「ベイズ統計の理論と方法」の4章前半の理解のためのノート

参考文献

1. **渡辺澄夫．ベイズ統計の理論と方法．コロナ社．2012．**
2. 松本幸夫．多様体の基礎．東京大学出版会．1988．
    - 多様体を勉強したことがなかったので言葉の定義の確認と1の分割の理解のために参照しました。但し，4章で肝心な特異点解消定理はより進んだ代数多様体という分野の話なのでこの本にはありません。
3. 土井正男．統計力学．朝倉書店．2006．
    - たまたま手元にありました。元々の状態密度の意味を参照しました（蛇足）。他の統計力学の本やもっと新しい本のことを自分は知りません。

このノートについて

参考文献 1. の4章前半（87〜109ページ）の理解のために記したものですが，参考文献 1. の内容をすべてなぞるものではありません。文章は筆者によるものです。参考文献の定理をそのまま引用している箇所を除き，誤りは筆者に帰属します。お気付きの点がございましたらお手数ですが Twitter またはその他の可能な連絡手段（勉強会の Slack など）にて筆者までご連絡ください。

- この文字色は重要や注意の意味です。
- この文字色は特に脇道にそれる話です。
- 見出しの下に以下のようにあったらその箇所は明確に書き途中です。

> 書き途中です。

目次

登場人物のおさらい

- **パラメータ空間に定義される量たち**

$L(w)$	平均対数損失	$L(w) \equiv -\int q(x) \log p(x\|w) dx$ 。つまり，真の分布 $q(x)$ と確率モデル $p(x\|w)$ の交差エントロピー（なのでサンプルは関係ない）。なので，これが小さいパラメータ w ほど，そのパラメータ w における確率モデルが真の分布に近いイメージ。この $L(w)$ の最小点 w_0 が「最適なパラメータ」といわれる。真の分布が確率モデルで実現可能な場合には $p(x\|w_0) = q(x)$ なので $L(w_0)$ は真の分布のエントロピーに等しくなる。
$L_n(w)$	経験対数損失	$L_n(w) \equiv -(1/n) \sum_{i=1}^n \log p(X_i\|w)$ 。つまり，サンプルによる経験分布と確率モデル $p(x\|w)$ の交差エントロピー。なのでサンプルに依存する確率変数。
$K(w)$	平均誤差	$K(w) \equiv \int q(x) f(x,w) dx$ 。つまり，真の分布 $q(x)$ 上での，対数尤度比（※） $f(x,w) \equiv \log p(x\|w_0)/p(x\|w)$ の平均（なのでサンプルは関係ない）。平均損失の最適なパラメータとの差 $L(w) - L(w_0)$ に等しい。最適なパラメータ w_0 においては任意の x で $f(x,w_0) = 0$ なので $K(w_0) = 0$ である。
$K_n(w)$	経験誤差	$K_n(w) \equiv (1/n) \sum_{i=1}^n f(X_i,w)$ 。つまり，サンプルによる経験分布上での，対数尤度比 $f(x,w)$ の平均。なのでサンプルに依存する確率変数。平均誤差同様，$L_n(w) - L_n(w_0)$ に等しく，$K_n(w_0) = 0$ である。

※ 平均対数損失 $L(w)$ の最小点が1点でなく，最適な確率分布が実質的にユニークでない場合は対数尤度比は w_0 の選び方に依存するので $f(x, w_0, w)$ とかくべきだが，3章と4章では最適な確率分布が実質的にユニークな場合を扱う。

- **予測分布の形にかかわる量たち**

すべてサンプルに依存する確率変数。

$Z_n(\beta)$	分配関数	$Z_n(\beta) \equiv \int_W \varphi(w) \prod_{i=1}^n p(X_i\|w)^\beta dw$ 。事後分布の分母。$Z_n(\beta)$ 自体は $n \to \infty$ で 0 に収束してしまうので，n を大きくしたときのベイズ推測の性質を知りたいときは，下の自由エネルギーの方が適している。 --- 以下のように式変形できる。 $Z_n(\beta) = \int_W \exp\big(\log \prod_{i=1}^n p(X_i\|w)^\beta\big)\varphi(w)dw$ $= \int_W \exp\big(\beta \sum_{i=1}^n \log p(X_i\|w)\big)\varphi(w)dw$ $= \int_W \exp\big(-n\beta L_n(w)\big)\varphi(w)dw$ $= \int_W \exp\big(-n\beta\big(L_n(w_0) + K_n(w)\big)\big)\varphi(w)dw$ $= \exp\big(-n\beta L_n(w_0)\big) \int_W \exp\big(-n\beta K_n(w)\big)\varphi(w)dw$ $\equiv \exp\big(-n\beta L_n(w_0)\big) Z_n^{(0)}(\beta)$ ここで定義した $Z_n^{(0)}(\beta) \equiv \int_W \exp\big(-n\beta K_n(w)\big)\varphi(w)dw$ を正規化された分配関数とよぶ。$K_n(w) \geqq 0$ なので，$0 < Z_n^{(0)}(\beta) \leqq 1$ 。
$F_n(\beta)$	自由エネルギー	$F_n(\beta) \equiv -\beta^{-1} \log Z_n(\beta)$ 。分配関数 $Z_n(\beta)$ の対数をとって逆温度の逆数をかけてマイナス1倍したもの。 --- $F_n^{(0)}(\beta) \equiv -\beta^{-1} \log Z_n^{(0)}(\beta)$ を正規化された自由エネルギーとよぶ。$0 \leqq F_n^{(0)}(\beta)$ 。また，正規化された分配関数の式の対数をとればわかるように， $F_n(\beta) = nL_n(w_0) + F_n^{(0)}(\beta)$ 。
$p(w\|X^n)$	事後分布	$p(w\|X^n) \equiv Z_n(\beta)^{-1}\varphi(w) \prod_{i=1}^n p(X_i\|w)^\beta$ 。ベイズ推測の予測分布とはこの事後分布上で確率モデルを平均したもの $p^*(x) \equiv \int_W p(x\|w)p(w\|X^n)dw$ である。 --- 正規化された分配関数 $Z_n^{(0)}(\beta)$ と経験誤差 $K_n(w)$ を用いて以下のようにかける。 $p(w\|X^n) = Z_n(\beta)^{-1} \exp\big(-n\beta\big(L_n(w_0) + K_n(w)\big)\big)\varphi(w)$ $= Z_n^{(0)}(\beta)^{-1} \exp\big(-n\beta K_n(w)\big)\varphi(w)$

- **予測分布のよさをあらわす量たち**

すべてサンプルに依存する確率変数。

G_n	汎化損失	$G_n \equiv -\int q(x) \log p^*(x)dx$ 。真の分布 $q(x)$ と予測分布 $p^*(x)$ の交差エントロピー（予測分布はサンプルに基づくのでサンプル依存）。 --- 以下のように式変形できる。 $G_n = -\int q(x) \log p^*(x)dx$ $= -\int q(x) \log\Big(\int_W p(x\|w)p(w\|X^n)dw\Big)dx$ $= -\int q(x) \log\Big(\int_W \exp\big(\log p(x\|w)\big)p(w\|X^n)dw\Big)dx$ $= -\int q(x) \log\Big(\int_W \exp\big(\log p(x\|w_0) - f(x,w)\big)p(w\|X^n)dw\Big)dx$ $= -\int q(x) \log\Big(\exp\big(\log p(x\|w_0)\big) \int_W \exp\big(-f(x,w)\big)p(w\|X^n)dw\Big)dx$ $= -\int q(x) \log p(x\|w_0)dx - \int q(x) \log\Big(\int_W \exp\big(-f(x,w)\big)p(w\|X^n)dw\Big)dx$ $= L(w_0) - \int q(x) \log\Big(\int_W \exp\big(-f(x,w)\big)p(w\|X^n)dw\Big)dx$ 上式の第2項 $G_n^{(0)} \equiv -\int q(x) \log\Big(\int_W \exp\big(-f(x,w)\big)p(w\|X^n)dw\Big)dx$ を汎化誤差とよぶ。汎化誤差は，真の分布が確率モデルで実現可能な場合には，真の分布と予測分布のカルバック・ライブラー情報量に等しい（その場合に限っては42ページの誤植が正しい）。 --- $\mathcal{G}_n(\alpha) \equiv \mathbb{E}_X\Big[\log \mathbb{E}_w\big[p(X\|w)^\alpha\big]\Big]$ と定義すると（これを汎化損失のキュムラント母関数とよぶ；サンプルに依存する確率的な関数），以下が成り立つ。 $\mathcal{G}_n(1) = \mathbb{E}_X\Big[\log \mathbb{E}_w\big[p(X\|w)\big]\Big] = \mathbb{E}_X\Big[\log\big[\int_W p(X\|w)p(w\|X^n)dw\big]\Big]$ $= \mathbb{E}_X\Big[\log p^*(X)\Big] = \int q(x) \log p^*(X)dx = -G_n$ また，任意の $\alpha > 0$ に対して，$\mathcal{G}_n(\alpha)$ が区間 $[0, \alpha]$ で3回微分可能ならば，テイラーの定理より以下を満たす $0 < \alpha^* < \alpha$ が存在する。 $\mathcal{G}_n(\alpha) = \mathcal{G}_n(0) + \alpha\mathcal{G}_n'(0) + (1/2)\alpha^2\mathcal{G}_n''(0) + (1/6)\alpha^3\mathcal{G}_n^{(3)}(\alpha^*)$

		したがって，汎化損失はある $0 < \alpha^* < 1$ を用いて以下のように展開できる。 $G_n = -\mathcal{G}_n(0) - \alpha\mathcal{G}_n'(0) - (1/2)\alpha^2\mathcal{G}_n''(0) - (1/6)\mathcal{G}_n^{(3)}(\alpha^*)$
T_n	経験損失	$T_n \equiv -(1/n)\sum_{i=1}^n \log p^*(X_i)$ 。サンプルによる経験分布と予測分布 $p^*(x)$ の交差エントロピー（なので経験分布的にも予測分布的にもサンプル依存）。 --- 汎化損失同様，以下のように式変形できる。 $T_n = L(w_0) - (1/n)\sum_{i=1}^n \log\Big(\int_W \exp\big(-f(X_i,w)\big)p(w\|X^n)dw\Big)$ 上式の第2項 $T_n^{(0)} \equiv -(1/n)\sum_{i=1}^n \log\Big(\int_W \exp\big(-f(X_i,w)\big)p(w\|X^n)dw\Big)$ を経験誤差（※ $K_n(w)$ も経験誤差なので名前がかぶっているが別物である）とよぶ。 --- $\mathcal{T}_n(\alpha) \equiv (1/n)\sum_{i=1}^n \log \mathbb{E}_w\big[p(X_i\|w)^\alpha\big]$ と定義すると（これを経験損失のキュムラント母関数とよぶ；サンプルに依存する確率的な関数），以下が成り立つ。 $\mathcal{T}_n(1) = (1/n)\sum_{i=1}^n \log \mathbb{E}_w\big[p(X_i\|w)\big] = (1/n)\sum_{i=1}^n \log\big[\int_W p(X_i\|w)p(w\|X^n)dw\big]$ $= (1/n)\sum_{i=1}^n \log p^*(X_i) = -T_n$ また，$\mathcal{T}_n(\alpha)$ が区間 $[0,1]$ で3回微分可能ならば，汎化損失同様，経験損失はある $0 < \alpha^* < 1$ を用いて以下のように展開できる。 $T_n = -\mathcal{T}_n(0) - \alpha\mathcal{T}_n'(0) - (1/2)\alpha^2\mathcal{T}_n''(0) - (1/6)\mathcal{T}_n^{(3)}(\alpha^*)$

3章までのあらすじ

全体的に以下を仮定します。

- パラメータ集合 W はユークリッド空間のコンパクトな部分集合（※）であり，かつ，開集合 $W' \supset W$ が存在して，$K(w)$ が W' 上の解析関数（※）であることにします（95ページ）。
  - ※ 「ユークリッド空間のコンパクトな部分集合である」は，「ユークリッド空間の閉集合であって，じゅうぶん大きな半径 $r > 0$ の球で覆うことができる」と同値です。
  - ※ 解析関数とは，定義域の各点においてその点の周りでのテイラー展開と一致するような関数のことです。
- 3章と4章では，対数尤度比 $f(x,w_0,w)$ が相対的に有限な分散をもつ（任意の w_0, w について，真の分布上での $f(x,w_0,w)$ の2乗の平均が1乗の平均の定数倍で抑えられる）とします（39ページ，41ページ）。

1章のあらすじ

- 「ベイズ推測する」とは「真の分布はおおよそ $p^*(x)$ だろうと考える」ということです。しかし，そういわれただけでは $p^*(x)$ は本当にいつも真の分布に近く（近いとは？）なってくれるのかとか，どんなときにどれくらい近くなってくれるのかとかが全然わかりません。この辺の理論的な根拠を示さないとベイズ推測の沽券に関わります。
  - 現実に統計的推測をするときは「真の分布」などというものはないことも多いですが，少なくとも真の分布があるときには（何らかの指標で）それに近づくことができないと統計的推測として駄目だと思います。
- そういう根拠として，「真の分布 $q(x)$ と予測分布 $p^*(x)$ の交差エントロピーである汎化損失 G_n が，サンプル数 n を大きくしていくほど，その確率モデル $p(x|w)$ で到達しうる下限値 $L(w_0)$ の近くで分布する」といえればひとまずはよい気がします。
  - 「サンプル数を大きくしていけば真の分布を再現することができる」は，統計的推測に期待する性質としてまっとうなものだと思います。頻度統計でもこれに該当する一致性（$n \to \infty$ でパラメータの推定量が真のパラメータに確率収束すること）という性質があり，よく知られる最尤推定や最小2乗推定は適当な条件下でそれを満たします。
  - 分布間の近さの指標として交差エントロピーは唯一無二のものではないですが，ここではこれを指標にします。
  - 結果的には上のことはちゃんといえて，3章だと70ページの定理3，4章だと114ページの定理12がそれです。

2章のあらすじ

- なので，汎化損失 $G_n \equiv -\mathbb{E}_X\Big[\log \mathbb{E}_w\big[p(X|w)\big]\Big]$ ってどう分布するんだろう，という話になります。この式のままだと G_n の分布がどんな形をしていて，その形がサンプル数 n にどう依存しているのか全然わかりません。そこでまず一番内側の期待値に注目すると，$p(X|w)$ という量について事後分布上での期待値 $\mathbb{E}_w$ をとっています。これを事後分布の平均（期待値）ではなくもっと違う分布形状の特徴，例えば分散などが出てくるように変形したいです。なぜなら，事後分布はサンプル数 n が大きくなるほど分散が小さくなりそうなので，分散のような分布形状の特徴にこそ n への依存性が出そうだからです（この文章は何となくかいたのでとりわけあやしいです）。ここで一般に，確率変数 Z に対して，$e^{\alpha Z}$ の期待値を

$$\mathbb{E}[e^{\alpha Z}] = \exp\left(\alpha\kappa_1 + \frac{1}{2}\alpha^2\kappa_2 + \frac{1}{3!}\alpha^3\kappa_3 + \cdots\right)$$

と展開できます（キュムラント展開）。ここで κ_k は Z の k 次キュムラントといって，Z の分布の形状を特徴づける値になっており，1 次キュムラントは Z の平均に等しく，2 次キュムラントは Z の分散に等しいです。なので，$Z = \log p(X|w)$ を事後分布上でキュムラント展開して $\alpha = 1$ とすれば

$$\mathbb{E}_w\left[p(X|w)\right] = \exp\left(\kappa_1 + \frac{1}{2}\kappa_2 + \frac{1}{3!}\kappa_3 + \cdots\right)$$

$$\log \mathbb{E}_w\left[p(X|w)\right] = \kappa_1 + \frac{1}{2}\kappa_2 + \frac{1}{3!}\kappa_3 + \cdots \qquad (*)$$

となり，分散 κ_2 を登場させることができます。一般に Z の k 次キュムラントは $\log \mathbb{E}\left[e^{\alpha Z}\right]$ を α で k 回微分して $\alpha = 0$ とすると得られます（キュムラント母関数）。なので，ここでのキュムラント母関数は $\log \mathbb{E}_w\left[p(X|w)^\alpha\right]$ ですが，汎化損失は $(*)$ にさらに真の分布上での期待値 $\mathbb{E}_X$ をとるので，今回はキュムラント母関数もあらかじめ $\mathcal{G}_n(\alpha) \equiv \mathbb{E}_X\left[\log \mathbb{E}_w\left[p(X|w)^\alpha\right]\right]$ と定義すれば

$$G_n = -\kappa_1 - \frac{1}{2}\kappa_2 - \frac{1}{3!}\kappa_3 - \cdots$$

と展開できます。

- ただ，キュムラント母関数をこう定義するならば，k 次キュムラントを求めるときに，これをこのまま k 回微分して $\alpha = 0$ とするのは誤りで，「$\mathbb{E}_X$ を一度はがして，その中身に k 回微分して $\alpha = 0$ として，$\mathbb{E}_X$ をとる」という手順によらなければならないと思います。しかし，α での微分と $\mathbb{E}_X$ が交換するならどのみち同じです。以下，交換するものとします。

- なお，汎化損失 G_n は対数尤度比をつかって $G_n = -\mathbb{E}_X\left[\log \mathbb{E}_w\left[\exp\left(-L(w_0) - f(X,w)\right)\right]\right]$ ともかけるので，キュムラント母関数も $\mathcal{G}_n(\alpha) = -\alpha L(w_0) + \mathbb{E}_X\left[\log \mathbb{E}_w\left[\exp\left(-\alpha f(X,w)\right)\right]\right]$ ともかけます。
- 具体的に $\mathcal{G}_n(\alpha)$ の 1 回微分と2 回微分を計算すると以下を得ます（α での微分と $\mathbb{E}_w$ も交換するものとします）。

$$\frac{d\mathcal{G}_n(\alpha)}{d\alpha} = -L(w_0) + \mathbb{E}_X\left[\frac{\mathbb{E}_w\left[-f(X,w)\exp\left(-\alpha f(X,w)\right)\right]}{\mathbb{E}_w\left[\exp\left(-\alpha f(X,w)\right)\right]}\right]$$

$$\frac{d^2\mathcal{G}_n(\alpha)}{d\alpha^2} = \mathbb{E}_X\left[\frac{\mathbb{E}_w\left[\left(-f(X,w)\right)^2\exp\left(-\alpha f(X,w)\right)\right]}{\mathbb{E}_w\left[\exp\left(-\alpha f(X,w)\right)\right]} - \left(\frac{\mathbb{E}_w\left[-f(X,w)\exp\left(-\alpha f(X,w)\right)\right]}{\mathbb{E}_w\left[\exp\left(-\alpha f(X,w)\right)\right]}\right)^2\right]$$

よって，1 次キュムラントと 2 次キュムラントはこうなります。

$$\kappa_1 = \left.\frac{d\mathcal{G}_n(\alpha)}{d\alpha}\right|_{\alpha=0} = -L(w_0) + \mathbb{E}_X\left[\mathbb{E}_w\left[-f(X,w)\right]\right] = -L(w_0) - \mathbb{E}_w\left[K(w)\right]$$

$$\kappa_2 = \left.\frac{d^2\mathcal{G}_n(\alpha)}{d\alpha^2}\right|_{\alpha=0} = \mathbb{E}_X\left[\mathbb{E}_w\left[f(X,w)^2\right] - \mathbb{E}_w\left[f(X,w)\right]^2\right]$$

このことから，汎化損失 G_n が以下のように，下限値 $L(w_0)$ と残りの項の和でかけることがわかります。

$$G_n = L(w_0) + \mathbb{E}_w\left[K(w)\right] - \frac{1}{2}\mathbb{E}_X\left[\mathbb{E}_w\left[f(X,w)^2\right] - \mathbb{E}_w\left[f(X,w)\right]^2\right] - \frac{1}{3!}\kappa_3 - \cdots$$

残りの項には事後分布上での平均誤差 $K(w)$ の期待値や対数尤度比 $f(X,w)$ の分散が出てきます。3 次や 4 次のキュムラントは事後分布上での $\log p(X|w)$ の分布の尖度や歪度に対応しますが，もし $n \to \infty$ でその分布が正規分布（正規分布は 3 次以上のキュムラントが 0）に近づくならば 3 次以上のキュムラントは 0 に近づきます。

- まとめると，ベイズ推測の沽券のために少なくとも示さないといけないことは，n を大きくすると

$$\mathbb{E}_w\left[K(w)\right] - \frac{1}{2}\mathbb{E}_X\left[\mathbb{E}_w\left[f(X,w)^2\right] - \mathbb{E}_w\left[f(X,w)\right]^2\right] - \frac{1}{3!}\kappa_3 - \cdots$$

が小さくなることです（これはサンプルの選び方に依存する確率変数ですが，n が大きいときはこの分布の平均が 0 に近くなってほしいし，分散も小さくなってほしいということです）。

3章のあらすじ

- 上のことを示さないといけないんですが，といっても事後分布がどんな形か全然わからないので事後分布上での期待値や分散などわかりません。事後分布は $p(w|X^n) \propto \exp\bigl(-n\beta L_n(w)\bigr)\varphi(w)$ でした。これをみると，$L_n(w)$ の部分が $(w-\mu)^\top \Sigma^{-1}(w-\mu) \cdots (**)$ という形をしてくれていれば正規分布になることに気付きます（但し Σ は正定値）。正規分布であれば期待値も分散も計算できそうです。$L_n(w)$ はサンプルの選び方に依存するので $L_n(w)$ が $(**)$ の形であると仮定するのは現実的ではないですが，$L(w)$ が $(**)$ の形であると仮定して，n がじゅうぶん大きければ $L_n(w)$ の形もだいたいこの形になるといえればいい気がします。また，パラメータ空間 W 上のどこでも $(**)$ の形でないといけないわけでもなくて，事後分布の確率密度が一番濃くなる点は（$n \to 0$ で）w_0 なので，確率密度が密集していく w_0 の周りでだけこの形であれば何とかなりそうです。最適なパラメータ w_0 がただ1点のみでない場合は事後分布が単峰の正規分布にならないので w_0 はただ1点と仮定してしまいましょう。$L(w)$ が w_0 の周りで $(**)$ の形に近似できることは，$L(w)$ のヘッセ行列が $w=w_0$ で正定値であることと同じです。

  いま仮定したこと： $L(w)$ がただ1つの最小点 w_0 をもち，w_0 でのヘッセ行列が正定値である（※）。

  ※ 2章で定義した言葉でいえば，「 $q(x)$ が $p(x|w)$ に対して正則である」。

  実際，こう仮定すれば，n が大きいときに事後分布が正規分布に近くなることを示すことができます。以下の手順によります。

  - 事後分布は $p(w|X^n) \propto \exp\bigl(-n\beta K_n(w)\bigr)\varphi(w)$ ともかけました（分布のピークで exp の中身が 0 になるように $L_n(w)$ をオフセットしただけです）。事後分布の密度が濃いところにだけ注目したいので，$K(w)$ が 0 に近いところだけ切り取ってしまいましょう。どうせ外側の確率は無視できると思います。内側の確率と外側の確率はそれぞれ以下です。

$$Z_n^{(1)}(\beta) \equiv \int_{K(w)<\epsilon} \exp\bigl(-n\beta K_n(w)\bigr)\varphi(w)dw$$

$$Z_n^{(2)}(\beta) \equiv \int_{K(w)\geqq\epsilon} \exp\bigl(-n\beta K_n(w)\bigr)\varphi(w)dw$$

    但し，ϵ は n の関数であって，$n \to \infty$ で 0 に近づくが，$n^{-1/2}$ よりは 0 に近づくのが遅いものにしてください。$\epsilon = n^{-1/4}$ とかでいいです。w_0 の周りだけ切り取りたいんですが，あまり小さく切り取りすぎると外側の確率が無視できなくなってくるのでこうします。
  - ちゃんと確かめると，上記の $Z_n^{(2)}(\beta)$ の方は $\exp(-\sqrt{n})$ より速く 0 に収束します（補題12）。
  - ちゃんと確かめると，上記の $Z_n^{(1)}(\beta)$ の方は以下のようになることがわかります（補題13）。

$$Z_n^{(1)}(\beta) = \int_{K(w)<\epsilon} \exp\left(-\frac{n\beta}{2}\left\|J^{1/2}\left(w-w_0-\frac{\hat{\xi}_n}{\sqrt{n}}\right)\right\|^2\right)dw \times \exp\left(\frac{\beta}{2}||\xi_n||^2\right) \times \varphi(w_0)(1+o_p(1))$$

  - だったら事後分布による平均は以下のようになりそうです。なります（補題15）。これで事後分布上での期待値や分散が計算できそうです。

$$\mathbb{E}_w[\cdot] = \frac{\displaystyle\int (\cdot)\exp\left(-\frac{n\beta}{2}\left\|J^{1/2}\left(w-w_0-\frac{\hat{\xi}_n}{\sqrt{n}}\right)\right\|^2\right)dw}{\displaystyle\int \exp\left(-\frac{n\beta}{2}\left\|J^{1/2}\left(w-w_0-\frac{\hat{\xi}_n}{\sqrt{n}}\right)\right\|^2\right)dw}(1+o_p(1))$$

  - 事後分布が正規分布に近づくなら，3 次以上のキュムラントも 0 に近づいていくはずです。確かめると，$n^{-k/2}$ と同じ速さで 0 に近づきます（補題17）。
- 以上から，汎化損失 G_n は n を大きくするほど $L(w_0)$ の近くで分布することが示されます（定理3）。

$$G_n = L(w_0) + \frac{1}{n}\left(\frac{d}{2\beta} + \frac{1}{2}||\xi_n||^2 - \frac{1}{2\beta}\mathrm{tr}(IJ^{-1})\right) + o_p\left(\frac{1}{n}\right)$$

  言い換えると，ベイズ推測の予測分布は，サンプル数 n を大きくするほど確率モデルで実現しうるベストな分布である $p(x|w_0)$ に近づくことが示されました。これならベイズ推測を安心して利用することができそうです。

4章前半のあらすじ

- と，一瞬安心した気がしたんですが，さっきの議論では $L(w)$ のヘッセ行列がただ1つの最小点 $w=w_0$ で正定値であるという仮定を置いていました。この仮定が満たされているかは現実には確かめようがありません（真の分布は知り得ないので）（もっとも，最初に仮定した「平均誤差が解析的」とか「相対的に有限な分散」とかも知り得ないと思いますが）。これではやっぱり安心して眠れない気がします。
- $L(w)=(w-\mu)^\top\Sigma^{-1}(w-\mu)$ と仮定することができないなら，パラメータ空間 W の方を歪めてしまうことで，つまり，何か座標変換することで，$L_u(u)=(u-\mu)^\top\Sigma^{-1}(u-\mu)$ のようにすることはできないでしょうか。オフセットした $K(w)$ の方でいうなら $K_u(u)=u^\top\Sigma^{-1}u$ のようにすることはできないでしょうか。
- 実は，$K_u(u)=u_1^{2k_1}u_2^{2k_2}\cdots u_d^{2k_d}$ とすることがいつもできます。以下の手順によります。
  - かなり自由に歪めることができる多様体という空間を用意して，パラメータはそこから上京してきたことにします。パラメータは田舎 $\mathcal{M}$ では u という住所（座標）だったんですが，都会 W に上京して $w=g(u)$ という住所（座標）になったということにします。
  - 実は，この本の仮定を満たすパラメータ空間 W と平均誤差 $K(w)$ にはいつも，「その多様体上の座標 u でなら $K_u(u)=u_1^{2k_1}u_2^{2k_2}\cdots u_d^{2k_d}$ となる」ような都合のよい多様体 $\mathcal{M}$ と座標変換 g が存在してくれます（特異点解消定理）。そんな多様体 $\mathcal{M}$ をパラメータたちの出身の田舎だと思って，帰省してもらえばよいです。

  u は $\mathcal{M}$ 上の d 次元座標（$\mathbb{R}^d$ の元 ※）で，$k=(k_1,k_2,\cdots,k_d)$ は少なくともどれか1つは 0 ではない非負整数の d 個組ですが，以降，表記の仕方の約束として，$u^{2k}\equiv u_1^{2k_1}u_2^{2k_2}\cdots u_d^{2k_d}$ とします。

  ※ 但し u の取りうる値は $\mathbb{R}^d$ 全体ではなく，g によって W にうつされる $\mathcal{M}$ 上の点の座標がとりうる値です。このような u の集合はコンパクトになります（95ページ）。
- そうなると事後分布 $p(w|X^n)\propto\exp\bigl(-n\beta K_n(w)\bigr)\varphi(w)$ がどうなるのか考えます。
  - 経験誤差 $K_n(w)$ は $n\to\infty$ で $K(w)$ に近づきそうですが，実際以下のようになります（定理7）。

$$nK_n(g(u))=nu^{2k}-\sqrt{n}u^k\xi_n(u)$$

  - 事前分布 $\varphi(w)$ はどうなるのでしょうか。3章では $q(x)$ が $p(x|w)$ に対して正則と仮定していたので $n\to\infty$ ではただ1点の w_0 での値しか事後分布に関わってきませんでしたが（というか正規化で消えましたが），いまは最適なパラメータ w_0 は複数あったり無数にあったりするかもしれないです。$\mathcal{M}$ の座標の世界で事前分布がどのような分布になっているかちゃんと考えないといけないです。ただ，せっかく都合のよい座標変換をするのなら，$\varphi(w)$ の分布の濃淡も込みで座標変換しておけばいい気がします。実際そのようにできます。
    - $w=g(u)$ という座標変換をすると点 u の周りの微小体積は u におけるヤコビ行列の絶対値 $|g'(u)|$ 倍に引きのばされます（重積分の変数変換）。なので，確率の釣り合いより，帰省先での事前分布の密度 $\varphi_u(u)$ は $\varphi(w)$ と以下の関係があります（g が単射のときのイメージです）。

$$\varphi_u(u)du=\varphi(w)|g'(u)|du=\varphi\bigl(g(u)\bigr)|g'(u)|du$$

      ただこれだと，$\varphi\bigl(g(u)\bigr)$ の形がよくわかりません。そこで実は，$K(w)$ に特異点解消定理を適用するときに，$\varphi(w)$ と同時に特異点を解消することができます（96ページ）。するとこうなります。

$$\varphi_u(u)du=|u^h|b(u)du$$

  - まとめると，帰省先では事後分布は以下に比例します。

$$p_u(u|X^n)\propto\exp\bigl(-n\beta u^{2k}+\sqrt{n}\beta u^k\xi_n(u)\bigr)|u^h|b(u)$$

- 正則でない場合にも事後分布がかけたのはいいんですが，ただこれは正規分布っぽさはありますが正規分布ではないのでどんな分布なのかよくわかりません。やりたいことは事後分布上での平均誤差 $K(w)$ の期待値を求めたり，対数尤度比 $f(X,w)$ の分散を求めたりすることでした。この積分ができればいいです。
  - 平均誤差は特異点解消により u の関数としては $K(g(u))=u^{2k}$ となっていました。
  - 対数尤度比は実は $f(X,g(u))=\sqrt{u^{2k}}a(x,u)$ とかけます（補題20）。
  - これらとさっきの事後分布をみると u^{2k} が出てきます。$t=u^{2k}$ と変数変換できれば，部分積分で期待値や分散が求められそうです（被積分関数が多項式と指数関数の積の形になるので）。変数を w から u に変換するのは特異点解消定理がよしなにやってくれたんですが，u から t に変換するのはこちらでちゃんとやらないといけないです。変数を変換するには，変換先のある微小体積に，変換前のどこの体積が入るのか確かめて，確率の釣り合いが取れるような変換先の密度を求めないといけません。$t=u^{2k}$ は単射ではないので，変換先のある微小体積には複数の微小体積が集まってくることになります。それらを考慮した点 t での密度（デルタ関数を用いて表現すると $\delta(t-u^{2k})$ に他なりません）を t の関数でかきくださなければなりません。頑張ってかきくだすと以下のようになります（定理8）。

$$\delta(t-u^{2k})|u^h|b(u)du = t^{\lambda-1}(-\log t)^{m-1}du^* + o\big(t^{\lambda-1}(-\log t)^{m-1}\big)$$

これで正則でない場合の事後分布上での期待値や分散をとる準備ができました。4章後半に続きます。

4章前半のノート

STEP1. 平均誤差が標準形になるようにパラメータを変換する

書き途中です。

パラメータ空間 W の方を歪めることで平均誤差 $K(w)$ を標準形にしたいです。まず，パラメータの帰省先 $\mathcal{M}$ を用意したいと思います。$\mathcal{M}$ は集合です。どんな構造をした集合にすべきかということになりますが，さすがに各点の周りに座標（つまり，帰省先でのパラメータの姿）はほしいです。ただの集合だと「この点の周り」という概念がないので，まず位相というものを入れます。

定義（開集合系，位相空間）

集合 $\mathcal{M}$ の部分集合族 $\mathcal{O}$ が以下の3つの条件を満たすとき， $\mathcal{O}$ は $\mathcal{M}$ の**開集合系**であるという。

1. $\emptyset, \mathcal{M} \in \mathcal{O}$
2. $O_1, O_2 \in \mathcal{O} \Rightarrow O_1 \cap O_2 \in \mathcal{O}$
3. 任意の集合 Λ に対し，各元 $\lambda \in \Lambda$ から $\mathcal{O}$ の元 O_λ への対応を与えたとき，$\bigcup_{\lambda\in\Lambda} O_\lambda \in \mathcal{O}$

集合 $\mathcal{M}$ にある開集合系 $\mathcal{O}$ が与えられているとき， 集合 $\mathcal{M}$ を $\mathcal{O}$ を開集合系とする**位相空間**といい，$\mathcal{O}$ を $\mathcal{M}$ の**位相**という。また，$\mathcal{O}$ の元を $\mathcal{M}$ の**開集合**という。

次に位相空間 $\mathcal{M}$ の各点の周り（各点を含む開集合）に d 次元の座標を割り当てますが，好き勝手な写像で座標を割り当てたらせっかく入れた位相が台無しなので位相が保たれるようにします。つまり，$\mathcal{M}$ の開集合の像が $\mathbb{R}^d$ の開集合になり，$\mathbb{R}^d$ の開集合の逆像も $\mathcal{M}$ の開集合になるようにします。

定義（連続写像）

X，Y を位相空間とする。写像 $f : X \to Y$ に対して，Y の任意の開集合 V の逆像 $f^{-1}(V)$ が X の開集合であるとき，f は**連続写像**であるという。

定義（同相写像）

X，Y を位相空間とする。写像 $f : X \to Y$ に対して，f が全単射で f も f^{-1} も連続写像であるとき，f は**同相写像**であるという。

定義（座標近傍系）

位相空間 $\mathcal{M}$ の開集合 U から $\mathbb{R}^d$ の開集合 V への写像 $\phi : U \to V$ が同相写像であるとき，(U, ϕ) の対を $\mathcal{M}$ の d **次元座標近傍（チャート；地図）**という（※）。また，d 次元座標近傍の族 $S = \{(U_\lambda, \phi_\lambda)\}_{\lambda\in\Lambda}$ が $\bigcup_{\lambda\in\Lambda} U_\lambda = \mathcal{M}$ を満たすとき（※），S を $\mathcal{M}$ の d **次元座標近傍系（アトラス；地図帳）**という。

※ このとき ϕ を $U \subset \mathcal{M}$ の**局所座標系**といい，$x \in U$ に対して $\phi(x)$ を**局所座標**といいます。

※ このように $\{U_\lambda\}_{\lambda\in\Lambda}$ が $\mathcal{M}$ 全体を覆っているとき，集合族 $\{U_\lambda\}_{\lambda\in\Lambda}$ は集合 $\mathcal{M}$ の**被覆**であるといいます。特に，U_λ が全て開集合である場合は**開被覆**といいます。

というわけで，パラメータの帰省先の集合 $\mathcal{M}$ は，位相空間であって d 次元座標近傍系が定義されているものであってほしい気がします。なお，ハウスドルフ性（後述）をもつ位相空間であって d 次元座標近傍系が定義されているものを**多様体**といいます。

※ 単に多様体といったときの定義はお手元の本により異なることがあります。

多様体なら，局所ごとに自由に座標を割り当てられるので，パラメータを多様体にとばして座標を上手く割り当てれば平均誤差 $K(w)$ を標準形にすることも実現できそうな気がします。実際，以下のような定理があります。

定理（特異点解消）

$K(w) \geqq 0$ を開集合 $W \subset \mathbb{R}^d$ 上の非負の値をとる解析関数とし，$K(w) = 0$ を満たす $w \in W$ が存在するとする。このとき，ある d 次元多様体 $\mathcal{M}$ と解析写像 g が存在して，$\mathcal{M}$ の局所座標ごとに（途中）

STEP2. 変換したパラメータ上の事後分布を求める

書き途中です。

STEP3. 事後分布による期待値をとるためさらに変数変換する

書き途中です。

事後分布に対して $t = u^{2k}$ という変数変換がしたいです。変換先の点 t の周りの微小体積には，$t = u^{2k}$ を満たす点 u たちの周りの微小体積が移されるので，そのような点たちにおける密度を抜き出して，変換に伴う微小体積の伸縮を考慮した比で足し合わせたものが変換後の密度になります。

ここで，他の関数にかけて積分することで，その関数のある点（たち）における値を抜き出すようなはたらきをするものがあります。デルタ関数といいます。デルタ関数といいますが関数ではないです。

定義（デルタ関数）

無限回微分できる任意の関数 $\varphi(x)$ に対して，以下が成り立つような $\delta(x)$ をデルタ関数という。

$$\int \delta(x)\varphi(x)dx = \varphi(0)$$

蛇足

- 力学で，注目している系の全エネルギーを，構成要素の座標と運動量の関数としてかいたものをハミルトン関数といいます。全ての変数を束ねて Γ として，$H(\Gamma)$ とかくことにします。この系のエネルギー E が一定に保たれているとすると，この系がとりうる状態 Γ は $H(\Gamma) = E$ を満たします。そんな風に，系の構成要素が何か運動するとき，状態 Γ がどんなルールを満たすかを取り扱うのが力学です。ただ「Γ はこの式を満たします」だけだと，たくさんの構成要素からなる系が全体としてどんな性質を帯びているかを考えにくいです。
- ので，統計力学では，たくさんの構成要素からなる系の状態 Γ がどう分布するかを取り扱います。
    - 例えば，エネルギー E が一定に保たれている系（ミクロカノニカルアンサンブルといいます）をある一定の長時間観察することにします。状態 Γ が観察される確率密度 $P(\Gamma)$ というものを考えたいです。まず，$H(\Gamma) = E$ を満たさない Γ が観察されることはないです。なのでそういう Γ は $P(\Gamma) = 0$ でいいです。他方，$H(\Gamma) = E$ を満たす Γ であれば観察されうります。Γ の空間内の曲線 $H(\Gamma) = E$ 上の点はすべて観察されうるとなると，この曲線上の確率密度をどうするべきかとなりますが，「曲線上の点は全部同じ確率密度でいいや」と仮定します（等重率の原理；等重率の原理が成り立つことを証明できる系もありますが，一般には証明できないので仮定とかきました）。「$H(\Gamma) = E$ を満たす Γ のみが実現し，複数あればどれも等しい確率で実現する」ような確率密度は

      $$P(\Gamma) \propto \delta\big(E - H(\Gamma)\big)$$

      に他ならないです（ミクロカノニカル分布）。これを用いると，この系において $F(\Gamma)$ という量が平均的にどんな値になるかを以下によって求めることができます。

      $$\langle F \rangle = \frac{\int F(\Gamma)\delta\big(E - H(\Gamma)\big)d\Gamma}{\int \delta\big(E - H(\Gamma)\big)d\Gamma}$$

      この分母を $W(E)$ とかいて状態密度とよびます。これは正規化定数です。$W(E)$ は $H(\Gamma) = E$ を満たす Γ において $H(\Gamma)$ の傾きが急なら小さくなるし，緩やかなら大きくなるし，$H(\Gamma) = E$ を満たす Γ が複数あれば足し合わされて大きくなります。$W(E)$ が大きい E の周辺では系が色々な状態をとりえるといったイメージです。ちなみに，状態密度の対数をエントロピーとよびます。

      $$S(E) = k_B \log W(E)$$

    - 系のエネルギーではなく温度 T が一定に保たれている場合（カノニカルアンサンブルといいます）は，確率密度は以下になります（カノニカル分布）。

      $$P(\Gamma) = \frac{\exp\big(-H(\Gamma)/(k_B T)\big)}{\int \exp\big(-H(\Gamma)/(k_B T)\big)d\Gamma}$$

この分母の正規化定数を $Z(T)$ とかいて分配関数とよびます。この積分は状態密度を用いて E の積分にすることもできます。

$$Z(T) = \int \left(\int \delta\big(E - H(\Gamma)\big) dE \right) \exp\left(-\frac{H(\Gamma)}{k_B T} \right) d\Gamma = \int W(E) \exp\left(-\frac{E}{k_B T} \right) dE$$

状態密度はエネルギー E を固定した系で状態 Γ がどれだけ密集しているかだったので，状態 Γ の積分をエネルギー E の積分にするときはその濃さを掛け算しないといけないといった感じです。この被積分関数はある E でピークをもちます。ちなみに，分配関数の対数をとって温度をかけたものを自由エネルギーとよびます。

$$F(T) = -k_B T \log\big(Z(T)\big)$$

ちなみに，よく $\beta = 1/(k_B T)$ と置き換えをします。

- 他方，ベイズ統計の一般理論では帰省先パラメータ u の空間上の事後分布を取り扱いたいです。これがベイズ統計におけるカノニカル分布です。$n \to \infty$ でピークは $u = 0$ です。この正規化定数が分配関数です。

$$p_u(u|X^n) \propto \exp\big(-n\beta u^{2k} + \sqrt{n}\beta u^k \xi_n(u)\big) |u^h| b(u)$$

これに対して以下が状態密度として出てきました。

$$\delta(t - nu^{2k}) |u^h| b(u) du$$

統計力学と見比べると，nu^{2k} が系のエネルギーのようです。