特異学習理論のまとめ

東京工業大学
渡辺澄夫

このファイルは何を説明しているか

このファイルは事後分布がガウス関数で近似できなくても成り立つ学習理論を説明しています。

講義やセミナーでは紹介しませんが、関心がある人は読んでみてください。

深層学習・混合正規分布・ニューラルネット・ボルツマンマシン・ベイズネットワークなどの構造を持つ統計モデルでは、尤度関数が正規分布で近似できないために従来の方法で扱うことはできませんでした。それらのモデルの予測精度を知ることができる理論は、いまのところここで述べるものの他にはありません。(2015年8月現在)。

なお、「特異学習理論」は正則な場合を特殊な場合として含みます。

すなわち、ここで述べることは事後分布が正規分布で近似できない場合を主として想定していますが、近似できる場合でも(従来のものより高精度で)成り立ちます。

☆　小さな青字で書いてある部分は説明のための文章であり、読まなくても理解できる場合には読む必要はありません。

第1章　記号と準備

☆　最初の章では、いくつかの記号の定義をします。

確率モデル p(x,w) のかわりに対数尤度比関数 f(x,w) を用いて
考える問題を定式化します。ここでは数学的に難しい点はありません。
記号や記法に慣れてください。

情報源・統計モデル・事前分布

データ $x \in R^N$, パラメータ $w \in W \subset R^d$

(1) 情報源 $q(x)$ ～ X および $X_1, X_2, \ldots, X_n$

(2) 統計モデル $p(x|w)$

(3) 事前分布 $\varphi(w)$

（注）以下では Π, Σ は全て i=1,2,…,n の積と和を表すものとします。

☆ 情報源のことを真の分布とも呼びます。統計モデルは学習モデルと呼ばれることもあります。事前分布がどんなものでも、以下で述べる定理は成り立ちます。
$X_1, X_2, \ldots, X_n$ のことを学習データあるいはサンプルと呼び、Xをテストデータと呼びます。

事後分布と自由エネルギーの定義

事後分布

$$E_w[\quad] = \frac{\int (\quad) \prod p(X_i|w)\,\varphi(w)\,dw}{\int \prod p(X_i|w)\,\varphi(w)\,dw}$$

自由エネルギー （＝－対数周辺尤度）

$$F = -\log \int \prod p(X_i|w)\,\varphi(w)\,dw$$

☆ 「データに対してFを最小にする(p,φ)が良いモデルである」という設計法が提案されています。ただし F の最小化は次ページの汎化損失の最小化とは等価ではありません。

汎化・学習・交差損失の定義

予測分布　$E_w[\, p(x|w)\,]$

汎化損失　$G_n = -E_x[\, \log E_w[\, p(X|w)\,]\,]$

学習損失　$T_n = -(1/n) \Sigma \log E_w[p(X_i|w)]$

交差損失　$C_n = (1/n) \Sigma \log E_w[1/p(X_i|w)]$

☆　ベイズ法において情報源を推測したものが予測分布です。
予測分布がどのくらい情報源を正しく推測しているか調べたいのです。
汎化損失は予測分布の未知のデータ対する誤差を表します。
学習損失は予測分布の学習データに対する誤差を表します。
交差損失は予測分布の学習データに対する交差検証の値を表します。

主目標

次の式を導出します。

目標. ある関数 w=g(u), 定数 λ と m、パラメータw_0、
ある確率過程 S(u) が存在して、n →∞のとき次式が成立。

$$\left(\frac{n^{\lambda}}{(\log n)^{m-1}} \right) \frac{\prod p(X_i|w)\, \varphi(w)\, dw}{\prod p(X_i|w_0)} \rightarrow S(u)\, du$$

このことから、自由エネルギー、汎化損失、学習損失、交差損失の挙動が解明できます。

☆　左辺は、データに依存する事後分布であり、この挙動を解明したいのです。
一方、右辺はデータに依存しない確率分布です。関数　w=g(u) 、定数λとm、
確率過程 S(u) は、すべてデータに依存しません。上式は法則収束を表します。
このことを利用すると統計学で必要になる問題を解明することができます。

平均の記号の2種類

未来のXについての平均の記号

$$E_x[\quad] = \int (\quad) q(x)dx$$

学習データに関する平均の記号

$$E[\quad] = \int\int \cdots \int (\quad) \prod q(x_i)dx_i$$

補題　$E[G_{n-1}]=E[C_n]$　①

☆　この補題は、C_n が交差損失であることからすぐに得られます。

対数損失関数と最適パラメータ集合の定義

対数損失関数の定義　$L(w) = -E_x[\log p(X|w)]$

最適パラメータ集合の定義　$W_0 = \{ w \in W \ ; \ L(w) \text{ 最小} \}$

仮定 :実質的に唯一　$(\forall w_0 \in W_0) \ \exists p_0(x) = p(x|w_0)$

$$\begin{cases} L_0 = -E_x[\log p_0(X)] & \text{とおく} \\ L_n = -(1/n)\Sigma \log p_0(X_i) & \text{とおく} \end{cases}$$

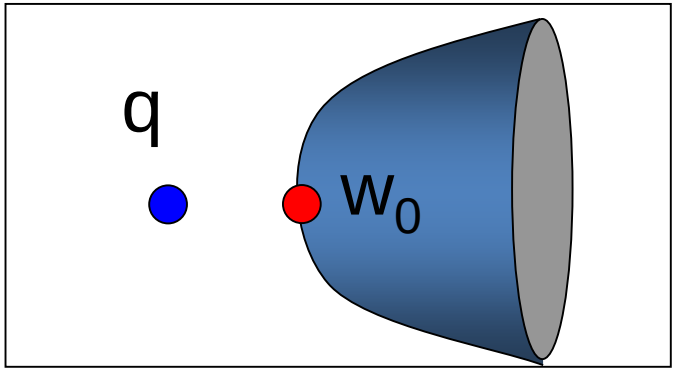
定義から $L_0 = E[L_n]$ が成り立つ

☆ 「データに対してFを最小にする(p,φ)が良いモデルである」という設計法が提案されています。ただし F の最小化は次ページの汎化損失の最小化とは等価ではありません。

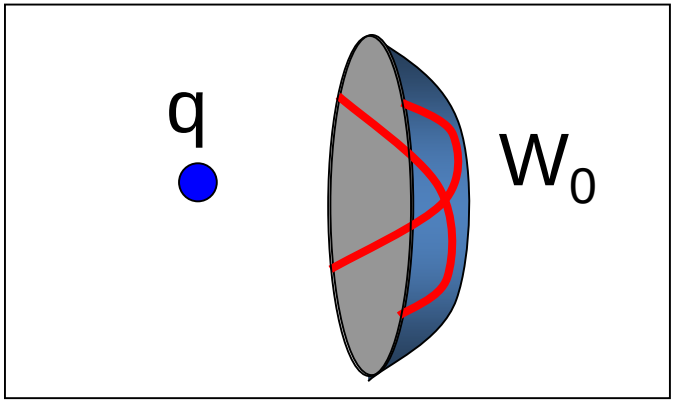
正則と特異の定義

W_0 の要素がひとつ w_0 だけであり、
かつヘッセ行列 $\nabla^2 L(w_0)$ が正則であるとき、
$q(x)$ は $p(x|w)$ に対して正則であるという。

そうでないとき特異であるという。



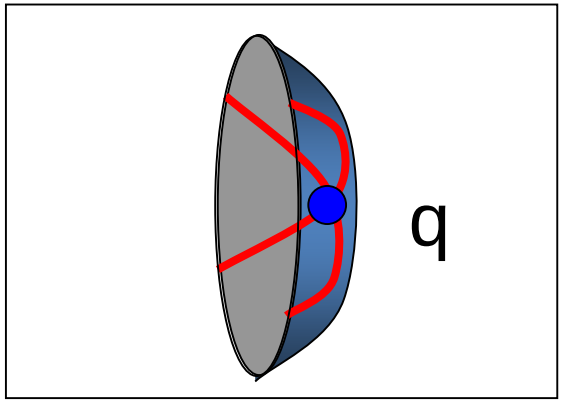
正則



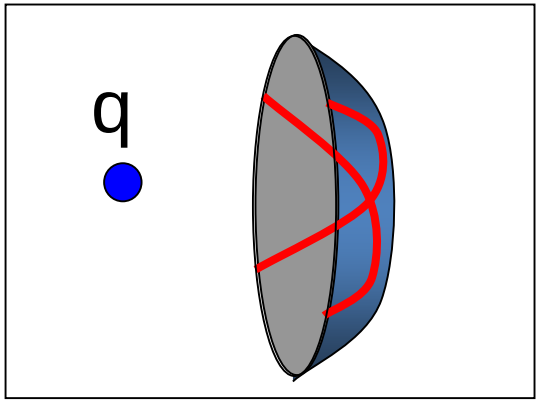
特異

実現可能の定義

条件　$q(x) = p_0(x)$ が成り立つとき、
$q(x)$ は $p(x|w)$ によって実現可能であるという。



実現可能



実現可能でない

関数 f(x,w) の定義

記号　$f(x,w) = \log(p_0(x)/p(x|w))$　を用いる。

このとき定義から　$p(x|w) = p_0(x) \exp(-f(x,w))$　である。

[定義]　ある定数　$\varepsilon>0$　が存在して

$$(\forall w) \quad E_x[\, f(X,w)\,] \geqq \varepsilon\, E_x[\, f(X,w)^2\,],$$

が成り立つとき、f(X,w) は相対的に有限な分散を持つという。

☆　f(x,w) を対数尤度比関数と呼びます。P(x,w) の代わりに f(x,w) で問題を書きなおすと、$n \to \infty$ のとき、$E_w[\, |f(x,w)|\,] \to 0$ を示すことができるので、解析しやすくなります。

定義の間に成り立つ包含関係

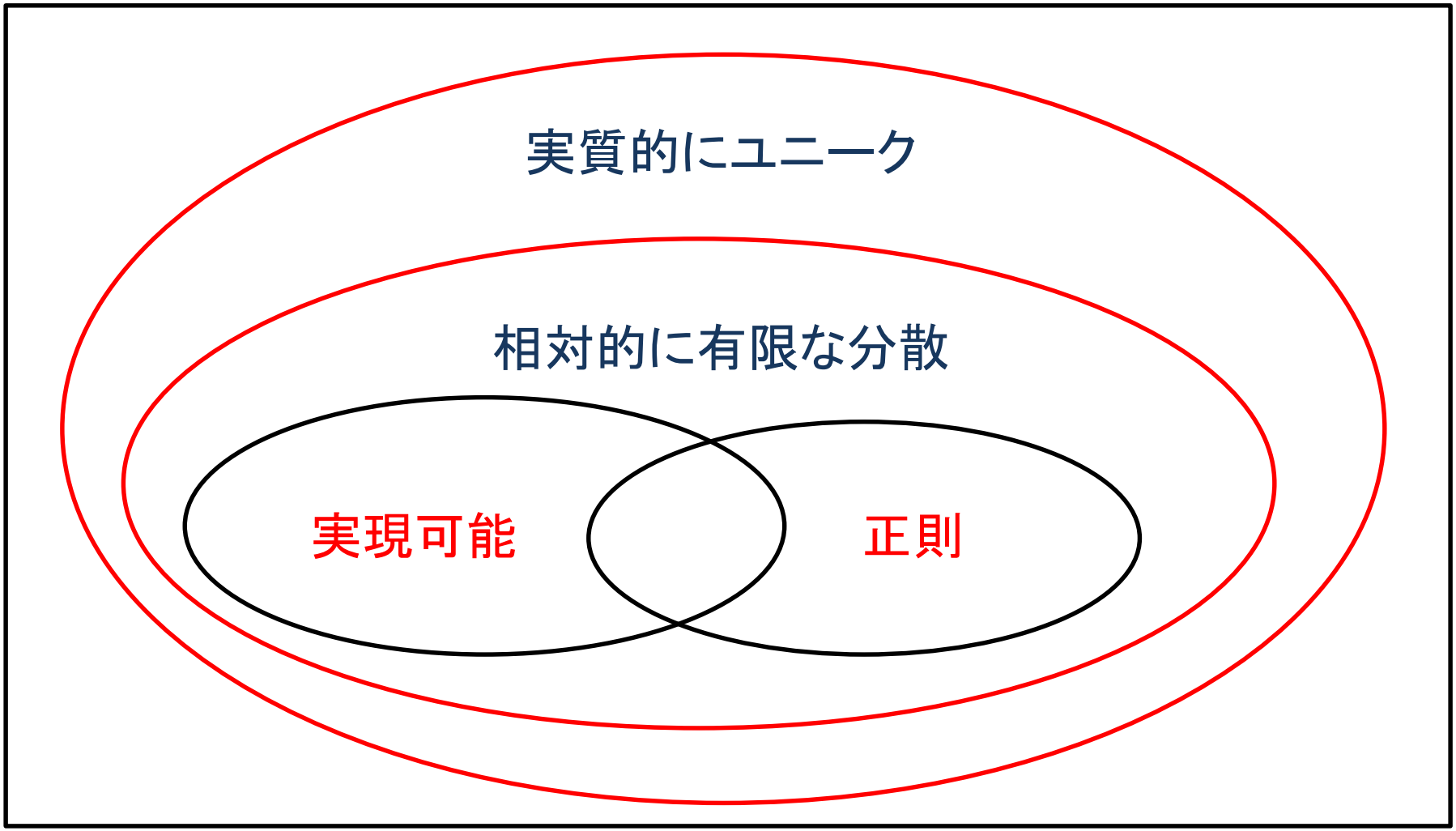
補題.　もしも q(x)が p(x|w)に対して正則なら
f(X,w) は相対的に有限な分散を持つ

補題.　もしも q(x) が p(x|w) により実現可能なら
f(X,w) は相対的に有限な分散を持つ

補題.　もしも f(x,w)が相対的に有限な分散を持てば
$p(x|w_0)$ は実質的に唯一。

☆　q(x) が p(x|w) に対して特異であり、かつ実現可能でないときは、
相対的に有限な分散を持つときとそうでないときとがある。
相対的に有限でないときは F, T_n, G_n, C_nの漸近挙動が変わりうる。
以下では相対的に有限な分散を持つことを仮定する。

条件の包含関係の図示



「p(x,w) → f(x,w) 」の等価変形(1)

$p(x|w) = p_0(x) \exp(-f(x,w))$ を用いると

事後分布，F, G_n,T_n,C_n は f(x,w) で書き直せる。

事後分布

$$E_w[\quad] = \frac{\int (\quad) \exp(- \Sigma f(X_i,w)) \varphi(w)\, dw}{\int \exp(- \Sigma f(X_i,w)) \varphi(w)\, dw}$$

自由エネルギー

$$F = nL_n - \log \int \exp(- \Sigma f(X_i,w)) \varphi(w)\, dw$$

記号 $K_n(w)$ と $K(w)$ の定義

$$f(x,w) = \log(p_0(x)/p(x|w))$$

$$K(w) = E_x[f(X,w)] \geqq 0$$

$$K_n(w) = (1/n)\ \Sigma\ f(X_i,w)$$

事後分布

$$E_w[\quad] = \frac{\int (\quad) \exp(-nk_n(w))\ \varphi(w)\ dw}{\int \exp(-nK_n(w))\ \varphi(w)\ dw}$$

自由エネルギー

$$F = nL_n - \log \int \exp(-nK_n(w))\ \varphi(w)\ dw$$

「p(x,w)→f(x,w)」の等価変形(2)

汎化損失

$$G_n = L_0 - E_x[\log E_w[\exp(-f(X,w))]]$$

学習損失

$$T_n = L_n - (1/n) \Sigma \log E_w[\exp(-f(X_i,w))]$$

交差損失

$$C_n = L_n + (1/n) \Sigma \log E_w[\exp(f(X_i,w))]]$$

☆ モデル p(x|w) が与えられたときの F, G_n,C_n,T_n を調べるかわりに関数 f(x,w) が与えられたときの F, G_n,C_n,T_n を調べればよいことがわかりました。次ページ以降では、その解析を行います。

第2章　平均関数と揺らぎ関数の挙動

ここでは、学習理論を展開するためにどうしても必要になる数学的な基盤を説明します。このファイルの目標は学習理論の全体的な構造を把握することですので、個々の数学的な要素について詳しい説明はしません。学習理論の全体構造が把握できた後に、さらに深く数学を学びたい人は、それぞれの問題が書かれている数学の本を読みましょう。

数学の定理が必要になる場所では、その定理の背景や証明までも説明するとあまりにも枝道が多くなるので、ここでは定理の記述を認めて進みましょう。証明を述べるかわりにその定理の例をあげます。

平均関数と揺らぎ関数に分ける

事後分布は $\exp(-nK_n(w))$ という形をしていますが分けます。

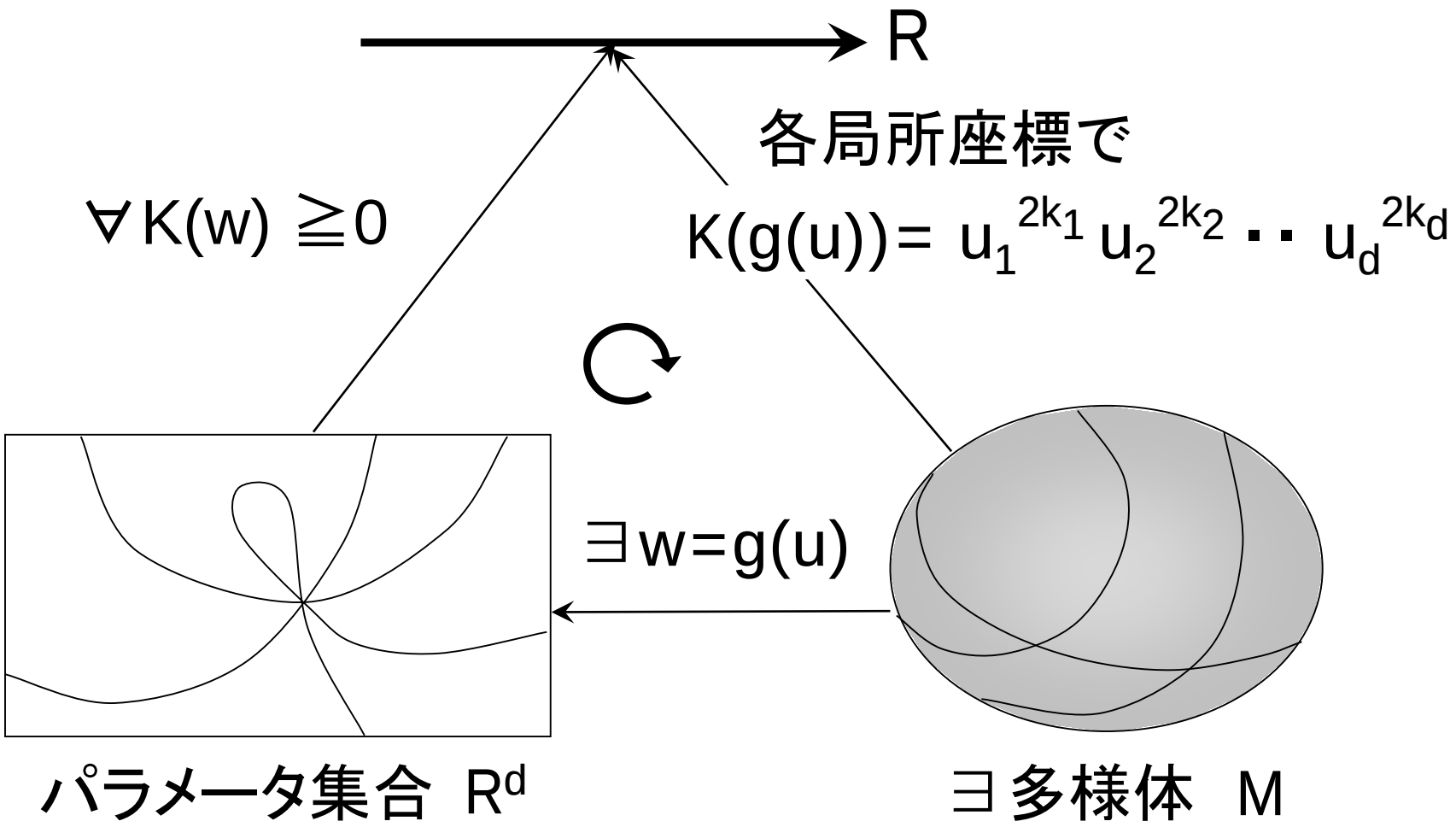
$$n\,K_n(w) = nK(w) - \sum_{i=1}^{n} \{\, K(w) - f(X_i,w) \,\}$$

平均の関数　　揺らぎの関数

事後分布の挙動を調べるために次の二つを考えます。

(1) $n \to \infty$のときの $\exp(-n\text{平均})$ の挙動
(2) $n \to \infty$ののときの揺らぎ関数の挙動

特異点解消定理： 任意の解析関数K(w)に対して、ある多様体Mと
ある解析関数　w=g(u)　が存在して、K(g(u))　は変数毎の積として書くことができます。



☆　特異点解消定理は証明は難しいですが、定理の記述は難しくはありません。
ここでは定理を認めて進み、証明したい人は数学科に入学しましょう。

対数閾値と多重度の定義

各局所座標で　$K(g(u)) = u_1^{2k_1} u_2^{2k_2} \cdots u_d^{2k_d}$

$= u^{2k}$　と書く

$|g(u)'|\varphi(g(u)) = b(u)\, u_1^{h_1} u_2^{h_2} \cdots u_d^{h_d}$

$= b(u)\, u^h$　と書く　$(b(u)>0)$

対数閾値　$\lambda = \min_{\text{局所座標}} \min_{j=1,2,\ldots,d} (h_j+1)/(2k_j)$

多重度　$m =$ 上記のminを与えるjの個数の最大値

☆　対数閾値は、高次元代数幾何学で大切な役割を果たすことが知られていますが、学習理論においては事後分布の挙動を定める主要な値であることがわかります。

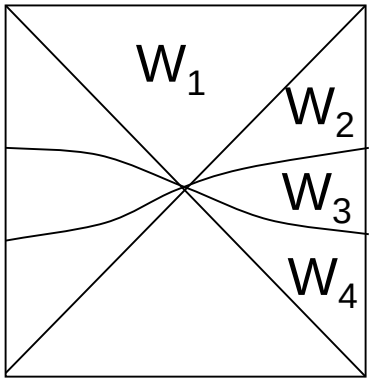
統計モデルの特異点解消の具体例

モデル　$y = a\, s(bx) + cx +$ 雑音

$s(x) = x + x^2$

情報源　$y = 0 +$雑音

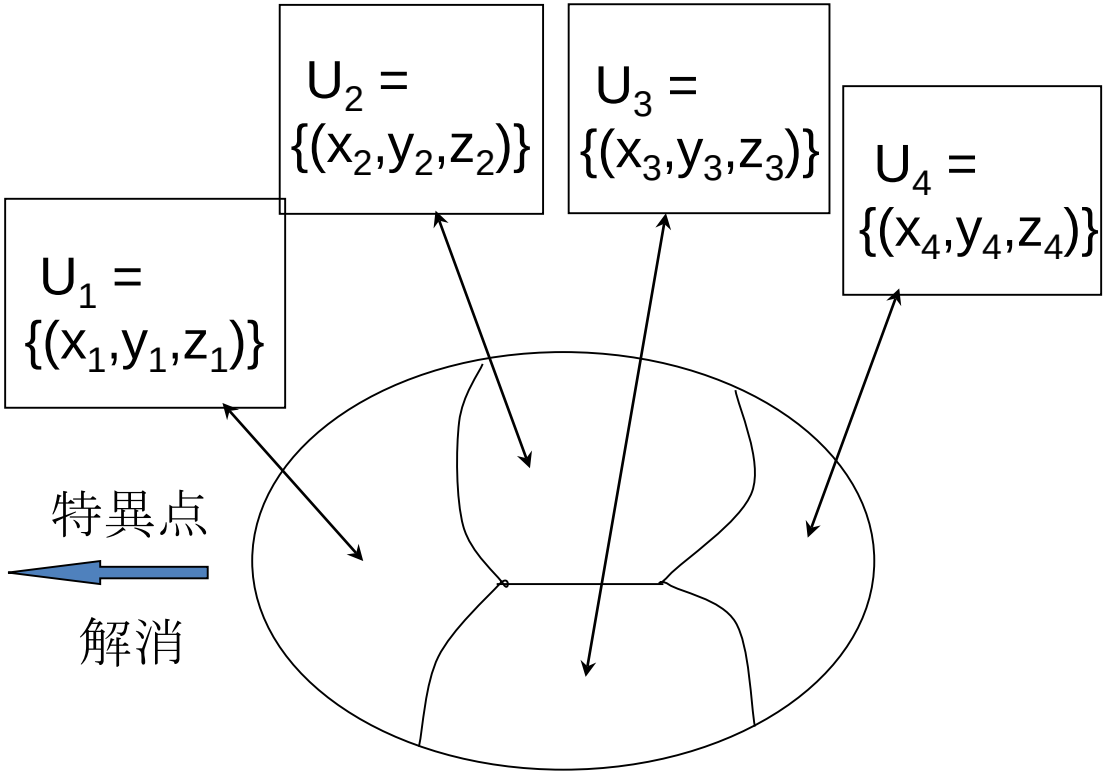
$K(a,b,c) = \{(ab+c)^2 + a^2b^4\}/2$



パラメータ空間 W は
ユークリッド空間

特異点

解消



パラメータ空間 U は4枚の座標系
のはり合わせでできる多様体

局所座標と対数閾値の具体例

$K(a,b,c) = \{ (ab+c)^2 + a^2b^4 \}/2$

$W_1 = \{ |a| \leqq |c| \}$ $\quad a = zx,\ b=y,\ \ c=z$

$W_2 = \{ |a| \geqq |c| ,\ |ab| \leqq |ab+c| \}$ $\quad a = x,\ \ b=yz,\ c = x(1-y)z$

$W_3 = \{ |a| \geqq |c| ,\ |ab+c| \leqq |ab^2| \}$ $\quad a = x,\ \ b=y,\ \ c=xy(yz-1)$

$W_4 = \{ |a| \geqq |c| ,\ |ab^2| \leqq |ab+c| \leqq |ab| \}$ $\quad a = x,\ \ b=yz,\ \ c=xyz(z-1)$

$$K(g(u)) = \begin{cases} z^2 \{ (xy+1)^2 + x^2y^4 \} \\ x^2z^2(1+y^4z^2) \\ x^2y^4(z^2+1) \\ x^2y^2z^4(y^2+1) \end{cases} \qquad |g'(u)| = \begin{cases} |z| \\ |zx| \\ |xy^2| \\ |xyz^2| \end{cases}$$

局所座標の $(\lambda, m) = (1,1), (1,2), (3/4,1), (3/4,1)$

全体の $(\lambda, m) = (3/4,1)$

☆ 「これと同様のことがいつでも必ずできる」ということが特異点解消定理です。
幾つかの統計モデルについては関数 g(u) が具体的に見出されています。

超関数(状態密度関数)の収束

補題.　ある 超関数 D(u) が存在して

$$\frac{n^{\lambda}}{(\log n)^{m-1}}\,\delta(\,t-nu^{2k}\,)u^{h}b(u) \to t^{\lambda-1}D(u)$$

D(u)の台は $g^{-1}(W_0)$に含まれる。

☆　この補題は超関数のメリン変換を用いて初等的に証明できますが
しかし、ここでその説明を始めると戻ってこれなくなる可能性が高いので
ここでは上記の補題を認めて進みましょう。
超関数を定義から理解したい人は次の本を読みましょう。
I.M.ゲルファント, G.E.シーロフ, 超関数入門I,II,　共立出版, 1964.

超関数の収束の具体例

$[0,1]^3$ 上の超関数について次が成立します。

$$\frac{n^{\lambda}}{(\log n)^{m-1}}\,\delta(\,t-nx^4y^6z^8\,)\,x^1y^2z^6 \longrightarrow \frac{1}{24}\,t^{\lambda-1}\,\delta(x)\delta(y)z^2$$

ここで　$\lambda = \min\{\,(1+1)/4,\ (2+1)/6,\ (6+1)/8\,\} = 1/2$

$m = 2$

☆ 「これと同様のことがいつでも必ず成り立つ」ということです。
上記の式は左辺も右辺も超関数であり、$n\to\infty$のとき、超関数の空間で収束するということを意味しています。 このことは計算はメンドウくさいのですが、初等的に証明できます。渡辺澄夫「ベイズ統計の理論と方法」コロナ社のP.108の定理8にも記載してあります。

揺らぎ関数の分解

仮定「相対的に有限な分散」 より

$$K(w) = E_x[\, f(X,w)\,] \geqq \varepsilon\, E_x[\, f(X,w)^2\,]$$

$K(g(u))=u^{2k}$ から、ある $a(x,u)$ が存在して

$$f(x,g(u)) = a(x,u)\, u^k$$

分解された関数の挙動

$$nK_n(g(u)) = \Sigma\, u^k\, a(X_i,u)$$

$$= nu^{2k} - n^{1/2}u^k \quad n^{-1/2}\Sigma\,\{\, u^k - a(X_i,u)\,\}$$

$$\equiv \text{経験過程}\ \xi_n(u)$$

基本形 $$nK_n(w) = nu^{2k} - n^{1/2}u^k\xi_n(u) \qquad ②$$

☆ 関数 $nK_n(w)$ を $n\to\infty$ で零に近づく項 u^k と、確率的に収束する項 ξ_n に分けて表すことができました。このことを用いて学習理論を作ることができます。

経験過程と法則収束

経験過程 $\xi_n(u) = n^{-1/2} \sum_{i=1}^{n} \{ u^k - a(X_i, u) \}$

確率過程に関する中心極限定理

法則収束 $\xi_n(u) \rightarrow \xi(u)$: 正規確率過程

(つまり$\mathcal{F}(\ )$ が関数空間上で有界連続なら $E[\mathcal{F}(\xi_n)] \rightarrow E_\xi[\mathcal{F}(\xi)]$)

☆ 各 u 毎に法則収束「 $\xi_n(u) \rightarrow \xi(u)$ 」することは普通の中心極限定理からすぐに得られることですが、学習理論においてはそれだけでは不十分で、汎関数 $\mathcal{F}$ にについての収束が必要になります。それを可能にするものが経験過程の理論です。経験過程の法則収束は学習理論において重要ですが、関数空間上の確率変数を扱う必要があり、この説明を始めると戻って来れない可能性が高いので、ここでは認めて進みましょう。経験過程を定義から理解したい人は次の本を読みましょう。
Aad W. van deer Vaart, et.al. Weak Convergence and Empirical Processes, Springer, 1996. なお、学習理論ではVC次元が必要になりますので、この本を読まれたかたは多いのではないかと思います。

事後分布の漸近挙動

以上で述べてきたことを統合することにより、事後分布の漸近挙動を次のように導出することができます。ここが学習理論の核心部分です。

$$\exp(-nK_n(w))\,\varphi(w)\,dw$$

$$= \exp(-nu^{2k} + n^{1/2}u^k\xi_n(u))\,\varphi(g(u))|g'(u)|\,du$$

$$= \int dt\,\delta(t-nu^{2k})\,u^h\,b(u)\exp(-t + t^{1/2}\xi_n(u))\,du$$

$$\rightarrow \frac{(\log n)^{m-1}}{n^{\lambda}} \int dt\,t^{\lambda-1}\exp(-t + t^{1/2}\xi(u))\,D(u)du$$

☆　特異点解消定理を用いてパラメータの空間を w から u に移行することにより超関数と経験過程の漸近挙動を、どちらも数学的に扱うことが可能になりました。

事後分布を二つに分けることができた

事後分布が定義する測度

$$\exp(-nK_n(w))\,\varphi(w)\,dw$$

$$= \frac{(\log n)^{m-1}}{n^{\lambda}} \int dt\, t^{\lambda-1} \exp(-t + t^{1/2}\xi(u))\, D(u)du$$

$n \to \infty$で
零になる速さ

確率的に揺らいでいる測度

☆ 事後分布の挙動が解明できたので、後は計算の問題になります。

第3章　自由エネルギーの挙動

事後分布の漸近挙動が得られましたので、自由エネルギーの挙動はそこからすぐに導出できます。

自由エネルギーの漸近挙動の導出

自由エネルギーの式

$$F = nL_n - \log \int \exp(- nK_n(w)) \varphi(w)\, dw$$

に事後分布の漸近挙動

$$\exp(-nK_n(w))\, \varphi(w)\, dw$$

$$= \frac{(\log n)^{m-1}}{n^{\lambda}} \int dt\, t^{\lambda-1} \exp(- t + t^{1/2}\xi(u))\, D(u)du$$

を代入すればよい。∫ は局所座標の和の積分で書けるが、
一番大きなオーダーのところだけ残る。次の定理が得られた。

定理1 : 自由エネルギーの漸近挙動

主定理.1.

f(X,w) が相対的に有限な分散を持つとする。

自由エネルギーの漸近挙動は

$$F = n L_n + \lambda \log n$$

$$-(m-1) \log\log n + O_p(1).$$

主定理1についての説明

主定理1は、自由エネルギーの理論的な挙動を示したものである。真の分布が学習モデルに対して正則であれば、λ＝d/2, m=1であるが、一般にはそうではない。

なお、一般に(λ,m) は真の分布に依存するので，真の分布がわからない場合には、主定理1を直接に使って自由エネルギーを計算することはできない。

なお、統計モデルの評価を行う際に自由エネルギーを数値的に計算する方法はいろいろあるが、真の分布が分かっている際に主定理1から理論値がわかるので、数値計算の正しさを調べることができる。

WBICと自由エネルギー

真の分布がわからない場合に自由エネルギーの近似値を求める方法としては次の方法がある。

$$E_w^{1/\log n}[\quad] = \frac{\int(\quad)\ \Pi\, p(X_i|w)^{1/\log n}\,\varphi(w)\,dw}{\int\ \Pi\, p(X_i|w)^{1/\log n}\,\varphi(w)\,dw}$$

WBIC= $E_w^{1/\log n}[\ -\Sigma \log p(X_i|w)\ ]$ とおくと

$$\mathrm{WBIC} = n L_n + \lambda \log n + O_p((\log n)^{1/2}).$$

第４章　汎化･学習･交差損失の挙動

汎化損失、学習損失、交差損失の間に成り立つ関係についても、
事後分布の漸近挙動から初等的に導出できるのですが、計算は少し
ヤヤコシイです。とは言っても未知数が4つの線形連立方程式を解く
だけなので数学的に難しい点はありません。

４つの連立方程式は、いろいろな場所から集めてくる必要があり、
①②③④という番号がついています。

繰り込まれた事後分布の定義

事後分布は $\exp(-nK_n(w))\,\varphi(w)\,dw$

$= \quad (n\text{の関数}) \times \int dt\, t^{\lambda-1} \exp(- t + t^{1/2}\xi(u))\, D(u)du$

定義：繰り込まれた事後分布による平均 $< \quad >$

$$= \frac{\int dt \int du D(u)(\quad) t^{\lambda-1}\exp(-t+t^{1/2}\xi(u))}{\int dt \int du D(u)\, t^{\lambda-1}\exp(-t+t^{1/2}\xi(u))}$$

補題 $< t > = \lambda+(1/2)<t^{1/2}\xi(u) >$ ③

(分子を t で部分積分すると得られる)

スケーリング関係の導出

関数 f(x,w) を w=g(u) を用いて変換したものは

$$f(x,g(u)) = a(x,u)\, u^k = a(x,u)\, (t/n)^{1/2}$$

$$u^{2k} = t/n$$

補題　任意の $s \geqq 0$ で、次の法則収束が成立

$$n^{s/2}\, E_w [\ f(x,w)^s\ ] \rightarrow < t^{s/2}\, a(x,u)^s >$$

$$n^s\, E_w [\ K(w)^s\ ] \rightarrow < t^s >$$

事後分布による f(x,w) の平均は繰り込まれた事後分布による平均で表すことができる。

特異ゆらぎの定義

定義. 特異ゆらぎを繰り込まれた分布で定義する

$$2\nu = E_\xi E_x[\langle t\, a(X,u)^2 \rangle - \langle t^{1/2} a(X,u) \rangle^2] \qquad ④$$

定義. 汎関数分散 を事後分布で定義する

$$V = \Sigma \{ E_w[(\log p(X_i|w))^2] - E_w[\log p(X_i|w)]^2 \}$$

記号　$f(x,w) = \log(p_0(x)/p(x|w))$

スケール関係　$f(x,g(u)) = a(x,u)\,(t/n)^{1/2}$

から次の補題が示される。

特異ゆらぎと汎関数分散の関係

補題　f(X,w) が相対的に有限な分散を持つとき、

$$\lim_{n \to \infty} E[\, V \,] = 2\nu$$

☆　真の分布が統計モデルで実現可能であり、かつ正則であれば $\nu = d/2$ です。
正則であって実現可能でない場合には　$\nu = \mathrm{tr}(IJ^{-1})/2$　になります。ここで I は
フィッシャー情報行列で J はLのヘシアンです。正則でない場合には ν の値はまだ
分かっておりません。有理数にならない場合もあると思います。

定理1：自由エネルギーの漸近挙動

主定理.2.

f(X,w) が相対的に有限な分散を持つとき

$$E[\,G_n\,] = L_0 + \lambda / n + o(1/n),$$

$$E[\,C_n\,] = L_0 + \lambda / n + o(1/n),$$

$$E[\,T_n\,] = L_0 + (\lambda - 2\nu) / n + o(1/n),$$

$$E[\,V\,] = 2\nu + o(1).$$

①②③④を組みあわせると証明できる。

☆　f(x,w) が相対的に有限な分散を持たない場合には、上記の定理が成り立たない場合があります。しかし、その場合でも $E[G_n]=E[T_n]+E[V]/n+o(1)$ は成立していることが多いです。もっと深い数学的構造があるものと予想されます。

(証明1)汎化損失の解析

$$G(\alpha) \equiv \alpha L_0 - E_x[\ \log E_w[\ \exp(-\alpha f(X,w))\ ]\ ]$$

と定義すると

$$G_n = G(1) = G(0) + G'(0) + G''(0)/2 + O_p(1/n^{3/2})$$

それぞれの平均値が計算できる。④を使うと

$$E[G(0)]=0$$
$$E[G'(0)]=L_0+EE_xE_w[f(X,w)]=L_0+E[E_w[K(w)]]$$
$$= L_0+ E[<t>/n]$$
$$E[G''(0)]= -EE_x\{ E_w[f(x,w)^2] - E_w[f(x,w)]^2 \} = -2\nu/n$$

従って

$$E[\ G_n\ ] = L_0+E[<t>]/n - \nu/n + O(1/n^{3/2})$$

(証明2)学習損失と交差損失

$$T(\alpha)= \alpha L_n - (1/n)\,\Sigma \log E_w[\exp(-\alpha f(X_i,w))]]$$

と定義すると

$$T_n = T(1) = T(0) + T'(0) + T''(0)/2 + O_p(1/n^{3/2})$$
$$C_n = -T(-1) = -T(0) + T'(0) - T''(0)/2 + O_p(1/n^{3/2})$$

それぞれの平均値が計算できる。②を用いて

$$E[T(0)]=0$$
$$E[T'(0)]=L_0+EE_w[(1/n)\Sigma f(X_i,w)]=L_0+EE_w[K_n(w)]$$
$$=L_0+ E[<t/n-t^{1/2}\xi(u)/n>]$$
$$E[T''(0)] = -E[(1/n)\Sigma\{ E_w[f(X_i,w)^2] - E_w[f(X_i,w)]^2 \}] = -2\nu/n$$

従って

$$E[T_n]= L_0+ E[<t/n-t^{1/2}\xi(u)/n>] - \nu/n + O(1/n^{3/2})$$
$$E[C_n]= L_0+ E[<t/n- t^{1/2}\xi(u)/n>] + \nu/n + O(1/n^{3/2})$$

（証明3）汎化損失と交差損失

①から $E[G_{n-1}]=E[C_n]$ が成り立つので

$$L_0+E<t>/n -\nu/n = L_0 + E<t/n-t^{1/2}\xi(u)/n> + \nu/n + O(1/n^{3/2})$$

これより　$E<t^{1/2}\xi(u)/n>=2\nu/n + O(1/n^{3/2})$　◎

③から　$E<t>/n = \lambda/n + E<t^{1/2}\xi(u) >/(2n)=(\lambda+\nu)/n+ O(1/n^{3/2})$　◎

この◎を$E[G_n],E[T_n],E[C_n]$に代入すると定理が得られる。

（証明終わり）

情報量規準への応用

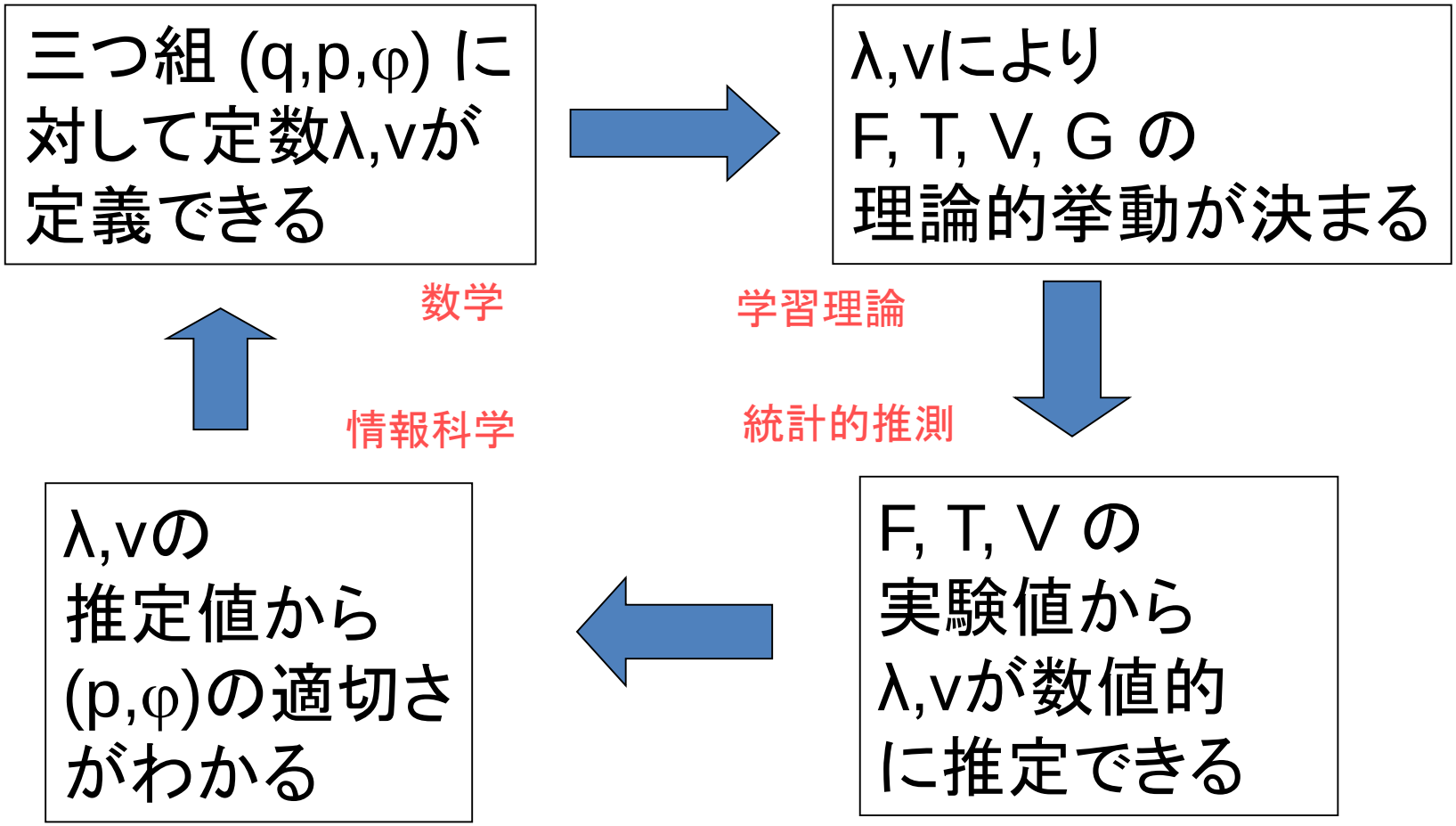
系.　WAIC = T+V/n　とおくと

$$E[G] = E[W] + O(1/n^2)$$

任意の組 $(q(x),p(x|w),\varphi(w))$ で成立。

☆　WAIC の統計学への応用については世界中で論じられるようになりました。広く研究されるようになったのはこのファイルの著者の努力によるものではなく、多くの統計学者の先生のお力によるものです。検索サイトやGithubに行って「WAIC criterion」で探してみてください。このファイルの著者が書いたものよりもずっと理解しやすいと評判になっている解説も幾つかあります。

まとめ


三つ組 (q,p,φ) に対して定数λ,νが定義できる
λ,νにより F, T, V, G の理論的挙動が決まる
学習理論
統計的推測
F, T, V の実験値からλ,νが数値的に推定できる
λ,νの推定値から(p,φ)の適切さがわかる
情報科学
数学

なぜ学習理論が作れたか

座標変換 w=g(u) をうまく選ぶことにより

事後分布の積分要素　$\exp(-nK_n(w))\ \varphi(w)\ dw$

＝　(nの関数)×（nに依存しない積分要素）

とすることができた（繰り込み可能にできた）から。

☆　確率モデルが区分的に解析的であれば、ここで述べたことと同じ方法が使えて事後分布が繰り込み可能であることを導出することができます。統計学上で重要な問題で事後分布が繰り込み可能でないケースを見つけた場合には論文を書いて発表してください。

☆　ここでは自由エネルギーと汎化損失の挙動を導出しましたが、統計学上で解析したい確率変数は他にもあると思います。それらの確率変数の挙動が導出できた場合にも、論文を書いて発表してください。