W8PRML 3.4-3.6

東工大　山本航平

目次

3.4 ベイズモデル比較

- モデルの選択や正則化パラメータの決定
  - 1章ではクロスバリデーションを用いてモデルの選択を行った
  - ここではベイズの立場から考える
- ここでは一般論を述べ，3.5では線形回帰の正則化パラメータの具体例を述べる

L個のモデルを比較

- L個のモデル$\{\mathcal{M}_i\}(i = 1, \ldots, L)$
  - このモデルはデータ$\mathcal{D}$上の確率分布
  - 目的値$\mathbf{t}$上の分布，入力$\mathbf{X}$は既知とする
- このモデルのどれかによってデータは生成されているが
  そのどれかはわからない

- このとき，データ集合が与えられた時の事後分布を評価する

$$p(\mathcal{M}_i|\mathcal{D}) \propto p(\mathcal{M}_i)p(\mathcal{D}|\mathcal{M}_i) \quad (3.66)$$

モデルに対する好み
（ここでは定数と仮定）

モデルエビデンス
(周辺尤度)

モデルエビデンス(周辺尤度)の評価

- ベイズ因子(＝エビデンスの比) $\dfrac{p(\mathcal{D}|\mathcal{M}_i)}{p(\mathcal{D}|\mathcal{M}_j)}$
- モデルの事後分布がわかると予測分布がわかる

$$p(t|\mathbf{x},\mathcal{D}) = \sum_{i=1}^{L} p(t|\mathbf{x},\mathcal{M}_i,\mathcal{D})p(\mathcal{M}_i|\mathcal{D}) \quad (3.67)$$

- この分布は混合分布である
  - 全体の予測分布が個々のモデルの予測分布の事後確率に関する重み付き平均で与えられる

＝モデル平均(の近似)を求めたい

モデル選択

- モデル平均のもっとも単純な近似

→一番もっともらしいモデルを選ぶ（＝モデル選択）

- エビデンスは以下で与えられる（$\mathbf{w}$はモデルのパラメータ）

$$p(\mathcal{D}|\mathcal{M}_i) = \int p(\mathcal{D}|\mathbf{w}, \mathcal{M}_i) p(\mathbf{w}|\mathcal{M}_i) \mathrm{d}\mathbf{w} \quad (3.68)$$

- モデルエビデンスはパラメータの事後分布を計算するときの分母

$$p(\mathbf{w}|\mathcal{D}, \mathcal{M}_i) = \frac{p(\mathcal{D}|\mathbf{w}, \mathcal{M}_i) p(\mathbf{w}|\mathcal{M}_i)}{p(\mathcal{D}|\mathcal{M}_i)} \quad (3.69)$$

モデルエビデンスの別の解釈

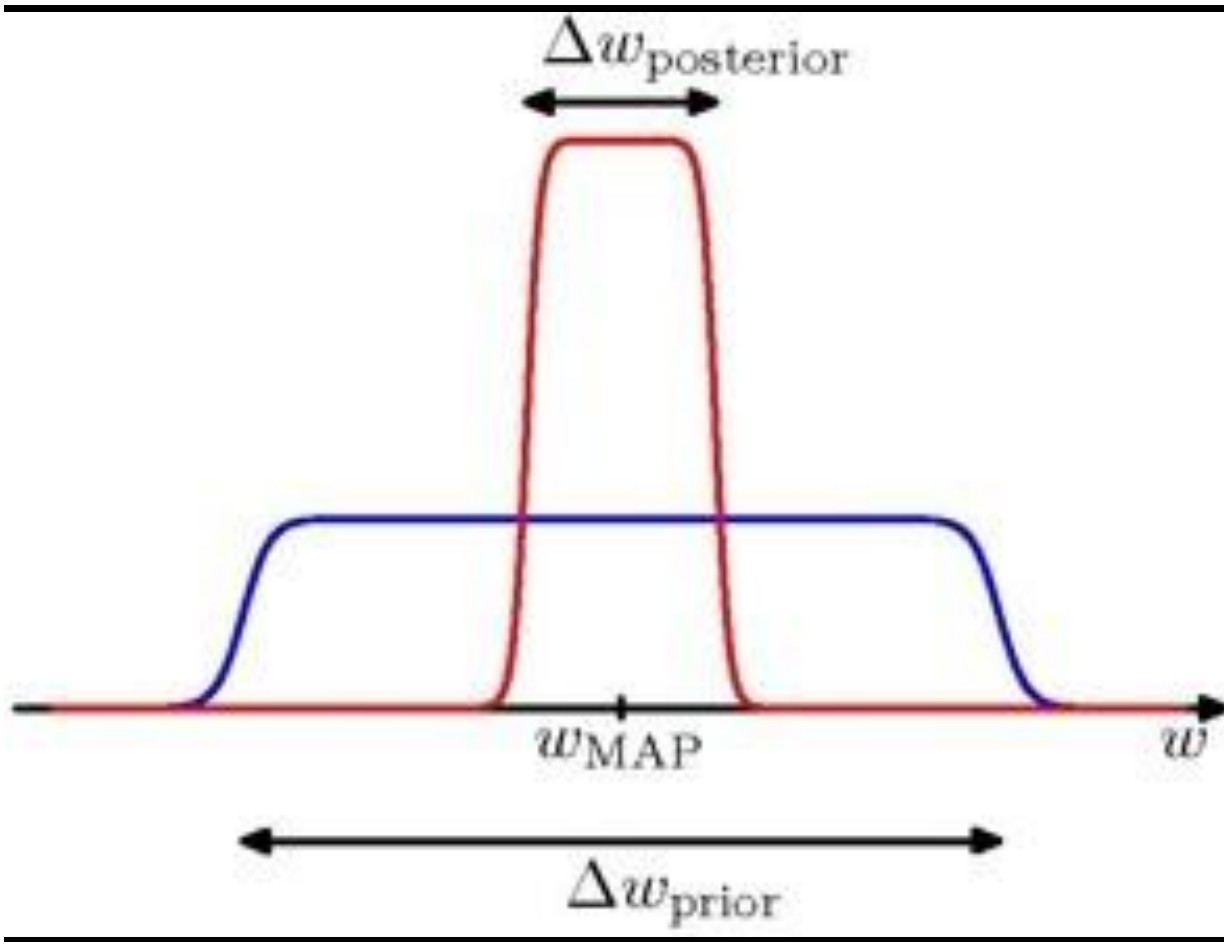
- パラメータがひとつの例 $p(\mathcal{D}) = \int p(\mathcal{D}|w)p(w)\mathrm{d}w$ (3.70')
- 仮定:
  - 事前確率は平坦で幅が $\Delta w_{\mathrm{prior}}$ ($p(w) = \frac{1}{\Delta w_{\mathrm{prior}}}$)
  - パラメータの事後分布が最頻値(w_{MAP})の近傍で尖っている
  - その幅が$\Delta w_{\mathrm{posterior}}$
- 仮定を用いると，掛け算で(3.70')を近似できる

$$p(\mathcal{D}) = \int p(\mathcal{D}|w)p(w)\mathrm{d}w \simeq p(\mathcal{D}|w_{\mathrm{MAP}})\frac{\Delta w_{\mathrm{posterior}}}{\Delta w_{\mathrm{prior}}} \quad (3.70)$$

- 対数をとる

$$\ln p(\mathcal{D}) \simeq \ln p(\mathcal{D}|w_{\mathrm{MAP}}) + \ln \frac{\Delta w_{\mathrm{posterior}}}{\Delta w_{\mathrm{prior}}} \quad (3.71)$$

近似の様子を図に示す



(3.71)の解釈

$$\ln p(\mathcal{D}) \simeq \ln p(\mathcal{D}|w_{\mathrm{MAP}}) + \ln \frac{\Delta w_{\mathrm{posterior}}}{\Delta w_{\mathrm{prior}}} \tag{3.71}$$

- 第一項
  - 一番尤もらしいパラメータによるデータへのフィッティング度
- 第二項
  - ペナルティ項
  - 事後分布がデータに強くフィットするようにパラメータを調整するとペナルティは大きくなる

M個のパラメータを含む場合

- 全てのパラメータにおける $\frac{\Delta w_{\text{posterior}}}{\Delta w_{\text{prior}}}$ の値が等しいとき，以下がなりたつ

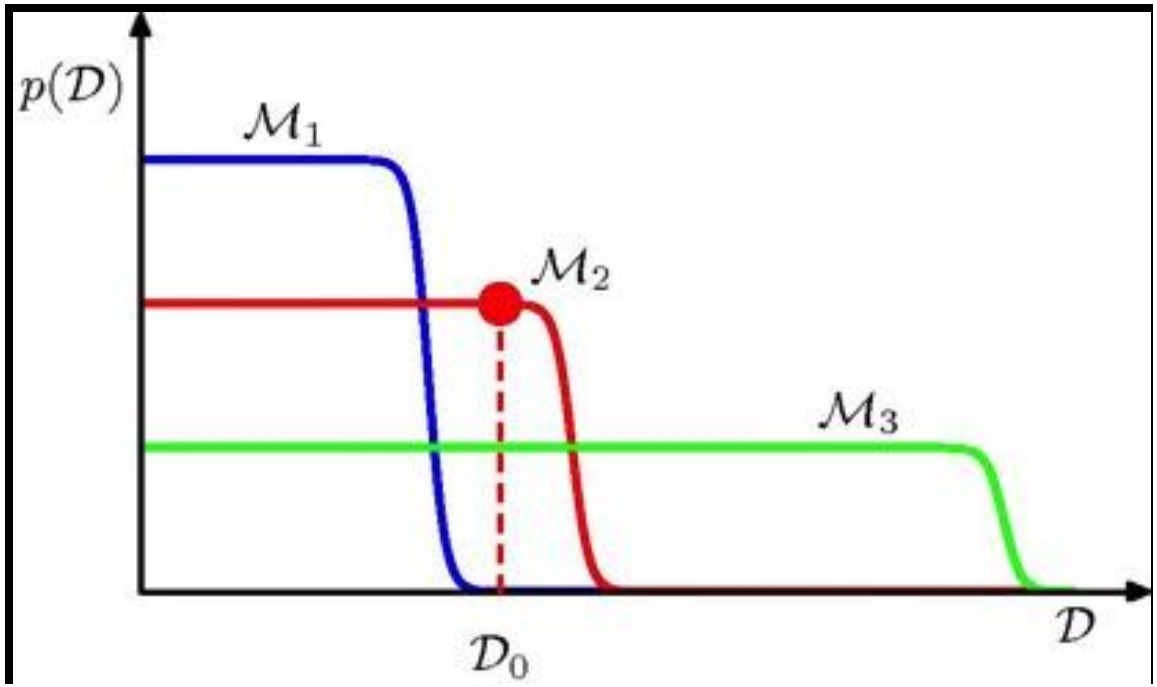
$$\ln p(\mathcal{D}) \simeq \ln p(\mathcal{D}|\mathbf{w}_{\text{MAP}}) + M \ln \left(\frac{\Delta w_{\text{posterior}}}{\Delta w_{\text{prior}}} \right) \qquad (3.72)$$

- モデルの複雑さを増すと…
  - 第一項は増加
  - 第二項は減少

バランスが大事

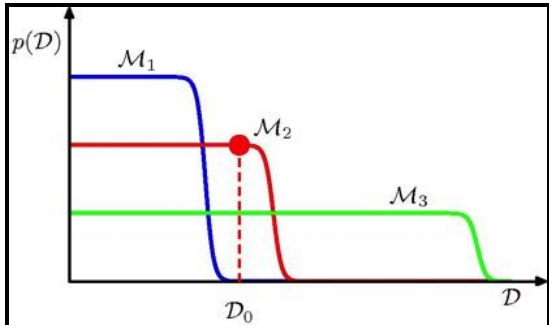
さらなる解釈

- 3つの，複雑さの異なるモデルを考えた時，なぜ周辺尤度最大化で中間の複雑さのモデルが選ばれるのか



横軸：
各データ集合空間

さらなる解釈



- データ$\mathcal{D}$の生成
  - パラメータのを事前分布$p(\mathbf{w})$に従って選択
  - それに基づいてデータを$p(\mathcal{D}|\mathbf{w})$からサンプリング
  - 単純なモデル（M_1）は多様性に乏しい（狭い）
  - 複雑なモデル（M_3）はデータの範囲が広い

- 単純なモデル→D_0を生成できない
- 複雑なモデル→D_0が選ばれることが相対的に少ない

ベイズモデル比較の枠組み

- ベイズモデル比較においては，
  モデル集合の中に真の分布が含まれていると仮定

- 2つのモデル$\mathcal{M}_1, \mathcal{M}_2$を考える（前者が真のモデル）
- 期待ベイズ因子

$$\int p(\mathcal{D}|\mathcal{M}_1) \ln \frac{p(\mathcal{D}|\mathcal{M}_1)}{p(\mathcal{D}|\mathcal{M}_2)} \mathrm{d}\mathcal{D} \qquad (3.73)$$

  - KLダイバージェンスの例
  - 2つの分布が等しい時にだけ0，それ以外は常に正
  - 従って，平均的には常に正しいモデルのベイズ因子のほうが大きい
- ベイズの枠組みでは
  - モデルの形に関して仮定をおいているのでテスト集合を用意して
    性能評価するべき

3.5 エビデンス近似

- 線形基底関数モデルを完全にベイズ的に取り扱う
  - 超パラメータα, βの導入
  - 通常のパラメータだけでなく，超パラメータも周辺化して予測

→全ての変数の上で完全に積分するのは難しい

- 周辺尤度関数を最大にするように超パラメータを決める
  - 経験ベイズ，第二種の最尤推定，一般化最尤推定，エビデンス近似

予測分布(αとβとw全部を周辺化)

$$p(t|\mathbf{t}) = \int\int\int p(t|\mathbf{w},\beta)p(\mathbf{w}|\mathbf{t},\alpha,\beta)p(\alpha,\beta|\mathbf{t})\mathrm{d}\mathbf{w}\mathrm{d}\alpha\mathrm{d}\beta \qquad (3.74)$$

- ただし,

$$p(t|\mathbf{x},\mathbf{w},\beta) = \mathcal{N}(t|y(\mathbf{x},\mathbf{w}),\beta^{-1}) \qquad (3.8)$$

$$p(\mathbf{w}|\mathbf{t}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_N,\mathbf{S}_N) \qquad (3.49)$$

- 事後分布 $p(\alpha,\beta|\mathbf{t})$ が $\hat{\alpha},\hat{\beta}$ の近傍で鋭く尖っている時

$$p(t|\mathbf{t}) \simeq p(t|\mathbf{t},\hat{\alpha},\hat{\beta}) = \int p(t|\mathbf{w},\hat{\beta})p(\mathbf{w}|\mathbf{t},\hat{\alpha},\hat{\beta})\mathrm{d}\mathbf{w} \qquad (3.75)$$

尖ってるα, βの場所を求める

- $\hat{\alpha}, \hat{\beta}$ の事後分布は

$$p(\alpha, \beta|\mathbf{t}) \propto p(\mathbf{t}|\alpha, \beta)p(\alpha, \beta) \qquad (3.76)$$

- 事前分布が平坦のとき
- エビデンス関数 $p(\mathbf{t}|\alpha, \beta)$を最大化することで, $\hat{\alpha}, \hat{\beta}$を求める

- 対数エビデンスを最大化する方法
  - EMアルゴリズム
  - 解析的に求める(3.5.2節)

3.5.1 エビデンス関数の評価

- エビデンス関数は以下の通り

$$p(\mathbf{t}|\alpha,\beta) = \int p(\mathbf{t}|\mathbf{w},\beta)p(\mathbf{w}|\alpha)\mathrm{d}\mathbf{w} \tag{3.77}$$

- これをごにょごにょします
- (3.11), (3.12), (3.52)よりエビデンス関数は以下の様に書ける

$$p(\mathbf{t}|\alpha,\beta) = \left(\frac{\beta}{2\pi}\right)^{\frac{N}{2}} \left(\frac{\alpha}{2\pi}\right)^{\frac{M}{2}} \int \exp\{-E(\mathbf{w})\}\mathrm{d}\mathbf{w} \tag{3.78}$$

- ただし，Mはパラメータwの次元数であり，

$$\begin{aligned} E(\mathbf{w}) =& \beta E_D(\mathbf{w}) + \alpha E_W(\mathbf{w}) \\ =& \frac{\beta}{2}\|\mathbf{t} - \Phi\mathbf{w}\|^2 + \frac{\alpha}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w} \end{aligned} \tag{3.79}$$

(3.78)の確認（演習3.17）

- 3.11, 3.12から $$p(\mathbf{t}|\mathbf{w},\beta) = \exp\left\{\frac{N}{2}\ln\frac{\beta}{2\pi} - \beta E_D(\mathbf{w})\right\}$$
$$= \left(\frac{\beta}{2\pi}\right)^{\frac{N}{2}} \exp\{-\beta E_D(\mathbf{w})\}$$
- 3.52から $$p(\mathbf{w}|\alpha) = \mathcal{N}(\mathbf{w}|\mathbf{0},\alpha^{-1}\mathbf{I}) = \left(\frac{\alpha}{2\pi}\right)^{\frac{M}{2}} \exp\left\{-\frac{1}{2}\mathbf{w}^{\mathrm{T}}(\alpha^{-1}\mathbf{I})^{-1}\mathbf{w}\right\}$$
- 従って

$$p(\mathbf{t}|\alpha,\beta) = \int p(\mathbf{t}|\mathbf{w},\beta)p(\mathbf{w}|\alpha)\mathrm{d}\mathbf{w}$$
$$= \left(\frac{\beta}{2\pi}\right)^{\frac{N}{2}} \left(\frac{\alpha}{2\pi}\right)^{\frac{M}{2}} \int \exp\left\{-\beta E_D(\mathbf{w}) - \alpha\frac{1}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w}\right\}$$

$E_W(\mathbf{W})$

$-E(\mathbf{w})$

(3.79)を平方完成(演習3.18)

$$E(\mathbf{w}) = E(\mathbf{m}_N) + \frac{1}{2}(\mathbf{w} - \mathbf{m}_N)^{\mathrm{T}}\mathbf{A}(\mathbf{w} - \mathbf{m}_N) \quad (3.80)$$

- ただし,

$$\mathbf{A} = \alpha\mathbf{I} + \beta\mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi} \quad (3.81)$$

$$E(\mathbf{m}_N) = \frac{\beta}{2}\|\mathbf{t} - \mathbf{\Phi}\mathbf{m}_N\|^2 + \frac{\alpha}{2}\mathbf{m}_N^{\mathrm{T}}\mathbf{m}_N \quad (3.82)$$

$$\mathbf{m}_N = \beta\mathbf{A}^{-1}\mathbf{\Phi}^{\mathrm{T}}\mathbf{t} \quad (3.84)$$

- また, $\mathbf{A} = \nabla\nabla E(\mathbf{w})$と対応している(ヘッセ行列)

- (3.54)から, $\mathbf{A} = \mathbf{S}_N^{-1}$が成り立つ→(3.84)は事後分布の平均

- 以上から(演習3.19help!)

$$\int \exp\{-E(\mathbf{w})\}\,\mathrm{d}\mathbf{w}$$
$$= \exp\{-E(\mathbf{m}_N)\} \int \exp\left\{\frac{1}{2}(\mathbf{w}-\mathbf{m}_N)^{\mathrm{T}}\mathbf{A}(\mathbf{w}-\mathbf{m}_N)\right\}$$
$$= \exp\{-E(\mathbf{m}_N)\}\,(2\pi)^{\frac{M}{2}}\,|\mathbf{A}|^{-\frac{1}{2}} \qquad (3.85)$$

- 対数周辺尤度は

$$\ln p(\mathbf{t}|\alpha,\beta) = \frac{M}{2}\ln\alpha + \frac{N}{2}\ln\beta - E(\mathbf{m}_N) - \frac{1}{2}\ln|\mathbf{A}| - \frac{N}{2}\ln 2\pi \qquad (3.86)$$

演習(3.18)

- まず展開

$$\frac{\beta}{2}\|\mathbf{t}-\mathbf{\Phi w}\|^2+\frac{\alpha}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w}$$

$$=\frac{\beta}{2}\left(\mathbf{t}^{\mathrm{T}}\mathbf{t}-2\mathbf{t}^{\mathrm{T}}\mathbf{\Phi w}+\mathbf{w}^{\mathrm{T}}\mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi w}\right)+\frac{\alpha}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w}$$

$$=\frac{1}{2}\left(\beta\mathbf{t}^{\mathrm{T}}\mathbf{t}-2\beta\mathbf{t}^{\mathrm{T}}\mathbf{\Phi w}+\mathbf{w}^{\mathrm{T}}(\beta\mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi}+\alpha\mathbf{I})\mathbf{w}\right)$$

A

- さらに式変形

$$\frac{1}{2}\left(\beta\mathbf{t}^{\mathrm{T}}\mathbf{t}-2\beta\mathbf{t}^{\mathrm{T}}\mathbf{\Phi w}+\mathbf{w}^{\mathrm{T}}\mathbf{A}\mathbf{w}\right)$$

$$=\frac{1}{2}\left(\beta\mathbf{t}^{\mathrm{T}}\mathbf{t}-2\beta\mathbf{t}^{\mathrm{T}}\mathbf{\Phi}\mathbf{A}^{-1}\mathbf{A w}+\mathbf{w}^{\mathrm{T}}\mathbf{A}\mathbf{w}\right)$$

$$=\frac{1}{2}\left(\beta\mathbf{t}^{\mathrm{T}}\mathbf{t}-2\mathbf{m}_N^{\mathrm{T}}\mathbf{A w}+\mathbf{w}^{\mathrm{T}}\mathbf{A}\mathbf{w}\right)$$

$$=\frac{1}{2}\left(\beta\mathbf{t}^{\mathrm{T}}\mathbf{t}-2\mathbf{m}_N^{\mathrm{T}}\mathbf{A w}+\mathbf{w}^{\mathrm{T}}\mathbf{A}\mathbf{w}+\mathbf{m}_N^{\mathrm{T}}\mathbf{A}\mathbf{m}_N-\mathbf{m}_N^{\mathrm{T}}\mathbf{A}\mathbf{m}_N\right)$$

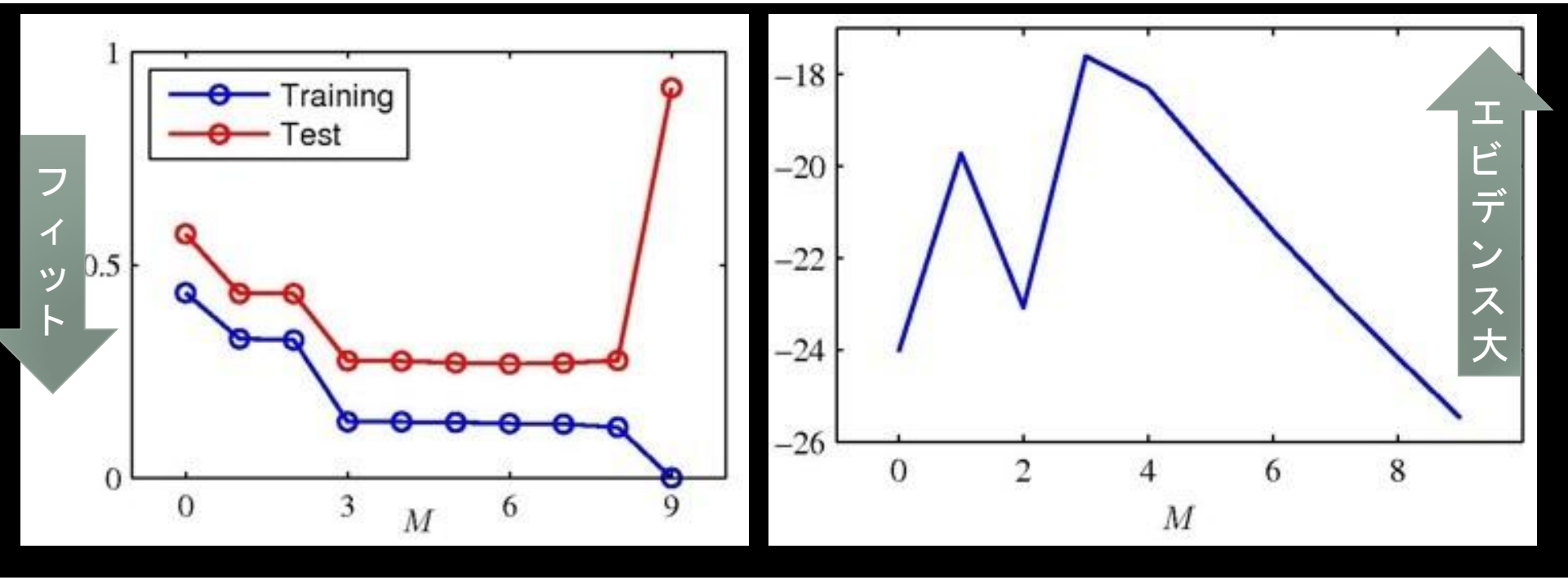
$$=\frac{1}{2}\left(\beta\mathbf{t}^{\mathrm{T}}\mathbf{t}-\mathbf{m}_N^{\mathrm{T}}\mathbf{A}\mathbf{m}_N\right)+\frac{1}{2}(\mathbf{w}-\mathbf{m}_N)^{\mathrm{T}}\mathbf{A}(\mathbf{w}-\mathbf{m}_N)$$

$(A^{-1})^{T}=A^{-1}$

- 第一項を更に変形

$$
\begin{aligned}
&\frac{1}{2}\left(\beta \mathbf{t}^{\mathrm{T}}\mathbf{t} - \mathbf{m}_N^{\mathrm{T}}\mathbf{A}\mathbf{m}_N\right) \\
&= \frac{1}{2}\left(\beta \mathbf{t}^{\mathrm{T}}\mathbf{t} - 2\mathbf{m}_N^{\mathrm{T}}\mathbf{A}\mathbf{m}_N + \mathbf{m}_N^{\mathrm{T}}\mathbf{A}\mathbf{m}_N\right) \\
&= \frac{1}{2}\left(\beta \mathbf{t}^{\mathrm{T}}\mathbf{t} - 2\mathbf{m}_N^{\mathrm{T}}\mathbf{A}\mathbf{A}^{-1}\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{t}\beta + \mathbf{m}_N^{\mathrm{T}}(\alpha\mathbf{I} + \beta\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi})\mathbf{m}_N\right) \\
&= \frac{1}{2}\left(\beta \mathbf{t}^{\mathrm{T}}\mathbf{t} - 2\mathbf{m}_N^{\mathrm{T}}\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{t}\beta + \beta\mathbf{m}_N^{\mathrm{T}}\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi}\mathbf{m}_N + \alpha\mathbf{m}_N^{\mathrm{T}}\mathbf{m}_N\right) \\
&= \frac{1}{2}\left(\beta(\mathbf{t} - \boldsymbol{\Phi}\mathbf{m}_N)^{\mathrm{T}}(\mathbf{t} - \boldsymbol{\Phi}\mathbf{m}_N) + \alpha\mathbf{m}_N^{\mathrm{T}}\mathbf{m}_N\right) \\
&= \frac{\beta}{2}\|\mathbf{t} - \boldsymbol{\Phi}\mathbf{m}_N\|^2 + \frac{\alpha}{2}\mathbf{m}_N^{\mathrm{T}}\mathbf{m}_N
\end{aligned}
$$

回帰の問題に戻ります



- 左図1.4, 右図3.14
- エビデンス最大はM=3（データを説明できる最も簡単なモデル）
  （よくデータにフィットしておりかつ複雑さが一番小さいから）

3.5.2 エビデンス関数の最大化

- $p(\mathbf{t}|\alpha,\beta)$をαに関して最大化する問題を定義
- まず $\ln|\mathbf{A}|$ のαに関する導関数を求めるために以下の固有ベクトル方程式を考える

$$\left(\beta\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi}\right)\mathbf{u}_i = \lambda_i\mathbf{u}_i \qquad (3.87)$$

- (3.81)より, $\mathbf{A}$は固有値$\alpha+\lambda_i$を持つ. したがって

$$\frac{\mathrm{d}}{\mathrm{d}\alpha}\ln|\mathbf{A}| = \frac{\mathrm{d}}{\mathrm{d}\alpha}\ln\prod_i(\lambda_i+\alpha) = \frac{\mathrm{d}}{\mathrm{d}\alpha}\sum_i\ln(\lambda_i+\alpha) = \sum_i\frac{1}{\lambda_i+\alpha} \qquad (3.88)$$

- (3.86)をαで微分して，0とおく

$$0 = \frac{M}{2\alpha} - \frac{1}{2}\mathbf{m}_N^{\mathrm{T}}\mathbf{m}_N - \frac{1}{2}\sum_i \frac{1}{\lambda_i + \alpha} \qquad (3.89)$$

- 両辺に2αをかけて整理

$$\alpha\mathbf{m}_N^{\mathrm{T}}\mathbf{m}_N = M - \alpha\sum_i \frac{1}{\lambda_i + \alpha} = \gamma \qquad (3.90)$$

- ΣにはM個の項が含まれるので

$$\gamma = \sum_i \frac{\lambda_i}{\lambda_i + \alpha} \qquad (3.91)$$

- 周辺尤度を最大にするαは

$$\alpha = \frac{\gamma}{\mathbf{m}_N^{\mathrm{T}}\mathbf{m}_N} \qquad (3.92)$$

αの求め方

1. （式(3.53, 3.91)）から $\mathbf{m}_N, \gamma$を始めに求める
   （これらはαに依存する）
2. (3.92)を使ってαを最推定
3. この仮定を収束するまで続ける
   - $\mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi}$ は変化しないので，固有値を最初に選ぶだけで良い

   - βに関しても同じ手順で求めることができる

βの場合

- ※λ_iがβに比例することに注意する $$\frac{\mathrm{d}\lambda_i}{\mathrm{d}\beta} = \frac{\lambda_i}{\beta}$$

$$\frac{\mathrm{d}}{\mathrm{d}\beta}\ln|\mathbf{A}| = \frac{\mathrm{d}}{\mathrm{d}\beta}\sum_i \ln(\lambda_i+\alpha) = \frac{1}{\beta}\sum_i \frac{1}{\lambda_i+\alpha} = \frac{\gamma}{\beta} \quad (3.93)$$

- よって

$$0 = \frac{N}{2\beta} - \frac{1}{2}\sum_{n=1}^{N}\left\{t_n - \mathbf{m}_N^{\mathrm{T}}\phi(\mathbf{x}_n)\right\}^2 \quad (3.94)$$

- 整理すると

$$\frac{1}{\beta} = \frac{1}{N-\gamma}\sum_{n=1}^{N}\left\{t_n - \mathbf{m}_N^{\mathrm{T}}\phi(\mathbf{x}_n)\right\}^2 \quad (3.95)$$

3.5.3 有効パラメータ数

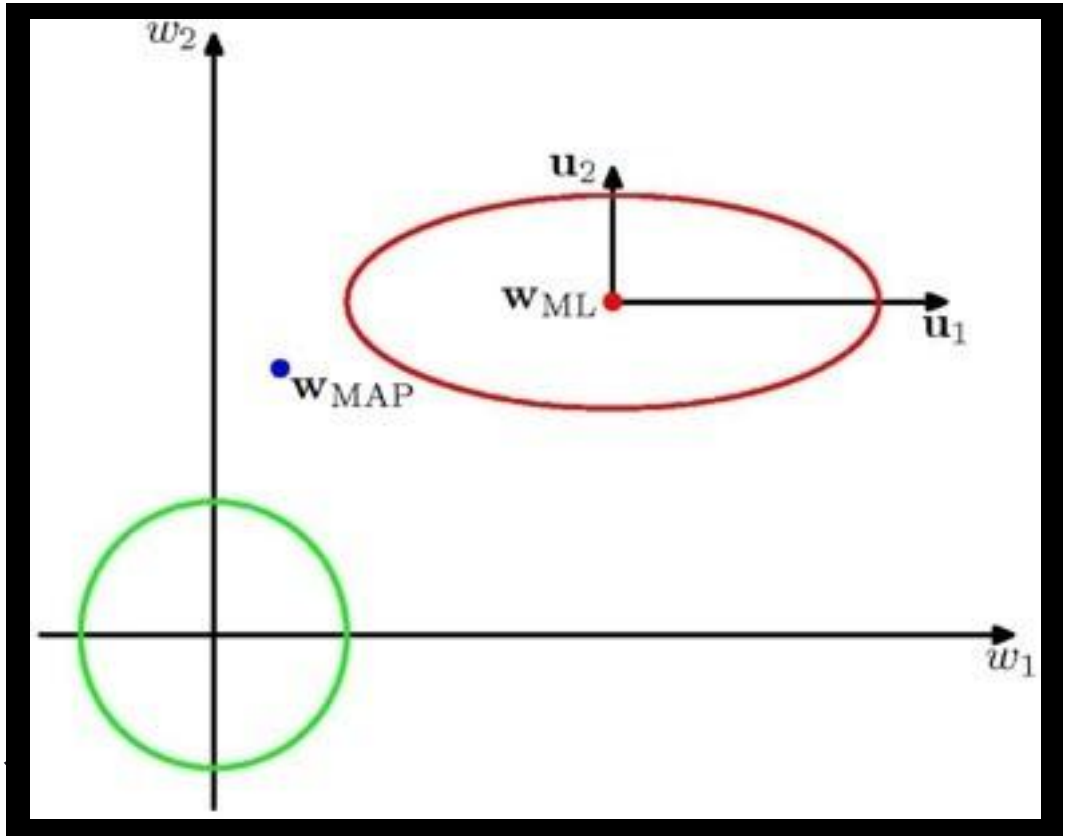
- ベイズ解αの解釈 $\alpha = \dfrac{\gamma}{\mathbf{m}_N^{\mathrm{T}}\mathbf{m}_N}$, $\gamma = \sum_i \dfrac{\lambda_i}{\lambda_i + \alpha}$
- 尤度関数と事前分布の等高線を描く（図3.15, 次のスライド）
  - 固有値は全て正（$\beta\mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi}$が正定値行列）なので

$$0 \leq \frac{\lambda_i}{\lambda_i + \alpha} \leq 1 \qquad 0 \leq \gamma \leq M$$

- $\lambda_i \gg \alpha \rightarrow \dfrac{\lambda_i}{\lambda_i + \alpha} \simeq 1$ →γによく影響する（well-determined）
- $\lambda_i \ll \alpha \rightarrow \dfrac{\lambda_i}{\lambda_i + \alpha} \simeq 0$→γにあまり影響しない

- したがって，γは有効(well-determined)なパラメータの数

図3.15



- 緑：事前分布
- 赤：尤度関数
- ここでは$\lambda_1 < \alpha < \lambda_2$

- λ_2はλ_1よりも影響が大きい

→λ_2のほうが最尤に近い

- ぼくにはよくわからない

β(3.95)の解釈

- 1変数xのガウス分布の分散の最尤推定値（式1.56より）

$$\sigma^2_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N}(x_n - \mu_{\mathrm{ML}})^2 \qquad (3.96)$$

- μ_{ML} はデータに含まれるノイズまでフィットしているので，上式はバイアスをもつ
→平均の推定に自由度の1つを使ってしまっている
したがって，分散の不偏推定量は

$$\sigma^2_{\mathrm{MAP}} = \frac{1}{N-1}\sum_{n=1}^{N}(x_n - \mu_{\mathrm{ML}})^2 \qquad (3.97)$$

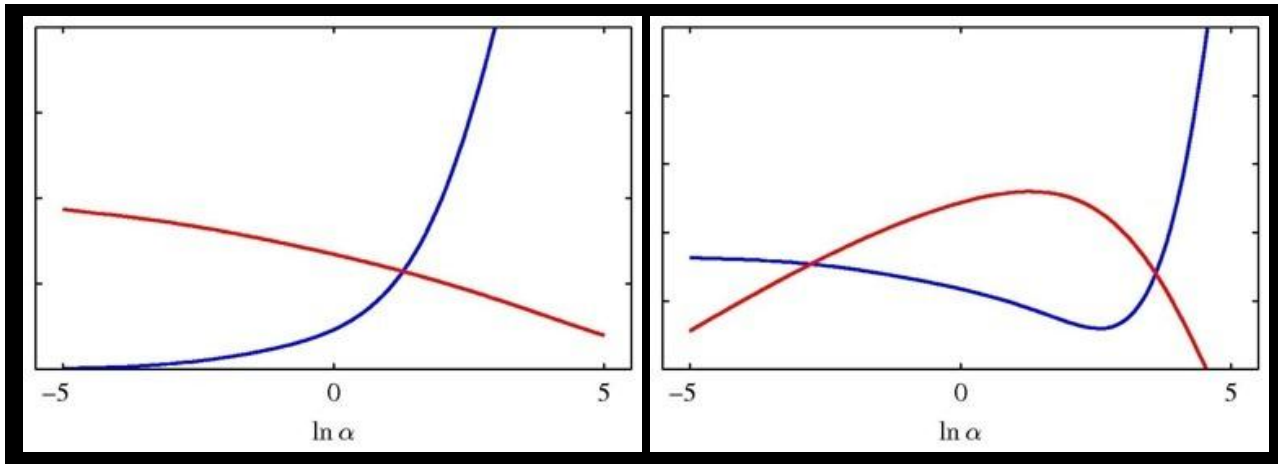
バイアスを取り
除くための”-1”

- 線形回帰の場合，有効なパラメータ数はγ
- 式(3.95)では，そのぶんの補正をおこなっている

$$\frac{1}{\beta} = \boxed{\frac{1}{N-\gamma}} \sum_{n=1}^{N} \left\{ t_n - \mathbf{m}_N^{\mathrm{T}} \phi(X_n) \right\}^2 \qquad (3.95)$$

例:
三角関数をガウス基底関数モデルで近似

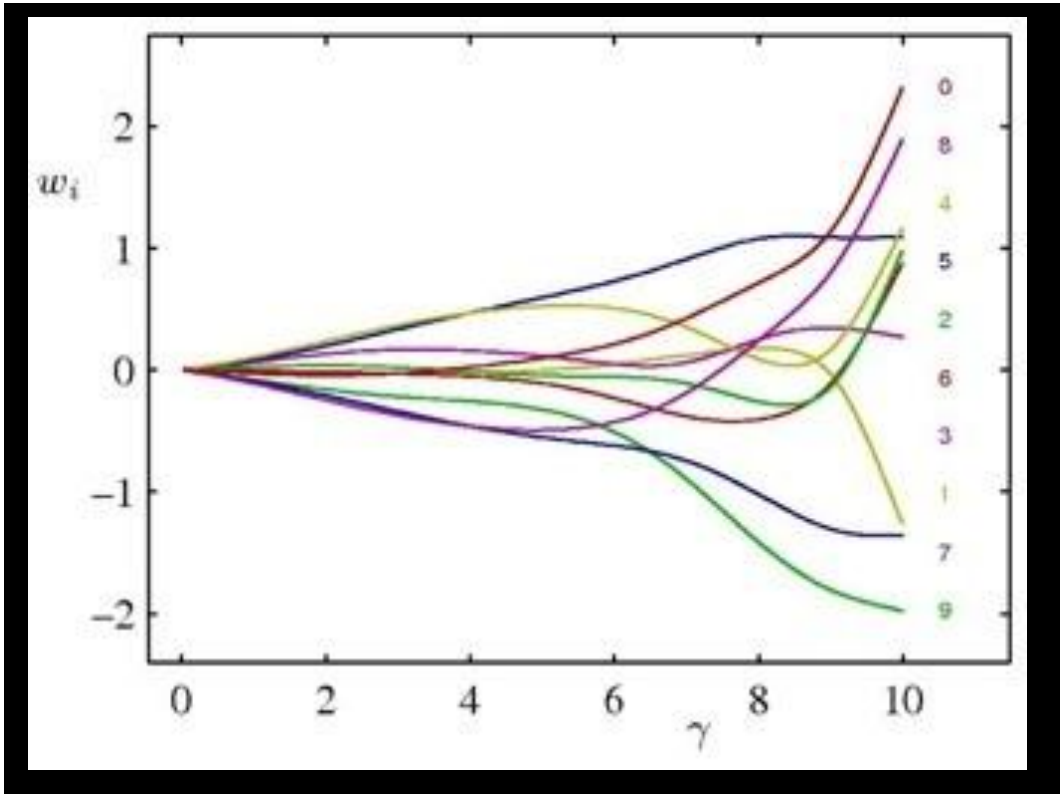
- 9個の基底関数を使う　M=10（9＋バイアスパラメータ）
- βを真の値11.1に設定したあとに，エビデンスの枠組みでαを決定する
- 左：青線→γ，赤線→$2E_W(\mathbf{m}_N)$，交点が最適なαの値
- 右：青線→テスト集合に対する誤差，赤線→対数エビデンス $\ln p(\mathbf{t}|\alpha,\beta)$
- 右図は，（左図の）交点の位置で誤差最小，
  エビデンス最大になっていることがわかる



各パラメータの値とγの関係

- 各グラフが各パラメータに対応
- αを$0 \leq \alpha \leq \infty$の範囲で変化させると，γが $0 \leq \gamma \leq \mathrm{M}$ の範囲で変化する

- αの値がパラメータの大きさを制御している



データ点の数がパラメータの数と比べて十分大きい時

- $N \gg M$ の時，全てのパラメータはwell-determined
  - 式(3.87)における $\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi}$ はデータ点に関する陰的な和を含んでいるため，データ集合のサイズとおもに固有値の数も増加するため
- このとき，$\gamma = M$
- すると，αとβの再推定方程式は

$$\alpha = \frac{M}{2E_W(\mathbf{m}_N)} \qquad \beta = \frac{N}{2E_D(\mathbf{m}_N)} \qquad (3.98), (3.99)$$

- ただし，$E_W(\mathbf{m}_N), E_D(\mathbf{m}_N)$ は式(3.25)(3.26)で定義した

$$E_W(\mathbf{m}_N) = \frac{1}{2}\mathbf{m}_N^{\mathrm{T}}\mathbf{m}_N \qquad E_D(\mathbf{m}_N) = \frac{1}{2}\sum_{n=1}^{N}\left\{t_n - \mathbf{w}^{\mathrm{T}}\phi(\mathbf{x}_n)\right\}^2$$

3.6 固定された基底関数の限界

- メリット
  - パラメータの線形性を仮定している
    →最小二乗問題の閉じた解が求まる
  - ベイズ推定の計算が簡単になる
  - 基底関数を適切に選べば，任意の非線形変換をモデル化できる
- デメリット
  - 訓練データの観測前に基底関数を固定する
  - 入力空間の次元数に対して，指数的に基底関数を増やす必要がある
- デメリットの解決法→実際のデータがもつ性質を利用する
  - 実データは限られた非線形多様体上に存在する
    - 入力空間中のデータがある場所にのみ基底関数を配置できる
  - 目標変数がデータ多様体中の少数の可能な方向にしか強く依存しない

3.86をβで微分

$$\ln p(\mathbf{t}|\alpha,\beta) = \frac{M}{2}\ln\alpha + \frac{N}{2}\ln\beta - E(\mathbf{m}_N) - \frac{1}{2}\ln|\mathbf{A}| - \frac{N}{2}\ln 2\pi \tag{3.86}$$

$$E(\mathbf{m}_N) = \frac{\beta}{2}\|\mathbf{t} - \Phi\mathbf{m}_N\|^2 + \frac{\alpha}{2}\mathbf{m}_N^{\mathrm{T}}\mathbf{m}_N$$

$$\mathbf{m}_N = \beta\mathbf{A}^{-1}\Phi^{\mathrm{T}}\mathbf{t}$$

- というかこれの微分→展開しただけだった