Watanabe 理論メモ

@phykm

2017 年 3 月 30 日

概要

渡辺澄夫氏による学習理論の基礎 [1] をやっていきするにあたってのメモ。漸近論に入る前に、まず各種概念にきちんと所属集合 (型) をつけて見通しを立てる。ひとまず [1] の 2 章までで登場する概念 (キュムラント母関数以外) に型を付ける。

1　学習理論全般の前提

統計的推測の状況を、ひとまずここでは「ある可測空間が与えられ、そこでの確率的トライアルが任意回行えるとしたときに、この確率分布を得るまたは近似するための方法と、その性能を解析する手段を得たい」という状況のこととしよう。重要なことは、確率変数のトライアルができるだけで、その分布を直接触ることはできない点である。目標とする系が古典的な確率測度で表現できる事自体は仮定する。

このトライアル自体を表現する可測空間を得なければ始まらない。(X, Σ_X) を目標とする可測空間、真の確率測度を μ としよう。任意回この確率空間のトライアルが可能であるということは、それ全体はこの (X, Σ) の可算直積による可測空間で書けることになる。$(X^\omega, \Sigma_X^\omega)$ をその可算直積可測空間とする。この Σ_X^ω は有限個の射影が可測になる最小 σ 代数で生成されることに注意。トライアルを有限回行うことで、その結果から (X, Σ_X) 上の測度を推定するのだから、推定アルゴリズムとは、$(X^\omega, \Sigma_X^\omega)$ のもつ自然なフィルトレーション

$$\mathcal{F}_i = (X^i, \Sigma_X^i) \times (X^\omega, \{\emptyset, X\}^\omega) \tag{1}$$

にしたがう、$G(X)$ 値確率過程ということになる。$G(X)$ は (X, Σ_X) 上の確率測度のなす可測空間で、$A \in \Sigma_X$ についての $\mathrm{ev}_A = (\lambda\mu.\mu(A))$ を可測にする最小 σ 代数を取るとする。したがって、n 回目までのトライアルによる推定 $\alpha_n : (X^\omega, \Sigma_X^\omega) \to G(X)$ は、X^ω の最初の (X^n, Σ_X^n) だけで決まる。

Definition 1.1. (X, Σ_X) 上の推定アルゴリズムとは、確率過程 $\{\alpha_n\}_n : (X^\omega, \Sigma_X^\omega, \{\mathcal{F}_n\}_n) \to G(X)$ とする。α_n は $\mathcal{F}_n$ について可測である。α_n を (X^n, Σ_X^n) 上の可測関数とみなした時に $\overline{\alpha_n}$ とする。例えば $\alpha_n(x_1, x_2 ... x_n ...) = \overline{\alpha_n}(x_1, x_2 ... x_n) \in G(X)$。以下で定義していく確率過程も同様の記法に従う。

推定アルゴリズムの結果確率的に得られる確率測度について、その真の分布との一致度を測るために通常 KL-divergence が用いられる。

Definition 1.2. 確率測度の空間 $G(X)$ 上、次で決まる正定値二項関数を *KL-divergence* と呼ぶ。

$$D(\mu/\nu) = \int d\mu \log \frac{d\mu}{d\nu} \tag{2}$$

ただし、$\frac{d\mu}{d\nu}$ は *Radon-Nikodim* 微分とし、$\mu \ll \nu$ でない場合はこれを無限大とする。

KL-div の解釈は、抽象的なものから具体的なものまで様々である。符号化に即して言うなれば、可測空間を有限として、ν に基づく対数符号長エンコードを、実際の生成分布が μ である情報源に対して行った場合の、平均符号長損失（最適な場合に比較した冗長さ）はちょうど KL-div である。「情報量」とは確率の対数であるということを受け入れれられるならば、KL-div は ν くりこみ済み μ の情報量といってもいい。ただしこうしたエントロピー理論に基づく解釈には限界があり、それはこの解釈がいずれも有限可測空間でしか正確な意味を持たないという事情による。エントロピーは有限測度空間においてのみ定義可能だから、KL-div はよりひろい定義域をもつ[*1]。よく知られているように、これは正定値ではあるが対称でないので、確率測度における「距離」概念としてどのくらい適切かどうかに関してはわからない部分がある。しかしとにかくこれを採用するとしよう。

さてこれを基準に推定アルゴリズムの誤差を示す確率過程が手に入る。

Definition 1.3. 次の確率過程を汎化誤差と呼ぶ。

$$\{E_n\}_n = D(\mu/\{\alpha_n\}_n(-)) : (X^\omega, \Sigma_X^\omega, \{\mathcal{F}_n\}_n) \to \overline{\mathbb{R}} \tag{3}$$

$$\overline{E_n} : (X^n, \Sigma_X^n) \to \overline{\mathbb{R}} \tag{4}$$

$$(x_1...x_n) \mapsto D(\mu/\overline{\alpha_n}(x_1...x_n)) = \int d\mu \log \frac{d\mu}{d\overline{\alpha_n}(x_1...x_n)} \tag{5}$$

これは誤差の名の通り、アルゴリズムが正常に稼働すれば、ゼロに近づくことが期待される確率変数で、何らかの極限で確定的に 0 になれば、KL-div の正定値性から、その学習アルゴリズムは正確に真の分布を推定したことがわかる。つまりこの確率過程を追跡することで、推定の進行程度がわかるのである。推定アルゴリズムによっては、完全にゼロにはならないかもしれないが、その場合はこの値が最小値に到達した事でそれを「最適」とみなして推定完了とする。

しかしこれは、目標分布 μ がわかっている時に、そこへの漸近性能をみるためには使えても、現実の状況では $x_1...x_n$ という結果を得ても計算することが出来ない。何度かのトライアルの結果に対して、この E_n を計算するには真の分布 μ が必要であり、そのようなことは当然叶えられないからである。そこでこの値に近い値を取ると期待される別の確率変数を追跡することを考える。

真の分布 μ および、α_n たちの像である予測分布を絶対連続に従えると期待できる基準測度 ν を一つ固定し、これについての微分エントロピー

$$S_\nu(\mu) = -\int d\mu \log \frac{d\mu}{d\nu} \tag{6}$$

を考えておく。

Definition 1.4. 次の確率過程を汎化損失と呼ぶ。

$$\{G_n\}_n = \{E_n + S_\nu(\mu)\}_n : (X^\omega, \Sigma_X^\omega, \{\mathcal{F}_n\}_n) \to \overline{\mathbb{R}} \tag{7}$$

$$\overline{G_n} : (X^n, \Sigma_X^n) \to \overline{\mathbb{R}} \tag{8}$$

$$(x_1...x_n) \mapsto D(\mu/\overline{\alpha_n}(x_1...x_n)) + S_\nu(\mu) = -\int d\mu \log \frac{d\overline{\alpha_n}(x_1...x_n)}{d\nu} \tag{9}$$

[*1] 微分エントロピーを持ち出す人がいるかもしれないが、あれはルベーグ測度に対する KL-div に相当する。ルベーグ測度は確率分布ですらないので、やはりこの解釈はできず、エントロピーの定義域をルベーグ測度の「くりこみ」で無理やり広げたものとみなせる。

つまり、微分エントロピーを加えることで被積分関数から μ を追放した。しかしこれも μ での平均を使っていることには変わりがない。そこで、μ での平均を経験平均に置き換えて次を定義する。

Definition 1.5. 次の確率過程を経験損失と呼ぶ。

$$\{T_n\}_n : (X^\omega, \Sigma_X^\omega, \{\mathcal{F}_n\}_n) \to \overline{\mathbb{R}} \tag{10}$$

$$\overline{T_n} : (X^n, \Sigma_X^n) \to \overline{\mathbb{R}} \tag{11}$$

$$(x_1 ... x_n) \mapsto -\frac{1}{n}\sum_{i=1}^{n} \log \frac{d\overline{\alpha_n}(x_1 ... x_n)}{d\nu}(x_i) \tag{12}$$

これならば、(α_n が近似的にでも計算できる限り)μ を知らなくても、トライアルの結果だけから計算することができる。実際に推定をさせるときに、サンプリングを行って、アルゴリズムをまわし、トライアルが増えるごとにこの T_n を計算する。T_n の挙動は G_n または定数の違いで E_n にある程度似たものになるはずで、この観察によって、推定アルゴリズムがどのくらいうまく行ったか、あるいは推定が完了したかどうかを判断することになる。これは早く小さくなればなるほどよい。ただしこれらは確率過程であるから確率的にゆらいでいるので、断定できるような量ではない。あくまで統計的挙動によって比較できる。

そうすると、推測における基本的な問題として次のような物が上がってくる。

- 具体的なアルゴリズム $\{\alpha_n\}_n$ をどのように設計するか。
- アルゴリズムや問題を固定したとき、汎化誤差 E_n、汎化損失 G_n、経験損失 T_n はどう振る舞い、どのくらい食い違うのか。
- これらの量の漸近挙動は何によって決まるのか。それはどうやって計算できるか。

したがって、統計的推測の理論とは、こうした疑問に対して（部分的にでも）解答をもたらすものを言う。

2 モデル

通常実用的な可測空間は、容易にその確率測度の空間全体が巨大になってしまうので、現実的には、その可測空間上の扱いやすい部分集合をとって、そこに目標の分布が所属するか、目標の分布へある程度接近できるようにし、かつそれがパラメトライズできるような状況を考える。ここで二つの宣言が考えられる。パラメータ空間を $\mathbb{R}$ 上のコンパクト可微分多様体 M、$U(M)$ を可微分構造の忘却とする。

- $U(M) \subset G(X)$ つまり、パラメータ空間は確率測度の部分集合それ自体とする。
- $P : U(M) \to G(X)$ つまり、パラメータ空間 M の点ごとに、確率測度を与えるとする。

おそらく、情報幾何学が想定しているのは前者の状況であり、学習理論が想定しているのは後者の状況である。情報幾何はフィッシャー計量のなすリーマン多様体を考えているので、この計量の正定値性や解析性が損なわれるのはよろしくない。だから、考察する確率測度の集合それ自体が多様体をなす状況を考え、その幾何について考察する。一方学習理論は、目標の確率分布を推測するにあたって、それに対してより複雑な構造を考えて、その結果として目標の確率変数が出てくるようにする。あくまでその構造の一側面として目標値が出るように組み立てられるのだから、単射性や尤度関数の正則性は保証されない。このメモは推定理論をメインにするので、後者を考える。

パラメータやその背景構造は、目的や推定したい確率変数によって大きくかわりうる。例えば既に対象とする確率変数をもつ系の、信頼できる理論的解析が進んでいて、あとはそのパラメータを決定するだけであれ

ば、そのパラメータ空間を M に取るだろう。逆に理論的解析までも推定器械に任せるために、汎用の複雑な関数近似器を使うかもしれない。

写像 $P : U(M) \to G(X)$ は、統計的な解析がある程度し易いように選ばれるとしよう。これは便宜的な仮定である。具体的には、適当な可微分性がほしい。前節で、経験損失が推定の進行を表す重要なシグナルであると知っているので、基準測度 ν についての、モデルの対数尤度

$$-\log \frac{dP(-)}{d\nu} \tag{13}$$

は可微分であって欲しい。

Definition 2.1. (X, Σ_X) へのモデルとは、コンパクト多様体（理論空間）M から $G(X)$（測度空間）への写像 P で、適当に定めた基準測度 ν についての対数尤度が可微分であるものとする。

このようなモデルを利用する推定アルゴリズムは複数ある。ベイズはその一つだが、最尤推定もそうである。

Definition 2.2. モデル $P : M \to G(X)$ における最尤推定とは、$\{\alpha_n\} : (X^\omega, \Sigma_X^\omega, \{\mathcal{F}_n\}_n) \to G(X)$ を次で与えるものとする。

$$\alpha_n(x_1, x_2, ...) = P(r) \tag{14}$$

$$r = \arg\max_q \sum_{i=1}^{n} -\log \frac{dP(q)}{d\nu}(x_i) \tag{15}$$

つまり、尤度:尤もらしさを最大化するパラメータのモデル分布を返す。

Definition 2.3. モデル $P : M \to G(X)$ におけるベイズ推定とは、M 上の事前確率分布 $\phi \in G(M)$ を用いて、$\{\alpha_n\} : (X^\omega, \Sigma_X^\omega, \{\mathcal{F}_n\}_n) \to G(X)$ を次で与えるものとする。

$$\alpha_n(x_1, x_2, ...) = \int P(r) d\phi[\prod_{i=1}^{n} P](-|x_1...x_n)(r) \tag{16}$$

$$\tag{17}$$

ただし、$\phi[\prod_{i=1}^n P]$ は $(M, \mathcal{B}(M)) \times (X^n, \Sigma_X^n)$ 上の確率測度で、

$$\phi[\prod_{i=1}^{n} P](N \times A_1 \times A_2 ... A_n) = \int_N d\phi(r) \prod_{i=1}^{n} P(r)(A_i) \tag{18}$$

とする。$\phi[\prod_{i=1}^n P](-|x_1...x_n)$ は、この $x_1...x_n$ 条件付き確率測度である[*2]。

ベイズと呼ばれる所以はもちろん事前分布との結合分布から、出た結果を条件付け反転し、事前分布を更新することで推定測度を与えることによる。結合確率測度と条件付き確率を関連付ける公式がベイズの定理と呼ばれるのでこれもそう呼ばれる。

Definition 2.4. モデル $P : M \to G(X)$ における MAP 推定とは、M 上の事前確率分布 $\phi \in G(M)$ のもとで、上記と同様の $\phi[\prod_{i=1}^n P](-|x_1...x_n)$ について、$\{\alpha_n\} : (X^\omega, \Sigma_X^\omega, \{\mathcal{F}_n\}_n) \to G(X)$ を次で与えるものと

[*2] このような点条件付きを行うには一般には Regularity が要るが、ここでは Regularity が成り立つような状況のみを考えているとする。

する。

$$\alpha_n(x_1...x_n) = P(r) \tag{19}$$

$$r = \arg\max_q \frac{d\phi[\prod_{i=1}^n P](-|x_1...x_n)}{d\phi}(q) \tag{20}$$

事前分布による *Radon-Nikodim* 微分を用いることに注意。[*3]

以下はベイズ推定を考えていく。

3　ベイズ推定とカノニカルアンサンブル

幾つかの概念を導入して、ベイズ推定を正確に述べなおす。大まかな話として、統計物理とベイズ推定には、事後分布の計算が、事前分布を基準とした相空間測度の上の、経験対数損失ランダムハミルトニアンによるカノニカルアンサンブルの、低温極限とのアナロジーとして理解できるという関係がある。したがって、極々大雑把には、このハミルトニアンの「基底状態」すなわち、最小値まわりの確率分布に収斂していくことが期待されるのだが、問題はこの最小値の近傍が特異である、つまりヘシアンが潰れていたり、最小値を取る集合が点でなかったりする可能性があることである。もしそうでなければ、ラプラス最急降下法の確率変数版のようなものを考えれば、この漸近挙動は比較的素直に計算できるが、特異であるときには特異点を解消してから解析をすすめなければ正確な挙動が計算できない。この手続きのために解析性の仮定と、代数幾何におけるいささかの技術的計算が要求される。いずれにせよ尤度関数が解析的とみなせる場合において、たとえ特異点近傍であっても推測の漸近挙動を計算する方法がある、というのが [1] の主な主張と思われる。

モデル $P : M \to G(X)$ の元で、いくつかの概念を定義する。

Definition 3.1. 次の $\overline{\mathbb{R}}$ 値関数を平均対数損失という。すなわち、真の分布に基づくモデルの平均逆尤度である。従ってこれは事前に知ることはできないが、存在するものとして扱う。

$$L : M \to \overline{\mathbb{R}} \tag{21}$$

$$q \in M \mapsto -\int d\mu \log \frac{dP(q)}{d\nu} \tag{22}$$

次の M 上の確率過程を経験対数損失という。明らかに経験対数損失の期待値は平均対数損失であり、確率過程としての前者は後者に漸近することが期待される。

$$\{L_n\}_n : M \to (X^\omega, \Sigma_X^\omega, \{\mathcal{F}_n\}_n) \to \overline{\mathbb{R}} \tag{23}$$

$$\overline{L_n} : M \to (X^n, \Sigma_X^n) \to \overline{\mathbb{R}} \tag{24}$$

$$q \in M \mapsto (x_1...x_n) \mapsto -\frac{1}{n}\sum_{i=1}^n \log \frac{dP(q)}{d\nu}(x_i) \tag{25}$$

経験損失は、平均損失と違い、実際のトライアルによって決まるので計算できる。[*4]

[*3] 度々ベイズ推定は、事前分布が「主観」確率とみなされうることについて批判される事があるようだが、それは確率密度関数しか見ていないからそう見えるだけで、以上のように MAP もまた事前分布に依存しているし、最尤推定も確率を密度関数にしてその尤度を最大化するために (X, Σ_X) 上の先験的な基準測度 ν を要求する点では同様に批判の対象になる。そもそも推定をしようとしているのだから、そのための全ての構造に先験的バイアスがあって当然である。

[*4] 連続分布のときに、可測関数が計算できるとはどういうことか？というのは大変悩ましいが、ここでは大雑把に未知の値、特に真の分布 μ を直接参照しないような所与の関数は計算可能と呼んでしまう。

さてベイズ推定における、$x_1...x_n$ 条件づけした M 上の事後分布 $\phi[\prod_{i=1}^n P](-|x_1...x_n)$ は、条件付き確率の定義から、各 $N \in \mathcal{B}(M)$ について (X^n, Σ_X^n) 上の RN 微分である。

$$\phi[\prod_{i=1}^n P](N|x_1...x_n) = \frac{d\phi[\prod_{i=1}^n P](N \times -)}{d\phi[\prod_{i=1}^n P](M \times -)}(x_1...x_n) \tag{26}$$

これを次のように書き換える。まず

$$\frac{d\phi[\prod_{i=1}^n P](N \times -)}{d\phi[\prod_{i=1}^n P](M \times -)} = \frac{\frac{d\phi[\prod_{i=1}^n P](N \times -)}{d\nu^n}}{\frac{d\phi[\prod_{i=1}^n P](M \times -)}{d\nu^n}} \tag{27}$$

$\phi[\prod_{i=1}^n P]$ の定義を思い出して、$P(q)$ ごとの RN 微分でこれを代行する。

$$\frac{\frac{d\phi[\prod_{i=1}^n P](N \times -)}{d\nu^n}}{\frac{d\phi[\prod_{i=1}^n P](M \times -)}{d\nu^n}} = \frac{\int_N d\phi(q) \prod_{i=1}^n \frac{dP(q)}{d\nu}(x_i)}{\int_M d\phi(q) \prod_{i=1}^n \frac{dP(q)}{d\nu}(x_i)} \tag{28}$$

これを経験対数損失で書き換える。

$$\frac{\int_N d\phi(q) \prod_{i=1}^n \frac{dP(q)}{d\nu}(x_i)}{\int_M d\phi(q) \prod_{i=1}^n \frac{dP(q)}{d\nu}(x_i)} = \frac{\int_N d\phi(q) \exp(-nL_n(q)(x_1...))}{\int_M d\phi(q) \exp(-nL_n(q)(x_1...))} \tag{29}$$

統計力学のアナロジーを強くするために、分母を規格化定数と考えよう。既に n が逆温度のように見えるが、後の便宜を図ってアドホックなパラメータ β を挿入しておく。$\beta = 1$ がこれまで通りの場合である。

Definition 3.2. 次の確率過程を分配関数と呼ぶ。

$$\{Z_{\beta,n}\}_n : (X^\omega, \Sigma_X^\omega, \{\mathcal{F}_n\}_n) \to \overline{\mathbb{R}} \tag{30}$$

$$Z_{\beta,n}(x_1...) = \int_M d\phi(q) \exp(-n\beta L_n(q)(x_1...)) \tag{31}$$

さらに以下を自由エネルギーと呼ぶ。

$$\{F_{\beta,n}\}_n : (X^\omega, \Sigma_X^\omega, \{\mathcal{F}_n\}_n) \to \overline{\mathbb{R}} \tag{32}$$

$$F_{\beta,n}(x_1...) = -\frac{1}{\beta} \log Z_{\beta,n}(x_1...) \tag{33}$$

この時、経験による条件付き事後分布は

$$\phi[\prod_{i=1}^n P](-|x_1...x_n) = \frac{1}{Z_{\beta,n}} \int_{(-)} d\phi(q) \exp(-n\beta L_n(q)(x_1...x_n)) \tag{34}$$

となる。もちろん、$x_1...x_n$ を Curry することでこれ自体が M 上の確率測度に値をとる確率分布である。一応、事後分布を定義しておこう。

Definition 3.3. 次の確率過程を事後分布と呼ぶ。

$$\{\psi_n\}_n : (X^\omega, \Sigma_X^\omega, \{\mathcal{F}_n\}_n) \to G(M) \tag{35}$$

$$\psi_n(x_1...) = \phi[\prod_{i=1}^n P](-|x_1...x_n) = \frac{1}{Z_{\beta,n}} \int_{(-)} d\phi(q) \exp(-n\beta L_n(q)(x_1...)) \tag{36}$$

この形式で見れば、系のサイズ n でインデックスされた確率過程ハミルトニアン L_n による ϕ バイアスのかかったカノニカルアンサンブルが、ベイズ推定の事後分布を与えていることがわかる。統計力学からの注意点として、ハミルトニアンと分配関数には定数加算/定数乗算の任意性があることがある。少なくとも分布を得るという目的において、ハミルトニアンを定数加えても、分配関数がその対数だけ乗算されるだけで、分布はかわらない。したがって、経験対数損失をハミルトニアンにとるとしつつも、目的に応じてその値を一様にアジャストすることは認められる。カノニカルアンサンブルの拡大体積極限であるから、ハミルトニアンの振る舞いがある程度「都合の良い」ものであれば、これらの値（特に事後分布）の挙動は、ハミルトニアンの最小値近傍の振る舞いのみで制御される。従って、結果得られる推定分布の汎化損失もまた、ハミルトニアンの最小値近傍の振る舞いが支配的となる。そして問題となるのが、この最小値を取る集合の形状、そしてそれの漸近収束速度への影響である。

モデルを固定すると、真の分布 μ に対して、そのモデルがどこまで接近できるかが決まる。どのような事前分布を考えても、この「最適な」分布はかわらないので、汎化損失や汎化誤差の挙動は、この「最適な」分布によって底を打つことになる。そこで、これまで導入した確率過程を、この「最適な」分布基準にアジャストしたものを考えておく。

Definition 3.4. $Q_0 = \{q \in M | L(q) = \min_r L(r)\}$ を最適集合とよぶ。最適集合の点 q_0 をとって、

$$K = L - L(q_0) = \int d\mu \log \frac{dP(q_0)}{dP(-)} \tag{37}$$

$$K_n = L_n - L(q_0) = (x_1...) \mapsto \frac{1}{n}\sum_{i=1}^{n} \log \frac{dP(q_0)}{dP(-)}(x_i) \tag{38}$$

を正規化平均対数損失、正規化経験対数損失という。もちろん前者はパラメータ多様体上の関数、後者はパラメータ多様体上の確率過程である。これに応じて、正規化分配関数、正規化自由エネルギーを

$$Z^0_{\beta,n} = (x_1...) \mapsto \int_M d\phi(q) \exp\left(-n\beta K_n(q)(x_1...)\right) \tag{39}$$

$$F^0_{\beta,n} = -\frac{1}{\beta} \log Z^0_{\beta,n} \tag{40}$$

とする。このときもちろん $F_n(\beta, *) = nL_n(x_0) + F^0_n(\beta, *)$ である。これらは共に *(β パラメトライズされた)* 確率過程である。以上で出現する

$$f : M \to X \to \overline{\mathbb{R}} \tag{41}$$

$$q \mapsto x \mapsto \log \frac{dP(x_0)}{dP(q)}(x) \tag{42}$$

はしばし対数尤度比関数と呼ばれる。

最初一般論を想定し、汎化損失と経験損失を、恣意的な基準測度 ν を用いて定義したが、一度推定アルゴリズムを固定すると、そのモデルが実現できる最適な分布 $P(q_0)$ が決まるので、（適当な絶対連続性が満たされる限りにおいて）これを基準にとる $(\nu = P(q_0))$ 事ができる。そこで次を定義しておこう。

Definition 3.5. 次の確率過程を正規化汎化損失と呼ぶ。

$$\{G_n^0\}_n : (X^\omega, \Sigma_X^\omega, \{\mathcal{F}_n\}_n) \to \overline{\mathbb{R}} \tag{43}$$

$$\overline{G_n^0} : (X^n, \Sigma_X^n) \to \overline{\mathbb{R}} \tag{44}$$

$$(x_1...x_n) \mapsto -\int d\mu \log \frac{d\overline{\alpha_n}(x_1...x_n)}{dP(q_0)} \tag{45}$$

次の確率過程を正規化経験損失と呼ぶ。

$$\{T_n^0\}_n : (X^\omega, \Sigma_X^\omega, \{\mathcal{F}_n\}_n) \to \overline{\mathbb{R}} \tag{46}$$

$$\overline{T_n^0} : (X^n, \Sigma_X^n) \to \overline{\mathbb{R}} \tag{47}$$

$$(x_1...x_n) \mapsto -\frac{1}{n}\sum_{i=1}^{n} \log \frac{d\overline{\alpha_n}(x_1...x_n)}{dP(q_0)} \tag{48}$$

ここで、

$$\frac{d\overline{\alpha_n}(x_1...x_n)}{dP(q_0)} = \frac{\int_M d\psi_n(x_1...)(q)P(q)}{dP(q_0)} = \int_M d\psi_n(x_1...)\exp(-f(q)) \tag{49}$$

なので、

$$G_n^0(x_1...) = -\int_X d\mu(x) \log \int_M d\psi_n(x_1...)(q)\exp(-f(q)(x)) \tag{50}$$

$$T_n^0(x_1...) = -\frac{1}{n}\sum_{i=1}^{n} \log \int_M d\psi_n(x_1...)(q)\exp(-f(q)(x)) \tag{51}$$

である。

4 To be added

参考文献

[1] 渡辺澄夫:ベイズ統計の理論と方法, コロナ社,2015