なぜ特異モデルに最尤推定は適していないか

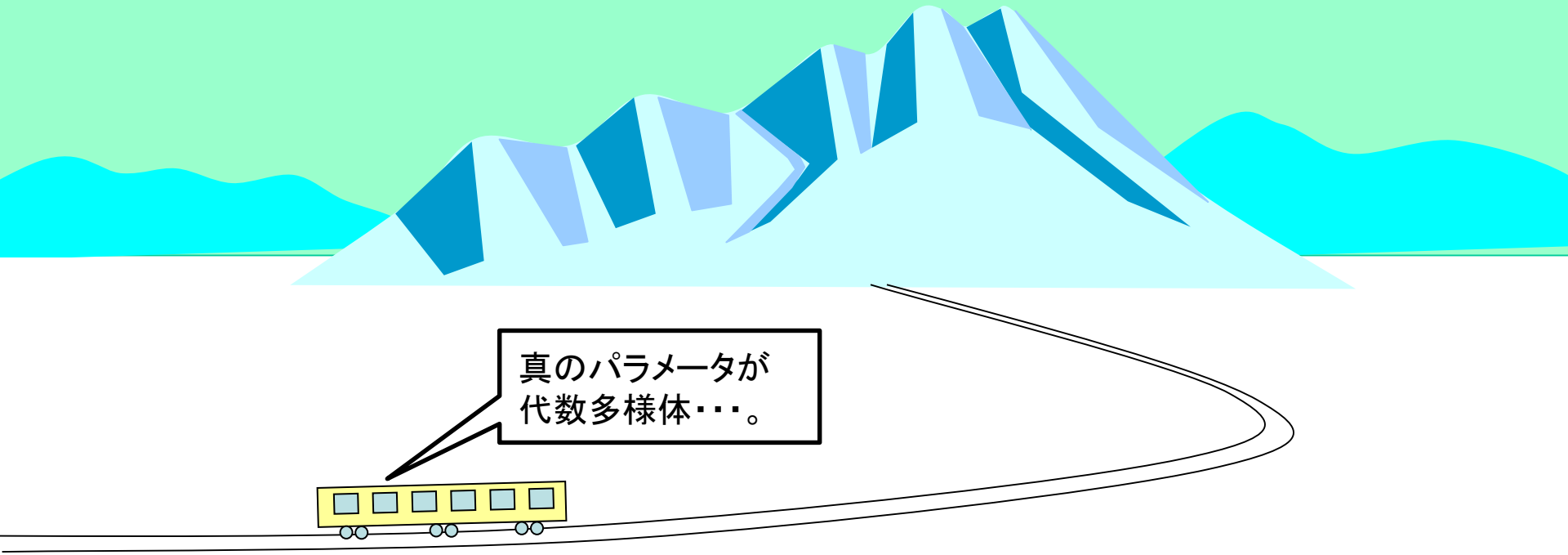
ベイズ統計の理論と方法　4.6節 の解説

このファイルでは、「ベイズ統計の理論と方法」の第４．６節を解説しています。

渡辺澄夫
東京工業大学

特異モデルで最尤推定の汎化誤差が大きくなる理由

学生のみなさんから、「なぜ、特異モデルで最尤推定を使うと汎化誤差が大きくなるのですか」という質問をいただきましたので、お答えします。



特異モデルと最尤推定

神経回路網や混合正規分布において、真 $q(x)$ に対してモデル $p(x|w)$ が冗長な場合を考えてみましょう。関数 $f(x)$ のノルム $\|f\|$ を

$$\|f\|^2 = \int |f(x)|^2 \, q(x) \, dx$$

と定義します。$s_i(x,w) = \partial_{w_i} \log p(x|w)$ とおいて 真とモデルが一致するパラメータの集合を W_0 とするとき、関数の集合

$$\{ \, f(x) \, ; \;\; f(x) = \lim_{w \to W_0} \; s_i(x,w) \, / \, \|s_i(\;\; ,w)\| \text{ の極限が存在 } \}$$

を含む最小の線形空間の次元は無限次元になります。ここで極限 $w \longrightarrow W_0$ は、極限が存在するような経路を任意に選んでよいという意味です。正則な場合は、有限次元（パラメータの次元と等しい）であることに注意してください。

特異モデルと最尤推定

モデルが複雑であるほど汎化損失が大きくなることはご存知ですね。

与えられたデータに対して対数尤度を最大にしようとすると、無限次元の空間の中から尤度を最大にする関数を探してしまうので学習損失は小さくなりますが
汎化損失は大きくなります。

真のパラメータが統計モデルの特異点でなくても、事後分布が正規分布で
近似できない状況では、同種の問題が起こります。

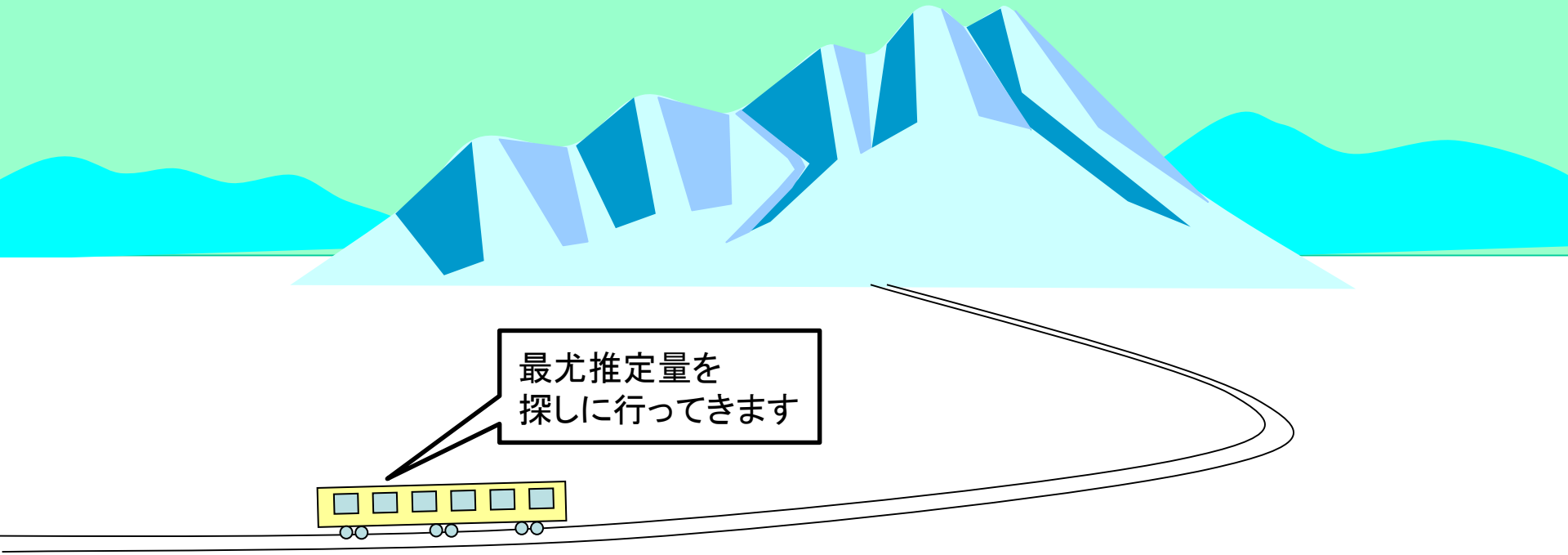
このため、ベイズ法と最尤法を比較すると、ベイズ法のほうが汎化損失が
小さくなります。

実験の例　$Y=a \tanh(bX)$ + 雑音　で真のパラメータが $(0.3,1)$ の場合

http://watanabe-www.math.dis.titech.ac.jp/users/swatanab/a_tanh_bx.mp4

最尤推定の旅

階層的な構造を持つモデルでは、最尤推定量は精度のよい予測を与えないことがわかりました。その理由を考えてみましょう。



4.6 章　最尤推定の汎化誤差と学習誤差

以下のファイルで述べることは、「ベイズ統計の理論と方法」4.6章に書かれています。

最尤推定のときの汎化誤差と学習誤差の漸近挙動を特異点解消定理を
用いて導出してみましょう。

確率モデル $p(x|w)$, 事前分布 $\varphi(w)$
パラメータの集合 W はコンパクトであるとする。（コンパクトでないときは難しい）。

真の分布 $q(x)$ から $p(x|w)$ までのKL情報量を最小にするパラメータを w_0 と書く。
真の分布 $q(x)$ から独立に $\{X_i\}$ が得られたとする。

パラメータがコンパクトでない場合には、解析関数の無限遠点の状況に依存する
ので、一般論はいまのところ難しいです。$Y=a \sin(bx)$ + 雑音　を考えれば
難しい理由がわかると思います。（注：この場合、$b \to \infty$を考えると学習損失は
確率1で零にできることに注意してください）。

4.6 章　最尤推定の汎化誤差と学習誤差

次の関数 $L(w)$ を最小にするパラメータを $\hat{w}$ と書く。

$$L(w) = (1/n) \sum_{i=1}^{n} \log (p(X_i|w_0) / p(X_i|w)) - 1/(n\beta) \log \varphi(w)$$

β=∞のとき最尤推定といい、β=1のとき事後確率最大化推定という。

学習誤差関数と汎化誤差関数をそれぞれ次式で定義する。（本とは定数分の差）。

$$L_n(w) = (1/n) \sum_{i=1}^{n} \log (p(X_i|w_0) / p(X_i|w))$$

$$L(w) = \int q(x) \log (p(x|w_0) / p(x|w)) dx$$

$\hat{w}$ を代入したものを学習誤差、汎化誤差とよび、n→∞での漸近挙動を考える。

L(w) の特異点を解消して考えることにします。

変数を w から u に。

特異点解消写像 $w=g(u)$ $(u \in \mathbf{R}^d)$ を用いて局所座標ごとに

$$n\, L_n(g(u)) = n\, u^{2k} - n^{1/2}\, u^k\, \xi_n(u).$$

ここで $k=(k_1,k_2,\ldots,k_r)$ ただしここでは $k_r>0$ とする(rはd以下の正整数)。

ここで $\xi_n(u)$ はある正規確率過程 $\xi(u)$ に法則収束する確率過程。
特異点解消定理を用いることで $\xi_n(u)$ を well-defined に定義することができる。

ベイズ法の場合にはこの座標 u を用いて漸近挙動が解明できる。

変数を w から u に。さらに (t,v) に

最尤法およびMAP法の場合にはさらに座標変換 を用いる。

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

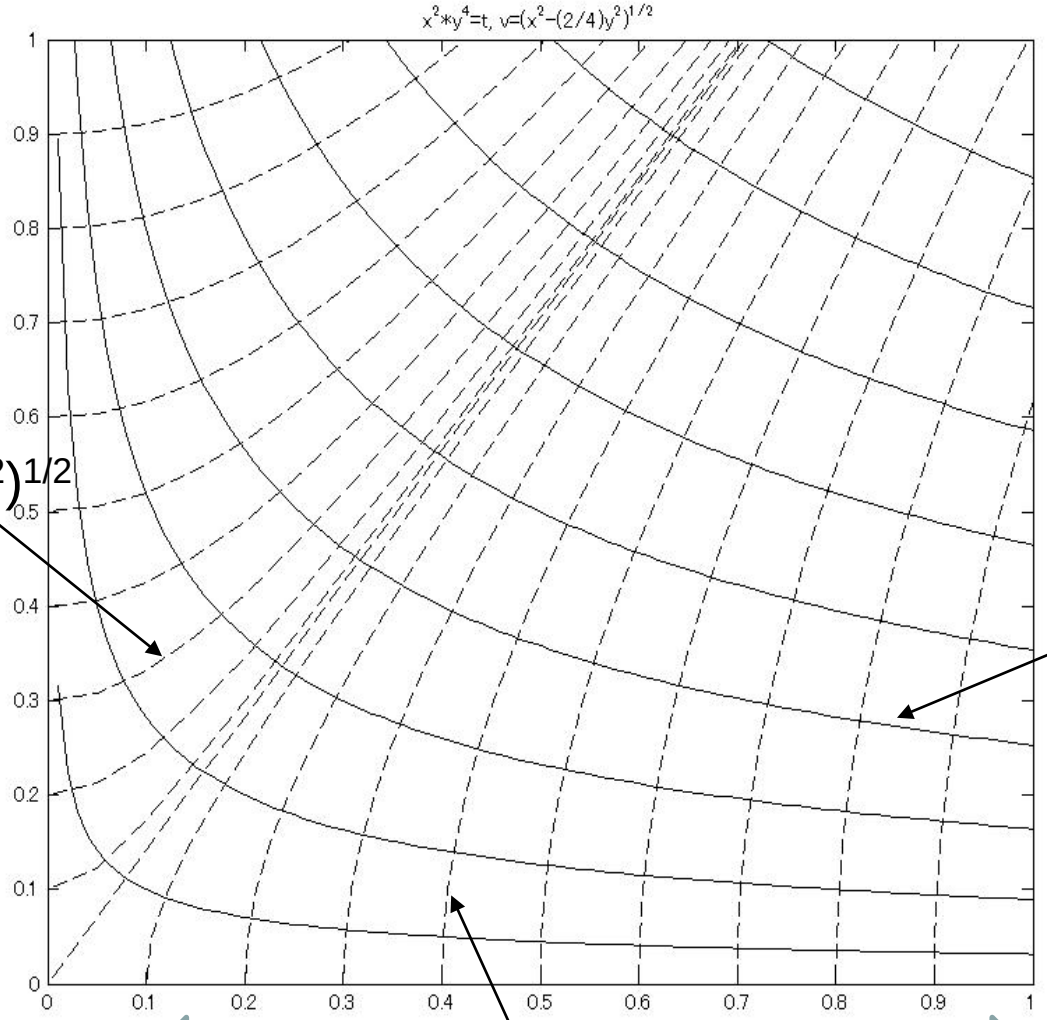
$$v_i = \begin{cases} (u_i^2 - (k_i/k_a)\, u_a^2)^{1/2} & (1 \leqq i \leqq r) \\ u_i & (r < i \leqq d) \end{cases}$$

ここで座標番号 a は、$u_i^2 \geqq (k_i/k_a)\, u_a^2$ となるように決める（u に依存する）。

定義から $v \in V = \{ v ; v_1 v_2 \cdots v_r = 0 \}$ が成り立つ.

この座標の意味は次ページに図示する。

例 $K(g(u)) = u_1^2 u_2^4$ のとき 変数 (t,v) の意味



$v_2=(u_2^2-(4/2)u_1^2)^{1/2}$

$t=u_1^2u_2^4$

$v_1=(u_1^2-(2/4)u_2^2)^{1/2}$

補題

補題 $\boldsymbol{k}=1/\{2(k_1+k_2+\cdots+k_r)\}$ とおくと $[0,1]^d$ 上連続微分可能な f(u) について

$$| f(t,v) - f(0,v) | \leqq C\, t^{\boldsymbol{k}} \quad \sup_{u\in[0,1]^d} \quad \max_{1\leqq j\leqq d} \quad |\partial f/\partial u_j(u)|$$

（証明）$u=(t,v)$, $u'=(0,v)$, $|\nabla f|= \sup \max |\cdots|$ とおく。平均値の定理より

$$| f(t,v) - f(0,v) | = |f(u) - f(u')| \leqq \| u-u' \| \, |\nabla f|$$

$$\leqq r \max |u_j-u'_j| \, |\nabla f| \leqq C \, |u^{2k}|^{\boldsymbol{k}} \, |\nabla f|$$

ここで最後の不等式は次のように導出する。

$j=a$ のときは $|u_j-u'_j| = u_a$. また $j\neq a$ のときは $u_a^2/k_a \leqq u_j^2/k_j$ なので

$|u_j-u'_j| = |u_j-(u_j^2-(k_j/k_a)u_a^2)^{1/2}| = (k_j/k_a)u_a^2 / |u_j + (u_j^2-(k_j/k_a)u_a^2)^{1/2}| \leqq$ 定数 u_a.

一般に $(u_a)^{2(k1+k2+\cdots+kr)} \leqq$ 定数 u^{2k} がなりたつので $|u_j-u'_j| \leqq$ 定数 $(u^{2k})^{\boldsymbol{k}}$.

(証明終わり）

最小化される関数と学習誤差・汎化誤差

座標 (t,v) で記述する。

$\psi(t,v) = - (1/\beta) \log \varphi(g(t,v))$ とおくと、最小化する関数はパラメータについての定数を除いて

$$L(t,v) = t - (t/n)^{1/2} \xi_n(t,v) + (1/n) \psi(t,v) \qquad ①$$

学習誤差関数と汎化誤差関数は

$$L_n(t,v) = t - (t/n)^{1/2} \xi_n(t,v) \qquad ②$$

$$L(t,v) = t \qquad ③$$

準備：補題より $|\xi_n(t,v) - \xi_n(0,v)| = o_p(1)$.

目標：①を最小にする (t,v) を求めて、そのときの②と③を計算する。

平方完成して場合わけする

①式を平方完成して

$$L(t,v) = [\, t^{1/2} - \xi_n(t,v) / 2n^{1/2} \,]^2 - \xi_n(t,v)^2 / 4n + (1/n)\, \psi(t,v) \qquad ④$$

(i) $\xi_n(t,v) \leqq 0$ のとき。(1/n) よりも大きなオーダーの最小化を考えると
定数以下のオーダーの変数 t^* が存在して $t^{1/2} = t^*/n^{1/2}$ である。④に代入して

$$L(t,v) = [t^* - \xi_n(0,v) / 2\,]^2 / n - \xi_n(0,v)^2 / 4n + (1/n)\, \psi(0,v) + o_p(1/n)$$

$\xi_n(t,v) \leqq 0$ より $\xi_n(0,v) \leqq o_p(1)$. 従って $L(t,v)$ を最小にする t^* は $o_p(1)$ である。$t^{1/2} = t^*/n^{1/2}$ を代入して

$$L(t^{*2}/n,v) = (1/n)\, \psi(0,v) + o_p(1/n)$$

この式を最小にする v^* は $v^* = \mathrm{argmin}\, \psi(0,v) + o_p(1)$ である。

$t = o_p(1) / n$, $v = \mathrm{argmin}\, \psi(0,v) + o_p(1)$ のとき

$L_n(t,v) = o_p(1)$, $L(t,v) = o_p(1)$

場合わけの2番目

④式をもう一度書くと

$$\boldsymbol{L}(t,v) = [\, t^{1/2} - \xi_n(t,v) / 2n^{1/2} \,]^2 - \xi_n(t,v)^2 / 4n + (1/n)\, \psi(t,v) \qquad ④$$

(ii) $\xi_n(t,v)>0$ のとき。(1/n) よりも大きなオーダーの最小化を考えると
定数以下のオーダーの項 t* が存在して $t^{1/2} - \xi_n(t,v) / 2n^{1/2} = t^*/n^{1/2}$ 。
そこで $t^{1/2} = \xi_n(t,v) / 2n^{1/2} + t^*/2n^{1/2}$ を代入すると

$$\boldsymbol{L}(t,v) = [t^*]^2 / n - \xi_n(0,v)^2 / 4n + (1/n)\, \psi(0,v) + o_p(1/n)$$

この式を最小にする t* は $o_p(1)$ である。従って

$$\boldsymbol{L}(t,v) = - \xi_n(0,v)^2 / 4n + (1/n)\, \psi(0,v) + o_p(1/n)$$

この式を最小にする v* は $v^*=\mathrm{argmin}\, \{- \xi_n(0,v)^2 / 4 + \psi(0,v)\} + o_p(1)$ である。

$v=\mathrm{argmin}\, \{- \xi_n(0,v)^2 / 4 + \psi(0,v)\} + o_p(1)$, $t = \xi_n(0,v)^2 / 2n + o_p(1)$ のとき

$$L_n(t,v) = - \xi_n(0,v)^2 / 4n + o_p(1/n), \quad L(t,v) = \xi_n(0,v)^2 / 4n + o_p(1/n)$$

主定理

パラメータ $\hat{u}$ はすべての局所座標の中で $L(g(u))$ を最小化するものである。
$\xi_n(0,v)$ の正負により場合わけが生じるが、下記のようにまとめることができる。

$\hat{u} = \mathrm{argmax}\ \{ \max_u(0,\xi_n(u))^2 / 4 - \psi(u)\} + o_p(1)$ が最尤推定量で

(ただし $\max_u$ は $K(g(u))=0$ を満たす u の集合の中での最大値)
学習誤差と汎化誤差は

$$L_n(\hat{u}) = - \max(0, \xi_n(\hat{u}))^2 / 4n + o_p(1/n),$$
$$L(\hat{u}) = \max(0, \xi_n(\hat{u}))^2 / 4n + o_p(1/n)$$

(注意) 変数 v に対する最適化は学習誤差を小さくするが、汎化誤差を大きくするように働いている。本来は推定する必要のないパラメータである v を学習誤差を最小にするために最適化することは汎化誤差を小さくする目的上では適切ではない。

最急降下法では　まず t について最適化され　次に v について最適化される。
これが過学習が生じる理由である。

確率過程の最大値の収束について

正規確率過程 $\xi(u)$ について $\max_u(0,\xi_n(u))^2$ は簡単な確率分布には従わず、
その平均値の具体的な値は一般には求めることができない。

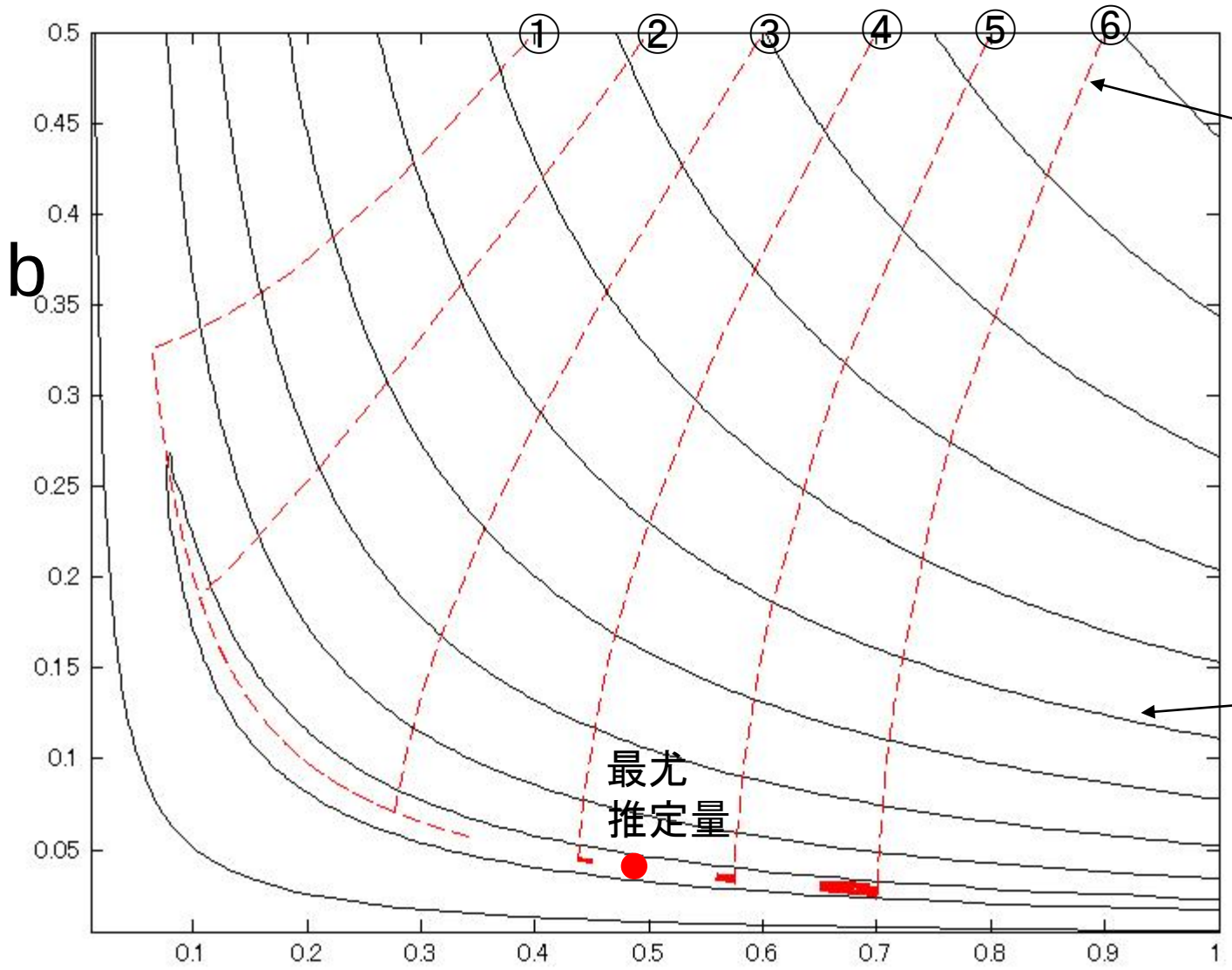
ただし $\xi_n(u) \to \xi(u)$ はコンパクト集合上の有界連続関数全体が作る関数空間
（一様ノルムで完備。可分でもある。）における法則収束である。最大値をとる
操作は、その空間上では連続であるから、法則収束 $\max_u(0,\xi_n(u))^2 \to \max_u(0,\xi(u))^2$
が成り立つ。すなわち、収束先の正規確率過程を使って最大値を表すことができる。

平均値の収束 $E\max_u(0,\xi_n(u))^2 \to E\max_u(0,\xi(u))^2$ を示すためには、もう
少し付加的な条件、たとえば $\max_u(0,\xi_n(u))^2$ が一様漸近可積分であることなど

が必要になる。一般に、その平均値は、正則な場合と比較してずっと大きな値になる。
（確率過程の最大値であるため）。

入力1中間1出力1の神経回路網

モデル: $Y = a \tanh(bx) + N(0,0.1^2)$,　真 (a=b=0) からデータ n=20 個を独立に取る。

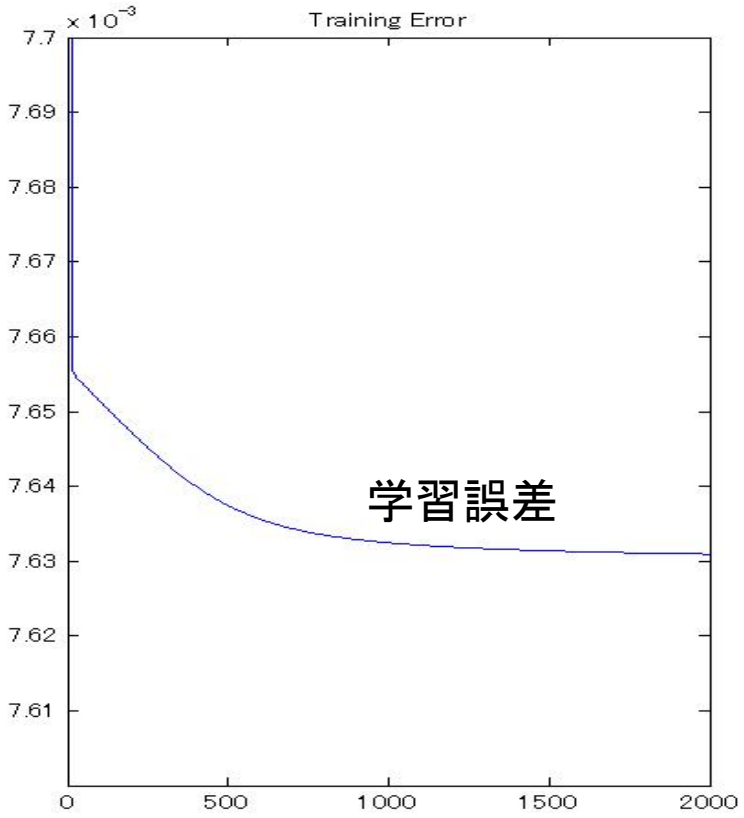


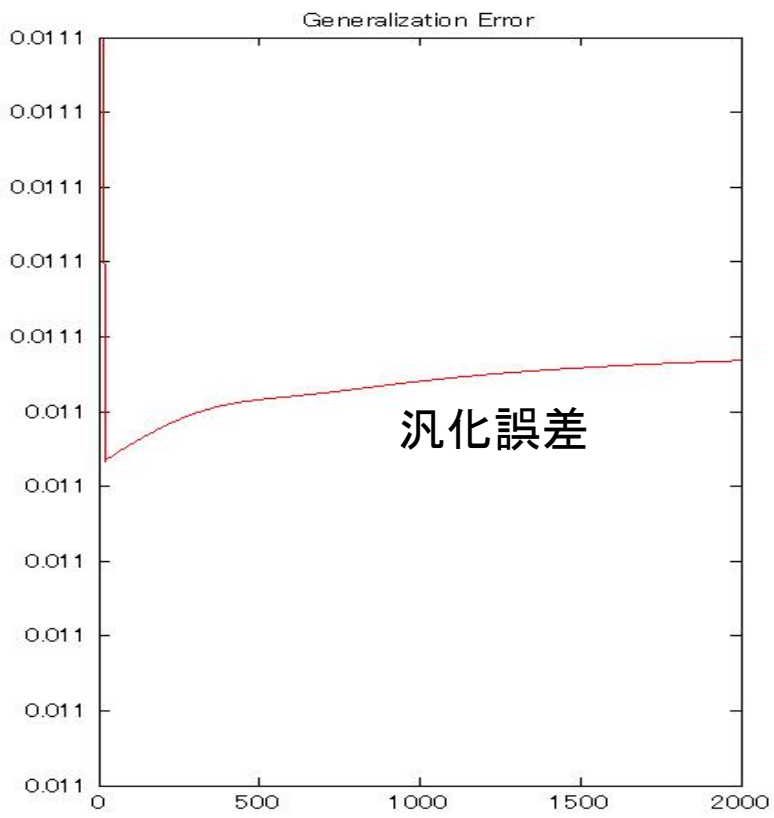
二乗誤差に
対する
最急降下法
の力学系を
初期パラメータ
を①から⑥
まで変化させて
描いたもの

二乗誤差
の等高線

簡単な例

前ページの①の学習における学習誤差と汎化誤差





この例では、すぐに過学習が始まるが、現実の問題では、過学習に見える領域を超えてから、なんども繰り返して学習が行われることがあるため過学習かどうかの判定はカンタンではない。

参考文献

Sumio Watanabe,
Algebraic geometry of singular learning machines and symmetry of generalization and training error. Neurocomputing. 67(1), pp.198-213,2005.