

# 代数幾何と学習理論 入門と新展開

2010年6月14日

情報論的学習理論と機械学習研究会講演資料  
解説文つき

渡辺 澄夫  
東京工業大学

# もくじ

- (1) 学習理論とは
- (2) 主定理
- (3) 対数閾値
- (4) 特異ゆらぎ
- (5) 現在の展開

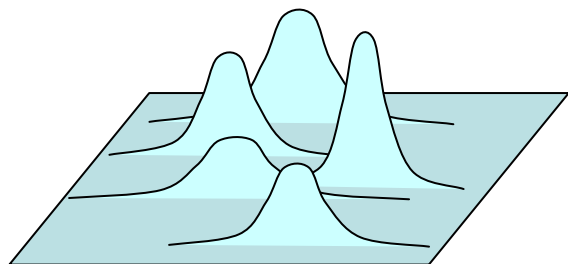
(解説) 一般の学習モデルを考えると、数学的にいちばん自然な方法を述べます。自然な方法は、強力な方法でもあります。

# 1

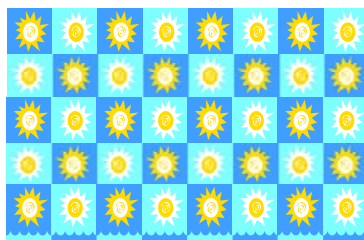
# 学習理論とは

(解説) 第1章では学習理論の枠組みを説明します。

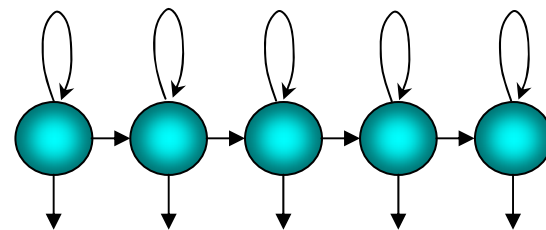
# 人工知能・情報工学・生物学で使われるモデル



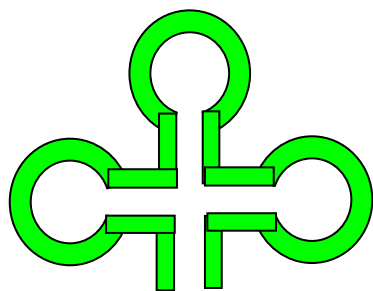
混合正規分布



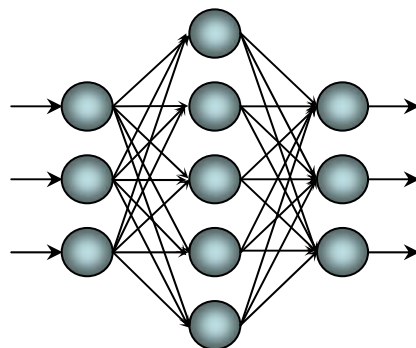
混合ベルヌーイ



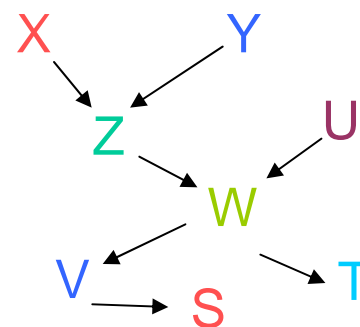
隠れマルコフ



確率文法

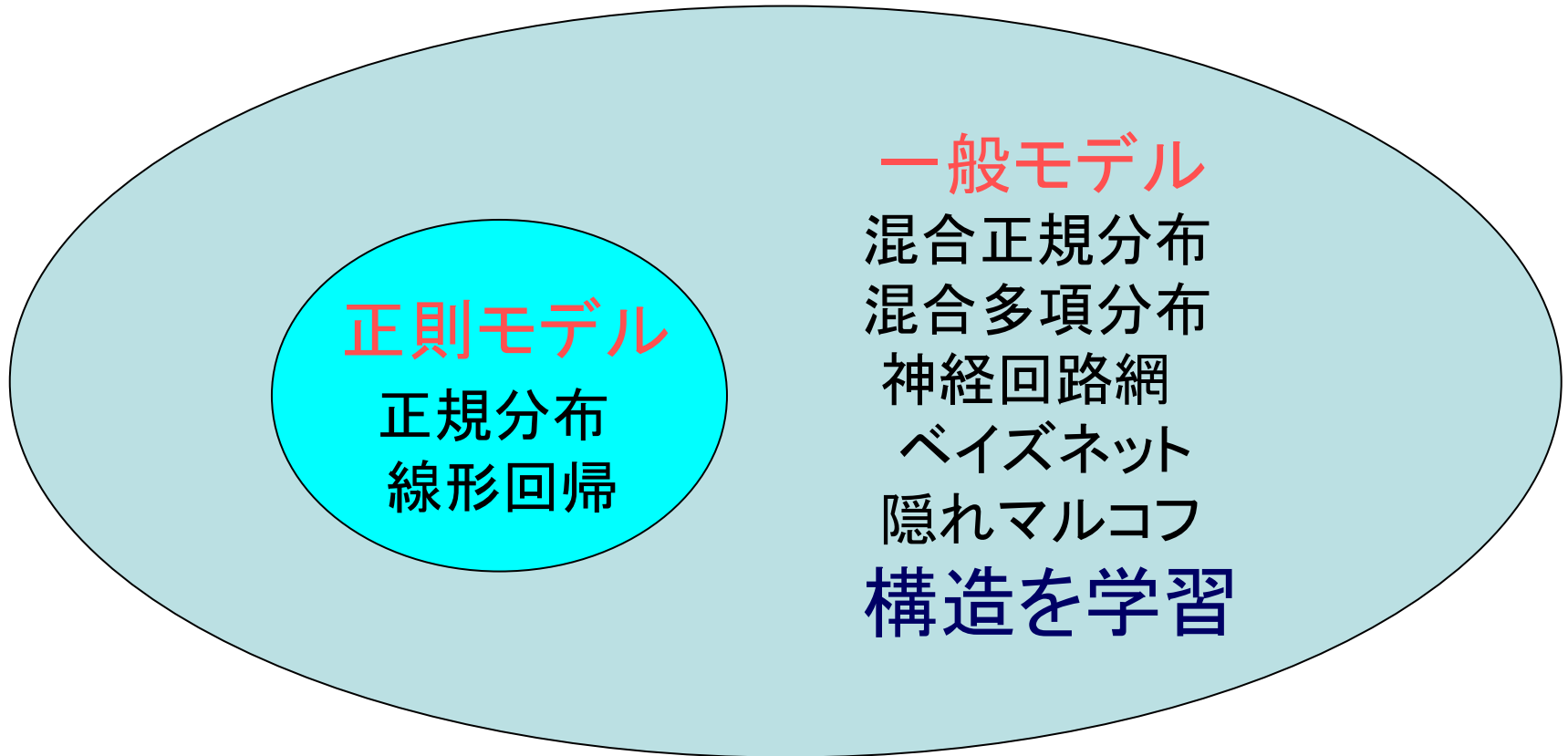


ニューラルネット



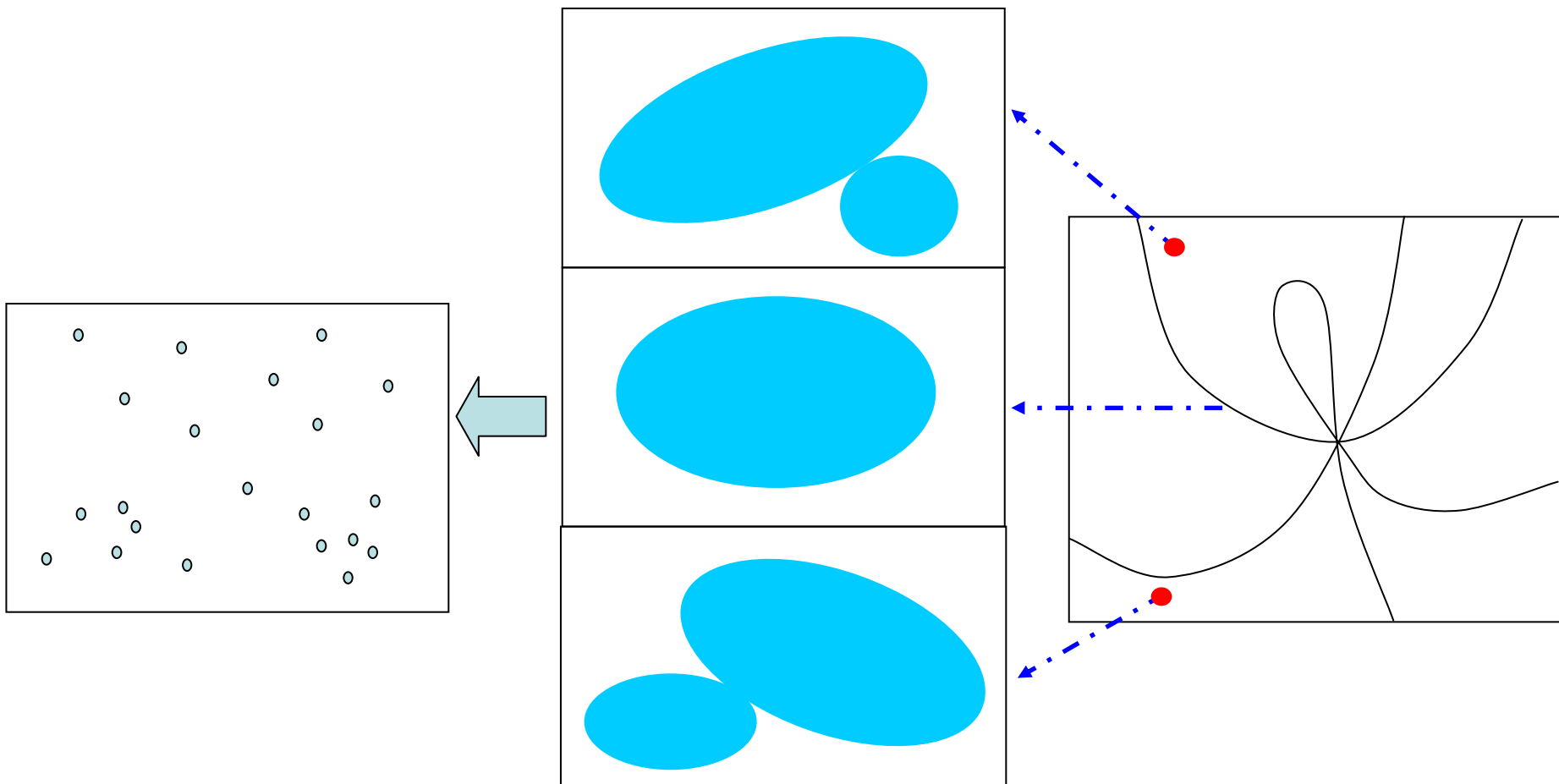
ベイズネット

# 正則モデルと一般モデル



(解説) データの背後にある構造を求める学習モデルは統計的に正則ではないので、普通の統計学は成立しません。AIC, BIC, MDLも成り立ちません。

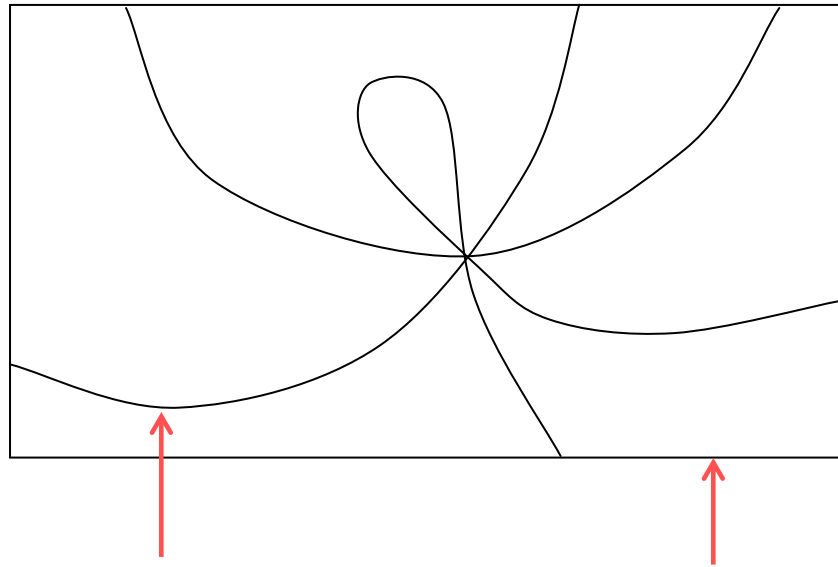
# 「例から構造を知る」



(解説) データから構造を知ることは、代数幾何学を考えると等価です。

# 一般モデルの特徴

モデル1  $\subset$  モデル2  $\subset$  モデル3  $\subset$  ...



モデル (k)  $\subset$  モデル(k+1)

(解説) 複雑なモデルの中で、シンプルなモデルは代数多様体になっています。

# 記号

$$x \in \mathbb{R}^N, \quad w \in W(\text{コンパクト}) \subset \mathbb{R}^d$$

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

(2) 学習モデル  $p(x|w)$

(3) 事前分布  $\varphi(w)$

(解説) 三組【情報源、学習モデル、事前分布】は、学習理論の基本要素です。

# 事後分布

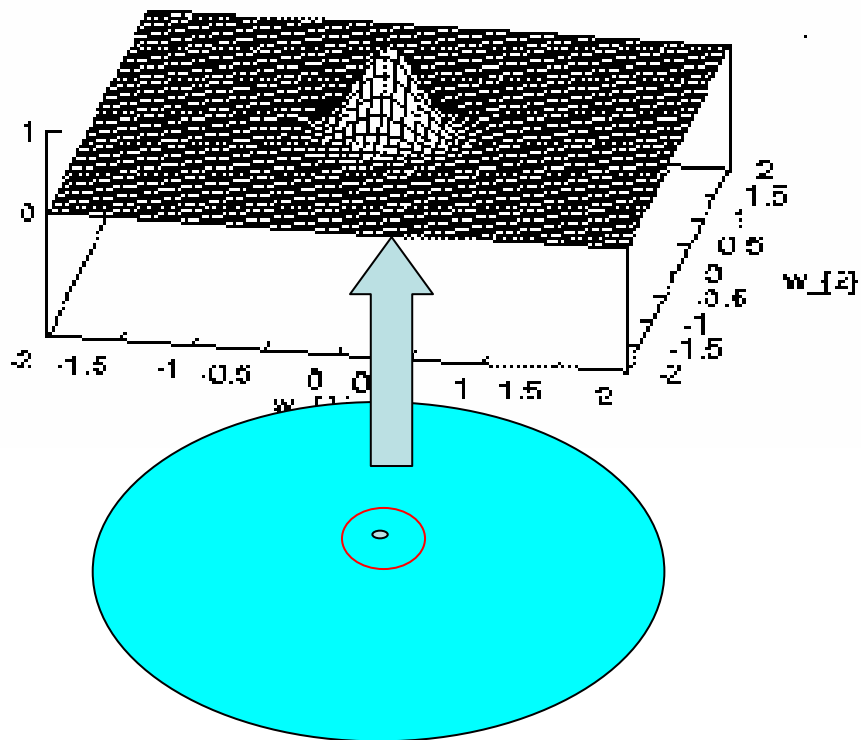
$0 < \beta < \infty$ , 普通は  $\beta = 1$

$$E_w[ \quad ] = \frac{\int ( \quad ) \prod_{i=1}^n p(X_i|w)^\beta \varphi(w) dw}{\int \prod_{i=1}^n p(X_i|w)^\beta \varphi(w) dw}$$

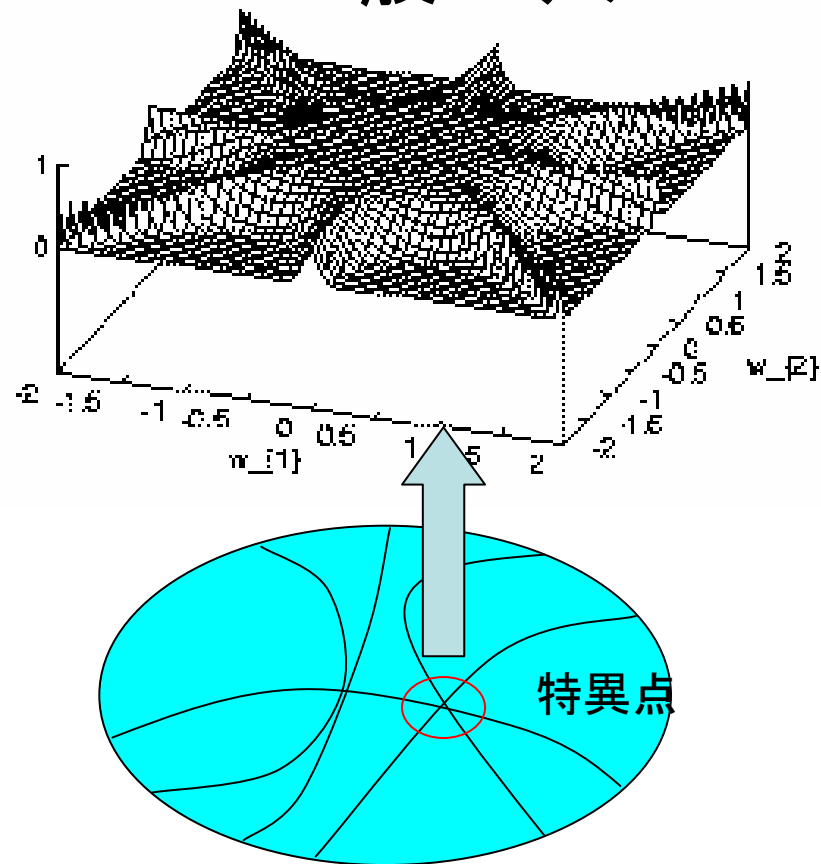
(解説)事後分布による平均操作をこの式で定義します。この式によって計算される平均値はサンプルの関数ですから確率変数です。

# 事後分布の様子

## 正則モデル



## 一般モデル



(解説) 一般のモデルでは事後分布は正規分布では近似できません。

# 2

## 主定理

(解説) 第2章では主定理を述べます。  
定数である $\lambda$ と $\nu$ の定義は3章4章で説明します。

# 記号

対数損失関数

$$L(w) = - E_x[ \log p(X|w) ]$$

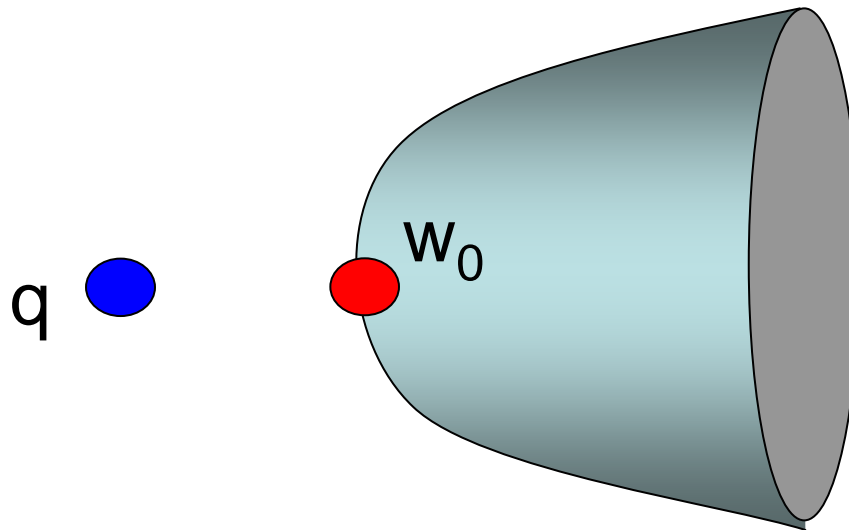
最適パラメータの集合 (= 解析的集合)

$$W_0 = \{ w \in W ; L(w) \text{ 最小} \}$$

(解説)関数  $L(w)$ は、情報理論における符号長、統計学における対数損失です。

# 定義 1

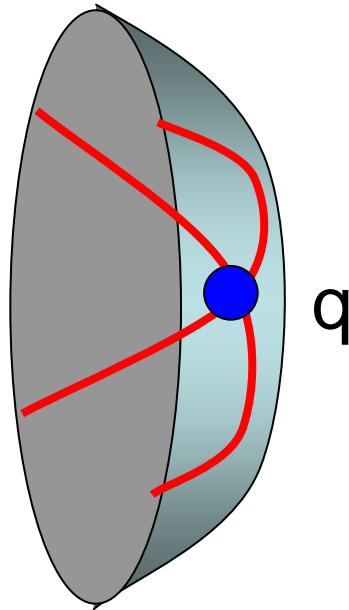
$W_0$  の要素がひとつ  $w_0$  だけであり、  
かつヘッセ行列  $\nabla^2 L(w_0)$  が正則であるとき、  
 $q(x)$  は  $p(x|w)$  に対して**正則**であるという。



(解説) 正則のときは正規分布の計算で計算できるので  
数学的には困難はありません。

## 定義 2

条件  $q(x) = p(x|w_0)$  を満たす  $w_0$  があるとき、 $q(x)$  は  $p(x|w)$  によって**実現可能**であるという。



(解説) 実現可能であっても正則でないことはしばしば起こります。  
正則でない場合に生じる現象は、これまでは解明されていませんでした。

## 定義 3

正則または実現可能のとき、ある  $w_0 \in W_0$  を用いて  $p_0(x) \equiv p(x|w_0)$  と書く。

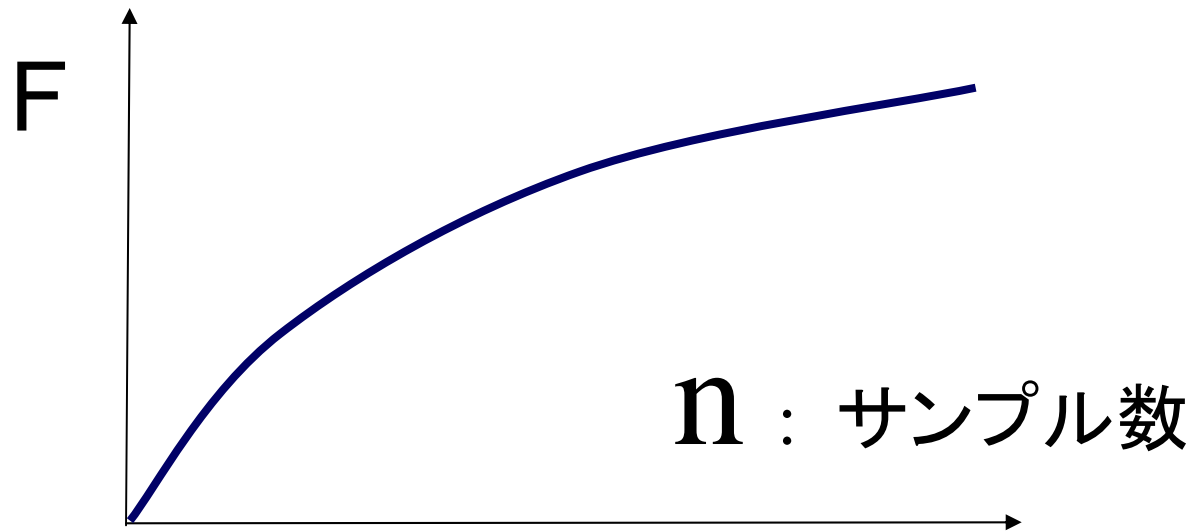
$$L_0 = - E_x[\log p_0(X)]$$

$$L_n = - (1/n) \sum_{i=1}^n \log p_0(X_i)$$

(解説) 情報源がモデルに含まれるとき、 $L_0$  は情報源のエントロピーであり、 $L_n$  は経験エントロピーです。一般に、 $L_0$  は  $L(w)$  の最小値です。

# 確率的複雑さ

$$F = -\log \int \prod_{i=1}^n p(X_i|w)^\beta \varphi(w) dw$$



(解説) 確率変数  $F$  はベイズ符号長であり、また、データに対する  $(p, \phi)$  の対数尤度でもあります。統計力学では自由エネルギーと呼ばれます。この量は情報理論・統計学・統計力学において第一義的に重要です。

## 主定理.1.

正則または実現可能であるとする。

$(q,p,\varphi)$  から決まる有理数  $\lambda$  と自然数  $m$  が存在して

確率的複雑さの漸近挙動は

$$F = n^\beta L_n + \lambda \log n \\ - (m-1) \log \log n + O_p(1).$$

(解説) 確率変数  $F$  の漸近挙動が解明されました。有理数  $\lambda$  が何であるかは以下のページで説明しています。一般に  $\lambda$  は  $(d/2)$  とは異なります。

# 汎化誤差と学習誤差

予測分布  $p^*(x) = E_w[ p(x|w) ]$

汎化損失  $G = - E_x[ \log p^*(X) ]$

学習損失  $T = -(1/n) \sum_{i=1}^n \log p^*(X_i)$

(解説) ベイズ学習とは、「予測分布はきっと情報源に近いだろう」と推測する学習法です。汎化損失が小さいほど推測精度は優れています。汎化損失は情報源を知らないと計算できません。学習損失はデータだけで計算できます。

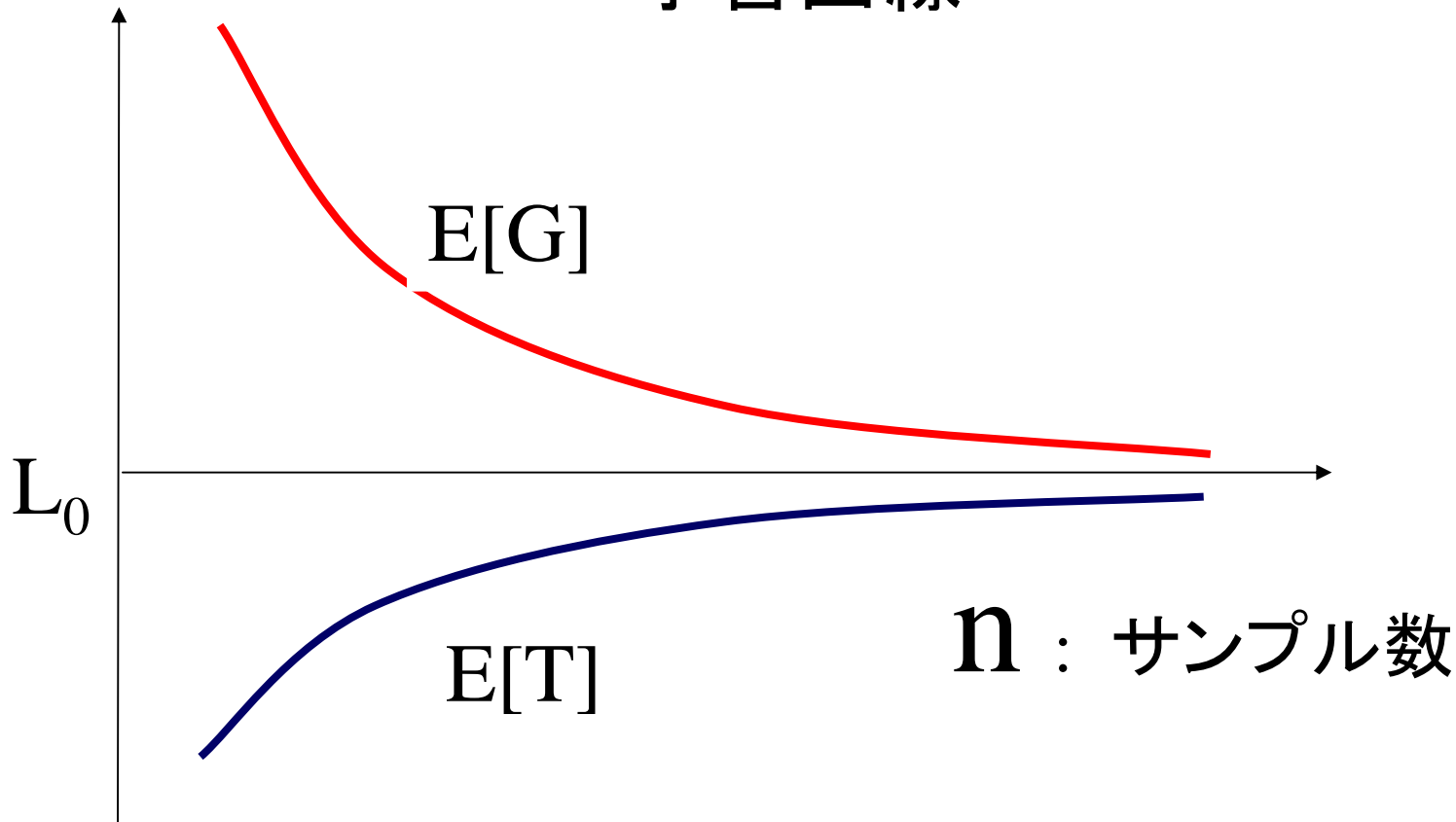
# 汎関数分散

## 定義. 汎関数分散

$$V = \sum_{i=1}^n \{ E_w[ (\log p(X_i|w) )^2] - E_w[ \log p(X_i|w) ]^2 \}$$

(解説) 汎関数分散は、代数幾何と学習理論の研究の成果として得られた量です。この量の重要性は、以下のページで説明されています。この量は、データさえあれば、情報源を知らなくても計算できます。実際の応用でも容易に計算できます。

# 学習曲線



(解説) 汎化損失と学習損失がサンプル数のどんな関数であるかが問題です。  
その関数のことを学習曲線といいます。

## 主定理.2.

正則または実現可能であるとする。

$(q, p, \varphi)$  から決まる定数  $\lambda$  と  $\nu$  が存在して

$$E[ G ] = L_0 + \{ (\lambda - \nu) / \beta + \nu \} / n + o(1/n),$$

$$E[ T ] = L_0 + \{ (\lambda - \nu) / \beta - \nu \} / n + o(1/n),$$

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

(解説) 学習曲線が解明されました。定数  $\lambda$  は主定理1の  $\lambda$  と同じです。  
定数  $\nu$  は特異ゆらぎと呼ばれるもので、以下のページに説明があります。

# 学習の状態方程式

系.  $WAIC = T + \beta V/n$  (広く使える規準)

$$E[G] = E[WAIC] + o(1/n)$$

$\beta = 1$  のとき

$$(G - L_0) + (WAIC - L_n) = 2\lambda/n + o_p(1/n)$$

(解説) WAICは情報源がわからなくてもデータだけから計算できます。WAICは、情報源がモデルに含まれなくも、正則でなくも、汎化損失と同じ平均を持ちますが、各サンプル毎の挙動は、汎化誤差と相反します。

## 正則な場合

$q(x)$  が  $p(x|w)$  に対して正則なら

$$\lambda = d/2$$

$$v = (1/2) \text{tr}( IJ^{-1} )$$

$$\left\{ \begin{array}{l} I = E_x[ \nabla \log p(x|w_0) \nabla \log p(x|w_0) ] \\ J = - \nabla^2 L(w_0) \end{array} \right.$$

確率収束  $V \rightarrow 2v/\beta$  が成立

(解説) 情報源がモデルに対して正則であるとき、WAICは漸近的にTICと等価です。情報源が正則かつ実現可能であるときはWAICは漸近的にAICと等価です。TIC では $J^{-1}$ の計算が揺らぎますが、WAICは、尤度関数の事後分布における分散なので、そのような揺れはありません。

# 3

## 対数関数

第3章では代数幾何の基本定理を導入し  
定数  $\lambda$  が代数幾何学において重要な量であることを示します。

# 定義

対数尤度比  $f(x, w) = \log(p_0(x)/p(x|w))$

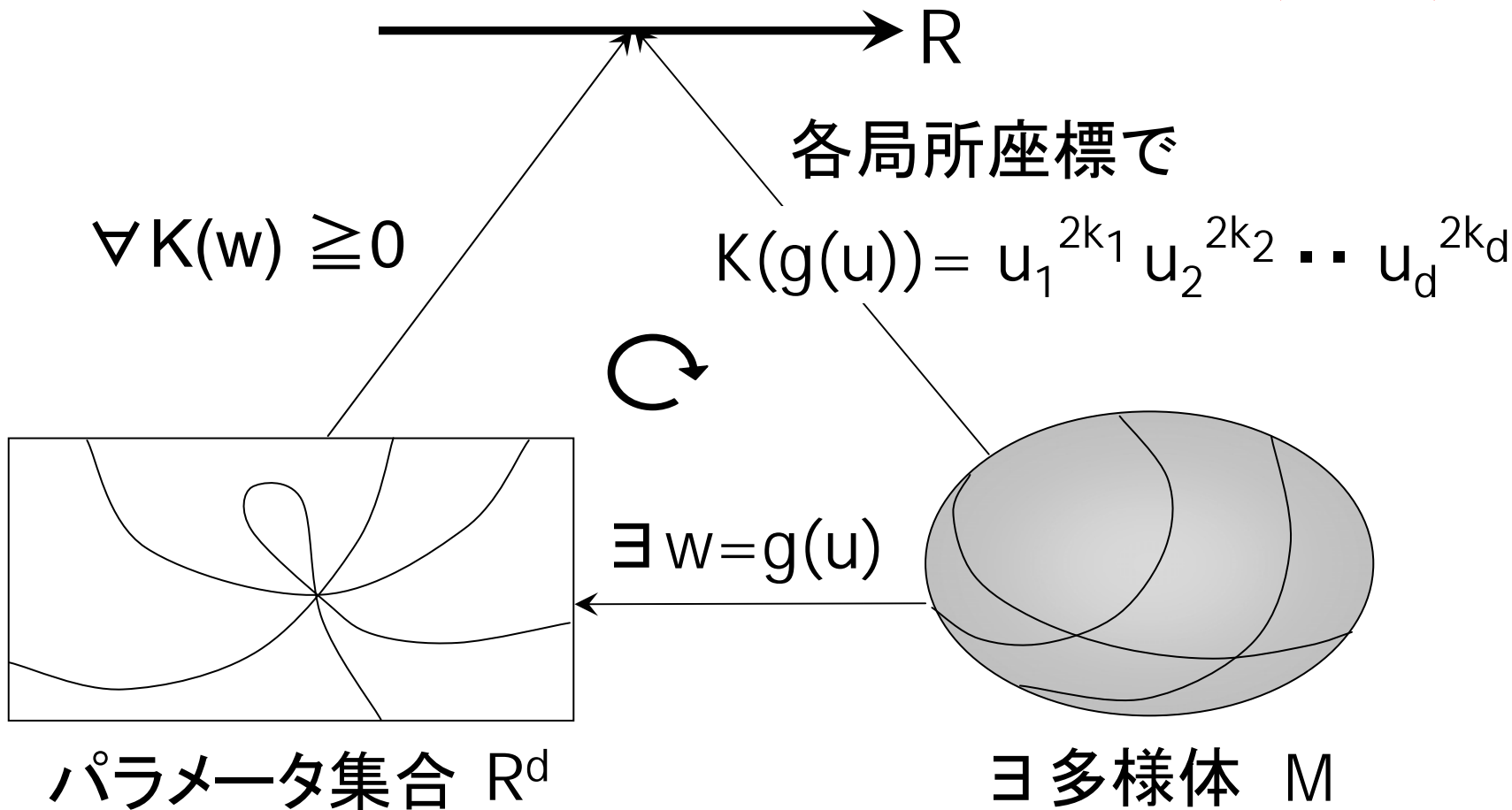
KL 情報量

$$K(w) = E_x[f(X, w)] = L(w) - L_0 \geq 0$$

(解説)ここから理論に入ります。集合 $W_0$ は $K(w)=0$ が成り立つことで特徴付けられます。集合 $W_0$ の幾何学を考えるために、関数 $K(w)$ に対して特異点の解消を行います。それが次のページです。

# 特異点の解消

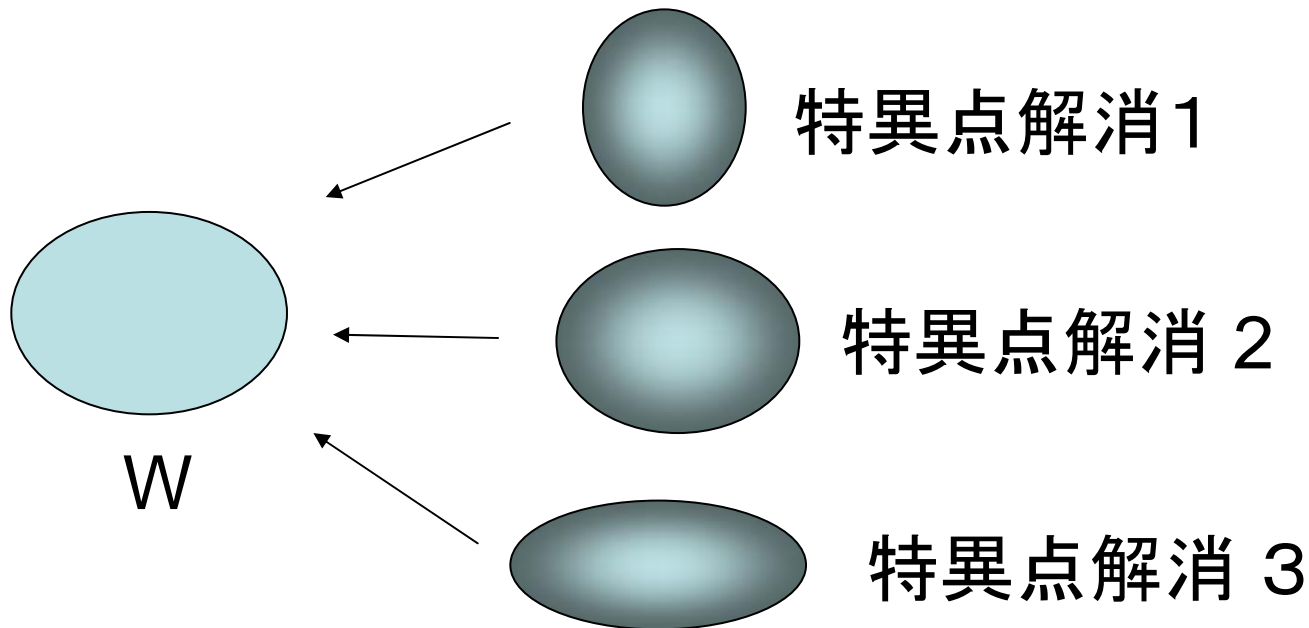
広中(1964)



(解説)これは代数幾何の基本定理です。K(w)について学習理論を作るのは難しいですが、K(g(u))については容易に学習理論を作ることができます。

# 双有理不変量

特異点の解消に対して定義される量が  
特異点の解消によらないとき、**双有理不変量**という。



(解説) 特異点解消は無限にあります。その中で一番基本的なものが  
どんな種類であるかを探すのが「極小モデル・プログラム」です。

# 対数閾値の定義

$$\begin{aligned} K(g(u)) &= u_1^{2k_1} u_2^{2k_2} \cdots u_d^{2k_d} \\ &= u^{2k} \quad \text{と書く} \end{aligned}$$

$$\begin{aligned} |g(u)'| \phi(g(u)) &= b(u) u_1^{h_1} u_2^{h_2} \cdots u_d^{h_d} \\ &= b(u) u^h \quad \text{と書く} \quad (b(u) > 0) \end{aligned}$$

## 対数閾値と重複度(定義)

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

$m =$  最小値を与える  $j$  の個数

# 超関数の漸近挙動

$$\delta(t/n - K(w)) \longleftrightarrow \delta(t/n - u^{2k})$$

補題. ある測度  $D(u)$  が存在して

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

$D(u)$ の台は  $W_0$  に含まれる。

(解説) 対数閾値は、超関数の漸近挙動を定めています。

## 対数閾値と学習理論

対数閾値  $\lambda$  は事後分布が縮む速さ.

$$\begin{aligned} e^{-nK(w)}\varphi(w)dw &= \int \delta(t/n - K(w))\varphi(w)e^{-t} dt dw/n \\ &= \int t^{\lambda-1}e^{-t} dt D(u)du \cdot (\log n)^{m-1}/n^\lambda \end{aligned}$$

(注)  $\lambda$  は双有理不変量.

(解説) 超関数の漸近挙動から事後分布の漸近挙動が導出されます。

# 4

## 特異ゆらぎ

第4章では事後分布における尤度比関数の分散を考え、特異ゆらぎ $\nu$ の定義を述べます。定数 $\nu$ は学習理論で発見された双有理不変量です。定数 $\lambda$ は、尤度比関数の平均値を表し、定数 $\nu$ は分散を表します。

## KL情報量・平均とゆらぎ

平均はわかった。ゆらぎの大きさは？

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

$$E[ K_n(w) ] = K(w)$$

事後分布  $\propto \exp( - n \beta K_n(w) ) \varphi(w)$

(解説) 対数閾値により  $\exp(-nK(w))$  の漸近挙動はわかりました。次に  $\exp(-nK_n(w))$  の漸近挙動を考えます。つまり揺らぎを考えます。

# 尤度比関数

正則あるいは実現可能のとき、ある  $\varepsilon > 0$  があって

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

$K(g(u))=u^{2k}$  から、ある  $a(x,u)$  が存在して

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

(解説) 特異点解消定理から、 $K(g(u))$  は、変数毎の積になるので、 $f(x,g(u))$  が  $u^k$  で割り切れることがわかります。この割り切れるという性質が非常に重要です。割り切れるから  $a(x,u)$  が存在して、漸近挙動がわかります。

## 尤度比関数と経験過程

$$\begin{aligned} nK_n(g(u)) &= \sum_{i=1}^n f(X_i, g(u)) = \sum_{i=1}^n u^k a(X_i, u) \\ &= nu^{2k} - n^{1/2}u^k \cdot \underbrace{n^{-1/2} \sum_{i=1}^n \{ u^k - a(X_i, u) \}}_{\equiv \text{経験過程 } \xi_n(u)} \end{aligned}$$

(解説)  $nK_n(g(u))$ を、平均関数と、そこからのゆれを表す関数に分けて書きました。ゆれを表す関数は経験過程になります。経験過程とは関数空間上の中心極限定理を満たす関数のことです。

# 経験過程と法則収束

## 関数空間上の中心極限定理

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

定理. 対数尤度比関数の漸近挙動

$$nK_n(g(u)) = nu^{2k} - n^{1/2}u^k \xi(u)$$

(解説) 関数  $nK_n(g(u))$  の挙動が解明されました。全ての統計学的な問題は  
この応用として解き明かすことができます。

## 特異ゆらぎ

極限事後分布は、 $\xi$  による平均  $\langle \quad \rangle$

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

---

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

定理.  $n(G-L_0)$ ,  $n(T-L_n)$ ,  $V$  は  $\xi_n$  のある汎関数として表され、対応する  $\xi$  の汎関数に法則収束.

(解説) 事後分布を、サンプル数と共に縮小する部分と一定になる部分とに分けて書くことができました。つまり、事後分布についての繰り込み理論を作ることができました。こうして、学習理論の観測値の挙動が解明できました。

# 特異ゆらぎと汎関数分散

定義. **特異ゆらぎ**

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

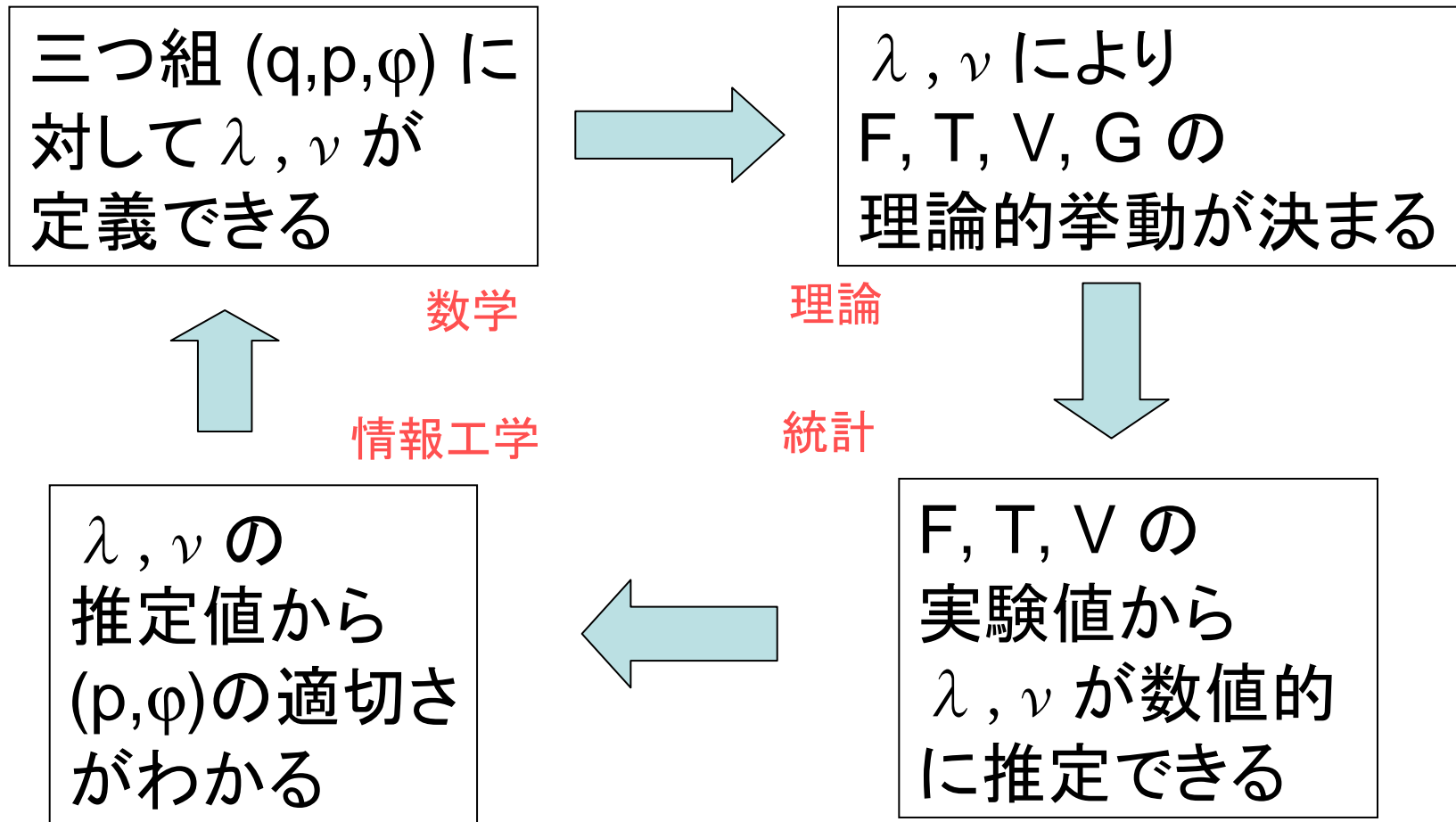
定理. 正則または実現可能のとき

$$\lim_{n \rightarrow \infty} E[ V ] = 2v / \beta$$

特異ゆらぎは双有理不変量である。

(解説) 特異ゆらぎは特異点解消を用いて定義されていますが、特異点解消に依存しない量になっていることがわかりました。

# わかったこと



(解説) 数学と情報工学の間に架橋が作られました。

# 5

## 現在の展開

代数幾何と学習理論の関係をめぐって非常に多くの研究が発展しています。

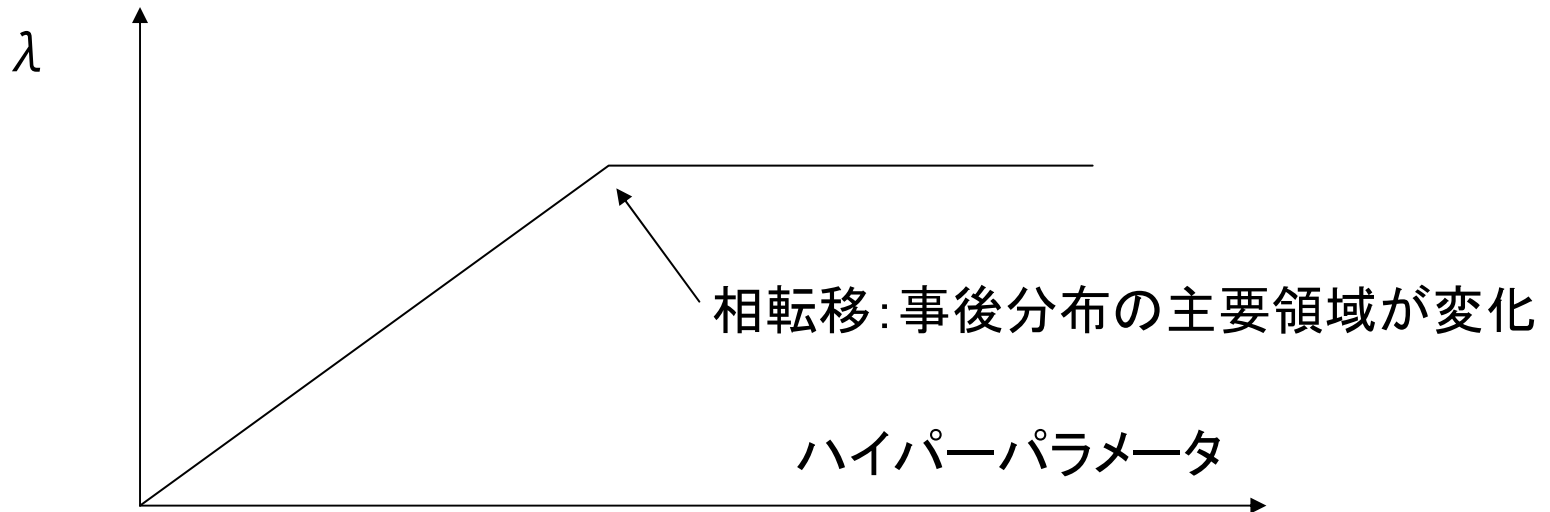
# 統計学として

	正則モデル	一般モデル
クラメル・ラオ	成立	無意味
最尤推定量	漸近正規	漸近正規でない
ベイズ事後分布	漸近正規	漸近正規でない
汎化 - 学習	$d/n$	$2\nu/n$
確率的複雑さ	$(d/2) \log n$	$\lambda \log n$
情報量規準	AIC	WAIC

(解説)この理論は統計学としても真に新しいものです。

# 情報工学として

( $p, \varphi$ ) がハイパーパラメータを持つとき  
 $\lambda$  はハイパーパラメータの関数になる  
→ 学習モデル設計の理論基盤



(解説) 対数閾値  $\lambda$  がモデルやハイパーパラメータによってどのように変化するかを理論的に解明することができます。

# 数学として

- $(W, W_0)$  の相対的な関係を表す
- 高次元代数幾何学において重要
- ベルンシュタイン・佐藤の多項式
- ジェットスキームから算出する方法
- 有限体の場合への一般化
- 計算機代数学の発展

(解説) 対数閾値  $\lambda$  は、純粹数学においても、現在、活発に研究されています。

# 学習モデルの対数閾値

- $q(x)$  が  $p(x|w)$  に対して正則なら  $\lambda = d/2$ . (1978, Schwarz)
- 混合正規分布(2003-, Yamazaki)
- 縮小ランク回帰(2005-, Aoyagi)
- ベイズネット・簡単な場合 (2005-, Rusakov ,Geiger)
- 代数統計学 (Drton, Lin, Sullivant, & Sturmfels)
- 木構造の任意のベイズネット (2010, Zwiernik )

(解説) 学習モデルの対数閾値について、多くのことが解明されています。

# 学習理論の発展

- 変分ベイズの確率的複雑さ (2005, K.Watanabe, T.Hosino)
- ハイパーパラメータについての相転移 (2005, K.Watanabe)
- 変分ベイズの汎化誤差 (2006, S. Nakajima)
- 対数閾値はMCMCの交換率を決めている (2007, Nagata)
- WAIC はクロスバリデーションと漸近等価 (2010, Watanabe)

(解説) 構造を持つ学習モデルについて、理論によって初めてわかったことがたくさんあります。学習理論の実験をするとき、とても参考になると思います。

# 結論

例から構造を学習 = 代数幾何 + 確率論

ここに書かれている結果は、日本ではあまり知られていませんが、海外では学生の人たちも知っている基本的なものです。

## 参考文献

渡辺澄夫, 代数幾何と学習理論、森北出版, 2006.

M. Drton, B. Sturmfels, S. Sullivant, Lectures on Algebraic Statistics, BirkHauser, 2009.