1　なんで分配関数を計算するのか？

第三章では、（正規化された）分配関数 $Z_n^{(0)}(\beta)$ の $n \to \infty$ 極限での振る舞いを調べるのだが、ベイズ統計の枠組みにおいてこのような量を計算する理由がわかりにくい。この資料では、分配関数の計算がなぜ必要なのかを説明する。分配関数は

$$Z_n^{(0)}(\beta) \coloneqq \int \exp(-n\beta K_n(w))\varphi(w)\, dw \tag{1}$$

と定義される。$K_n(w)$ は経験誤差関数と呼ばれる量であり、

$$K_n(w) \coloneqq -\frac{1}{n}\sum_{i=1}^{n} f(X_i, w) = -\frac{1}{n}\sum_{i=1}^{n} \ln \frac{p(X_i|w_0)}{p(X_i|w)} \tag{2}$$

である。パラメータ w が確率変数であることから、$K_n(w)$ も確率変数となる。

事後分布による期待値計算と分配関数

分配関数の振る舞いを解明することは、事後分布による積分計算の挙動を解明することと等価である。なぜなら、事後分布 $p(w|X^n)$ は

$$p(w|X^n) \propto \varphi(w)\prod_{i=1}^{n} p(X_i|w)^{\beta} \propto \varphi(w)\exp(-n\beta K_n(w)) \tag{3}$$

と変形できるためである。仮に $K_n(w)$ が $w = w_0$ で唯一の極小値を持っているとすると、期待値計算では $w = w_0$ 近傍の値が主要な寄与を持つことになる。第三章の前半では、このような予想が実際に成り立つことを示し、w_0 近傍以外の効果がどれくらいのオーダーで 0 に収束するかを調べる。

第三章の前半は、汎化損失 G_n を求めるために必要な事後分布による期待値計算について、Gauss 積分で近似できること（鞍点近似を適用できること）を示すことが目的だと思われる。$K_n(w)$ の確率的な振る舞い（すなわち $f(X, w)$ の確率的な振る舞い）を調べることで、鞍点近似の妥当性の評価を行う。

三章前半のゴール

汎化損失 G_n が $n \to \infty$ の極限で以下の振る舞いをする：

$$G_n = L(\omega_0) + \frac{1}{n}\left(\frac{d}{2\beta} + \frac{1}{2}|\xi|^2 - \frac{1}{2\beta}\operatorname{tr}\left(IJ^{-1}\right)\right) + \mathrm{o_p}\left(\frac{1}{n}\right)。 \tag{4}$$

ここで汎化損失 G_n はどのような量と結びついていたかを復習する。確率変数 $f(w)$ に対して、事後分布による期待値を

$$\langle f(w)\rangle_{w|X^n} \coloneqq \int f(w)p(w|X^n)\, dw \tag{5}$$

と定義する。このとき、汎化損失は以下のように表される：

$$G_n = -\mathbb{E}_X\left[\ln\left(\langle p(X|w)\rangle_{w|X^n}\right)\right]。 \tag{6}$$

分配関数の計算を行うことができれば、$\langle f(X,w)\rangle_{w|X^n}$ の振る舞い、すなわち $\langle p(X|w)\rangle_{w|X^n}$ の振る舞いを直接評価できる。しかし、それが対数関数の中に入っているため、G_n を直接計算することは難しそうだ。途中まで計算を進めてみると、

$$\begin{aligned} G_n &= -\mathbb{E}_X\Big[\ln\Big(\langle p(X|w)\rangle_{w|X^n}\Big)\Big] = -\mathbb{E}_X\left[\ln p(X|w_0) - \ln\left\langle\frac{p(X|w_0)}{p(X|w)}\right\rangle_{w|X^n}\right] \\ &= -\mathbb{E}_X[\ln(p(X|w_0))] + \mathbb{E}_X\left[\ln\left\langle e^{f(X,w)}\right\rangle_{w|X^n}\right] \end{aligned} \tag{7}$$

という形になる。右辺第一項は $L(w_0)$ そのものである。第二項の計算が難しいのだが、ログをなくすためには Taylor 展開 $\ln(1+x) = x - x^2/2 + \cdots$ を適用すれば良いことに気づく。確率変数 X に対して、X^2 までの項の寄与を考えると

$$\ln\left\langle e^X\right\rangle = \ln\left\langle 1 + X + \frac{X^2}{2}\right\rangle = \left\langle X + \frac{X^2}{2}\right\rangle - \frac{1}{2}\left\langle X + \frac{X^2}{2}\right\rangle^2 = \langle X\rangle + \frac{1}{2}\Big(\left\langle X^2\right\rangle - \langle X\rangle^2\Big) \tag{8}$$

が成り立つので、

$$\ln\left\langle e^{f(X,w)}\right\rangle_{w|X^n} = \langle f(X,w)\rangle_{w|X^n} + \frac{1}{2}\mathbb{V}[f(X,w)] \tag{9}$$

という形になる。$\mathbb{V}[f(X,w)]$ は分散である。次章で示すのだが、$G_n = -\mathcal{G}_n(1)$ の関係式を用いた計算は、上記の Taylor 展開を用いた計算と全く同じ結果を得るらしい。その結果を先に述べておく：

$$G_n = L(w_0) + \frac{1}{n}\left(\langle nK(w)\rangle_{w|X^n} - \frac{1}{2}\mathbb{E}_X[\mathbb{V}[nf(X,w)]]\right) + \mathrm{o_p}\left(\frac{1}{n}\right)。 \tag{10}$$

ここで

$$K(w) = \mathbb{E}_X[f(X,w)], \qquad \mathbb{V}[f(X,w)] = \left\langle f(X,w)^2\right\rangle_{w|X^n} - \langle f(X,w)\rangle^2_{w|X^n} \tag{11}$$

であるため、$\mathcal{G}_n(1)$ を用いた計算は式 (9) と同じ形（なぜか分散の符号が逆）になっている。さて、ここまで計算できれば、あとは $f(X,w)$ の確率的な振る舞いがどのようなものかを解析すれば良い。

2 Taylor 展開の正当化

上で行った Taylor 展開は、第二章で得られた結果を再現しているようである。第二章で議論したように、とある性質の良い関数 $\mathcal{G}_n(\alpha)$ を用いて

$$G_n = -\mathcal{G}_n(1) \tag{12}$$

と計算できることがわかっている。しかし、$\mathcal{G}_n(1)$ の値を直接計算することは難しい。そこで、$\alpha = 0$ 周りの Taylor 展開を考えると、一般の α に対して

$$\mathcal{G}_n(\alpha) = \mathcal{G}_n(0) + \sum_{k=1}^{\infty}\frac{1}{n!}\mathcal{G}_n^{(k)}(0)(\alpha-0)^k \tag{13}$$

と表される（と $\mathcal{G}$ に仮定を置いている）。ただし、$\mathcal{G}_n^{(k)}$ は $\mathcal{G}_n$ の α による k 階微分である。また、$\mathcal{G}_n(0) = 0$ となることは簡単に示せる。正則理論の枠組み[*1] を考えている範囲では、高次の微分項がほとんどゼロにな

[*1] 事後分布が正規分布で近似できるという特別な場合の理論であり、具体的には 3 つの仮定を置くことで成り立つ：
1. 真の分布に対して最適なパラメータが一つである。すなわち、$W_0 = \{w_0\} \subset W$ である。
2. $L(w)$ は、その Hessian が $w = w_0$ で正定値行列となる。
3. サンプルの数（データセットの数）n が非常に大きい。非常に大きいとは、確率過程 $\eta_n(w)$ が法則収束しているとみなせる程度の大きさである。

ることが示される：

$$\forall k \geq 3, \quad \left|\mathcal{G}_n^{(k)}(0)\right| = \mathrm{o_p}\left(\frac{1}{n}\right) 。 \tag{14}$$

したがって、

$$G_n = -\mathcal{G}_n(1) = -\mathcal{G}_n^{(0)}(0) - \frac{1}{2}\mathcal{G}_n^{(1)}(0) + \mathrm{o_p}\left(\frac{1}{n}\right) 。 \tag{15}$$

という関係式を用いて G_n を求められる。さらに任意の k について、$\mathcal{G}_n^{(k)}(0)$ は

$$l_k(X) \coloneqq \left\langle (\ln p(X|w))^k \right\rangle_{w|X^n} = (-1)^k \left\langle (f(X,w) - \ln p(X|w_0))^k \right\rangle_{w|X^n} \tag{16}$$

で定義される $l_k(X)$ を用いて計算可能である。これを用いると、

$$\mathcal{G}_n^{(1)}(0) = \mathbb{E}_X\left[\langle f(X,w)\rangle_{w|X^n}\right] \tag{17}$$

$$\mathcal{G}_n^{(2)}(0) = \mathbb{E}_X[\mathbb{V}[f(X,w)]] \tag{18}$$

$$\tag{19}$$

となる。したがって式 (15) の右辺を計算すると、

$$G_n = L(w_0) + \frac{1}{n}\left(\langle nK(w)\rangle_{w|X^n} - \frac{1}{2}\mathbb{E}_X[\mathbb{V}[nf(X,w)]]\right) + \mathrm{o_p}\left(\frac{1}{n}\right) \tag{20}$$

という結果を得る。これは式 (9) と（分散の符号を除いて）同じである。

3　分配関数の計算

G_n を計算するために最後に必要なことは、$f(X,w)$ の確率的な振る舞いを明らかにすることである。そのためには、事後分布による積分を楽に評価できるようにする必要がある。

$$p(w|X^n) \propto \varphi(w)\exp(-n\beta K_n(w)) \tag{21}$$

であったため、分配関数

$$Z_n^{(0)}(\beta) = \int \exp(-n\beta K_n(w))\varphi(w)\,dw \tag{22}$$

の性質を明らかにすれば、期待値計算を評価できると考えられる。積分をサボりたいため、被積分関数の寄与が最も大きい部分、すなわち

$$\omega^* = \arg\min_w K_n(w) = \arg\min_w L(w) = w_0 \tag{23}$$

の ϵ- 近傍を考えてみよう。正則理論の枠組みでは、

$$\nabla^2 K_n(w)\big|_{w=w_0} \tag{24}$$

が正定値行列という仮定を入れていた。したがって図 1 のような状況になっており、w_0 の ϵ- 近傍が最も積分に寄与する。実際に、ϵ- 近傍の外側の寄与はほとんどないことが示される。さらに、ϵ- 近傍の内側では、被積分関数を Gaussian として扱えることが示されるため、$\mathbb{R}^d$ 全体の積分に置き換えて、解析的に計算できるようになる。これが分配関数の計算ストーリーである。

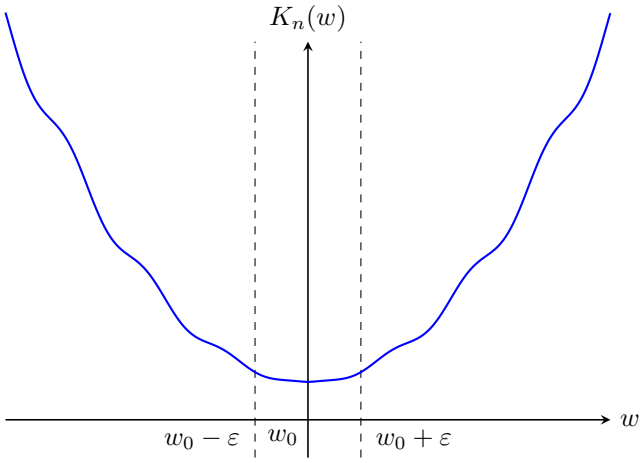

図 1　関数 $K_n(w)$ とその極小点 w_0 の ε- 近傍

4　分配関数の計算のイメージ

積分を一番サボる方法は、被積分関数が最も大きい場所に対する短冊近似（抜き出す区間を $\Delta_n(w_0)$ とする。この区間は関数のピーキー性を見ながら調節する必要があるため、n に依存して良い）を実行することである：

$$Z_n^{(0)}(\beta) = \int \exp(-n\beta K_n(w))\varphi(w)\,dw \approx \underbrace{\exp(-n\beta K_n(w_0))}_{=1}\varphi(w_0)\Delta_n(w_0) \tag{25}$$

しかしこれでは精度が悪すぎて、n に対する振る舞いの情報が落ちてしまっている。$n \to \infty$ に近づくにつれてどのような挙動をするか見たいので、少しだけ w_0 周りの寄与を足すことにする。これが第三章で行っている鞍点近似のお気持ちである。その結果として、

$$Z_n^{(0)}(\beta) \approx \exp\left(\frac{\beta}{2}|\xi_n|^2\right)\varphi(w_0)\underbrace{\left(\frac{2\pi}{n\beta}\right)^{d/2}\det\{J\}^{-1/2}}_{=\Delta_n(w_0)} \tag{26}$$

という形になる。確率変数 ξ_n は $n \to \infty$ で 0 に確率収束するような性質を持っている（これが大事なのに地味に示せない。。。）ため、

$$Z_n^{(0)}(\beta) \approx \left[1 + \underbrace{\frac{\beta}{2}|\xi_n|^2 + \cdots}_{\to 0\,(n\to\infty)}\right]\varphi(w_0)\Delta_n(w_0) \tag{27}$$

となり、$n \to \infty$ の極限で短冊近似の結果を再現する。

5　第三章の主要な結果

第三章で得られる最も主要な主張はおそらく、正則理論の仮定のもとでは事後分布が正規分布で近似できる（正規分布に法則収束する）ということである。なぜなら、ベイズ理論の目的は G_n を計算することであり、G_n は事後分布の期待値計算を用いて計算できるからである。

事後分布が正規分布に法則収束する

正則理論の仮定の元では、事後分布 $p(w|X^n)$ は、$J = \nabla^2 L(w_0)$ を用いて、平均 ω_0、分散 $(nJ)^{-1}$ の正規分布に法則収束する：

$$p(w|X^n) \xrightarrow[n\to\infty]{\mathrm{d}} \mathcal{N}(\omega_0, (nJ)^{-1}) = \frac{1}{\sqrt{(2\pi)^d \det\{(nJ)^{-1}\}}} \exp\left(\frac{1}{2}(\omega-\omega_0)^\top nJ(\omega-\omega_0)\right)。 \quad (28)$$

この結果に関連する議論をしているのが 3.4 節までである。3.5 節では、ベイズ推測以外の方法（事後確率最大化法、最尤推定法、平均プラグイン法）を用いた場合の $f(X, w)$ の振る舞いを調べているようだった。3.6 節では、事後分布が正規分布で近似できるという仮定の元で、サンプルだけから汎化損失 G_n を数値的に推定する方法を示している。

サンプルだけから汎化損失を推定する

汎化損失はその定義から、サンプルだけでは計算することができないが、サンプルだけから計算可能な正則の情報量基準を

$$\mathrm{RIC} := T_n + \frac{1}{n}\mathrm{tr}\big(IJ^{-1}\big) \quad (29)$$

と定義すると、$G_n - L(w_0)$ と $\mathrm{RIC} - L_n(w_0)$ は平均と分散が漸近的に一致することが示される。後者の値はサンプルだけから計算できるため、汎化損失の推定値として用いることができる。（たぶん）