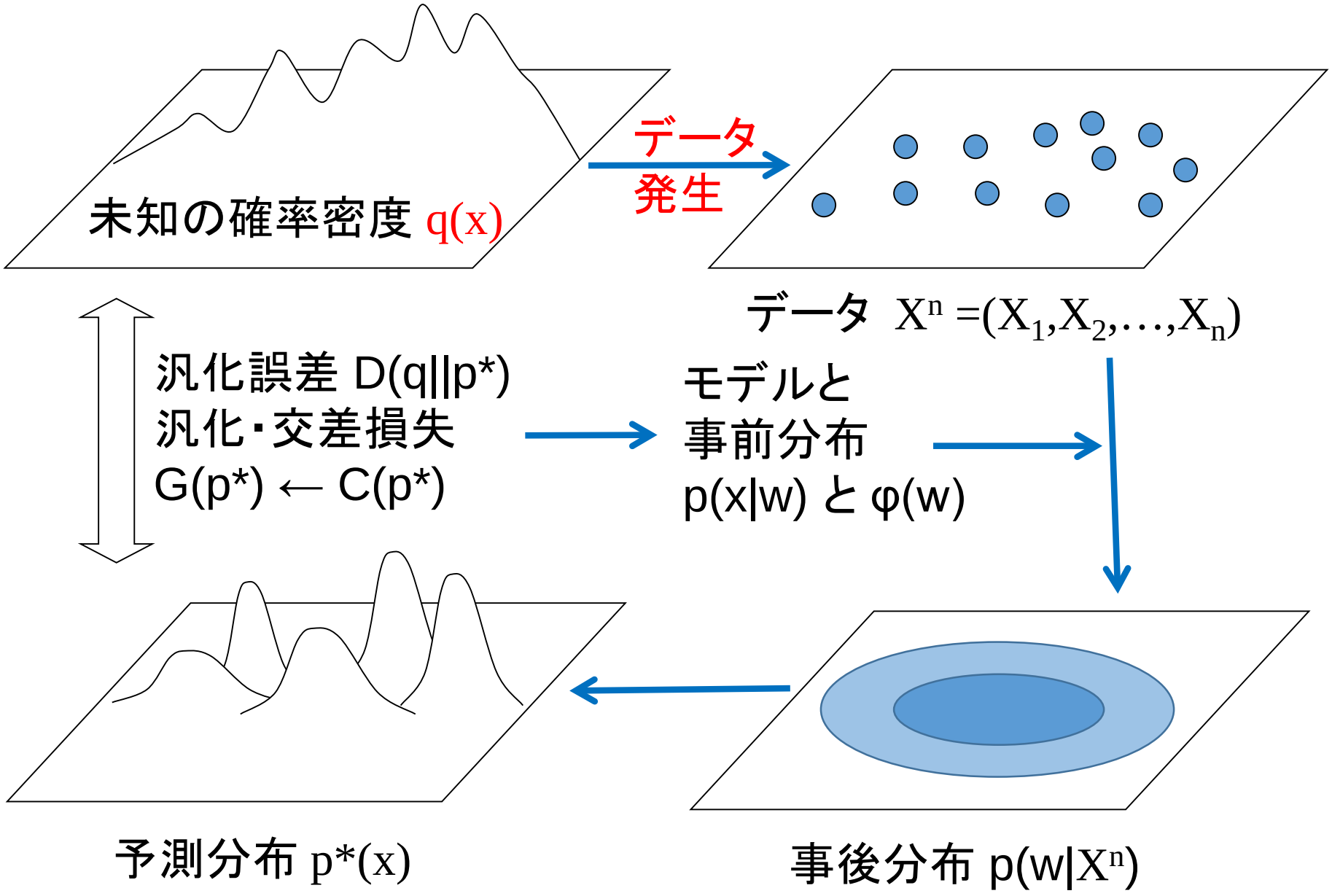
学習理論の練習　7

復習

復習:ベイズ法


未知の確率密度 q(x)
データ
発生
データ $X^n=(X_1,X_2,\ldots,X_n)$
汎化誤差 D(q||p*)
汎化・交差損失
G(p*) ← C(p*)
モデルと
事前分布
p(x|w) と φ(w)
予測分布 p*(x)
事後分布 p(w|X^n)

まだ学んでいないが、いずれ学ぶかもしれないこと

「交差損失を最小化したらうまくいくのでは」という提案がなされた。

「交差損失を最小化したら本当に良いのか」はわからないので

理論的に調べていくことになる （⇒研究へ）。

(1) モデルと事前分布 $p(x|w)$ と $\varphi(w)$ を固定する。

(2) X^n は独立で確率密度 $q(x)$ を持つ。

(3) $G(p^*)$ と $C(p^*)$ は確率変数 $X^n=(X_1,X_2,\ldots,X_n)$ から実数への関数。

⇒ $G(p^*)$ と $C(p^*)$ は実数に値をとる確率変数である。

統計的学習理論の目標の例:

$(q(x),p(x|w),\varphi(w))$ を固定したとき、確率変数 $G(p^*)$ と $C(p^*)$ の挙動を解明したい。特に $n\to\infty$ での挙動を導出したい。

周辺尤度

周辺尤度

定義. $X^n=\{X_1,X_2,\ldots,X_n\}$ を独立な確率変数の集合とする。事前分布 $\varphi(w)$ と統計モデル $p(x|w)$ の周辺尤度を

$$Z(X^n) = \int \varphi(w) \prod_i p(X_i|w)\, dw$$

と定義する。「周辺尤度が大きいほどデータに対して統計モデルと事前分布が適切である」という考え方 (I.J.Good) が提案されている。定義より

$$\int Z(x_1,x_2,\ldots,x_n)\, dx_1 dx_2 \ldots dx_n = 1.$$

$Z(x_1,x_2,\ldots,x_n)$ は $(x_1,x_2,\ldots,x_n)$ の(推測された)確率密度関数である。

注意。$(x_1,x_2,\ldots,x_n)$の真の密度関数は $q(x^n) = q(x_1)q(x_2)\cdots q(x_n)$.

なぜ周辺尤度の最大化が提案されたのか

(1) もしもパラメータ w が $\varphi(w)$ から発生し X^n が $\Pi_i\, p(x_i|w)$ から発生したとすると $Z(X^n)$ は「 $\varphi(w)$と $p(x|w)$ が与えられたときの X^n の確率密度」。

$$Z(X^n) = P(X^n|p,\varphi).$$

(2) 集合 $\{ \varphi(w), p(x|w) \}$ の上の事前確率を $P(p,\varphi)$ と書く。

(3) X^nが得られたとき $\{ \varphi(w), p(x^n|w) \}$ の事後確率は

$$P(p,\varphi|X^n) \propto P(X^n|p,\varphi)P(p,\varphi) = Z(X^n)\, P(p,\varphi).$$

(4) $P(p,\varphi)$＝一定であれば、$Z(X^n)$の最大化は事後確率の最大化。

(5)「 $Z(X^n)$ を最大化すれば良い」とは限らないので、この方法の良さについては理論で調べることになる。

自由エネルギー

自由エネルギー、マイナス対数周辺尤度、ベイズ符号長を次式で定義する。

$$F(X^n) = -\log Z(X^n)$$

このとき「周辺尤度が大きい⇔自由エネルギーが小さい」が成り立つ。

注意。$F(X^n)$ を用いる場合 $\int\varphi(w)dw=1$ が必要です。

注意。交差損失は $\int\varphi(w)dw=1$ が成り立たなくても使える。

注意。自由エネルギー最小化は交差損失最小化とは異なる。

注意。実際、交差損失の最小化と自由エネルギーの最小化では
選ばれるモデルや事前分布は、一般には同じではない。
（たまたま同じになることはある）。

まだ学んでいないがいずれ学ぶかもしれないこと

「周辺尤度を最大化したらうまくいくのでは」という提案がなされた。

「周辺尤度を最大化したら本当に良いのか」はわからないので

理論的に調べていくことになる （⇒研究へ）。

(1) モデルと事前分布 $p(x|w)$ と $\varphi(w)$ を固定する。

(2) X^n は独立で確率密度 $q(x)$ を持つ。

(3) $F(X^n)$ は確率変数 $X^n=(X_1,X_2,\ldots,X_n)$ から実数への関数。

$\Rightarrow F = F(X^n)$ は実数に値をとる確率変数である。

統計的学習理論の目標の例:

$(q(x),p(x|w),\varphi(w))$ を固定したとき、確率変数 $F = F(X^n)$ の挙動を

解明したい。特に $n \to \infty$ での挙動を導出したい。

例

x, w ∈**R** とする。

統計モデル $p(x|w) = 1/(2\pi)^{1/2} \exp(-(x-w)^2/2)$

事前分布　　$\varphi(w) = 1/(2\pi)^{1/2} \exp(-w^2/2)$

データを $\{X_1,X_2,\ldots,X_n\}$ とする。$X^* = 1/(n+1)\ \Sigma\ X_i$, $Y = 1/(n+1)\ \Sigma\ X_i^2$ と書くと

$$\varphi(w) \prod_{i=1}^{n} p(X_i|w) = 1/(2\pi)^{(n+1)/2} \exp(- (n+1)\, w^2/2 + (n+1)\, X^* \cdot w - (n+1)Y/2)$$

$$= 1/(2\pi)^{(n+1)/2} \exp \{ - (n+1)/2 (w - X^*)^2 + (n+1)/2 [(X^*)^2 - Y] \}$$

$$Z(X^n) = 1/\{ (2\pi)^{n/2} (n+1)^{1/2} \} \exp \{ (n+1)/2 [(X^*)^2 - Y] \}$$

$$F(X^n) = (1/2) \log(n+1) + (n/2) \log (2\pi) + (n+1)/2 [Y - (X^*)^2]$$

公式： $a>0$ のとき $\int \exp(-au^2)du = (\pi/a)^{1/2}$

周辺尤度の数学的性質

自由エネルギーとKL情報量

定理. $S= -\int q(x)\log q(x)dx$ とする. $E[\quad]$ を学習データについての平均とすると

$$D(q(x^n) \| Z(x^n)) = E[F_n] - nS,$$

証明:

$$D(q(x^n)\|Z(x^n)) = \int\int \cdots \int q(x^n) \log \frac{q(x^n)}{Z(x^n)} dx_1\, dx_2 \cdots dx_n$$

$$= E[\log q(X^n)] - E[\log Z(X^n)] = - n S + E[F_n].$$

自由エネルギーと汎化損失

汎化損失　$G_n = -\int q(x) \log p(x|X^n)\, dx = E_X[\, -\log p(X|X^n)\,]$.

定理.　$E[\,F_n\,] = \sum_{i=0}^{n-1} E[G_j]$,　　$E[G_{n-1}] = E[F_n] - E[F_{n-1}]$.

証明. 定義から $F_0=0$. 任意の $j>0$ について

$$-\log p(X_{j+1}|X^j) = -\log\left\{ \int \varphi(w)\, p(X_{j+1}|w) \prod_{i=1}^{j} p(X_i|w)\, dw \,/\, Z(X^j) \right\}$$

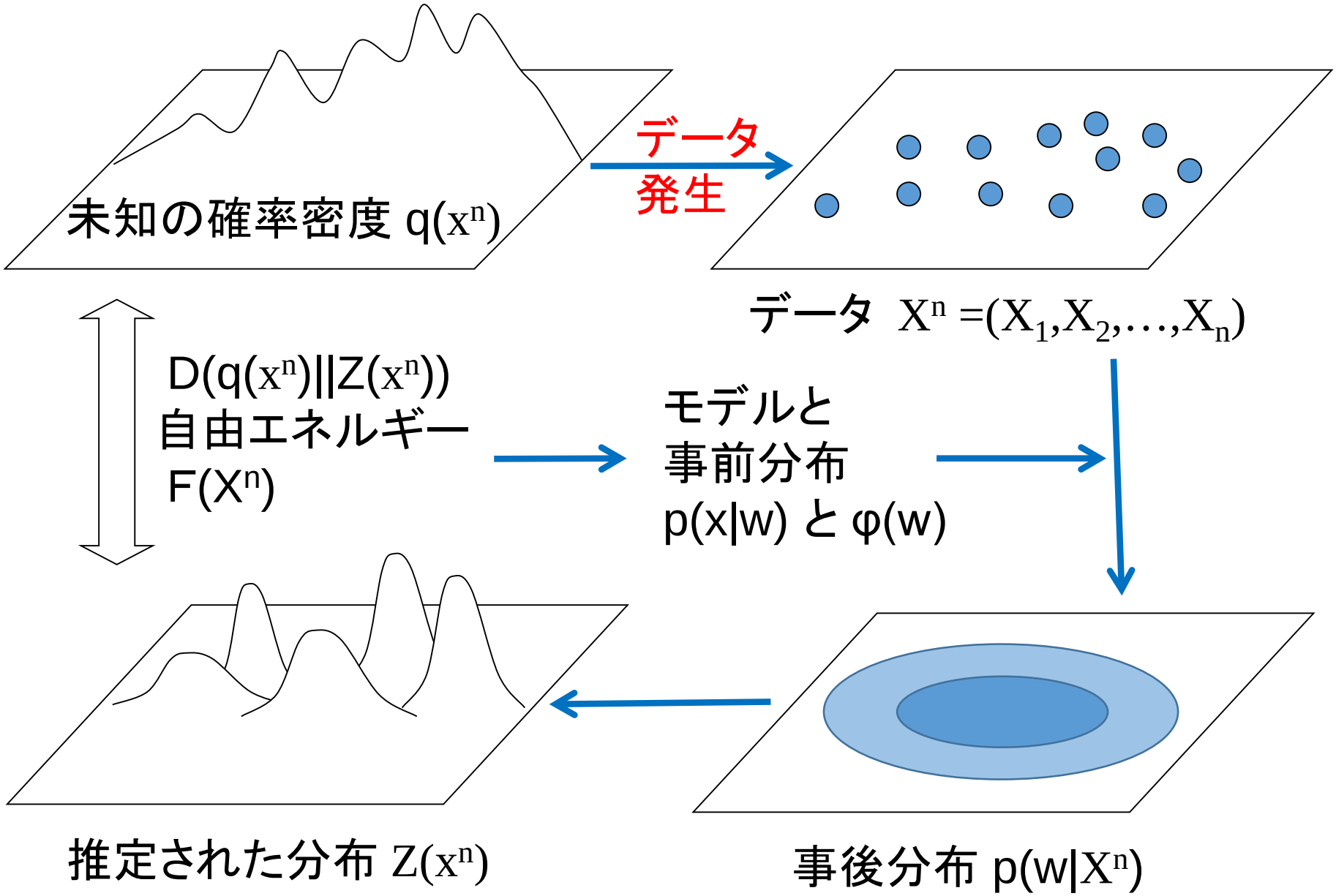
$$= F_{j+1} - F_j.$$

従って　　$E[\,G_j\,] = E[\,-\log p(X_{j+1}|X^j)\,] = E[F_{j+1}\,] - E[F_j]$.

数学的帰納法により定理が証明できた。

（注意） $G_0 = -\log \int p(x|w)\, \varphi(w)\, dw$ と定義した。

ベイズ法


未知の確率密度 q(x^n)
データ
発生
データ X^n =(X_1,X_2,…,X_n)
D(q(x^n)||Z(x^n))
自由エネルギー
F(X^n)
モデルと
事前分布
p(x|w) と φ(w)
推定された分布 Z(x^n)
事後分布 p(w|X^n)

ラプラス近似法

記号 f と φ を $\mathbf{R}^d$ から実数 **R** への関数とする。以下の条件を仮定する。

(1) $f(w) \geqq 0$ は3回連続微分可能で $w=w_0$ でのみ最小値 0 をとる。

(2) 任意の w について $(\nabla^2 f_{jk})(w) = (\partial^2 / \partial w_j \partial w_k) f(w)$ は正定値行列。

(3) $\varphi(w) \geqq 0$ は $\int \varphi(w) dw = 1$ を満たし、$\varphi(w_0) > 0$.

定理1（ラプラス）. 正の整数 n について $Z(n) = \int \exp(- n f(w)) \varphi(w) \, dw$ とおくと $n \to \infty$ の極限において

$$- \log Z(n) = (d/2) \log n + (1/2) \log \det \nabla^2 f(w_0) - \log \varphi(w_0) - (d/2) \log (2\pi) + o(1).$$

漸近挙動

データX^n、事前分布 $\varphi(w)$ 、統計モデル $p(x|w)$ が与えられたとき

周辺尤度を $Z(X^n) = \int \varphi(w) \prod_i p(X_i|w)\, dw$ とおく。

自由エネルギーを $F(X^n) = -\mathrm{Log}\, Z(X^n)$ とおく。

経験誤差関数（= -対数尤度）を $H(w) = -(1/n) \sum_i \log p(X_i|w)$ とおく。

最尤推定量 w^* は $H(w)$ を最小にするパラメータである。

定理２．ラプラス近似が使えるとき次式が成立する。

$$F(X^n) = nH(w^*) + (d/2) \log n + (1/2) \log \det(\nabla^2 H(w^*))$$

$$- \log \varphi(w^*) - (d/2) \log (2\pi) + o_p(1).$$

(注意） 深層学習などではラプラス近似が使えません（⇒研究）。

多項分布の例

{A,B,C} の目が確率 (w_1,w_2,w_3)で出るサイコロの確率分布(多項分布）は

変数を $x \in \{(1,0,0),(0,1,0),(0,0,1)\}$ とすると $p(x|w) = (w_1)^{x1}(w_2)^{x2}(w_3)^{x3}$.

事前分布を集合 $\{w=(w_1,w_2,w_3)\ ;\ w_i \geqq 0,\ w_1+w_2+w_3=1\}$ 上の一様分布

$\varphi(w)=\Gamma(3)$ とする。試行 n 回で{A,B,C} の出た回数を(n_1,n_2,n_3) とすると、

自由エネルギーの厳密値は（*この例では計算できる*）

$$F(X^{\mathbf{n}}) = \log \Gamma(n+3) - \Sigma_{j=1}^{3} \log \Gamma(n_j+1) - \log \Gamma(3).$$

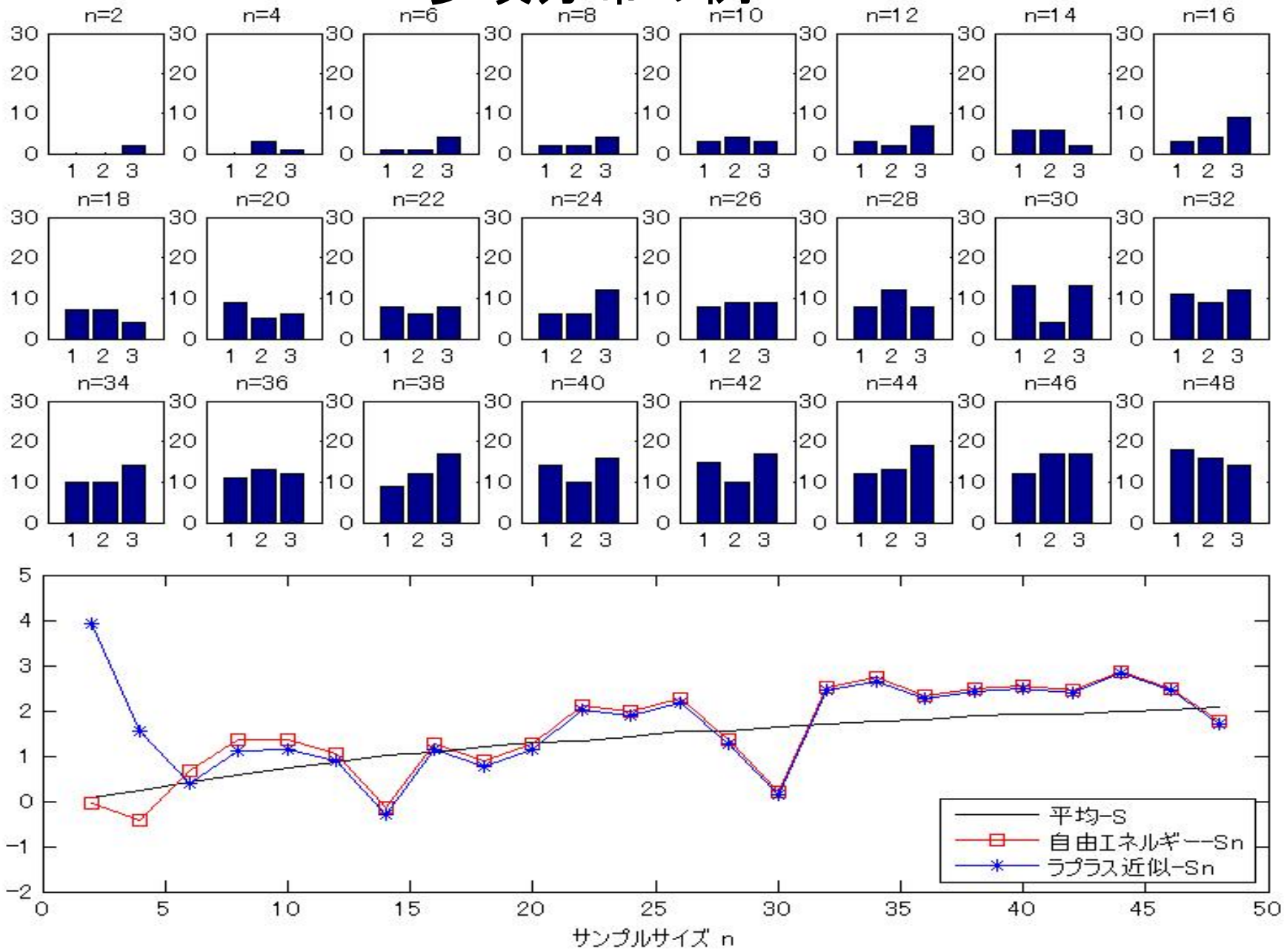
一方、ラプラス近似を用いると $w^*=(n_1/n,n_2/n,n_3/n)$ でパラメータは2個であり

$$F(X^{\mathbf{n}}) \fallingdotseq nH(w^*) + \log n + (1/2) \log \det(\nabla^2 H(w^*)) - \log \Gamma(3) - \log (2\pi).$$

ここで $H(w) = -\Sigma_{j=1}^{3} (n_j/n) \log w_j$ である。真のエントロピーは $S_n=H(w_0)$で

$$\nabla^2 H(w^*) = \begin{pmatrix} n_1/(n\,w_1^2) + n_3/(nw_3^2) & n_3/(nw_3^2) \\ n_3/(nw_3^2) & n_2/(n\,w_2^2) + n_3/(nw_3^2) \end{pmatrix}$$

多項分布の例


n=2
n=4
n=6
n=8
n=10
n=12
n=14
n=16
n=18
n=20
n=22
n=24
n=26
n=28
n=30
n=32
n=34
n=36
n=38
n=40
n=42
n=44
n=46
n=48
平均-S
自由エネルギーSn
ラプラス近似-Sn
サンプルサイズ n

定理1の証明

関数 $\varepsilon = 1/n^{2/5}$ とおく。$Z(n) = Z_1(n) + Z_2(n)$ と分解する。ただし

$$Z_1(n) = \int_{\|w-w_0\|<\varepsilon} \exp(- n f(w)) \varphi(w)\, dw,$$

$$Z_2(n) = \int_{\|w-w_0\|\geqq\varepsilon} \exp(- n f(w)) \varphi(w)\, dw.$$

関数 f(w) は w=w0 で最小値をとるので $f(w_0)=0, \nabla f(w_0)=0.$

$$f(w) = (1/2)(w-w_0) \cdot \nabla^2 f(w_0) (w-w_0) + O(\|w-w_0\|^3).$$

まず$Z_2(n)$を考える。$\nabla^2 f(w)$ は正定値行列なので最小固有値は正値。

集合 $\|w-w_0\|\geqq\varepsilon$ では $n f(w)\geqq c\, n^{1/5}$ となる定数 $c>0$ が存在する。

$$0 \leqq Z_2(n) \leqq \exp(- c\, n^{1/5}).$$

証明つづき

次に $Z_1(n)$を考える。$I(w)$ は正定値行列なので最小固有値は正値。

集合 $\|w-w_0\|<\varepsilon$ では

$$n f(w) = (n/2)(w-w_0)\cdot \nabla^2 f(w_0)(w-w_0) + O(n^{-1/5})$$

$$\varphi(w) = \varphi(w0) + O(n^{-1/2})$$

が成り立つので

$$Z_1(n) = \int_{\|w-w_0\|<\varepsilon} \exp(- n f(w)) \varphi(w)\, dw$$

$$= \int_{\|w-w_0\|<\varepsilon} \exp(- (n/2)(w-w_0)\cdot \nabla^2 f(w_0)(w-w_0)+O(n^{-1/5})) (\varphi(w_0)+O(n^{-1/2}))\, dw$$

証明つづき

変数を変換して $u = n^{1/2}(w-w_0)$ とおくと $du = n^{d/2}\,dw$ であり、積分領域は

集合 $\|w-w_0\| < 1/n^{2/5}$ から集合 $\|u\| < n^{1/5}$ に変換されるので　$Z_1(n)$ は

$$\int_{\|w-w_0\|<\varepsilon} \exp\left(-(n/2)(w-w_0)\cdot\nabla^2 f(w_0)(w-w_0) + O(n^{-1/5})\right)\left(\varphi(w_0)+O(n^{-1/2})\right) dw$$

$$= 1/n^{d/2} \int_{\|u\|< n1/5} \exp\left(-(1/2)u\cdot\nabla^2 f(w_0)\,u + O(n^{-1/5})\right)\left(\varphi(w_0)+O(n^{-1/2})\right) du$$

$$= 1/n^{d/2}\ \varphi(w_0)\left\{ \int \exp\left(-(1/2)u\cdot\nabla^2 f(w_0)\,u\right)\ du + o(1)\right\}$$

$$= 1/n^{d/2}\ \varphi(w_0)\left\{ (2\pi)^{d/2} / (\det \nabla^2 f(w_0))^{1/2} + o(1)\right\}$$

公式

$$\int \exp(-u\cdot Su/2)du$$

$$= (2\pi)^{d/2} / (\det S)^{1/2}$$

以上より $-\log Z(n) = -\log(Z_1(n) + Z_2(n))$ であるから定理が得られた（証明終）。