PRML 10.4 - 10.6

宮澤　彬

総合研究大学院大学 博士前期

miyazawa-a@nii.ac.jp

August 7, 2015
(modified: September 9, 2015)

はじめに

- このスライドの LuaLaTeX のソースコードは `https://github.com/pecorarista/documents` にあります.
- 教科書とは若干異なる表記をしている場合があります.
- 10.4.1 と 10.7 は間に合いませんでした.

潜在変数とパラメータ

今までモデルの中で観測値 (observed variable) と隠れ変数 (hidden variable) を区別してきた．これからは更に以下のような区別を導入する．

- **潜在変数** (latent variable) Z
  観測値集合の大きさに従って数が増える（**外延的変数**）
  例：ガウス混合モデルのインジケータ変数 z_{kn}
- **パラメータ** (parameter) θ
  観測値集合の大きさに関わらず数が固定（**内包的変数**）
  例：ガウス混合モデルの平均 μ_k，精度 Λ_k，混合比 π_k

指数型分布族

独立に同分布に従うデータの集合 $X := \{x_1, \ldots, x_N\}$ とそれに対応する潜在変数の集合 $Z := \{z_1, \ldots, z_N\}$ があるとする. これらの同時分布が自然パラメータ η を使った以下の指数型分布族で表せるとする.

$$p(X, Z|\eta) = \prod_{n=1}^{N} h(x_n, z_n) g(\eta) \exp\left(\eta' u(x_n, z_n)\right). \tag{10.113}$$

また η は共役事前分布

$$p(\eta|\nu_0, \chi_0) = f(\nu_0, \chi_0) g(\eta)^{\nu_0} \exp(\nu_0 \eta' \chi_0)$$

に従うものとする.

指数型分布族について復習

第 2.4 節で指数型分布族とその共役事前分布について学んだ．次のような形をした指数型確率分布

$$p\left(x|\eta\right) = h\left(x\right) g\left(\eta\right) \exp\left(\eta' u\left(x\right)\right) \tag{2.194}$$

について

$$p\left(\eta|\chi,\nu\right) = f\left(\chi,\nu\right) g\left(\eta\right)^{\nu} \exp\left(\nu\eta'\chi\right) \tag{2.229}$$

という形の共役事前分布が存在する．データ $\mathrm{X} = \{x_1, \ldots, x_n\}$ が与えられたとき，尤度は

$$p\left(X|\eta\right) = \left(\prod_{n=1}^{N} h\left(x_n\right)\right) g\left(\eta\right)^{N} \exp\left(\eta' \sum_{n=1}^{N} u\left(x_n\right)\right) \tag{2.227}$$

となる．

指数型分布族について復習

事後分布は

$$p(\eta|X,\chi,\nu) \propto p(X|\eta)\,p(\eta|\chi,\nu)$$
$$\propto g(\eta)^{\nu+N} \exp\left(\eta' \left(\sum_{n=1}^{N} u(x_n) + \nu\chi\right)\right) \qquad (2.230)$$

と計算できる．この式から，事前分布のパラメータ ν は，有効な事前の仮想観測値の数と解釈できる．ただし，仮想観測値では，十分統計量 $u(x)$ の代わりに，χ が与えられる．

指数型分布族の変分近似

指数型分布族を変分分布近似することを考える．これまでと同様に，周辺分布の対数 $\log p(X)$ を

$$\log p(X) = \mathcal{L}(q) + \mathrm{KL}(q\|p),$$

$$\mathcal{L}(q) = \int q(Z,\eta) \log \left(\frac{p(X,Z|\eta)\, p(\eta)}{q(Z,\eta)} \right) (d\eta dZ),$$

$$\mathrm{KL}(q\|p) = -\int q(Z,\eta) \log \left(\frac{p(Z|X,\eta)\, p(\eta)}{q(Z,\eta)} \right) (d\eta dZ)$$

と分解し，$\mathcal{L}$ を q について最大化する．

KL タイバージェンスの最小化

f と g を確率密度関数とする．区間 $(0, \infty)$ において $\log x \leq x - 1$ が成り立つ．ここで $x := f/g$ とすると $f \log g - f \log f \leq g - f$ となる．積分の線型性と単調性から

$$\int_X f \log g d\mu - \int_X f \log f d\mu \leq \int_X g d\mu - \int_X f d\mu$$

が成り立つ．確率密度関数の性質から $\int_X f d\mu = \int_X g d\mu = 1$ なので

$$\int_X f \log g d\mu \leq \int_X f \log f d\mu$$

を得る．ほとんど確実に $f = g$ のとき上式の等号が成り立つ．また等号が成り立つのはそのときに限る．

KL タイバージェンスの最小化の補足

$[0, \infty]$ 値 $\mathfrak{M}$-可測関数 f について $\int_X f(x)\,\mu(dx) = 0$ が成り立つならば，$\mu(\{x \in X\,;\, f(x) \neq 0\}) = 0$ が成り立つ．

証明 $E_n := \{x \in X\,;\, f(x) \geq 1/n\}$ とおくと

$$\bigcup_{n=1}^{\infty} E_n = \{x \in X\,;\, f(x) > 0\}$$

である．仮定 $\int_X f(x)\,\mu(dx) = 0$ と $f(x) \geq 0$ から，任意の n について

$$0 = \int_X f(x)\,\mu(dx) \geq \int_{E_n} f(x)\,\mu(dx) \geq \frac{1}{n}\mu(E_n) \geq 0$$

となる．すなわち各 n で $\mu(E_n) = 0$ が成り立つ．ゆえに測度の性質から

$$\mu\left(\bigcup_{n=1}^{\infty} E_n\right) \leq \sum_{n=1}^{\infty} \mu(E_n) = 0$$

であることが分かる． ∎

指数型分布族の変分近似

$\mathcal{L}$ の最大化に戻る.

計算を進めるため，変分分布が潜在変数とパラメータで分けられる，すなわち $q(Z, \eta) = q(Z)\, q(\eta)$ と分解できると仮定する.

$$
\begin{aligned}
\mathcal{L}(q) &= \int q(Z, \eta) \log \left(p(X, Z|\eta)\, p(\eta) \right) (d\eta dZ) \\
&\quad - \int q(Z, \eta) \log q(Z, \eta)\, (d\eta dZ) \\
&= \int q(Z) \left(\int q(\eta) \log p(X, Z|\eta)\, d\eta \right) dZ \\
&\quad - \int q(Z) \log q(Z)\, dZ \\
&\quad - \int q(\eta) \left(\log q(\eta) - \log p(\eta) \right) d\eta
\end{aligned}
$$

指数型分布族の変分近似

したがって最適な $q(Z)$ は

$$\begin{aligned}\log q^{\star}(Z) &= \int q(\eta) \log p(X, Z|\eta)\, d\eta + \text{const.} \\ &= \mathbb{E}_{\eta}\left[\log p(X, Z|\eta)\right] + \text{const.} \\ &= \sum_{n=1}^{N} \left(\log h(x_n, z_n) + \mathbb{E}\left[\eta'\right] u(x_n, z_n)\right) + \text{const.} \quad (10.115)\end{aligned}$$

を満たさなければならない．この式の右辺に注目すると，各 n ごとに独立な項の和に分解できるので $q^{\star}(Z) = \prod_{n=1}^{N} q^{\star}(z_n)$ となる．よって指数をとって

$$q^{\star}(z_n) = h(x_n, z_n)\, g(\mathbb{E}[\eta]) \exp\left(\mathbb{E}\left[\eta'\right] u(x_n, z_n)\right) \quad (10.116)$$

を得る．ただし $g(\mathbb{E}[\eta])$ は正則化のため，指数型分布族の標準的な形に合わせて付加したものである．

指数型分布族の変分近似

次に $q(\eta)$ について最大化する．$\mathcal{L}$ は

$$\begin{aligned}\mathcal{L}(q) = &\int q(\eta)\left(\int q(Z)\log p(X,Z|\eta)\,dZ\right)d\eta \\ &- \int q(Z)\log q(Z)\,dZ \\ &- \int q(\eta)\left(\log q(\eta) - \log p(\eta)\right)d\eta\end{aligned}$$

と表せるので

$$\begin{aligned}\log q^{\star}(\eta) &= \log p(\eta|\nu_0,\chi_0) + \mathbb{E}_Z\left[\log p(X,Z|\eta)\right] + \text{const.} \qquad (10.117)\\ &= \nu_0 \log g(\eta) + \nu_0\eta'\chi_0 \\ &\quad + \sum_{n=1}^{N}\left(\log g(\eta) + \eta'\mathbb{E}_{z_n}\left[u(x_n,z_n)\right]\right) + \text{const.} \qquad (10.118)\end{aligned}$$

となる．

指数型分布族の変分近似

指数をとって

$$q^{\star}(\eta) = f(\nu_N, \chi_N)\, g(\eta)^{\nu_N} \exp\left(\nu_N \eta' \chi_N\right) \tag{10.119}$$

を得る．ただし，

$$\nu_N = \nu_0 + N \tag{10.120}$$

$$\nu_N \chi_N = \nu_0 \chi_0 + \sum_{n=1}^{N} \mathbb{E}_{z_n}\left[u(x_n, z_n)\right] \tag{10.121}$$

とした．

指数型分布族の変分近似

$q^{\star}(z_n)$ と $q^{\star}(\eta)$ の解には相互に依存関係があるので，二段階の繰り返しで解く．

変分 E ステップ

$$q(z_n) \leftarrow h(x_n, z_n)\, g(\mathbb{E}[\eta]) \exp\left(\mathbb{E}[\eta']\, u(x_n, z_n)\right)$$

$$q(\eta) \leftarrow f(\nu_N, \chi_N)\, g(\eta)^{\nu_N} \exp\left(\nu_N \eta' \chi_N\right)$$

where

$$\mu_n \leftarrow \mathbb{E}_{z_n}[u(x_n, z_n)] = \int q(z_n)\, u(x_n, z_n)\, dz_n$$

$$\nu_N \chi_N \leftarrow \nu_0 \chi_0 + \sum_{n=1}^{N} \mu_n$$

変分 M ステップ

$$\eta \leftarrow \mathbb{E}[\eta] = \int q(\eta)\, \eta d\eta$$

局所的変分推論

10.1 節や 10.2 節では，事後分布の近似を直接求めた.

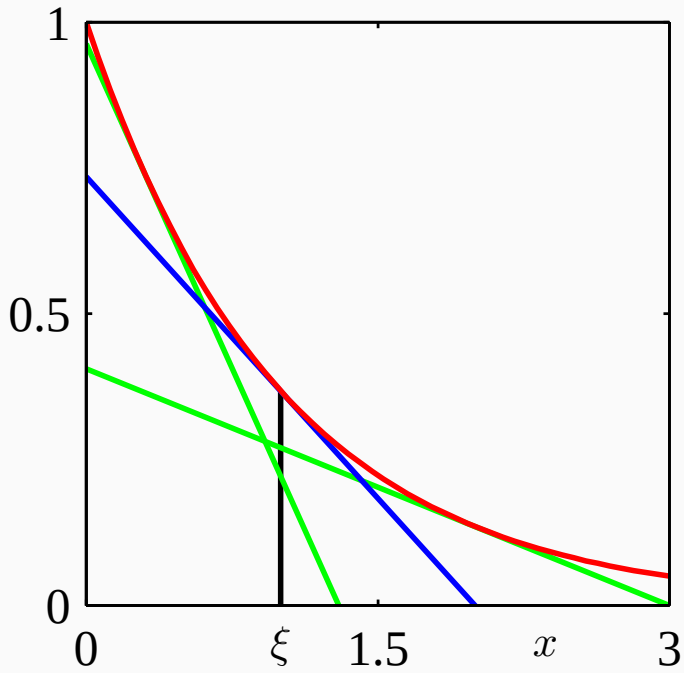
10.5 節と 10.6 節では各変数の上からあるいは下から近似を使う方法を学ぶ.

凸関数の下限の表示

関数 $f(x) = \exp(-x)$ の点 $(\xi, f(\xi))$ における接線の方程式を求めると

$$y(x) = -\exp(-\xi)(x - \xi) + \exp(-\xi) \tag{10.126}$$

である.



凸関数の下限の表示

ここで $\eta := -\exp(-\xi)$ とすると $\xi = -\log(-\eta)$ なので

$$y(x, \eta) = \eta x - \eta + \eta \log(-\eta) \tag{10.127}$$

である．凸関数の性質から，f の接線はグラフの下にくるので [1]，

$$f(x) = \max_{\eta} \{\eta x - \eta + \eta \log(-\eta)\} \tag{10.128}$$

と表せる．

---

1 任意の $\lambda \in (0, 1)$ をとると

$$f(\lambda x + (1-\lambda)\xi) \leq \lambda f(x) + (1-\lambda) f(\xi)$$

$$\frac{f(\lambda x + (1-\lambda)\xi) - f(\xi)}{\lambda} \leq f(x) - f(\xi)$$

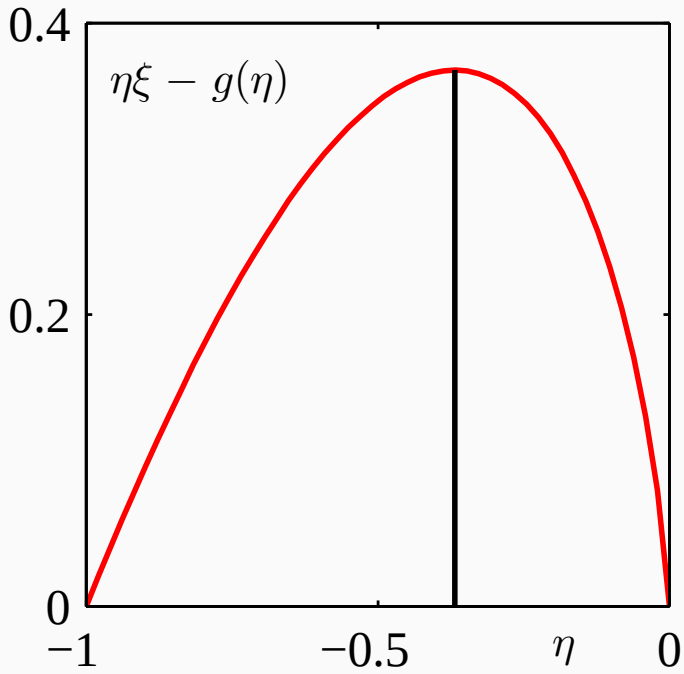
が成り立つ．よって $\lambda \to +0$ として

$$\nabla f(x) \cdot (x - \xi) \leq f(x) - f(\xi)$$

を得る．

凸関数の下限の表示

新たなパラメータ η を導入するという代償は払ったが，一次関数による下から近似 $y(x, \eta)$ を得た.



これは一般に**凸双対性**の理論で説明されることである.

共役関数

X を Banach 空間とし，X^* を X の共役空間[2] とする．このとき関数 $f : X \to [-\infty, +\infty]$ に対し，$f^* : X^* \to [-\infty, +\infty]$ を

$$f^*(\phi) := \sup\{\phi(x) - f(x) \ ; \ x \in X\}$$

で定め，f の**共役関数** (conjugate function) や **Fenchel 双対** (Fenchel conjugate)[3] などと呼ばれる．

また f の**双共役関数** (biconjugate function) $f^{**} : X \to [-\infty, +\infty]$ を

$$f^{**}(x) := \sup\{\phi(x) - f^*(\phi) \ ; \ \phi \in X^*\}$$

と呼ぶ．

2 X 上の有界線型汎関数全体が作る線型空間．
3 Werner Fenchel /(de) ˈvɛʁnɐ ˈfɛnçəl/ (1905 - 1988)

共役関数

共役関数は非凸関数についても考えることができるが，多くの場合は proper[4] な凸関数の共役関数を考える．

凹関数については共役関数の定義において $\sup$ を $\inf$ に置き換えた**凹共役関数** (concave conjugate function) を考える．

$$f_* (\phi) := \inf \{\phi (x) - f (x) \ ; \ x \in X\} .$$

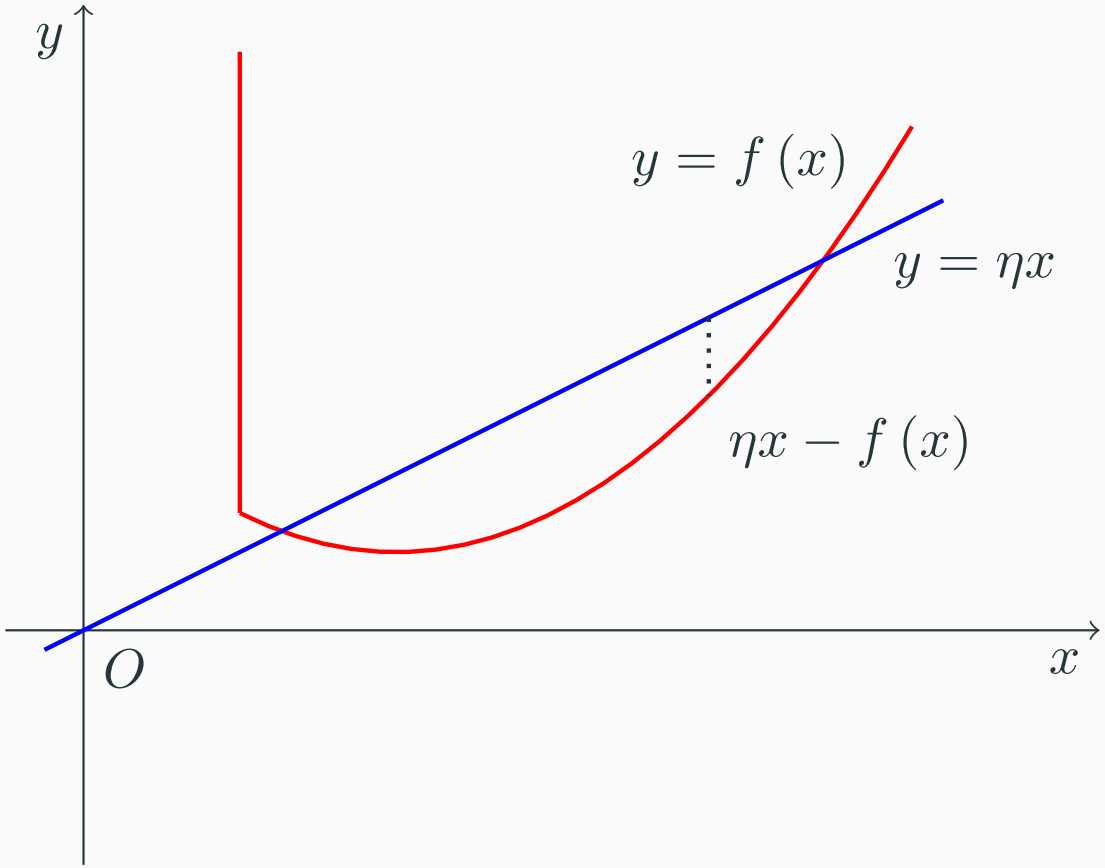
凹双共役関数 (concave biconjugate function) も同様に考えられる．

$$f_{**} (x) := \inf \{\phi (x) - f_* (\phi) \ ; \ \phi \in X^*\} .$$

4 関数 $f : X \to (-\infty, \infty]$ が **proper** であるとは $\operatorname{dom} f = \{x \in X \ ; \ f (x) < \infty\}$ が空でないということである．つまり ∞ にべったり張り付かないということ．

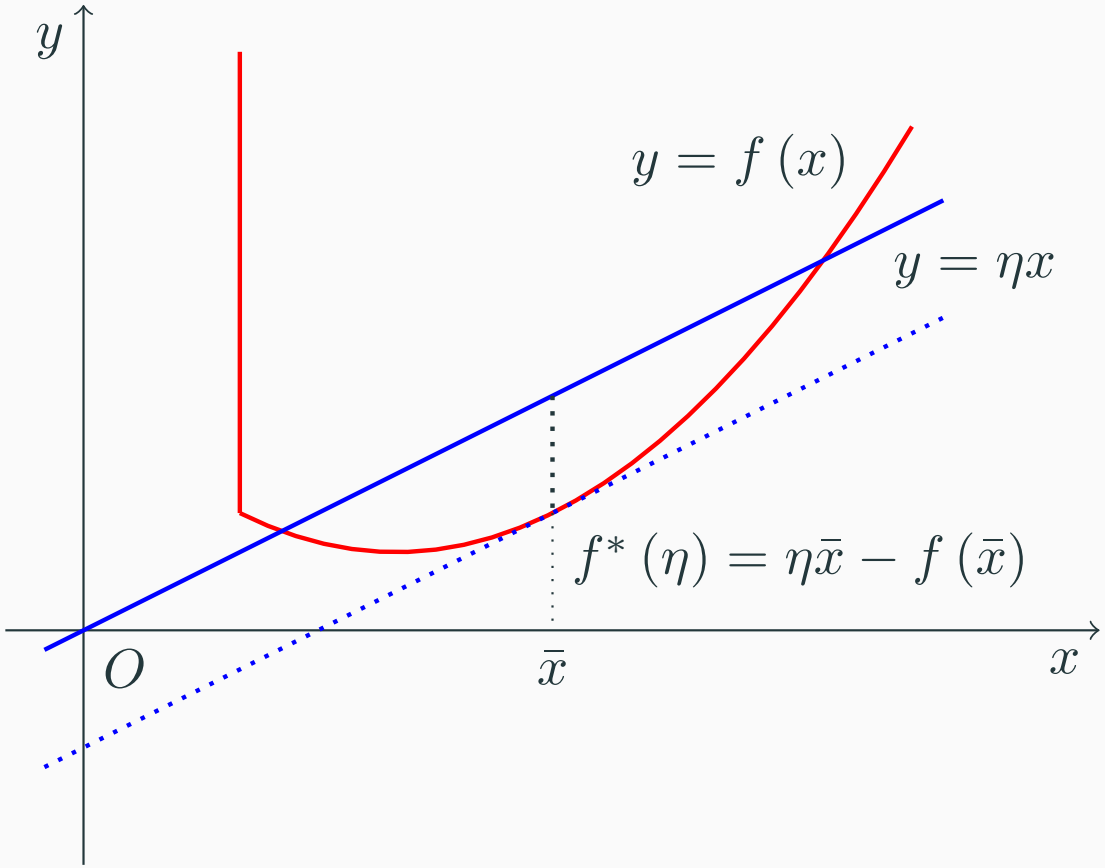
共役関数

$$f^{*}(\eta) = \sup\{\eta x - f(x) \ ; \ x \in \mathbb{R}\}$$



共役関数

$$f^*(\eta) = \sup\{\eta x - f(x) \ ; \ x \in \mathbb{R}\}$$



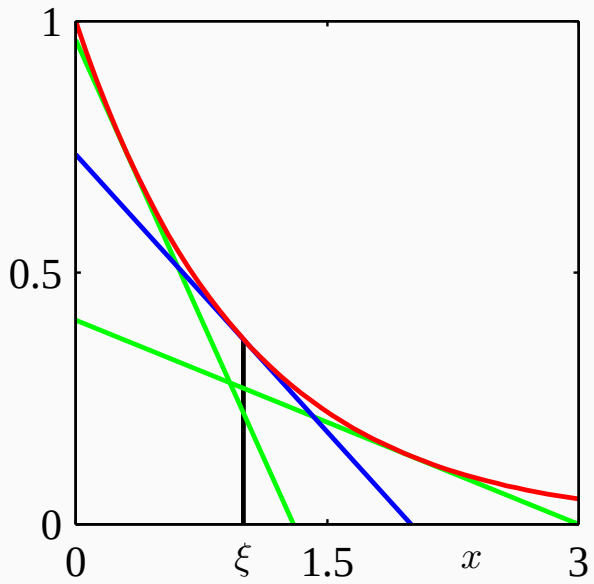
共役関数

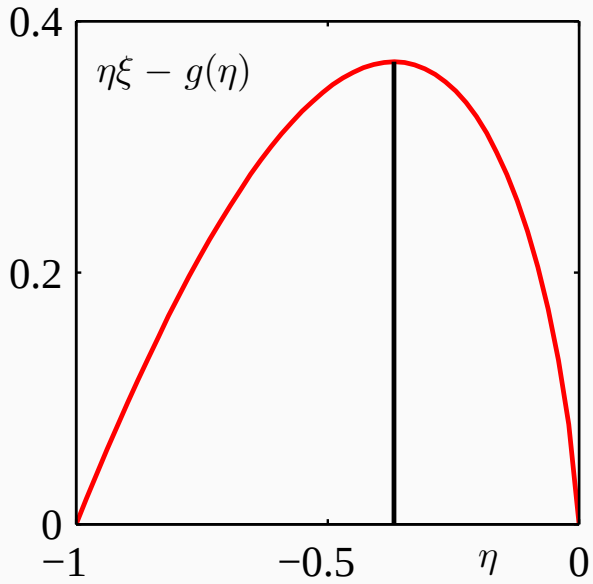
例として $f(x) := \exp(-x)$ の共役関数 $f^*(\eta) = \sup_x \{\eta x - f(x)\}$ を陽に求めてみる．$\{\cdot\}$ の中を x で微分して 0 と置くと

$$\eta + \exp(-x) = 0$$

$$x = -\log(-\eta)$$

となる．したがって $f^*(x) = \eta - \eta \log(-\eta)$ を得る．





共役関数

次に双共役関数 $f^{**}(x) = \sup_\eta \{\eta x - f^*(\eta)\}$ を陽に求めてみる．$\{\cdot\}$ の中を η で微分して 0 と置くと

$$\frac{d}{d\eta}(\eta x - \eta + \eta \log(-\eta)) = 0$$

$$x - 1 + \log(-\eta) + 1 = 0$$

$$\eta = -\exp(-x)$$

となる．ゆえに

$$\begin{aligned} f^{**}(x) &= -\exp(-x)\, x + \exp(-x) + \exp(-x)(-x) \\ &= \exp(-x) \\ &= f(x) \end{aligned}$$

を得る．つまり $f = f^{**}$ が成り立つ．これは Fenchel 変換によって情報が失われていないということである．

一般に f を proper で下半連続な凸関数とすると $f = f^{**}$ が成り立つ．

ロジスティックシグモイド関数

次にパターン認識でよく現れるロジスティックシグモイドの共役関数を求める.

$$\sigma(x) = \frac{1}{1+\exp(-x)} \tag{10.134}$$

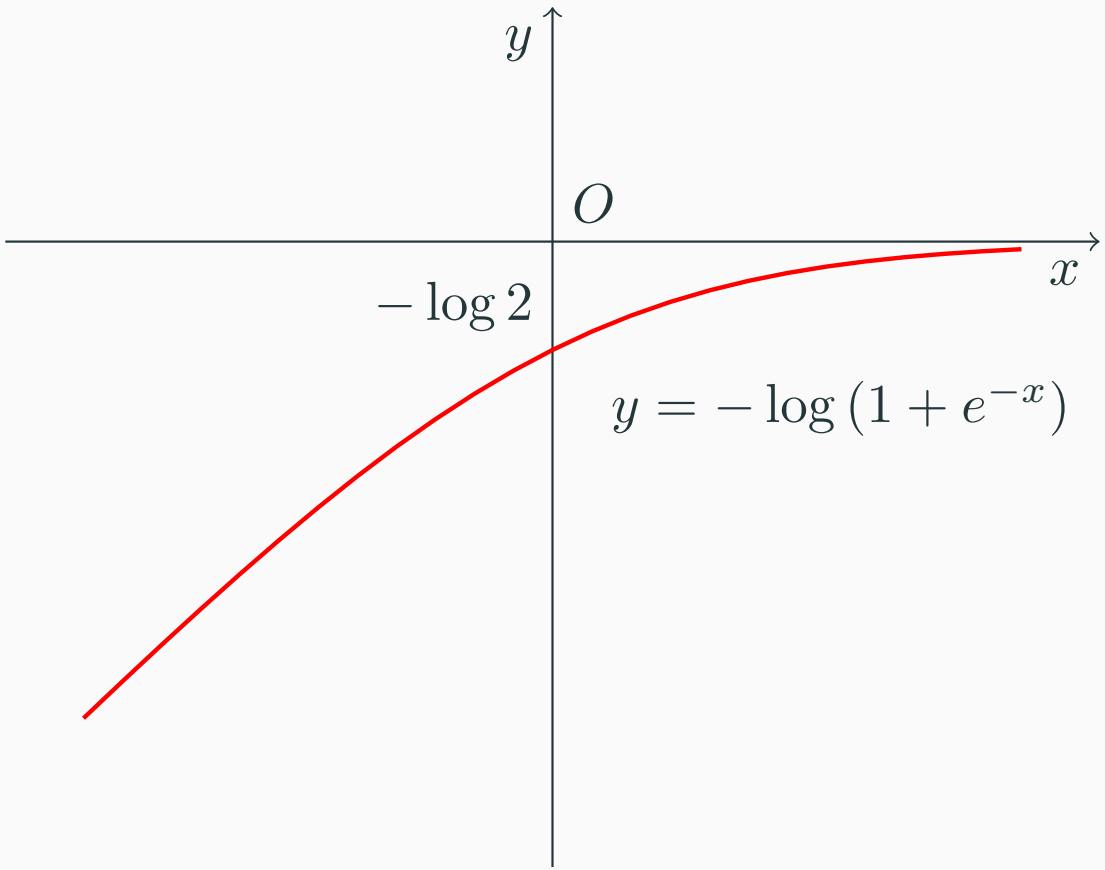
これは凸でも凹でもないが，対数をとると凹になる．$u(x) := \log \sigma(x)$ とする.

$$u(x) = -\log\left(1+e^{-x}\right)$$

$$\frac{d}{dx}u(x) = -\frac{-e^{-x}}{1+e^{-x}} = \frac{1}{e^{x}+1} > 0$$

$$\frac{d^2}{dx^2}u(x) = -\frac{e^{x}}{(e^{x}+1)^2} < 0$$

共役関数



ロジスティックシグモイド関数の上からの評価

関数 u は凹なので，凹共役 $u_*(\eta) = \inf_x \{\eta x - u(x)\}$ を考える．$\{\cdot\}$ の中を x で微分して 0 と置くと

$$\eta = \frac{1}{e^x + 1}$$

$$x = \log(1-\eta) - \log\eta$$

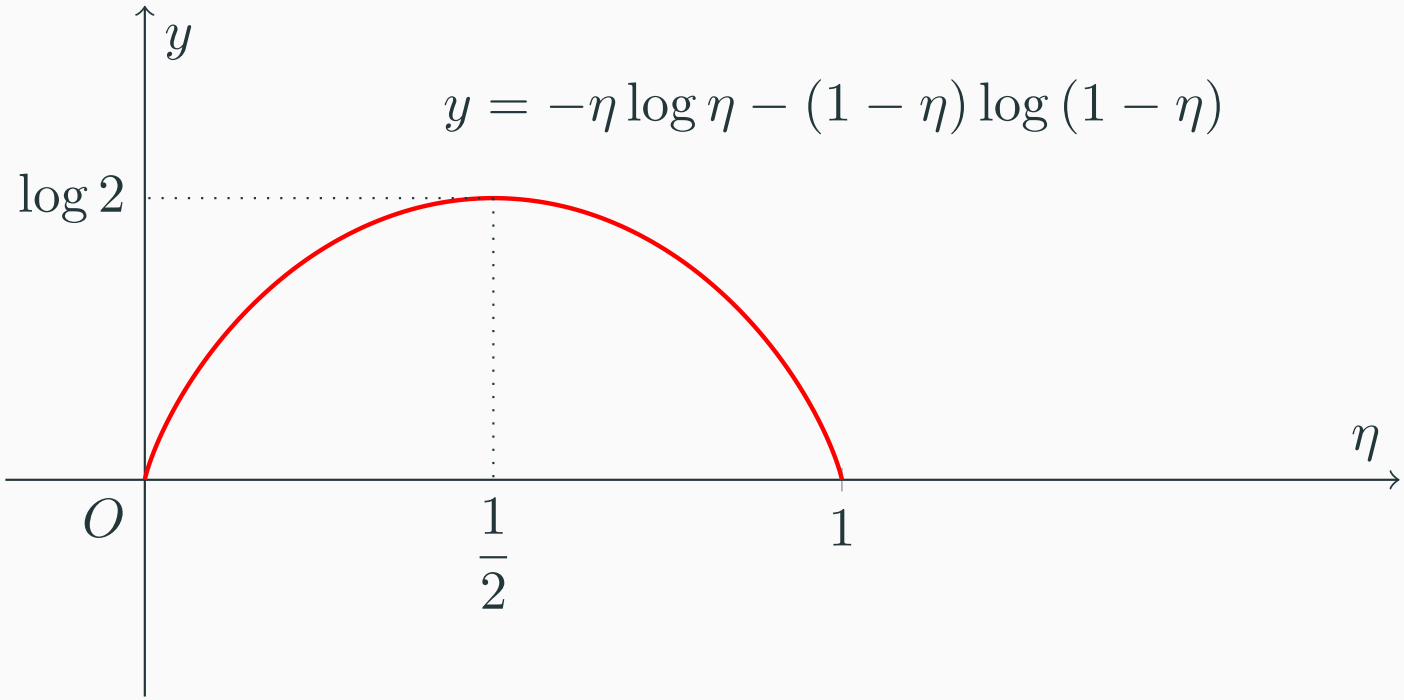
となる．したがって

$$u_*(\eta) = \eta\log(1-\eta) - \eta\log\eta + \log\left(1 + \frac{\eta}{1-\eta}\right)$$

$$= -\eta\log\eta - (1-\eta)\log(1-\eta)$$

を得る．これは 2 値エントロピー関数になっている．

２値エントロピー関数



ロジスティックシグモイド関数の上からの評価

凹共役 $u_*(\eta)$ を使えば $u(x) = \log \sigma(x)$ を

$$\log \sigma(x) \leq \eta x - u_*(\eta) \tag{10.136}$$

と上から抑えられる．指数をとって

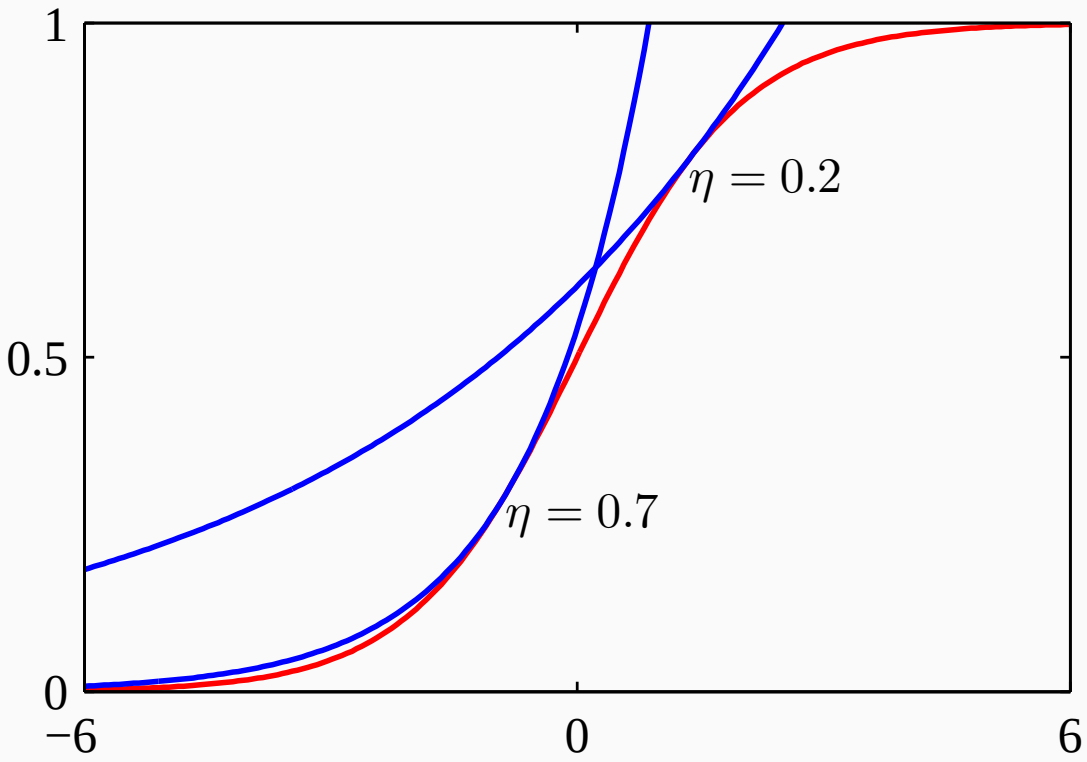
$$\sigma(x) \leq \exp(\eta x - u_*(\eta)) \tag{10.136}$$

となる．

ロジスティックシグモイド関数の上からの評価

今までの計算によって，私達はロジスティックシグモイド関数の上からの評価を得た.

$$\sigma(x) \leq \exp(\eta x + \eta \log \eta + (1-\eta) \log(1-\eta))$$



ロジスティックシグモイド関数の下からの評価

下からの評価を得るため，まず u を少し変形しておく．

$$\begin{aligned} u(x) &= -\log\left(1+e^{-x}\right) \\ &= -\log\left(e^{-x/2}\left(e^{x/2}+e^{-x/2}\right)\right) \\ &= \frac{x}{2} - \log\left(2\cosh\frac{x}{2}\right) \end{aligned}$$

右辺の第 2 項に着目する．ここで $v := x^2$,
$f(v) := -\log\left(2\cosh\left(\sqrt{v}/2\right)\right)$ と置いて解析を進める．

ロジスティックシグモイド関数の下からの評価

いま定義した $f(v) = -\log(2\cosh(\sqrt{v}/2))$ は凸関数である.

$$\begin{aligned}\frac{df}{dv}(v) &= -\frac{2\sinh(\sqrt{v}/2)}{2\cosh(\sqrt{v}/2)}\frac{1}{4\sqrt{v}} \\ &= -\frac{1}{4\sqrt{v}}\tanh\frac{\sqrt{v}}{2}\end{aligned}$$

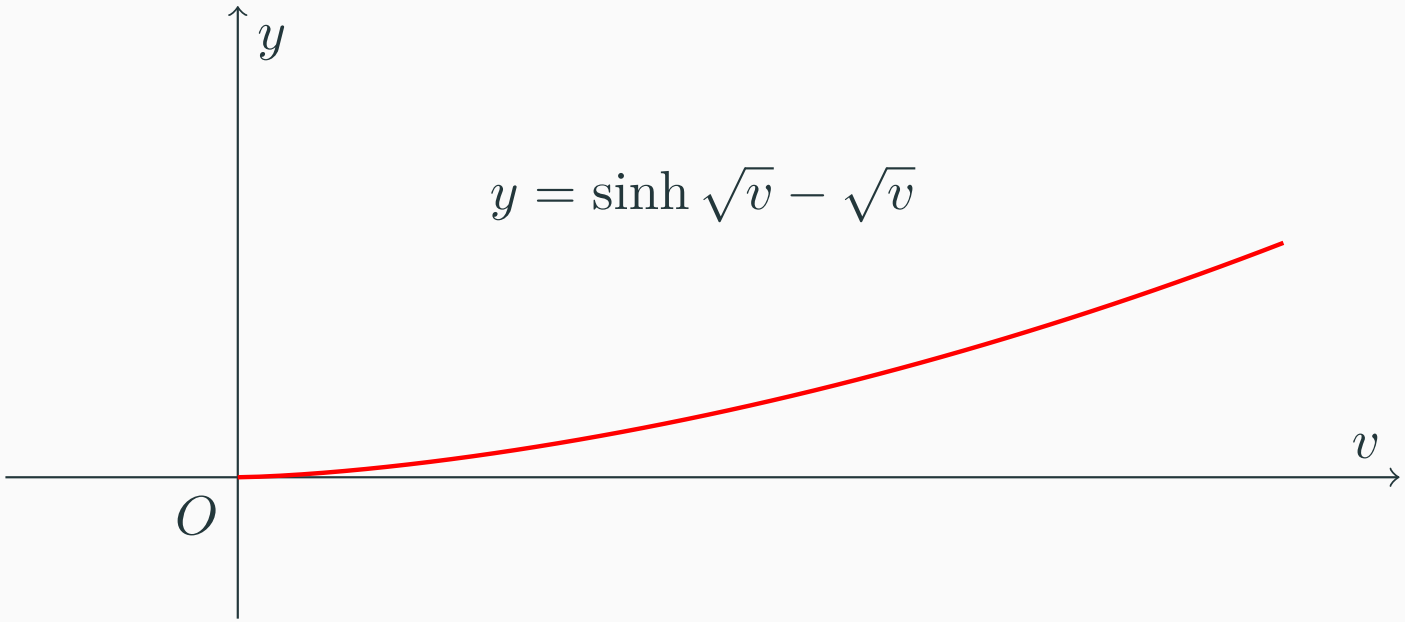
$$\begin{aligned}\frac{d^2f}{dv^2}(v) &= -\frac{1}{8v\sqrt{v}}\tanh\frac{\sqrt{v}}{2} + \frac{1}{4\sqrt{v}}\frac{1}{\cosh^2(\sqrt{v}/2)}\frac{1}{4\sqrt{v}} \\ &= \frac{1}{16v\sqrt{v}\cosh^2(\sqrt{v}/2)}\left(2\sinh\frac{\sqrt{v}}{2}\cosh\frac{\sqrt{v}}{2} - \sqrt{v}\right) \\ &= \frac{\sinh\sqrt{v} - \sqrt{v}}{16v\sqrt{v}\cosh^2(\sqrt{v}/2)} \geq 0\end{aligned}$$

ロジスティックシグモイド関数の下からの評価

念のため $g(v) := \sinh\sqrt{v} - \sqrt{v}$ が $v \geq 0$ の範囲で非負であることを確認する．導関数は

$$\frac{dg}{dv}(x) = \frac{\cosh\sqrt{v} - 1}{2\sqrt{v}} \geq 0$$

であり，$\sinh 0 - 0 = 0$ であるから，確かにこの範囲では $g(v) \geq 0$ が成り立つ．



ロジスティックシグモイド関数の下からの評価

凸関数の性質から接線がグラフの下にくるので

$$f(v) \geq f\left(\xi^2\right) + \frac{df}{dv}\left(\xi^2\right)\left(v - \xi^2\right)$$

$$f\left(x^2\right) - f\left(\xi^2\right) \geq -\frac{1}{4\xi}\tanh\frac{\xi}{2}\left(x^2 - \xi^2\right)$$

が成り立つ．教科書に合わせて $\lambda(\xi) := (4\xi)^{-1}\tanh(\xi/2)$ と置く．上式を使うと

$$\begin{aligned}\log\sigma(x) - \log\sigma(\xi) &= \frac{x}{2} + f\left(x^2\right) - \left(\frac{\xi}{2} + f\left(\xi^2\right)\right) \\ &\geq \frac{x-\xi}{2} - \lambda(\xi)\left(x^2 - \xi^2\right)\end{aligned}$$

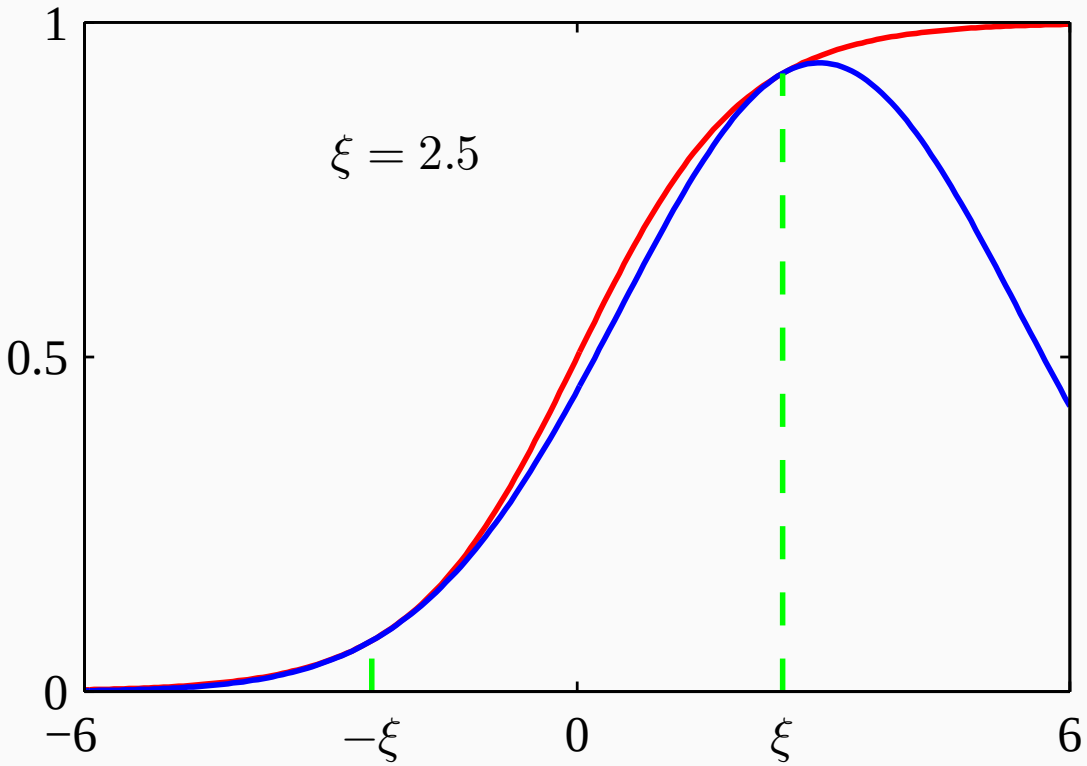
であり，両辺の指数をとって以下を得る．

$$\sigma(x) \geq \sigma(\xi)\exp\left(\frac{x-\xi}{2} - \lambda(\xi)\left(x^2 - \xi^2\right)\right). \tag{10.144}$$

ロジスティックシグモイド関数の下からの評価

今までの計算によって，私達はロジスティックシグモイド関数の下からの評価を得た.

$$\sigma(x) \geq \sigma(\xi) \exp\left(\frac{x-\xi}{2} - \lambda(\xi)\left(x^2 - \xi^2\right)\right)$$



どうやって使うのか

得られた不等式を積分計算に使ってみる．ロジスティックシグモイド関数 σ とガウス確率密度 p について以下の積分

$$I = \int \sigma(a)\, p(a)\, da \tag{10.145}$$

を計算したいとする．こういった計算は予測分布を求めるときに必要になる．不等式 $\sigma(a) \geq f(a, \xi)$ が成り立っているとすると (10.145) は

$$I \geq \int f(a, \xi)\, p(a)\, da = F(\xi) \tag{10.146}$$

と評価できる．もし F を最大化できれば I の良い近似になる．

変分事後分布

ロジスティック回帰モデルに変分近似を使ってみる．目的変数は $t \in \{0, 1\}$ とし，表記を簡単にするため $a = w'\phi$ とする．このとき尤度は

$$
\begin{aligned}
p(t|w) &= \sigma(a)^t (1 - \sigma(a))^{1-t} \\
&= \left(\frac{1}{1+e^{-a}}\right)^t \left(1 - \frac{1}{1+e^{-a}}\right)^{1-t} \\
&= \left(\frac{\frac{1}{1+e^{-a}}}{1 - \frac{1}{1+e^{-a}}}\right)^t \left(1 - \frac{1}{1+e^{-a}}\right) \\
&= e^{at} \frac{e^{-a}}{1+e^{-a}} = e^{at} \sigma(-a) \qquad (10.148)
\end{aligned}
$$

となる．

変分事後分布

前の節で計算した下限を使って評価する．(10.144) により

$$\begin{aligned} p(t|w) &= e^{at}\sigma(-a) \\ &\geq e^{at}\sigma(\xi)\exp\left(-\frac{a+\xi}{2}-\lambda(\xi)\left(a^2-\xi^2\right)\right) \end{aligned} \tag{10.151}$$

となる．

変分事後分布

したがって観測値の系列 $\mathbf{t}$ が得られたとき

$$\begin{aligned} p(\mathbf{t}, w) &= p(\mathbf{t}|w)\, p(w) \\ &= p(w) \prod_{n=1}^{N} p(t_n|w) \\ &\geq p(w)\, h(w, \xi), \end{aligned} \tag{10.152}$$

ただし

$$\begin{aligned} h(w, \xi) := \prod_{n=1}^{N} \sigma(\xi_n) \exp\Bigg(& w'\phi_n t_n - \frac{w'\phi_n + \xi_n}{2} \\ & - \lambda(\xi_n)\left((w'\phi_n)^2 - \xi_n^2 \right) \Bigg) \end{aligned} \tag{10.153}$$

となる.

変分事後分布

(10.153) の対数をとると

$$\log\left(p\left(\mathbf{t}|w\right)p\left(w\right)\right) \geq \log p\left(w\right) + \log h\left(w,\xi\right)$$

$$= \log p\left(w\right) + \sum_{n=1}^{N}\left(\log\sigma\left(\xi_n\right) + w'\phi_n t_n - \frac{w'\phi_n + \xi_n}{2} - \lambda\left(\xi_n\right)\left(\left(w'\phi_n\right)^2 - \xi_n^2\right)\right) \quad (10.154)$$

となる．右辺に事前分布 $p\left(w\right) = \mathcal{N}\left(w|m_0, S_0\right)$ を代入すると

$$-\left(w - m_0\right)' S_0^{-1}\left(w - m_0\right) + \sum_{n=1}^{N}\left(w'\phi_n\left(t_n - \frac{1}{2}\right) - \lambda\left(\xi_n\right)w'\left(\phi_n\phi_n'\right)w\right) + \text{const.} \quad (10.155)$$

となる．

変分事後分布

(10.155) の形から，変分事後分布 $q(w)$[5] は適当なガウス分布 $\mathcal{N}(w|m_N, S_N)$ で表せることが分かる．2 次の項に着目すると精度は

$$S_N^{-1} = S_0^{-1} + 2\sum_{n=1}^{N} \lambda(\xi_n)\phi_n\phi_n'$$

と分かる．1 次の項に着目すると平均 m_N は

$$m_N = S_N\left(S_0^{-1}m_0 + \sum_{n=1}^{N}\left(t_n - \frac{1}{2}\right)\phi_n\right)$$

である．つまり $q(w) = \mathcal{N}(w|m_N, S_N)$ である．

こうして私達はラプラス近似のように事後分布のガウス分布近似を得た．今回はさらに変分パラメータ $\{\xi_n\}_n$ が加わって柔軟になっているため，より高い精度が期待できる．

---

5 $q(w)$ は同時確率 $p(\mathbf{t}, w) = p(\mathbf{t}|w)\,p(w)$ を近似するもので，結果として事後分布 $p(w|\mathbf{t})$ を近似するものになるのだった．

変分パラメータの最適化

変分事後分布が $q(w) = \mathcal{N}(w|m_N, S_N)$ となることは分かった．平均 m_N と分散 S_N はどちらも ξ に依存しているので，ξ の最適化を考えなければならない．いつも通り周辺尤度の下からの近似を考えよう．

$$\begin{aligned} \log p(\mathbf{t}) &= \log \int p(\mathbf{t}|w)\, p(w)\, dw \\ &\geq \log \int h(w, \xi)\, p(w)\, dw = \mathcal{L}(\xi) \end{aligned} \tag{10.159}$$

この後の方法には二通りある．

1. w を潜在変数とみなして EM アルゴリズムを使う．
2. w に対する積分を計算し，ξ を直接最大化する．

まずは一つ目の EM アルゴリズムを使う方法から見ていく．

変分パラメータの最適化（EM アルゴリズム）

M ステップの導出については次ページ以降で説明する.

E ステップ

$$q(w) \leftarrow \mathcal{N}(w|m_N, S_N)$$

`where`

$$S_N^{-1} \leftarrow S_0^{-1} + 2\sum_{n=1}^{N} \lambda(\xi_n)\phi_n\phi_n'$$

$$m_N \leftarrow S_N\left(S_0^{-1}m_0 + \sum_{n=1}^{N}\left(t_n - \frac{1}{2}\right)\phi_n\right)$$

M ステップ

$$\xi_n^2 \leftarrow \phi_n'\mathbb{E}[ww']\phi_n = \phi_n'(S_N + m_N m_N')\phi_n$$

変分パラメータの最適化（EMアルゴリズム）

Mステップではξの新しい値を求めるため，Eステップで計算した事後分布$q(w)$を使って

$$\begin{aligned} Q\left(\xi, \xi^{\mathrm{old}}\right) &:= \mathbb{E}\left[\log\left(h\left(w, \xi\right) p\left(w\right)\right)\right] \\ &= \int q\left(w\right) \log\left(h\left(w, \xi\right) p\left(w\right)\right) dw \end{aligned}$$

を計算する．ξ_nに依存する項のみに着目すると

$$\begin{aligned} Q\left(\xi, \xi^{\mathrm{old}}\right) = \sum_{n=1}^{N} \Bigg(& \log \sigma\left(\xi_n\right) - \frac{\xi_n}{2} - \lambda\left(\xi_n\right)\left(\phi_n' \mathbb{E}\left[w w'\right] \phi_n - \xi_n^2\right) \Bigg) \\ & + \text{const.} \end{aligned} \tag{10.161}$$

となる．

変分パラメータの最適化（EMアルゴリズム）

以下を計算する.

$$\frac{d}{d\xi_n}\left(\log\sigma\left(\xi_n\right)-\frac{\xi_n}{2}-\lambda\left(\xi_n\right)\left(\phi_n'\mathbb{E}\left[ww'\right]\phi_n-\xi_n^2\right)\right).$$

ロジスティックシグモイド関数の微分の公式

$$\frac{d\sigma}{da}=\sigma\left(1-\sigma\right) \tag{4.88}$$

を使うと

$$\frac{d}{d\xi}\log\left(\sigma\left(\xi\right)\right)=\frac{\sigma\left(\xi\right)\left(1-\sigma\left(\xi\right)\right)}{\sigma\left(\xi\right)}=1-\sigma\left(\xi\right)$$

が成り立つ.

変分パラメータの最適化（EMアルゴリズム）

また $\lambda(\xi)$ については以下のように変形できる.

$$
\begin{aligned}
\lambda(\xi) &= \frac{1}{4\xi}\tanh\frac{\xi}{2} \\
&= \frac{1}{4\xi}\frac{e^{\xi/2}-e^{-\xi/2}}{e^{\xi/2}+e^{-\xi/2}} \\
&= \frac{1}{4\xi}\frac{1-e^{-\xi}}{1+e^{-\xi}} \\
&= \frac{2}{4\xi}\left(\frac{1}{1+e^{-\xi}}-\frac{1/2+e^{-\xi}/2}{1+e^{-\xi}}\right) \\
&= \frac{1}{2\xi}\left(\sigma(\xi)-\frac{1}{2}\right). \qquad (10.150)
\end{aligned}
$$

変分パラメータの最適化（EMアルゴリズム）

よって

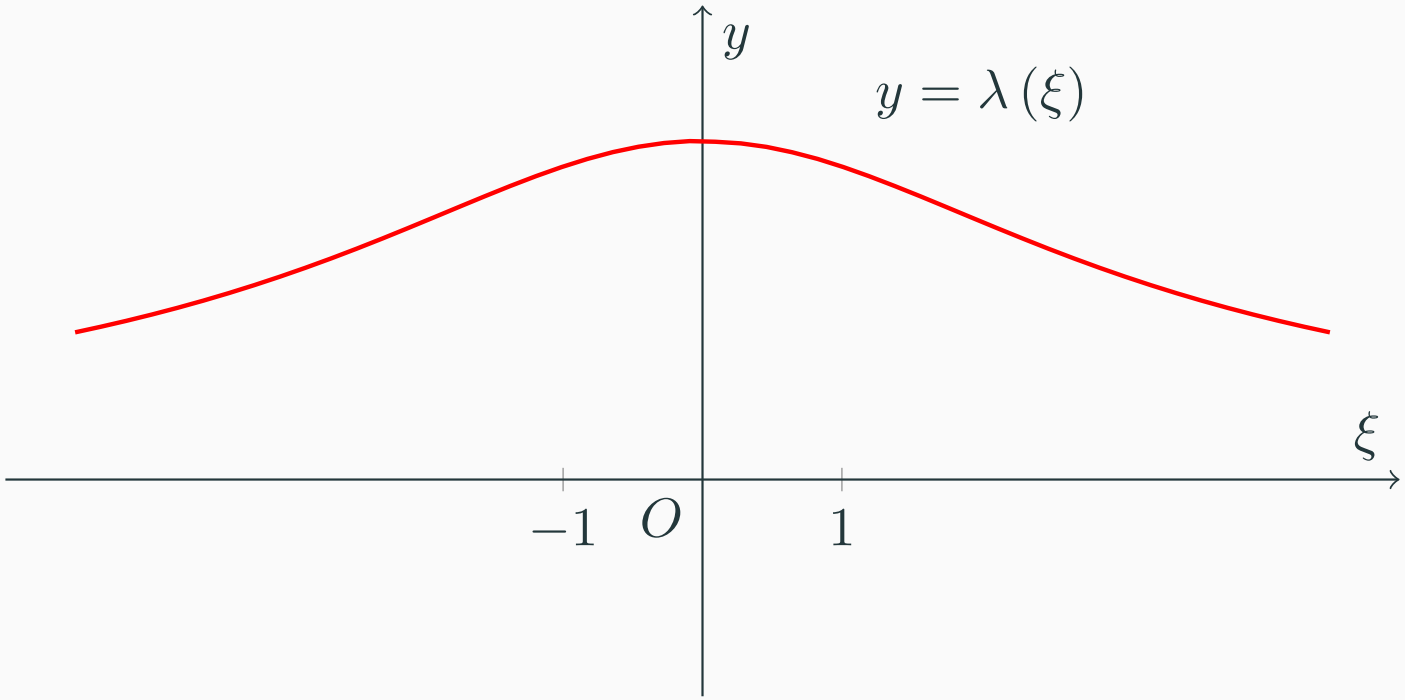
$$\begin{aligned}
&\frac{d}{d\xi_n}\left(\log\sigma\left(\xi_n\right)-\frac{\xi_n}{2}-\lambda\left(\xi_n\right)\left(\phi_n'\mathbb{E}\left[ww'\right]\phi_n-\xi_n^2\right)\right)\\
&=1-\sigma\left(\xi_n\right)-\frac{1}{2}+\frac{d\lambda}{d\xi_n}\left(\xi_n\right)\left(\phi_n'\mathbb{E}\left[ww'\right]\phi_n-\xi_n^2\right)+\sigma\left(\xi_n\right)-\frac{1}{2}\\
&=\frac{d\lambda}{d\xi_n}\left(\xi_n\right)\left(\phi_n'\mathbb{E}\left[ww'\right]\phi_n-\xi_n^2\right)
\end{aligned}$$

となる．停留条件を求めたいので

$$\frac{d\lambda}{d\xi_n}\left(\xi_n\right)\left(\phi_n'\mathbb{E}\left[ww'\right]\phi_n-\xi_n^2\right)=0 \tag{10.162}$$

とする．

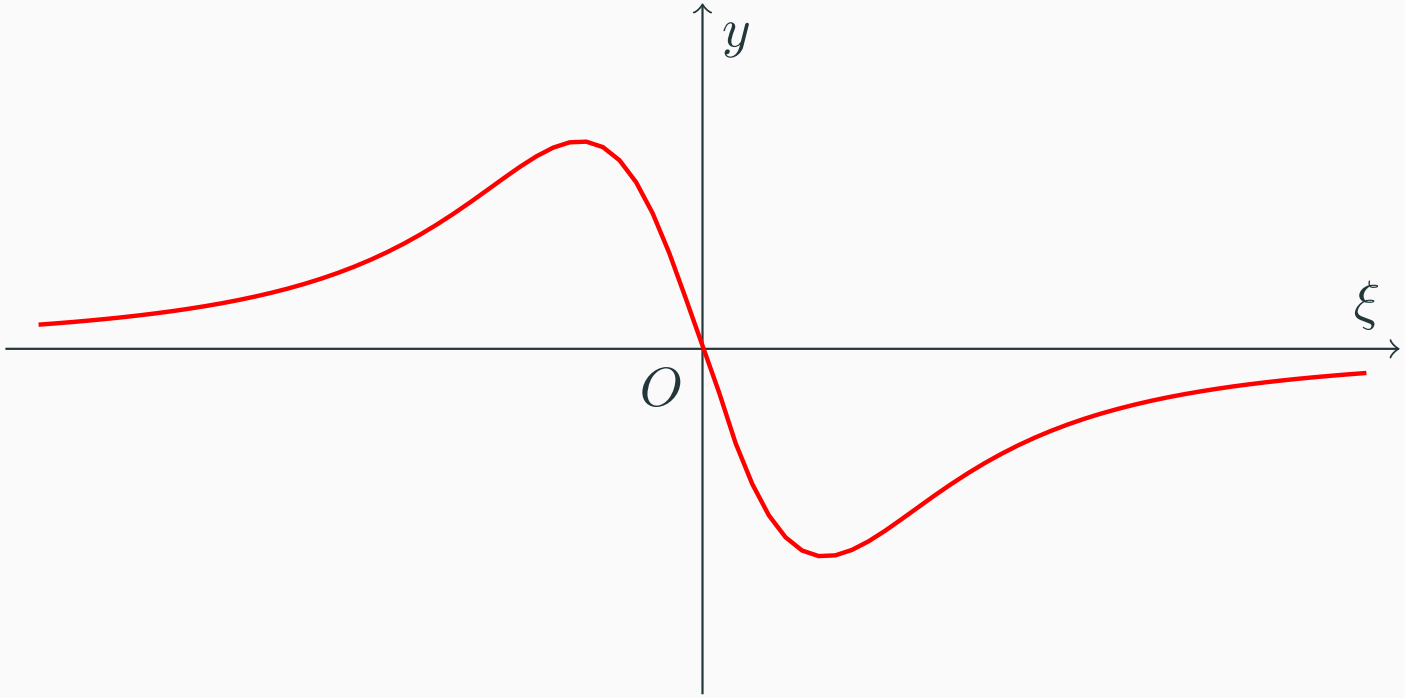
$y = \lambda(\xi)$ のグラフ



$$\lambda(0) = \lim_{\xi \to 0} \frac{\sigma(\xi) - \sigma(0)}{\xi - 0} = \left. \frac{d\sigma}{d\xi} \right|_{\xi=0}$$

$$= \sigma(0)(1 - \sigma(0)) = \frac{1}{4}$$

$y = (d\lambda/d\xi)(\xi)$ のグラフ



$$\frac{d\lambda}{d\xi}(\xi) = -\frac{1}{\xi^2}\left(\sigma(\xi) - \frac{1}{2}\right) - \frac{1}{\xi}\sigma(\xi)(1 - \sigma(\xi))$$

変分パラメータの最適化（EMアルゴリズム）

$\xi \neq 0$ の範囲では $(d\lambda/d\xi)(\xi) \neq 0$ なので

$$(\xi_n^{\text{new}})^2 = \phi_n' \mathbb{E}[ww'] \phi_n$$

となる．分散の定義，期待値の性質と $q(w) = \mathcal{N}(w|m_N, S_N)$ より

$$\begin{aligned} S_N &= \mathbb{E}\left[(w - m_N)(w - m_N)'\right] \\ &= \mathbb{E}[ww'] - m_N \mathbb{E}[w]' - \mathbb{E}[w] m_N' + m_N m_N' \\ &= \mathbb{E}[ww'] - m_N m_N' \end{aligned}$$

であるから

$$(\xi_n^{\text{new}})^2 = \phi_n' (S_N + m_N m_N') \phi_n \tag{10.163}$$

を得る．

変分パラメータの最適化（直接計算）

二つ目の方法は，$\mathcal{L}(\xi)$ を計算する方法である．計算で求められた $\mathcal{L}$ の式を ξ で微分することにより，$\mathcal{L}$ を最大化するような ξ を求めるのである．

$$\begin{aligned}\mathcal{L}(\xi) = \sum_{n=1}^{N}\left(\log\sigma(\xi_n) - \frac{\xi_n}{2} + \lambda(\xi_n)\,\xi_n^2\right) \\ + \log\int \frac{dw}{(2\pi)^m\sqrt{|S_0|}}\exp\left(-\frac{1}{2}(w-m_0)'\,S_0^{-1}\,(w-m_0)\right) \\ \exp\left(w'\left(\sum_{n=1}^{N}\left(t_n-\frac{1}{2}\right)\phi_n\right) - w'\left(\sum_{n=1}^{N}\lambda(\xi_n)\,\phi_n\phi_n'\right)w\right)\end{aligned}$$

変分パラメータの最適化（直接計算）

積分記号の中の指数関数の引数に着目すると

$$
\begin{aligned}
&-\frac{1}{2}w'S_0^{-1}w + w'\left(m_0S_0^{-1} + \sum_{n=1}^{N}\left(t_n - \frac{1}{2}\right)\phi_n\right) \\
&\quad -\frac{1}{2}m_0S_0^{-1}m_0 - w'\left(\sum_{n=1}^{N}\lambda\left(\xi_n\right)\phi_n\phi_n'\right)w \\
&= -\frac{1}{2}w'S_0^{-1}w + w'S_N^{-1}m_N - \frac{1}{2}m_0'S_0^{-1}m_0 - \frac{1}{2}w'\left(S_N^{-1} - S_0^{-1}\right)w \\
&= -\frac{1}{2}\left(w - m_N\right)'S_N^{-1}\left(w - m_N\right) + \frac{1}{2}m_N'S_N^{-1}m_N - \frac{1}{2}m_0'S_0^{-1}m_0
\end{aligned}
$$

変分パラメータの最適化（直接計算）

$$
\begin{aligned}
\mathcal{L}(\xi) = & \sum_{n=1}^{N} \left(\log \sigma(\xi_n) - \frac{\xi_n}{2} + \lambda(\xi_n) \xi_n^2 \right) \\
& + \log \int \sqrt{\frac{|S_N|}{|S_0|}} \frac{1}{\sqrt{|S_N|}} \exp \left(-\frac{1}{2} (w - m_N)' S_N^{-1} (w - m_N) \right) \\
& \qquad \exp \left(\frac{1}{2} m_N' S_N^{-1} m_N - \frac{1}{2} m_0' S_0^{-1} m_0 \right) dw \\
= & \frac{1}{2} \log \frac{|S_N|}{|S_0|} + \frac{1}{2} m_N' S_N^{-1} m_N - \frac{1}{2} m_0' S_0^{-1} m_0 \\
& + \sum_{n=1}^{N} \left(\log \sigma(\xi_n) - \frac{\xi_n}{2} + \lambda(\xi_n) \xi_n^2 \right) \qquad (10.164)
\end{aligned}
$$

変分パラメータの最適化（直接計算）

(10.164) の各項の ξ_n についての導関数を計算しておこう．まず $\log|S_N|$ を ξ_n で微分することを考える．公式 $(d/dt)\,|A(t)| = \mathrm{tr}\left(A^{-1}(t)\,\dot{A}(t)\right)$ を使う．

$$\begin{aligned}\frac{\partial}{\partial \xi_n}\log|S_N| &= -\frac{\partial}{\partial \xi_n}\log\left|S_N^{-1}\right| \\ &= -\mathrm{tr}\left(S_N\frac{\partial S_N^{-1}}{\partial \xi_n}\right) \\ &= -\mathrm{tr}\left(S_N\left(2\frac{d\lambda}{d\xi_n}(\xi_n)\,\phi_n\phi'\right)\right)\end{aligned}$$

対称行列の跡に関する公式 $x'Ax = \mathrm{tr}\,(Axx')$ を使って次を得る．

$$\frac{\partial}{\partial \xi_n}\log|S_N| = -2\frac{d\lambda}{d\xi_n}(\xi_n)\,\phi_n' S_N \phi_n.$$

公式の導出に関しては，このスライドの補足 1 と補足 2 を参照．

変分パラメータの最適化（直接計算）

次に $m_N' S_N^{-1} m_N$ を ξ_n で微分する．$m_N = S_N v_N$ と置く．(C.21) の公式 $\partial A^{-1}/\partial x = -A^{-1}(\partial A/\partial x)A^{-1}$ などを使って

$$\begin{aligned}
\frac{\partial}{\partial \xi_n}\left(m_N' S_N^{-1} m_N\right) &= \frac{\partial}{\partial \xi_n}\left(v_N' S_N' S_N^{-1} S_N v_N\right) \\
&= v_N' \left(\frac{\partial S_N}{\partial \xi_n}\right)' v_N \\
&= -v_N' S_N' \frac{\partial S_N^{-1}}{\partial \xi_n} S_N v_N \\
&= -2\frac{d\lambda}{d\xi_n}(\xi_n)\, m_N' \left(\phi_n \phi_n'\right) m_N \\
&= -2\frac{d\lambda}{d\xi_n}(\xi_n)\, \phi_n' m_N m_N' \phi_n
\end{aligned}$$

を得る．

変分パラメータの最適化（直接計算）

最後に総和記号の中を ξ_n で微分する.

$$
\begin{aligned}
&\frac{d}{d\xi_n}\left(\log\sigma\left(\xi_n\right)-\frac{\xi_n}{2}+\lambda\left(\xi_n\right)\xi_n^2\right)\\
&=1-\sigma\left(\xi_n\right)-\frac{1}{2}+\frac{d\lambda}{d\xi_n}\left(\xi_n\right)\xi_n^2+2\lambda\left(\xi_n\right)\xi_n\\
&=\frac{1}{2}-\sigma\left(\xi_n\right)+\frac{d\lambda}{d\xi_n}\left(\xi_n\right)\xi_n^2+\left(\sigma\left(\xi_n\right)-\frac{1}{2}\right)\\
&=\frac{d\lambda}{d\xi_n}\left(\xi_n\right)\xi_n^2
\end{aligned}
$$

変分パラメータの最適化（直接計算）

以上の結果から $\partial \mathcal{L}/\partial \xi_n = 0$ と置くと

$$\frac{d\lambda}{d\xi_n}(\xi_n)\left(\xi_n^2 - \phi_n' S_N \phi_n - \phi_n' m_N m_N' \phi_n\right) = 0$$

$$\frac{d\lambda}{d\xi_n}(\xi_n)\left(\xi_n^2 - \phi_n'\left(S_N + m_N m_N'\right)\phi_n\right) = 0$$

となり，EM アルゴリズムと全く同じ結果になる.

超パラメータの推論

ベイズロジスティック回帰モデルにおいて今まで w を定める超パラメータ α は既知の定数としてきたが，α もデータから推測できたらうれしい．以下でその方法を説明する．

w の事前分布として，以下の等方ガウス分布を仮定する [6]．

$$
\begin{aligned}
p(w|\alpha) &= \mathcal{N}\left(w \middle| 0, \alpha^{-1} I\right) \qquad (10.165)\\
&= \frac{1}{(2\pi)^{M/2} \left|\alpha^{-1} I\right|^{1/2}} \exp\left(-\frac{\alpha}{2} w'w\right)\\
&= \left(\frac{\alpha}{2\pi}\right)^{M/2} \exp\left(-\frac{\alpha}{2} w'w\right)
\end{aligned}
$$

6 w の次元が M であるという記述が見つけられなかったけど多分それであってるはず．

超パラメータの推論

共役超事前分布 $p(\alpha)$ はガンマ分布

$$p(\alpha) = \mathrm{Gamma}(\alpha|a_0, b_0) \tag{10.166}$$
$$= \frac{1}{\Gamma(a_0)} b_0^{a_0} \alpha^{a_0-1} e^{-b_0\alpha}$$

とする.

このモデルの周辺尤度は

$$p(\mathbf{t}) = \iint p(w, \alpha, \mathbf{t})\, dw d\alpha \tag{10.167}$$

である. ただし

$$p(w, \alpha, \mathbf{t}) = p(\mathbf{t}|w)\, p(w|\alpha)\, p(\alpha) \tag{10.168}$$

である.

超パラメータの推論

いつも通り周辺尤度の対数 $\log p(\mathbf{t})$ を以下のように分解する.

$$\log p(\mathbf{t}) = \mathcal{L}(q) + \mathrm{KL}(q\|p). \tag{10.169}$$

ここで

$$\mathcal{L}(q) = \iint q(w,\alpha) \log\left(\frac{p(w,\alpha,\mathbf{t})}{q(w,\alpha)}\right) dwd\alpha \tag{10.170}$$

$$\mathrm{KL}(q\|p) = -\iint q(w,\alpha) \log\left(\frac{p(w,\alpha|\mathbf{t})}{q(w,\alpha)}\right) dwd\alpha \tag{10.171}$$

である. このままでは $\mathcal{L}$ の最大化の計算が進められないので, またいつものように下から近似する.

$$\begin{aligned}\log p(\mathbf{t}) &\geq \mathcal{L}(q) \geq \widetilde{\mathcal{L}}(q,\xi) \\ &= \iint q(w,\alpha) \log\left(\frac{h(w,\xi)\, p(w|\alpha)\, p(\alpha)}{q(w,\alpha)}\right) dwd\alpha \end{aligned} \tag{10.172}$$

超パラメータの推論

変分分布がパラメータと超パラメータに分解できると仮定しよう.

$$q(w, \alpha) = q(w)\, q(\alpha) \tag{10.173}$$

これで $\mathcal{L}$ の最大化に取り組むことができる.

$$\begin{aligned}
\mathcal{L}(q) = & \iint q(w)\, q(\alpha) \log\left(h(w, \xi)\, p(w|\alpha)\, p(\alpha)\right) d\alpha dw \\
& - \int q(w) \left(\int q(\alpha) \log\left(q(w)\, q(\alpha)\right) d\alpha \right) dw \\
= & \int q(w) \left(\int q(\alpha) \log\left(h(w, \xi)\, p(w|\alpha)\, p(\alpha)\right) d\alpha \right) dw \\
& - \int q(w) \log q(w)\, dw - \int q(\alpha) \log q(\alpha)\, d\alpha
\end{aligned}$$

超パラメータの推論

KL ダイバージェンスの最小化条件を使って

$$\begin{aligned}\log q(w) &= \int q(\alpha) \log \left(h(w, \xi)\, p(w|\alpha)\, p(\alpha) \right) d\alpha + \text{const.} \\ &= \log h(w, \xi) + \mathbb{E}_\alpha \left[\log p(w|\alpha) \right] + \text{const.}\end{aligned}$$

となる．$\log h(w, \xi)$ に (10.153) を，$\log p(w|\alpha)$ に (10.165) を代入して次式を得る．

$$\begin{aligned}\log q(w) = &-\frac{\mathbb{E}[\alpha]}{2} w'w \\ &+ \sum_{n=1}^{N} \left(\left(t_n - \frac{1}{2} \right) w'\phi_n - \lambda(\xi_n)\, w'\phi_n \phi' w \right) + \text{const.}\end{aligned}$$

これは w の二次関数なので $q(w)$ はガウス分布であることが分かる．

超パラメータの推論

したがって $q(w) = \mathcal{N}(w|\mu_N, \Sigma_N)$ とすれば

$$\Sigma_N^{-1} = \mathbb{E}[\alpha] I + 2 \sum_{n=1}^{N} \lambda(\xi_n) \phi_n \phi_n', \tag{10.176}$$

$$\Sigma_N^{-1} \mu_N = \sum_{n=1}^{N} \left(t_n - \frac{1}{2} \phi_n \right) \tag{10.175}$$

が成り立つ.

超パラメータの推論

次に α の変分分布について計算する．$\mathcal{L}$ は以下のように変形できる.

$$\mathcal{L}(q) = \int q(\alpha) \left(\log p(\alpha) + \int q(w) \log p(w|\alpha)\, dw \right) d\alpha$$
$$- \int q(\alpha) \log q(\alpha)\, d\alpha + \text{const.}$$

よって KL ダイバージェンスの最小化条件を使って

$$\log q(\alpha) = \mathbb{E}_w \left[\log p(w|\alpha) \right] + \log p(\alpha) + \text{const.}$$

となる.

超パラメータの推論

$\log p\left(w|\alpha\right)$ に (10.165) を，$\log p\left(\alpha\right)$ に (10.166) を代入して

$$\begin{aligned}\log q\left(\alpha\right) &= \frac{M}{2}\log\alpha - \frac{\alpha}{2}\mathbb{E}\left[w'w\right] + \left(a_0 - 1\right)\log\alpha - b_0\alpha + \text{const.} \\ &= \left(\frac{M}{2} + a_0 - 1\right)\log\alpha - \left(b_0 + \frac{1}{2}\mathbb{E}\left[w'w\right]\right)\alpha + \text{const.}\end{aligned}$$

を得る．これはガンマ分布の対数の形になっているので

$$a_N := a_0 + \frac{M}{2}, \tag{10.178}$$

$$b_N := b_0 + \frac{1}{2}\mathbb{E}\left[w'w\right] \tag{10.179}$$

と置くと

$$q\left(\alpha\right) = \text{Gamma}\left(\alpha|a_N, b_N\right) = \frac{1}{\Gamma\left(a_N\right)} a_N^{b_N}\,\alpha^{a_N - 1} e^{-b_N\alpha} \tag{10.177}$$

となる．

超パラメータの推論

最後に ξ の推定を行う方法を考える．$\widetilde{\mathcal{L}}(q, \xi)$ の ξ に関連する項に着目すると

$$\begin{aligned}\mathcal{L}(q) &= \iint q(w)\, q(\alpha) \log h(w, \xi)\, d\alpha dw + \text{const.} \\ &= \int q(w) \log h(w, \xi)\, dw + \text{const.}\end{aligned}$$

となる．これはと同じ形なので前の結果から

$$\left(\xi_n^{\text{new}}\right)^2 = \phi_n' \left(\Sigma_N + \mu_N \mu_N'\right) \phi_n \tag{10.181}$$

以上で $q(w)$, $q(\alpha)$, ξ を再推定する方程式が得られた．

行列式の微分

行列式の微分の公式を導く．A を n 次正則行列とし，その各行を $a_i\ (i = 1, \ldots, n)$ と表す．

$$\frac{d}{dt}\left|A\left(t\right)\right|$$

$$= \lim_{t\to 0}\frac{1}{t}\left(\left|\begin{array}{c} a_1\left(t+h\right) \\ a_2\left(t+h\right) \\ \vdots \\ a_n\left(t+h\right) \end{array}\right| - \left|\begin{array}{c} a_1\left(t\right) \\ a_2\left(t\right) \\ \vdots \\ a_n\left(t\right) \end{array}\right|\right)$$

$$= \lim_{t\to 0}\frac{1}{t}\left(\left|\begin{array}{c} a_1\left(t+h\right) \\ a_2\left(t+h\right) \\ \vdots \\ a_n\left(t+h\right) \end{array}\right| - \left|\begin{array}{c} a_1\left(t\right) \\ a_2\left(t+h\right) \\ \vdots \\ a_n\left(t+h\right) \end{array}\right| + \left|\begin{array}{c} a_1\left(t\right) \\ a_2\left(t+h\right) \\ \vdots \\ a_n\left(t+h\right) \end{array}\right| - \left|\begin{array}{c} a_1\left(t\right) \\ a_2\left(t\right) \\ \vdots \\ a_n\left(t\right) \end{array}\right|\right)$$

行列式の微分

$$
\begin{aligned}
&= \lim_{t\to 0} \frac{1}{t} \left(\begin{vmatrix} a_1(t+h) - a_1(t) \\ a_2(t+h) \\ \vdots \\ a_n(t+h) \end{vmatrix} \right) \\
&\quad + \lim_{t\to 0} \frac{1}{t} \left(\begin{vmatrix} a_1(t) \\ a_2(t+h) \\ \vdots \\ a_n(t+h) \end{vmatrix} - \begin{vmatrix} a_1(t) \\ a_2(t) \\ \vdots \\ a_n(t) \end{vmatrix} \right) \\
&= \begin{vmatrix} \dot{a}_1(t) \\ a_2(t) \\ \vdots \\ a_n(t) \end{vmatrix} + \lim_{t\to 0} \frac{1}{t} \left(\begin{vmatrix} a_1(t) \\ a_2(t+h) \\ \vdots \\ a_n(t+h) \end{vmatrix} - \begin{vmatrix} a_1(t) \\ a_2(t) \\ \vdots \\ a_n(t) \end{vmatrix} \right)
\end{aligned}
$$

行列式の微分

$$
\begin{aligned}
&= \begin{vmatrix} \dot{a}_1(t) \\ a_2(t) \\ \vdots \\ a_n(t) \end{vmatrix} + \lim_{t\to 0} \frac{1}{t} \left(\begin{vmatrix} a_1(t) \\ a_2(t+h) \\ \vdots \\ a_n(t+h) \end{vmatrix} - \begin{vmatrix} a_1(t) \\ a_2(t) \\ \vdots \\ a_n(t+h) \end{vmatrix} \right) \\
&\quad + \lim_{t\to 0} \frac{1}{t} \left(\begin{vmatrix} a_1(t) \\ a_2(t) \\ \vdots \\ a_n(t+h) \end{vmatrix} - \begin{vmatrix} a_1(t) \\ a_2(t) \\ \vdots \\ a_n(t) \end{vmatrix} \right) \\
&= \begin{vmatrix} \dot{a}_1(t) \\ a_2(t) \\ \vdots \\ a_n(t) \end{vmatrix} + \begin{vmatrix} a_1(t) \\ \dot{a}_2(t) \\ \vdots \\ a_n(t) \end{vmatrix} + \lim_{t\to 0} \frac{1}{t} \left(\begin{vmatrix} a_1(t) \\ a_2(t) \\ \vdots \\ a_n(t+h) \end{vmatrix} - \begin{vmatrix} a_1(t) \\ a_2(t) \\ \vdots \\ a_n(t) \end{vmatrix} \right)
\end{aligned}
$$

行列式の微分

上記の計算を繰り返すことで以下の公式を得る.

$$\frac{d}{dt}\left|A\left(t\right)\right| = \sum_{i=1}^{n} \begin{vmatrix} a_1\left(t\right) \\ \vdots \\ \dot{a}_i\left(t\right) \\ \vdots \\ a_n\left(t\right) \end{vmatrix} = \sum_{i=1}^{n}\sum_{k=1}^{n} \dot{a}_{ik}\Delta_{ik}$$

これを使いさらに簡潔な表現を得ることができる. 次頁で説明する.

行列式の微分

A を n 次正則行列 A とし，その (i,j) 成分の余因子を $\Delta_{ij} := (-1)^{i+j} |A_{ij}|$，余因子行列を $\widetilde{A} = (\Delta_{ji})_{ij}$ と表す．このとき第 i 行に関する余因子展開を行えば

$$\frac{d}{dt} |A(t)| = \sum_{i=1}^{n} \begin{vmatrix} a_1(t) \\ \vdots \\ \dot{a}_i(t) \\ \vdots \\ a_n(t) \end{vmatrix} = \sum_{i=1}^{n} \sum_{k=1}^{n} \dot{a}_{ik} \Delta_{ik}$$

となる．

$$\mathrm{tr}\left(\widetilde{A}(t)\dot{A}(t)\right) = \mathrm{tr}\left(\left(\sum_{k=1}^{n} \Delta_{ki}\dot{a}_{kj}\right)_{ij}\right) = \sum_{i=1}^{n}\sum_{k=1}^{n} \Delta_{ki}\dot{a}_{ki}$$

なので $(d/dt)|A(t)| = \mathrm{tr}\left(\widetilde{A}(t)\dot{A}(t)\right)$ を得る．

行列式の対数の微分

正則行列について $\widetilde{A}/|A| = A^{-1}$ が成り立つことと，いま導いた $(d/dt)|A(t)| = \mathrm{tr}\left(\widetilde{A}(t)\dot{A}(t)\right)$ を使うと

$$\begin{aligned}\frac{d}{dt}\log|A(t)| &= \frac{1}{|A(t)|}\frac{d}{dt}|A(t)| \\ &= \frac{1}{|A(t)|}\mathrm{tr}\left(\widetilde{A}(t)\dot{A}(t)\right) \\ &= \frac{1}{|A(t)|}\mathrm{tr}\left(\widetilde{A}(t)\dot{A}(t)\right) \\ &= \mathrm{tr}\left(A^{-1}(t)\dot{A}(t)\right)\end{aligned}$$

を得る.

跡と二次形式

公式 $\mathrm{tr}\,(Axx') = x'Ax$ を示す．そのために，まずは $\mathrm{tr}\,(AB) = \mathrm{tr}\,(BA)$ を示す．

$$\begin{aligned}\mathrm{tr}\,(AB) &= \sum_{i=1}^{m}\sum_{k=1}^{n} a_{ik}b_{ki} \\ &= \sum_{i=1}^{m}\sum_{k=1}^{n} b_{ki}a_{ik} \\ &= \mathrm{tr}\,(BA)\end{aligned}$$

したがって

$$\begin{aligned}\mathrm{tr}\,(Axx') &= \mathrm{tr}\,(x'\,(Ax)) \\ &= \mathrm{tr}\,(x'Ax) \\ &= x'Ax\end{aligned}$$

が成り立つ．

参考文献

Bogachev, V. I. (2007). Measure theory, volume 1. Springer Science & Business Media.

Borwein, J. M. and Lewis, A. S. (2010). Convex analysis and nonlinear optimization: theory and examples. Springer Science & Business Media.

Golberg, M. (1972). The derivative of a determinant. American Mathematical Monthly, page 1124–1126.

ビショップ, C. M. (2008). パターン認識と機械学習 下. シュプリンガージャパン.