脳を学ぶ上で重要な数学シリーズ 計算論編

後藤 優仁

2021 年 6 月 7 日

目次

1 はじめに 1

2 情報理論 2
2.1 エントロピー 3
2.2 KL 距離 8
2.3 条件付きエントロピー 10
2.4 相互情報量 11

1 はじめに

さて，この数学シリーズの中でもこの advanced はかなり異質で，筆者が脳神経科学を研究する上で関わっていった様々な (あまり一般的ではない) 議論を展開するために必要になった数学的知識をまとめる場です．筆者は脳活動の非線形ダイナミクスの機能的な役割を研究しています．脳は多数の非線形素子が結合した大自由度の力学系とみなすことができ，多様なダイナミクスを示します．計算論的神経科学の観点で，神経系が示す同期，自発活動，誘発活動ダイナミクス，ノイズ誘発同期，... といった様々な現象を捉え，その情報処理メカニズムの解明を試みています．

こうした作業には，信号処理は勿論，非線形ダイナミクス（つまり力学系），情報理論，複雑系，機械学習... 多種多様な数学的，理論的知識が求められます．この勉強については終わりがなく，本当に役に立つのかも分かりません．ただ確実に言えるのは，

「○○をやっている時の脳は△△領域で〜〜 ms 後に × Ｈｚの波が同期している！！」

なんてことだけ見ていても，脳を理解する事は出来ないということです．その活動が何故大事なのか，その活動を通してどんな処理をしているのか... こうしたところまで理解してこその科学だと，筆者含め計算論的神経科学者たちは考えています．

実験的にデータを集めるだけでは理解に至らず，その背景で何が行われているのか，どんな表現がなされているのか．そんなところまで考えないとだよね，というスタンスです．その必要性が分からん，そんなに大

事だろうか．そう思う人はユニークな思考実験的論文があるので是非読んでみてください [1][2].
何をもって脳の理解とするか？
この問に答えるのが，有名なマーの 3 レベルという概念？お話？ [3] です．

- 計算理論
- 表現とアルゴリズム
- ハードウェアによる実装

この 3 段階を踏み，相互に対応付ける事が脳の理解に大切だ，とする話です．計算理論は，我々が脳を使ってどんな「計算」を行っているのか，行っているべきなのか，といった議論．自由エネルギー原理だとかの話，すなわち脳が採用している戦略を考える所です．ハードウェアによる実装は，多くの神経科学者がやっているように脳のどの部分でどんな活動が起きていてといった解剖学・生理学的知見．最後に表現とアルゴリズムは，ハードウェアの実装を使っていかに計算理論で提案された処理を実行するのか，になります．

彼に言わせれば，これまで主流の神経科学はハードウェアの実装ばかりだったわけですね．計算理論と表現に関する議論は，無論ありましたがあまり活発ではなかった．

ここから先，どう考えるかは個人の自由だと思います．計算理論や表現についても考えていこうとするか，そんなに色々手を出しても回収しきれないと見切りをつけるか，あるいは他の人が結び付けてくれる事を期待して実験データを提供するに集中するのか...

筆者は，計算理論と実装を結ぶ，表現の研究者になりたいと考えた次第です．一番勉強する事が多いような気もしますが，楽しんでやっていきます．

長くなりましたが，本書はそんなモチベーションのもと，計算理論やアルゴリズムについて学習したことをまとめていくものにします．なので神経に本当に役立つのか，理解が正しいのか，様々な問題があると思いますが，まあ教科書ではなく筆者のノートだと思って見てください．結構やってみると楽しいです．また本稿はその性質上，随所で本や論文を引用しながら議論を展開していきます．筆者の拙い理解での説明では不十分だったり不適切だったりすることも少なくないはずなので，気になるところは適宜参照してください．

2　情報理論

近年は神経科学に情報理論の議論を輸入するのが流行りになっている気がします．脳波の解析もだし，情報処理の理論もそうだし，いろんなとこで見るのでとりあえず勉強．関係する研究は以下とか

- 自由エネルギー原理 [4]
- Phase Amplitude Coupling の評価．[5]
- 相互情報量

- トランスファーエントロピー

*まだちゃんと引っ張ってきてない

2.1 エントロピー

はじめにエントロピーの考え方を導入しましょう．まずは離散確率変数 x を考えます．観測者がこの変数に対するある値を観測したとき，どれだけの情報量，surprise を得られるのか．これを考える概念がエントロピーです.
直観的に，起きそうもない事象が得られたら情報量は大きいし，その逆も然りですよね．宝くじで 1 等があたるのはめちゃくちゃびっくりする，つまり情報量大きいけど，参加賞的なのもらっても何も思いません．つまり情報量は確率分布 $p(x)$ に依存していて，その値によって定まる単調な関数 $h(x)$ といえます.

また，2 つの事象 x, y を考えたとき，これらが独立なら両方を観測したときの情報量は別個に観測したときの情報量の和と等しい (式 1) はずです．宝くじが当たる事による surprise によって，帰りに頭に雷が落ちてくる事によって生じる surprise が小さくみたいになることはないですよね？あるかもな．ないって事にしてください.
よって以下の式 (1) が成り立ちます.

$$h(x, y) = h(x) + h(y) \tag{1}$$

次に，これらの事象の同時確率についても単純に積で求められます (式 2)．宝くじが当たり，かつ頭に雷が降ってくる確率です.

$$p(x, y) = p(x)p(y) \tag{2}$$

もし式 (1, 2) が分からないようなら基本的な確率が出来てないので，statistics.pdf で勉強してみてください.

さて，ここはちょっとテクいです.

この二つの関係から，x と y の確率をかける操作をしたものに対して何らかの処理をしたものが，何らかの処理をした x と y の和になっているので，関数 $h(\cdot)$ は対数をとっている事が分かり（ほら，掛け算って対数だと足し算じゃん？），

$$h(x) = -logp(x) \tag{3}$$

がいえます．log (x) は単調増加で，確率 p(x) は常に 0 から 1 の範囲をとるため，$h(x)$ の値を常に正にするため符号を反転させている事に注意です．ここで対数の底に 2 を採用するのが情報理論での一般の使い方で，その場合は $h(x)$ の単位は bit になるようです.
これを使って，信号のサイズ，情報量 (bit) を算出してるわけですね．あとデータの圧縮なんかにも関係するぽいですがそこまでは知らないし触れませんし触れられません.

次に，この値を分布全体に適用する事を考えます．つまり，確率分布そのものが与える情報量です．その指標として，確率変数 x の分布 $p(x)$ に関して $h(x) = -logp(x)$ の期待値をとることで，情報量の平均を定義します．確率変数の期待値の一般的な計算です．これも分からなければやはり statistics に行ってください．

$$H[x] = -\Sigma_x p(x) logp(x) \tag{4}$$

式 (4) に定義する量をシャノンエントロピーといいます [6]．意外と簡単ですね．もっと難しいと思ってました．

ついでにこれを連続変数にすると微分エントロピー (式 5) が求まります．

$$H[x] = -\int p(x) logp(x) dx \tag{5}$$

さて，次に当然浮かぶ疑問は，どんな確率分布だとどんなエントロピーが算出されるのか，です．
ちゃんと数学的に証明することも出来るっぽいけど面倒だしそこにあんま興味ないので，simulation してみます．

とりあえず一様分布で確認してみます．0.1 から 1 の値をとる，サンプル数 10 の一様分布 (図 2.1) のエントロピーを算出します (コード 1)．

Listing 1　エントロピーの計算

```
1 x = [0.1:0.1:1];
2 px = zeros(1,10)+0.1;
3 H = - sum(x .* log(px));
```

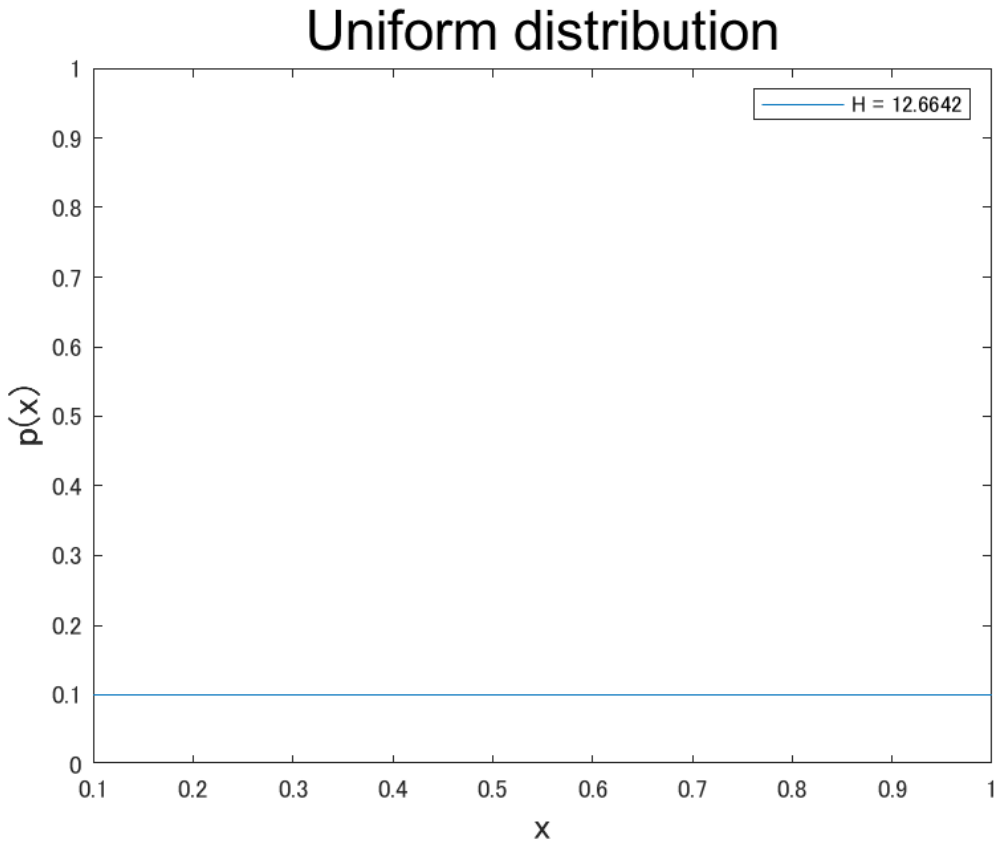

図 1　0.1 が 10 この一様分布

エントロピーは 12.7 でした．次に，同じ一様分布でもサンプル数が多いとどうなるのか試します．先ほどと同じ条件の，100 個のデータ (図 2.1) です．

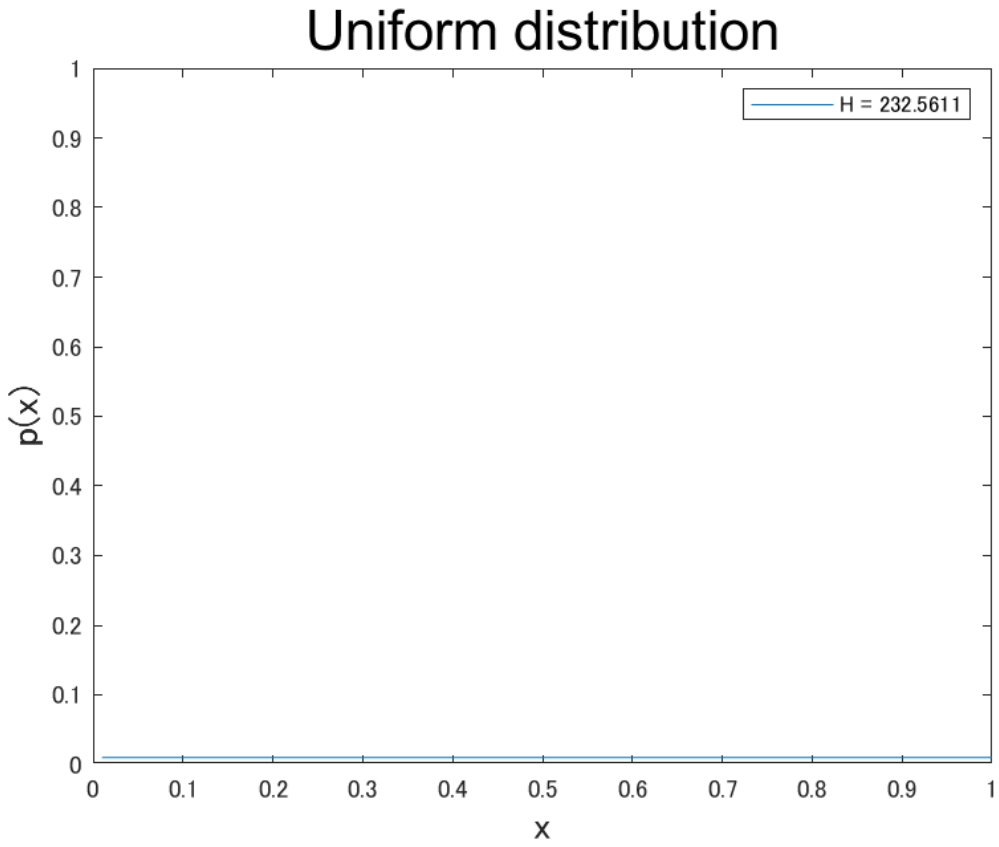

図 2　0.01 が 100 この一様分布

　エントロピーは 232．5．大きくなりましたね．データ数に応じてエントロピー自体は大きくなるぽいです．

次に山を持たせた分布で見てみます．

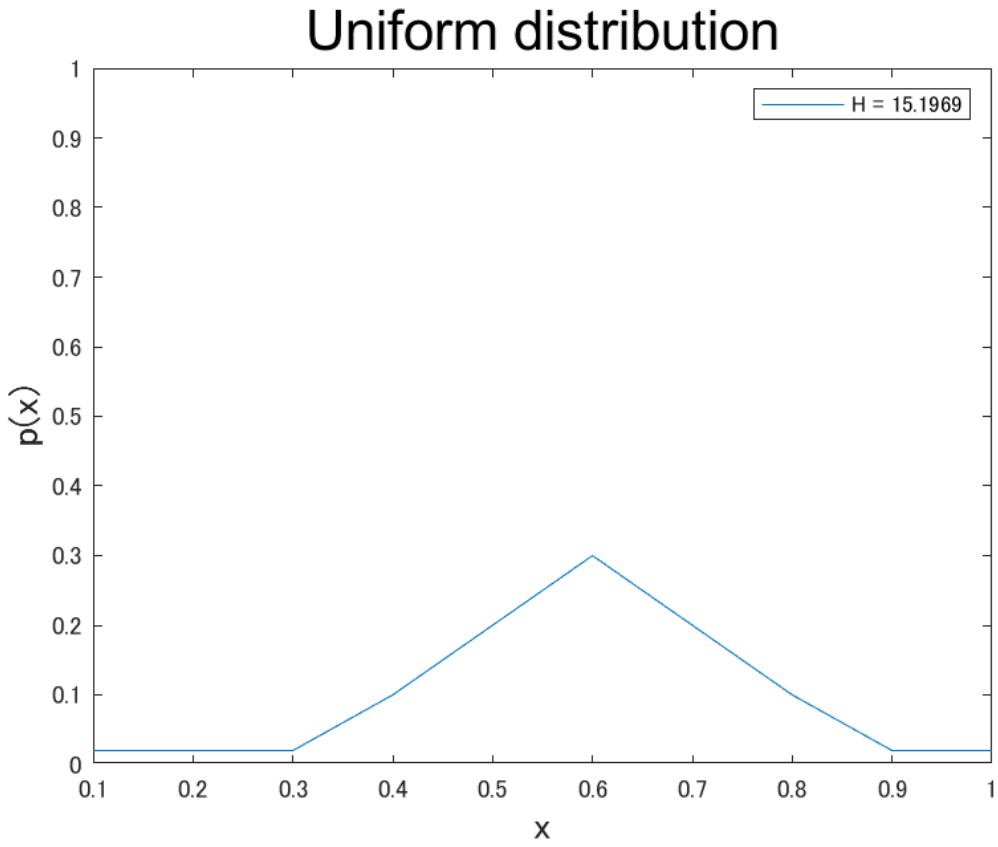

図 3　適当につくった山あり分布

エントロピーは 22.3．一様分布に比べるとかなり小さいです．本当はもっと山の位置動かしたり急峻にさせてみたりと比較したいですが，飽きたので結論．(離散) エントロピーは確率分布 $p(x)$ が一様分布に近付く程大きくなり，一様分布の時に最大になります．$x * log(x)$ なのでまあ，考えればそうかなって感じ．証明は結構めんどくさいぽいです．

ちなみに微分エントロピーの場合はガウス分布が最もエントロピー高いようです．

他にも見れる性質があって，まず $p(x)$ が 0 は困ります．log にかけた時に計算がこわれるので．無限に吹っ飛びます．

あと，一様分布の比較から分かるようにデータ数が多いほどエントロピーも増大するぽいですね．これもまあ普通に総和とってるんだから当たり前か?

2.2 KL 距離

さて，このエントロピーがどんな事に使えるのか考えていきます．エントロピーは分布の特徴を表す量になっていたわけなので，これを使うと二つの分布の比較，なんてことも出来ることになります．直観的には，全く同じ特徴の分布同士ならそのエントロピーに差はないし，違う分布なら差がある，という感じです．

式にしてみましょう．まず，微分エントロピー $(-\int p(x)logp(x))$ は確率変数 x の分布 $p(x)$ の元での期待値でした．なら，ここで新しい分布 $q(x)$ を考えたとき，仮にこの分布が同一 $(p(x) = q(x))$ であれば，

$$\int p(x)logq(x)dx - \int p(x)logp(x)dx = 0 \tag{6}$$

が成り立つ事になります．同じ分布の元で考えた同じ確率変数の期待値だから，当たり前です．分布 $p(x)$ の元で見た $q(x)$ の期待値が，分布 $p(x)$ の元で見た $p(x)$ の期待値と等しい，ということです．逆にこの分布が異なるものであるほど，この計算の結果は大きな値を取る事になります．

てなわけで，この量をちゃんと正負の調整した上で，「分布 $p(x), q(x)$ の相対エントロピー，あるいはカルバック-ライブラー距離，またはカルバックライブラーダイバージェンス」として以下の式で定義します [6].

$$\mathbb{D}_{KL}(p||q) = -\int p(x)logq(x)dx - (-\int p(x)logp(x)dx) \tag{7}$$

$$= -\int p(x)log\frac{q(x)}{p(x)}dx \tag{8}$$

一寸ややこしく見えますが，基本的には式 (6) を log について整理しただけです．
簡単ですね．

KL 距離の性質ですが，まず $\mathbb{D}_{KL}(p||q) \geq 0$ です．距離だし．
等号が成り立つのは分布 $p(x), q(x)$ が等しいときのみです．

それから $\mathbb{D}_{KL}(p||q) \neq \mathbb{D}_{KL}(q||p)$ なことにも気を付けてください．分布 $p(x)$ の元で見た $q(x)$ の期待値と，分布 $q(x)$ の元でみた $p(x)$ の期待値とは別物ですからね．

それから，対数なので KLD は以下のような表記のこともあります．一緒です．log の計算の性質を思い出してください．割り算は引き算です (?)

$$\mathbb{D}_{KL}(p||q) = \int p(x)log\frac{p(x)}{q(x)}dx \tag{9}$$

留意してください．

あと，KL “距離” と日本語で呼んでいますが厳密には距離じゃないので注意が必要です．というのも，KLD は以下に示す距離の公理 [7] を満たしていないからです．

距離の公理
2 点 A,B が与えられたとき, 実数 $d(A,B)$ を与える規則で, 次の性質を満たすものを距離という.

- $d(A,B) \geq 0$
- $d(A,B) = 0 \leftrightarrow A = B$
- $d(A,B) = d(B,A)$
- $d(A,B) + d(B,C) \geq d(A,C)$

このうち，KLD が満たしていないのはなんでしょう？
そう，３つめの対称性ですね！ $\mathbb{D}_{KL}(p||q) \neq \mathbb{D}_{KL}(q||p)$ でした．

あと４つ目，三角不等式も怪しいと思うんですよね．呼んでた資料とかでは特に対称性のとこだけネチネチと言われてましたが，三角不等式はどうなんでしょう？

直観的には微妙だと思ってて，だから KLD の値を単純に比較したりだとかの議論は出来ない気がしている．

じゃあ KLD はどう使うんだよって話ですが，最小化したい量として導入してるのが多い気がします．
つまり，予測分布を真の分布に近づけたい，だとかですね．この時に最小化する，分布と分布との距離として使われる量です．
あとは，一様分布と得られたデータ分布との距離を測る，なんて使い方もありました [5]．この場合は正規化的なのして，$\mathbb{D}_{KL}(P||U)$(where U is the uniform dist) を 0-1 の値にして使ってましたね．

いまのとこ個人的に分からないのは，$\mathbb{D}_{KL}(A||B)$ と $\mathbb{D}_{KL}(C||D)$ の値を比較した議論 (たとえば，A-B は C-D の 3 倍離れている！) なんてのは出来るのかなってところです．
ユークリッドなら自明に出来ると思うんですけど，これだとなんか出来ない気がする．三角不等式も怪しいし．
どうなんでしょう？今後の課題になってます．

余談ですが，式 (9) の右辺第一項，$-\int p(x)logq(x)dx$ は交差エントロピー $H(p,q)$ とも言います．分布 $p(x)$ の元で見た $q(x)$ の期待値なので，$p(x)$ の分布を想定したとき，$q(x)$ がどれだけ予測しにくいかとも捉えられます．

これだけでも，交差エントロピー $H(p,q)$ は正解値と推定値の比較なんかの用途で使えるようです．

じゃあ KLD と何が違うのか，というと，ここからは個人的な予想ですが...
問題なのは $p(x)$ 自体の分布が既にもってる情報量，つまり $H(p)$ なんだと思います．
交差エントロピーは計算式をみれば分かるように，$p(x)$ 自体のエントロピーの影響を受けた数値になってしまうため，なんというか「どれくらい外れているか」の指標に使うにはフェアじゃない気がします．

なので交差エントロピーの値から，$p(x)$ 自体が持っているエントロピーの値を差し引いた量が知りたいわけですね．そうすると式 (9) は

$$\mathbb{D}_{KL}(p||q) = H(p,q) - H(p) \tag{10}$$

とも捉えられますね．あってるのかな？

2.2.1 他の分布間距離

確率分布同士の距離を測る指標は $\mathbb{D}_{KL}$ だけでなく，他にも以下のようなのがあるっぽいです [8]．

$$\chi^2(Q||P) := \Sigma_{i=1}^{k} \frac{(p_i - q_i)^2}{p_i} \qquad \chi^2 \text{ 統計量} \tag{11}$$

$$L_1(Q||P) := \int |Q(x) - P(x)|dx \qquad L_1 \text{ ノルム} \tag{12}$$

$$L_2(Q||P) := \int \{Q(x) - P(x)\}^2 dx \qquad L_2 \text{ ノルム} \tag{13}$$

$$I_K(Q||P) := \int \{\sqrt{Q(x)} - \sqrt{P(x)}\}^2 dx \qquad \text{ヘリンジャー距離} \tag{14}$$

$$\mathbb{D}(Q||P) := \int f(\frac{Q(x)}{P(x)})Q(x)dx \qquad \text{f-ダイバージェンス} \tag{15}$$

$$I_\lambda(Q||P) := \int \{(\frac{Q(x)}{P(x)})^\lambda - 1\}Q(x)dx \qquad \text{一般化情報量} \tag{16}$$

$$\mathbb{D}_{KL}(Q||P) := \int log(\frac{Q(x)}{P(x)})Q(x)dx \qquad \text{KL 情報量} \tag{17}$$

どれがどんな時にどう使われるとかは調べてないです，でもこれでいくと上 3 つはよく見る気がする．でもまあ全体的に似てるポイですね．なんとなく哲学というか考え方はどれも似たりよったりな気がします．

2.3 条件付きエントロピー

エントロピーは式 (5) に示す量でしたが，これを条件付き確率 $p(x,y)$ に拡張して考えます．今 x が既知である場合，同時分布 $p(x,y)$ について y を特定するための情報は $-logp(y|x)$ なので（これはいいよね？条件付き確率です），その合計は

$$H(y|x) = -\iint p(y,x)logp(y|x)dydx \tag{18}$$

で表され，これは条件付きエントロピーといいます [6][8]．さらにこれを使えば

$$H(x,y) = H(y|x) + H(x) \tag{19}$$

と書けますね！エントロピーは対数なので，確率の乗法を意味しています．つまり x と y の同時分布を記述する情報量は，x 単体の情報量と x が与えられた元での y の情報量との和になるわけですね．

2.4 相互情報量

KL 距離は分布と分布の距離を測れる便利な指標でした．
これを使った，これまた便利そうな指標の一つが相互情報量です．二つの変数 x, y を考えて，こいつらの同時分布 $p(x,y)$ が得られたとします．この時，この変数２人の間にどんな関係があるのか確認したくなりますよね．他人なのか，それとも親密な関係なのか... まあつまり独立かどうかです．

さて，ＫＬ距離はこの独立性の検証的な使い方が可能で，それがまさに相互情報量の計算です．式 (20) を見た方が早いでしょう [8]．

$$MI(x,y) = H(x) + H(y) - H(x,y) \tag{20}$$

x と y が独立であった場合の同時確率の情報量と独立でないときの情報量の離れ具合を見るわけですね．例のごとく対数なので，要は $p(x)p(y)$ と $p(x,y)$ です．これは KLD を使えば式 (21) のように表せます．

$$MI(x,y) := \mathbb{D}_{KL}(p(x,y)||p(x)p(y)) = -\iint p(x,y) log \frac{p(x)p(y)}{p(x,y)} dxdy \tag{21}$$

あら簡単．変数 x, y の同時分布と周辺分布積との KLD を見るだけですね．KLD なので，両者が同じ，つまり x と y が独立である時に限って 0 になる量ってわけですね．

てことは，x と y がずぶずぶの関係であるほど値が大きくなるわけだから，y の値を知る事によって x の不確実性が減った度合を表すと言えそうです [6]．

KLD 同様，符号反転で以下の表記 (式 22) もあります．

$$MI(x,y) := \iint p(x,y) log \frac{p(x,y)}{p(x)p(y)} dxdy \tag{22}$$

参考文献

[1] Lazepnik, Y. (2002). “Can a biologist fix a radio?-Or, what I learned while studying apoptosis.” Cancer Cell, 2(3), 179-182.

[2] Jonas, Eric, and Konrad Paul Kording. (2017). ”Could a neuroscientist understand a microprocessor?.” PLoS computational biology 13.1.

[3] Marr. (1982). “Vision.”

[4] Friston, K. (2010). “The free-energy principle: A unified brain theory?” Nature Reviews Neuroscience. Volume 11, Issue 2, February 2010, 127-138

[5] Adriano B. L. Tort, et al. (2010). “Measuring Phase-Amplitude Coupling Between Neuronal Oscillations of Different Frequencies” Journal of Neurophysiolosy 104:1195-1210.
[6] Christopher M. Bishop. (2006). “Pattern Recognition and Machine Learning”
[7] Wikipedia
[8] yumaloop. “Kullback-Leibler Divergence について まとめる” https://yul.hatenablog.com/entry/2019/01/07/152738