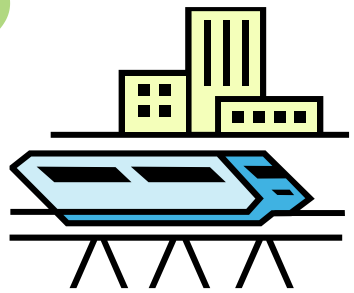
ベイズ統計入門　⑨

目標　　一般理論への準備

東京工業大学
渡辺澄夫

旅の地図

(1) ベイズ統計の定義
(2) 密度と条件つき密度
(3) 混合正規分布＋ギブスサンプラー
(4) 神経回路網＋ランジュバン方程式
(5) 真とモデルの関係
(6) 正則モデルの漸近理論
(7) AIC と BIC
(8) ハイパーパラメータ最適化
(9) 一般モデルの漸近理論
(10) 一般モデルの漸近理論
(11) 一般モデルの選択
(12) 条件つき独立 高次元
(13) 階層ベイズ
(14) 相転移
(15) まとめ



1　復習

記号

w_0 : $K(q(x)||p(x|w))$ を最小にする w はひとつではないが

$p_0(x) = p(x|w_0)$ は w_0 の選び方に依存しない関数

$f(x,w) = \log(p_0(x)/p(x|w))$ 対数密度比関数

$K(w) = E_x[f(X,w)] \geqq 0$ 平均対数尤度比関数

$K_n(w) = (1/n) \Sigma_{i=1}^{n} f(X_i,w)$ 経験対数尤度比関数

自由エネルギーと汎化損失

自由エネルギーは $L_n = -(1/n)\Sigma_{i=1}^{n} \log p_0(X_i)$ とおくと

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

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

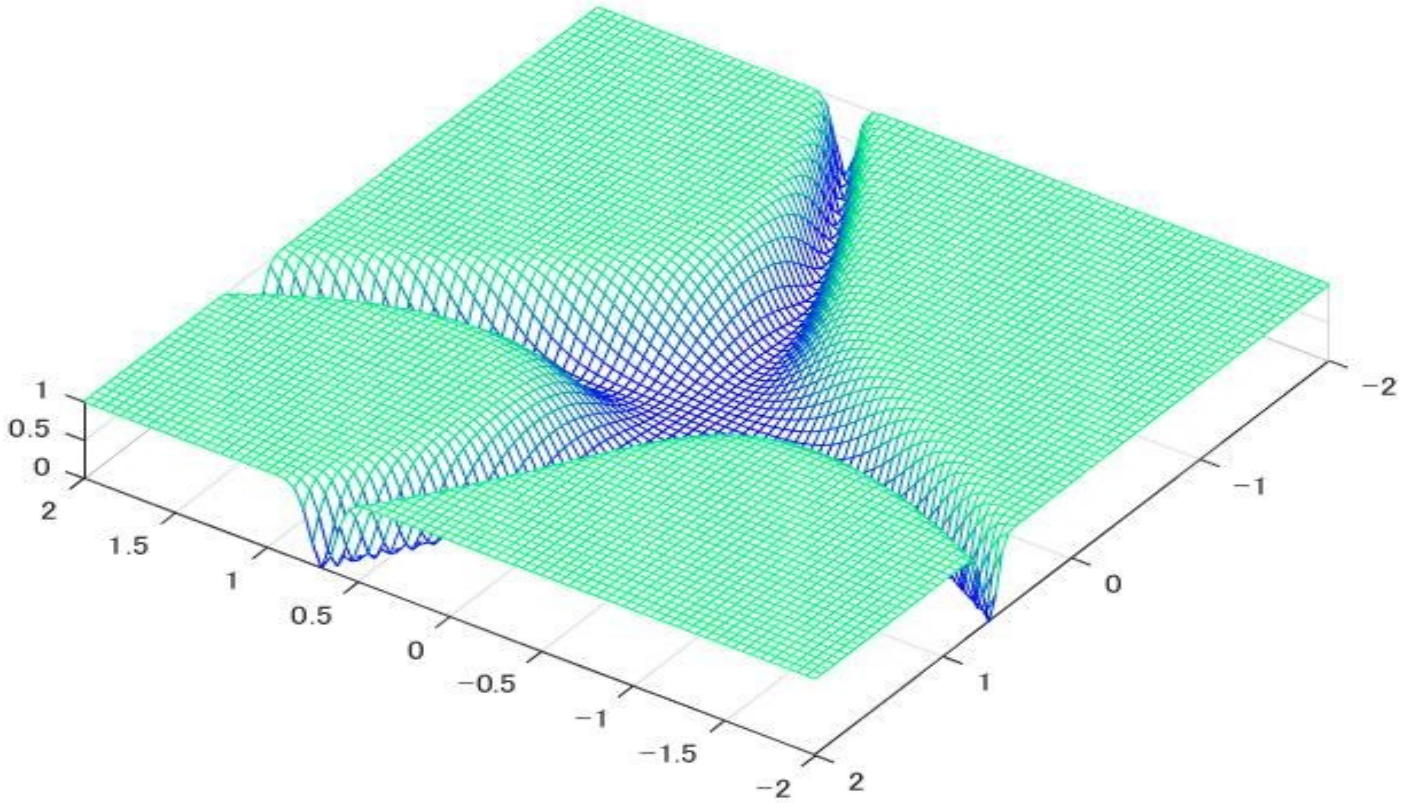
学習損失　$T_n = L_n - (1/n)\Sigma \log E_w[\exp(-f(X_i,w))]$

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

WAIC　$W_n = T_n + (1/n)\Sigma V_w[f(X_i,w)]$

目標

関数　$K(w) = K(q(x)||p(x|w_0))$　の零点集合が代数的集合あるいは解析的集合であるときの理論を作りたい。



２　準備：特異点解消定理

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

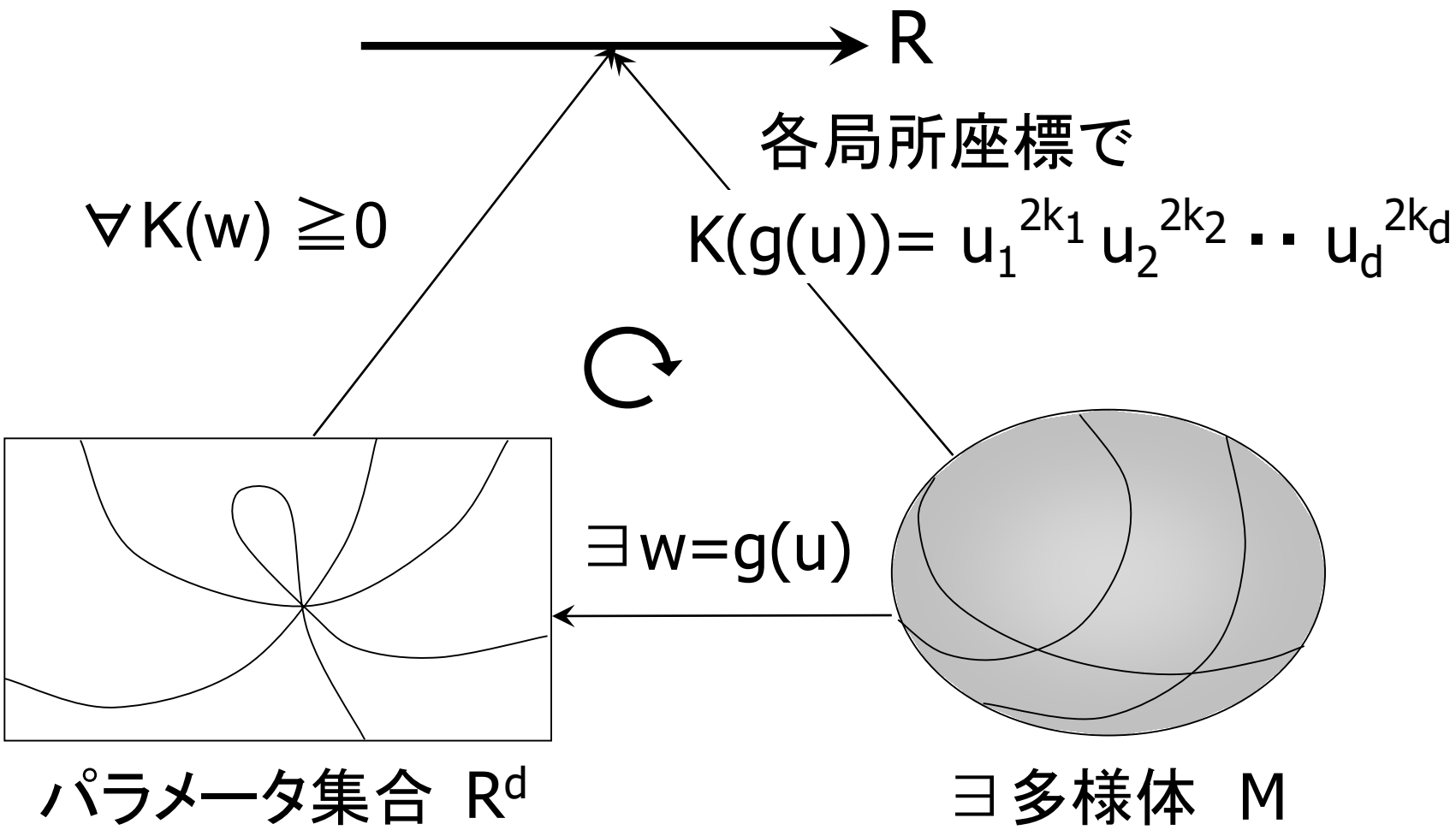
相対的に有限な分散を持つ ⇒ 揺らぎが平均でバウンドできる

事後分布の挙動を調べるために次の二つを考えます。

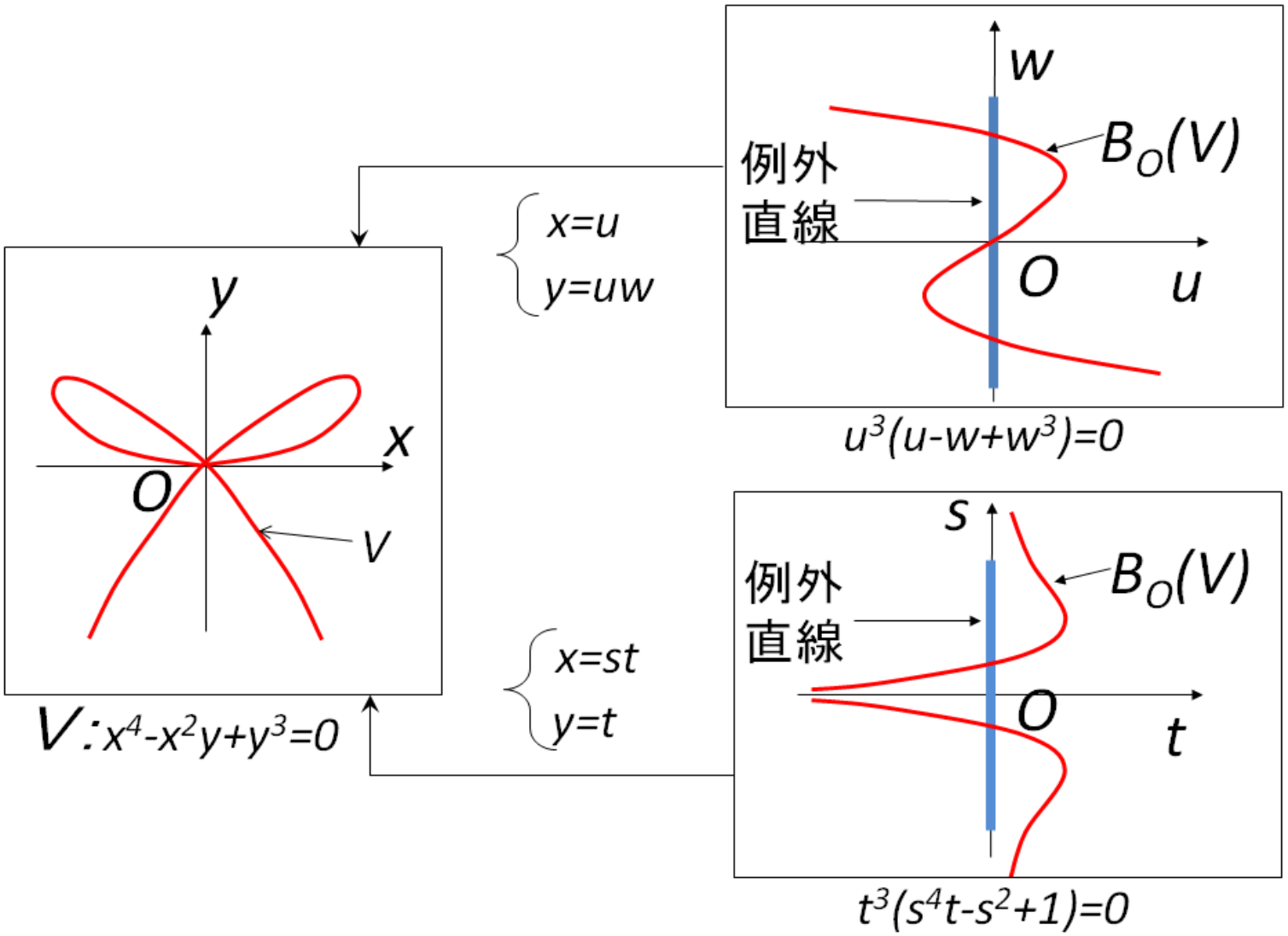
(1) $n \to \infty$ のときの $\exp(-n K(w))$ の挙動
(2) $n \to \infty$ ののときの揺らぎ関数の挙動

準備1:特異点解消定理

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。



特異点解消の例


w
例外
直線
$B_O(V)$
O
u
$u^3(u-w+w^3)=0$
x=u
y=uw
y
x
O
V
$V: x^4-x^2y+y^3=0$
x=st
y=t
s
例外
直線
$B_O(V)$
O
t
$t^3(s^4t-s^2+1)=0$

事後分布の局所性

コンパクトサポート関数の有限和（1の分割） $\varphi(w) = \Sigma_k \varphi_k(w)$ を使うと、任意の関数 $\psi(w)$ について

$$\int \psi(w)\, p(w|X^n)\, dw = (1/Z) \Sigma_k \int \psi(w)\, \varphi_k(w) \prod_{i=1}^{n} p(X_i|w)\, dw$$

パラメータ上の積分は、有限個の局所的な積分の和で書ける。局所的な積分は、局所座標ごとに原点を選びなおすことにより次の式で書ける（ということが特異点解消定理）。

$$K(g(u)) = u_1^{2k_1} u_2^{2k_2} \cdots u_d^{2k_d}$$

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

統計学の理論を作るとき、常にこのような座標表現が取れることを使うことができるので、自由エネルギーや汎化誤差の漸近挙動を導出できます。

（注意） ベイズ入門では、特異点解消定理についてこれ以上のことは必要になりません。

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、統計学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。

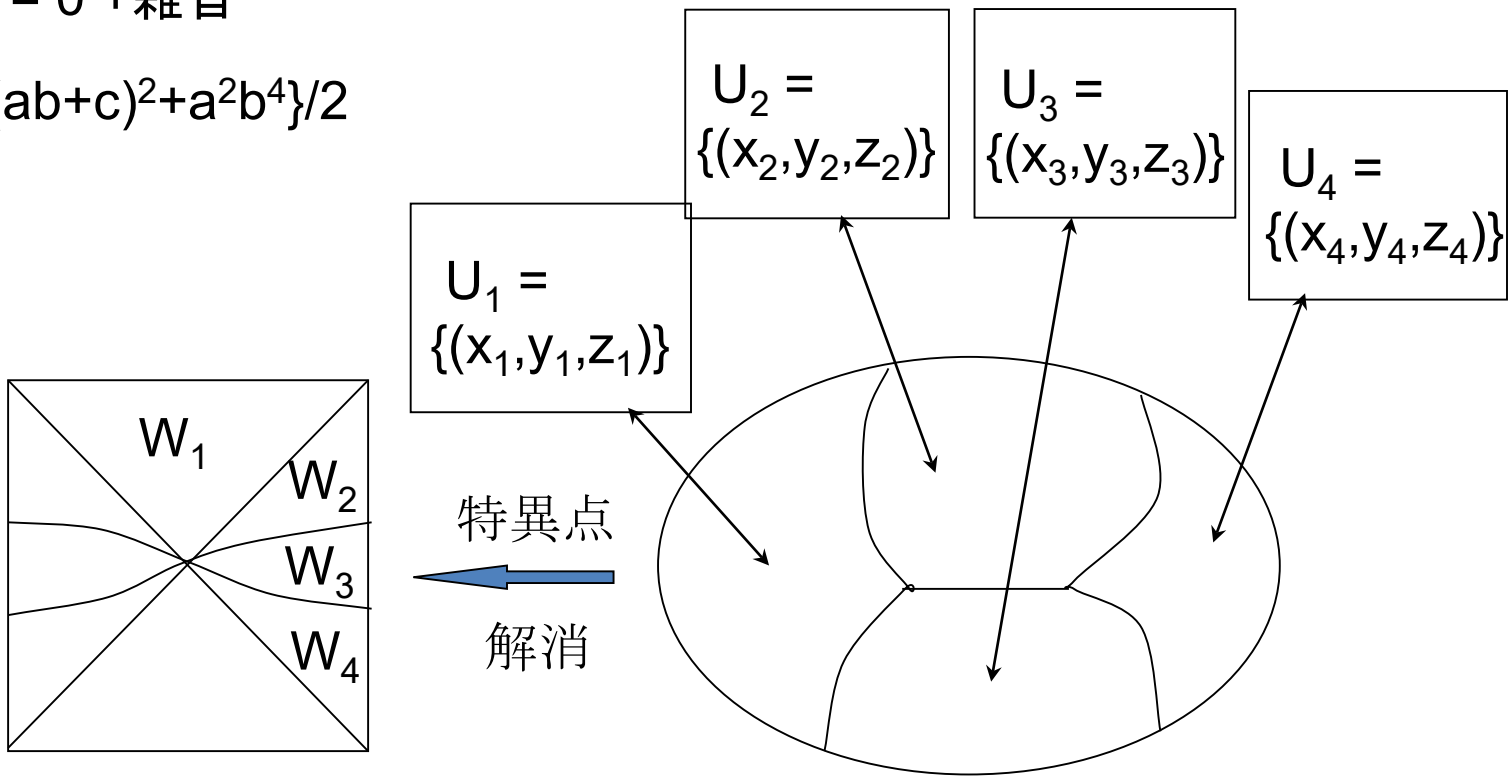
統計モデルの特異点解消の具体例

統計モデル　$y = a\, s(bx) + cx +$ 雑音

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

真の分布　$y = 0 +$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

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。

３　準備：超関数の漸近挙動

準備２：超関数と漸近挙動

超関数とは 「関数 $\psi(w)$ から実数への連続線形汎関数 $F(\psi)$ 」のこと

（連続性を定義するために関数空間に位相を入れておく必要がありますが
ベイズ入門では、そのための基礎は扱わないことにします）。

事後分布は超関数 「$\psi(w) \rightarrow \int \psi(w)\, p(w|X^n)\, dw$ 」 です。

☆ n が大きくなると事後分布は集合 $\{w; K(w)=0\}$ の上だけで零でない
超関数 $\delta(t-K(w))$ $(t=0)$ に近づいていくと考えられますが、
$\delta(t- K(w))$ は $t=0$ のとき超関数として well-defined ではありません。
そこで $t \rightarrow +0$における超関数 $\delta(t- K(w))$ の発散の挙動を考えましょう。

特異点解消定理を用いれば、必ず $K(g(u))=u^{2k}$ とできますので、
その場合だけを扱うことができれば十分です。

定義：超関数が零にならない集合の閉包をその超関数の台（サポート）
と呼びます。

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

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

n に依存する超関数　→　n に依存しない超関数

ここで D(u)の台は $g^{-1}(W_0)$に含まれる。

☆　上記の式は「左辺も右辺も超関数であり、n→∞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。この補題は計算はメンドウですがメリン変換を用いて初等的に証明できます。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。次の本に詳しい記載があります。
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　$\lambda = \min\{\ (1+1)/4,\ (2+1)/6,\ (6+1)/8\ \} = 1/2$

$m = 2$

上記は次のことと同じことを言っています：任意の関数 $\psi(x,y,z)$ について

$$\frac{n^{\lambda}}{(\log n)^{m-1}} \int_{[0,1]^3} \psi(x,y,z)\ \delta(\ t-\ nx^4y^6z^8\ )\ x^1y^2z^6\ dxdydz \longrightarrow \frac{1}{24}\ t^{\lambda-1} \int \psi(0,0,z)\ z^2\ dz$$

☆　超関数 $\delta(\ t-\ nx^4y^6z^8\ )\ x^1y^2z^6$ は 集合 $\{(x,y,z);\ xyz=0\}$ に含まれるサポートを持つ超関数に収束することがわかりました。収束先の超関数のサポートは $\{(x,y,z);x=y=0\}$ なので、集合 $\{(x,y,z);\ xyz=0\}$ に含まれますが等しくはないことに注意してください。

４　準備：経験過程

準備3： 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

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

☆ 一変数の場合、因数定理 「f(x) が f(a)=0 を満たせば f(x) は (x-a) で割り切れる」がなりたちます。しかし
多変数では「f(x,y) が f(a,b)=0を満たせば f(x,y)は (x-a)(y-b)で割り切れる」は成り立ちません。一方、$K(g(u))=u^{2k}$ は各変数 $u_1, u_2, \ldots, u_d$ ごとに一変数の因数定理を適用できる形をしているので、上記のような割り算ができます。もとのパラメータ w のままで考えると上記のような変形はできないことに注意してください。

経験対数尤度比関数の分解

どんな経験対数尤度比関数も w=g(u) をうまく選ぶことにより

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

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

$$\equiv \text{経験過程}\ \xi_n(u)$$

経験対数尤度比の基本形　$nK_n(g(u)) = nu^{2k} - n^{1/2}\, u^k\, \xi_n(u)$　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。このことを用いて統計学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、統計学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統計学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、学習の数理を研究したい場合、この本を読む必要が高い人が多いのではないかと思います。

経験過程の具体例

実数 u と標準正規分布に従う確率変数 $\{X_i\}$ があったとき

$$\xi_n(u) = n^{-1/2} \sum_{i=1}^{n} \sin(u X_i)$$

は、$n \to \infty$のとき、各 u ごとに正規分布に法則収束しますが、関数としては正規確率過程に分布収束します。