Kullback-Leibler 情報量と Sanov の定理

黒木玄

2016 年 6 月 16 日作成*

https://genkuroki.github.io/documents/20160616KullbackLeibler.pdf

古い版　古い版を以下の場所で公開した:

https://genkuroki.github.io/documents/20160616KullbackLeibler/

Ver.0.1 は 10 ページしかなかった.

目 次

*最新版は下記 URL からダウンロードできる. 飽きるまで継続的に更新と訂正を続ける予定である. 6 月 16 日 Ver.0.1(10 頁). 数時間かけて 10 頁ほど書いた. ((中略)) 9 月 12 日 Ver.0.21. この更新記録を大幅に削除した. 更新の歴史については公開した古い版を参照して欲しい. 9 月 14 日 Ver.0.22(71 頁). Poisson 分布から Kullback-Leibler 情報量や多項分布の中心極限定理を出すことに関する第 1.9 節を追加した. 9 月 14 日 Ver.0.22a: 古い版へのリンクの場所を変えた. 9 月 14 日 Ver.0.23(70 頁): 第 1 節の最初に Stirling の公式の簡単な証明を付け加えた. なぜか allowdisplaybreaks を入れたら 1 頁減った. 9 月 29 日 Ver.0.24(72 頁): Mcmillan の不等式とその応用に関する第 10 節を追加した. 10 月 10 日 Ver.0.24a(73 頁): 第 10 節の終わりの方を微小に修正. 2017 年 4 月 28 日 Ver.0.25(73 頁): 第 10.2 節における「エントロピー」を「情報量」に直した. 6 月 3 日 Ver.0.25a(73 頁): Mathtodon で教えて頂いた誤植を訂正した. ありさん、どうもありがとうございました. 6 月 11 日 Ver.0.25b(73 頁): リンク先を少し変えた. 6 月 12 日 Ver.0.25c(73 頁): 微修正. 6 月 13 日 Ver.0.25d(73 頁): 微修正. 6 月 17 日 Ver.0.25e(73 頁): 微修正.

0 はじめに

このノートは次のノートの続編である:

「ガンマ分布の中心極限定理と Stirling の公式」というタイトルの雑多なノート
https://genkuroki.github.io/documents/20160501StirlingFormula.pdf

このノートで使用する Stirling の公式についてはそのノートを見て欲しい. この雑多なノートは「タイトルにいつわりあり」の雑多な内容のノートになっている.

このノートの目標は Kullback-Leibler 情報量 (相対エントロピーの -1 倍) および Boltzmann 因子 $\exp(-\sum_\nu \beta_\nu f_\nu(k))$ で記述されるカノニカル分布が必然的に出て来る理由を説明することである[1]. 最初の方では直観的な説明を重視し, 数学的に厳密な議論は行なわない. 測度論の詳細が必要な議論もしない[2]. 第 3, 4 節において可能な範囲内で数学的に厳密な証明を行なう.

以下の文献などを参考にした.

参考文献

[1] Csiszar, Imre. A simple proof of Sanov's theorem. Bull Braz Math Soc, New Series 37(4), 453–459, 2006.
http://www.emis.ams.org/journals/em/docs/boletim/vol374/v37-4-a2-2006.pdf

[2] Csiszár, Imre. Axiomatic characterizations of information measures. Entropy, 2008, 10, 261–273. http://www.mdpi.com/1099-4300/10/3/261/pdf

[3] Cover, M. Thomas and Thomas, Joy A. Elements of Information Theory. Second Edition, John Wiley & Sons, Inc., 2006, xxiii+748 pages. (Google で検索)

[1] インターネット上での日本語による検索結果を眺めたところ, Kullback-Leibler 情報量 (相対エントロピーの -1 倍) について「2 つの確率分布の "距離" を表わす量」「2 つの確率分布の違いを表わす量」のように説明しただけですませているものが目立ち, Kullback-Leibler 情報量が自然に出て来るシンプルな理由を十分に説明しているものを見付けることができなかったのでこの解説ノートを書くことにした. Kullback-Leibler 情報量が必然的に出て来る理由は多項分布の $n \to \infty$ での漸近挙動に Kullback-Leibler 情報量が自然に出て来るからである. そのことから, $n \to \infty$ のときの経験分布の挙動を Kullback-Leibler 情報量で記述可能になる. その結果の数学的に厳密な定式化は Sanov の定理と呼ばれている. この解説ノートを書いたもう一つの理由は, Boltzmann 因子, カノニカル分布が出て来る理由を多項分布の $n \to \infty$ での漸近挙動 (もしくは Sanov の定理) に基づいて分かり易く説明している日本語の解説をインターネット上に見付けることができなかったことである. この解説ノートでは Boltzmann 因子 $e^{-\beta E_i}$ が出て来る理由も詳しく説明する.

[2] 主に有限集合上の確率分布を扱う.

[4] Dembo, Amir and Zeitouni, Ofer. Large Deviations Techniques and Applications. Stochastic Modelling and Applied Probability (formerly: Applications of Mathematics), 38, Second Edition, Springer, 1998, 396 pages. (Google で検索)

[5] Ellis, Richard, S. The theory of large deviations and applications to statistical mechanics. Lecture notes for École de Physique Les Houches, August 5–8, 2008, 123 pages.
http://people.math.umass.edu/~rsellis/pdf-files/Les-Houches-lectures.pdf

[6] 奥村晴彦. R で楽しむ統計 (Wonderful R 1). 共立出版 (2016/9/8), 208 頁.
https://github.com/okumuralab/RforFun

[7] Sanov, I. N. On the probability of large deviations of random variables. English translation of Matematicheskii Sbornik, 42(84):1, pp. 11–44. Institute of Statistics Mimeograph Series No. 192, March, 1958.
http://www.stat.ncsu.edu/information/library/mimeo.archive/ISMS_1958_192.pdf

[8] Suyari, Hiroki. Mathematical structure derived from the q-multinomial coefficient in Tsallis statistics. arXiv:cond-mat/0401546

[9] Suyari, Hiroki and Scarfone, Antonio Maria. α-divergence derived as the generalized rate function in Tsallis statistics. 信学技報, vol. 114, no. 138, IT2014-16, pp. 25–30, 2014 年 7 月. http://www.ieice.org/ken/paper/201407178BPp/

[10] 田崎晴明. 統計力学 I, II. 新物理学シリーズ, 培風館 (2008/12), 合計 525 ページ.
https://www.amazon.co.jp/dp/4563024376
https://www.amazon.co.jp/dp/4563024384

[11] Tim van Erven and Peter Harremoës. Rényi divergence and Kullback-Leibler divergence. arXiv:1206.2459

[12] Ramon van Handel. Lecture 3: Sanov's theorem. Stochastic Analytic Seminar (Princeton University), Blog Article, 10 October 2013.
https://blogs.princeton.edu/sas/2013/10/10/lecture-3-sanovs-theorem/

[13] Vasicek, Oldrich Alfonso. A conditional law of large numbers. Ann. Probab., Volume 8, Number 1 (1980), 142–147.
http://projecteuclid.org/euclid.aop/1176994830

1 多項分布から Kullback-Leibler 情報量へ

多項分布に Stirling の公式を単純に代入するだけで自然かつ容易に Kullback-Leibler 情報量 (もしくはその -1 倍の相対エントロピー) が現われることを説明したい.

準備 (Stirling の公式) Stirling の公式はガンマ函数について知っていれば以下のような計算で容易に証明される. 実際, $x = n + \sqrt{n}\,y = n(1 + y/\sqrt{n})$ による置換積分を実行すると

$$n! = \Gamma(n+1) = \int_0^\infty e^{-x}x^n\,dx = n^n e^{-n}\sqrt{n}\int_{-\sqrt{n}}^\infty e^{-\sqrt{n}\,y}(1+y/\sqrt{n})^n\,dy$$

となり,

$$\begin{aligned}\log(e^{-\sqrt{n}\,y}(1+y/\sqrt{n})^n) &= -\sqrt{n}\,y + n\log(1+y/\sqrt{n})\\ &= -\sqrt{n}\,y + n\left(y/\sqrt{n} - y^2/(2n) + O\left(1/(n\sqrt{n}\,)\right)\right)\\ &= -y^2/2 + O\left(1/\sqrt{n}\right)\end{aligned}$$

なので, $n \to \infty$ とすると[3],

$$\int_{-\sqrt{n}}^\infty e^{-\sqrt{n}\,y}(1+y/\sqrt{n})^n\,dy \longrightarrow \int_{-\infty}^\infty e^{-y^2/2}\,dy = \sqrt{2\pi}.$$

これで次の Stirling の公式が得られた:

$$n! = n^n e^{-n}\sqrt{2\pi n}(1+o(1)), \qquad \log n! = n\log n - n + \frac{1}{2}\log n + \log\sqrt{2\pi} + o(1).$$

1.1 母集団分布が q_i の多項分布

$q_i \geqq 0$, $\sum_{i=1}^r q_i = 1$ とする. 1 回の独立試行で状態 i が確率 q_i で得られる状況を考える. $q = (q_1, \ldots, q_r)$ を**母集団分布**と呼ぶことにする. そのような試行を n 回繰り返したとき, 状態 i が生じた回数を k_i と書く (k_i は確率変数である). そのとき状態 i が生じた割合 k_i/n (これを**経験分布**と呼ぶことにする) が $n \to \infty$ でどのように振る舞うかを調べよう.

これは, サイコロ (歪んでいてもよい) を n 回ふったときの i の目が出た割合の分布 (経験分布) が $n \to \infty$ でどのように振る舞うかを調べる問題だと言ってよい.

大数の法則によって $n \to \infty$ で $k_i/n \to q_i$ となるが, 後で条件付き確率を考えたいので母集団分布から離れた分布が経験分布として現われる確率がどのように減衰するかを知りたい. 第 2 節では条件付き確率を考えることによって Boltzmann 因子が得られることを説明する.

我々はこれから母集団分布 $q = (q_1, \ldots, q_r)$ を任意に固定し, 経験分布 $(k_1/n, \ldots, k_r/n)$ の確率分布を考え, その $n \to \infty$ での様子を調べることになる.

n 回の独立試行で状態 i が k_i 回得られる確率は, $\sum_{i=1}^r k_i = n$ のとき

$$\frac{n!}{k_1!\cdots k_r!}q_1^{k_1}\cdots q_r^{k_r} \tag{$*$}$$

になり, 他のとき 0 になる (多項分布).

$p_i \geqq 0$, $\sum_{i=1}^r p_i = 1$ と仮定する. n 回の独立試行で状態 i が得られた割合 k_i/n がほぼ p_i になるとき, 経験分布はほぼ p_i になると言うことにする.

---

[3] $e^{-\sqrt{n}\,y}(1+y/\sqrt{n})^n$ は $y < 0$ で n について単調増加, $y > 0$ で n について単調減少であることから極限と積分の交換を証明できる. Lebesgue の収束定理を使えば容易だが, 使わなくても容易である.

1.2 サンプルサイズを大きくしたときの多項分布の漸近挙動

$n \to \infty$ のとき経験分布がほぼ p_i になる確率がどのように振る舞うかを知りたい. そこで $n \to \infty$ のとき, k_i たちが

$$k_i = np_i + O(\log n) = np_i \left(1 + O\left(\frac{\log n}{n}\right)\right) \tag{$**$}$$

を満たしていると仮定し, 上の確率 $(*)$ がどのように振る舞うかを調べよう. この仮定のもとで $\log(k_i/n) = \log p_i + O((\log n)/n)$ が成立することに注意せよ[4].

Stirling の公式と $\sum_{i=1}^r k_i = n$ より

$$\begin{aligned}
&\log n! = n\log n - n + O(\log n) = \sum_{i=1}^r k_i \log n - \sum_{i=1}^r k_i + O(\log n),\\
&\log k_i! = k_i \log k_i - k_i + O(\log k_i) = k_i \log k_i - k_i + O(\log n),\\
&\log q_i^{k_i} = k_i \log q_i.
\end{aligned}$$

これらを上の確率 $(*)$ の対数に代入すると k_i の項はキャンセルする. さらに $(**)$ を代入すると次が得られる:

$$\begin{aligned}
\log\left(\frac{n!}{k_1!\cdots k_r!} q_1^{k_1}\cdots q_r^{k_r}\right) &= -n\sum_{i=1}^r \frac{k_i}{n}\left(\log\frac{k_i}{n} - \log q_i\right) + O(\log n)\\
&= -n\sum_{i=1}^r p_i(\log p_i - \log q_i) + O(\log n)\\
&= -n\sum_{i=1}^r p_i \log\frac{p_i}{q_i} + O(\log n).
\end{aligned}$$

同様の計算を区分求積法を用いた高校レベルの計算で実行することもできる (第 1.7 節).

1.3 Kullback-Leibler 情報量と相対エントロピーの定義

第 1.2 節の結果は

$$D(p||q) = \sum_{i=1}^r p_i \log\frac{p_i}{q_i}$$

とおくと次のように書き直される:

$$\log\left(\frac{n!}{k_1!\cdots k_r!} q_1^{k_1}\cdots q_r^{k_r}\right) = -nD(p||q) + O(\log n).$$

左辺は経験分布 k_i/n がほぼ p_i になる確率の対数を意味していることに注意せよ. $D(p||q)$ を **Kullback-Leibler 情報量** (カルバック・ライブラー情報量) もしくは **Kullback-Leibler divergence** と呼ぶ. Kullback-Leibler 情報量の -1 倍

$$S(p||q) = -D(p||q) = -\sum_{i=1}^r p_i \log\frac{p_i}{q_i}$$

[4] Taylor 展開 $\log(1+x) = x - x^2/2 + x^3/3 - x^4/4 + \cdots$ より.

を**相対エントロピー**と呼ぶことにする. 相対エントロピーは本質的に n が大きなときの「母集団分布が q_i のとき経験分布がほぼ p_i となる確率の対数の n 分の 1」である.

対数を取る前の公式は次の通り:

(n 回の独立試行で経験分布がほぼ p_i になる確率) $= \exp(-nD(p||q) + O(\log n))$.

もしも $D(p||q) > 0$ ならば, n を十分に大きくすれば $O(\log n)$ の項は $nD(p||q)$ の項と比較して無視できる量になるので, この確率は $\exp(-nD(p||q))$ の部分でほぼ決まっていると考えてよい.

1.4 Kullback-Leibler 情報量の基本性質

Kullback-Leibler 情報量 $D(p||q) = \sum_{i=1}^{r} p_i \log(p_i/q_i)$ は函数 $f(x) = x \log x$ を用いて, $D(p||q) = \sum_{i=1}^{r} f(p_i/q_i)q_i$ と表わされるので, $D(p||q)$ の $p = (p_1, \ldots, p_r)$ の函数としての性質を調べるためには函数 $f(x) = x \log x$ の性質を調べればよい. $f'(x) = \log x + 1$, $f''(x) = 1/x > 0$ なので函数 $f(x)$ は下に狭義凸である. ゆえに函数 $f(x)$ はその接線の函数で下から押さえられる. 特に $f(x) \geqq f(1) + f'(1)(x-1) = x - 1$ (等号の成立と $x = 1$ は同値). ゆえに

$$D(p||q) = \sum_{i=1}^{r} f\left(\frac{p_i}{q_i}\right) q_i \geqq \sum_{i=1}^{r} \left(\frac{p_i}{q_i} - 1\right) q_i = 0,$$

等号の成立は $p_i = q_i$ $(i = 1, \ldots, r)$ と同値.

さらに $f(x)$ が下に狭義凸であることより, $D(p||q)$ も p の函数として下に狭義凸であることもわかる.

このように Kullback-Leibler 情報量の値は 0 以上になり, 最小値 0 が実現することと分布 p_i が母集団分布 q_i に等しくなることは同値である. ゆえに, 分布 p_i が母集団分布 q_i に等しくないとき, $D(p||q) > 0$ となるので, 経験分布がほぼ p_i になる確率は $n \to \infty$ で n について指数函数的に 0 に収束する. したがって, $n \to \infty$ で経験分布 k_i/n は母集団分布 q_i に近付く. これは**大数の法則**の成立を意味している.

Kullback-Leibler 情報量は母集団分布 q_i のもとで分布 p_i が経験分布としてどれだけ確率的に実現し難いかを表わしている. 異なる分布が実現する確率の比は $n \to \infty$ で Kullback-Leibler 情報量の差の $-n$ 倍の指数函数のように振る舞う. ゆえに Kullback-Leibler 情報量がほんの少しでも違っていれば, Kullback-Leibler 情報量がより大きな方の分布は相対的にほとんど生じないということもわかる. ゆえに, ある条件を課して分布 p_i が生じる条件付き確率を考える場合には, 課した条件のもとで Kullback-Leibler 情報量が最小になる分布に経験分布は (条件付き確率の意味で) 近付くことになる (**条件付き大数の法則**, **条件付き極限定理**). この法則を**最小 Kullback-Leibler 情報量の原理**と呼ぶ. n が非常に大きなとき, ある条件のもとで経験的に実現される分布は課した条件のもとで Kullback-Leibler 情報量が最小の分布になる.

相対エントロピーは Kullback-Leibler 情報量の -1 倍だったので, 条件付きで分布 p_i が経験的に生じる確率を考える場合には課した条件のもとで相対エントロピーが最大になる分布に経験分布が近付くことになる. この言い換えを**最大相対エントロピーの原理**と呼ぶ. n が大きなとき、ある条件のもとで経験的に実現される分布は課した条件のもとで相対エントロピーが最大になるような分布である.

補足. 説明の簡素化のために条件 B が成立しているとき条件 A が常に成立していると仮定する. このとき, 条件 A のもとで条件 B が成立する確率 (条件付き確率) は, 条件 B が成立する確率を条件 A が確率で割ったものと定義される. このように条件付き確率は確率の商で定義される. だから, 確率の商が $n \to \infty$ でどのように振る舞うかを確認できれば, 条件付き確率がどのように振る舞うかがわかる. 上の議論ではこの考え方を使った.

1.5 二項分布の場合の計算例

$r = 2, q_1 = q, q_2 = 1-q$ の「コイン投げ」(もしくは「丁半博打」) の場合を考える. この場合に多項分布は二項分布になる. このとき, $p_1 = p, p_2 = 1-p$ とおくと, Kullback-Leibler 情報量は次のように表わされる:

$$D(p||q) = p \log \frac{p}{q} + (1-p) \log \frac{1-p}{1-q}.$$

これは $p = q$ で最小値 0 になり, p が q から離れれば離れるほど大きくなる. Kullback-Leibler 情報量は分布の経験的な生じ難さを表わす量なので q から遠い p ほど経験的に生じ難くなる. しかも p が経験的に生じる確率は $n \to \infty$ で $\exp(-nD(p||q) + O(\log n))$ と振る舞う. ゆえに, 複数の p の生じる確率を比較すると, $D(p||q)$ が相対的に大きな p が生じる確率は $n \to \infty$ で比の意味で相対的に 0 に近付く. 以上を踏まえた上で次の問題について考えよう.

問題 n は非常に大きいと仮定する. n 回のコイン投げの結果表が出た割合が a 以上になったとする. このとき表の割合はどの程度になるだろうか?

大数の法則より, $n \to \infty$ で表の割合は q に近付く. ゆえに $0 \leqq a < q$ のとき, 表の割合が a 以上であるという条件は $n \to \infty$ で常に実現することになる. だから, $0 \leqq a < q$ のとき, 表の割合が a 以上の場合に制限**しても**, n が大きければ表の割合はほぼ q に等しくなっていると考えられる.

問題は $q < a \leqq 1$ の場合である. そのとき, n が大きくなればなるほど, 表の割合が a 以上になる確率は 0 に近付く. 上の問題は表の割合が a 以上になる場合に制限したときに表の割合がほぼ p になる確率 (条件付き確率) がどのように振る舞うかという問題になる. この場合には上で計算した Kullback-Leibler 情報量が役に立つ. $p \geqq a$ という条件のもとでの $D(p||q)$ の最小値は $p = a$ で実現される. ゆえに条件付き大数の法則より, $n \to \infty$ で経験分布は $p = a$ に近付く. $q < a \leqq 1$ のとき, 表の割合が a 以上の場合に制限**すると**, n が大きければ表の割合はほぼ a に等しくなっていると考えられる.

以上の結果から以下の公式が成立していることもわかる:

$$\lim_{n\to\infty} \frac{1}{n} \log \sum_{k/n \geqq a} \binom{n}{k} q^k (1-q)^{n-k} = -\inf_{p \geqq a} D(p||q) = \begin{cases} -D(q||q) = 0 & (0 \leqq a \leqq q), \\ -D(a||q) & (q < a \leqq 1). \end{cases}$$

対数を使わない形式でこの公式を書き下すと,

$$\sum_{k/n \geqq a} \binom{n}{k} q^k (1-q)^{n-k} = \exp\left(-n \inf_{p \geqq a} D(p||q) + o(n) \right).$$

左辺は表の割合が a 以上になる確率である. $n \to \infty$ のとき確率には $D(p||q)$ が最小になる分布だけが強く効いて来る.

1.6 max-plus 代数への極限や Laplace の方法との関係

実数または $-\infty$ の a, b に対して演算

$$(a, b) \mapsto \max\{a, b\}, \qquad (a, b) \mapsto a + b$$

を考えたもの (半環 (semiring), 半体 (semifield) と呼ばれている) を **max-plus 代数**と呼ぶ. (max-plus 代数は**超離散化**や **tropical mathematics** や各種**正値性を扱う問題**などに登場する重要な "代数" である. 体は加減剰余が自由にできる "代数" のことであるが, 半体は加乗除は自由にできるが引算は自由にできない "代数" のことである. 引算が自由にできなくても意味のある面白い数学を作れる.)

大雑把には, max は 0 以上の実数の足算に対応しており, $+$ は掛算に対応していて, $-\infty$ は足算の単位元 0 に対応している. その対応は log を取って極限を取ることによって与えられる. すなわち, 次の公式が成立している:

$$\lim_{n\to\infty} \frac{1}{n} \log(e^{na} + e^{nb}) = \max\{a, b\}, \qquad \lim_{n\to\infty} \frac{1}{n} \log(e^{na} e^{nb}) = a + b.$$

後者は自明である. 前者の公式は次のようにして確かめられる. $a \geqq b$ と仮定すると, $b - a \leqq 0$ となるので, $e^{n(b-a)}$ は有界になり,

$$\frac{1}{n} \log(e^{an} + e^{nb}) = \frac{1}{n} \log\left(e^{na}\left(1 + e^{n(b-a)}\right)\right) = a + \frac{1}{n} \log\left(1 + e^{n(b-a)}\right) \to a \quad (n \to \infty)$$

となる. これで前者の公式も示された.

より一般に次が成立している:

$$\lim_{n\to\infty} \frac{1}{n} \log \sum_{i=1}^{r} \exp(na_i + O(\log n)) = \max\{a_1, \ldots, a_r\}.$$

このように $\exp(na_i + O(\log n))$ のように振る舞う量の和の対数の $1/n$ 倍では $n \to \infty$ のとき最大の a_i の部分のみが効いて来る. 対数を使わない方の公式を書き下すと,

$$\sum_{i=1}^{r} \exp(na_i + O(\log n)) = \exp(n \max\{a_1, \ldots, a_r\} + o(n)) \qquad (n \to \infty).$$

これは積分の場合の Laplace の方法の類似であるとみなされる.

積分の場合は次の通り. 適切な設定のもとで次が成立している:

$$\int_\alpha^\beta \exp\Big(-nf(x) + O(\log n)\Big)\, dx = \exp\left(-n \inf_{\alpha\leqq x\leqq\beta} f(x) + o(n)\right) \qquad (n \to \infty).$$

$f(x)$ が $\alpha < x = x_0 < \beta$ で一意的な最小値を持ち, $f''(x_0) > 0$ ならば,

$$\int_\alpha^\beta e^{-nf(x)} g(x)\, dx = e^{-nf(x_0)} g(x_0) \sqrt{\frac{2\pi}{nf''(x_0)}}\,(1 + o(1)) \qquad (n \to \infty).$$

このような漸近挙動の計算の仕方は **Laplace の方法**と呼ばれている.

1.7 区分求積法による高校レベルの計算で KL 情報量を出す方法

多項分布の $n \to \infty$ での漸近挙動を以下のようにして, 区分求積法を使った高校数学っぽい方法で調べることもできる.

$q_i \geqq 0$, $\sum_{i=1}^r q_i = 1$ とし, 非負の整数 a, b_i は $\sum_{i=1}^r b_i = a$ をみたしているとし,

$$p_i = \frac{b_i}{a} = \frac{Nb_i}{Na}$$

とおく. このとき

$$\lim_{N\to\infty} \frac{1}{Na} \log \left(\frac{(Na)!}{(Nb_1)! \cdots (Nb_r)!} q_1^{Nb_1} \cdots q_r^{Nb_r} \right) = -\sum_{i=1}^r p_i \log \frac{p_i}{q_i}. \tag{$*$}$$

これの右辺は相対エントロピー (Kullback-Leibler 情報量の -1 倍) である. すなわち

$$\lim_{N\to\infty} \left(\frac{(Na)!}{(Nb_1)! \cdots (Nb_r)!} q_1^{Nb_1} \cdots q_r^{Nb_r} \right)^{1/(Na)} = \frac{1}{(p_1/q_1)^{p_1} \cdots (p_r/q_r)^{p_r}}.$$

区分求積法でこれを証明してみよう. 公式 $(*)$ を示せばよい. $N \to \infty$ のとき

$$\begin{aligned}
&\frac{1}{Na} \log \left(\frac{(Na)!}{(Nb_1)! \cdots (Nb_r)!} q_1^{Nb_1} \cdots q_r^{Nb_r} \right) \\
&= \frac{1}{Na} \left(\sum_{k=1}^{Na} \log k - \sum_{i=1}^r \sum_{k=1}^{Nb_i} \log k + \sum_{i=1}^r Nb_i \log q_i \right) \\
&= \frac{1}{Na} \left(\sum_{k=1}^{Na} \log \frac{k}{Na} - \sum_{i=1}^r \sum_{k=1}^{Nb_i} \log \frac{k}{Na} + \sum_{i=1}^r Nb_i \log q_i \right) \\
&= \frac{1}{Na} \sum_{k=1}^{Na} \log \frac{k}{Na} - \sum_{i=1}^r \frac{1}{Na} \sum_{k=1}^{Nb_i} \log \frac{k}{Na} + \sum_{i=1}^r p_i \log q_i \\
&\to \int_0^1 \log x \, dx - \sum_{i=1}^r \int_0^{p_i} \log x \, dx + \sum_{i=1}^r p_i \log q_i \\
&= [x \log x - x]_0^1 - \sum_{i=1}^r [x \log x - x]_0^{p_i} + \sum_{i=1}^r p_i \log q_i = -\sum_{i=1}^r p_i \log \frac{p_i}{q_i}.
\end{aligned}$$

2 つ目の等号で括弧の内側に $Na \log(Na) - \sum_{i=1}^r Nb_i \log(Na) = 0$ を挿入した. それによって区分求積法を適用できる形に変形できた.

以上の結果は次が成立することを意味している: $N \to \infty$ のとき

$$(Na \text{ 回の試行で経験分布が } p_i = b_i/a \text{ になる確率})^{1/(Na)} \to \frac{1}{(p_1/q_1)^{p_1} \cdots (p_r/q_r)^{p_r}}.$$

1.8 Kullback-Leibler 情報量と多項分布の中心極限定理の関係

この部分節は連続ツイート

https://twitter.com/genkuroki/status/773390919450132481

の内容をまとめ直したものである.

基本になる公式の導出 $q_i, p_i > 0$, $\sum_{i=1}^r q_i = \sum_{i=1}^r p_i = 1$, k_i は正の整数で $\sum_{i=1}^r k_i = n$ であるとする. 多項分布における確率

$$\frac{n!}{k_1! \cdots k_r!} q_1^{k_1} \cdots q_r^{k_r}$$

が, $k_i = np_i + \varepsilon_i$, $p_i = q_i + x_i/\sqrt{n}$, $\varepsilon_i = o(\sqrt{n})$ のとき[5], $n \to \infty$ でどのように振る舞うかを調べたい. そこで階乗 $n!$, $k_i!$ に Stirling の公式

$$n! = n^n e^{-n} \sqrt{2\pi n}(1 + O(1/n)), \qquad k_i! = k_i^{k_i} e^{-k_i} \sqrt{2\pi k_i}(1 + O(1/n))$$

を代入すると

$$\frac{n!}{k_1! \cdots k_r!} q_1^{k_1} \cdots q_r^{k_r} = \frac{n^n e^{-n} \sqrt{2\pi n}(1 + O(1/n))}{k_1^{k_1} e^{-k_1} \sqrt{2\pi k_1} \cdots k_r^{k_r} e^{-k_r} \sqrt{2\pi k_r}}.$$

分子の e^{-n} と分母の e^{-k_i} たちは $\sum_{i=1}^r k_i = n$ よりキャンセルして消える. $n^n = n^{k_1} \cdots n^{k_r}$ と $k_i = (k_i/n)n$ を代入して整理すると

$$\frac{n!}{k_1! \cdots k_r!} q_1^{k_1} \cdots q_r^{k_r} = \left(\frac{k_1/n}{q_1}\right)^{-k_1} \cdots \left(\frac{k_r/n}{q_r}\right)^{-k_r} \frac{1 + O(1/n)}{\sqrt{(2\pi n)^{r-1}(k_1/n) \cdots (k_r/n)}}. \qquad (*)$$

この公式が以下の議論の基本になる. この公式 $(*)$ をよく眺めれば多項分布の中心極限定理と Kullback-Leibler 情報量の関係がわかる[6].

多項分布の中心極限定理 多項分布の多次元正規分布による近似を得るためには

$$k_i = np_i + \varepsilon_i, \qquad p_i = q_i + \frac{x_i}{\sqrt{n}}, \qquad \varepsilon_i = o(\sqrt{n})$$

を $(*)$ に代入すればよい[7]. $\sum_{i=1}^r k_i = n$ と $\sum_{i=1}^r q_i = \sum_{i=1}^r p_i = 1$ より, $\sum_{i=1}^r \varepsilon_i = 0$, $\sum_{i=1}^r x_i = 0$ となり,

$$\begin{aligned}
&\frac{k_i}{n} = q_i \left(1 + \frac{x_i}{\sqrt{n}\, q_i} + \frac{\varepsilon_i}{nq_i}\right) = q_i(1 + o(1)), \qquad k_i = nq_i + \sqrt{n}\, x_i + \varepsilon_i, \\
&\log \left(\frac{k_i/n}{q_i}\right)^{-k_i} = -(nq_i + \sqrt{n}\, x_i + \varepsilon_i) \log \left(1 + \frac{x_i}{\sqrt{n}\, q_i} + \frac{\varepsilon_i}{nq_i}\right) \\
&\quad = -(nq_i + \sqrt{n}\, x_i + \varepsilon_i) \left(\frac{x_i}{\sqrt{n}\, q_i} + \frac{\varepsilon_i}{nq_i} + \frac{1}{2}\left(\frac{x_i}{\sqrt{n}\, q_i} + \frac{\varepsilon_i}{nq_i}\right)^2 + o\left(\frac{1}{n}\right)\right) \\
&\quad = -\sqrt{n}\, x_i - \varepsilon_i + \frac{x_i^2}{2q_i} + o(1)
\end{aligned}$$

なので, $\sum_{i=1}^r (\sqrt{n}\, x_i + \varepsilon_i) = 0$ より,

$$\frac{n!}{k_1! \cdots k_r!} q_1^{k_1} \cdots q_r^{k_r} = \frac{\exp\left(\frac{1}{2} \sum_{i=1}^r \frac{x_i^2}{q_i}\right)}{\sqrt{(2\pi n)^{r-1} q_1 \cdots q_r}} \times (1 + o(1)).$$

[5] 実際には $|\varepsilon_i| \leqq 1/2$ (特に $\varepsilon_i = O(1)$) に取れる.
[6] 公式 $(*)$ に n が大きなときの多項分布の様子に関する情報がほぼすべて含まれていると考えてよい.
[7] 実際には $|\varepsilon_i| \leqq 1/2$ に取れる.

ゆえに $dk_i = \sqrt{n}\,dx_i$ たちを両辺に沿えると,

$$\frac{n!}{k_1!\cdots k_r!}q_1^{k_1}\cdots q_r^{k_r}\,dk_1\cdots dk_{r-1} = \frac{\exp\left(\frac{1}{2}\sum_{i=1}^{r}\frac{x_i^2}{q_i}\right)}{\sqrt{(2\pi)^{r-1}q_1\cdots q_r}}\,dx_1\cdots dx_{r-1}\times(1+o(1)).$$

$\sum_{i=1}^{r} k_i = 1$, $\sum_{i=1}^{r} x_i = 0$ であることに注意せよ. この結果は多項分布が n が大きいときに多次元正規分布で近似できることを意味している (多項分布の中心極限定理).

KL 情報量の導出 Kullback-Leibler 情報量を得るためには, $(*)$ の両辺の対数を取って,

$$k_i = np_i + o(n) = np_i(1+o(1))$$

を代入して $o(n)$ の項を無視すればよい. $k_i = np_i + o(n) = np_i(1+o(1))$ のとき

$$\log\left(\frac{k_i/n}{q_i}\right)^{-k_i} = -(np_i+o(n))\log\left(\frac{p_i(1+o(1))}{q_i}\right) = -np_i\log\frac{p_i}{q_i} + o(n)$$

でかつ $\log n = o(n)$ なので

$$\log\left(\frac{n!}{k_1!\cdots k_r!}q_1^{k_1}\cdots q_r^{k_r}\right) = -n\sum_{i=1}^{r}p_i\log\frac{p_i}{q_i} + o(n) = -nD(p||q) + o(n).$$

ここで $D(p||q)$ は Kullback-Leibler 情報量である.

KL 情報量と多項分布の中心極限定理の関係 もしもぴったり $k_i = np_i$ ならば Kullback-Leibler 情報量は $(*)$ の右辺の因子

$$\left(\frac{k_1/n}{q_1}\right)^{-k_1}\cdots\left(\frac{k_r/n}{q_r}\right)^{-k_r}$$

の対数の $-1/n$ 倍に一致する. $k_i = np_i + o(n)$ の場合には, Kullback-Leibler 情報量はこの因子の対数の $-1/n$ 倍の $n\to\infty$ での極限に一致する. そして多項分布の多次元正規分布による近似の指数函数部分もこの因子から得られるのであった. したがって多項分布を近似する多次元正規分布の指数函数部分の対数の -1 倍

$$\frac{1}{2}\sum_{i=1}^{r}\frac{x_i^2}{q_i}$$

は Kullback-Leibler 情報量から得られるはずである. Kullback-Leibler 情報量 $D(p||q)$ は $p_i = q_i$ のとき最小値 0 になるのであった. その点で $nD(p||q)$ を Taylor 展開した結果の 2 次の部分から多項分布を近似する多次元正規分布の確率密度函数の指数函数部分の対数の -1 倍が得られる. そのことを確認しよう. まず $\sum_{i=1}^{r} x_i = 0$ と制限せずに $nD(p||q)$ に $p_i = q_i + x_i/\sqrt{n}$ を代入して, x_i たちについて展開すると

$$nD(p||q) = \sum_{i=1}^{r}(nq_i + \sqrt{n}\,x_i)\log\left(1+\frac{x_i}{\sqrt{n}\,q_i}\right)$$

$$= \sum_{i=1}^{r} (nq_i + \sqrt{n}\, x_i) \left(\frac{x_i}{\sqrt{n}\, q_i} - \frac{x_i^2}{2nq_i^2} + o\left(\frac{1}{n}\right) \right)$$
$$= \sum_{i=1}^{r} \left(\sqrt{n}\, x_i + \frac{x_i^2}{2q_i} + o(1) \right).$$

ここで $\sum_{i=1}^{r} x_i = 0$ を使うと,

$$nD(p||q) = \frac{1}{2} \sum_{i=1}^{r} \frac{x_i^2}{q_i} + o(1).$$

このことから多項分布の多次元正規分布による近似は p_i たちが q_i たちに近いときに Kullback-Leibler 情報量の Taylor 展開の 3 次以上の項を無視することに相当することがわかる.

Pearson のカイ 2 乗統計量との関係　多項分布における

$$\sum_{i=1}^{n} \frac{((i \text{ の個数の観測値}) - (i \text{ の個数の期待値}))^2}{(i \text{ の個数の期待値})} = \sum_{i=1}^{r} \frac{(k_i - nq_i)^2}{nq_i}$$

を Pearson のカイ 2 乗統計量と呼ぶ. これに $k_i = nq_i + \sqrt{n}\, x_i + o(\sqrt{n})$ を代入して整理すると,

$$\sum_{i=1}^{r} \frac{(k_i - nq_i)^2}{nq_i} = \sum_{i=1}^{r} \frac{x_i^2}{q_i} + o(1).$$

ゆえに n が大きいとき, Pearson のカイ 2 乗統計量は多項分布を近似する多次元正規分布にしたがう x_i たちに関する $\sum_{i=1}^{r} x_i^2/q_i$ で近似される. 多項分布を近似する多次元正規分布の確率密度函数の指数函数部分は $\exp\left(-(1/2)\sum_{i=1}^{r} x_i^2/q_i\right)$ の形をしているのであった. このことから, n が大きいとき, Pearson のカイ 2 乗統計量はカイ 2 乗分布に近似的にしたがうことがわかる. 多項分布 (r 項分布) を近似する多次元正規分布の確率密度函数は条件 $\sum_{i=1}^{r} x_i = 0$ で定義される $r-1$ 次元の台を持つので, そのカイ 2 乗分布の自由度は $r-1$ になる.

注意 (カイ 2 乗分布とは). $Z_1, \ldots, Z_s$ は標準正規分布にしたがう独立な確率変数であるとする. 各々の Z_i について $f(Z_i)$ の期待値は

$$E[f(Z_i)] = \int_{\mathbb{R}} f(z_i) \frac{e^{-z_i^2/2}}{\sqrt{2\pi}}\, dz_i$$

と表わされる. このとき確率変数 $Y = \sum_{i=1}^{s} Z_i^2$ がしたがう確率分布を自由度 s の**カイ二乗分布**と呼ぶ. カイ二乗分布における期待値は次のように表わされる:

$$E[f(Y)] = \int_0^{\infty} f(y) \frac{e^{-y/2} y^{s/2-1}}{\Gamma(s/2) 2^{s/2}}\, dy.$$

実際,

$$E[f(Y)] = \text{const.} \int_{\mathbb{R}^s} f\left(\sum_{i=1}^{s} z_i^2 \right) e^{-\sum_{i=1}^{s} z_i^2/2}\, dz_1 \cdots dz_s$$

$$= \text{const.} \int_0^\infty f(r^2) e^{-r^2/2} r^{s-1}\, dr = \text{const.} \int_0^\infty f(y) e^{-y/2} y^{s/2-1}\, dy.$$

2 つ目の等号で $r^2 = \sum_{i=1}^s z_i^2$ とおき, 積分を r と球面上の積分に書き変えた. 球面の面積は半径の $s-1$ 乗に比例するので r^{s-1} の因子が出る. 球面上の積分を実行し, 出て来た定数を const. に繰り込んだ. 3 つ目の等号で $y = r^2$ とおいた. 定数倍 const. は全確率の総和が 1 になるという条件から自動的に決まる.

$A = [a_{ij}]$ は固有値がすべて正の s 次実対称行列であるとし, その逆行列を $A^{-1} = [b_{ij}]$ と書くことにする. $(X_1, \ldots, X_s)$ は (s 次元の台を持つ) 確率密度函数

$$\frac{\exp\left(-\frac{1}{2}\sum_{i,j=1}^s b_{ij} x_i x_j\right)}{\sqrt{\det(2\pi A)}}$$

が定める多次元正規分布に従う確率変数であるとする. このとき, 確率密度函数の指数函数部分の対数の -2 倍に対応する確率変数

$$Y = \sum_{i,j=1}^s b_{ij} X_i X_j$$

も自由度 s のカイ二乗分布にしたがう. すなわち, **台の次元が s の多次元正規分布の確率密度函数の指数函数部分の対数の -2 倍に対応する確率変数は自由度 s のカイ二乗分布にしたがう.** このことからカイ 2 乗分布は一般の多次元正規分布に付随する (最も) 基本的な確率分布であると考えることができる.

上の Y が自由度 s のカイ二乗分布にしたがう理由は以下の通り. 本質的に正値実対称行列の線形代数である. 実対称行列 A はある直交行列 $U = [u_{ij}]$ と対角行列 $D = \mathrm{diag}(\alpha_1, \ldots, \alpha_s)$ によって $A = UDU^{-1} = UDU^T$ と表わされる. A の固有値 α_i はすべて正なので $\sqrt{D} = \mathrm{diag}(\sqrt{\alpha_1}, \ldots, \sqrt{\alpha_s})$, $C = U\sqrt{D}$ とおくと, $A = CC^T$ となる. そのとき $A^{-1} = (C^{-1})^T C^{-1}$ なので

$$Y = \sum_{i,j=1}^s b_{ij} X_i X_j = X^T A^{-1} X = (C^{-1}X)^T (C^{-1}X).$$

ここで X は確率変数 X_i を第 i 成分とする列ベクトルである. すなわち, 確率変数の列ベクトル $Z = [Z_i]$ を $Z = C^{-1}X$ と定めると,

$$Y = Z^T Z = \sum_{i=1}^s Z_i^2.$$

Z_i たちが独立で各々が標準正規分布にしたがうことを示せれば, Y が自由度 s のカイ 2 乗分布にしたがうことがわかる. そのためには Z_i たちの分散共分散行列が単位行列になることを示せば十分である. Z_i たちの分散共分散行列の定義は $E[ZZ^T]$ であり, X_i たちの分散共分散行列は $E[XX^T] = A = CC^T$ なので

$$E[ZZ^T] = E[(C^{-1}X)(C^{-1}X)^T] = C^{-1}E[XX^T](C^T)^{-1} = C^{-1}CC^T(C^T)^{-1} = E.$$

これで示すべきことがすべて示された. □

1.9 Poisson 分布と Kullback-Leibler 情報量の関係

この部分節は文献 [6] の p.80 におけるポアソン分布を使った説明の数学的正当化に関する

https://twitter.com/genkuroki/status/775549559405555713

以降の連続ツイートの内容をまとめ直したものである.

Poisson 分布 $\lambda > 0$ であるとする. 確率変数 K がパラメーター λ の Poisson 分布にしたがうとは, $k = 0, 1, 2, \dots$ に対する $K = k$ の確率が

$$P(K = k) = e^{-\lambda}\frac{\lambda^k}{k!}$$

となっていることである. 確率の総和が 1 になることは

$$\sum_{k=0}^{\infty} P(K = k) = e^{-\lambda}\sum_{k=0}^{\infty}\frac{\lambda^k}{k!} = e^{-\lambda}e^{\lambda} = 1$$

と確かめられる. Poisson 分布は二項分布の極限として得られる. $0 < q < 1$ に対する二項分布の確率 $n!/(k!(n-k)!)\cdot q^k(1-q)^{n-k}$ に $q = \lambda/n$ を代入すると,

$$\frac{n(n-1)\cdots(n-k+1)}{k!}\left(\frac{\lambda}{n}\right)^k\left(1-\frac{\lambda}{n}\right)^{n-k} = \left(1-\frac{\lambda}{n}\right)^n\frac{\lambda^k}{k!}\times A_n,$$
$$A_n = \frac{(1-0/n)(1-1/n)\cdots(1-(k-1)/n)}{(1-\lambda/n)^k} \to 1 \qquad (n\to\infty)$$

となる. これで, 二項分布は $nq = \lambda$ が一定のままで $n \to \infty$ とすると Poisson 分布に収束することがわかった. Poisson 分布の確率 $P(K = k)$ は単位時間で k 回のイベントが起こる確率であると解釈される. Poisson 分布の期待値 $E[K]$ と分散 $E[K^2] - E[K]^2$ は次のように計算される:

$$E[K] = \sum_{k=0}^{\infty} kP(K = k) = e^{-\lambda}\sum_{k=0}^{\infty} k\frac{\lambda^k}{k!} = e^{-\lambda}\sum_{k=1}^{\infty}\frac{\lambda^k}{(k-1)!} = e^{-\lambda}\lambda e^{\lambda} = \lambda,$$
$$E[K(K-1)] = e^{-\lambda}\sum_{k=2}^{\infty}\frac{\lambda^k}{(k-2)!} = e^{-\lambda}\lambda^2 e^{\lambda} = \lambda^2$$
$$E[K^2] - E[K]^2 = E[K] + \lambda^2 - \lambda^2 = E[K] = \lambda.$$

Poisson 分布のモーメント母函数 $E[e^{tK}$ は次のように計算される:

$$E[e^{tK}] = e^{-\lambda}\sum_{k=0}^{\infty} e^{tk}\frac{\lambda^k}{k!} = e^{-\lambda}e^{\lambda e^t} = e^{\lambda(e^t-1)}.$$

特に Poisson 分布はパラメーター λ について再生性を持つ[8]. ゆえに, 中心極限定理より, K_λ がパラメーター λ の Poisson 分布にしたがう確率変数であるとき, $(K_\lambda - \lambda)/\sqrt{\lambda}$ は

[8] 独立な確率変数 X, Y がそれぞれパラメーター λ_X, λ_Y の Poisson 分布にしたがうとき, $X + Y$ はパラメーター $\lambda_X + \lambda_Y$ の Poisson 分布にしたがう.

$\lambda \to \infty$ で標準正規分布にしたがう確率変数に弱収束する. したがって, 独立な確率変数たち K_{i,λ_i} がそれぞれパラメーター λ_i の Poisson 分布にしたがうとき,

$$\sum_{i=1}^{s} \frac{(K_{i,\lambda_i} - \lambda_i)^2}{\lambda_i}$$

は λ_i たちを同意に大きくする極限で自由度 s のカイ 2 乗分布にしたがう確率変数に弱収束する. $O_i = K_{i,\lambda_i}$ を観測度数と解釈し, $E_i = \lambda_i$ を期待度数と解釈すると, 上の量は

$$\sum_{i=1}^{s} \frac{(O_i - E_i)^2}{E_i}$$

と書き直され, よく Pearson のカイ二乗統計量として使われるスタイルになる.

Poisson 分布の中心極限定理 Poisson 分布の中心極限定理の Stirling の定理を使った直接的証明の概略を説明しよう. $f(x)$ は $\mathbb{R}$ 上の実数値有界連続函数であるとする. K_λ はパラメーター λ の Poisson 分布にしたがう確率変数であるとし, $X_\lambda = (K_\lambda - \lambda)/\sqrt{\lambda}$ とおく. このとき

$$\begin{aligned}
E[f(X_\lambda)] &= e^{-\lambda} \sum_{k=0}^{\infty} f\left(\frac{k-\lambda}{\sqrt{\lambda}}\right) \frac{\lambda^k}{k!} = \sum_{x \in \mathbb{Z}_{\geqq -\lambda/\sqrt{\lambda}}} f(x) \frac{\lambda^{\lambda+\sqrt{\lambda}x} e^{-\lambda}}{(\lambda+\sqrt{\lambda}\,x)!} \\
&= \sum_{x \in \mathbb{Z}_{\geqq -\lambda/\sqrt{\lambda}}} f(x) \frac{\lambda^{\lambda+\sqrt{\lambda}x} e^{-\lambda}(1+O(1/\lambda)))}{(\lambda+\sqrt{\lambda}\,x)^{\lambda+\sqrt{\lambda}x} e^{-\lambda-\sqrt{\lambda}x} \sqrt{2\pi(\lambda+\sqrt{\lambda}\,x)}} \\
&= \sum_{x \in \mathbb{Z}_{\geqq -\lambda/\sqrt{\lambda}}} f(x) \left(1+\frac{x}{\sqrt{\lambda}}\right)^{-(\lambda+\sqrt{\lambda}x)} e^{\sqrt{\lambda}x} \frac{1+O(1/\lambda)}{\sqrt{2\pi(1+x/\sqrt{\lambda})}} \frac{1}{\sqrt{\lambda}} \\
&= \sum_{x \in \mathbb{Z}_{\geqq -\lambda/\sqrt{\lambda}}} f(x) \frac{e^{-x^2/2}(1+O(1/\sqrt{\lambda}))}{\sqrt{2\pi(1+x/\sqrt{\lambda})}} \frac{1}{\sqrt{\lambda}} \\
&\longrightarrow \int_{-\infty}^{\infty} f(x) \frac{e^{-x^2/2}}{\sqrt{2\pi}}\, dx \qquad (\lambda \to \infty).
\end{aligned}$$

3 つ目の等号で Stirling の公式を使った. 5 つ目の等号で

$$\begin{aligned}
&-(\lambda+\sqrt{\lambda}\,x)\log\left(1+\frac{x}{\sqrt{\lambda}}\right) + \sqrt{\lambda}\,x = -(\lambda+\sqrt{\lambda}\,x)\left(\frac{x}{\sqrt{\lambda}} - \frac{x^2}{2\lambda} + O\left(\frac{1}{\lambda\sqrt{\lambda}}\right)\right) + \sqrt{\lambda}\,x \\
&= -\sqrt{\lambda}\,x + \frac{x^2}{2} - x^2 + \sqrt{\lambda}\,x + O\left(\frac{1}{\sqrt{\lambda}}\right) = -\frac{x^2}{2} + O\left(\frac{1}{\sqrt{\lambda}}\right)
\end{aligned}$$

を使った.

Poisson 分布から多項分布へ $q_i > 0$, $\sum_{i=1}^{r} q_i = 1$ と仮定する. 多項分布における確率は

$$\frac{n!}{k_1! \cdots k_r!} q_1^{k_1} \cdots q_r^{k_r} = n! \prod_{i=1}^{r} \frac{q_i^{k_i}}{k_i!}$$

と書ける．ただし, k_i たちの動く範囲は $\sum_{i=1}^r k_i = n$ に制限されている．したがって, $\lambda_i = nq_i$ とおくと,

$$n! \prod_{i=1}^r \frac{q_i^{k_i}}{k_i!} = \frac{n!}{n^n e^{-n}} \prod_{i=1}^r \left(e^{-nq_i} \frac{(nq_i)^{k_i}}{k_i!} \right) = \frac{n!}{n^n e^{-n}} \prod_{i=1}^r \left(e^{-\lambda_i} \frac{\lambda_i^{k_i}}{k_i!} \right).$$

この式中の積の因子はパラメーター $\lambda_i = nq_i$ の Poisson 分布の確率の式の形をしている．このことから, 多項分布は独立な Poisson 分布の積を $\sum_{i=1}^r k_i = n$ で制限したものになっていることがわかる.

単独の Poisson 分布から KL 情報量の各項へ　単独の Poisson 分布から Kullback-Leibler 情報量の「各項」が出て来ることを説明しよう．パラメーター λ の Poisson 分布の確率に Stirling の公式を適用すると,

$$e^{-\lambda} \frac{\lambda^k}{k!} = \frac{e^{-\lambda} \lambda^k}{k^k e^{-k} \sqrt{2\pi k}(1 + O(1/k))} = \left(\frac{k}{\lambda} \right)^{-k} e^{k-\lambda} \frac{1 + O(1/k)}{\sqrt{2\pi k}}.$$

この形の式が基本になる．この式の右辺の最初の二つの因子の積の対数は

$$\log \left(\left(\frac{k}{\lambda} \right)^{-k} e^{k-\lambda} \right) = -k \log \frac{k}{\lambda} + k - \lambda.$$

これはほとんど Kullback-Leibler 情報量の各項の形をしている．さらにそれらしく見えるようにするためには, $k = np$, $\lambda = nq$ を代入して

$$-k \log \frac{k}{\lambda} + k - \lambda = -n \left(p \log \frac{p}{q} - (p - q) \right)$$

と変形すればよい．右辺の括弧の中に Kullback-Leibler 情報量の各項の形の式が現れている．$p \log(p/q)$ は $p = q$ で最小にならないが, そこから $p - q$ を引いた結果は $p = q$ で最小になる．Poisson 分布の中心極限定理は $-k \log(k - \lambda) + k - \lambda$ の $k = \lambda$ における Taylor 展開

$$-k \log \frac{k}{\lambda} + k - \lambda = -\frac{1}{1 \cdot 2} \frac{(k-\lambda)^2}{\lambda} + \frac{1}{2 \cdot 3} \frac{(k-\lambda)^3}{\lambda^2} - \frac{1}{3 \cdot 4} \frac{(k-\lambda)^4}{\lambda^3} + \cdots$$

を二次の項までで切る近似をすればよい．より正確に述べると, この Taylor 展開に $k = \lambda + \sqrt{\lambda}\, x$ を代入して, $\lambda \to \infty$ の極限を考える．$k = \lambda + \sqrt{\lambda}\, x$ のとき

$$-k \log \frac{k}{\lambda} + k - \lambda \longrightarrow -\frac{x^2}{2} \qquad (\lambda \to \infty).$$

このことから, $k = \lambda + \sqrt{\lambda}\, x$ のとき, $k/\lambda = 1 + x/\sqrt{\lambda}$, $dk = \sqrt{\lambda}\, dx$ より

$$e^{-\lambda} \frac{\lambda^k}{k!}\, dk = \left(\frac{k}{\lambda} \right)^{-k} e^{k-\lambda} \frac{1 + O(1/k)}{\sqrt{2\pi k}}\, dk \longrightarrow \frac{e^{-x^2}}{\sqrt{2\pi}}\, dx \qquad (\lambda \to \infty)$$

となることもわかる．これで Poisson 分布の中心極限定理は, Poisson 分布から得られる KL 情報量の各項の Taylor 展開を二次までで切る近似をすることによって得られることがわかった．$-k \log(k/\lambda) + k - \lambda$ の $k = \lambda$ における Taylor 展開の二次の項に現われる $(k - \lambda)^2/\lambda$ は $O =$ (観測度数) $= k$, $E =$ (期待度数) $= \lambda$ とおくと

$$\frac{(k-\lambda)^2}{\lambda} = \frac{(\text{(観測度数)} - \text{(期待度数)})^2}{\text{(期待度数)}} = \frac{(O-E)^2}{E}$$

と統計学の教科書でよく見るスタイルで書ける．Poisson 分布の中心極限定理から, E が大きなとき, $(O - E)^2/E$ は近似的に自由度 1 のカイ二乗分布に従うことがわかる.

複数の Poisson 分布の積から KL 情報量へ 単独の Poisson 分布から Kullback-Leibler 情報量の各項が得られたので, 複数の Poisson 分布の積から Kullback-Leibler 情報量そのものが得られる. 単独の Poisson 分布に関する計算より

$$\prod_{i=1}^{r}\left(e^{-\lambda_i}\frac{\lambda_i^{k_i}}{k_i!}\right)=\prod_{i=1}^{r}\left(\left(\frac{k_i}{\lambda_i}\right)^{-k_i}e^{k_i-\lambda_i}\frac{1+O(1/k_i)}{\sqrt{2\pi k_i}}\right).$$

そして, $k_i=np_i$, $\lambda_i=nq_i$ とおくと

$$\begin{aligned}\log\prod_{i=1}^{r}\left(\left(\frac{k_i}{\lambda_i}\right)^{-k_i}e^{k_i-\lambda_i}\right)&=-\sum_{i=1}^{r}\left(k_i\log\frac{k_i}{\lambda_i}+k_i-\lambda_i\right)\\&=-n\sum_{i=1}^{r}\left(p_i\log\frac{p_i}{q_i}-(p_i-q_i)\right).\end{aligned}$$

ゆえに $\sum_{i=1}^{r}p_i=\sum_{i=1}^{r}q_i$ と仮定すると

$$\log\prod_{i=1}^{r}\left(\left(\frac{k_i}{\lambda_i}\right)^{-k_i}e^{k_i-\lambda_i}\right)=-n\sum_{i=1}^{r}p_i\log\frac{p_i}{q_i}.$$

右辺はちょうど Kullback-Leibler 情報量の $-n$ 倍の形をしている. Kullback-Leibler 情報量は多項分布の確率の $n\to\infty$ での漸近挙動から得られるのであった. そのことを「多項分布が Poisson 分布の積を $\sum_{i=1}^{r}k_i=n$ という条件で制限することによって得られること」を用いて示すこともできる. Poisson 分布の積と多項分布の関係より, $q_i>0$, $\sum_{i=1}^{r}q_i=1$, $\lambda_i=nq_i$, $\sum_{i=1}^{r}k_i=n$ とすると,

$$\begin{aligned}\frac{n!}{k_1!\cdots k_r!}q_1^{k_1}\cdots q_r^{k_r}&=\frac{n!}{n^ne^{-n}}\prod_{i=1}^{r}\left(e^{-\lambda_i}\frac{\lambda_i^{k_i}}{k_i!}\right)\\&=\sqrt{2\pi n}(1+O(1/n))\prod_{i=1}^{r}\frac{(k_i/\lambda_i)^{-k_i}e^{k_i-\lambda_i}}{\sqrt{2\pi k_i}(1+O(1/k_i))}.\end{aligned}$$

Stirling の公式を用いた. この式が基本になる. さらに $k_i=np_i$ とおくと,

$$\frac{n!}{k_1!\cdots k_r!}q_1^{k_1}\cdots q_r^{k_r}=\frac{1+O(1/n)}{\sqrt{(2\pi n)^{r-1}\prod_{i=1}^{r}p_i}}\prod_{i=1}^{r}\left(\frac{p_i}{q_i}\right)^{-np_i}.$$

したがって,

$$\log\left(\frac{n!}{k_1!\cdots k_r!}q_1^{k_1}\cdots q_r^{k_r}\right)=-n\sum_{i=1}^{r}p_i\log\frac{p_i}{q_i}+O(\log n).$$

右辺に Kullback-Leibler 情報量 $D(p||q)=\sum_{i=1}^{r}p_i\log(p_i/q_i)$ が現れた.

複数の Poisson 分布の積から多項分布の中心極限定理へ 上の設定をそのまま引き継ぐ.

$$\frac{n!}{k_1!\cdots k_r!}q_1^{k_1}\cdots q_r^{k_r}=\sqrt{2\pi n}(1+O(1/n))\prod_{i=1}^{r}\frac{(k_i/\lambda_i)^{-k_i}e^{k_i-\lambda_i}}{\sqrt{2\pi k_i}(1+O(1/k_i))},$$

$$\log((k_i/\lambda_i)^{-k_i}e^{k_i-\lambda_i}) = -k_i\log(k_i/\lambda_i) + k_i - \lambda_i = -\frac{1}{2}\frac{(k_i-\lambda_i)^2}{\lambda_i} + \frac{1}{6}\frac{(k_i-\lambda_i)^3}{\lambda_i^2} - \cdots$$

より,

$$k_i = \lambda_i + \sqrt{\lambda_i}\,x_i = nq_i + \sqrt{nq_i}\,x_i, \qquad x_i = \frac{k_i-\lambda_i}{\sqrt{\lambda_i}} = \frac{k_i - nq_i}{\sqrt{nq_i}}$$

とおくと, $dk_i = \sqrt{nq_i}\,dx_i$ なので

$$\frac{n!}{k_1!\cdots k_r!}q_1^{k_1}\cdots q_r^{k_r}\,dk_1\cdots dk_{r-1} \longrightarrow \frac{\exp\left(-\frac{1}{2}\sum_{i=1}^r x_i^2\right)}{\sqrt{(2\pi)^{r-1}q_r}}\,dx_1\cdots dx_{r-1} \qquad (n\to\infty).$$

このように Poisson 分布を経由して多項分布の中心極限定理を出すこともできる.

以上のような見方をすると, **多項分布よりも Poisson 分布の方が基本的な分布である**ように見えて来る. 多項分布は複数の Poisson 分布の積を $\sum_{i=1}^r k_i = n$ という条件で制限することによって得られる.

2 条件付き大数の法則から Boltzmann 因子へ

条件付き大数の法則 (最小 Kullback-Leibler 情報量の原理, 最大相対エントロピーの原理) から Boltzmann 因子で記述される分布が自然に得られることを説明したい.

2.1 問題の設定

母集団分布が $q = (q_1, \ldots, q_r)$ の多項分布の設定に戻る.

n 回の独立試行によって各々の i について状態 i が生じた割合 k_i/n がほぼ p_i に等しいとき, 経験分布がほぼ $p = (p_1, \ldots, p_r)$ に等しくなると言うことにする. その確率について

$$(n \text{ 回で経験分布がほぼ } p \text{ になる確率}) = \exp(-nD(p||q) + O(\log n)) \qquad (n\to\infty)$$

が成立しているのであった.

次の問題を考える: 分布 $p = (p_1, \ldots, p_r)$ に s 個の条件

$$\sum_{i=1}^r f_{\nu,i}p_i \approx c_\nu \qquad (\nu = 1, 2, \ldots, s) \tag{$*$}$$

を課す. ただし, $\mathbb{R}^r$ のベクトルたち $(1,1,\ldots,1), (f_{\nu,1},\ldots,f_{\nu,r})$ $(\nu = 1,\ldots,s)$ は一次独立であると仮定しておく. 経験分布がこの条件を満たす分布 p にほぼ等しい場合に制限したとき, 経験分布の確率分布は $n\to\infty$ でどのように振る舞うか?

たとえば, 状態 i のエネルギーが E_i の場合に

$$\sum_{i=1}^r E_i p_i \approx U$$

という条件 (すなわちエネルギーの経験的平均値がほぼ U に等しくなっているという条件) を課したとき, 経験分布が $n\to\infty$ でどのように振る舞うか?

たとえば, サイコロを振って i の目が出たら, 賞金を E_i ペリカもらえるとき,

$$\sum_{i=1}^{r} E_i p_i \approx U$$

という条件 (すなわち 1 回あたりの賞金の経験的平均値がほぼ U ペリカに等しくなっているという条件) を課したとき, 経験分布が $n \to \infty$ でどのように振る舞うか?

以上の 2 つの例では $s = 1$ である. 複数の条件を課せば $s > 1$ となる.

2.2 Boltzmann 因子の導出

条件 $(*)$ のもとでの経験分布の条件付き確率は $n \to \infty$ で, 条件 $\sum_{i=1}^{r} p_i = 1$ と条件 $(*)$ のもとで Kullback-Leibler 情報量 $D(p||q) = \sum_{i=1}^{r} p_i \log(p_i/q_i)$ が最小値になる分布 $p = (p_1, \ldots, p_r)$ に集中することになる.

その条件付き最小値問題を解くために Lagrange の未定乗数法を使おう. (Kullback-Leibler 情報量が p について下に狭義凸な函数であったことを思い出そう.) そのために

$$L = \sum_{i=1}^{r} p_i \log \frac{p_i}{q_i} + (\lambda - 1)\left(\sum_{i=1}^{r} p_i - 1\right) + \sum_{\nu=1}^{s} \beta_\nu \left(\sum_{i=1}^{r} f_{\nu,i} p_i - c_\nu\right)$$

とおく. ここで $\lambda - 1$, β_ν が未定乗数である. 未定乗数と p_i で L を偏微分した結果がすべて 0 になるという方程式

$$0 = \frac{\partial L}{\partial \lambda} = \sum_{i=1}^{r} p_i - 1, \tag{1}$$

$$0 = \frac{\partial L}{\partial \beta_\nu} = \sum_{i=1}^{r} f_{\nu,i} p_i - c_\nu \qquad (\nu = 1, \ldots, s), \tag{2}$$

$$0 = \frac{\partial L}{\partial p_i} = \log \frac{p_i}{q_i} + \lambda + \sum_{\nu=1}^{s} \beta_\nu f_{\nu,i} \qquad (i = 1, \ldots, r) \tag{3}$$

を解けばよい. (3) より,

$$p_i = \exp\left(-\lambda - \sum_{\nu=1}^{s} \beta_\nu f_{\nu,i}\right) q_i$$

これを (1) に代入すると,

$$Z := e^{\lambda} = \sum_{i=1}^{r} e^{-\sum_{\nu=1}^{s} \beta_\nu f_{\nu,i}} q_i, \qquad p_i = \frac{1}{Z} e^{-\sum_{\nu=1}^{s} \beta_\nu f_{\nu,i}} q_i \tag{4}$$

となることがわかる. この Z は**分配函数**と呼ばれる. このように p_i と $Z = e^{\lambda}$ は β_ν たちの函数になっている. β_ν たちは (4) を (2) に代入することによって決定される. $\exp\left(-\sum_{\nu=1}^{s} \beta_\nu f_{\nu,i}\right)$ を **Boltzmann 因子**と呼ぶことにする. Boltzmann 因子は母集団分布 q_i と条件付きの経験分布 p_i がどれだけ異なるかを記述している. このようにして求められた分布 p_i を**カノニカル分布**と呼ぶことにする.

条件 $(*)$ が成立している場合に制限した場合の経験分布は, $n \to \infty$ で以上で求めた分布 $p = (p_1, \ldots, p_r)$ に近付く (条件付き大数の法則より). n が巨大ならば経験分布はカノニカル分布の形をしているとしてよい.

たとえば $s=1$, $f_{1,i}=E_i$, $c_1=U$, $\beta_1=\beta$ のとき,

$$p_i=\frac{1}{Z}e^{-\beta E_i}q_i,\qquad Z=\sum_{i=1}^{r}e^{-\beta E_i}q_i,\qquad -\frac{\partial\log Z}{\partial\beta}=\frac{1}{Z}\sum_{i=1}^{r}E_ie^{-\beta E_i}q_i=U.$$

これらの公式は q_i たちが互いにすべて等しい場合には統計力学における Boltzmann 因子を用いた確率分布の記述に一致している.

カノニカル分布に対する相対エントロピー $S(p||q)=-D(p||q)=-\sum_{i=1}^{r}p_i\log(p_i/q_i)$ の別の表示を求めよう: $\log(p_i/q_i)=-\sum_{\nu=1}^{s}\beta_\nu f_{\nu,i}-\log Z$, $\sum_{i=1}^{r}p_i=1$, $\sum_{i=1}^{r}f_{\nu,i}p_i=c_\nu$ なので

$$S(p||q)=\sum_{\nu=1}^{s}\beta_\nu c_\nu+\log Z.$$

たとえば $s=1$, $f_{1,i}=E_i$, $c_1=U$, $\beta_1=\beta$ のとき

$$S(p||q)=\beta U+\log Z.$$

自由エネルギー F を $F=-\beta^{-1}\log Z$ と定義すると,

$$S(p||q)=\beta(U-F)$$

この公式は, Boltzmann 定数が含まれていない点を除けば, 統計力学を知っている人達にとってお馴染みの公式だろう[9].

2.3 母分布が連続型の場合から連続型の指数型分布族が得られること

母集団分布が確率密度函数 $q(x)$ で与えられている場合を考えよう. この場合には n 回の独立試行の結果得られる経験分布の確率密度函数がほぼ $p(x)$ になる確率の対数の $1/n$ 倍は $n\to\infty$ で

$$S(p||q)=-D(p||q)=-\int p(x)\log\frac{p(x)}{q(x)}\,dx$$

に近付くと考えられる. 分布 $p(x)$ に以下の条件を課す:

$$\int f_\nu(x)p(x)\,dx=c_\nu\qquad(\nu=1,\dots,s).$$

前節と同様にして, この条件のもとで $D(p||q)$ を最小にする確率密度函数 $p(x)$ を求めると次のようになることがわかる:

$$\begin{aligned}
p(x)&=\frac{1}{Z}e^{-\sum_{\nu=1}^{s}\beta_\nu f_\nu(x)}q(x),\\
Z&=\int e^{-\sum_{\nu=1}^{s}\beta_\nu f_\nu(x)}q(x)\,dx,\\
-\frac{\partial\log Z}{\partial\beta_\nu}&=\frac{1}{Z}\int f_\nu(x)e^{-\sum_{\nu=1}^{s}\beta_\nu f_\nu(x)}q(x)\,dx=c_\nu.
\end{aligned}$$

このようにな形の連続型確率分布の族を**連続型の指数型分布族**と呼ぶ. 積分が和の場合には**離散型の指数型分布族**と呼ばれる.

たとえば以下の確率分布はすべて指数型分布族に含まれている.

[9] Boltzmann 定数が 1 になる単位系を採用することもできる.

二項分布: $0<\theta<1$ のとき, $-\beta=\log\theta-\log(1-\theta)$ とおくと, $k=0,1,\dots,n$ について

$$p_k=\binom{n}{k}\theta^k(1-\theta)^{n-k}=\frac{e^{-\beta k}q_k}{Z},\qquad q_k=\binom{n}{k}\frac{1}{2^n},\qquad Z=\frac{1}{2^n(1-\theta)^n}.$$

この場合と条件付き大数の法則の関係については例 4.3 も参照せよ.

多項分布: $\theta_i\geqq 0$, $\theta_r>0$, $\sum_{i=1}^r\theta_i=1$ であるとし, $-\beta_i=\log\theta_i-\log\theta_r$ とおくと, $k_1+\cdots+k_r=n$ のとき

$$p_{k_1,\dots,k_r}=\frac{n!}{k_1!\cdots k_r!}\theta_1^{k_1}\cdots\theta_r^{k_r}=\frac{e^{-\sum_{i=1}^{r-1}\beta_i k_i}q_{k_1,\dots,k_r}}{Z},$$
$$q_{k_1,\dots,k_r}=\frac{n!}{k_1!\cdots k_r!}\frac{1}{r^n},\qquad Z=\frac{1}{r^n\theta_r^n}$$

正規分布:

$$p(x)=\frac{e^{-(x-\mu)^2/(2\sigma^2)}}{\sqrt{2\pi}}=\frac{e^{-(1/(2\sigma^2))x^2+(\mu/\sigma^2)x}}{Z},\qquad Z=e^{\mu^2/(2\sigma^2)}\sqrt{2\pi\sigma^2}.$$

$\mu=0$, $\sigma=1$ の場合については第 2.4 節も参照して欲しい. 正規分布の確率密度函数 $p(x)$ は平均 μ と分散 σ^2 を指定したときに, すなわち $\int_{\mathbb{R}}p(x)\,dx=1$,

$$\int_{\mathbb{R}}x\,p(x)\,dx=\mu,\qquad \int_{\mathbb{R}}x^2\,p(x)\,dx=\sigma^2+\mu^2$$

という条件のもとで, エントロピー

$$S(p)=-\int_{\mathbb{R}}p(x)\log p(x)\,dx$$

が最大になる $p(x)$ として特徴付けられる.

Gamma 分布: $x>0$ において

$$p(x)=\frac{e^{-x/\tau}x^{\alpha-1}}{\tau^\alpha\Gamma(\alpha)}=\frac{e^{-x/\tau+(\alpha-1)\log x}}{Z},\qquad Z=\tau^\alpha\Gamma(\alpha).$$

Gamma 分布の確率密度函数 $p(x)$ は $\int_{\mathbb{R}}p(x)\,dx=1$,

$$\int_0^\infty x\,p(x)\,dx=c_1,\qquad \int_0^\infty(\log x)\,p(x)\,dx=c_2$$

という条件のもとでエントロピー

$$S(p(x)]=-\int_0^\infty p(x)\log p(x)\,dx$$

が最大になる $p(x)$ として特徴付けられる. 以下も同様である.

第二種 Beta 分布: $x>0$ において

$$p(x)=\frac{1}{B(\alpha,\beta)}\frac{x^{\alpha-1}}{(1+x)^{\alpha+\beta}}=\frac{e^{(\alpha-1)\log x-(\alpha+\beta)\log(1+x)}}{Z},\qquad Z=B(\alpha,\beta).$$

自由度 n の t 分布を $1/\sqrt{n}$ でスケールしたもの: 自由度 n の t 分布の確率密度は

$$\rho(t)\,dt = \frac{1}{c_n}\left(1+\frac{t^2}{n}\right)^{-(n+1)/2} dt, \qquad c_n = \sqrt{n}B(1/2,n/2) = \frac{\sqrt{n\pi}\,\Gamma(n/2)}{\Gamma((n+1)/2)}$$

であった. $p(x)\,dx = \rho(\sqrt{n}\,x)\,d(\sqrt{n}\,x)$, $\beta = (n+1)/2$ とおくと

$$p(x) = \frac{1}{Z}\frac{1}{(1+x^2)^{(n+1)/2}} = \frac{e^{-\beta\log(1+x^2)}}{Z}, \qquad Z = B(1/2,n/2).$$

第一種 Beta 分布: $0<x<1$ について

$$p(x) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)} = \frac{e^{(\alpha-1)\log x+(\beta-1)\log(1-x)}}{Z}, \quad Z = B(\alpha,\beta).$$

Poisson 分布:

$$p_k = \frac{e^{-\lambda}\lambda^k}{k!} = \frac{e^{-(\log\lambda)k}q_k}{Z}, \quad q_k = \frac{e}{k!}, \quad Z = e^{\lambda+1}.$$

2.4 標準正規分布の導出例

例として $s=1$, $f_1(x)=x^2$, $c_1=1$, $q(x)=1$ の場合にどうなるかを計算してみよう[10]. この場合に上の結果は, n 回の独立試行の結果得られた x^2 の経験的期待値 $(x_1^2+\cdots+x_n^2)/n$ について

$$\frac{x_1^2+\cdots+x_n^2}{n} = 1$$

という条件を課したとき, $n\to\infty$ で x の経験的分布がどうなるかを求めることに等しい. 上の公式を使うと

$$p(x) = \frac{1}{Z}e^{-\beta x^2}, \qquad Z = \int_{\mathbb{R}} e^{-\beta x^2}\,dx = \sqrt{\pi}\beta^{-1/2}, \qquad -\frac{\partial\log Z}{\partial\beta} = \frac{1}{2\beta} = 1.$$

ゆえに $\beta=1/2$, $Z=\sqrt{2\pi}$, $p(x)=e^{-x^2/2}/\sqrt{2\pi}$ となる. すなわち $n\to\infty$ で得られる分布は標準正規分布になる.

この結果は $\mathbb{R}^n$ 内の半径の 2 乗が n の原点を中心とする $n-1$ 次元球面上の一様分布の 1 次元部分空間への射影が $n\to\infty$ で標準正規分布に収束することを意味している. すなわち次の公式が成立している:

$$\lim_{n\to\infty}\int_{\sqrt{n}\,S^{n-1}} f(x_1)\,\mu_n(dx) = \int_{\mathbb{R}} f(x)\frac{e^{-x^2/2}}{\sqrt{2\pi}}\,dx.$$

ここで $\sqrt{n}\,S^{n-1}$ は原点を中心とする半径 $\sqrt{n}$ の $n-1$ 次元球面

$$\{\,(x_1,\ldots,x_n)\in\mathbb{R}^n \mid x_1^2+\cdots+x_n^2 = n\,\}$$

を表わし, μ_n はその上の一様確率分布であり, $f(x_1)$ の x_1 は球面上の点 $(x_1,\ldots,x_n)$ の射影である. この極限の公式は通常の多変数の微積分の計算で直接に確認できる[11].

以上の計算例を見れば, 指数型分布族に属する他の確率分布がどのような条件を課したときに自然に現われるかも理解できると思う.

[10] $q(x)=1$ なのでこの場合に $q(x)$ は確率密度函数にならない. しかし, 以下の計算の結論は正しい.

[11] 次の雑多なノートの Maxwell-Boltzmann 則の節にその直接的な計算が書いてある.
http://www.math.tohoku.ac.jp/~kuroki/LaTeX/20160501StirlingFormula.pdf

3 多項分布の場合の Sanov の定理

多項分布の場合の Sanov の定理の主張を明確に述べて厳密に証明しておくことにする. Stirling の公式さえ使わない易しい証明を紹介する. この節の証明はブログ記事 [12] で解説されている証明と本質的に同じものである. そのブログには参考になる解説がたくさんある.

3.1 Sanov の定理の主張

有限集合 $\{1,2,\ldots,r\}$ 上の確率分布全体の集合を $\mathcal{P}$ と書く:

$$\mathcal{P} = \{\, p=(p_1,\ldots,p_r) \in \mathbb{R}^r \mid p_1,\ldots,p_r \geqq 0,\ p_1+\cdots+p_r = 1 \,\}.$$

$\mathcal{P}$ は $r-1$ 次元の閉単体である. たとえば $r=3$ のとき $\mathcal{P}$ は正三角形になる.

確率分布 $q=(q_1,\ldots,q_r) \in \mathcal{P}$ を任意に取って固定する. 確率変数 $X_1, X_2, \ldots$ は集合 $\{1,2,\ldots,r\}$ に値を持つ確率変数列であり, 独立で同分布 $q=(q_1,\ldots,q_r)$ にしたがっていると仮定する. $q=(q_1,\ldots,q_r)$ を**母集団分布**と呼ぶ.

集合 A に対してその元の個数を $\#A$ と書き, 条件 A が満たされる確率を $P(A)$ と書くことにする. (後で条件 A のもとでの B の条件付き確率を $P(B|A)$ と書く.)

各々の $i=1,\ldots,r$ に対して $X_1,\ldots,X_n$ に含まれる i の個数が k_i 個になる確率は

$$P\Big(\#\{\, k=1,2,\ldots,n \mid X_k = i \,\} = k_i \text{ for each } i=1,\ldots,r\Big) = \frac{n!}{k_1!\cdots k_r!} q_1^{k_1}\cdots q_r^{k_r}$$

となる. 可能な $(k_1,\ldots,k_r)$ の組合せは $k_i = 0,1,\ldots,n$, $k_1+\cdots+k_r=n$ を満たしていなければいけない. このような $(k_1,\ldots,k_r)$ に対する $(k_1/n,\ldots,k_r/n)$ 全体の集合を $\mathcal{P}_n \subset \mathcal{P}$ と書くことにする:

$$\mathcal{P}_n = \left\{ \left(\frac{k_1}{n},\ldots,\frac{k_r}{n}\right) \,\middle|\, k_i = 0,1,\ldots,n,\ k_1+\cdots+k_r = n \right\}.$$

このとき $\mathcal{P}_n$ の元の個数は $(n+1)^r$ 以下になる. ($\#\mathcal{P}_n \leqq (n+1)^r$ を後で自由に利用する.) $X_1,\ldots,X_n$ に対応する $\mathcal{P}_n$ の元 $P_n = (k_1/n,\cdots,k_r/n)$ を**経験分布**と呼ぶ. 経験分布 P_n は $\mathcal{P}_n$ に値を持つ確率変数である.

確率分布の組 $(p,q) \in \mathcal{P}^2$ の函数 $D(p||q)$ を次のように定める:

$$D(p||q) = \sum_{i=1}^{r} p_i \log \frac{p_i}{q_i}.$$

p_i や q_i が 0 になる場合には $0\log 0 = 0$, $-\log 0 = \infty$ という約束のもとで値を定めておく. $D(p||q)$ を **Kullback-Leibler 情報量**もしくは **Kullback-Leibler ダイバージェンス**と呼ぶ.

定理 3.1 (Sanov). 以上の設定のもとで以下が成立している[12]:

[12] $\liminf$, $\limsup$ について第 9 節に簡単な解説を書いておいた.

(1) A が $\mathcal{P}$ の開部分集合ならば

$$\liminf_{n\to\infty} \frac{1}{n} \log P(P_n \in A) \geqq - \inf_{p\in A} D(p||q).$$

(2) A が $\mathcal{P}$ の部分集合ならば[13]

$$\limsup_{n\to\infty} \frac{1}{n} \log P(P_n \in A) \leqq - \inf_{p\in A} D(p||q).$$

(3) $\mathcal{P}$ の部分集合 A の開核の閉包が A を含むならば

$$\lim_{n\to\infty} \frac{1}{n} \log P(P_n \in A) = - \inf_{p\in A} D(p||q).$$

このように経験分布の $n \to \infty$ での漸近挙動は Kullback-Leibler 情報量 $D(p||q)$ の inf で記述される. □

例 3.2 (二項分布の場合). $r = 2$ とし, $q_1 = q$, $q_2 = 1 - q$, $p_1 = p$, $p_2 = 1 - p$ とおくと,

$$D(p||q) = p \log \frac{p}{q} + (1 - p) \log \frac{1 - p}{1 - q}.$$

これは $p = q$ のとき最低値 0 になり, p が q から離れるとこれの値は減少する.

$0 \leqq a < b \leqq 1$ であるとし, $A = (a, b) = (a$ から b までの開区間) とおく. このとき

$$P(P_n \in A) = \sum_{a<k/n<b} \binom{n}{k} q^k (1 - q)^{n-k}$$

なので

$$\lim_{n\to\infty} \frac{1}{n} \log \sum_{a<k/n<b} \binom{n}{k} q^k (1 - q)^{n-k} = - \inf_{a<p<b} D(p||q) = \begin{cases} -D(b||q) & (b < q), \\ -D(q||q) = 0 & (a \leqq q \leqq b), \\ -D(a||q) & (q < a) \end{cases}$$

となる. これが Sanov の定理の非自明な応用の最も簡単な場合である. □

3.2 Sanov の定理の証明の準備

次の補題が後で Stirling の公式の代わりに使われる.

補題 3.3. 非負の整数 k, l に対して

$$\frac{l!}{k!} \geqq k^{l-k}.$$

[13] 我々が扱っている場合には A は任意の部分集合であっても問題ない. しかし, 無限次元の場合には A は閉部分集合だと仮定することが重要になるらしい.

証明. $l \geqq k$ のとき

$$\frac{l!}{k!} = (k+1)(k+2)\cdots l \geqq k^{l-k}.$$

$l \leqq k$ のとき

$$\frac{l!}{k!} = \frac{1}{(l+1)(l+2)\cdots k} \geqq \frac{1}{k^{k-l}} = k^{l-k}.$$

これで示すべきことが示された. □

次の補題が証明できれば Sanov の定理の証明は易しい. 次の補題の証明には Stirling の公式を使わない.

補題 3.4. 任意の $p \in \mathcal{P}_n$ に対して

$$\frac{1}{(n+1)^r} e^{-nD(p||q)} \leqq P(P_n = p) \leqq e^{-nD(p||q)}.$$

証明. $p = (p_1, \ldots, p_r) = (k_1/n, \ldots, k_r/n) \in \mathcal{P}_n$ のとき,

$$-nD(p||q) = -\sum_{i=1}^{r} k_i \log p_i + \sum_{i=1}^{r} k_i \log q_i,$$

$$e^{-nD(p||q)} = \frac{q_1^{k_1} \cdots q_r^{k_r}}{p_1^{k_1} \cdots p_r^{k_r}}, \qquad P(P_n = p) = \frac{n!}{k_1! \cdots k_r!} q_1^{k_1} \cdots q_r^{k_r}.$$

ゆえに, この補題の結論は次と同値である:

$$\frac{1}{(n+1)^r} \leqq \frac{n!}{k_1! \ldots k_r!} p_1^{k_1} \cdots p_r^{k_r} \leqq 1.$$

上からの評価の方 (右側の不等式) は多項分布の知識より自明である. (多項分布における確率が 1 以下であることを意味しているに過ぎない.) 以下で下からの評価 (左側の不等式) を証明しよう.

$l_i = 0, 1, \ldots, n$, $l_1 + \cdots + l_r = n$ と仮定する. このとき, $p_i = k_i/n$ なので

$$\frac{n!}{l_1! \ldots l_r!} p_1^{l_1} \cdots p_r^{l_r} \leqq \frac{n!}{k_1! \ldots k_r!} p_1^{k_1} \cdots p_r^{k_r} \tag{$*$}$$

が成立しているはずである. なぜならば多項分布において確率が最大になるのは経験分布 (今の場合は l_i/n) が母集団分布 (今の場合は $p_i = k_i/n$) に等しくなるときだからである. 実際, 補題 3.3 より,

$$\frac{(\text{右辺})}{(\text{左辺})} = \frac{l_1!}{k_1!} \cdots \frac{l_r!}{k_r!} \cdot k_1^{k_1-l_1} \cdots k_r^{k_r-l_r} \geqq k_1^{l_1-k_1} \cdots k_r^{l_r-k_r} \cdot k_1^{k_1-l_1} \cdots k_r^{k_r-l_r} = 1.$$

これで $(*)$ が証明された. ゆえに, 多項定理より

$$1 = \sum_{l_1+\cdots+l_r=n} \frac{n!}{l_1! \ldots l_r!} p_1^{l_1} \cdots p_r^{l_r} \leqq (n+1)^r \frac{n!}{k_1! \ldots k_r!} p_1^{k_1} \cdots p_r^{k_r}$$

両辺を $(n+1)^r$ で割れば下からの評価が得られる. □

注意 3.5. 以上の結果の一部分は以下のように拡張される.

$f(n)$ は $f(0)=0$ を満たす非負の整数 n の単調増加函数であるとし,

$$f(n)! = f(1)f(2)\cdots f(n), \qquad f(0)! = 1$$

と定める. これを f 階乗と呼ぶことにする. このとき非負の整数 k, l に対して,

$$\frac{f(l)!}{f(k)!} \geqq f(k)^{l-k}.$$

実際, $l \geqq k$ のとき

$$\frac{f(l)!}{f(k)!} = f(k+1)f(k+2)\cdots f(l) \geqq f(k)^{l-k}$$

となり, $l \leqq k$ のとき

$$\frac{l!}{k!} = \frac{1}{f(l+1)f(l+2)\cdots f(k)} \geqq \frac{1}{f(k)^{k-l}} = f(k)^{l-k}.$$

$k_i \in \mathbb{Z}_{\geqq 0}$ $p_i = f(k_i)/f(n)$ とおくと, $l_i \in \mathbb{Z}_{\geqq 0}$ のとき,

$$\frac{f(n)!}{f(l_1)!\cdots f(l_r)!} p_1^{l_1}\cdots p_r^{l_r} \leqq \frac{f(n)!}{f(l_1)!\cdots k_r)!} p_1^{k_1}\cdots p_r^{k_r}.$$

なぜならば

$$\begin{aligned}\frac{(\text{右辺})}{(\text{左辺})} &= \frac{f(l_1)!}{f(k_1)!}\cdots\frac{f(l_r)!}{f(k_r)!}\cdot f(k_1)^{k_1-l_1}\cdots f(k_r)^{k_r-l_r} \\ &\geqq f(k_1)^{l_1-k_1}\cdots f(k_r)^{l_r-k_r}\cdot f(k_1)^{k_1-l_1}\cdots f(k_r)^{k_r-l_r} = 1.\end{aligned}$$

$f(n)$ が特別な場合にはこの類似をさらにたどることができると思われる. □

3.3 Sanov の定理の証明

定理 3.1 の証明. 下からの評価 (1) を示そう. A は有限集合 $\{1,2,\ldots,r\}$ 上の確率分布全体の空間 $\mathcal{P}$ (これは $r-1$ 次元単体になる) の開部分集合であるとする. $\bigcup_{n=1}^{\infty}\mathcal{P}_n = \mathcal{P}\cap\mathbb{Q}^r$ は $\mathcal{P}$ の中で稠密である. A は $\mathcal{P}$ の開部分集合なので分布列 $p_n \in \mathcal{P}_n \cap A$ で

$$\lim_{n\to\infty} D(p_n||q) = \inf_{p\in A} D(p||q)$$

をみたすものを取れる. 以上の状況で

$$P(P_n \in A) = \sum_{p\in\mathcal{P}_n\cap A} P(P_n = p) \geqq P(P_n = p_n) \geqq \frac{1}{(n+1)^r} e^{-nD(p_n||q)}.$$

最後の不等号で補題 3.4 の下からの評価を使った. これより

$$\frac{1}{n}\log P(P_n \in A) \geqq -D(p_n||q) - \frac{r}{n}\log(n+1)$$

となることがわかる. したがって, $n \to \infty$ とすることによって,

$$\liminf_{n\to\infty} \frac{1}{n} \log P(P_n \in A) \geqq - \inf_{p\in A} D(p||q).$$

これで (1) が証明された.

上からの評価 (2) を示そう. A は有限集合 $\{1, 2, \ldots, r\}$ 上の確率分布全体の空間 $\mathcal{P}$ の任意の部分集合であるとする. このとき

$$P(P_n \in A) = \sum_{p\in \mathcal{P}_n \cap A} P(P_n = p) \leqq \sum_{p\in \mathcal{P}_n \cap A} e^{-nD(p||q)} \leqq (n+1)^r e^{-n \inf_{p\in A} D(p||q)}.$$

最初の不等号で補題 3.4 の上からの評価を使った. これより

$$\frac{1}{n} \log P(P_n \in A) \leqq - \inf_{p\in A} D(p||q) + \frac{r}{n} \log(n+1)$$

となることがわかる. したがって, $n \to \infty$ とすることによって,

$$\limsup_{n\to\infty} \frac{1}{n} \log P(P_n \in A) \leqq - \inf_{p\in A} D(p||q).$$

これで (2) が証明された.

(3) を示そう. A の開核を B と書き, B の閉包を C と書き, $A \subset C$ と仮定する. $B \subset A \subset C$ より $-\inf_{p\in B} D(p||q) \leqq -\inf_{p\in A} D(p||q) \leqq -\inf_{p\in C} D(p||q)$. C が B の閉包であること $D(p||q)$ が p の連続函数であることより, $-\inf_{p\in C} D(p||q) = -\inf_{p\in B} D(p||q)$. ゆえに $-\inf_{p\in B} D(p||q) = -\inf_{p\in A} D(p||q) = -\inf_{p\in C} D(p||q)$. したがって (1),(2) から (3) が導かれる.

これで定理 3.1 が証明された. □

注意 3.6. 以上の証明では階乗に関する Stirling の近似公式を使っていない. 証明で本質的に使った事柄は次の二つだけである.

(1) 上からの評価のために次の事実を使った:
$p_i \geqq 0$, $p_1 + \cdots + p_r = 1$ のとき

$$\frac{n!}{k_1! \cdots k_r!} p_1^{k_1} \cdots p_r^{k_r} \leqq 1 \quad (k_i \in \mathbb{Z}_{\geqq 0},\ k_1 + \cdots + k_r = n).$$

これは多項分布において「確率は 1 以下であること」を意味している. それを意味する不等式は, 左辺を k_i たちを動かして足し上げた結果が多項定理より 1 になること

$$\sum_{k_1+\cdots+k_r=n} \frac{n!}{k_1! \cdots k_r!} p_1^{k_1} \cdots p_r^{k_r} = (p_1 + \cdots + p_r)^r = 1$$

から, ただちに得られる.

(2) 下からの評価のために次の事実を使った:
$k_i \in \mathbb{Z}_{\geqq 0}$, $k_1 + \cdots + k_r = n$, $p_i = k_i/n$ のとき,

$$\frac{n!}{l_1! \cdots l_r!} p_1^{l_1} \cdots p_r^{l_r} \leqq \frac{n!}{k_1! \cdots k_r!} p_1^{k_1} \cdots p_r^{k_r} \quad (l_i \in \mathbb{Z}_{\geqq 0},\ l_1 + \cdots + l_r = n)$$

これは多項分布において「確率が最大になるのは分布が母集団分布に等しくなるときであること」を意味している. その不等式は次の易しい不等式 (k, l の大小関係によらずに成立している) からただちに得られる:

$$\frac{l!}{k!} \geqq k^{l-k} \qquad (k, l \in \mathbb{Z}_{\geqq 0}).$$

実際, この不等式を使うと, $p_i = k_i/n$ より

$$\frac{(\text{右辺})}{(\text{左辺})} = \frac{l_1!}{k_1!} \cdots \frac{l_r!}{k_r!} \frac{k_1^{k_1}}{k_1^{l_1}} \cdots \frac{k_r^{k_r}}{k_r^{l_r}} \geqq k_1^{l_1-k_1} \cdots k_r^{l_r-k_r} \cdot k_1^{k_1-l_1} \cdots k_r^{k_r-l_r} = 1.$$

以上の 2 つの結果は多項分布について知っていれば当然知っているはずの事柄である. たったそれだけの事実から多項分布版の Sanov の定理は証明されるのである.

補題 3.4 の証明を逆にたどって Kullback-Leibler 情報量が出て来るところまでの議論を繰り返そう.

$k_i \in \mathbb{Z}_{\geqq 0}$, $k_1 + \cdots + k_r = n$, $p_i = k_i/n$ と仮定する. 上の (2) を l_i 達について足し上げることによって

$$1 = \sum_{l_1+\cdots+l_r=n} \frac{n!}{l_1! \cdots l_r!} p_1^{l_1} \cdots p_r^{l_r} \leqq (n+1)^r \frac{n!}{k_1! \cdots p_r!} p_1^{k_1} \cdots p_r^{k_r}.$$

これの両辺を $(n+1)^r$ で割って得られる不等式と上の (1) を合わせると

$$\frac{1}{(n+1)^r} \leqq \frac{n!}{k_1! \cdots k_r!} p_1^{k_1} \cdots p_r^{k_r} \leqq 1$$

を得る. $q_i \in \mathbb{Z}_{\geqq 0}$, $q_1 + \cdots + q_r = 1$ であるとし, この不等式全体を $p_1^{k_1} \cdots p_r^{k_r}$ で割って, $q_1^{k_1} \cdots q_r^{k_r}$ をかけると

$$\frac{1}{(n+1)^r} \frac{q_1^{k_1} \cdots q_r^{k_r}}{p_1^{k_1} \cdots p_r^{k_r}} \leqq \frac{n!}{k_1! \cdots k_r!} q_1^{k_1} \cdots q_r^{k_r} \leqq \frac{q_1^{k_1} \cdots q_r^{k_r}}{p_1^{k_1} \cdots p_r^{k_r}}.$$

$k_i = np_i$ より, この時点ですでに Kullback-Leibler 情報量

$$D(p||q) = \sum_{i=1}^{r} p_i \log \frac{p_i}{q_i}$$

が見えている:

$$\log \frac{q_1^{k_1} \cdots q_r^{k_r}}{p_1^{k_1} \cdots p_r^{k_r}} = \log \left(\left(\frac{q_1}{p_r} \right)^{p_1} \cdots \left(\frac{q_r}{p_r} \right)^{p_r} \right)^n = -nD(p||q).$$

したがって

$$\frac{1}{(n+1)^r} e^{-nD(p||q)} \leqq \frac{n!}{k_1! \cdots k_r!} q_1^{k_1} \cdots q_r^{k_r} \leqq e^{-nD(p||q)}.$$

この不等式が補題 3.4 の結論であった. そしてこの不等式を用いて多項分布の $n \to \infty$ での様子を調べればただちに Sanov の定理 (定理 3.1) が得られるのであった. □

4 Sanov の定理を使ったカノニカル分布の導出

第 3 節の記号をそのまま引き継ぐ. たとえば $\mathcal{P}$ は有限集合 $\{1,2,\dots,r\}$ 上の確率分布 $p=(p_1,\dots,p_r)$ 全体の集合であるとし, 母集団分布 $q=(q_1,\dots,q_r)\in\mathcal{P}$ を任意に取って固定する. n 回の独立試行の結果, 状態 i が生じた回数を k_i と書くと, 状態 i の生じた割合は k_i/n である. 経験分布 $P_n=(k_1/n,\dots,k_r/n)$ は $\mathcal{P}$ に値を持つ確率変数になる[14].

4.1 分配函数とエネルギーの期待値

$E=(E_1,\dots,E_r)\in\mathbb{R}^r$ であるとし,

$$E_1=\cdots=E_a<E_{a+1}\leqq\cdots\leqq E_{r-b}<E_{r-b+1}=\cdots=E_r$$

かつ $q_1,q_r>0$ であると仮定しておく (あとで分配函数の対数凸性などを保証するための仮定). E_i たちを状態 i の**エネルギー**と呼ぶ[15]. $\beta\in\mathbb{R}$ に対して分布 $p(\beta)=(p_1(\beta),\dots,p_r(\beta))\in\mathcal{P}$ と函数 $Z(\beta)$ を

$$p_i(\beta)=\frac{e^{-\beta E_i}q_i}{Z(\beta)},\qquad Z(\beta)=\sum_{i=1}^{r}e^{-\beta E_i}q_i,$$

によって定める. さらに函数 $U(\beta)=\langle E\rangle_\beta$ を

$$U(\beta)=\langle E\rangle_\beta=\sum_{i=1}^{r}E_ip_i(\beta)=-\frac{\partial}{\partial\beta}\log Z(\beta)$$

と定める. β を**逆温度**と呼び, $e^{-\beta E_i}$ を **Boltzmann 因子**と呼び, $p(\beta)$ を**カノニカル分布**と呼び, 函数 $Z(\beta)$ を**分配函数**と呼び, 函数 $U(\beta)$ を**エネルギーの期待値**と呼ぶ.

$\log Z(\beta)$ は β に関する下に狭義凸な函数である. なぜならば

$$\left(\frac{\partial}{\partial\beta}\right)^2\log Z(\beta)=\frac{Z''(\beta)Z(\beta)-Z'(\beta)^2}{Z(\beta)^2}$$

であり, $a_i=e^{-\beta E_i}q_i\geqq 0$ とおくと, 最初の方の仮定から $a_1,a_r>0$ かつ $E_1<E_r$ なので

$$\begin{aligned}Z''(\beta)Z(\beta)-Z'(\beta)^2&=\sum_{i,j}E_i^2a_ia_j-\sum_{i,j}E_ia_iE_ja_j\\&=\frac{1}{2}\sum_{i,j}(E_i^2+E_j^2)a_ia_j-\frac{1}{2}\sum_{i,j}2E_iE_ja_ia_j=\frac{1}{2}\sum_{i,j}(E_i-E_j)^2a_ia_j>0\end{aligned}$$

となり, ゆえに

$$\left(\frac{\partial}{\partial\beta}\right)^2\log Z(\beta)>0$$

となるからである. したがって, エネルギーの期待値

$$U(\beta)=-\frac{\partial}{\partial\beta}\log Z(\beta)$$

[14] サイコロの目のように試行ごとに値が確率的に変化する変数を確率変数と呼ぶ. 経験分布 P_n の値は n 回の独立試行をやり直すごとに確率的に変化するので, P_n は確率変数だとみなされる.

[15] ギャンブルが好きな人はエネルギーをサイコロで i の目が出たときにもらえる賞金だと思ってよい.

は β の狭義単調減少函数である.

次に $U(\beta)$ の値の様子を調べよう. まず $p(0) = q$ より

$$U(0) = \sum_{i=1}^{r} E_i q_i.$$

次に $\beta \to \infty$ のとき

$$U(\beta) = \frac{\sum_i E_i e^{-\beta E_i} q_i}{\sum_i e^{-\beta E_i} q_i} = \frac{e^{-\beta E_1} \sum_i E_i e^{-\beta(E_i - E_1)} q_i}{e^{-\beta E_1} \sum_i e^{-\beta(E_i - E_1)} q_i} \to \frac{e^{-\beta E_1} \sum_{i=1}^{a} E_i q_i}{e^{-\beta E_1} \sum_{i=1}^{a} q_i} = E_1.$$

最後に $\beta \to -\infty$ のとき

$$U(\beta) = \frac{\sum_i E_i e^{-\beta E_i} q_i}{\sum_i e^{-\beta E_i} q_i} = \frac{e^{-\beta E_r} \sum_i E_i e^{\beta(E_r - E_i)} q_i}{e^{-\beta E_r} \sum_i e^{\beta(E_r - E_i)} q_i} \to \frac{e^{-\beta E_r} \sum_{i=r-b+1}^{r} E_i q_i}{e^{-\beta E_r} \sum_{i=r-b+1}^{r} q_i} = E_r.$$

以上によって, $E_r \geqq U \geqq E_1$ と $-\infty \leqq \beta \leqq \infty$ は $U = U(\beta)$ によって一対一に対応していることがわかる.

4.2 条件付き確率分布のカノニカル分布への収束

経験分布 $p = (p_1, \ldots, p_r) \in \mathcal{P}$ について, 条件 $\sum_{i=1}^{r} E_i p_i \approx U(\beta)$ のもとで, $n \to \infty$ のとき条件付き確率分布がカノニカル分布 $p(\beta)$ に収束することを示したい.

以下では, 数学的に厳密な取り扱いをするために, 条件 $\sum_{i=1}^{r} E_i p_i \approx U(\beta)$ の代わりに, 任意に $a > 0$ を取って以下のように条件を課す:

- $\beta \geqq 0$ のとき, 条件 $U(\beta) - a \leqq \sum_{i=1}^{r} E_i p_i \leqq U(\beta)$ を課す.
- $\beta \leqq 0$ のとき, 条件 $U(\beta) \leqq \sum_{i=1}^{r} E_i p_i \leqq U(\beta) + a$ を課す.

後で $a > 0$ の取り方は議論の本質に無関係であることがわかる. この条件のもとでの条件付き確率を考えるために $\{1, 2, \ldots, r\}$ 上の確率分布全体の集合 $\mathcal{P}$ の部分集合 A を

$$A = \begin{cases} \{\, p \in \mathcal{P} \mid U(\beta) - a \leqq \sum_{i=1}^{r} E_i p_i \leqq U(\beta) \,\} & (\beta \geqq 0), \\ \{\, p \in \mathcal{P} \mid U(\beta) \leqq \sum_{i=1}^{r} E_i p_i \leqq U(\beta) + a \,\} & (\beta \leqq 0) \end{cases}$$

と定める. 条件 $P_n \in A$ のもとでの条件付き確率

$$P(P_n \in B | P_n \in A) = \frac{P(P_n \in A \cap B)}{P(P_n \in A)} \qquad (B \subset \mathcal{P})$$

が $n \to \infty$ でカノニカル分布 $p(\beta)$ に集中することを Sanov の定理 (定理 3.1) を使って証明したい. そのために, 任意に $\varepsilon > 0$ を取って, $\mathcal{P}$ の部分集合 B を次のように定める:

$$B = \{\, p \in \mathcal{P} \mid ||p - p(\beta)|| < \varepsilon \,\}.$$

ここで $||\cdot||$ は Euclid ノルムである. B は $p(\beta)$ の ε 開近傍である. 以上の設定のもとで, $n\to\infty$ で条件付き確率分布がカノニカル分布 $p(\beta)$ に集中することを意味する

$$P(P_n\in B|P_n\in A)\to 1 \qquad (n\to\infty) \tag{$*$}$$

を示すことが以下の目標である.

Kullback-Leibler 情報量 $D(p||q)$ の定義を $\mathcal{P}$ の部分集合 C に

$$D(C||q)=\inf_{p\in C} D(p||q)$$

と拡張しておく. Sanov の定理より, $\mathcal{P}$ の部分集合 C の開核の閉包が C を含むとき

$$P(P_n\in C)=\exp(-nD(C||q)+o(n)).$$

上で定めた $\mathcal{P}$ の部分集合 A, B, $A\cap B$ の開核の閉包はそれぞれ A, B, $A\cap B$ を含む. さらに B の A での補集合 $B'=A\setminus B$ も同様である. ゆえに

$$P(P_n\in B'|P_n\in A)=\frac{P(P_n\in B')}{P(P_n\in A)}=\exp(-n(D(B'||q)-D(A||q))+o(n)).$$

これが $n\to\infty$ で 0 に収束することと目標である $(*)$ は同値である.

もしも条件 $p\in A$ のもとで $p=p(\beta)$ が $D(p||q)$ が唯一の最小点になるならば, $B'=A\setminus B$ の閉包に $p(\beta)$ が含まれないことより, $D(B'||q)>D(A||q)=D(p(\beta)||q)$ となり, $n\to\infty$ で $P(P_n\in B'|P_n\in A)\to 0$ となることがわかる.

$D(p||q)$ は p の函数として下に狭義凸であり, A は $\mathcal{P}$ の凸部分集合なので, 条件 $p\in A$ のもとでの $D(p||q)$ が $p=p(\beta)$ で最小になるならば, $p=p(\beta)$ は唯一の最小点になる. ゆえに条件 $p\in A$ のもとで $D(p||q)$ が $p=p(\beta)$ で最小になることを示せば $(*)$ の証明が終了する. 以下でそのことを証明しよう.

カノニカル分布 $p(\beta)$ は

$$\sum_{i=1}^r E_i p_i(\beta)=U(\beta)$$

を満たしているので, $p(\beta)\in A$ である. さらに[16]

$$\begin{aligned} D(p(\beta)||q)&=\sum_{i=1}^r p_i(\beta)\log\frac{p_i(\beta)}{q_i}=\sum_{i=1}^r p_i(\beta)\log\frac{e^{-\beta E_i}}{Z(\beta)}\\ &=\sum_{i=1}^r p_i(\beta)(-\beta E_i-\log Z(\beta))=-\beta U(\beta)-\log Z(\beta). \end{aligned} \tag{\%}$$

これが条件 $p\in A$ のもとでの $D(p||q)$ の最小値であることを示したい. すなわち $p\in A$ のとき $D(p||q)\geqq D(p(\beta)||q)$ となることを示したい.

$p\in A$ と仮定する. このとき, A の定義より, $\beta\geqq 0$ のとき $\sum_{i=1}^r E_i p_i\leqq U(\beta)$ となり, $\beta\leqq 0$ のとき $\sum_{i=1}^r E_i p_i\geqq U(\beta)$ となるので, β の符号によらずに

$$\beta\sum_{i=1}^r E_i p_i\leqq \beta U(\beta). \tag{\#}$$

[16]相対エントロピー $S(p||q)=-D(p||q)$ を用いて公式を書き直すと $S(p||q)=\beta U(\beta)+\log Z(\beta)$ になる. この手の公式は統計力学を知っている人達にはお馴染みのものだろう.

が成立している. 第 4.1 節の計算より, $\beta > 0$ と $U(\beta) < \sum_{i=1}^r E_i q_i$ は同値であり, $\beta < 0$ と $U(\beta) > \sum_{i=1}^r E_i q_i$ は同値である. 集合 A を定義するときに用いた $a > 0$ は以下の議論には関係しない.

Kullback-Leibler 情報量 $D(p||q)$ は以下のように変形される:

$$\begin{aligned} D(p||q) &= \sum_{i=1}^r p_i \log \frac{p_i}{q_i} = \sum_{i=1}^r p_i \log \left(\frac{p_i}{p_i(\beta)} \frac{p_i(\beta)}{q_i} \right) = \sum_{i=1}^r p_i \log \frac{p_i}{p_i(\beta)} + \sum_{i=1}^r p_i \log \frac{p_i(\beta)}{q_i} \\ &= D(p||p(\beta)) + \sum_{i=1}^r p_i \log \frac{e^{-\beta E_i}}{Z(\beta)} = D(p||p(\beta)) + \sum_{i=1}^r p_i(-\beta E_i - \log Z(\beta)) \\ &= D(p||p(\beta)) - \beta \sum_{i=1}^r E_i p_i - \log Z(\beta). \end{aligned}$$

ゆえに, 不等式 (#) とカノニカル分布の Kullback-Leibler 情報量 $D(p(\beta)||q)$ の表示 (%) と Kullback-Leibler 情報量が常に 0 以上であることより,

$$D(p||q) \geqq D(p||p(\beta)) - \beta U(\beta) - \log Z(\beta) = D(p||p(\beta)) + D(p(\beta)||q) \geqq D(p(\beta)||q).$$

これで条件 $p \in A$ のもとで $D(p||q)$ は $p = p(\beta)$ で最小になることがわかった. 目標の $(*)$ が証明された.

注意 4.1 (不等式 (#) について). 以上の議論は本質的に不等式 (#) の仮定だけに基づく.

統計力学の文脈では β は絶対温度の Boltzmann 定数倍の逆数を意味する. その場合には $\beta > 0$ となるので不等式 (#) は

$$(\text{エネルギーの平均値}) = \sum_{i=1}^r E_i p_i \leqq U(\beta)$$

を意味する. この型の条件でカノニカル分布が特徴付けられることについては田崎 [10] の第 9-2-1 節 (p.319) も参照せよ.

カノニカル分布が経験的に自然に得られることを示すためには, $\sum_{i=1}^r E_i p_i \approx U(\beta)$ という強い条件を仮定する必要はなく, 不等式 (#) を仮定するだけでよい. この点についてもう少し詳しくコメントしておく.

第 4.1 節で計算した通り, $U(0)$ は母集団分布でのエネルギーの期待値 $\sum_{i=1}^r E_i q_i$ になる. そして $\beta > 0$ を大きくすると $U(\beta)$ は最小エネルギー準位 $\min\{E_1, \ldots, E_r\}$ に近付き, $\beta < 0$ を小さくすると $U(\beta)$ は最大エネルギー準位 $\max\{E_1, \ldots, E_r\}$ に近付く. $\beta = \infty$ で状態は最小エネルギー状態 (基底状態) にはりつくようになり, $\beta = -\infty$ で状態は最大エネルギー状態にはりつくようになる.

統計力学において β は絶対温度の逆数であり, $\beta = \infty$ は絶対零度に対応し, $\beta = 0$ は絶対温度無限大に対応している. 我々が扱っている場合には β は負にもなりえる. その場合には対応する絶対温度も負の値になる. 絶対温度の高さを逆温度 β の低さで測ることにすれば, 負の絶対温度は絶対温度無限大よりも高温であるとみなされる.

我々が扱っているのは次のような状況であると考えられる.

r 種類の目が出るルーレットを回して, i の目が出たら賞金を E_i ペリカもらえるゲームを考える. ($E_i < 0$ の場合には $|E_i|$ ペリカ支払うことにすればギャンブルになる.) i の目が出る確率は q_i であるとする. そのようなゲーム 1 回あたりの賞金の期待値は $U_0 = \sum_{i=1}^r E_i q_i$

になる. 大数の法則よりそのようなゲームをたくさん繰り返せば 1 回あたりの賞金の平均値は U_0 に近付く.

まず $U > U_0$ であると仮定する. ゲームをたくさん繰り返して (回数は n 回とする). ゲーム 1 回あたりの賞金の平均値が U 未満で終わったならば, 時間を巻き戻して何度でも n 回分のゲームをやり直せると仮定する. そのようにしてゲーム 1 回あたりの賞金の平均値がゲーム自体の期待値である U_0 より大きい U 以上になったら時間を巻き戻すのを止める. このとき, n 回のゲーム中 i の目が出た割合 p_i は (n が大きなとき) どのような値になる可能性が高いだろうか?

ゲーム 1 回あたりの賞金の平均値が U_0 から離れれば離れるほどそのような状況が生じる確率は下がるので, ゲーム 1 回あたりの賞金の平均値はほぼ U (すなわち U よりほんの少し大きな数値) になってしまう可能性が高いだろう.

そのときの i の目が出た割合 p_i を計算すると, U が定める β に対応するカノニカル分布 $p_i(\beta)$ に近くなる可能性が高いというのがこの節において数学的にきちんと証明したことである.

この場合には $U > U_0$ なので $\beta < 0$ となる. すなわち必要ならば時間を巻き戻すことによって, ゲーム自体の賞金期待値よりも高い賞金を求めると, 対応する絶対温度は負の値になってしまうと解釈される.

絶対温度が正の値の状況を作り出すには, $U < U_0$ であると仮定し, 必要ならば時間を巻き戻して, ゲーム 1 回あたりの賞金の平均値が U 以下になるようにすればよい. そのとき, ゲーム 1 回あたりの賞金の平均値が U_0 から離れれば離れるほどそのような状況が生じる確率は下がるので, ゲーム 1 回あたりの賞金の平均値はほぼ U になる可能性が高く, i の目が出た割合は U に対応するカノニカル分布 $p_i(\beta)$ に近くなる可能性が高い. この場合には $U < U_0$ なので $\beta > 0$ となり, 絶対温度は正の値になる.

つまり, 時間を巻き戻して, ゲーム自体の期待値よりも低い賞金を得るようにすると絶対温度は正の値になるとされるのである.

以上の説明を読めばカノニカル分布の導出で使った不等式 (#) の向きがどのように自然であるかがわかると思う. 上の議論と第 1.5 節の最後の方の極限の計算と比較してみよ. 第 1.5 節では「丁半博打」のケースを扱っていると考えられる. □

4.3 まとめと二項分布もカノニカル分布の例になっていること

以上の結果は次のようにまとめられる.

定理 4.2 (条件付き大数の弱法則, 条件付き極限定理). 母集団分布は $q = (q_1, \ldots, q_r) \in \mathcal{P}$ であるとする. $X_1, X_2, \ldots$ は独立で同分布 q にしたがう $\{1, 2, \ldots, r\}$ に値を持つ確率変数列であるとする. $X_1, X_2, \ldots, X_n$ の中に含まれる i の個数を k_i と書き, $P_n = (k_1/n, \ldots, k_r/n)$ とおく. P_n は分布の集合 $\mathcal{P}$ に値を持つ確率変数になる. $E_i \in \mathbb{R}$ は第 4.1 節の通りとする. $E_1 < U < E_r$ であるとし, $\beta \in \mathbb{R}$, $p(\beta) = (p_1(\beta), \ldots, p_r(\beta)) \in \mathcal{P}$, $Z(\beta)$ を以下の条件で定める:

$$p_i(\beta) = \frac{e^{-\beta E_i} q_i}{Z(\beta)}, \quad Z(\beta) = \sum_{i=1}^{r} e^{-\beta E_i} q_i, \quad -\frac{\partial}{\partial \beta} \log Z(\beta) = \sum_{i=1}^{r} E_i p_i(\beta) = U.$$

$p(\beta)$ を**カノニカル分布**と呼ぶ. $0 < a \leqq \infty$ とし, 分布の集合 $A_U \subset \mathcal{P}$ を

$$A_U = \begin{cases} \{\, p \in \mathcal{P} \mid U - a \leqq \sum_{i=1}^r E_i p_i \leqq U \,\} & (\beta \geqq 0), \\ \{\, p \in \mathcal{P} \mid U \leqq \sum_{i=1}^r E_i p_i \leqq U + a \,\} & (\beta \leqq 0) \end{cases}$$

と定める. このとき $\sum_{i=1}^r E_i p_i(\beta) = U$ なので $p(\beta) \in A_U$ である. 任意に $\varepsilon > 0$ を取り, $p(\beta)$ の A_U における ε 開近傍を $B_\varepsilon(p(\beta))$ と書く. このとき, $n \to \infty$ で

$$P(P_n \in B_\varepsilon(p(\beta)) | P_n \in A_U) = \frac{P(P_n \in B_\varepsilon(p(\beta)))}{P(P_n \in A_U)} \to 1.$$

すなわち経験分布 P_n は $n \to \infty$ でカノニカル分布 $p(\beta)$ に (確率) 収束する. □

例 4.3. 確率 θ に対応する一般の二項分布が確率 $1/2$ の二項分布を母集団分布とする場合のカノニカル分布として自然に現われることを説明しよう. この例は母集団分布が一様分布でない場合のカノニカル分布の簡単な例になっている.

有限集合 $\{0, 1, \ldots, r\}$ に値を持つ確率変数 X は確率 $1/2$ に対応する対称な二項分布にしたがうと仮定する:

$$P(X = i) = \binom{r}{i} \frac{1}{2^r} \qquad (i = 0, 1, \ldots, r).$$

$X_1, X_2, \ldots$ は独立で X と同じ分布を持つ確率変数の列であるとする. $X_1, X_2, \ldots, X_n$ の中に含まれる i の個数を k_i と書き, $P_n = (k_0/n, k_1/n, \ldots, k_r/n)$ とおくと, P_n は $\{0, 1, \ldots, r\}$ 上の確率分布に値を持つ確率変数になる. 大数の法則より, 何も条件を付けずに $n \to \infty$ とすると, P_n は対称な二項分布に近付く.

$E_i = i$ の場合にカノニカル分布が何になるかを計算してみよう. そのとき, 分配函数は二項定理より

$$Z(\beta) = \sum_{i=0}^r e^{-\beta E_i} q_i = \sum_{i=0}^r \binom{r}{i} \left(\frac{e^{-\beta}}{2}\right)^i \left(\frac{1}{2}\right)^{r-i} = \frac{(e^{-\beta} + 1)^r}{2^r}$$

となるので, カノニカル分布は

$$p_i(\beta) = \frac{e^{-\beta E_i} q_i}{Z(\beta)} = \binom{r}{i} \frac{e^{-\beta i}}{(e^{-\beta} + 1)^r} = \binom{r}{i} \left(\frac{e^{-\beta}}{e^{-\beta} + 1}\right)^i \left(\frac{1}{e^{-\beta} + 1}\right)^{r-i}$$

と二項分布になる. つまり,

$$p_i(\beta) = \binom{r}{i} \theta^i (1 - \theta)^{r-i}, \qquad \theta = \frac{e^{-\beta}}{e^{-\beta} + 1}.$$

このとき,

$$-Z'(\beta) = \frac{r e^{-\beta} (e^{-\beta} + 1)^{r-1}}{2^r}$$

なので, "エネルギーの期待値" は

$$U(\beta) = -\frac{\partial}{\partial \beta} Z(\beta) = \frac{-Z'(\beta)}{Z(\beta)} = \frac{r e^{-\beta}}{e^{-\beta} + 1} = r\theta$$

と確率 θ に対応する二項分布における i の期待値になる.

確率 $1/2$ に対応する対称な二項分布についてはコイン投げの状況を想像すると分かり易いだろう. コインを投げたとき表になる確率が $1/2$ である状況を考える. そのようなコインを r 回投げて表の出た回数 i と表が出た割合 i/r を記録する行為を n 回繰り返したとしよう. ただし n は非常に大きいとする. 記録には表の出た回数 i と表が出た割合 i/r がそれぞれ n 個ずつ記録されている. その記録を見ると次が成立していたとする[17]:

$$(\text{表の出た割合 } i/r \text{ の平均値}) = \sum_{i=0}^{r} \frac{i}{r} p_i \approx \theta. \tag{$*$}$$

このとき $p_i = k_i/n$ たちはどのような値になっている可能性が高いだろうか?

「経験分布 $p = (p_0, p_1, \ldots, p_r)$ は確率 θ に対応する二項分布にほぼ等しくなっている可能性が高い」というのが, 定理 4.2 をすぐ上の計算に適用したときの結論になる.

上で計算したカノニカル分布 $p(\beta)$ (確率 θ に対応する二項分布) は, 定理 4.2 より,

$$(\text{表が出た回数 } i \text{ の期待値}) = \sum_{i=0}^{r} i p_i = \sum_{i=0}^{r} E_i p_i \approx U(\beta) = r\theta$$

を満たす分布 $p = (p_0, p_1, \ldots, p_r)$ に制限した場合の経験分布 P_n が $n \to \infty$ で近付く先になっている. この条件は上の $(*)$ と同値である. 表の出る割合が θ になるという条件で制限を付ければ確率 θ に対応する二項分布が経験分布として自然に現われる.

結論. 表の出る確率が $1/2$ の公平なコインを r 回投げて表の出た回数を数えることをたくさん繰り返し, r 回中表の出た回数が大量に記録されたリストを作ったとする. そのときもしもその記録において r 回中表の出た割合の平均値が θ になっているならば[18], そのリストにはあたかも「表の出る確率が θ の二項分布の記録が残っている」かのように見えてしまうことになる. □

注意 4.4. 例 4.3 の議論を一般化するとほぼ自明に以下のようなことが成立していることがわかる. 母集団分布 q_i は最初から E_i に関するカノニカル分布の形をしていると仮定する:

$$q_i = \frac{e^{-\beta_0 E_i} q_{0,i}}{Z_0}, \qquad Z_0 = \sum_{i=1}^{r} e^{-\beta_0 E_i} q_{i,0}.$$

ここで $q_{0,i} \geqq 0$, $\sum_{i=1}^{r} q_{0,i} = 1$, $E_1, \ldots, E_r \in \mathbb{R}$ の最大値 E_1 と最大値 E_r は異なり, $q_1, q_r > 0$ であると仮定する. このとき, この母集団分布と E_i たちに対応するカノニカル分布 $p(\beta)$ も母集団分布と同じ形のカノニカル分布になる:

$$p_i(\beta) = \frac{e^{-(\beta_0+\beta)E_i} q_{0,i}}{Z_0 Z(\beta)}, \qquad Z_0 Z(\beta) = Z_0 \sum_{i=1}^{r} e^{-\beta E_i} q_i = \sum_{i=1}^{r} e^{-(\beta_0+\beta)E_i} q_{0,i}.$$

二項分布は離散型の指数型分布族に含まれているので例 4.3 のようなことが成立するのである. 他の指数型分布族に含まれる確率分布についても同様のことが成立する. □

[17] 大数の (弱) 法則より, θ が $1/2$ から離れている確率は $n \to \infty$ で 0 に近付く. ここではそのような稀なケースが生じた場合を想定している.

[18] θ が $1/2$ から離れると, n が大きなときそのようなリストが得られる確率はほぼ 0 になるが, そのような稀な状況が生じてしまった場合についても考えている.

5　付録: Kullback-Leibler 情報量に関する不等式

Cover-Thomas [3] は情報理論に関する有名な教科書である. 情報量とエントロピー一般に関する詳しい解説を読みたい人はその本を参照すればよい. 以下では主にこの教科書を参照しながら, Kullback-Leibler 情報量が「距離」のような性質を持っていることを意味する不等式を扱う.

5.1　準備: Jensen の不等式

函数 $f(X)$ を数 $E[f(X)]$ に対応させる汎函数 $E[\,\cdot\,]$ は以下の条件を満たしていると仮定する:

(1) 線形性: 函数 $f(X)$, $g(X)$ と数 α, β に対して,
$$E[\alpha f(X) + \beta g(X)] = \alpha E[f(X)] + \beta E[g(X)].$$

(2) 短調性: $f(X) \geqq g(X)$ が常に成立しているならば $E[f(X)] \geqq E[g(X)]$.

(3) 規格化条件: $E[1] = 1$.

規格化条件と短調性より, 数 α に対して $E[\alpha] = \alpha$ となることがわかる. このような $E[\,\cdot\,]$ を期待値汎函数と呼ぶ. たとえば $p_i \geqq 0$, $\sum_{i=1}^r p_i = 1$ のとき, $E[f(X)] = \sum_{i=1}^r f(x_i)p_i$ は期待値汎函数である. 他にも $\rho(x) \geqq 0$, $\int_a^b \rho(x)\,dx = 1$ のとき, $E[f(X)] = \int_a^b f(x)\rho(x)\,dx$ も期待値汎函数である.

このとき以下の不等式が成立している (**Jensen の不等式**):

- $f(X)$ が上に凸ならば $E[f(X)] \leqq f(E[X])$.
- $f(X)$ が下に凸ならば $E[f(X)] \geqq f(E[X])$.

以下で前者のみを証明しよう. 後者は $-f(X)$ に前者を適用すれば得られる.

前者の証明. 函数 $f(X)$ は上に凸であると仮定し, $\mu = E[X]$ とおく. このとき上に凸な函数 $f(X)$ の $X = \mu$ での "接線" を $a(X-\mu) + f(\mu)$ と書くと,

$$f(X) \leqq a(X-\mu) + f(\mu)$$

となるので,

$$E[f(X)] \leqq E[a(X-\mu) + f(\mu)] = a(E[X]-\mu) + f(\mu) = f(E[X]).$$

2 つ目の等号で期待値汎函数の短調性を使い, 2 つ目の等号でその線形性と規格化条件を使った[19]. □

注意 5.1. 上の証明から, $f(X)$ が上に狭義凸ならば, X の分布が $\mu = E[f(X)]$ に集中していない限り (ほとんど確実に $f(X) = \mu$ が成立していない限り), 等号を含まない強い不等式が成立していることがわかる. □

[19] $E[f(X)] = \sum_{i=1}^r f(x_i)p_i$ の場合の Jensen の不等式は r に関する数学的帰納法で証明することもできるが, 期待値汎函数の公理だけを使って証明する方が不等式が成立する理由が分かりやすいと思う.

5.2 対数和不等式とその応用

次の不等式はたとえば [3], p.31, Theorem 2.7.1 (Log sum inequality) に書いてある.

対数和不等式 0 以上の a_i, b_i に対して,

$$\sum_{i=1}^{n} a_i \log \frac{a_i}{b_i} \geqq A \log \frac{A}{B}, \qquad A = \sum_{i=1}^{n} a_i, \quad B = \sum_{i=1}^{n} b_i.$$

等号の成立と a_i/b_i が互いにすべて等しいことは同値である. いつものように $0 \log 0 = 0$, $a \log(a/0) = \infty$ と約束しておく.

証明. $a_i > 0$, $b_i > 0$ の場合のみを証明すれば十分である. (一般の場合はその場合の極限として証明される.) $f(x) = x \log x$ とおくと, $f'(x) = \log x + 1$, $f''(x) = 1/x$ なので $f(x)$ は $x > 0$ で下に狭義凸である. これに Jensen の不等式を適用しよう. そのために $q_i = b_i/B$ とおく. このとき

$$\begin{aligned} \frac{1}{B} \sum_{i=1}^{n} a_i \log \frac{a_i}{b_i} &= \sum_{i=1}^{n} \frac{b_i}{B} \frac{a_i}{b_i} \log \frac{a_i}{b_i} = \sum_{i=1}^{n} q_i f\left(\frac{a_i}{b_i}\right) \\ &\geqq f\left(\sum_{i=1}^{n} q_i \frac{a_i}{b_i}\right) = f\left(\sum_{i=1}^{n} \frac{b_i}{B} \frac{a_i}{b_i}\right) = f\left(\frac{A}{B}\right) = \frac{A}{B} \log \frac{A}{B}. \end{aligned}$$

$f(x)$ の狭義凸性より, 等号が成立することと a_i/b_i が互いにすべて等しいことが同値であることもわかる. □

この不等式を使えば特に p_i, q_i が非負でそれぞれの総和が 1 のとき

$$D(p||q) = \sum_{i=1}^{r} p_i \log \frac{p_i}{q_i} \geqq 1 \log \frac{1}{1} = 0$$

が得られる (**Kullback-Leibler 情報量の非負性**). さらに, 集合 $\{1, 2, \ldots, r\}$ の分割

$$\{1, 2, \ldots, r\} = A_1 \sqcup \cdots \sqcup A_s$$

に対して, $\{ A_1, \ldots, A_s \}$ 上の確率分布 $P = (P_1, \ldots, P_s)$, $Q_j = (Q_1, \ldots, Q_s)$ を

$$P_j = \sum_{i \in A_j} p_i, \qquad Q_j = \sum_{i \in A_j} q_i$$

と定めると, 対数和不等式より

$$D(p||q) = \sum_{j=1}^{s} \sum_{i \in A_j} p_i \log \frac{p_i}{q_i} \geqq \sum_{j=1}^{s} P_i \log \frac{P_j}{Q_j} = D(P||Q).$$

要するに, 細部の情報を忘れると **Kullback-Leibler 情報量は小さくなる**.

5.3 Kullback-Leibler 情報量で L^1 距離を上からおさえられこと

集合 $\{1,2,\dots,r\}$ 上の確率分布 $p=(p_1,\dots,p_r)$, $q=(q_1,\dots,q_r)$ のあいだの L^1 距離 $||p-q||_{L^1}$ を

$$||p-q||_{L^1}=\sum_{i=1}^{r}|p_i-q_i|$$

と定める. 確率分布 p に関する確率を

$$p(A)=\sum_{i\in A}p_i \qquad (A\subset\{1,2,\dots,r\})$$

と書くと,

$$||p-q||_{L^1}=2(p(A)-q(A)),\qquad A=\left\{\,i\in\{1,2,\dots,r\}\,\middle|\,p_i\geqq q_i\,\right\}. \tag{\#}$$

なぜならば

$$\begin{aligned}||p-q||_{L^1}&=\sum_{i\in A}(p_i-q_i)+\sum_{i\in A^c}(q_i-p_i)=p(A)-q(A)+q(A^c)-p(A^c)\\&=p(A)-q(A)+(1-q(A))-(1-p(A))=2(p(A)-q(A)).\end{aligned}$$

以上の記号を以下においてそのまま用いる.

KL 情報量で L^1 距離を上からおさえられること:

$$D(p||q)\geqq\frac{1}{2}||p-q||_{L^1}^2. \tag{$*$}$$

この不等式を証明したい.

$\mathbf{r=2}$ の場合の $(*)$ の証明. $0<a<1$, $0<b<1$ のとき

$$a\log\frac{a}{b}+(1-a)\log\frac{1-a}{1-b}\geqq 2(a-b)^2$$

となることを示せばよい. そのために左辺から右辺を引いた結果を $f(b)$ と書く:

$$f(b)=a\log\frac{a}{b}+(1-a)\log\frac{1-a}{1-b}-2(a-b)^2.$$

このとき

$$f'(b)=-\frac{a}{b}+\frac{1-a}{1-b}-4(b-a)=(b-a)\left(\frac{1}{b(1-b)}-4\right).$$

$b(1-a)\leqq 1/4$ より, $1/(b(1-b))-4\geqq 0$ となる. ゆえに $f(b)$ の符号は $b-a$ の符号に等しい. すなわち $f(b)$ は $b<a$ で単調減少し, $b>a$ で単調増加する. したがって $f(p)=0$ なので $f(b)\geqq 0$ となることがわかる. □

一般の場合の $(*)$ の証明. 集合 A は $(\#)$ の通りであるとし, 集合 $\{1,2,\dots,r\}$ の分割 $\{A,A^c\}$ 上の確率分布 $P=(a,1-a)$, $Q=(b,1-b)$ を $a=p(A)$, $b=q(A)$ と定める. このとき細部の情報を忘れると Kullback-Leibler 情報量が小さくなること (第 5.2 節) より

$$D(p||q)\geqq D(P||Q)\geqq 2(a-b)^2=2(p(A)-q(A))^2=\frac{1}{2}||p-q||_{L^1}^2.$$

ここで2つ目の不等号で上で証明した $r=2$ の場合の結果を使い, 最後に $(\#)$ を使った. □

5.4 Pithagorian theorem

$\mathcal{P} = \{\, p = (p_1, \dots, p_r) \in \mathbb{R}^r_{\geqq 0} \mid p_1 + \cdots + p_r = 1 \,\}$ とおき, $\mathcal{P}$ を集合 $\{1, 2, \dots, r\}$ 上の確率分布全体の集合とみなす. $\mathcal{P}$ は $r-1$ 次元単体である.

次の不等式は [3], p.367, Theorem 11.6.1 にある.

Pythagorian theorem E は $\mathcal{P}$ の凸閉部分集合であるとし, $q \in \mathcal{P} \setminus E$ であるとする. $p^* = (p_1^*, \dots, p_r^*) \in E$ は $D(p||q)$ を E 上で最小化する p であるとする:

$$D(p^*||q) = \min_{p \in E} D(p||q).$$

このとき

$$D(p||q) \geqq D(p||p^*) + D(p^*||q) \qquad (p \in E).$$

この不等式と Kullback-Leibler 情報量の非負性より, $D(p||q) \to D(p^*||q)$ ならば $p \to p^*$ となることが導かれる.

Pythegorian theorem の証明. p^* と p を通る直線上の Kullback-Leibler 情報量の値の p^* における微係数を見ればこの不等式が証明される.

$t \in \mathbb{R}$ に対して

$$p(t) = (p_1(t), \dots, p_r(t)) = tp + (1-t)p^*, \qquad p_i(t) = tp_i + (1-t)p_i^*,$$
$$f(t) = D(p(t)||q) = \sum_{i=1}^r (tp_i + (1-t)p_i^*) \log \frac{tp_i + (1-t)p_i^*}{q_i}$$

とおく. このとき

$$\begin{aligned} f'(t) &= \sum_{i=1}^r \left((p_i - p_i^*) \log \frac{tp_i + (1-t)p_i^*}{q_i} + (p_i - p_i^*) \right) \\ &= \sum_{i=1}^r (p_i - p_i^*) \log \frac{tp_i + (1-t)p_i^*}{q_i}. \end{aligned}$$

2 つ目の等号で $\sum_{i=1}^r p_i = \sum_{i=1}^r p_i^* = 1$ となることを使った.

$p(0) = p^* \in E$, $p(1) = p \in E$ であり, E は凸だったので $p(t) \in E$ $(0 \leqq t \leqq 1)$. p^* は $D(p||q)$ を E 上で最小化する p だったので, $f'(0) \geqq 0$ となる. ゆえに

$$\begin{aligned} 0 \leqq f'(0) &= \sum_{i=1}^r (p_i - p_i^*) \log \frac{p_i^*}{q_i} = \sum_{i=1}^r p_i \log \left(\frac{p_i}{q_i} \frac{p_i^*}{p_i} \right) - \sum_{i=1}^r p_i^* \log \frac{p_i^*}{q_i} \\ &= \sum_{i=1}^r p_i \log \frac{p_i}{q_i} - \sum_{i=1}^r p_i \log \frac{p_i}{p_i^*} - \sum_{i=1}^r p_i^* \log \frac{p_i^*}{q_i} = D(p||q) - D(p||p^*) - D(p^*||q). \end{aligned}$$

これで示したい不等式が示された. □

6 付録: Cramér の定理

完璧に論理的に厳密な解説をするつもりはないので厳密な証明に興味がある人は注意して欲しい. 目標は統計力学の formulation との対応を明瞭になるようなスタイルで Cramér の定理の証明の概略を説明することである.

6.1 Cramér の定理の設定と主張

H は平均値を持つ確率変数であるとし, その平均値を

$$U_0 = E[H]$$

と書く. H のモーメント母函数[20]:

$$Z(\beta) = E[e^{-\beta H}].$$

は正の幅を持つ区間上で有限の値を持つと仮定する. $U(\beta)$ を次のように定める:

$$U(\beta) = \frac{E[He^{-\beta H}]}{Z(\beta)} = -\frac{\partial}{\partial\beta}\log Z(\beta).$$

H_1, H_2 独立な確率変数達で H と同じ分布にしたがうならば, $Z(\beta)$ が有限になる開区間において,

$$\left(\frac{\partial}{\partial\beta}\right)^2 \log Z(\beta) = -\frac{\partial}{\partial\beta}U(\beta) = \frac{Z''(\beta)Z(\beta) - Z'(\beta)^2}{Z(\beta)^2},$$
$$Z''(\beta)Z(\beta) - Z'(\beta)^2 = \frac{1}{2}E[(H_1 - H_2)^2 e^{-\beta(H_1+H_2)}] \geqq 0$$

となるので, $\log Z(\beta)$ は下に凸な函数になり, $U(\beta)$ は単調減少函数になる. 特に, $U(0) = U_0 = E[H]$ となり, $\beta \geqq 0$ のとき $U(\beta) \leqq U_0$ となり, $\beta \leqq 0$ のとき $U(\beta) \geqq U_0$ となる.

$\log Z(\beta)$ の Legendre 変換 $S(u)$ とその -1 倍 $D(u)$ を次のように定める:

$$S(u) = \inf_{\beta\in\mathbb{R}}(\beta u + \log Z(\beta)),$$
$$D(u) = -S(u) = \sup_{\beta\in\mathbb{R}}(-\beta u - \log Z(\beta)) \qquad (u \in \mathbb{R}).$$

$S(u)$ は u に関する一次函数 (特に下に凸な函数) の族の各点ごとの下限で定義された函数なので, $S(u)$ は上に凸な函数になる. $0u - \log Z(0) = 0$ なので

$$S(u) \leqq 0, \qquad D(u) \geqq 0 \qquad (u \in \mathbb{R})$$

となる. さらに $Z(\beta)$ が有限な開区間において,

$$\frac{\partial}{\partial\beta}(\beta u + \log Z(\beta)) = u - U(\beta)$$

なので, $U(\beta) = u$ となる $\beta = \beta(u)$ が存在するならば,

$$S(u) = \beta(u)u + \log Z(\beta(u)), \qquad S(U(\beta)) = \beta U(\beta) + \log Z(\beta)$$

となる. 簡単のためこのような $\beta = \beta(u)$ が常に取れると仮定する. これらの公式より $S(u)$ は**カノニカル分布の相対エントロピー**に対応し, $D(u) = -S(u)$ は**カノニカル分布の Kullback-Leibler 情報量**に対応することがわかる. 特に

$$S(U(0)) = 0, \qquad D(U(0)) = 0$$

[20]確率論の教科書では, 通常, 確率変数 H のモーメント母函数を $M(t) = E[e^{tH}]$ と定義するが, ここでは統計力学との対応を見易くするために $t = -\beta$ とおき, 統計力学における分配函数と同じ記号 $Z(\beta)$ をモーメント母函数に割り振った.

となるので, $S(u)$ は $u = U(0) = E[H]$ で最大値 0 になり, $D(u)$ は同点で最小値 0 になる. $S(u)$ (もしくは $D(u)$) は $u \geqq U(0) = E[H]$ で単調減少 (もしくは単調増加) し, $u \leqq U(0) = E[H]$ で単調増加 (もしくは単調減少) する.

$U(\beta)$ は単調減少函数なので, $u \leqq U(0) = E[H]$ のとき $\beta(u) \geqq 0$ となり, $u \geqq U(0) = E[H]$ のとき $\beta(u) \leqq 0$ となる. ゆえに次が成立している:

$$S(u) = \begin{cases} \inf_{\beta \geqq 0}(\beta u + \log Z(\beta)) & (u \leqq U(0) = E[H]), \\ \inf_{\beta \leqq 0}(\beta u + \log Z(\beta)) & (u \geqq U(0) = E[H]). \end{cases} \tag{\#}$$

$D(u)$ についても同様の表示が存在する. u が H の平均値 $U(0) = E[H]$ 以上と以下の場合で上限もしくは下限を取る範囲 β の範囲を半分に制限できる.

定理 6.1 (Cramér の定理). 以上の設定のもとで, $H_1, H_2, \ldots$ は独立同分布な確率変数列であり, H と同じ分布にしたがうと仮定する. このとき以下が成立している:

(1) F が $\mathbb{R}$ の閉部分集合ならば

$$\limsup_{n\to\infty} \frac{1}{n} \log P\left(\frac{1}{n}\sum_{k=1}^{n} H_k \in F\right) \leqq \sup_{u\in F} S(u) = -\inf_{u\in F} D(u).$$

(2) G が $\mathbb{R}$ の開部分集合ならば

$$\liminf_{n\to\infty} \frac{1}{n} \log P\left(\frac{1}{n}\sum_{k=1}^{n} H_k \in G\right) \geqq \sup_{u\in G} S(u) = -\inf_{u\in G} D(u).$$

(3) A が $\mathbb{R}$ の部分集合であり, A の開核 G の閉包 F が A を含み, $\sup_{u\in G} S(u) = \sup_{u\in F} S(u)$ が成立しているならば

$$\lim_{n\to\infty} \frac{1}{n} \log P\left(\frac{1}{n}\sum_{k=1}^{n} H_k \in A\right) = \sup_{u\in A} S(u) = -\inf_{u\in A} D(u).$$

このように, 標本平均 $(H_1 + \cdots + H_n)/n$ が集合 A に含まれる確率の $n \to \infty$ での漸近挙動はカノニカル分布の相対エントロピーに対応する量 $S(u)$ (もしくはカノニカル分布の Kullback-Leibler 情報量に対応する量 $D(u)$) の A における上限 (もしくは下限) で記述される. □

6.2 Cramér の定理の証明

補題 6.2 (Cramér の定理の本質的部分). 第 6.1 節の設定のもとで以下が成立している.

(1) u が H の平均値 $U(0) = E[H]$ 以上と以下の場合において, それぞれ

$$\frac{1}{n} \log P\left(\frac{1}{n}\sum_{k=1}^{n} H_k \geqq u\right) \leqq S(u) \qquad (u \geqq U(0) = E[H]),$$

$$\frac{1}{n} \log P\left(\frac{1}{n}\sum_{k=1}^{n} H_k \leqq u\right) \leqq S(u) \qquad (u \leqq U(0) = E[H]).$$

この結果を適用するときに, $u \geqq U(0) = E[H]$ で $S(u)$ は単調減少し, $u \leqq U(0) = E[H]$ で $S(u)$ は単調増加することに注意せよ.

(2) 任意の $\delta > 0$ に対して

$$\liminf_{n\to\infty} \frac{1}{n} \log P\left(\frac{1}{n}\sum_{k=1}^{n} H_k \in (u-\delta, u+\delta)\right) \geqq S(u).$$

(1) の上からの評価は特別なアイデア抜きに容易に証明される. (2) の下からの評価は “カノニカル分布に関する大数の弱法則” を使って証明される. □

証明. (1) の上からの評価を証明しよう. まず, $u \leqq U(0) = E[H]$ と仮定する. $H_1 + \cdots + H_n \leqq nu$ のとき 1 で他のとき 0 になる函数を $1_{H_1+\cdots+H_n\leqq nu}$ と書くと, $\beta \geqq 0$ のとき

$$\begin{aligned} P\left(\frac{1}{n}\sum_{k=1}^{n} H_k \leqq u\right) &= E[1_{H_1+\cdots+H_n\leqq nu}] \\ &\leqq E\left[1_{H_1+\cdots+H_n\leqq nu} e^{-\beta(H_1+\cdots+H_n-nu)}\right] \\ &\leqq E\left[e^{-\beta(H_1+\cdots+H_n-nu)}\right] \\ &= e^{n\beta u} Z(\beta)^n = e^{n(\beta u + \log Z(\beta))}. \end{aligned}$$

ゆえに $u \leqq U(0) = E[H]$, $\beta \geqq 0$ のとき

$$\frac{1}{n} \log P\left(\frac{1}{n}\sum_{k=1}^{n} H_k \leqq u\right) \leqq \beta u + \log Z(\beta).$$

したがって第 6.1 節の (#) より

$$\frac{1}{n} \log P\left(\frac{1}{n}\sum_{k=1}^{n} H_k \leqq u\right) \leqq \inf_{\beta\geqq 0}(\beta u + \log Z(\beta)) = S(u).$$

次に, $u \geqq U(0) = E[H]$ と仮定して, 上と同様の議論を行なう. $H_1 + \cdots + H_n \geqq nu$ のとき 1 で他のとき 0 になる函数を $1_{H_1+\cdots+H_n\geqq nu}$ と書くと, $\beta \leqq 0$ のとき

$$\begin{aligned} P\left(\frac{1}{n}\sum_{k=1}^{n} H_k \geqq u\right) &= E[1_{H_1+\cdots+H_n\geqq nu}] \\ &\leqq E\left[1_{H_1+\cdots+H_n\geqq nu} e^{-\beta(H_1+\cdots+H_n-nu)}\right] \\ &\leqq E\left[e^{-\beta(H_1+\cdots+H_n-nu)}\right] \\ &= e^{n\beta u} Z(\beta)^n = e^{n(\beta u + \log Z(\beta))}. \end{aligned}$$

ゆえに $u \geqq U(0) = E[H]$, $\beta \leqq 0$ のとき

$$\frac{1}{n} \log P\left(\frac{1}{n}\sum_{k=1}^{n} H_k \geqq u\right) \leqq \beta u + \log Z(\beta).$$

したがって第 6.1 節の (#) より

$$\frac{1}{n} \log P\left(\frac{1}{n}\sum_{k=1}^{n} H_k \geqq u\right) \leqq \inf_{\beta\leqq 0}(\beta u + \log Z(\beta)) = S(u).$$

これで (1) の上からの評価が証明された.

(2) の下からの評価を証明しよう. (2) は “カノニカル分布に関する大数の弱法則” から導かれる. 確率変数 H の確率分布は $\mathbb{R}$ 上の確率測度 μ が定める確率分布にしたがっているとする. 確率測度 μ_β を

$$\mu_\beta(dx) = \frac{e^{-\beta x}\mu(dx)}{Z(\beta)}$$

と定め, この確率測度の定める確率分布を**カノニカル分布**と呼ぶことにする. カノニカル分布に関する期待値と確率をそれぞれ $E_\beta[\ ]$, $P_\beta(\ )$ と書く. 確率変数 H のカノニカル分布に関する平均は

$$E_\beta[H] = \frac{E[He^{-\beta H}]}{Z(\beta)} = U(\beta)$$

になる. 以下では $u = U(\beta)$, $\delta > 0$ と仮定する. このとき $S(u)$ の定義より,

$$S(u) = \beta u + \log Z(\beta).$$

δ 以下の任意の $\varepsilon > 0$ を取る. カノニカル分布に関する大数の弱法則より

$$\lim_{n\to\infty} P_\beta\left(\frac{1}{n}\sum_{k=1}^{n} H_k \in (u-\varepsilon, u+\varepsilon)\right) = 1.$$

そして, カノニカル分布での確率と母集団分布での確率のあいだに以下の関係がある:

$$\begin{aligned}
P_\beta\left(\frac{1}{n}\sum_{k=1}^{n} H_k \in (u-\varepsilon, u+\varepsilon)\right) &= \frac{E[1_{H_1+\cdots+H_n\in(nu-n\varepsilon,nu+n\varepsilon)}\, e^{-\beta(H_1+\cdots+H_n)}]}{Z(\beta)^n} \\
&\leqq Z(\beta)^{-n} E[1_{H_1+\cdots+H_n\in(nu-n\varepsilon,nu+n\varepsilon)}\, e^{-n\beta u+n|\beta|\varepsilon}] \\
&= e^{-n(\beta u+\log Z(\beta)-|\beta|\delta)} P\left(\frac{1}{n}\sum_{k=1}^{n} H_k \in (u-\varepsilon, u+\varepsilon)\right) \\
&\leqq e^{-n(\beta u+\log Z(\beta)-|\beta|\delta)} P\left(\frac{1}{n}\sum_{k=1}^{n} H_k \in (u-\delta, u+\delta)\right).
\end{aligned}$$

以上の結果を合わせると

$$P\left(\frac{1}{n}\sum_{k=1}^{n} H_k \in (u-\delta, u+\delta)\right) \geqq e^{n(\beta u+\log Z(\beta)-|\beta|\varepsilon)}(1+o(1)) \qquad (n\to\infty).$$

ゆえに両辺の対数の $1/n$ 倍の $n\to\infty$ での極限を取ることによって次を得る:

$$\liminf_{n\to\infty} \frac{1}{n}\log P\left(\frac{1}{n}\sum_{k=1}^{n} H_k \in (u-\delta, u+\delta)\right) \geqq \beta u + \log Z(\beta) - |\beta|\varepsilon = S(u) - |\beta|\varepsilon.$$

$\varepsilon > 0$ はいくらでも小さくできるので,

$$\liminf_{n\to\infty} \frac{1}{n}\log P\left(\frac{1}{n}\sum_{k=1}^{n} H_k \in (u-\delta, u+\delta)\right) \geqq S(u).$$

これで (2) も示された. □

定理 6.1 の証明. (1) の上からの評価を証明しよう. F は $\mathbb{R}$ の閉部分集合であるとし,

$$F_+ = \{\, u \in F \mid u \geqq U(0) = E[H] \,\}, \qquad F_- = \{\, u \in F \mid u \leqq U(0) = E[H] \,\}$$

とおく. これらも $\mathbb{R}$ の閉部分集合なので, F_- の最大値 u_- と F_+ の最小値 u_+ が存在する. $S(u)$ は $u \geqq U(0) = E[H]$ で単調減少し, $u \leqq U(0) = E[H]$ で単調増加するので

$$\sup_{u\in F_+} S(u) = S(u_+), \quad \sup_{u\in F_-} S(u) = S(u_-), \quad \sup_{u\in F} S(u) = \max\{S(u_+), S(u_-)\}.$$

ゆえに補題 6.2 (1) より,

$$\begin{aligned}
&\frac{1}{n}\log P\left(\frac{1}{n}\sum_{k=1}^n H_k \in F_+\right) \leqq \frac{1}{n}\log P\left(\frac{1}{n}\sum_{k=1}^n H_k \geqq u_+\right) \leqq S(u_+),\\
&\frac{1}{n}\log P\left(\frac{1}{n}\sum_{k=1}^n H_k \in F_-\right) \leqq \frac{1}{n}\log P\left(\frac{1}{n}\sum_{k=1}^n H_k \leqq u_-\right) \leqq S(u_-),\\
&P\left(\frac{1}{n}\sum_{k=1}^n H_k \in F\right) \leqq e^{nS(u_+)} + e^{nS(u_-)} \leqq 2e^{n\sup_{u\in F} S(u)}.
\end{aligned}$$

したがって

$$\limsup_{n\to\infty} \frac{1}{n}\log P\left(\frac{1}{n}\sum_{k=1}^n H_k \in F\right) \leqq \sup_{u\in F} S(u).$$

これで (1) が示された.

(2) の下からの評価を証明しよう. G は $\mathbb{R}$ の開部分集合であると仮定する. 任意に $\varepsilon > 0$ を取ると, ある $u \in G$ で

$$S(v) \geqq \sup_{u\in G} S(u) - \varepsilon$$

を満たすものが存在する. G は開部分集合なので, ある $\delta > 0$ で $(v-\delta, v+\delta) \subset G$ を満たすものを取れる. このとき, 補題 6.2 (2) より

$$\begin{aligned}
\liminf_{n\to\infty} \frac{1}{n}\log P\left(\frac{1}{n}\sum_{k=1}^n H_k \in G\right) &\geqq \liminf_{n\to\infty} \frac{1}{n}\log P\left(\frac{1}{n}\sum_{k=1}^n H_k \in (v-\delta, v+\delta)\right)\\
&\geqq S(v) \geqq \sup_{u\in G} S(u) - \varepsilon.
\end{aligned}$$

したがって

$$\liminf_{n\to\infty} \frac{1}{n}\log P\left(\frac{1}{n}\sum_{k=1}^n H_k \in G\right) \geqq \sup_{u\in G} S(u).$$

これで (2) も示された.

最後に (3) を示そう. A は $\mathbb{R}$ の部分集合であるとし, A の開核を G と書き, G の閉包を F と書く. $A \subset F$ と仮定する. このとき $G \subset A \subset F$ なので

$$\sup_{u\in G} S(u) \leqq \liminf_{n\to\infty} \frac{1}{n}\log P\left(\frac{1}{n}\sum_{k=1}^n H_k \in G\right) \leqq \liminf_{n\to\infty} \frac{1}{n}\log P\left(\frac{1}{n}\sum_{k=1}^n H_k \in A\right)$$

$$\leqq \limsup_{n\to\infty} \frac{1}{n} \log P\left(\frac{1}{n}\sum_{k=1}^{n} H_k \in A\right) \leqq \limsup_{n\to\infty} \frac{1}{n} \log P\left(\frac{1}{n}\sum_{k=1}^{n} H_k \in F\right) \leqq \sup_{u\in F} S(u),$$

$$\sup_{u\in G} S(u) \leqq \sup_{u\in A} S(u) \leqq \sup_{u\in F} S(u).$$

ゆえに $\sup_{u\in G} S(u) = \sup_{u\in F} S(u)$ ならば

$$\lim_{n\to\infty} \frac{1}{n} \log P\left(\frac{1}{n}\sum_{k=1}^{n} H_k \in A\right) = \sup_{u\in A} S(u).$$

これで示すべきことがすべて示された. □

6.3 カノニカル分布の相対エントロピーとの関係

$q = (q_1, \ldots, q_r)$ は有限集合 $\{1, \ldots, r\}$ 上の確率分布であるとする:

$$q_i \geqq 0, \qquad q_1 + \cdots + q_r = 1.$$

確率変数 H すなわち有限集合 $\{1, \ldots, r\}$ 上の函数 $H(i) = E_i \in \mathbb{R}$ を考える. このとき

$$Z(\beta) = E[e^{-\beta H}] = \sum_{i=1}^{r} e^{-\beta E_i} q_i.$$

カノニカル分布 $p(\beta) = (p_1(\beta), \ldots, p_r(\beta))$ は

$$p_i(\beta) = \frac{e^{-\beta E_i} q_i}{Z(\beta)}$$

と定義され,

$$U(\beta) = E_\beta[H] = \sum_{i=1}^{r} E_i p_i(\beta) = \frac{1}{Z(\beta)} \sum_{i=1}^{r} E_i e^{-\beta E_i} q_i.$$

ゆえに

$$\begin{aligned} S(U(\beta)) &= \beta U(\beta) + \log Z(\beta) = -(-\beta U(\beta) - \log Z(\beta)) \\ &= -\sum_{i=1}^{r} p_i(\beta)(-\beta E_i - \log Z(\beta)) = -\sum_{i=1}^{r} p_i(\beta) \log \frac{p_i(\beta)}{q_i} = S(p(\beta)||q). \end{aligned}$$

すなわち, $\log Z(\beta)$ の Legendre 変換で定義された $S(u)$ に $u = U(\beta)$ を代入した結果はカノニカル分布の相対エントロピーに一致する.

6.4 ガンマ分布の場合の例

確率変数 H が形状 $\alpha > 0$, スケール $\tau > 0$ のガンマ分布にしたがうとは

$$E[f(H)] = \frac{1}{\Gamma(\alpha)\tau^\alpha} \int_0^\infty f(x) e^{-x/\tau} x^{\alpha-1}\, dx$$

が成立することである. このとき $\beta > -1/\tau$ ならば

$$Z(\beta) = E[e^{-\beta H}] = \frac{1}{\Gamma(\alpha)\tau^\alpha}\int_0^\infty e^{-(1+\tau\beta)x/\tau}x^{\alpha-1}\,dx = \frac{1}{(1+\tau\beta)^\alpha},$$
$$U(\beta) = \frac{(1+\tau\beta)^\alpha}{\Gamma(\alpha)\tau^\alpha}\int_0^\infty e^{-(1+\tau\beta)x/\tau}x^{\alpha}\,dx = \frac{(1+\tau\beta)^\alpha}{\Gamma(\alpha)\tau^\alpha}\frac{\Gamma(\alpha+1)\tau^{\alpha+1}}{(1+\tau\beta)^{\alpha+1}} = \frac{\tau\alpha}{1+\tau\beta} > 0.$$

ここでよく使われる次の公式を使った[21]:

$$\int_0^\infty e^{-cx}x^{s-1}\,dx = \frac{\Gamma(s)}{c^s} \qquad (s, c > 0).$$

ゆえに

$$S(U(\beta)) = \beta U(\beta) + \log Z(\beta) = \frac{\tau\alpha\beta}{1+\tau\beta} - \alpha\log(1+\tau\beta)$$
$$= \alpha - \frac{\alpha}{1+\tau\beta} - \alpha\log(1+\tau\beta).$$

これに $U(\beta) = u > 0$ と

$$1+\tau\beta = \frac{\tau\alpha}{u}, \qquad \beta = \frac{\alpha}{u} - \frac{1}{\tau}$$

を代入すると,

$$S(u) = \alpha - \frac{u}{\tau} - \alpha\log\frac{\tau\alpha}{u} = \alpha - \alpha\log\alpha - \left(\frac{u}{\tau} - \alpha\log\frac{u}{\tau}\right).$$

$S(u)$ は $u = U(0) = \tau\alpha$ で最大値 0 になる.

$H_1, H_2, \ldots$ が独立同分布な確率変数列で各々が H と同じ形状 α, スケール τ のガンマ分布にしたがうとき, ガンマ分布の再生性より[22], $H_1 + \cdots + H_n$ は形状 $n\alpha$, スケール τ の

[21] 使った公式はガンマ函数の定義から置換積分によって容易に示される. $U(\beta)$ に関する公式は $U(\beta) = -(\partial/\partial\beta)\log Z(\beta)$ を使えばより簡単に得られる.

[22] $H_1 + \cdots + H_n$ のモーメント母函数はは H のモーメント母函数の n 乗 $Z(\beta)^n = (1+\tau\beta)^{-n\alpha}$ に等しい. これは形状 $n\alpha$, スケール τ のガンマ分布のモーメント母函数に一致する. このことから $H_1 + \cdots + H_n$ が形状 $n\alpha$, スケール τ のガンマ分布にしたがうことがわかる. 直接的な計算によってもそのことを示せる. H, K は独立な確率変数であり, それぞれ形状 α, β, スケール τ のガンマ分布にしたがうならば

$$\begin{aligned}
E[f(H+K)] &= \frac{1}{\Gamma(\alpha)\Gamma(\beta)\tau^{\alpha+\beta}}\int_0^\infty\left(\int_0^\infty f(x+y)e^{-(x+y)/\tau}x^{\alpha-1}y^{\beta-1}\,dy\right)dx \\
&= \frac{1}{\Gamma(\alpha)\Gamma(\beta)\tau^{\alpha+\beta}}\int_0^\infty\left(\int_x^\infty f(t)e^{-t/\tau}x^{\alpha-1}(t-x)^{\beta-1}\,dt\right)dx \\
&= \frac{1}{\Gamma(\alpha)\Gamma(\beta)\tau^{\alpha+\beta}}\int_0^\infty f(t)e^{-t/\tau}\left(\int_0^y x^{\alpha-1}(t-x)^{\beta-1}\,dx\right)dy \\
&= \frac{1}{\Gamma(\alpha)\Gamma(\beta)\tau^{\alpha+\beta}}\int_0^\infty f(t)e^{-t/\tau}\left(\int_0^1 t^{\alpha-1}u^{\alpha-1}t^{\beta-1}(1-u)^{\beta-1}t\,du\right)dy \\
&= \frac{B(\alpha,\beta)}{\Gamma(\alpha)\Gamma(\beta)\tau^{\alpha+\beta}}\int_0^\infty f(t)e^{-t/\tau}t^{\alpha+\beta-1}\,dt \\
&= \frac{1}{\Gamma(\alpha+\beta)\tau^{\alpha+\beta}}\int_0^\infty f(t)e^{-t/\tau}t^{\alpha+\beta-1}\,dt
\end{aligned}$$

2 つ目の等号で $y = t - x$ とおいて y から t に積分変数を置換し, 4 つ目の等号で $x = tu$ とおいて x から u に積分変数を変換した. これより $H + K$ が形状 $\alpha + \beta$, スケール τ のガンマ分布にしたがうことがわかる. ガンマ分布は形状について再生性を持つと言う.

ガンマ分布にしたがうので,

$$E\left[f\left(\frac{1}{n}\sum_{k=1}^{n}H_k\right)\right]=\frac{1}{\Gamma(n\alpha)\tau^{n\alpha}}\int_0^\infty f\left(\frac{y}{n}\right)e^{-y/\tau}y^{n\alpha-1}\,dy.$$

ゆえに $0\leqq a<b$ のとき

$$P\left(a<\frac{1}{n}\sum_{k=1}^{r}H_k<b\right)=\frac{1}{\Gamma(n\alpha)\tau^{n\alpha}}\int_{na}^{nb}e^{-y/\tau}y^{n\alpha-1}\,dy=\frac{n^{n\alpha}}{\Gamma(n\alpha)}\int_{a/\tau}^{b/\tau}e^{-nx}x^{n\alpha-1}\,dx.$$

2 つ目の等号で $y=n\tau x$ とおいた. Cramér の定理より,

$$\lim_{n\to\infty}\left(\frac{1}{n}\log\frac{n^{n\alpha}}{\Gamma(n\alpha)}\int_{a/\tau}^{b/\tau}e^{-nx}x^{n\alpha-1}\,dx\right)=\sup_{a<u<b}S(u)=\begin{cases}S(b) & (b/\tau<\alpha),\\ 0 & (a/\tau\leqq\alpha\leqq b/\tau),\\ S(a) & (\alpha<a/\tau).\end{cases}$$

Stirling の公式より $-\log\Gamma(n\alpha)=-n\alpha\log n+n(\alpha-\alpha\log\alpha)+o(n)$ なので, これは次の公式と同値である ($A=a/\tau$, $B=b/\tau$):

$$\begin{aligned}&\lim_{n\to\infty}\frac{1}{n}\log\left(\int_A^B e^{-nx}x^{n\alpha-1}\,dx\right)\\ &\quad=\sup_{a/\tau<x<b/\tau}(\alpha\log x-x)=\begin{cases}\alpha\log B-B & (B<\alpha),\\ \alpha\log\alpha-\alpha & (A\leqq\alpha\leqq B),\\ \alpha\log A-A & (\alpha<A).\end{cases}\end{aligned}$$

この公式は

$$e^{-nx}x^{n\alpha-1}=\exp\left(n(\alpha\log x-x)+o(n)\right)$$

に Laplace の方法を適用することによって直接に示される.

6.5 Sanov の定理が拡張された Cramér の定理の特別な場合であること

以下の説明は厳密ではない.

まず, 複数の逆温度を持つ場合に拡張された Cramér の定理について大雑把に説明しよう. $\mathbb{R}^r$ の標準内積を $\langle\ ,\ \rangle$ と書く. X は $\mathbb{R}^r$ に値を持つ確率変数であるとし, $\beta=(\beta_1,\dots,\beta_r)\in\mathbb{R}^r$ に関する分配函数と Massieu 函数を

$$Z(\beta)=E[e^{-\langle\beta,X\rangle}],\qquad \Psi(\beta)=\log Z(\beta)$$

と定める. $\Psi(\beta)$ は $\beta\in\mathbb{R}^r$ に関する下に凸な函数になる. 実際 $a=(a_1,\dots,a_r)\in\mathbb{R}^r$ に対して $\partial_a=\sum_{i=1}^r a_i\partial/\partial\beta_i$, $X_a=\langle a,X\rangle$ とおくと

$$\partial_a^2\Psi(\beta)=\frac{E[X_a^2e^{\langle\beta,X\rangle}]\,E[e^{-\langle\beta,X\rangle}]-E[X_ae^{-\langle\beta,X\rangle}]^2}{Z(\beta)^2}=\frac{E[(X_a-Y_a)^2e^{\langle\beta,X+Y\rangle}]}{2Z(\beta)^2}\geqq 0.$$

ここで Y は X と独立で X と同じ分布にしたがう確率変数である. (もしも X_a が定数になるならば $\partial_a^2\Psi(\beta)=0$ となることに注意せよ.) $p\in\mathbb{R}^r$ の函数 $S(p)$ を次のように定める ($\Psi(\beta)$ の Legendre 変換):

$$S(p)=\inf_{\beta\in\mathbb{R}^r}(\langle\beta,p\rangle+\Psi(\beta)).$$

$\langle\beta,p\rangle+\Psi(\beta)$ を β_i で偏微分した結果を 0 とおくと

$$p_i=-\frac{\partial}{\partial\beta_i}\Psi(\beta)=-\frac{\partial}{\partial\beta_i}\log Z(\beta)=\frac{E[X_ie^{-\langle\beta,X\rangle}]}{Z(\beta)}=:p_i(\beta).$$

ここで X_i は X の第 i 成分である (確率変数列中の i 番目の確率変数ではないことに注意せよ). これで $p(\beta)=(p_1(\beta),\dots,p_r(\beta))$ が定まった. この条件を満たす β を $\beta(p)=(\beta_1(p),\dots,\beta_r(p))$ と書く. (この辺はかなり荒っぽい議論になっている. 実際にはそれなりに細かい注意が必要になる.) このとき

$$S(p(\beta))=\langle\beta,p(\beta)\rangle+\Psi(\beta),\qquad S(p)=\langle\beta(p),p\rangle+\Psi(\beta(p)).$$

$X^{(1)},X^{(2)},\dots$ は独立同分布の $\mathbb{R}^r$ 確率変数列であるとし, X と同じ確率分布にしたがうと仮定し, $A\subset\mathbb{R}^r$ とする. このとき適切な条件のもとで

$$\lim_{n\to\infty}\frac{1}{n}\log P\left(\frac{1}{n}\sum_{k=1}^n X^{(k)}\in A\right)=\sup_{p\in A}S(p)$$

が成立する (拡張された Cramér の定理).

Sanov の定理は拡張された Cramér の定理の特別な場合であることを説明しよう.

有限集合 $\{1,2,\dots,r\}$ 上の確率分布全体の集合を $\mathcal{P}\subset\mathbb{R}^r$ と書く:

$$\mathcal{P}=\{\,p=(p_1,\dots,p_r)\in\mathbb{R}^r\mid p_1,\dots,p_r\geqq 0,\ p_1+\cdots+p_r=1\,\}.$$

ベクトル $e_i\in\mathbb{R}^r$ を第 i 成分だけが 1 で他が 0 であると定める. このとき $e_i\in\mathcal{P}$ である. $q=(q_1,\dots,q_r)\in\mathbb{R}^r$ を任意に固定し, X は確率 q_i で値 e_i になる確率変数であるとする. この場合に上で説明した拡張された Cramér の定理を適用すると Sanov の定理が得られる. そのためにはこの場合に

$$S(p)=-\sum_{i=1}^r p_i\log\frac{p_i}{q_i}=S(p||q)$$

となることを示せばよい. この場合には

$$Z(\beta)=\sum_{i=1}^r e^{-\langle\beta,e_i\rangle}q_i=\sum_{i=1}^r e^{-\beta_i}q_i,\qquad\Psi(\beta)=\log Z(\beta).$$

この $\Psi(\beta)$ の Legendre 変換が相対エントロピー $S(p||q)$ になることを示したい. 上で説明したように,

$$p_i=p_i(\beta)=-\frac{\partial}{\partial\beta_i}\Psi(\beta)=-\frac{\partial}{\partial\beta_i}\log Z(\beta)=\frac{e^{-\beta_i}q_i}{Z(\beta)}.$$

なので, 対数を取ることによって

$$\Psi(\beta) = \log Z(\beta) = \sum_{i=1}^{r} p_i \log Z(\beta)$$
$$= \sum_{i=1}^{r} p_i(-\beta_i + \log q_i - \log p_i) = -\langle \beta, p \rangle - \sum_{i=1}^{r} p_i \log \frac{p_i}{q_i}.$$

ゆえに, 上の方で説明した $S(p)$ に関する公式を使うと,

$$S(p) = \langle \beta, p \rangle + \Psi(\beta) = -\sum_{i=1}^{r} p_i \log \frac{p_i}{q_i} = S(p||q).$$

このように Cramér の定理の線形空間に値を持つ確率変数への拡張の特別な場合として Sanov の定理が得られる.

6.6 $\Psi(\beta) = \log \sum_{i=1}^{r} e^{-\beta_i} q_i$ の Legendre 変換は相対エントロピー

有限集合 $\{1, 2, \ldots, r\}$ 上の確率分布全体の集合を $\mathcal{P} \subset \mathbb{R}^r$ と書く:

$$\mathcal{P} = \{\, p = (p_1, \ldots, p_r) \in \mathbb{R}^r \mid p_1, \ldots, p_r \geqq 0,\ p_1 + \cdots + p_r = 1 \,\}.$$

簡単のため, 確率分布 $q = (q_1, \ldots, q_r) \in \mathcal{P}$ は $q_i > 0$ $(i = 1, \ldots, r)$ を満たしていると仮定する. (以下の計算は $q \in \mathcal{P}$ であっても, q_i たちがすべて正の実数ならば成立している.)

多変数の分配函数 $Z(\beta)$ と Massieu 函数 $\Psi(\beta)$ と確率分布 $p(\beta) = (p_1(\beta), \ldots, p_r(\beta)) \in \mathcal{P}$ を次のように定める:

$$Z(\beta) = \sum_{i=1}^{r} e^{-\beta_i} q_i, \qquad \Psi(\beta) = \log Z(\beta), \qquad p_i(\beta) = -\frac{\partial}{\partial \beta_i} \Psi(\beta) = \frac{e^{-\beta_i} q_i}{Z(\beta)}$$

ここで $\beta = (\beta_1, \ldots, \beta_r) \in \mathbb{R}^r$ である. これらの基本性質をまとめておこう.

$\Psi(\beta)$ は前節で一般の場合に証明しておいたように下に凸な函数である.

$e = (1, \ldots, 1) \in \mathbb{R}^r$ とおく. このとき

$$Z(\beta + \lambda e) = e^{-\lambda} Z(\beta), \qquad \Psi(\beta + \lambda e) = \Psi(\beta) - \lambda, \qquad p(\beta + \lambda e) = p(\beta).$$

特に確率分布 $p(\beta)$ は β の e 方向の変化で一定である.

任意の確率分布 $p = (p_1, \ldots, p_r) \in \mathcal{P}$ に対して, $\beta \in \mathbb{R}^r$ を

$$p_i = e^{-\beta_i - \lambda} q_i \qquad (\lambda \in \mathbb{R} \text{ は任意定数})$$

によって定めると, 上に述べたことより $p = p(\beta)$ となる:

$$e^{-\beta_i} q_i = e^{\lambda} p_i, \qquad Z(\beta) = \sum_{i=1}^{r} e^{-\beta_i} q_i = \sum_{i=1}^{r} e^{\lambda} p_i = e^{\lambda}, \qquad p_i(\beta) = \frac{e^{-\beta_i} q_i}{Z(\beta)} = p_i.$$

ゆえに $p(\beta)$ の像は $\mathcal{P}$ 全体に一致する.

$p = (p_1, \ldots, p_r) \in \mathbb{R}^r$ に対して $\beta \in \mathbb{R}^r$ の函数 $f(\beta)$ を

$$f(\beta) = \langle \beta, p \rangle + \Psi(\beta)$$

と定める. $f(\beta)$ は下に凸な函数なので, もしも $f(\beta)$ が最小値を持つならば, その最小点で $f(\beta)$ の偏微分係数はすべて 0 になる. $f(\beta)$ の偏導函数は

$$\frac{\partial}{\beta_i} f(\beta) = \frac{\partial}{\partial \beta_i} \sum_{i=1}^r \beta_i p_i + \frac{\partial}{\partial \beta_i} \Psi(\beta) = p_i - p_i(\beta)$$

となるので, $f(\beta)$ が最小値を持つための必要十分条件は $p \in \mathcal{P}$ となることである. それ以外のとき $f(\beta)$ は最小値を持たず $\inf_{\beta \in \mathbb{R}^r} f(\beta) = -\infty$ となる.

したがって, $\Psi(\beta)$ の Legendre 変換 $S(p)$ を

$$S(p) = \inf_{\beta \in \mathbb{R}^r} (\langle \beta, p \rangle + \Psi(\beta))$$

と定めると, $S(p)$ の定義域は $\mathcal{P}$ に一致する (定義域の外での値は $-\infty$ になると考える). $p \in \mathcal{P}$ のとき, $p = p(\beta(p))$ を満たす $\beta(p) \in \mathbb{R}^r$ を任意に取ると ($\beta(p)$ の取り方には e 方向の平行移動に関する不定性がある),

$$S(p) = \langle \beta(p), p \rangle + \Psi(\beta(p)).$$

$\beta = \beta(p)$ を上のように $p_i = e^{-\beta_i - \lambda} q_i$ のように取ると

$$\beta_i = -\log \frac{p_i}{q_i} - \lambda, \qquad Z(\beta) = e^{\lambda}, \qquad \Psi(\beta) = \lambda$$

なので

$$S(p) = \sum_{i=1}^r \left(-\log \frac{p_i}{q_i} - \lambda \right) p_i + \lambda = -\sum_{i=1}^r p_i \log \frac{p_i}{q_i} = S(p||q).$$

このように $\Psi(\beta)$ の Legendre 変換は相対エントロピー $S(p||q)$ に一致する. $S(p)$ は上に凸な函数になる.

逆に $S(p)$ の Legendre 変換

$$F(\beta) = \sup_{p \in \mathcal{P}} (-\langle \beta, p \rangle + S(p))$$

が $\Psi(\beta)$ に一致することも直接の計算によって確認しておこう (下半連続な凸函数については常にこうなる). 函数 $g(p)$ を

$$g(p) = -\langle \beta, p \rangle + S(p)$$

と定める. $g(p)$ は p に関して上に凸な函数である. $g(p)$ の $\mathcal{P}$ 上での最大値問題を考えたい. Lagrange の未定乗数法を使うために p と λ の函数 L を

$$L = g(p) - (\lambda - 1) \left(\sum_{i=1}^r p_i - 1 \right) = -\langle \beta, p \rangle + S(p) - (\lambda - 1) \left(\sum_{i=1}^r p_i - 1 \right)$$

と定める. このとき

$$-\frac{\partial L}{\partial \lambda} = \sum_{i=1}^r p_i - 1,$$

$$\frac{\partial L}{\partial p_i} = -\beta_i - \log\frac{p_i}{q_i} - 1 - (\lambda - 1) = -\beta_i - \log\frac{p_i}{q_i} - \lambda.$$

これより, $\sum_{i=1}^r p_i = 1$ という条件のもとで $g(p)$ が最大になるのは

$$e^\lambda = \sum_{i=1}^r e^{-\beta_i} q_i = Z(\beta), \quad p_i = e^{-\lambda-\beta_i} q_i = \frac{e^{-\beta_i} q_i}{Z(\beta)}, \quad \log Z(\beta) = \lambda = -\beta_i - \log\frac{p_i}{q_i}$$

のときであることがわかる. このとき,

$$\Psi(\beta) = \log Z(\beta) = \lambda = \sum_{i=1}^r p_i \lambda = \sum_{i=1}^r p_i \left(-\beta_i - \log\frac{p_i}{q_i}\right) = -\langle \beta, p\rangle + S(p) = F(\beta).$$

すなわち $S(p)$ の Legendre 変換 $F(\beta)$ は $\Psi(\beta)$ に一致する.

以上の結果は次のようにまとめられる.

命題 6.3. $q = (q_1, \ldots, q_r)$ の成分はすべて正の実数であるとする. $\mathbb{R}^r$ 上の下に凸な函数 $\Psi(\beta)$ と $\mathcal{P}$ 上の上に凸な函数 $S(p)$ を次のように定める:

$$\Psi(\beta) = \log \sum_{i=1}^r e^{-\beta_i} q_i, \qquad S(p) = -\sum_{i=1}^r p_i \log\frac{p_i}{q_i}.$$

これらは互いに相手の Legendre 変換になっている. すなわち

$$S(p) = \inf_{\beta\in\mathbb{R}^r} (\langle \beta, p\rangle + \Psi(\beta)), \qquad \Psi(\beta) = \sup_{p\in\mathcal{P}} (-\langle \beta, p\rangle + S(p)).$$

が成立している. □

注意 6.4. 上の命題の書き換え. $\mathcal{P}$ 上の函数 $D(p)$ を

$$D(p) = \sum_{i=1}^r p_i \log\frac{p_i}{q_i}$$

と定めると, $D(p)$ も $\Psi(\beta)$ と同様に下に凸な函数になり,

$$D(p) = \sup_{\beta\in\mathbb{R}^r} (-\langle \beta, p\rangle - \Psi(\beta)), \qquad \Psi(\beta) = \sup_{p\in\mathcal{P}} (-\langle \beta, p\rangle - D(p))$$

が成立している. □

7 付録: 統計力学との関係?

この節では数学的に厳密な議論をするつもりはない. 十分に理解していることを書くつもりもない. このノートの内容と “標準的な” 統計力学の関係について, 筆者が理解を深めるために書いたラフなスケッチを以下に記録しておく.

2016 年 7 月 13 日: まず, Cramér の定理の一般化に関する第 7.1 節を書いた.

2016 年 7 月 14 日: 統計力学の教科書におけるカノニカル分布の導出の仕方を紹介した第 7.2 節を書いた.

2016 年 7 月 21 日: 統計力学の教科書におけるカノニカル分布の導出の説明をやり直すための第 7.3 節を書いた.

7.1 パラメーターに関する分配函数の漸近挙動を仮定した場合

確率変数の独立同分布の仮定を確率変数のあるパラメーターに関する漸近挙動に関する仮定に置き換えることによって, Cramér の定理に関する第 6 節とほぼ同じ議論を繰り返そう.

H_λ はパラメーター λ を持つ確率変数であるとし, H_λ の確率分布は $\mathbb{R}$ 上の確率測度 μ_λ で記述されているとする:

$$E[f(H_\lambda)] = \int_{\mathbb{R}} f(x)\,\mu_\lambda(dx).$$

パラメーター λ は系のサイズ (例えば体積 V) を表わし, H_λ は系の全エネルギーを表わしていると考える.

分配函数 $Z(\beta,\lambda)$ とその対数 $\Psi(\beta,\lambda)$ (Massieu 函数) を次のように定義する:

$$Z(\beta,\lambda) = E[e^{-\beta H_\lambda}], \qquad \Psi(\beta,\lambda) = \log Z(\beta,\lambda).$$

さらに $\mathbb{R}$ 上の確率測度 (カノニカル分布)

$$\mu_{\beta,\lambda}(dx) = \frac{e^{-\beta x}\,\mu_\lambda(dx)}{Z(\beta,\lambda)}$$

の定める確率と期待値をそれぞれ $P_\beta(\ \ ), \langle\ \ \rangle_\beta = E_\beta[\ \ ]$ と書く.

$\Psi(\beta,\lambda) = \log Z(\beta,\lambda)$ は $\lambda\to\infty$ で

$$\Psi(\beta,\lambda) = \lambda(\psi(\beta)+\eta_\lambda(\beta)), \qquad \eta_\lambda(\beta) = o(1), \quad \eta'_\lambda(\beta) = o(1), \quad \eta''_\lambda(\beta) = o(1)$$

と振る舞うと仮定する. このとき

$$\langle H_\lambda\rangle_\beta = \frac{E[H_\lambda e^{-\beta H_\lambda}]}{Z(\beta,\lambda)} = -\frac{\partial}{\partial\beta}\Psi(\beta,\lambda) = -\lambda(\psi'(\beta)+o(1))$$

なので

$$u(\beta) = -\psi'(\beta)$$

とおくと

$$\left\langle \frac{H_\lambda}{\lambda}\right\rangle_\beta = u(\beta)+o(1)\to u(\beta) \qquad (\lambda\to\infty).$$

さらに, 確率測度 $\mu_{\beta,\lambda}$ に関する H_λ の分散は, $Z = Z(\beta.\lambda)$ と書くと,

$$\begin{aligned} 0 \leqq \Big\langle (H_\lambda - \langle H_\lambda\rangle_\beta)^2\Big\rangle_\beta &= \langle H_\lambda^2\rangle_\beta - (\langle H_\lambda\rangle_\beta)^2 = \frac{Z_{\beta\beta}Z - (Z_\beta)^2}{Z^2} \\ &= \left(\frac{\partial}{\partial\beta}\right)^2 \Psi(\beta,\lambda) = \lambda(\psi''(\beta)+o(1)) \end{aligned}$$

に等しい. (この公式より $\Psi(\beta,\lambda)$ が β の函数として下に凸なこともわかる. 以下では $\psi(\beta)$ も下に凸であると仮定する.) ゆえに

$$\left(\frac{H_\lambda}{\lambda}\ \text{の}\ \mu_{\beta,\lambda}\ \text{に関する分散}\right) = \frac{\psi''(\beta)}{\lambda} + o\left(\frac{1}{\lambda}\right) = O\left(\frac{1}{\lambda}\right)\to 0 \qquad (\lambda\to\infty).$$

以上をまとめると, 確率測度 $\mu_{\beta,\lambda}$ (カノニカル分布) のもとでの H_λ/λ が "大数の法則" を満たしていることがわかる. すなわち, 確率測度 $\mu_{\beta,\lambda}$ のもとでの H_λ/λ の分布は $\lambda\to\infty$ で

$u(\beta) = -\psi'(\beta)$ に集中し, λ が大きいとき, その分散 (ゆらぎの大きさの 2 乗) は $\psi''(\beta)/\lambda$ 程度になる.

$\Psi(\beta, \lambda)$ は β の函数として下に凸なので, $\langle H_\lambda \rangle_\beta = -\Psi_\beta(\beta, \lambda)$ は β について単調減少函数である. $\psi(\beta)$ も下に凸になると仮定したので, $u(\beta) = -\psi'(\beta)$ も単調減少函数になる.

以下では簡単のため $\beta \geqq 0$ と仮定し, $u = u(\beta) = -\psi'(\beta) \leqq u(0)$ とおき,

$$s(u) = \beta u + \psi(\beta)$$

と定める. $U = U(\beta, \lambda) = \langle H_\lambda \rangle_\beta$,

$$S(U, \lambda) = \beta U + \Psi(\beta, \lambda)$$

とおくと, $\lambda \to \infty$ において $U = \lambda(u + o(1))$, $\Psi(\beta, \lambda) = \lambda(\psi(\beta) + o(1))$ なので,

$$S(U, \lambda) = \lambda(\beta u + \psi(\beta) + o(1)) = \lambda s(u) + o(\lambda) \qquad (\lambda \to \infty).$$

さらに,

$$\mu_\lambda(dx) = q_\lambda(x)\, dx$$

のとき,

$$p_{\beta,\lambda}(x) = \frac{e^{-\beta x} q_\lambda(x)}{Z(\beta, \lambda)}$$

とおくと,

$$\mu_{\beta,\lambda}(dx) = \frac{e^{-\beta x} q_\lambda(x)}{Z(\beta, \lambda)}\, dx = p_{\beta,\lambda}(x)\, dx$$

なので, $S(U, \lambda)$ は次のように変形される:

$$\begin{aligned} S(U(\beta, \lambda), \lambda)) &= \int_{\mathbb{R}} (\beta x + \log Z(\beta, \lambda)) p_{\beta,\lambda}(x)\, dx \\ &= -\int_{\mathbb{R}} \log\left(\frac{e^{-\beta x}}{Z(\beta, \lambda)}\right) p_{\beta,\lambda}(x)\, dx = -\int_{\mathbb{R}} p_{\beta,\lambda}(x) \log\left(\frac{p_{\beta,\lambda}(x)}{q_\lambda(x)}\right) dx. \end{aligned}$$

すなわち $S(U, \lambda)$ はカノニカル分布の相対エントロピーであり, $s(u)$ はサイズ $\lambda \to \infty$ におけるカノニカル分布の相対エントロピー密度 (1 サイズあたりの相対エントロピー) である.

まず $P(H_\lambda/\lambda \leqq u)$ の $\lambda \to \infty$ における上からの評価を示そう.

$$\begin{aligned} P\left(\frac{H_\lambda}{\lambda} \leqq u\right) &= E[1_{H_\lambda \leqq \lambda u}] \leqq E\left[1_{H_\lambda \leqq \lambda u} e^{-\beta(H_\lambda - \lambda u)}\right] \\ &\leqq E[e^{-\beta(H_\lambda - \lambda u)}] = e^{\lambda\beta u} Z(\beta, \lambda) = e^{\lambda\beta u + \Psi(\beta, \lambda)}. \end{aligned}$$

ここで $1_{H_\lambda \leqq \lambda u}$ は $H_\lambda \leqq \lambda u$ のとき 1 になり, 他のとき 0 になる函数を表わしている. 1 つ目の不等号では $1_{H_\lambda \leqq \lambda u} \leqq 1_{H_\lambda \leqq \lambda u} e^{-\beta(H_\lambda - \lambda u)}$ ($\beta \geqq 0$ より) を使い, 2 つ目の不等号では $1_{H_\lambda \leqq \lambda u} e^{-\beta(H_\lambda - \lambda u)} \leqq e^{-\beta(H_\lambda - \lambda u)}$ を使った. ゆえに $\Psi(\beta, \lambda) = \lambda(\psi(\beta) + o(1))$ より

$$\log P\left(\frac{H_\lambda}{\lambda} \leqq u\right) = \lambda(\beta u + \psi(\beta) + o(1)) = \lambda(s(u) + o(1)).$$

両辺を λ で割って $\lambda \to \infty$ とすると

$$\limsup_{\lambda\to\infty} \frac{1}{\lambda} P\left(\frac{H_\lambda}{\lambda} \leqq u\right) \leqq s(u).$$

次に下からの評価を示そう. $0 < \varepsilon \leqq \delta$ と仮定する. 確率測度 $\mu_{\beta,\mu}$ に関する H_λ/λ の分布は $\lambda \to \infty$ で $u = u(\beta)$ に集中するので ("大数の法則"),

$$P_\beta\left(\frac{H_\lambda}{\lambda} \in (u-\varepsilon, u+\varepsilon)\right) = e^{o(1)} \to 1 \qquad (\lambda \to \infty).$$

さらに, $\beta \geqq 0$ より,

$$\begin{aligned}
P_\beta\left(\frac{H_\lambda}{\lambda} \in (u-\varepsilon, u+\varepsilon)\right) &= \frac{E\left[1_{H_\lambda \in (\lambda u - \lambda\varepsilon, \lambda u + \lambda\varepsilon)}\, e^{-\beta H_\lambda}\right]}{Z(\beta,\lambda)} \\
&\leqq Z(\beta,\lambda)^{-1} E\left[1_{H_\lambda \in (\lambda u - \lambda\varepsilon, \lambda u + \lambda\varepsilon)}\, e^{-\lambda\beta u + \lambda\beta\varepsilon}\right] \\
&= e^{-\lambda\beta u + \lambda\beta\varepsilon - \Psi(\beta,\lambda)} P\left(\frac{H_\lambda}{\lambda} \in (u-\varepsilon, u+\varepsilon)\right) \\
&= e^{-\lambda\beta u + \lambda\beta\varepsilon - \lambda(\psi(\beta)+o(1))} P\left(\frac{H_\lambda}{\lambda} \in (u-\varepsilon, u+\varepsilon)\right) \\
&\leqq e^{-\lambda(s(u) - \beta\varepsilon + o(1))} P\left(\frac{H_\lambda}{\lambda} \in (u-\delta, u+\delta)\right).
\end{aligned}$$

したがって $\lambda \to \infty$ において

$$P\left(\frac{H_\lambda}{\lambda} \in (u-\delta, u+\delta)\right) \geqq e^{\lambda(s(u) - \beta\varepsilon + o(1)) + o(1)} = e^{\lambda(s(u) - \beta\varepsilon) + o(\lambda)}.$$

両辺の大数の $1/\lambda$ 倍の $\lambda \to \infty$ での極限を取ることによって

$$\liminf_{\lambda\to\infty} \frac{1}{\lambda} P\left(\frac{H_\lambda}{\lambda} \in (u-\delta, u+\delta)\right) \geqq s(u) - \beta\varepsilon.$$

$\varepsilon > 0$ はいくらでも小さくできるので,

$$\liminf_{\lambda\to\infty} \frac{1}{\lambda} P\left(\frac{H_\lambda}{\lambda} \in (u-\delta, u+\delta)\right) \geqq s(u).$$

以上によって, 第 6 節で証明した Cramér の定理の本質的部分に対応する不等式が得られた. よって, 第 6 節と同様の議論を繰り返すことによって, 次が成立していることがわかる:

$$\lim_{\lambda\to\infty} \frac{1}{\lambda} P\left(\frac{H_\lambda}{\lambda} \leqq u\right) = s(u).$$

このようにエネルギー密度 H_λ/λ が $u = u(\beta)$ 以下の確率の $\lambda \to \infty$ での漸近挙動は, カノニカル分布の相対エントロピー密度 $s(u)$ で記述される.

しかし, これだけだとあまりわかった気になれない. 標準的な統計力学に書いてあるカノニカル分布の導出との関係はどうなっているのだろうか?

7.2 統計力学の教科書におけるカノニカル分布の導出 (1)

以下の議論は本質的に田崎 [10], pp. 105–106 と同じ議論のつもりである. ただし, 我々は「等確率の原理」(等重率の原理) を仮定せずに議論を進めているので,「状態数」を「確率」に置き換えて考える. その結果エントロピーではなく, 相対エントロピーを考えることになる. (等確率の原理の仮定のもとでは状態数で確率が決まる. エントロピーは状態数の対数であり, 相対エントロピーは確率の対数である.)

$H_{N,V}$ は体積 V と粒子の個数 N に依存する系の全エネルギーを意味する確率変数であるとする (これは物理のノートではなく, 数学のノートなので等確率の原理を仮定しない).

$U_0(N,V) = E[H_{N,V}]$ とおく. $n = N/V$ が一定の条件のもとで $V \to \infty$ とするとき, $U_0(N,V) = Vu_0(n) + o(V)$ が成立していると仮定する. ($U_0(N,V) = \infty$, $u_0(n) = \infty$ であってもよい.) $V \to \infty$ で大数の法則 $P(|H_{N,V}/V - u_0(n)| \leqq \varepsilon) \to 1$ $(\varepsilon > 0)$ が成立しているならば, $H_{N,V}/V$ の分布は $u_0(n)$ に集中する. したがって U が $U_0(N,V)$ から (もしくは u が $u_0(n)$ から) 離れれば離れるほど確率が小さくなる傾向になり, 十分小さな U と u について確率は単調増加函数になる. (物理的に典型的な状況では状態数 (したがって確率) はエネルギーの単調増加函数になる.)

相対エントロピー $S(U,N,V)$ をエネルギー $H_{N,V}$ が U 以下になる確率の対数と定める:

$$S(U,N,V) = \log P(H_{N,V} \leqq U) \qquad (U \leqq U_0(N,V)).$$

$S(U,N,V)$ は U に関する単調増加函数になる. よくある状況ではエネルギーが U 程度になる確率は V が大きなとき $P(H_{N,V} \leqq U)$ でよく近似される ([10], p. 105 の「δ を消す議論」を見よ. さらにこのノートの第 2 節 と第 4 節の議論を比較してみよ.). エネルギー密度 u と個数密度 n を次のように定める:

$$u = \frac{U}{V}, \qquad n = \frac{N}{V}.$$

相対エントロピーは以下を満たしていると仮定する. ([10], p. 76 の (3.2.29) 式も見よ.)

仮定 7.1. 相対エントロピー $S(U,N,V)$ は U に関して上に凸な函数であると仮定する. さらに $S(U,N,V)$ は $V \to \infty$ で以下の漸近挙動を持つ:

$$S(U,N,V) = V\, s(u,n) + \eta(u,n,V), \qquad u = \frac{U}{V}, \quad n = \frac{N}{V}$$

と $\eta(u,n,V)$ を定めると,

$$\eta(u,n,V) = o(V), \qquad \eta_u(u,n,V) = o(V) \qquad (V \to \infty).$$

特に

$$\lim_{V\to\infty} \frac{1}{V} S(U,N,V) = \lim_{V\to\infty} \frac{1}{V} \log P(H_{N,V} \leqq U) = s(u,n)$$

が成立しており, $U = Vu$, $\partial/\partial U = V^{-1}\partial/\partial u$ より,

$$S_U(U,N,V) = \frac{1}{V}(V s_u(u,n) + \eta_u(u,n,V)) = s_u(u,n) + o(1). \qquad (\sharp)$$

さらに $s_u(u,n)$ の u に関する連続性も仮定しておく. $s(u,n)$ を (体積無限大の極限における) **相対エントロピー密度**と呼ぶ. $s(u,n)$ も u について上に凸な単調増加函数になると仮定する. 逆温度函数 $\beta(u,n) \geqq 0$ を

$$\beta(u,n) = s_u(u,n) = S_U(U,N,V) + o(1)$$

と定める. **第 7.1 節の議論は固定された $\lambda = V$ のケースで以上の仮定が (ほぼ) 成立しているような設定を与えていると考えられる.** □

田崎 [10], pp. 105–106 の議論によれば, カノニカル分布は本質的に

$$\lim_{V\to\infty} \frac{P(H_{N,V} \leqq U - E)}{P(H_{N,V} \leqq U)} = e^{-\beta(u,n)E} \qquad (*)$$

を示すことによって導出される. これは次と同値である:

$$\log \frac{P(H_{N,V} \leqq U - E)}{P(H_{N,V} \leqq U)} = -\beta(u,n)E + o(1) \qquad (V \to \infty).$$

これは次のように示される. 平均値の定理より, $0 < \theta < 1$ を満たすある θ が存在して,

$$\begin{aligned}
\log \frac{P(H_{N,V} \leqq U - E)}{P(H_{N,V} \leqq U)} &= S(U-E,N,V) - S(U,N,V) \\
&= -E\, S_U(U - \theta E, N, V) \\
&= -E \left(s_u\left(u - \frac{\theta E}{V}, n\right) + o(1) \right) \\
&= -E\, s_u(u,n) + o(1) \\
&= -\beta(u,n)E + o(1). \qquad (V \to \infty).
\end{aligned}$$

2 つ目の等号で平均値の定理を用い, 3 つ目の等号で (♯) を使い, 4 つ目の等号で $s_u(u,n)$ が連続であるという仮定を使った.

$(*)$ の形式で導出された「Boltzmann 因子」は以下のように解釈される.

熱浴と注目する系が接触しているような系を考える. 注目する系のサイズは一定であるとし, 熱浴のサイズが無限大になる極限を考える. 熱浴の側に注目して各種確率を計算する. U は系全体のエネルギーであり, E は注目する系のエネルギーだとすると, $U - E$ は熱浴のエネルギーになる. 熱浴のエネルギーが $U - E$ 程度になる確率は V が大きなとき $P(H_{N,V} \leqq U - E)$ で近似される. 注目する系のエネルギーが E になる確率は $P(H_{N,V} \leqq U - E)$ に比例するだろう. したがって $(*)$ より, 注目する系のエネルギーが E である確率は V が大きなとき近似的に Boltzmann 因子 $e^{-\beta(u,n)E}$ に比例する.

しかし, これでもまだわかった気になれない.

以上の解釈のより正確な定式化はどうなっているのだろうか?

十分な理解に達するためには「熱浴を含めた全体の系と注目する部分系」という設定を直接扱う必要がありそうだ.

第 2.4 節における標準正規分布の導出は本質的に統計力学における Maxwell-Boltzmann 分布の導出に等しい. その議論における「全体の系」は半径 $\sqrt{n}$ の $n-1$ 次元球面上の一様分布 (全エネルギー一定という条件で定義される高次元球面上の一様分布) であり, 「注目する部分系」はその分布の 1 次元部分空間への射影であり, 「熱浴」は残りの自由度である. そして全体の系のサイズを大きくする極限が $n \to \infty$ の極限に対応する. この設定を一般化しなければいけない.

7.3 統計力学の教科書におけるカノニカル分布の導出 (2)

前節の議論をやり直す.

以下, $H, H_k, H^{\mathrm{res}}, H^{\mathrm{tot}}$ などはすべて確率変数であるとする.

熱浴 (thermal reservoir, heat bath) に接している注目する系を考える. 熱浴の体積を表わすパラメーターを V と書く. 粒子数 N は省略して書くことにする.

注目する系のエネルギーを H と書き, 熱浴のエネルギーを H_V^{res} と書くと, 注目する系と熱浴を合わせた全体の系のエネルギーは $H_V^{\mathrm{tot}} = H + H_V^{\mathrm{res}}$ と表わされると仮定する.

熱浴と注目する系を合わせた全体の系において,

$$P(H = E_i \text{ and } H_V^{\mathrm{tot}} \leqq U) = P(H = E_i)\, P(H_V^{\mathrm{res}} \leqq U - E_i) \tag{1}$$

が成立していると仮定する.

熱浴と無関係に注目する系は確率 q_i で状態 i になり, 状態 i のエネルギーは E_i であるとする. H は確率 q_i で値 E_i になる確率変数である:

$$P(H = E_i) = q_i. \tag{2}$$

等重率の原理 (等確率の原理) を仮定した場合にはすべての状態 i が等確率で実現されると考えるが, 応用先を物理的な統計力学に限定したくないので等重率の原理を仮定しない. (以下の議論が示すように Boltzmann 因子の導出に等重率の原理は関係しない.)

熱浴は $V \to \infty$ で次の漸近挙動を満たしていると仮定する:

$$S^{\mathrm{res}}(U, V) := \log P(H_V^{\mathrm{res}} \leqq U) = V s^{\mathrm{res}}(u) + o(V), \qquad u = \frac{U}{V}.$$

$S^{\mathrm{res}}(U, V)$ を熱浴のエントロピーと呼び, $u = U/V$ をエネルギー密度と呼び, $s^{\mathrm{res}}(u)$ を熱浴のエントロピー密度と呼ぶ (C^1 級と仮定). $o(V)$ の部分の u による導函数達も $o(V)$ であると仮定する. このとき, $U = Vu$ より $\partial/\partial U = (1/V)\partial/\partial u$ なので, 平均値の定理より, 各 E ごとにある θ で $0 < \theta < 1$ を満たすものが存在して,

$$\begin{aligned}
\log P(H_V^{\mathrm{res}} \leqq U - E) &= S^{\mathrm{res}}(U - E, V) = S^{\mathrm{res}}(U, V) - E\frac{\partial S^{\mathrm{res}}}{\partial U}(U - \theta E, V) \\
&= S^{\mathrm{res}}(U, V) - E\frac{1}{V}\frac{\partial}{\partial u}\left(V s^{\mathrm{res}}\left(u - \frac{\theta E}{V}\right) + o(V)\right) \\
&= S^{\mathrm{res}}(U, V) - E\frac{\partial s^{\mathrm{res}}}{\partial u}(u) + o(1).
\end{aligned}$$

すなわち, **熱浴の逆温度** $\beta(u)$ を

$$\beta(u) = \frac{\partial s^{\mathrm{res}}}{\partial u}(u)$$

と定めると,

$$P(H_V^{\mathrm{res}} \leqq U - E) = P(H_V^{\mathrm{res}} \leqq U)\, e^{-\beta(u)E + o(1)} \qquad (V \to \infty).$$

以上の仮定と結論を, (1),(2) を使ってまとめると,

$$P(H = E_i \text{ and } H_V^{\mathrm{tot}} \leqq U) = P(H_V^{\mathrm{res}} \leqq U)\, q_i e^{-\beta(u)E_i + o(1)}.$$

さらに

$$P(H_V^{\mathrm{tot}} \leqq U) = \sum_i P(H = E_i \text{ and } H_V^{\mathrm{tot}} \leqq U) = P(H_V^{\mathrm{res}} \leqq U) \sum_i q_i e^{-\beta(u)E_i + o(1)}.$$

したがって, 条件付き確率を

$$P(\, H = E_i \mid H_V^{\mathrm{tot}} \leqq U\,) = \frac{P(H = E_i \text{ and } H_V^{\mathrm{tot}} \leqq U)}{P(H^{\mathrm{tot}} \leqq U)}$$

と書くと,

$$P(\, H = E_i \mid H_V^{\mathrm{tot}} \leqq U\,) = \frac{q_i e^{-\beta(u)E_i + o(1)}}{\sum_j q_j e^{-\beta(u)E_j + o(1)}} \to \frac{q_i e^{-\beta(u)E_i}}{\sum_j q_j e^{-\beta(u)E_j}} \qquad (V \to \infty).$$

すなわち, 全体のエネルギーが U 以下であるという条件のもとで, 注目する系のエネルギーが E_i である条件付き確率の熱浴の体積無限大での極限は次のようになる:

$$\lim_{V\to\infty} P(\, H = E_i \mid H_V^{\mathrm{tot}} \leqq U\,) = \frac{q_i e^{-\beta(u)E_i}}{\sum_j q_j e^{-\beta(u)E_j}}.$$

このようにして得られる注目する系に関する確率分布を逆温度 $\beta(u)$ の**カノニカル分布**と呼ぶ. 熱浴の漸近挙動から得られる因子 $e^{-\beta(u)E_i}$ を **Boltzmann 因子**と呼ぶ.

例 7.2. $H_1, H_2, \ldots$ は独立同分布な確率変数で穏当な条件を満たすものであるとし, $V = n = 1, 2, 3, \ldots$ とおき, $H_n^{\mathrm{tot}} = H_1 + H_2 + \cdots + H_n$, $H = H_1$, $H_n^{\mathrm{res}} = H_2 + \cdots + H_n$ とおくと, Sanov の定理もしくは Cramér の定理によって, 以上の議論の設定を満たしている. この場合の熱浴は独立同分布な試行の繰り返しだと解釈される.

上の方の議論は独立同分布の試行の繰り返しとみなせるような特殊な場合でなくても, 熱浴の体積無限大での漸近挙動に関する一般的な前提から Boltzmann 因子 $e^{-\beta(u)E_i}$ で記述されるカノニカル分布が普遍的に得られることを示している. □

8 付録: 他の種類のエントロピーについて

8.1 自由エネルギーや Massieu 函数との関係

注意 8.1 (モーメント母函数とキュムラント母函数). 確率分布 q_i のもとで確率変数 $X : i \mapsto X_i$ のモーメント母函数 $M_X(t)$ は

$$M_X(t) = \sum_{i=1}^{r} e^{tX_i} q_i$$

と定義される. これは $X = E$, $t = -\beta$ のとき分配函数

$$Z(\beta) = \sum_{i=1}^{r} e^{-\beta E_i} q_i$$

に一致する. 確率論の教科書に書いてあるモーメント母函数 (積率母函数) は分配函数と本質的に同じものだと思ってよい. 確率論の教科書によればモーメント母函数の対数

$$K_X(t) = \log M_X(t)$$

は確率変数 X のキュムラント母函数 (cumulant generating function) と呼ばれている. 自由エネルギーの定義

$$F(\beta) = -\frac{1}{\beta}\log Z(\beta)$$

は本質的にキュムラント母函数の定義に一致している. より正確には逆温度 β で割る前の

$$\mathcal{F}(\beta) = \log Z(\beta) \qquad \text{(より正確には右辺はその Boltzmann 定数倍)}$$

の方がキュムラント母函数の直接の対応物になる. こちらの $\mathcal{F}(\beta)$ は **Massieu 函数**と呼ばれている. □

8.2 相対 Rényi エントロピー

注意 8.2 (相対 Rényi エントロピー). 2 つの確率分布 $p = (p_1, \ldots, p_r)$, $q = (q_1, \ldots, q_r)$ に対して, **相対 Rényi エントロピー** $S_\beta(p||q)$ が

$$S_\beta(p||q) = -\frac{1}{\beta-1}\log\sum_{i=1}^{r}\left(\frac{p_i}{q_i}\right)^\beta q_i = -\frac{1}{\beta-1}\log\sum_{i=1}^{r}p_i^\beta q_i^{1-\beta}$$

と定義される. これの $\beta-1$ 倍を β で微分すると

$$\frac{\partial}{\partial\beta}((\beta-1)S_\beta(p||q)) = -\frac{\sum_{i=1}^r p_i^\beta q_i^{1-\beta}\log(p_i/q_i)}{\sum_{i=1}^r p_i^\beta q_i^{1-\beta}}$$

なので, さらに $\beta = 1$ とすると,

$$\left.\frac{\partial}{\partial\beta}\right|_{\beta=1}((\beta-1)S_\beta(p||q)) = -\sum_{i=1}^{r}p_i\log\frac{p_i}{q_i} = S(p||q)$$

と相対エントロピーが出て来る. ゆえに

$$S_1(p||q) := \lim_{\beta\to1} S_\beta(p||q) = S(p||q).$$

相対 Rényi エントロピーは相対エントロピーのワンパラーメーター変形になっていると考えられる. $q_i = 1$ の場合の Rényi エントロピーの定義を知っていれば相対 Rényi エントロピーの定義は誰でも容易に思い付くと思われる.

相対 Rényi エントロピーの定義は分配函数

$$Z(\beta;p,q) = \sum_{i=1}^{r}\left(\frac{p_i}{q_i}\right)^\beta q_i = \sum_{i=1}^{r}e^{-\beta E_i}q_i, \qquad E_i = -\log\frac{p_i}{q_i}$$

に付随する自由エネルギー $F(\beta:p,q)$ と Massieu 函数 $\mathcal{F}(\beta;p,q)$ の定義

$$\begin{aligned} &F(\beta;p,q) = -\beta^{-1}\log Z(\beta;p,q), \\ &\mathcal{F}(\beta;p,q) = \log Z(\beta;p,q) \qquad \text{(Boltzmann 定数倍は略)} \end{aligned}$$

と本質的に同じである:

$$(\beta-1)S_\beta(p||q) = \beta F(\beta;p,q) = -\mathcal{F}(\beta;p,q) = -\log Z(\beta;p,q).$$

Rényi divergence (相対 Rényi エントロピーの -1 倍) の基本性質のまとめが [11] にある.

$(\beta-1)S_\beta(p||q) = -\log Z(\beta;p,q)$ は β の函数として上に凸である:

$$\left(\frac{\partial}{\partial\beta}\right)^2(-\log Z(\beta;p,q)) = -\frac{\sum_{i,j=1}^r (E_i-E_j)^2 e^{-\beta(E_i+E_j)}q_iq_j}{2Z(\beta)^2} \leqq 0$$

(等号成立は $p_i = q_i$ $(i=1,\dots,r)$ と同値).

そして, $(\beta-1)S_\beta(p||q) = -\log Z(\beta;p,q)$ の $\beta=1$ での値が $-\log Z(1;p,q) = -\log 1 = 0$ であることと, $(\beta-1)S_\beta(p||q) = -\log Z(\beta;p,q)$ の $\beta=1$ での微係数が相対エントロピー $S(p||q)$ に等しいという上の計算結果より,

$$(\beta-1)S_\beta(p||q) \leqq (\beta-1)S(p||q).$$

右辺は左辺の接線の式である. □

8.3 相対 Tsallis エントロピー

注意 8.3 (相対 Tsallis エントロピー). 確率分布 $p=(p_1,\dots,p_r), q=(q_1,\dots,q_r)$ に対して, $Z(\beta;p,q)$ を次のように定める:

$$Z(\beta;p,q) = \sum_{i=1}^r e^{-\beta E_i}q_i = \sum_{i=1}^r \left(\frac{p_i}{q_i}\right)^\beta q_i = \sum_{i=1}^r p_i^\beta q_i^{1-\beta}, \qquad E_i = -\log\frac{p_i}{q_i}.$$

各 E_i は 2 つの確率分布 p と q の各 i ごとの違いを表わしている. カノニカル分布 $p(\beta) = (p_1(\beta),\dots,p_r(\beta))$ を

$$p_i(\beta) = \frac{e^{-\beta E_i}q_i}{Z(\beta;p,q)} = \frac{p_i^\beta q_i^{1-\beta}}{Z(\beta;p,q)}$$

と定めると, 逆温度 β は $q_i = p_i(0)$ と $p_i = p_i(1)$ を補間するパラメーターになっている. このとき, 相対 Rényi エントロピー $S_\beta(p||q)$ は

$$S_\beta(p||q) = \frac{\log Z(\beta;p,q)}{1-\beta} = \frac{1}{1-\beta}\log\sum_{i=1}^r p_i^\beta q_i^{1-\beta}$$

と表わされ, 相対エントロピー $S(p||q)$ は

$$S(p||q) = -\left.\frac{\partial}{\partial\beta}\right|_{\beta=1}\log Z(\beta;p,q) = -\left.\frac{\partial}{\partial\beta}\right|_{\beta=1} Z(\beta;p,q) = -\sum_{i=1}^r p_i\log\frac{p_i}{q_i}$$

と表わされる. 2 つ目の等号で $Z(1;p,q)=1$ を使った.

次の演算を x に関する q 差分作用素と呼ぶ:

$$D_{x,q}f(x) = \frac{f(x)-f(qx)}{(1-q)x}.$$

$q\to 1$ で q 差分 $D_{x,q}f(x)$ は微分 $\partial f(x)/\partial x$ に収束する.

上の相対エントロピーの式の $\log Z(\beta;p,q)$ ではなく $Z(\beta;p,q)$ を用いた表示における β に関する微分を q 差分で置き換えることによって[23], **相対 Tsallis エントロピー**が次のように定義される[24] (q 差分の q を次の式では α と書く):

$$T_\alpha(p||q) = -\,D_{\beta,\alpha}|_{\beta=1}\, Z(\beta;p,q) = -\frac{Z(1;p,q)-Z(\alpha;p,q)}{1-\alpha} = -\frac{1-\sum_{i=1}^r p_i^\alpha q_i^{1-\alpha}}{1-\alpha}.$$

$\alpha\to 1$ で α 差分は通常の微分に収束するので, 相対 Tsallis エントロピーは相対エントロピーに収束する. そのことは

$$T_\alpha(p||q) = -\sum_{i=1}^r \frac{(p_i/q_i)-(p_i/q_i)^\alpha}{1-\alpha} q_i, \qquad \lim_{\alpha\to 1}\frac{x-x^\alpha}{1-\alpha} = \frac{\alpha}{\partial\alpha}\Big|_{\alpha=1} x^\alpha = x\log x.$$

より, 直接にも確かめられる. 相対 Tsallis エントロピーは相対エントロピーの定義における $x\log x$ を $(x-x^\alpha)/(1-\alpha)$ で置き換えたものだと言える. 相対 Tsallis エントロピーを相対 Rényi エントロピーで次のように表わすこともできる:

$$T_\beta(p||q) = \frac{Z(\beta;p,q)-1}{1-\beta} = \frac{\exp((1-\beta)S_\beta(p||q))-1}{1-\beta}.$$

逆に相対 Rényi エントロピーを相対 Tsallis エントロピーによって

$$S_\beta(p||q) = \frac{\log Z(\beta;p,q)}{1-\beta} = \frac{\log(1+(1-\beta)T_\beta(p||q))}{1-\beta}$$

と表わすこともできる. 相対 Tsallis エントロピーと相対 Rényi エントロピーの違いは $x-1$ と $\log x = \log(1+(x-1))$ の違いであると考えることもできる.

以上のように, 相対エントロピー, 相対 Rényi エントロピー, 相対 Tsallis エントロピーはどれも分配函数 $Z(\beta;p,q)$ からの派生物である. □

注意 (相対 Tsallis エントロピーと相対 Rényi エントロピーの関係). 相対 Tsallis エントロピー $T_\beta(p||q)$ と相対 Rényi エントロピー $S_\beta(p||q)$ はどちらも

$$Z_\beta(p||q) = \sum_{i=1}^r p_i^\beta q_i^{1-\beta}$$

を用いて

$$T_\beta(p||q) = -\frac{Z_\beta(p||q)-1}{\beta-1}, \qquad S_\beta(p||q) = -\frac{\log Z_\beta(p||q)}{\beta-1}.$$

と表わされる. $\beta>1$ (もしくは $\beta<1$) の場合はどちらも $Z_\beta(p||q)$ の単調減少函数 (もしくは単調増加函数) なので, それらを最大化することは $Z_\beta(p||q)$ を最小化 (もしくは最大化) することと同値になる. さらに

$$e_{\beta-1}(x) = (1+(\beta-1)x)^{1/(\beta-1)}, \qquad \ell_{\beta-1}(x) = \frac{x^{\beta-1}-1}{\beta-1}$$

[23]筆者は 2016 年 6 月 22 日の段階でその必然性をまったく理解できていない.

[24]筆者は (相対)Tsallis エントロピーの定義の必然性をまったく理解していない. (相対)Rényi エントロピーは本質的に分配函数の対数 (自由エネルギー, Massieu 函数) なのでそのようなものを考えることの必然性を納得できるが, (相対)Tsallis エントロピーについてはよくわからない.

が互いに相手の逆函数になることより,

$$Z_\beta(p||q)^{1/(\beta-1)} = e_{\beta-1}(-T_\beta(p||q)) = \exp(-S_\beta(p||q)).$$

この意味で相対 Tsallis エントロピーと相対 Rényi エントロピーの違いはちょうど $e_{\beta-1}(x)$ と $\exp(x)$ の違いになっている. もちろん, この事実は最初から

$$T_\beta(p||q) = -\ell_{\beta-1}\left(Z_\beta(p||q)^{1/(\beta-1)}\right), \qquad S_\beta(p||q) = -\log\left(Z_\beta(p||q)^{1/(\beta-1)}\right)$$

と書いておけば自明なのであるが. □

注意 8.4 (負値性). $p_i, q_i \geqq 0$, $\sum_{i=1}^r p_i = \sum_{i=1}^r = q_i = 1$ であるとし, $\beta > 0$ であると仮定する. 相対エントロピー, 相対 Rényi エントロピー, 相対 Tsallis エントロピーはそれぞれ

$$\begin{aligned} S(p||q) &= -\sum_{i=1}^r p_i \log \frac{p_i}{q_i} \\ S_\beta(p||q) &= -\frac{1}{\beta-1}\log \sum_{i=1}^r p_i^\beta q_i^{1-\beta}, \\ T_\beta(p||q) &= -\frac{\sum_{i=1}^r p_i^\beta q_i^{1-\beta} - 1}{\beta-1} \end{aligned}$$

と定義されたのであった. $\ell_{\beta-1}(x)$ を

$$\ell_{\beta-1}(x) = \frac{x^{\beta-1}-1}{\beta-1}$$

と定めると, 相対 Tsallis エントロピーは

$$T_\beta(p||q) = -\sum_{i=1}^r p_i \ell_{\beta-1}\left(\frac{p_i}{q_i}\right)$$

と表わされる. $S(p||q)$, $S_\beta(p||q)$, $T_\beta(p||q)$ がすべて 0 以下であることを示そう.

相対エントロピー $f(x) = x\log x$ とおくと, $f'(x) = \log x + 1$, $f''(x) = 1/x$ より $f(x)$ は下に凸な函数であり, $f(1) = 0$, $f'(1) = 1$ より, $f(x) \geqq x - 1$ となる. ゆえに

$$S(p||q) = -\sum_{i=1}^r f\left(\frac{p_i}{q_i}\right) q_i \leqq -\sum_{i=1}^r \left(\frac{p_i}{q_i} - 1\right) q_i = 0.$$

相対 Tsallis エントロピー $g(x) = x\ell_{\beta-1}(x)$ とおくと, $g'(x) = \ell_{\beta-1}(x) + x^{\beta-1}$, $g''(x) = \beta x^{\beta-2}$ と $\beta > 1$ より, $g(x)$ は下に凸な函数であり, $g(1) = 0$, $g'(1) = 1$ より, $g(x) \geqq x - 1$ となる. ゆえに

$$T_\beta(p||q) = -\sum_{i=1}^r g\left(\frac{p_i}{q_i}\right) q_i \leqq -\sum_{i=1}^r \left(\frac{p_i}{q_i} - 1\right) q_i = 0. \tag{8.1}$$

相対 Rényi エントロピー $\beta > 1$ という仮定より, $S_\beta(p||q) \leqq 0$ を示すためには

$$\sum_{i=1}^{r} p_i^\beta q_i^{1-\beta} = \sum_{i=1}^{r} \left(\frac{p_i}{q_i}\right)^\beta q_i \geqq 1$$

を示せばよい. $h(x) = x^\beta$ とおくと, $h'(x) = \beta x^{\beta-1}$, $h''(x) = \beta(\beta-1)x^{\beta-2}$ と $\beta > 1$ より, $h(x)$ は下に凸な函数であり, $h(1) = 1$, $h'(1) = \beta$ より $h(x) \geqq 1 + \beta(x-1)$ となる. ゆえに

$$\sum_{i=1}^{r} \left(\frac{p_i}{q_i}\right)^\beta q_i \geqq \sum_{i=1}^{r} \left(1 + \beta\left(\frac{p_i}{q_i} - 1\right)\right) q_i = 1$$

これで示すべきことがすべて示された. 以上の議論において Jensen の不等式を使えばほんの少しだけ近道できる. □

8.4 加法性 (示量性) について

注意 8.5 (加法性について). $\nu = 1, 2$ に対する有限集合 $R_\nu = \{1, 2, \ldots, r_\nu\}$ 上の確率分布 $p_\nu = (p_{\nu,1}, \ldots, p_{\nu,r_\nu})$, $p_\nu = (p_{\nu,1}, \ldots, p_{\nu,r_\nu})$ に対して, 相対エントロピーと相対 Rényi エントロピーは

$$S(p_\nu||q_\nu) = -\sum_{i=1}^{r_\nu} p_{\nu,i} \log \frac{p_{\nu,i}}{q_{\nu,i}},$$
$$S_\beta(p_\nu||q_\nu) = \frac{\log Z_\beta(p_\nu||q_\nu)}{1-\beta}, \qquad Z_\beta(p_\nu||q_\nu) = \sum_{i=1}^{r_\nu} p_{\nu,i}^\beta q_{\nu,i}^{1-\beta}$$

となる. 直積集合 $R_1 \times R_2 = \{\, (i,j) \mid i \in R_1,\ j \in R_2 \,\}$ 上の確率分布が $(i,j) \mapsto p_{1,i}p_{2,j}$, $(i,j) \mapsto q_{1,i}q_{2,j}$ によって定義される. この直積集合上の確率分布の組に対する相対エントロピーと相対 Rényi エントロピーの定義を書き下すと次のようになる:

$$S(p_1, p_2||q_1, q_2) = -\sum_{i,j} p_{1,i}p_{2,j} \log \frac{p_{1,i}p_{2,j}}{q_{1,i}q_{2,j}},$$
$$S_\beta(p_1, p_2||q_1, q_2) = \frac{\log Z_\beta(p_1, p_2||q_1, q_2)}{1-\beta}, \qquad Z_\beta(p_1, p_2||q_1, q_2) = \sum_{i,j} (p_{1,i}p_{2,j})^\beta (q_{1,i}q_{2,j})^{1-\beta}.$$

このとき次の**加法性**が成立している:

$$S(p_1, p_2||q_1, q_2) = S(p_1||q_2) + S(p_2||q_2), \qquad S_\beta(p_1, p_2||q_1, q_2) = S_\beta(p_1||q_2) + S_\beta(p_2||q_2).$$

後者は

$$Z_\beta(p_1, p_2||q_1, q_2) = Z_\beta(p_1||q_1)\, Z_\beta(p_2||q_2)$$

と同値である. 証明は以下の通り:

$$\begin{aligned} S(p_1, p_2||q_1, q_2) &= -\sum_{i,j} p_{1,i}p_{2,j} \log \frac{p_{1,i}}{q_{1,i}} - \sum_{i,j} p_{1,i}p_{2,j} \log \frac{p_{2,j}}{q_{2,j}} \\ &= -\sum_{i} p_{1,i} \log \frac{p_{1,i}}{q_{1,i}} - \sum_{j} p_{2,j} \log \frac{p_{2,j}}{q_{2,j}} = S(p_1||q_1) + S(p_2||q_2), \end{aligned}$$

$$Z_\beta(p_1,p_2||q_1,q_2)=\sum_{i,j}(p_{1,i}p_{2,j})^\beta(q_{1,i}q_{2,j})^{1-\beta}=\sum_{i,j}p_{1,i}^\beta q_{1,i}^{1-\beta}\cdot p_{2,j}^\beta q_{2,j}^{1-\beta}$$
$$=\sum_i p_{1,i}^\beta q_{1,i}^{1-\beta}\cdot\sum_j p_{2,j}^\beta q_{2,j}^{1-\beta}=Z_\beta(p_1||q_1)\,Z_\beta(p_2||q_2).$$

相対 Tsallis エントロピー

$$T_\beta(p_\nu||q_\nu)=\frac{Z_\beta(p_\nu||q_\nu)-1}{1-\beta},\qquad T_\beta(p_1,p_2||q_1,q_2)=\frac{Z_\beta(p_1,p_2||q_1,q_2)-1}{1-\beta}$$

は加法性を満たしていないが,

$$T_\beta(p_1,p_2||q_1,q_2)=T_\beta(p_1||q_1)+T_\beta(p_2||q_2)+(1-\beta)T_\beta(p_1||q_1)\,T_\beta(p_2||q_2)$$

を満たしている. 証明は次の通り:

$$\begin{aligned}
&T_\beta(p_1,p_2||q_1,q_2)-T_\beta(p_1||q_1]-T_\beta(p_2||q_2)\\
&=\frac{Z_\beta(p_1||q_1)\,Z_\beta(p_2||q_2)-1-(Z_\beta(p_1||q_1)-1)-(Z_\beta(p_2||q_2)-1)}{1-\beta}\\
&=\frac{Z_\beta(p_1||q_1)\,Z_\beta(p_2||q_2)-Z_\beta(p_1||q_1)-Z_\beta(p_2||q_2)+1}{1-\beta}\\
&=\frac{(Z_\beta(p_1||q_1)-1)(Z_\beta(p_2||q_2)-1)}{1-\beta}\\
&=(1-\beta)\,T_\beta(p_1||q_1)\,T_\beta(p_2||q_2).
\end{aligned}$$

これは q 数 $(x)_q=(1-q^x)/(1-q)$ に関する公式

$$(x+y)_q=(x)_q+(y)_q+(q-1)\,(x)_q\,(y)_q$$

に似ている. Tsallis エントロピー[25] に関係した数学的構造に関する議論および文献については [8] を参照せよ. そこでは $n\to\infty$ の漸近挙動に Tsallis エントロピーが現われる多項係数の類似物が扱われている. □

8.5 相対 Tsallis エントロピーを漸近挙動に含む多項分布の拡張 (1)

Suyari [8] は多項**係数**のある拡張の $n\to\infty$ の漸近挙動が Tsallis エントロピーで記述されることを示した. 以下ではその結果を多項**分布**と**相対** Tsallis エントロピーの場合に拡張する[26]. 以下では細かいことを気にせずに大雑把な素描を行なう.

以下ではパラメーター h とパラメーター β のあいだには

$$h=\beta-1,\qquad h+1=\beta$$

という関係があると仮定する. パラメーター $h=\beta-1$ による指数函数の拡張 $e_h(x)$ と対数函数の拡張 $\ell_h(x)$ を

$$e_h(x)=(1+hx)^{1/h},\qquad \ell_h(x)=\frac{x^h-1}{h}$$

[25] 相対 Tsallis エントロピーの定義で $q_i=1$ とすると Tsallis エントロピーの定義が得られる.
[26] ほぼ自明な拡張でしかない.

と定める. これらは互いに相手の逆函数になっており, $h \to 0$ ($\beta \to 1$) で通常の指数函数と対数函数に収束する. この記号法を使うと, 第 8.3 節で導入した相対 Tsallis エントロピー $T_\beta(p||q)$ を次のように表わせる[27]:

$$T_\beta(p||q) = -\sum_{i=1}^{r} p_i\, \ell_h\left(\frac{p_i}{q_i}\right).$$

なぜならば

$$\begin{aligned} -\sum_{i=1}^{r} p_i\, \ell_h\left(\frac{p_i}{q_i}\right) &= -\frac{1}{h}\sum_{i=1}^{r}\left(p_i\left(\frac{p_i}{q_i}\right)^h - p_i\right) = -\frac{1}{\beta-1}\sum_{i=1}^{r}(p_i^\beta q_i^{1-\beta} - p_i) \\ &= -\frac{\sum_{i=1}^{r} p_i^\beta q_i^{1-\beta} - 1}{\beta-1} = \frac{1-\sum_{i=1}^{r} p_i^\beta q_i^{1-\beta}}{\beta-1} = T_\beta(p||q). \end{aligned}$$

これが漸近挙動に現われるような多項分布の拡張を構成することが以下の目標である.

通常の多項分布における確率は

$$a(n;k) = \frac{n!}{k_1\cdots k_r} q_1^{k_1}\cdots q_r^{k_r} \qquad (k=(k_1,\ldots,k_r),\ k_i \geqq 0,\ k_1+\cdots+k_r = n)$$

であり, これの対数は次のように表わされる:

$$\log a(n;k) = \sum_{\nu=1}^{n} \log\nu - \sum_{i=1}^{r}\sum_{\nu_i=1}^{k_i} \log\left(\frac{\nu_i}{q_i}\right).$$

これを拡張して次の条件によって $\ell_h(a_h(n;k))$ を定義する:

$$\ell_h(a_h(n;k)) = \sum_{\nu=1}^{n} \ell_h(\nu) - \sum_{i=1}^{r}\sum_{\nu_i=1}^{k_i} \ell_h\left(\frac{\nu_i}{q_i}\right).$$

このとき, 右辺の $\ell_h(x) = x^h/h - 1/h$ の中の $-1/h$ の部分はキャンセルして消えるので,

$$\ell_h(a_h(n;k)) = \frac{1}{h}\left(\sum_{\nu=1}^{n}\nu^h - \sum_{i=1}^{r}\sum_{\nu_i=1}^{k_i}\left(\frac{\nu_i}{q_i}\right)^h\right) = \frac{1}{h}\left(\sum_{\nu=1}^{n}\nu^h - \sum_{i=1}^{r} q_i^{-h}\sum_{\nu_i=1}^{k_i}\nu_i^h\right).$$

$h>0$ ($\beta>1$) と仮定し, $n\to\infty$ で k_i/n はほぼ一定 $k_i = np_i + O(1) = n(p_i + O(1/n))$ という条件を仮定する. $n\to\infty$ のとき $\ell_h(a_h(n;k))$ がどのように振る舞うかを知りたい. $h>0$ と仮定したので,

$$\sum_{\nu=1}^{r}\nu^h = \frac{n^{h+1}}{h+1} + O(n^h) = \frac{n^\beta}{\beta} + O(n^{\beta-1}) \qquad (n\to\infty).$$

さらに $k_i = np_i + O(1)$ と仮定したので,

$$\sum_{\nu_i=1}^{k_i}\nu_i^h = \frac{(np_i)^{h+1}}{h+1} + O(n^h) = \frac{n^\beta}{\beta}p_i^\beta + O(n^{\beta-1}).$$

[27] これを相対 Tsallis エントロピーの定義だと思ってもよい. この形式の定義では, 相対 Tsallis エントロピー $T_\beta(p||q)$ が相対エントロピー $S(p||q) = -\sum_{i=1}^{r} p_i\log(p_i/q_i)$ のパラメーター $h=\beta-1$ による拡張になっていることが分かり易い.

ゆえに

$$\ell_h(a_h(n;k)) = \frac{n^\beta}{\beta}\frac{1-\sum_{i=1}^r p_i^\beta q_i^{1-\beta}}{\beta-1} + O(n^{\beta-1}) = \frac{n^\beta}{\beta}T_\beta(p||q) + O(n^{\beta-1}) \qquad (n\to\infty).$$

これが**目標としていた結果**である[28]. この結果は多項分布の漸近挙動

$$\log a(n;k) = \log\left(\frac{n!}{k_1!\cdots k_r!}q_1^{k_1}\cdots q_r^{k_r}\right) = nS(p||q) + O(\log n)$$

の拡張になっている.

8.6 相対 Tsallis エントロピーを漸近挙動に含む多項分布の拡張 (2)

Suyari-Scarfone [9] に第 8.6 節 とは異なる多項分布の拡張の仕方が書いてあったので以下で紹介することにする. 第 8.6 節の $a_h(n;k)$ 以外の記号を引き継ぐ. 文献 [9] の q と第 8.6 節の $h=\beta-1$ は $h=1-q$, $h+1=\beta=2-q$ によって対応している.

函数 $e_h(x)$ とその逆函数を $\ell_h(x)$ を

$$e_h(x) = (1+hx)^{1/h} > 0 \quad \left(x > -\frac{1}{h}\right), \qquad \ell_h(x) = \frac{x^h-1}{h} > -\frac{1}{h} \quad (x>0)$$

と定めると相対 Tsallis エントロピーは

$$T_\beta(p||q) = \frac{1-\sum_{i=1}^r p_i^\beta q_i^{1-\beta}}{\beta-1} = \frac{1}{h}\left(1-\sum_{i=1}^r p_i^{h+1}q_i^{-h}\right) = -\sum_{i=1}^r p_i\ell_h\left(\frac{p_i}{q_i}\right)$$

と表わせるのであった.

多項分布における確率は

$$a(n;k) = \frac{n!}{k_1\cdots k_r}q_1^{k_1}\cdots q_r^{k_r} \qquad (k=(k_1,\ldots,k_r),\ k_i\geqq 0,\ k_1+\cdots+k_r=n)$$

であり, これの対数は次のように表わされる:

$$\log a(n;k) = \sum_{\nu=1}^n \log\nu - \sum_{i=1}^r\left(\sum_{\nu_i=1}^{k_i}\log\nu_i + k_i\log q_i^{-1}\right).$$

この節ではこれを拡張して $\ell_h(a_h(n;k))$ を次のように定義する[29]:

$$\ell_h(a_h(n;k)) = \sum_{\nu=1}^n \ell_h(\nu) - \sum_{i=1}^r\left(\sum_{\nu_i=1}^{k_i}\ell_h(\nu_i) + \frac{k_i^{h+1}}{h+1}\ell_h(q_i^{-1})\right).$$

これは次のように書き直される:

$$\ell_h(a_h(n;k)) = \frac{1}{h}\left(\sum_{\nu=1}^n \nu^h - \sum_{i=1}^r\left(\sum_{\nu_i=1}^{k_i}\nu_i^h + \frac{k_i^{h+1}}{h+1}(q_i^{-h}-1)\right)\right).$$

[28] この結果 (もしくはその拡張) を使えば相対 Tsallis エントロピーの場合に関する Sanov の定理の拡張を証明できるだろう. しかし, 2016 年 6 月 23 日の時点で筆者はパラメーター $h=\beta-1$ による多項分布の拡張 $a_h(n;k)$ が出て来る必然性を理解できていない.

[29] 2016 年 7 月 6 日現在, この定義の必然性を筆者は理解できていない.

$h > 0$ ($\beta > 1$) と仮定し, $n \to \infty$ で k_i/n はほぼ一定 $k_i = np_i + O(1) = n(p_i + O(1/n))$ という条件を仮定する. $n \to \infty$ のとき $\ell_h(a_h(n;k))$ がどのように振る舞うかを知りたい. $h > 0$ と仮定したので,

$$\sum_{\nu=1}^{r} \nu^h = \frac{n^{h+1}}{h+1} + O(n^h) = \frac{n^\beta}{\beta} + O(n^{\beta-1}) \qquad (n \to \infty).$$

さらに $k_i = np_i + O(1)$ と仮定したので,

$$\sum_{\nu_i=1}^{k_i} \nu_i^h = \frac{(np_i)^{h+1}}{h+1} + O(n^h) = \frac{n^\beta}{\beta} p_i^\beta + O(n^{\beta-1}),$$
$$\frac{k_i^{h+1}}{h+1} = \frac{(np_i)^{h+1}}{h+1} + O(n^h) = \frac{n^\beta}{\beta} p_i^\beta + O(n^{\beta-1}).$$

そして $q_i^{-h} = q_i^{1-\beta}$ である. ゆえに, $n \to \infty$ において,

$$\ell_h(a(n;k)) = \frac{1}{h}\left(\frac{n^\beta}{\beta} - \sum_{i=1}^{r} \frac{n^\beta}{\beta} p_i^\beta q_i^{1-\beta}\right) + O(n^{\beta-1}) = \frac{n^\beta}{\beta} T_\beta(p||q) + O(n^{\beta-1}).$$

これが目標としていた結果である. この結果は多項分布の漸近挙動

$$\log a(n;k) = \log\left(\frac{n!}{k_1! \cdots k_r!} q_1^{k_1} \cdots q_r^{k_r}\right) = nS(p||q) + O(\log n)$$

の拡張になっている.

8.7 Csiszár の f-divergence

他にもたくさん文献があるのだが, Csiszár [2] に詳しい参考文献欄がある.

$f(x)$ は $0 < x < \infty$ で下に凸な非負値函数であり, $f(1) = 0$ (最小値) であると仮定する. 有限集合 $\{1, 2, \ldots, r\}$ 上の確率分布 $p = (p_1, \ldots, p_r)$, $q = (q_1, \ldots, q_r)$ に対して, q から p への f-divergence $D_f(p||q)$ が

$$D_f(p||q) = \sum_{i=1}^{r} f\left(\frac{p_i}{q_i}\right) q_i$$

と定義される.

たとえば $f(x) = x \log x$ のとき, f-divergence は Kullback-Leibler divergence

$$D(p||q) = \sum_{i=1}^{r} p_i \log\left(\frac{p_i}{q_i}\right)$$

に一致する. たとえば

$$f(x) = x\ell_h(x) = x\frac{x^h - 1}{h} = \frac{x^\beta - x}{\beta - 1}, \qquad h = \beta - 1$$

のとき, f-divergence は

$$D_f(p||q) = \sum_{i=1}^{r} \frac{(p_i/q_i)^\beta - (p_i/q_i)}{\beta - 1} q_i = \sum_{i=1}^{r} \frac{p_i^\beta q_i^{1-\beta} - p_i}{\beta - 1} = \frac{\sum_{i=1}^{r} p_i^\beta q_i^{1-\beta} - 1}{\beta - 1} = -T_\beta(p||q)$$

と Tsallis divergence (相対 Tsallis エントロピーの -1 倍) に一致する. 他の様々な相対情報量が f-divergence の特別な場合になっている.

対数和不等式の一般化　$\{1,2,\dots,r\}$ の部分集合 A に対して, 確率分布 p,q における A の確率をそれぞれ

$$p(A)=\sum_{i\in A}p_i,\qquad q(A)=\sum_{i\in A}q_i$$

と定義する. $A_1,\dots,A_s$ は集合 $\{1,2,\dots,r\}$ の分割であるとし, 集合 $\{A_1,A_2,\dots,A_s\}$ 上の確率分布 $P=(P_1,\dots,P_s)$, $Q=(Q_1,\dots,Q_r)$ を $P_j=p(A_j)$, $Q_j=q(A_j)$ と定める. 第 5.2 節では対数和不等式から Kullback-Leibler 情報量について

$$D(p||q)\geqq D(P||Q)$$

という不等式が成立していることを示した. この不等式は**細部の情報を忘れると情報量は小さくなること**を意味している. f-divergence についても同様の不等式

$$D_f(p||q)\geqq D_f(p||q)$$

が成立していることを下に凸な函数 $f(x)$ に関する Jensen の不等式を使って示せる:

$$\begin{aligned}D_f(p||q)&=\sum_{j=1}^{s}\sum_{i\in A_j}f\left(\frac{p_i}{q_i}\right)q_i=\sum_{j=1}^{s}Q_j\sum_{i\in A_j}f\left(\frac{p_i}{q_i}\right)\frac{q_i}{Q_j}\\&\geqq\sum_{j=1}^{s}Q_jf\left(\sum_{i\in A_j}\frac{p_i}{q_i}\frac{q_i}{Q_j}\right)=\sum_{j=1}^{s}f\left(\frac{P_j}{Q_j}\right)Q_j=D_f(P||Q).\end{aligned}$$

特に $s=1$, $A_1=\{1,2,\dots,r\}$ の場合を考えると $P_1=Q_1=1$, $f(1)=0$ より $D_f(P||Q)=0$ となるので

$$D_f(p||q)\geqq 0.$$

他にも Kullback-Leibler 情報量と同様の多くの性質を f-divergence が満たしていることを示せる.

9　付録: 上極限と下極限に関する簡単な解説

上極限 $\limsup$ と下極限 $\liminf$ は収束先として $\pm\infty$ を許せば常に収束するので, 収束するかどうかわからない実数列の漸近挙動を調べるときにとても便利である.

数学科の学生であれば上極限と下極限についても講義で習っていてよく知っているだろうが, 他学科の出身者は詳しく習ったことがないかもしれない. だから, この付録で上極限と下極限について簡単に解説しておくことにした.

9.1　上極限と下極限の定義

$a_1,a_2,\dots$ は実数列であるとする.

$a_n,a_{n+1},a_{n+2},\dots$ の上限 $\sup_{k\geqq n}a_k$ を

$$\sup_{k\geqq n}a_k=(\text{すべての }a_n,a_{n+1},a_{n+2},\dots\text{ 以上の }\alpha\text{ の中で最小のもの}).$$

ただし α は実数または ∞ であるとする[30]:

$$\sup_{k\geqq n} a_k = \min\{\, \alpha \in \mathbb{R}\cup\{\infty\} \mid a_k \leqq \alpha \ (k \geqq n)\,\}.$$

一般により小さな実数の集合の上限は小さくなるので n に関する数列 $\sup_{k\geqq n} a_k$ は単調減少数列になる. したがって, 数列 $\sup_{k\geqq n} a_k$ は $n\to\infty$ で実数または $\pm\infty$ に収束する[31]. その収束先を実数列 a_n の上極限 (limit superior) と呼び, 次のように表わす:

$$\limsup_{n\to\infty} a_n = \lim_{n\to\infty} \sup_{k\geqq n} a_k.$$

同様に下極限 (limit inferior) を次のように定義する:

$$\inf_{k\geqq n} a_k = \max\{\, \alpha \in \mathbb{R}\cup\{-\infty\} \mid a_k \geqq \alpha \ (k \geqq n)\,\},$$

$$\liminf_{n\to\infty} a_n = \lim_{n\to\infty} \inf_{k\geqq n} a_k.$$

上限 sup は下限 inf 以上なので上極限と下極限は次の不等式を満たしている:

$$\liminf_{n\to\infty} a_n \leqq \limsup_{n\to\infty} a_n.$$

実数列 a_n が収束するならば, $\sup_{k\geqq n} a_k$ と $\inf_{k\geqq n} a_k$ の差は 0 に収束するので,

$$\liminf_{n\to\infty} a_n = \limsup_{n\to\infty} a_n = \lim_{n\to\infty} a_n$$

が成立する. 逆に $\limsup_{n\to\infty} a_n$ と $\liminf_{n\to\infty} a_n$ が一致するならば実数列 a_n はそれらと同じ値に収束することもわかる.

例 9.1 (上極限と下極限の例). 以下が成立していることを定義に基づいて確認してみよ:

$$\limsup_{n\to\infty}(-1)^n = 1, \qquad \liminf_{n\to\infty}(-1)^n = -1,$$
$$\limsup_{n\to\infty}\left((-1)^n n\right) = \infty, \qquad \liminf_{n\to\infty}\left((-1)^n n\right) = -\infty,$$
$$\limsup_{n\to\infty}\left((-1)^n\left(1+2^{-n}\right)\right) = 1, \qquad \liminf_{n\to\infty}\left((-1)^n\left(1+2^{-n}\right)\right) = -1.$$

これらの上極限と下極限を図を描いて確認すれば上極限と下極限の概念を直観的に理解できると思う. □

9.2 上極限と下極限の使い方

上極限と下極限の典型的な使い方について説明しよう. 数列 a_n の $n\to\infty$ での様子を知りたいとしよう. 数列 a_n が $n\to\infty$ で収束する量 B_n, C_n によって

$$B_n \leqq a_n \leqq C_n$$

[30] 上限の存在は実数の連続性によって保証される. 上限が常に存在することを実数の連続性そのものだと思ってもよい.

[31] 収束することも実数の連続性によって保証される.

と評価されたとする (はさみうち!). このとき $n \to \infty$ で次が成立する:

$$\lim_{n\to\infty} B_n \leqq \liminf_{n\to\infty} a_n \leqq \limsup_{n\to\infty} a_n \leqq \lim_{n\to\infty} C_n.$$

この評価は実数列 a_n が収束していなくても成立する. もしも B_n と C_n が同じ値に収束するならば, この不等式より a_n もそれらと同じ値に収束することがわかる.

以上の議論のパターンを知っていれば, a_n が収束するとは限らない弱い条件のもとで上極限と下極限に関する不等式を示しておいて, 追加の強い条件のもとで a_n が収束することを示せるようになる. Sanov の定理の定式化と証明に関する第 3 節の議論はまさにそのようなタイプの議論の典型例になっている.

10 Mcmillan の不等式と平均符号長

以下の解説は連続ツイート

https://twitter.com/genkuroki/status/781366321665630209
https://twitter.com/genkuroki/status/781403859168002048

の内容をまとめたものである.

10.1 Mcmillan の不等式

集合 A に対して A の元を並べてできる語全体の集合を A^* と書くことにする:

$$A^* = \{\, a_1a_2\dots a_l \mid a_1, a_2, \dots, a_l \in A,\ l = 0, 1, 2, \dots \}.$$

$a = a_1a_2\cdots a_l$ の長さ l を $\ell(a) = l$ と書くことにする. A^* には語の連結で別の語を作る操作によって自然に積が定まる.

r は正の整数であるとし, b は 2 以上の整数であるとする. 集合 S, T はそれぞれ r 個, b 個の元からなる集合であるとする: $S = \{s_1, \dots, s_r\}$, $T = \{t_1, \dots, t_b\}$ と定める. 任意に与えられた長さが 1 以上の語 $w_1, \dots, w_r \in T^*$ に対して, 写像 $C : S^* \to T^*$ を

$$C(s_{i_1} \cdots s_{i_m}) = w_{i_1} \cdots w_{i_m}$$

によって定めることができる. この C を S の b 文字による符号化と呼ぶ.

符号化 $C : S^* \to T^*$ が単射のとき, C は**一意復号可能**であるという.

定理 10.1 (Mcmillan の不等式). C は一意復号可能な S の b 文字による符号化であるとする. このとき S の元 s_i の符号化 $C(s_i) = w_i$ の長さを l_i と書くと,

$$\sum_{i=1}^{r} b^{-l_i} \leqq 1$$

が成立している.

証明. $a = \sum_{i=1}^{r} b^{-l_i}$ とおく. $a \leqq 1$ を示せばよい. l は $l_1, \ldots, l_r$ の最大値であるとし, n は正の整数であるとする. このとき

$$a^n = \sum_{i_1,\ldots,i_n=1}^{r} b^{-(l_{i_1}+\cdots+l_{i_n})} = \sum_{k=1}^{nl} N_k b^{-k}.$$

ここで N_k は $l_{i_1} + \cdots + l_{i_n} = k$ を満たす $s = s_{i_1} \cdots s_{i_n}$ 全体の個数である. そのとき $w = C(s) = w_{i_1} \cdots w_{i_n}$ は b 種類の文字を k 個並べてできる語になるので, そのような w の個数は b^k 個以下である. 仮定より $C : S^* \to T^*$ は単射なので, 上のような s の個数も b^k 個以下になる. ゆえに

$$a^n = \sum_{k=1}^{nl} N_k b^{-k} \leqq \sum_{k=1}^{nl} b^k b^{-k} \leqq \sum_{k=1}^{nl} 1 = nl.$$

したがって $n \to \infty$ で

$$a = (nl)^{1/n} = e^{(1/n)\log(nl)} \to e^0 = 1$$

となる. これで $a \leqq 1$ が示された. □

10.2 平均符号長への応用

$p_i, q_i \geqq 0$, $\sum_{i=1}^{r} p_i = \sum_{i=1}^{r} q_i = 1$ であるとする. Kullback-Leibler 情報量は 0 以上になるのであった (Gibbs の情報不等式):

$$D(p||q) = \sum_{i=1}^{r} p_i \log \frac{p_i}{q_i} \geqq 0.$$

この不等式は下に凸な函数 $f(x) = x \log x$ に Jensen の不等式を適用すればただちに証明される:

$$D(p||q) = \sum_{i=1}^{r} q_i f\left(\frac{p_i}{q_i}\right) \geqq f\left(\sum_{i=1}^{r} q_i \frac{p_i}{q_i}\right) = f(1) = 0.$$

$a_i \geqq 0$, $\sum_{i=1}^{t} a_i = a \leqq 1$ のとき, $q_i = a_i/a$ すなわち $a_i = aq_i$ とおいてこの不等式を使うと

$$\sum_{i=1}^{r} p_i \log\left(\frac{p_i}{a_i}\right) = \sum_{i=1}^{r} p_i \log\left(\frac{p_i}{aq_i}\right) = D(p||q) - \log a \geqq 0.$$

この不等式は次のように書き直される:

$$a_i \geqq 0,\ \sum_{i=1}^{r} a_i \leqq 1 \implies -\sum_{i=1}^{r} p_i \log a_i \geqq -\sum_{i=1}^{r} p_i \log p_i. \qquad (*)$$

両辺を $\log b$ で割れば対数の底を任意の $b > 1$ にしてもこの不等式が成立していることがわかる.

C は $S = \{s_1, \ldots, s_r\}$ の b 文字による符号化であり, 一意復号可能であると仮定し, $w_i = C(s_i)$ の長さを l_i と書く. Mcmillan の不等式より

$$\sum_{i=1}^{r} b^{-l_i} \leqq 1$$

なので, $a_i = b^{-l_i}$ に対数の底が b の場合の不等式 $(*)$ を適用することによって, 不等式

$$\sum_{i=1}^{r} p_i l_i \geqq -\sum_{i=1}^{r} p_i \log_b p_i$$

が得られる. この不等式の左辺を S 上の確率分布 p_i に関する**平均符号長**と呼ぶ. 右辺は S 上の確率分布 p_i の (対数の底が b の場合の) 情報量 (Shannon 情報量[32]) である. すなわち, “ソース” S 上の確率分布が p_i のとき, 一意復号可能な符号化の平均符号長を “ソース” S 上の確率分布 p_i の情報量未満にすることは不可能である.

一般に S 上の分布 $a_i \geqq 0$, $\sum_{i=1}^{r} a_i \leqq 1$ に対して, $l_i = -\log_b a_i$ を事象 s_i の**情報量**と呼ぶ. $b = 2$ のとき事象 s_i は l_i bit の情報量を持つと言う. 情報量 $l_i = -\log_b a_i$ は s_i を b 種類の文字で表わしたときに何文字必要であるかを意味していると解釈することができる. このとき, 対数の底が b の Kullback-Leibler 情報量における確率分布 q を確率分布とは限らない分布 a に拡張したもの

$$D_b(p||a) = \sum_{i=1}^{r} p_i \log_b \left(\frac{p_i}{a_i}\right) = \sum_{i=1}^{r} p_i(-\log_b a_i) - \sum_{i=1}^{r} p_i(-\log_b p_i) \geqq 0$$

は分布 a_i に対応する方法で (文字 s_i を長さ $l_i = -\log a_i$ で) 一意複合化可能に符号化するときに[33], 「情報量よりも無駄に増えた平均符号長の大きさ」を意味していると解釈される.

このように Kullback-Leibler 情報量には「情報量からの平均符号長の増分」という解釈が可能である.

Kullback-Leibler 情報量 $D(p||a)$ は「平均符号長の長さで測ったソース p そのものの情報量が符号化 a によってどれだけ増えるか」を意味しているとも考えられる.

[32] Shannon エントロピーと呼ぶことも多い.
[33] そのとき Mcmillan の定理より, $\sum_{i=1}^{r} a_i \leqq 1$ となる.