PRML 10.5-10.6

9/14/2018

酒井 一徳

目次

10.5　局所的変分推論法

モチベーション

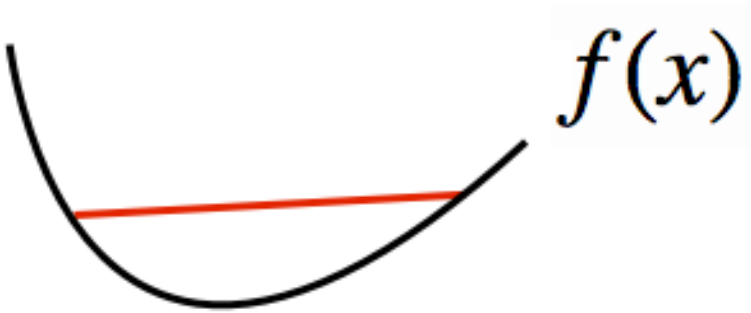
これまで

- すべての変数における事後分布を直接近似。

ここから

- 各変数や変数群における関数の近似。
  - 全体の計算可能な近似を得る。

凸関数

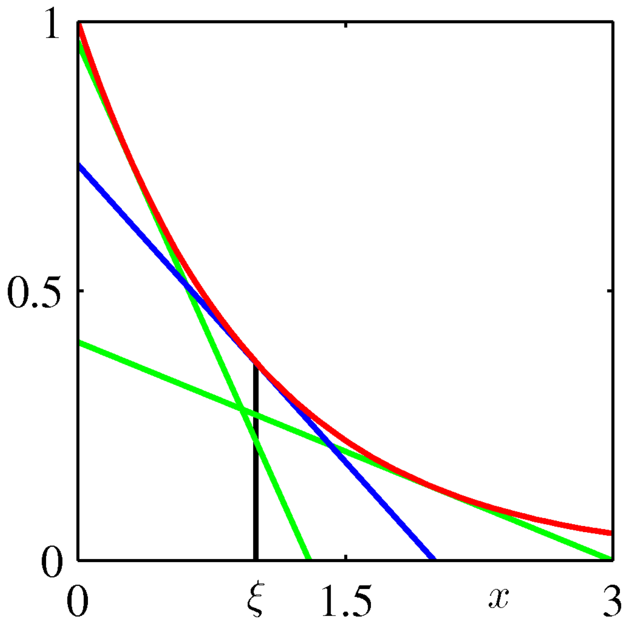


広義: $f(tx + (1-t)y) \leq tf(x) + (1-t)f(y)$

狭義: $f(tx + (1-t)y) < tf(x) + (1-t)f(y)$

※wikiより

凸関数の接線は下界



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

$x = \xi$ における接線は、

$$y(x) = \exp(-\xi) - \exp(-\xi)(x - \xi)$$

$$y(x) \leqslant f(x) \text{ (等号成立は } x = \xi \text{)}$$

$y(x)$ はパラメータ ξ を持つ一次関数

凸関数の下界による近似

$\eta \equiv -exp(-\xi)$ とすると接線は、

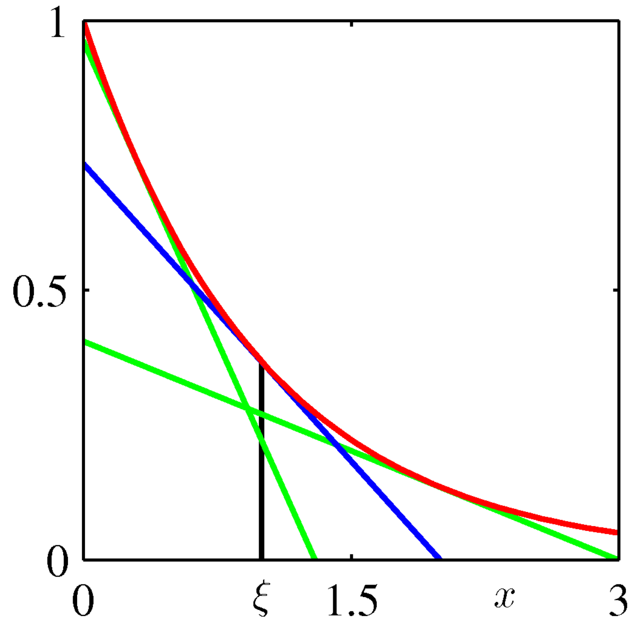
$$y(x,\eta) = \eta x - \eta + \eta \ln(-\eta).$$

(η の変更)=(接線の変更) を意味し、

$y(x,\eta) \leqslant f(x)$ であるので、

$$f(x) = \max_{\eta}\{\eta x - \eta + \eta \ln(-\eta)\}.$$

xは固定



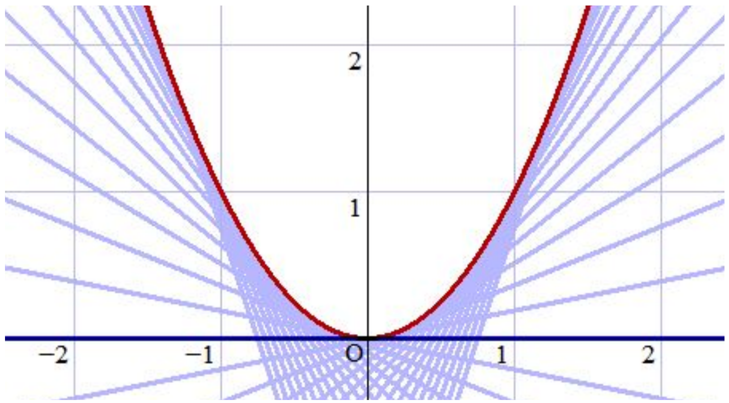
一次関数で近似に成功。
ただし、変分パラメータ η の最適化が必要。

凸双対性(凸共役性)

- ルジャンドル変換

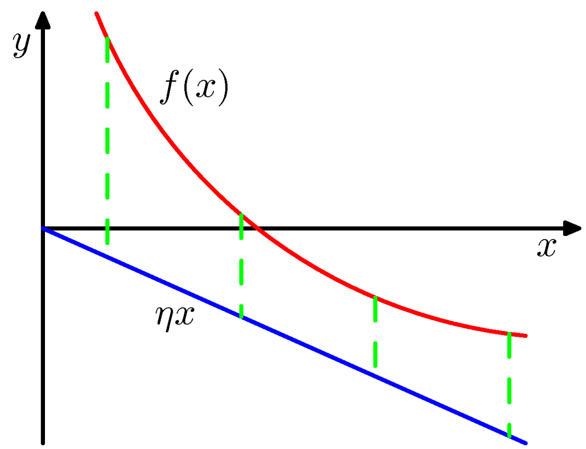
関数の変数変換手法。
凸関数を接線で表現できることに基づく(らしい)。

$$f^*(\eta) = \max_x\{\eta x - f(x)\}$$



- 再度変換

$$f^{**}(x) = \max_\eta\{x\eta - f^*(\eta)\}$$



凸関数の時、 $f^{**} = f$

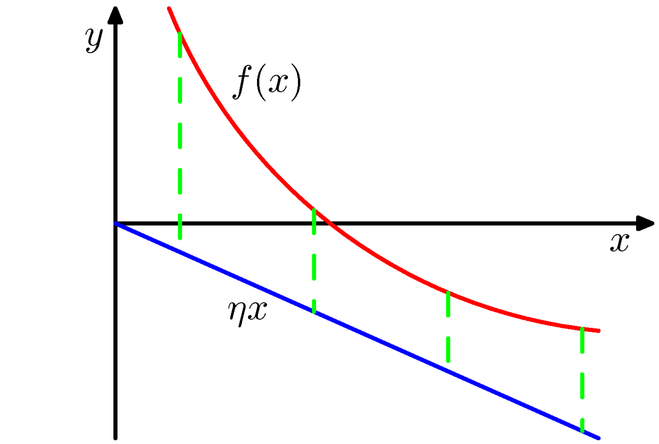
凸双対性(凸共役性)

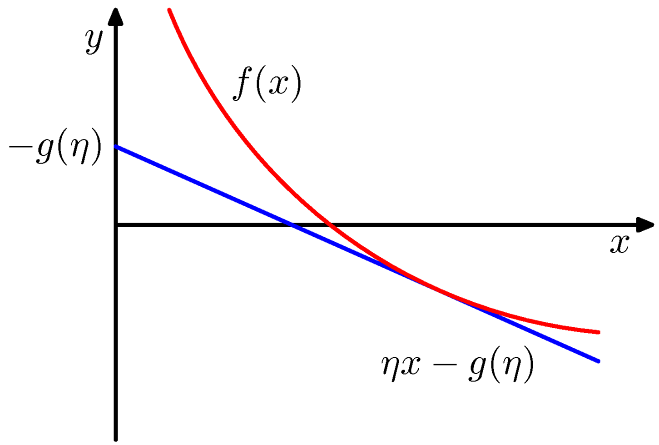
凸な関数	$f(x)$
接線	$\eta x - g(\eta)$
切片	$g(\eta)$

η を固定

$$g(\eta) = \max_x \{\eta x - f(x)\}$$

凹関数



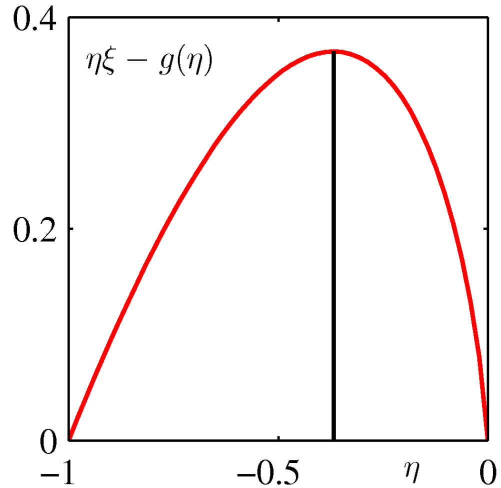


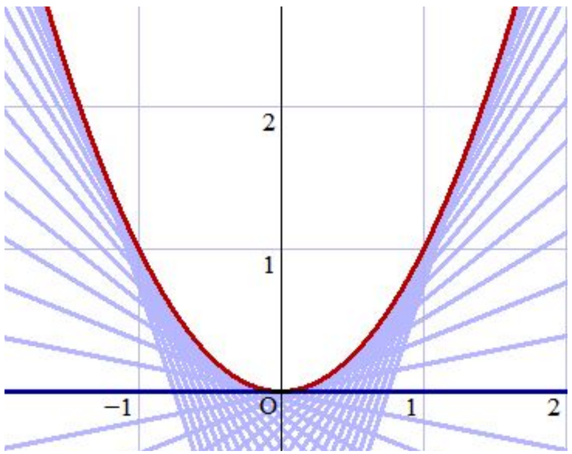
凸双対性(凸共役性)

x を固定

$$f(x) = \max_{\eta}\{x\eta - g(\eta)\}$$

凹関数





凸双対性(凸共役性)の具体例

- $f(x) = \exp(-x)$ の共役関数 $f^*(\eta) = \max_x \{\eta x - f(x)\}$を求める。

  微分し、0とおくことで、

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

よって共役関数は、

$$f^*(\eta) = \eta - \eta \ln(-\eta).$$

凸双対性(凸共役性)の具体例

- さらにその共役関数 $f^{**}(x) = \max_{\eta}\{x\eta - f^*(\eta)\}$ を求める。
  微分し、0とおくことで、

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

- 代入して元の関数が得られる。

$$f^{**}(x) = \exp(-x) = f(x)$$

凸関数と凹関数

- 凸関数なら次式で下界が、

$$f(x) = \max_{\eta}\{x\eta - g(\eta)\}$$

$$g(\eta) = \max_{x}\{\eta x - f(x)\}$$

- 凹関数なら次式で上界が得られる。

$$f(x) = \min_{\eta}\{x\eta - g(\eta)\}$$

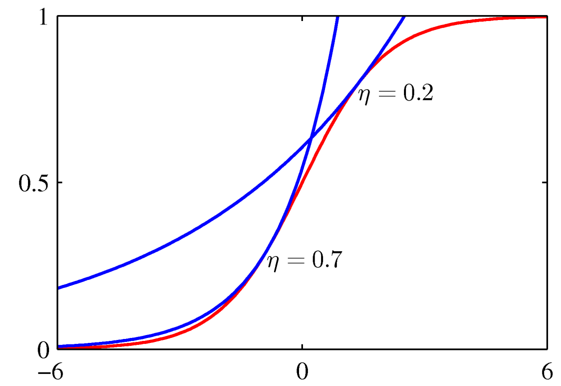
$$g(\eta) = \min_{x}\{\eta x - f(x)\}$$

- **凸でも凹でもない関数は？**

ロジスティックシグモイド関数の共役関数

- **可逆**な変換で凸か凹にする。

$$\sigma(x) = \frac{1}{1+e^{-x}}$$

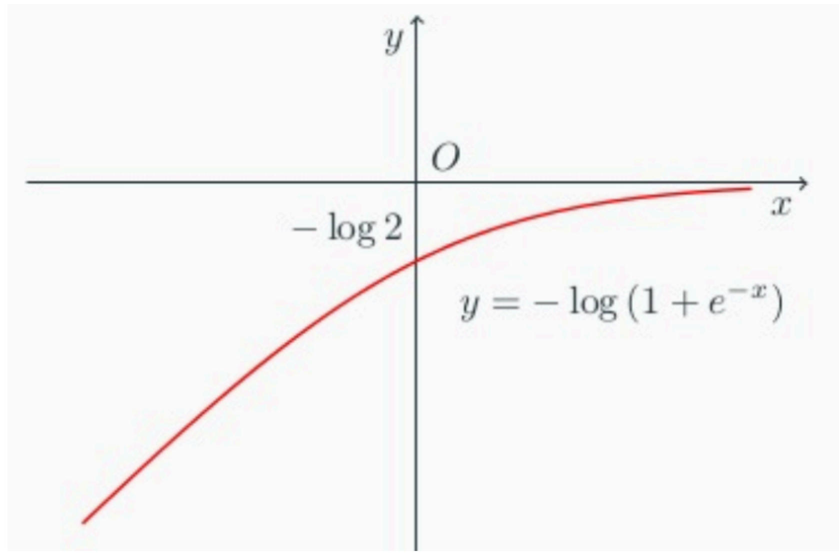


- 凸でも凹でもないが、対数を取ると凹になる。

$$\ln \sigma(x) = -\ln(1+e^{-x})$$

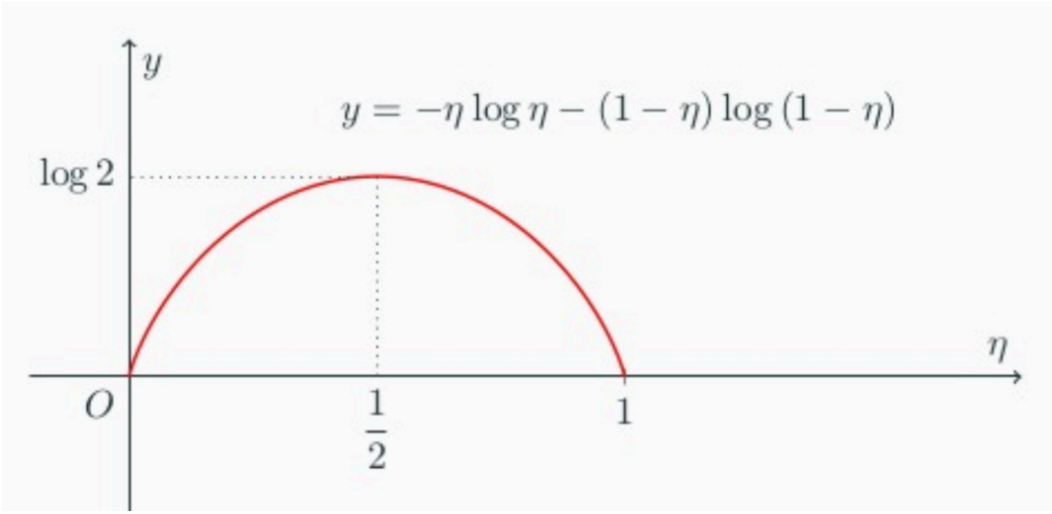
$$\frac{d}{dx}\ln \sigma(x) = \frac{1}{1+e^{x}} > 0$$

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



ロジスティックシグモイド関数の共役関数

$$g(\eta) = \min_x\{\eta x - f(x)\} = -\eta \ln \eta - (1-\eta)\ln(1-\eta)$$

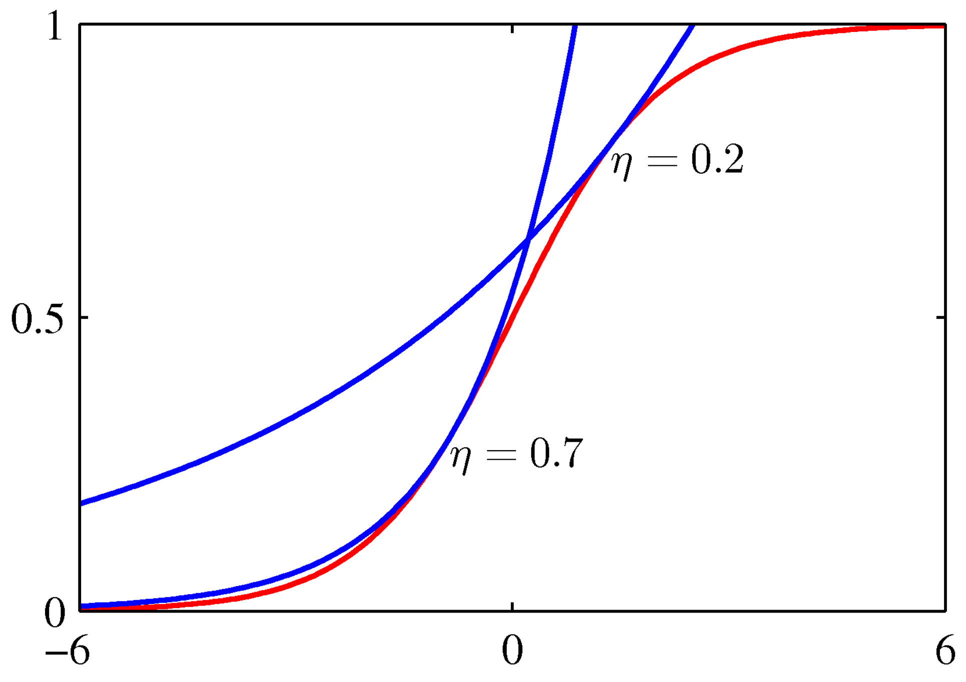


二値エントロピー関数

シグモイド関数の上界

$$\ln \sigma(x) \leqslant \eta x - g(\eta)$$

$$\sigma(x) \leqslant \exp(\eta x - g(\eta))$$



シグモイド関数の下界

➢ ガウス分布の関数形で下から抑えることもできる。

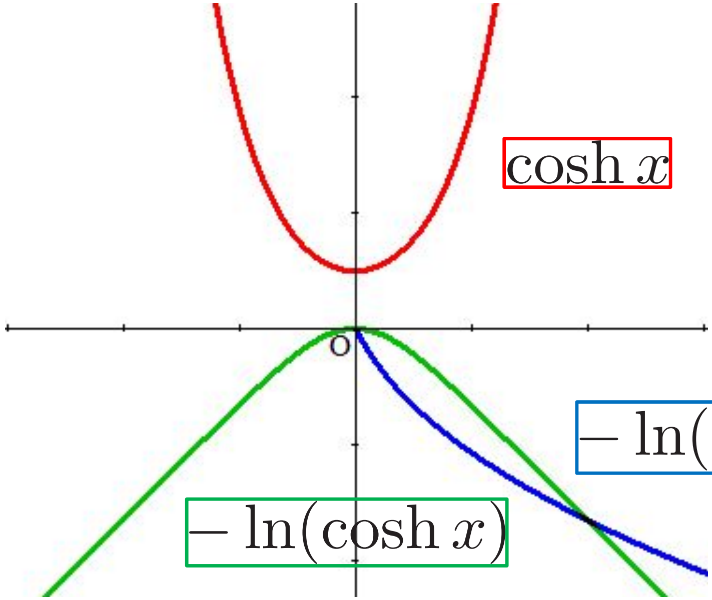
$$\ln \sigma(x) = -\ln\left(1+e^{-x}\right) = -\ln\left\{e^{-x/2}\left(e^{x/2}+e^{-x/2}\right)\right\}$$
$$= x/2 - \ln\left(e^{x/2}+e^{-x/2}\right)$$

$$\boxed{\cosh x = \frac{e^{x}+e^{-x}}{2}}$$

第二項に着目し、

$$f(x) = -\ln\left(2\cosh\left(\frac{x}{2}\right)\right)$$

x^2 について上記式は凸関数である。

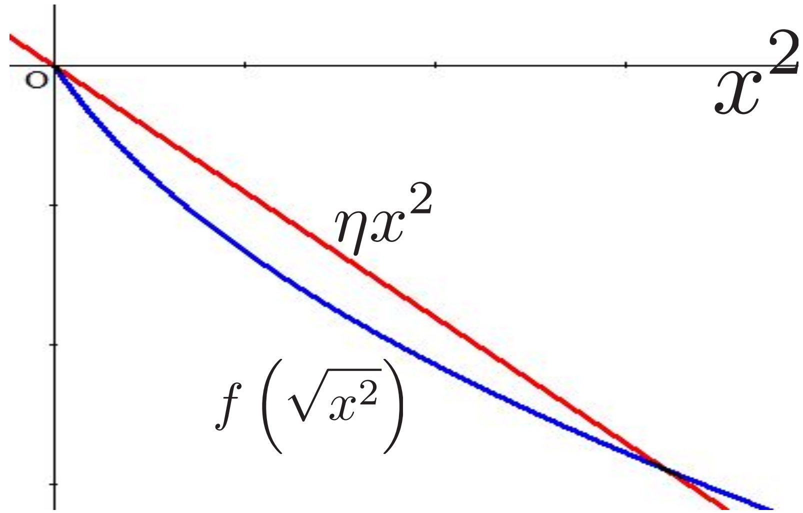


シグモイド関数の下界

共役関数: $g(\eta) = \max_{x^2}\left\{\eta x^2 - f\left(\sqrt{x^2}\right)\right\}$

変数x^2の凹関数

停留点の条件は $\quad 0 = \eta + \frac{1}{4x}\tanh\left(\frac{x}{2}\right).$



上記を満たす傾きηの接線の接点の値xをξとし以下を定義する。

$$\lambda(\xi) \equiv -\eta = \frac{1}{4\xi}\tanh\left(\frac{\xi}{2}\right)$$

シグモイド関数の下界

- よって共役関数は、

$$g(\lambda(\xi)) = -\lambda(\xi)\xi^2 - f(\xi) = -\lambda(\xi)\xi^2 + \ln\left(e^{\xi/2} + e^{-\xi/2}\right)$$

- さらにその共役関数は、

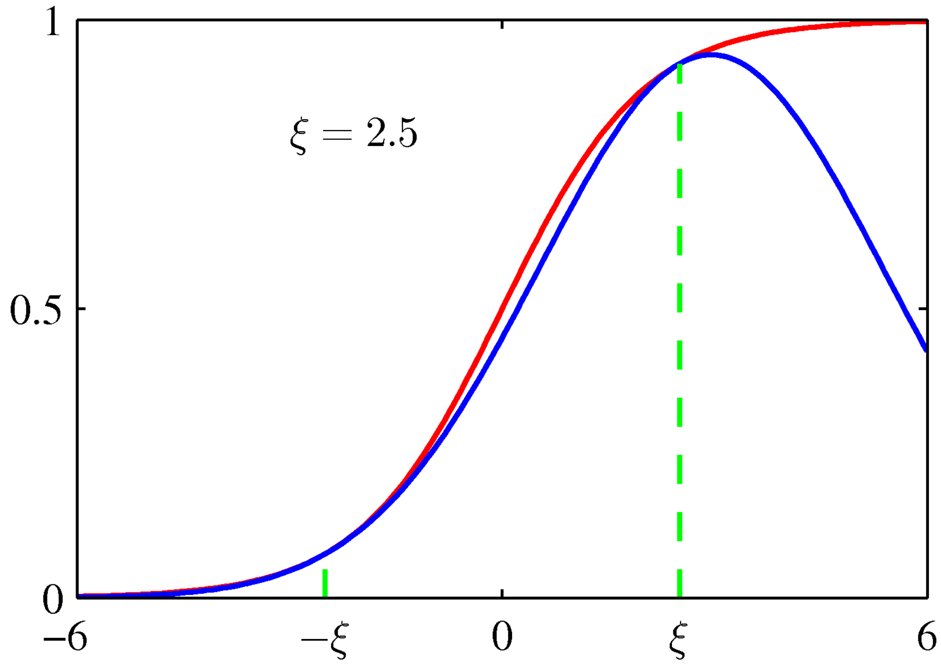
$$f(x) = \max_{\lambda(\xi)}\{-x^2\lambda(\xi) - g(\lambda(\xi))\}$$

- 上記から次が言える。

$$f(x) \geqslant -\lambda(\xi)x^2 - g(\lambda(\xi)) = -\lambda(\xi)x^2 + \lambda(\xi)\xi^2 - \lambda(\xi)(x^2 - \xi^2)$$

シグモイド関数の下界

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



利用法

- 例えば、次の計算をしたい。

$$I = \int \sigma(a)p(a)\mathrm{d}a$$

ガウス確率密度

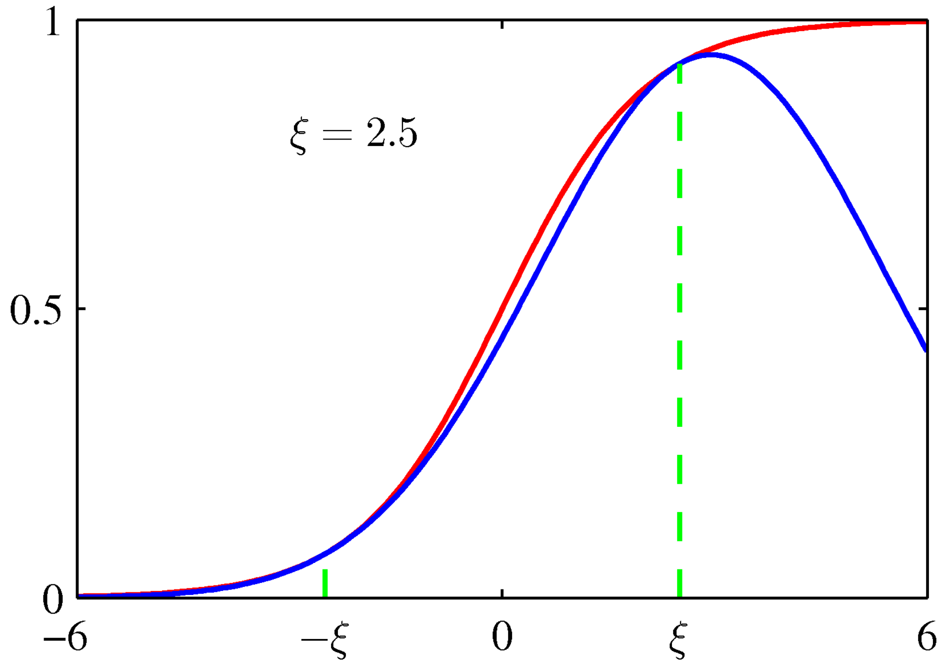
- $\sigma(a) \geqslant f(a,\xi)$ が成り立っていれば、

$$I \geqslant \int f(a,\xi)p(a)\mathrm{d}a = F(\xi)$$

- F を最大化することでより良い I の近似が得られる。

利用法

- 最適化した時の下界は一般に正確でない。
  - 選んだξはあるaに依存するため。



10.6 変分ロジスティック回帰

変分近似とラプラス近似

- 局所的変分法をベイズロジスティック回帰で試そう。
- ラプラス近似同様に事後分布をガウス分布で近似する。

- ラプラス近似より高い精度。
  - （データが相対的に少ない時？）
- モデルエビデンスの厳密な下界として与えられる明確な目的関数を最適化する。

10.6.1 変分事後分布

ベイズロジスティック回帰

➢ 周辺尤度

$$p(\mathbf{t}) = \int p(\mathbf{t}|\mathbf{w})p(\mathbf{w})\mathrm{d}\mathbf{w} = \int \left[\prod_{n=1}^{N} p\left(t_n|\mathbf{w}\right)\right] p(\mathbf{w})\mathrm{d}\mathbf{w}$$

➢ 前節から

$$\boxed{a = \mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}}$$

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

尤度の下界

再掲 $\sigma(x) \geqslant \sigma(\xi)\exp\left\{(x-\xi)/2 - \lambda(\xi)\left(x^2-\xi^2\right)\right\}$

- 前節から

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

- **観測値ごとに変分パラメータを用意する。**

同時分布の下界

$$p(\mathbf{t}, \mathbf{w}) = p(\mathbf{t}|\mathbf{w})p(\mathbf{w}) \geqslant h(\mathbf{w}, \boldsymbol{\xi})p(\mathbf{w})$$

正規化されていない

where

$$h(\mathbf{w}, \xi) = \prod_{n=1}^{N} \sigma(\xi_n) \exp\{\mathbf{w}^{\mathrm{T}} \boldsymbol{\phi}_n t_n - \left(\mathbf{w}^{\mathrm{T}} \boldsymbol{\phi}_n + \xi_n\right)/2 - \lambda(\xi_n)\left(\left[\mathbf{w}^{\mathrm{T}} \boldsymbol{\phi}_n\right]^2 - \xi_n^2\right)\}$$

- 事後分布は不等式の左辺を正規化する必要がある。

同時分布の下界

$$\ln\{p(\mathbf{t}|\mathbf{w})p(\mathbf{w})\} \geqslant \ln p(\mathbf{w}) + \sum_{n=1}^{N}\Big\{\ln\sigma\left(\xi_n\right) + \mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}_n t_n$$
$$- \left(\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}_n + \xi_n\right)/2 - \lambda\left(\xi_n\right)\left(\left[\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}_n\right]^2 - \xi_n^2\right)\Big\}$$

- 事前分布 $p(\mathbf{w}) = \mathcal{N}\left(\mathbf{w}|\mathbf{m}_0, \mathbf{S}_0\right)$ を代入。
- wに関して整理。

$$-\frac{1}{2}\left(\mathbf{w}-\mathbf{m}_0\right)^{\mathrm{T}}\mathbf{S}_0^{-1}\left(\mathbf{w}-\mathbf{m}_0\right)$$
$$+\sum_{n=1}^{N}\left\{\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}_n\left(t_n - 1/2\right) - \lambda\left(\xi_n\right)\mathbf{w}^{\mathrm{T}}\left(\boldsymbol{\phi}_n\boldsymbol{\phi}_n^{\mathrm{T}}\right)\mathbf{w}\right\} + \ \mathrm{const.}$$

変分事後分布

- wの2次と1次の項にそれぞれ着目。
- 事後分布の変分近似は、以下のガウス分布。

$$q(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_N, \mathbf{S}_N)$$

ただし、

$$\mathbf{S}_N^{-1} = \mathbf{S}_0^{-1} + 2\sum_{n=1}^{N} \lambda\left(\xi_n\right) \boldsymbol{\phi}_n \boldsymbol{\phi}_n^{\mathrm{T}}$$

$$\mathbf{m}_N = \mathbf{S}_N \left(\mathbf{S}_0^{-1}\mathbf{m}_0 + \sum_{n=1}^{N} \left(t_n - 1/2\right) \boldsymbol{\phi}_n \right)$$

ラプラス近似と変分近似

ラプラス近似

$$q(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{w}_{\mathrm{MAP}}, \mathbf{S}_N)$$

$$\begin{aligned}\mathbf{S}_N &= -\nabla\nabla \ln p(\mathbf{w}|\mathbf{t}) \\ &= \mathbf{S}_0^{-1} + \sum_{n=1}^{N} y_n (1 - y_n) \boldsymbol{\phi}_n \boldsymbol{\phi}_n^{\mathrm{T}}\end{aligned}$$

$$y_n = \sigma\left(\mathbf{w}^{\mathrm{T}} \boldsymbol{\phi}_n\right)$$

変分近似

$$q(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_N, \mathbf{S}_N)$$

$$\mathbf{m}_N = \mathbf{S}_N \left(\mathbf{S}_0^{-1}\mathbf{m}_0 + \sum_{n=1}^{N} (t_n - 1/2)\, \boldsymbol{\phi}_n\right)$$

$$\mathbf{S}_N^{-1} = \mathbf{S}_0^{-1} + 2\sum_{n=1}^{N} \lambda(\xi_n)\, \boldsymbol{\phi}_n \boldsymbol{\phi}_n^{\mathrm{T}}$$

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

10.6.2 変分パラメータの最適化

変分パラメータの最適化

- 周辺尤度の最大化で変分パラメータの決定を行う。

$$\ln p(\mathbf{t}) = \ln \int p(\mathbf{t}|\mathbf{w})p(\mathbf{w})\mathrm{d}\mathbf{w} \geqslant \ln \int h(\mathbf{w}, \xi)p(\mathbf{w})\mathrm{d}\mathbf{w} = \mathcal{L}(\xi)$$

- 決定方法は二通り。

1. wを潜在変数とみなしてEMアルゴリズム。
2. wに対する積分を解析的に計算し、直接最大化。

変分パラメータの最適化(EMアルゴリズム)

- Eステップ

変分パラメータに初期値を与え、事後分布を計算。

$$S_N^{-1} \leftarrow S_0^{-1} + 2\sum_{n=1}^{N} \lambda\left(\xi_n^{\mathrm{old}}\right) \boldsymbol{\phi}_n \boldsymbol{\phi}_n^{\mathrm{T}}$$

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

- Mステップ

$$\left(\xi_n^{\mathrm{new}}\right)^2 \leftarrow \boldsymbol{\phi}_n^{\mathrm{T}} \mathbb{E}\left[\mathbf{w}\mathbf{w}^{\mathrm{T}}\right] \boldsymbol{\phi}_n = \boldsymbol{\phi}_n^{\mathrm{T}} \left(S_N + \mathbf{m}_N \mathbf{m}_N^{\mathrm{T}} \right) \boldsymbol{\phi}_n$$

Mステップの導出は次頁

Mステップの導出

- 次式の期待完全データ対数尤度の最大化。

$$Q\left(\boldsymbol{\xi}, \boldsymbol{\xi}^{\text{ old}}\right) = \mathbb{E}[\ln h(\mathbf{w}, \boldsymbol{\xi}) p(\mathbf{w})]$$
$$= \int q(\mathbf{w}) \log(h(\mathbf{w}, \boldsymbol{\xi}) p(\mathbf{w})) d\mathbf{w}$$

- 変分パラメータのみに着目。

$$Q\left(\boldsymbol{\xi}, \boldsymbol{\xi}^{\text{ old}}\right) = \sum_{n=1}^{N} \left\{ \ln \sigma\left(\xi_n\right) - \xi_n / 2 - \lambda\left(\xi_n\right) \left(\boldsymbol{\phi}_n^{\mathrm{T}} \mathbb{E}\left[\mathbf{w}\mathbf{w}^{\mathrm{T}}\right] \boldsymbol{\phi}_n - \xi_n^2 \right) \right\}$$
$$+ \ \text{const}$$

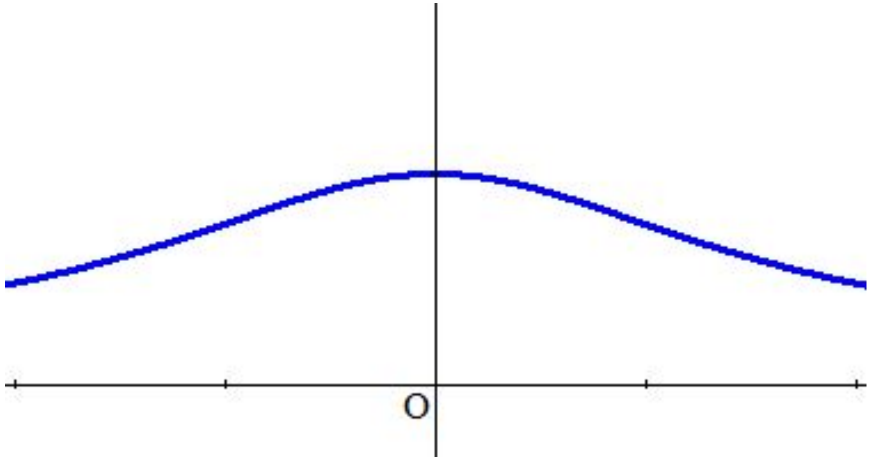
Mステップの導出

$$
\begin{aligned}
&\frac{\mathrm{d}}{\mathrm{d}\xi_n}\left(\ln \sigma\left(\xi_n\right)-\xi_n/2-\lambda\left(\xi_n\right)\left(\boldsymbol{\phi}_n^{\mathrm{T}}\mathbb{E}\left[\mathbf{w}\mathbf{w}^{\mathrm{T}}\right]\boldsymbol{\phi}_n-\xi_n^2\right)\right)\\
&=1-\sigma\left(\xi_n\right)-\frac{1}{2}-\frac{d\lambda}{d\xi_n}\left(\xi_n\right)\left(\boldsymbol{\phi}_n^{\mathrm{T}}\mathbb{E}\left[\mathbf{w}\mathbf{w}^{\mathrm{T}}\right]\boldsymbol{\phi}_n-\xi_n^2\right)+\sigma\left(\xi_n\right)-\frac{1}{2}\\
&=-\lambda'(\xi_n)\left(\boldsymbol{\phi}_n^{\mathrm{T}}\mathbb{E}\left[\mathbf{w}\mathbf{w}^{\mathrm{T}}\right]\boldsymbol{\phi}_n-\xi_n^2\right)
\end{aligned}
$$

- よって停留条件は、

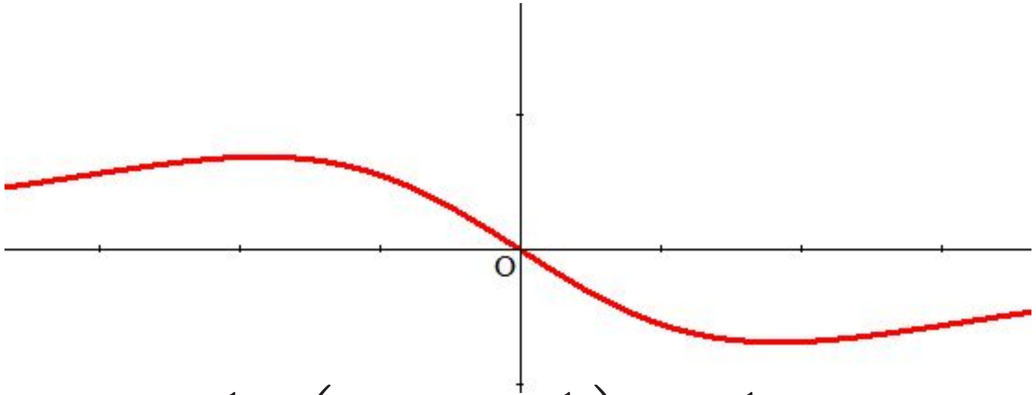
$$
0=\lambda'\left(\xi_n\right)\left(\boldsymbol{\phi}_n^{\mathrm{T}}\mathbb{E}\left[\mathbf{w}\mathbf{w}^{\mathrm{T}}\right]\boldsymbol{\phi}_n-\xi_n^2\right)
$$

Mステップの導出



$$\lambda(\xi) = \frac{1}{2\xi}\left(\sigma(\xi) - \frac{1}{2}\right)$$
$$= \frac{1}{4\xi}\tanh\frac{\xi}{2}$$

Mステップの導出



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

$\xi \neq 0$ の範囲で $\lambda'(\xi) \neq 0$ なので(らしいので)。

$$\left(\xi_n^{\mathrm{new}}\right)^2 = \boldsymbol{\phi}_n^{\mathrm{T}} \mathbb{E}\left[\mathbf{w}\mathbf{w}^{\mathrm{T}}\right] \boldsymbol{\phi}_n$$

Mステップの導出

$q(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_N, \mathbf{S}_N)$ なので、

$$\mathbb{E}\left[\mathbf{w}\mathbf{w}^{\mathrm{T}}\right] = \mathbf{S}_N + \mathbf{m}_N\mathbf{m}_N^{\mathrm{T}}$$

➢ よって更新式は

$$\left(\xi_n^{\mathrm{new}}\right)^2 = \boldsymbol{\phi}_n^{\mathrm{T}}\left(S_N + \mathbf{m}_N\mathbf{m}_N^{\mathrm{T}}\right)\boldsymbol{\phi}_n$$

EMアルゴリズム再掲

- Eステップ

変分パラメータに初期値を与え、事後分布を計算。

$$S_N^{-1} \leftarrow S_0^{-1} + 2\sum_{n=1}^{N} \lambda\left(\xi_n^{\mathrm{old}}\right) \boldsymbol{\phi}_n \boldsymbol{\phi}_n^{\mathrm{T}}$$

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

- Mステップ

$$\left(\xi_n^{\mathrm{new}}\right)^2 \leftarrow \boldsymbol{\phi}_n^{\mathrm{T}} \mathbb{E}\left[\mathbf{w}\mathbf{w}^{\mathrm{T}}\right] \boldsymbol{\phi}_n = \boldsymbol{\phi}_n^{\mathrm{T}} \left(S_N + \mathbf{m}_N \mathbf{m}_N^{\mathrm{T}} \right) \boldsymbol{\phi}_n$$

変分パラメータの最適化(直接計算)

- 解析的に計算をする。

$$
\begin{aligned}
\mathcal{L}(\boldsymbol{\xi}) =& \sum_{n=1}^{N} \left(\log \sigma\left(\xi_n\right) - \frac{\xi_n}{2} + \lambda\left(\xi_n\right)\xi_n^2 \right) \\
& + \ln \frac{1}{(2\pi)^{D/2}} \int \sqrt{\frac{|\mathbf{S}_N|}{|\mathbf{S}_0|}} \frac{1}{\sqrt{|\mathbf{S}_N|}} \exp\left(-\frac{1}{2}\left(\mathbf{w}-\mathbf{m}_N\right)^{\mathrm{T}} \mathbf{S}_N^{-1}\left(\mathbf{w}-\mathbf{m}_N\right) \right) \\
& \exp\left(\frac{1}{2}\mathbf{m}_N^{\mathrm{T}}\mathbf{S}_N^{-1}\mathbf{m}_N - \frac{1}{2}\mathbf{m}_0^{\mathrm{T}}\mathbf{S}_0^{-1}\mathbf{m}_0 \right) \mathrm{d}\mathbf{w} \\
=& \frac{1}{2}\ln\frac{|\mathbf{S}_N|}{|\mathbf{S}_0|} + \frac{1}{2}\mathbf{m}_N^{\mathrm{T}}\mathbf{S}_N^{-1}\mathbf{m}_N - \frac{1}{2}\mathbf{m}_0^{\mathrm{T}}\mathbf{S}_0^{-1}\mathbf{m}_0 \\
& + \sum_{n=1}^{N} \left(\ln \sigma\left(\xi_n\right) - \frac{\xi_n}{2} - \lambda\left(\xi_n\right)\xi_n^2 \right) \qquad (10.164)
\end{aligned}
$$

変分パラメータの最適化(直接計算)

(10.164) を変分パラメータについて微分。

$$\frac{\partial}{\partial x}\left(\mathbf{A}^{-1}\right) = -\mathbf{A}^{-1}\frac{\partial \mathbf{A}}{\partial x}\mathbf{A}^{-1} \quad \text{(C.21)}$$

$$\frac{\partial}{\partial x}\ln|\mathbf{A}| = \mathrm{Tr}\left(\mathbf{A}^{-1}\frac{\partial \mathbf{A}}{\partial x}\right) \quad \text{(C.22)}$$

一項目について、

$$\begin{aligned}
\frac{\partial}{\partial \xi_n}\ln|\mathbf{S}_N| &= \frac{1}{2}\,\mathrm{Tr}\left(\mathbf{S}_N^{-1}\frac{\partial \mathbf{S}_N}{\partial \xi_n}\right) \\
&= \frac{1}{2}\,\mathrm{Tr}\left(\mathbf{S}_N^{-1}\left(-\mathbf{S}_N\frac{\partial \mathbf{S}_N^{-1}}{\partial \xi_n}\mathbf{S}_N\right)\right) \\
&= \frac{1}{2}\,\mathrm{Tr}\left(\mathbf{S}_N^{-1}\left(-\mathbf{S}_N 2\lambda'\left(\xi_n\right)\boldsymbol{\phi}_n\boldsymbol{\phi}_n^{\mathrm{T}}\mathbf{S}_N\right)\right)
\end{aligned}$$

変分パラメータの最適化(直接計算)

二項目について、

$$\begin{aligned}\frac{\partial}{\partial \xi_n}\left(\frac{1}{2}\mathbf{m}_N^{\mathrm{T}}\mathbf{S}_N^{-1}\mathbf{m}_N\right) &= \frac{\partial}{\partial \xi_n}\left(\frac{1}{2}\mathbf{a}_N^{\mathrm{T}}\mathbf{S}_N^{\mathrm{T}}\mathbf{S}_N^{-1}\mathbf{S}_N\mathbf{a}_N\right) \\ &= \frac{\partial}{\partial \xi_n}\left(\frac{1}{2}\mathbf{a}_N^{\mathrm{T}}\mathbf{S}_N\mathbf{a}_N\right) \\ &= \frac{1}{2}\left(\mathbf{a}_N^{\mathrm{T}}\left(-\mathbf{S}_N^{\mathrm{T}}2\lambda'\left(\xi_n\right)\phi_n\phi_n^{\mathrm{T}}\mathbf{S}_N\right)\mathbf{a}_N\right) \\ &= \frac{1}{2}\mathrm{Tr}\left(-\mathbf{a}_N\mathbf{a}_N^{\mathrm{T}}\left(\mathbf{S}_N 2\lambda'\left(\xi_n\right)\phi_n\phi_n^{\mathrm{T}}\mathbf{S}_N\right)\right)\end{aligned}$$

三項目について、

$$\frac{\partial}{\partial \xi_n}\left(\ln\sigma\left(\xi_n\right) - \frac{1}{2}\xi_n + \lambda\left(\xi_n\right)\xi_n^2\right) = \lambda'\left(\xi_n\right)\xi_n^2$$

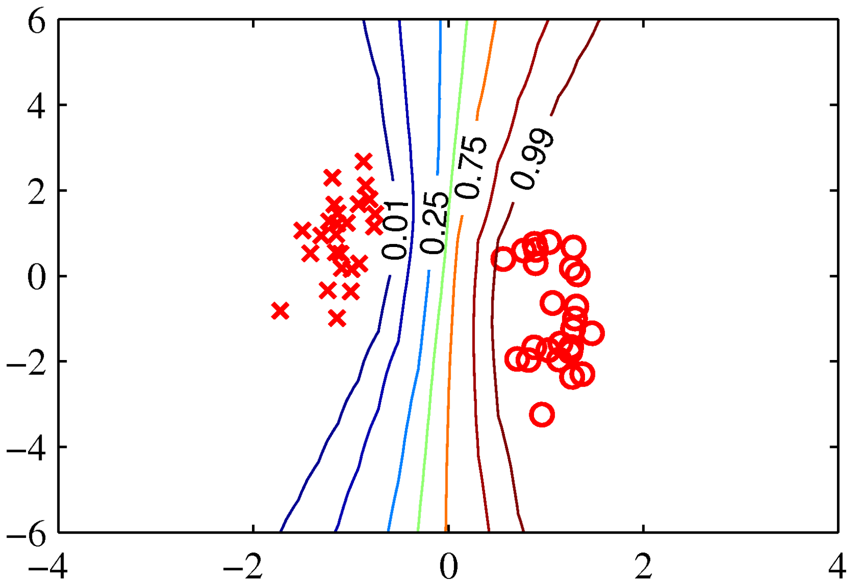
変分パラメータの最適化(直接計算)

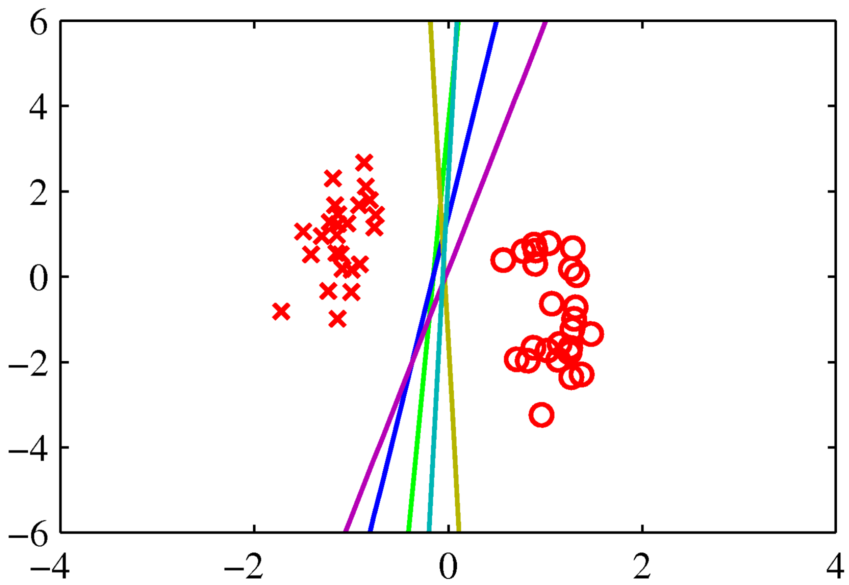
$$-\frac{1}{2}\operatorname{Tr}\left(\left(\mathbf{S}_N^{-1}+\mathbf{a}_N\mathbf{a}_N^{\mathrm{T}}\right)\mathbf{S}_N 2\lambda'\left(\xi_n\right)\phi_n\phi_n^{\mathrm{T}}\mathbf{S}_N\right)+\lambda'\left(\xi_n\right)\xi_n^2=0$$

- 整理して、

$$\begin{aligned}\xi_n^2 &= \boldsymbol{\phi}_n^{\mathrm{T}}\mathbf{S}_N\left(\mathbf{S}_N^{-1}+\mathbf{a}_N\mathbf{a}_N^{\mathrm{T}}\right)\mathbf{S}_N\boldsymbol{\phi}_n \\ &= \boldsymbol{\phi}_n^{\mathrm{T}}\left(\mathbf{S}_N+\mathbf{m}_N\mathbf{m}_N^{\mathrm{T}}\right)\boldsymbol{\phi}_n\end{aligned}$$

予測分布と分離境界





10.6.3 超パラメータの推論

よりbayesianに

- ここまでは事前分布の超パラメータは既知の定数。

  ⇒ 超パラメータもデータから決定しよう!!

- 事前分布に次の等方ガウス分布を仮定する。

$$p(\mathbf{w}|\alpha) = \mathcal{N}(\mathbf{w}|\mathbf{0}, \alpha^{-1}\mathbf{I})$$

- 共役超事前分布はガンマ分布。

$$p(\alpha) = \mathrm{Gam}(\alpha|a_0, b_0)$$

周辺尤度

$$p(\mathbf{t}) = \iint p(\mathbf{w}, \alpha, \mathbf{t})\mathrm{d}\mathbf{w}\mathrm{d}\alpha$$

where

$$p(\mathbf{w}, \alpha, \mathbf{t}) = \underbrace{p(\mathbf{t}|\mathbf{w})}_{\text{sigmoid}}\underbrace{p(\mathbf{w}|\alpha)}_{\text{gauss}}\underbrace{p(\alpha)}_{\text{gamma}}$$

- 解析的に計算不可。
  - 局所的変分法と大域的変分法でなんとかしよう。

大域的変分法

- 変分分布 $q(\mathbf{w}, \alpha)$ を導入。

$$\ln p(\mathbf{t}) = \mathcal{L}(q) + \mathrm{KL}(q\|p)$$

where

$$\mathcal{L}(q) = \iint q(\mathbf{w}, \alpha) \ln\left\{\frac{p(\mathbf{w}, \alpha, \mathbf{t})}{q(\mathbf{w}, \alpha)}\right\} \mathrm{d}\mathbf{w}\mathrm{d}\alpha$$

$$\mathrm{KL}(q\|p) = -\iint q(\mathbf{w}, \alpha) \ln\left\{\frac{p(\mathbf{w}, \alpha|\mathbf{t}))}{q(\mathbf{w}, \alpha)}\right\} \mathrm{d}\mathbf{w}\mathrm{d}\alpha$$

- まだ尤度関数の関数形がよろしくない。
  - 局所的変分法。

局所的変分法

$$p(\mathbf{t}|\mathbf{w}) \geqslant h(\mathbf{w}, \boldsymbol{\xi})$$

- 上記の前節の内容から、下記の下界の下界が得られる。

$$\begin{aligned}\ln p(\mathbf{t}) \geqslant \mathcal{L}(q) \geqslant \tilde{\mathcal{L}}(q, \boldsymbol{\xi}) \\ = \iint q(\mathbf{w}, \alpha) \ln \left\{ \frac{h(\mathbf{w}, \boldsymbol{\xi}) p(\mathbf{w}|\alpha) p(\alpha)}{q(\mathbf{w}, \alpha)} \right\} \mathrm{d}\mathbf{w} \mathrm{d}\alpha\end{aligned}$$

- 次の分解を仮定する。

$$q(\mathbf{w}, \alpha) = q(\mathbf{w}) q(\alpha)$$

各因子の最適解

$$\ln q_j^\star(\mathbf{Z}_j) = \mathbb{E}_{i\neq j}[\ln p(\mathbf{X},\mathbf{Z})] + \text{ const.}$$

- 上記を用いてまずは $q(\mathbf{w})$ について考える。

$$\begin{aligned}\ln q(\mathbf{w}) &= \mathbb{E}_\alpha[\ln\{h(\mathbf{w},\boldsymbol{\xi})p(\mathbf{w}|\alpha)p(\alpha)\}] + \text{ const} \\ &= \ln h(\mathbf{w},\boldsymbol{\xi}) + \mathbb{E}_\alpha[\ln p(\mathbf{w}|\alpha)] + \text{ const.}\end{aligned}$$

wについて

- 今までの結果から、

$$\begin{aligned}\ln q(\mathbf{w}) = &-\frac{\mathbb{E}[\alpha]}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w} \\ &+\sum_{n=1}^{N}\left\{(t_n - 1/2)\,\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}_n - \lambda(\xi_n)\,\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}_n\boldsymbol{\phi}_n^{\mathrm{T}}\mathbf{w}\right\} + \text{ const.}\end{aligned}$$

各因子の最適解

- 平方完成を行い、次を得る。

$$q(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\boldsymbol{\mu}_N, \boldsymbol{\Sigma}_N)$$

where

$$\Sigma_N^{-1} = \mathbb{E}[\alpha]\mathbf{I} + 2\sum_{n=1}^{N} \lambda(\xi_n)\,\boldsymbol{\phi}_n\boldsymbol{\phi}_n^{\mathrm{T}}$$

$$\Sigma_N^{-1}\boldsymbol{\mu}_N = \sum_{n=1}^{N} (t_n - 1/2)\,\boldsymbol{\phi}_n$$

各因子の最適解

- 次に $q(\alpha)$について考える。

$$\ln q(\alpha) = \mathbb{E}_{\mathbf{w}}[\ln p(\mathbf{w}|\alpha)] + \ln p(\alpha) + \text{ const.}$$

αについて

- 代入して整理する。

$$\ln q(\alpha) = \frac{M}{2}\ln\alpha - \frac{\alpha}{2}\mathbb{E}\left[\mathbf{w}^{\mathrm{T}}\mathbf{w}\right] + (a_0 - 1)\ln\alpha - b_0\alpha + \text{ const.}$$

これはガンマ分布の対数となっている。

各因子の最適解

- よって整理して以下を得る。

$$q(\alpha) = \mathrm{Gam}(\alpha | a_N, b_N) = \frac{1}{\Gamma(a_N)} a_N^{b_N} \alpha^{a_N - 1} e^{-b_N \alpha}$$

where

$$a_N = a_0 + \frac{M}{2}$$

$$b_N = b_0 + \frac{1}{2} \mathbb{E}_{\mathbf{w}} \left[\mathbf{w}^{\mathrm{T}} \mathbf{w} \right]$$

変分パラメータの最適化

- 前節と同様に変分パラメータについても最適化が必要。

$$\begin{aligned}\tilde{\mathcal{L}}(q,\boldsymbol{\xi}) &= \iint q(\mathbf{w},\alpha)\ln\left\{\frac{h(\mathbf{w},\xi)p(\mathbf{w}|\alpha)p(\alpha)}{q(\mathbf{w},\alpha)}\right\}\mathrm{d}\mathbf{w}\mathrm{d}\alpha \\ &= \iint q(\mathbf{w},\alpha)\ln\left\{h(\mathbf{w},\xi)p(\mathbf{w}|\alpha)p(\alpha)\right\}\mathrm{d}\mathbf{w}\mathrm{d}\alpha + \mathrm{const} \\ &= \int q(\mathbf{w})\ln h(\mathbf{w},\xi)\mathrm{d}\mathbf{w} + \ \mathrm{const}.\end{aligned}$$

- これは(10.160)と同じ形。
  - 前の結果を流用できる。

変分パラメータの最適化

➢ よって再推定の式は以下である。

$$\left(\xi_n^{\mathrm{new}}\right)^2 = \boldsymbol{\phi}_n^{\mathrm{T}} \left(\boldsymbol{\Sigma}_N + \boldsymbol{\mu}_N \boldsymbol{\mu}_N^{\mathrm{T}}\right) \boldsymbol{\phi}_n$$