Kullback-Leibler 情報量に関する解説

黒木玄

2016 年 6 月 16 日作成*

http://www.math.tohoku.ac.jp/~kuroki/LaTeX/20160616KullbackLeibler.pdf

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．6 月 17 日 Ver.0.2(16 頁)．区分求積法による高校レベルの方法に関する付録 4.1 と多項分布の場合の Sanov の定理の厳密に証明するための第 3 節を追加した．そこで紹介した証明は階乗に関する Stirling の公式さえ使わない極めて初等的な証明である．6 月 18 日 Ver.0.2.1. 小さな追加と訂正．

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
http://www.math.tohoku.ac.jp/~kuroki/LaTeX/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 $\exp(-\sum_\nu \beta_\nu f_\nu(k))$ で記述される確率分布が必然的に出て来る理由を説明することである. 数学的に厳密な議論は基本的にしない[1].

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] 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)

[3] 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

[4] 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

[5] 田崎晴明. 統計力学 I. 新物理学シリーズ, 培風館 (2008/12), 284 ページ.
https://www.amazon.co.jp/dp/4563024376

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

[7] 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] 多項分布版の Sanov の定理を厳密に証明している第 3 節は例外である.

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.

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. $(q_1, \ldots, q_r)$ 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 回ふって目 i の出た割合の分布 (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 $k_i/n \to q_i$ となるのだが, 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 $(q_1, \ldots, q_n)$ を任意に固定した状況で, 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.

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[2].

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

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

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

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 4.1 節).

1.3 Kullback-Leibler 情報量と相対エントロピーの定義

上の結果は

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

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

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

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:

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

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 $D[p|q]$ の $p=(p_1,\ldots,p_r)$ の函数としての性質は函数 $f(x)=x\log(x/q)=x(\log x-\log q)$ $(x>0)$ の性質を調べればわかる．$f'(x)=\log x-\log q+1$, $f''(x)=1/x>0$ 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 $x=q$ で

の接線の函数 x で下から押さえられる. すなわち $f(x) \geqq f(q) + f'(q)x = x - q$ (等号の成立と $x = q$ は同値). ゆえに

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

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

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, 課した条件もとで Kullback-Leibler 情報量が最小になる分布に条件が課された経験分布は近付くことになる (**条件付き大数の法則**). 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:

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

対数を使わない方の公式を書き下すと,

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

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 (semifiled) 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, $-\infty$ は掛算の単位元 1 に対応している. その対応は log を取って極限を考え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:

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

後者は明らかな公式である. 前者の公式は次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 $1/n$ 倍には $n\to\infty$ 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,

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

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.

適切な設定のもとで次が成立している:

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

$f(x)$ が $x=x_0$ で一意的な最大値を持ち, $f''(x_0)>0$ 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.

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

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

という条件を課す. ただし, $\mathbb{R}^r$ のベクトル $(1,1,\dots,1), (f_{\nu,1},\dots,f_{\nu,r})$ $(\nu=1,\dots,s)$ 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

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

という条件を課す場合には, エネルギーの経験期待値がほぼ U に等しくなっている場合に制限したときに, 経験分布が $n\to\infty$ でどのように振る舞うかを調べることになる.

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,

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

という条件を課す場合には, 1 回あたりの賞金の経験期待値がほぼ U ペリカに等しくなっている場合に制限したときに, 経験分布が $n\to\infty$ でどのように振る舞うかを調べることになる.

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 $K[p|q] = \sum_{i=1}^r p_i \log(p_i/q_i)$ が最低値になる分布 $p=(p_1,\dots,p_r)$ に集中することになる.

その条件付き最低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 の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

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

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

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

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 **Boltsmann 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の 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 **Gibbs 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 p_i は Gibbs 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,

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

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.

Gibbs 分布に対する相対エントロピー $S[p|q] = -K[p|q] = -\sum_{i=1}^{r} p_i \log(p_i/q_i)$ 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

$$S[p|q] = \beta U + \log Z.$$

これらの公式は, 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.

2.3 母分布が連続型の場合から連続型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

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

に近付くと考えられる. 分布 $p(x)$ に以下の条件を課す:

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

前節と同様にして, この条件のもとで $K[p|q]$ 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.

多項分布 $k_1 + \cdots + k_r = n$ のとき, $\beta_i = -\log q_i$ とおくと

$$
\begin{aligned}
&p_{k_1,\ldots,k_r} = \frac{n!}{k_1! \cdots k_r!} q_1^{k_1} \cdots q_r^{k_r} = \frac{e^{-\sum_{i=1}^{r} \beta_i k_i} q_{k_1,\ldots,k_r}}{Z}, \\
&q_{k_1,\ldots,k_r} = \frac{n!}{k_1! \cdots k_r!} \frac{1}{r^n}, \quad Z = \frac{1}{r^n}
\end{aligned}
$$

正規分布

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

Gamma 分布 $x > 0$ において

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

第二種 Beta 分布 $x > 0$ において

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

自由度 1 の t 分布 (Cauchy 分布)

$$
p(x) = \frac{1}{\pi} \frac{1}{1+x^2} = \frac{e^{-\log(1+x^2)}}{Z}, \quad Z = \pi.
$$

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[3]. 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 $\sqrt{n}\,S^{n-1}$ は半径 $\sqrt{n}$ の $n-1$ 次元球面であり, μ_n はその上の一様確率分布であり, $f(x_1)$ の x_1 は x_1 は球面上の点 $(x_1,\ldots,x_n)$ の射影である. この最後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[4].

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.

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 [6] 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.

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.

[3] $q(x)=1$ 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.

[4] 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

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.

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

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

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

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

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.) $X_1,\dots,X_n$ に対応する $\mathcal{P}_n$ の元を P_n と書き, P_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:

(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[5]

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

(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,

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

[5]確率分布全体の空間 $\mathcal{P}$ が有限次元ならば A は任意の部分集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.

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, $A = (a, b)$ 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

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

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

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

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,

$$\begin{aligned} -nD[p|q] &= -\sum_{i=1}^{r} k_i \log p_i + \sum_{i=1}^{r} k_i \log q_i, \\ e^{-nD[p|q]} &= \frac{1}{p_1^{k_1} \cdots p_r^{k_r}} q_1^{k_1} \cdots q_r^{k_r}, \\ P(P_n = p) &= \frac{n!}{k_1! \cdots k_r!} q_1^{k_1} \cdots q_r^{k_r}. \end{aligned}$$

ゆえに, この補題の結果は次と同値である:

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

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. 以下では下からの評価 (左側の不等式) を証明しよう.

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

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

が成立しているはずである. 実際, 補題 3.3 より,

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

ゆえに, 多項定理より

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

両辺を $(n+1)^r$ で割れば下からの評価が得られる. □

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 $\{1, 2, \dots, r\}$ 上の確率分布全体の空間 $\mathcal{P}$ 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 $\{1, 2, \dots, r\}$ 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

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

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 $\{1, 2, \dots, r\}$ 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

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

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 3.5. 以上の証明では階乗に関する Sirling の近似公式を使っていない. 以上の証明で本質的に使った事実は次の二つである.

(1) 上からの評価のために次の事実を使った:
$q_i \geqq 0$, $q_1 + \cdots + q_r = 1$ のとき

$$\frac{n!}{k_1! \cdots k_r!} q_1^{k_1} \cdots q_r^{k_r} \leqq 1 \quad (k_i \in \mathbb{Z}_{\geqq 0},\ k_1 + \cdots + k_r = n).$$

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. その不等式は, 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:
$k_i \in \mathbb{Z}_{\geqq 0}$, $k_1 + \cdots + k_r = n$, $k_i/n = q_i$ のとき,

$$\frac{n!}{l_1! \cdots l_r!} q_1^{l_1} \cdots q_r^{l_r} \leqq \frac{n!}{k_1! \cdots q_r!} q_1^{k_1} \cdots q_r^{k_r} \quad (l_i \in \mathbb{Z}_{\geqq 0},\ l_1 + \cdots + l_r = n)$$

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:

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

この不等式は k, l の大小関係とは無関係に常に成立している.

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実である. 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. □

4 付録

4.1 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

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

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

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