x).$$

ここで，確率変数 $Y_i(t)$ は $X(t)$ に対して独立であり，条件つき確率 $W[y_i(t)|x(t)]$ の値はすべての i と t に関して同一である．さらに本稿では，この問題をもっとも単純な 2 値系列で議論していく．つまり，データ系列 $\{X(t)\}$ と，それを一定のノイズレベルで計測した系列 $\{Y_i(t)\}$ はすべて 2 値系列であると仮定される．したがって，確率行列は次のようにパラメータ化できる．

$$W(y|x)=\begin{cases}1-p & (y=x)\\ p & (y\neq x).\end{cases}$$

ここで，$p\in[0,1/2]$ は計測におけるノイズのレベルを意味し，アルファベットは $\mathcal{X}=\mathcal{Y}$ と選択されている．さらに，簡単のため，$P(x)=1/2$ がいつも成立すると仮定しよう．これは，全く冗長性のないランダムな情報源を計測していることに対応する．

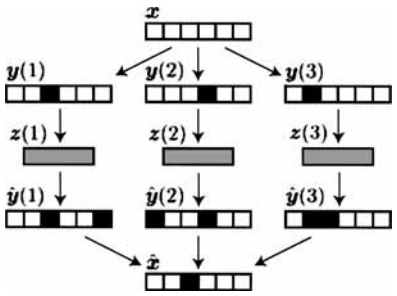

図 1. システムモデルの概念図．ここに描かれているのは，合計伝送率 $R=2$，センサー数 $L=3$，つまり各センサーにおける伝送率が $m/n=2/3$ のネットワークである．簡単のため $n=6$，$m=4$ とした．

符号化の段階では，センサー i が計測系列 $\{y_i(t)\}_{t=1}^{\infty}$ から長さ n のブロック $\boldsymbol{y}_i=[y_i(1),\ldots,y_i(n)]^T$ を切り取り，$\mathcal{Z}$ 上で定義された長さ m のブロック $\boldsymbol{z}_i=[z_i(1),\ldots,z_i(m)]^T$ にブロックごと符号化する（図 1）．以後，ブール代数の表記にならい，$\mathcal{X}=\{0,1\}$，したがって $\mathcal{Y}=\mathcal{Z}=\{0,1\}$ とする．いま，$\hat{\boldsymbol{y}}_i$ をこのブロックの復号系列で，圧縮系列の長さ $m(<n)$ が既与だとする．本稿では，比較のため，次に述べる二通りの分散符号化を検討する．(1) 各エージェントは独立にレート・歪み関数を達成する符号化を行うと仮定する．(2) 各エージェントは独立に準最適な低密度符号を実装している．本稿でいう低密度符号化では，$n\times m$ 型行列の 2 値行列 A_i を準備し，m ビットの系列 $\boldsymbol{z}_i=[z_i(1),\ldots,z_i(m)]^T$ が線形復号条件

$$\hat{\boldsymbol{y}}_i=A_i\boldsymbol{z}_i \pmod 2, \tag{2.1}$$

と忠実度規範

$$D=\frac{1}{n}d_{\mathrm{H}}(\boldsymbol{y}_i,\hat{\boldsymbol{y}}_i)$$

を満足するときに符号語（のひとつ）として定義する（Murayama and Okada, 2003）．ここで，ハミング距離 $d_{\mathrm{H}}(\cdot,\cdot)$ が歪み測度として採用されている．また，式 (2.1) では 2 を法とした加法を用いていることに注意．今，行列 A_i の各行にそれぞれ K 個，各列に C 個だけ非ゼロ要素の 1 が存在するように作成したとする．このとき，有限でしかも通常は小さい値を持つ K と C によって，低密度符号の符号族が指定されることになる．ここで，パラメータ K の値が非常に大きくなると低密度符号はレート・歪み関数を達成することが知られている．そのため，低密度符号化における $K\to\infty$ の極限を構成的に議論できるのなら，レート・歪み関数の存在を仮定した情報理論の分析と整合的な結論を与えるはずである．

復号・推定の段階では，データセンターは L 個の符号語の系列 $\boldsymbol{z}_1,\ldots,\boldsymbol{z}_L$ を回収することになる．符号語の長さはすべて m なので，合計の伝送率は $R=L\times m/n$ となる．そのため，この枠組みでは，データセンターは同一程度の歪みを持つ復号系列 $\hat{\boldsymbol{y}}_1,\ldots,\hat{\boldsymbol{y}}_L$ を提供する交換可能なセンサーを配置していることになる．最後に，推定系列 $\hat{\boldsymbol{x}}=[\hat{x}(1),\ldots,\hat{x}(n)]^T$ の第 t 番目のビットは復号系列の対応する L 個のビットの多数決によって計算される（MacKay, 2003）：

$$\hat{x}(t)=\begin{cases}0 & (\hat{y}_1(t)+\cdots+\hat{y}_L(t)\le L/2)\\ 1 & (\hat{y}_1(t)+\cdots+\hat{y}_L(t)>L/2).\end{cases} \tag{2.2}$$

よって，システム全体の性能は多数決 (2.2) によるビット誤り率の期待値 $P_{\mathrm{e}}=\Pr[x\neq\hat{x}]$ によって定義するのが自然である．本稿では，分散化のレベルをシステムの「戦略」と解釈して，次

の 2 つの選択肢を考える. (1) 無限個のセンサーで系列を無限に圧縮する: $L\to\infty$. (2) R 個のセンサーで系列を圧縮しない: $L=R$. 前者の戦略では各センサーに配分される伝送率はゼロに収束し，後者の戦略では符号化を行わないで通信をすることになる．しかし，一般には，どちらの戦略がある特定の系列の推定に適しているのかを決定するのは難しい．つまり，どちらの戦略がより小さいビット誤り率の期待値 $P_{\rm e}$ を与えるのかが自明ではないのである．実際，レート・歪み符号を用いることによって，データセンターはより多くの数のセンサーを利用することが可能になる．しかし，同時に各センサーが提供する復号系列の歪みはより大きくなるだろう．最適な分散化レベルの選択は，計測におけるノイズレベル p と通信における合計伝送率 R に依存して決定されるはずである．

3. システムサイズ効果

まず最初に，本稿で解説する情報科学的あるいは物理科学的アプローチによって得られるシステムサイズ効果の分析結果をあらかじめ要約する．簡単のため $K=1,2$ の低密度符号族と，$K\to\infty$ の極限を議論する．前章で触れたとおり，$K\to\infty$ の極限はシステムの理論限界を与えるレート・歪み関数に対応することに注意したい．今，計測におけるノイズレベルを p, そして通信における合計伝送率が有限の実数 R だとする．$L\to\infty$ の極限では，データセンターの推定におけるビット誤り率の期待値は

$$P_{\rm e}(p,R)=\int_{-\infty}^{-(1-2p)c_g\sqrt{R}}\frac{dr}{\sqrt{2\pi}}\exp\left(-\frac{r^2}{2}\right) \tag{3.1}$$

となる．ここで，低密度符号族に依存する定数は

$$c_g=\begin{cases}1 \quad (K=1)\\ \frac{1}{\sqrt{2}}\left\{\frac{2\ln 2}{\sqrt{\alpha}}+\frac{\sqrt{\alpha}}{2}[1-\langle\tanh^2 x\rangle_{\pi(x)}^2]-\sigma^2\sqrt{\alpha}[1-\langle\tanh^2 x\rangle_{\pi(x)}]+\frac{2}{\sqrt{\alpha}}\langle\ln\cosh x\rangle_{\pi(x)}\right\}\ (K=2)\\ \sqrt{2\ln 2}\quad (K\to\infty)\end{cases}$$

と求めることができる．特に有限の $K=2$ の場合は，スケール変換された秩序変数の分散:

$$\sigma^2=\alpha\,\langle\hat{x}^2\rangle_{\hat{\pi}(\hat{x})}, \tag{3.2}$$

とエントロピー消失条件:

$$\begin{aligned}0=&\frac{2\ln 2}{\alpha}-\frac{1}{2}\left[1-\langle\tanh^2 x\rangle_{\pi(x)}^2\right]+\sigma^2\left[1-\langle\tanh^2 x\rangle_{\pi(x)}\right]\\ &+2\langle\tanh^2 x\rangle_{\pi(x)}\langle x\ \mathrm{sech}^2 x\tanh x\rangle_{\pi(x)}-2\sigma^2\langle x\ \mathrm{sech}^2 x\tanh x\rangle_{\pi(x)}\\ &+\frac{2}{\alpha}\langle\ln\cosh x\rangle_{\pi(x)}-\frac{2}{\alpha}\langle x\tanh x\rangle_{\pi(x)}\end{aligned} \tag{3.3}$$

によって解析的に記述されている．スケール変換された分散 σ^2 とスケール不変なパラメータ α の値は，連立方程式 (3.2), (3.3) を数値的に解くことによって求めることができる．ただし，次のような略記法を用いた.

$$\begin{aligned}\langle\,\cdot\,\rangle_{\pi(x)}&=\int_{-\infty}^{\infty}\frac{dx}{\sqrt{2\pi\sigma^2}}\exp\left[-\frac{x^2}{2\sigma^2}\right](\,\cdot\,),\\ \langle\,\cdot\,\rangle_{\hat{\pi}(\hat{x})}&=\int_{-1}^{+1}\frac{d\hat{x}}{\sqrt{2\pi\sigma^2}}(1-\hat{x}^2)^{-1}\times\exp\left[-\frac{(\tanh^{-1}\hat{x})^2}{2\sigma^2}\right](\,\cdot\,).\end{aligned}$$

よって，式 (3.1) を与えられた p と R に対して数値的に評価するのは容易である.

さらに，有限の合計伝送率 R が与えられたとき，推定系列の相対的品質がノイズの大きさ p にどのように依存するのかを議論しよう．図 2，図 3 および図 4 には，デシベル (dB) 単位で測られた $P_{\mathrm{e}}(p,R)$ の典型的な挙動を示している．ただし，ここでは R を整数に制限し，参照レベルは次のように設定した．

$$P_{\mathrm{e}}^{(0)}(p,R)=\begin{cases}\displaystyle\sum_{l=0}^{(R-1)/2}\binom{R}{l}(1-p)^l p^{R-l} & (R \text{ is odd})\\ \displaystyle\sum_{l=0}^{R/2-1}\binom{R}{l}(1-p)^l p^{R-l}+\frac{1}{2}\binom{R}{R/2}(1-p)^{R/2}p^{R/2} & (R \text{ is even}).\end{cases} \tag{3.4}$$

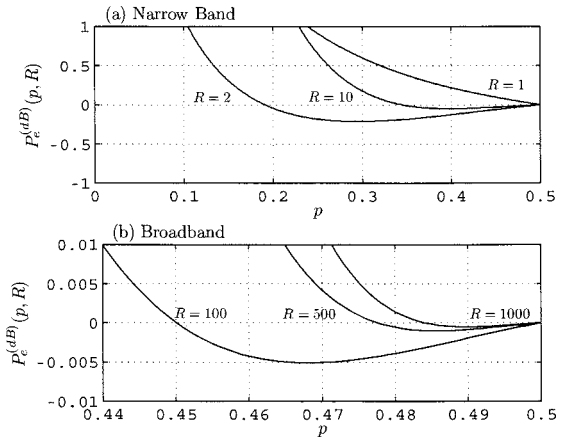

図 2. $K=1$ の単純量子化による分散符号化の数値解析．デシベル (dB) 単位で測った参照レベル $P_{\mathrm{e}}^{(0)}(p,R)$ に対する $P_{\mathrm{e}}(p,R)$ の相対的大きさを測った．(a) 合計伝送率 R の小さいナローバンド回線．(b) 合計伝送率 R の大きいブロードバンド回線．

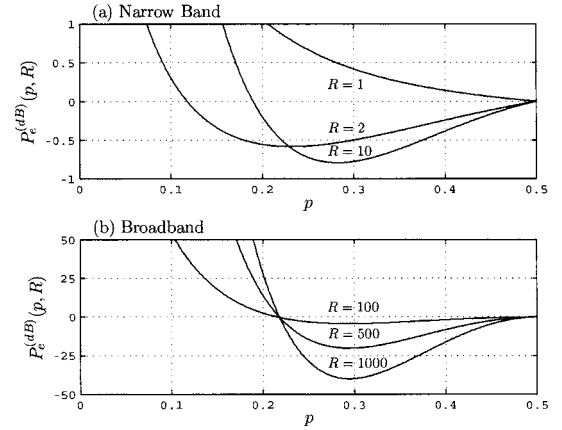

図 3. $K=2$ の低密度符号による分散符号化の数値解析．デシベル (dB) 単位で測った参照レベル $P_{\mathrm{e}}^{(0)}(p,R)$ に対する $P_{\mathrm{e}}(p,R)$ の相対的大きさを測った．(a) 合計伝送率 R の小さいナローバンド回線．(b) 合計伝送率 R の大きいブロードバンド回線．

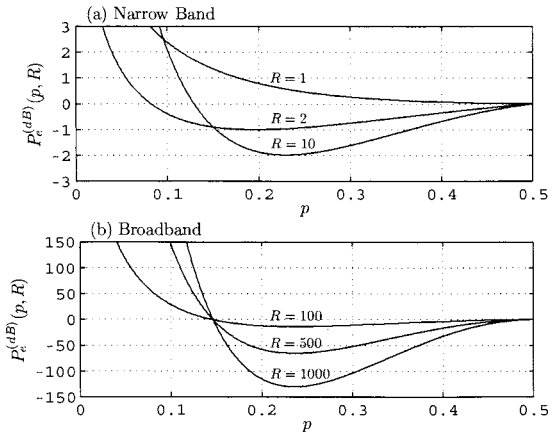

図 4. $K \to \infty$ の極限に対応するレート・歪み関数による分散符号化の数値解析．デシベル (dB) 単位で測った参照レベル $P_{\mathrm{e}}^{(0)}(p,R)$ に対する $P_{\mathrm{e}}(p,R)$ の相対的大きさを測った．(a) 合計伝送率 R の小さいナローバンド回線．(b) 合計伝送率 R の大きいブロードバンド回線．

参照レベル (3.4) はセンサー数 L を合計伝送率 R に一致させたときの P_{e} であり，これはセンサーが系列を全く圧縮しないシナリオに相当している．このとき，デシベル単位でのビット誤り率の期待値は

$$P_{\mathrm{e}}^{(\mathrm{dB})}(p,R) = 10 \log \frac{P_{\mathrm{e}}(p,R)}{P_{\mathrm{e}}^{(0)}(p,R)}, \tag{3.5}$$

と定義する．ただし，対数 log の底は 10 にとった．この単位でビット誤り率を測ることにすると，$P_{\mathrm{e}}(p,R)$ が参照レベル $P_{\mathrm{e}}^{(0)}(p,R)$ と同じになるときゼロになる．定義 (3.5) より，デシベルで測った量は負の値をとる可能性がある．そのようなときは，測定している分散化レベルでのビット誤り率の期待値が，参照レベルのものより小さくなっていることを意味している．数値解析によると，合計伝送率 R が小さいとき (ナローバンド) は，整数 R の偶奇性に強く依存したビット誤り率の挙動が観測できる (図 2 (a)，図 3 (a) および図 4 (a))．ここで，本来なら実数でも定義されている合計圧縮率を整数に限定しているのは，比較している参照レベルを自然に導入したいからである．特に，$R=2$ のケースでは，最も小さな閾値の値を持っているのがわかる．この閾値 p_c は，ここを超えると分散化の極限 $L \to \infty$ が参照レベル $L=R$ より大きな情報利得をもたらすことを意味する．しかし，$R=1$ のケースは特別で，このような閾値 p_c が存在していないのは興味深い．これとは対照的に，合計伝送率 R が大きいとき (ブロードバンド) は，デシベル単位で測ったビット誤り率の期待値の差 $P_{\mathrm{e}}^{(\mathrm{dB})}(p,R)$ が，定性的には安定した挙動を示す．$K=1$ では，R が大きくなる極限で p_c が $1/2$ に漸近していくようである (図 2 (b))．これは，分散化による情報利得が消滅することを示唆する．$K=2$ と $K \to \infty$ の極限では，R が大きくなるにしたがって前述の閾値 p_c がそれぞれ $1/2$ より小さいある値に収束していく様子がうかがえる (図 3 (b)，図 4 (b))．

本章の結果より，次のことが主張できる．つまり，レート・歪み符号による不可逆圧縮の自由度を各センサー i に与えるとき，それを利用した計測精度の向上が見込めるノイズ領域が存在する．それは，特徴的な閾値 p_c を超えた高ノイズ領域であり，この区間 $[p_c, 1/2]$ では「数」の効果が「質」の効果を凌駕していると解釈できる．逆に，低ノイズ領域 $[0, p_c]$ では，「質」の

効果が「数」の効果を凌駕しているので符号化による分散化の利得は得にくい．この結果は，計測と通信をうまく干渉させることで，システム全体として相乗効果が享受できる事実を理論的に示している．

4. 解析方法の解説と要約

4.1 情報科学的アプローチ

まず，レート・歪み理論における Shannon の定理を紹介する（Cover and Thomas, 1991）．この定理は情報源符号化定理（Shannon の第一定理），通信路符号化定理（Shannon の第二定理）に次ぐ第三の Shannon の定理であり，それは不可逆圧縮における限界記述長を与える．いま，ビット誤り率 D を許してデータを圧縮するとしよう．このとき，データの圧縮率が $r(D)$ より大きい限り，ビット誤り率が D より大きくならない符号化の方法が存在する．この限界の圧縮率 $r(D)$ を歪み D の関数とみなし，レート・歪み関数と呼ぶ．特に，単純な情報源のクラスでは，簡単に $r(D)$ が構成できることが知られている．仮に，n ビットのデータ $\boldsymbol{y}_i=[y_i(1),\ldots,y_i(n)]^T$ が m ビットの符号語 $\boldsymbol{z}_i=[z_i(1),\ldots,z_i(m)]^T$ に圧縮され，それが復号されて系列 $\hat{\boldsymbol{y}}_i=[\hat{y}_i(1),\ldots,\hat{y}_i(n)]^T$ を得たとしよう．簡単のため，元のデータ系列が全くのランダム系列であり，冗長性を利用した圧縮が不可能な場合を考えることにする（一般化は容易である）．すると，圧縮率を $r=m/n$ で定義して，復元におけるビット誤り率をハミング距離で測ることにすると，上述のレート・歪み関数は

$$r(D)=\begin{cases}1-h(D) & (0\le D\le 1/2)\\ 0 & (\text{otherwise}),\end{cases} \tag{4.1}$$

と求まる．ただし，$h(\cdot)$ は 2 値エントロピー関数であり，次のように定義された．

$$h(D)=-D\log_2 D-(1-D)\log_2(1-D).$$

このレート・歪み関数（4.1）は本稿で扱うシステム・モデルの分析にも便利な数学的道具である．

以後，センサー数が限りなく大きくなる極限を想定し，レート・歪み関数 $r(D)$ の $(D,r)=(1/2,0)$ 近傍について議論していく．関数（4.1）の連続性により，$D\in[0,1/2)$ においてテイラー展開すると

$$\begin{aligned}r(D)&=1-h(D)\\&=\frac{2}{\ln 2}\left(\frac{1}{2}-D\right)^2+\mathcal{O}\left(\left(\frac{1}{2}-D\right)^2\right)\end{aligned}$$

となる．ただし，ランダウの記号 $\mathcal{O}(\cdot)$ はその引数より高次の無限小を意味する（Lang, 1986）．ここで，関係式 $R/L=m/n$ を考慮すれば，センサー数 L と歪み D の間には漸近的に

$$\frac{R}{L}\approx\frac{2}{\ln 2}\left(\frac{1}{2}-D\right)^2 \tag{4.2}$$

が成立する．

一方，各センサーが独立に計測系列の符号化を行うと仮定すると，歪みに由来するビット誤りは Bernoulli 試行でモデル化できる（Chung, 2000）．そのため，復号系列 $\hat{\boldsymbol{y}}_i$ のビット誤り率は次のように求まる．

$$e=\Pr[x(t)\neq\hat{y}_i(t)]=p(1-D)+(1-p)D.$$

よって，推定系列 $\hat{\boldsymbol{x}}$ のビット誤り率の期待値は累積二項分布

$$P_{\mathrm{BER}}(e,L)=\begin{cases}B\left(\dfrac{L-1}{2}:e,L\right) & (L \text{ is odd})\\ B\left(\dfrac{L}{2}-1:e,L\right)+\dfrac{1}{2}b\left(\dfrac{L}{2}:e,L\right) & (L \text{ is even})\end{cases}$$

で記述できる (Hays, 1994). ただし，簡単のため

$$B(L':e,L)=\sum_{l=0}^{L'}b(l:e,L),$$
$$b(l:L,q)=\binom{L}{l}(1-e)^l e^{L-l}$$

と略記した．ここで, 整数 l は $\hat{\boldsymbol{y}}(t)$ における反転していない要素の合計を表し, 特に項 $(1/2)b(L/2:e,L)$ は $l=L/2$ となったときのランダムな推量を意味する．もちろん，記号 $\binom{L}{l}$ は L 個から l 個を選ぶ組み合わせの総数である.

では，センサーの数 L が十分に大きいとしよう．すると，累積二項分布は

$$\begin{aligned}P_{\mathrm{BER}}(e,L)&\approx B\left(\frac{L}{2}:e,L\right)\\&=\sum_{l=0}^{L/2}\binom{L}{l}(1-e)^l e^{L-l},\end{aligned}$$

と近似できる．さらに，統計学における基本的定理によると，二項分布と正規分布には

$$\begin{aligned}P_{\mathrm{e}}(p,R)&=\lim_{L\to\infty}P_{\mathrm{BER}}(e,L)\\&=\int_0^{L/2}du\ \mathrm{N}(L(1-e),Le(1-e))\end{aligned}\tag{4.3}$$

という関係式が成立する．ただし，$\mathrm{N}(X,Y)$ は平均 X で分散 Y の正規分布を表している (Hays, 1994). 積分 (4.3) を標準正規分布の形式にするには，測度を $r=(u-L(1-e))/\sqrt{Le(1-e)}$, $dr=du/\sqrt{Le(1-e)}$ と置き換えればよいことが知られている．その結果，ビット誤り率は

$$P_{\mathrm{e}}(p,R)=\lim_{L\to\infty}\int_{-\sqrt{L}}^{-r_c}dr\ \mathrm{N}(0,1)$$

と求まる．ただし r_c は次式を満たす.

$$r_c=\frac{L\left(\dfrac{1}{2}-e\right)}{\sqrt{Le(1-e)}}\approx 2\sqrt{L}(1-2p)\left(\frac{1}{2}-D\right).\tag{4.4}$$

この関係式は与えられた L,p,D の値に対しては常に成立し，符号化の個性は D の値に集約される．結局，分散化利得の理論限界は (4.2) と (4.4) を連立して

$$P_{\mathrm{e}}(p,R)=\int_{-\infty}^{(1-2p)\sqrt{2\ln 2R}}dr\ \mathrm{N}(0,1)\tag{4.5}$$

と求まる．これは前章の (3.1) において，$K\to\infty$ とした極限の式である.

4.2 物理科学的アプローチ

最近，統計力学の方法を利用して低密度符号の性能を分析する方法が確立された (Murayama and Okada, 2003). ここでは，この方法を用いて低密度符号の理論限界を導出し，分散符号化によ

る効果をスケーリング理論的な処方箋にしたがって計算してみる(Murayama and Davis, 2006).まず第一に,アルファベット $\mathcal{Z}=\{0,1\}$ を統計力学で多用するアルファベット $\mathcal{S}=\{+1,-1\}$ に翻訳する.すると,表現の整合性を維持するため,$\mathcal{Z}=\{0,1\}$ 上で定義される「加法」も,$\mathcal{S}=\{+1,-1\}$ 上で定義される「乗法」に翻訳する必要がある.例えば,$z_i(s)+z_i(s')\ (\mathrm{mod}\ 2)$ という数式は,$\sigma_i(s)\times\sigma_i(s')\in\mathcal{S}$ と書き換わる.同様にして,$y_i(t)$ を $J_i(t)$ に書き換えることができる.ただ簡単のため,L 個のセンサーを区別するための指標 i は省略することにする.以後,Sourlas の処方箋(Sourlas, 1989)に従い,Gibbs-Boltzmann 分布

$$\Pr[\boldsymbol{\sigma}]=\frac{\exp\left[-\beta H(\boldsymbol{\sigma}|\boldsymbol{J})\right]}{Z(\boldsymbol{J})} \tag{4.6}$$

を計算する.ただし,分配関数(規格化定数)

$$Z(\boldsymbol{J})=\sum_{\boldsymbol{\sigma}} e^{-\beta H(\boldsymbol{\sigma}|\boldsymbol{J})}$$

とハミルトニアン(エネルギー関数)

$$H(\boldsymbol{\sigma}|\boldsymbol{J})=-\sum_{s_1<\ldots<s_K}\mathcal{A}_{s_1\ldots s_K}J[t(s_1,\ldots,s_K)]\sigma(s_1)\cdots\sigma(s_K) \tag{4.7}$$

を適切に定義して用いた.ここで,時系列の指標 $t(s_1,\ldots,s_K)$ は,符号語の指標 $s_1,\ldots,s_K$ の集合に対応した t の値を指定し,パリティ検査条件(2.1)を満足させている.さらに,相互作用の希釈性を表現している対称テンソル $\mathcal{A}_{s_1\ldots s_K}$ の各要素は,指標集合 $(s_1,\ldots,s_K)$ の組み合わせに依存して 0 か 1 の値をとる.この符号化では指標 s に対して C 個の 1 がランダムに選択されるので,$\sum_{s_2,\ldots,s_K}\mathcal{A}_{ss_2\ldots s_K}=C$ が成立している.このとき,復号系列はひとつの指標 s に対して C 個のビットを持つが,これは K 個のビットを符号語から抽出していることになる.よって,符号化のレートは $R/L=K/C$ となっている.また,ハミルトニアン(4.7)が復号したときのエラー $[1-J[t(s_1,\ldots,s_K)]\cdot\sigma(s_1)\cdots\sigma(s_K)]/2$ を記録していることも容易に理解できる.

さらに統計力学によると,客観的な観測にかかる測定量(オブザーバブル)は自由エネルギーを利用することで解析的に計算が可能になっている.ここで,自由エネルギーとは,次式のように定義される関数である.

$$f=-\frac{1}{\beta}\langle\ln Z(\boldsymbol{J})\rangle_{\mathcal{A},\boldsymbol{J}}. \tag{4.8}$$

ここで,β は Gibbs-Boltzmann 分布(4.6)のパラメータで「逆温度」と呼ばれる.記号 $\langle\cdot\rangle_{\mathcal{A},\boldsymbol{J}}$ は配位平均を意味する.このため,自由エネルギーを計算するためには,分配関数 $Z(\boldsymbol{J})$ の対数に関する配位平均 $\langle\cdot\rangle_{\mathcal{A},\boldsymbol{J}}$ を実行する必要がある.しかし,これは数学的に困難な課題なので,いわゆるレプリカ法が利用される(Dotsenko, 2001).つまり,次の恒等式を利用して,実行が困難な分配関数の対数 $\ln Z(\boldsymbol{J})$ に関する平均操作をより簡単な分配関数 $Z(\boldsymbol{J})$ のべき乗の平均操作に帰着させるのである.

$$\langle\ln Z(\boldsymbol{J})\rangle_{\mathcal{A},\boldsymbol{J}}=\lim_{n\to 0}\frac{\langle Z(\boldsymbol{J})^n\rangle_{\mathcal{A},\boldsymbol{J}}-1}{n}.$$

こうして,自由エネルギー(4.8)が解析的に計算できると,低密度符号化における平均歪み D も次の関係式より直ちに求まることが知られている(Murayama and Okada, 2003).

$$D=\frac{1}{2}\left[1+f+\beta\frac{\partial f}{\partial\beta}\right].$$

よって,自由エネルギー(4.8)の $L\to\infty$ での振舞いを分析できれば,関係式(4.4)より P_{e} の評価が可能となる.以下,この処方箋にしたがった結果を要約し,統計力学による方法の汎用性

と情報理論の結果との整合性を確認しよう．計算の詳細は，参考文献(Murayama and Davis, 2006)などに記載されている．

$K=1$ の符号

$K=1$ の符号は自由エネルギーの厳密解が求まる．簡単な考察により，確率変数 x が標準正規分布 $p(x)$ にしたがうとして

$$-\beta f=\frac{n}{m}\left\langle \ln\left[2\cosh\left(\beta x\sqrt{\frac{m}{n}}\right)\right]\right\rangle_{p(x)}$$

となる．ここで，恒等式

$$\ln\left[2\cosh\left(\beta x\sqrt{\frac{m}{n}}\right)\right]=\beta|x|\sqrt{\frac{m}{n}}+\ln\left(1+e^{-2\beta|x|\sqrt{m/n}}\right)$$

を利用すれば，$\beta\to\infty$ の極限として $f=-\sqrt{n/m}$ と評価できる．これは，(3.1)の関数形と $c_g=1$ を意味する．

$K=2$ の符号

$K\geq 2$ の場合には符号語の生成にビット間相関が発生するので，スピングラス理論を用いた解析を実行する必要がある（Murayama and Davis, 2006）．しかし，$K=2$ の符号では，比較的容易に自由エネルギーの形式

$$\begin{aligned}\sqrt{C}f=&-\frac{2\ln 2}{\sqrt{\alpha}}-\frac{\sqrt{\alpha}}{2}\left[1-\langle\tanh^2 x\rangle^2_{\pi(x)}\right]\\&+\sigma^2\sqrt{\alpha}\left[1-\langle\tanh^2 x\rangle_{\pi(x)}\right]-\frac{2}{\sqrt{\alpha}}\langle\ln\cosh x\rangle_{\pi(x)}\end{aligned}$$

が導ける．ここで，秩序変数 $\hat{\pi}(\hat{x})$ の分散 $\sigma^2=\alpha\,\langle\hat{x}^2\rangle_{\hat{\pi}(\hat{x})}$ の定義式とエントロピー消失条件(3.3)が成立する．また，各平均操作も前述のように定義した．この結果から(3.1)の関数形と c_g の値が自己コンシステントに導ける．ただし，これは平均場近似の精度内での記述である．

$K\to\infty$ の符号

$K\to\infty$ の漸近論では，積分に有意な項だけを抽出する操作が容易になっている．結局，次の方程式に議論は帰着する．

$$\begin{aligned}\sqrt{L}f&=-\frac{\sqrt{\alpha_c}}{2}-\frac{R}{\sqrt{\alpha_c}}\ln 2,\\0&=-\frac{1}{2}+\frac{R}{\alpha_c}\ln 2.\end{aligned}$$

ここで，$\alpha_c=\beta^2 L$ という変数を定義した．これより直ちに(3.1)という関数形と，$c_g=\sqrt{2\ln 2}$ という値が求まる．これは前章で紹介したレート・歪み関数を前提とした議論と一致している．

5. おわりに

本稿では，低密度符号による分散符号化を題材に，情報理論と統計力学の整合性を検討した．その結果，大自由度観測系に特徴的な現象を数理的に発見することに成功し，しかも情報理論と統計力学が矛盾しない結果を与える事実を確認できた．このように，特定の分野における可解モデルを詳細に分析することで，一見異なる手法の整合性を検証できることは興味深い．また，最適戦略が転移するノイズレベル p_c の存在は，伝送率を任意に設定した場合の一般論へ発展できる可能性を示唆している．このように，システムモデルを極端に単純化する見返りとして，非自明な現象の存在が理論的に予見できるのが物理科学的アプローチの特徴である．

参 考 文 献

Berger, T., Zhang, Z. and Viswanathan, H. (1996). The CEO problem, *IEEE Transactions on Information Theory*, **42**, 887–902.

Chung K. L. (2000). *A Course in Probability Theory*, Academic Press, New York.

Cover, T. M. and Thomas, J. A. (1991). *Elements of Information Theory*, John Wiley & Sons, New Jersey.

Dotsenko, V. (2001). *Introduction to the Replica Theory of Disordered Statistical Systems*, Cambridge University Press, Cambridge.

Hays, W. (1994). *Statistics*, Wadsworth Publishing, Belmont.

Lang, S. (1986). *A First Course in Calculus*, Springer-Verlag, Berlin.

MacKay, D. J. C. (2003). *Information Theory, Inference and Learning Algorithms*, Cambridge University Press, Cambridge.

Murayama, T. and Davis, P. (2006). Rate distortion codes in sensor networks: A system-level analysis, *Advances in Neural Information Processing Systems*, **18**, 931–938.

Murayama, T. and Okada, M. (2003). Rate distortion function in the spin glass state: A toy model, *Advances in Neural Information Processing Systems*, **15**, 423–430.

Sourlas, N. (1989). Spin-glass models as error-correcting codes, *Nature*, **339**, 693–695.

232 Proceedings of the Institute of Statistical Mathematics Vol. 57, No. 2, 221–232 (2009)

Statistical Mechanics of Sensing and Coding

Tatsuto Murayama

NTT Communication Science Laboratories, NTT Corporation

Today very many and small sensing devices are produced and connected to the computer network. This can be considered as a new kind of information infrastructure. Therefore the potential demand for advanced and robust communication techniques for such systems seems to be increasing. In particular, the theoretical bound for such communications has crucial importance. In this paper, we consider a new approach to the problem, based on notions and techniques of information theory and statistical mechanics.

Key words: Sensor networks, coding, statistical mechanics.