

PRML輪講3.3～3.6

澤田研究室 M2

小林祐也

3.3 ベイズ線形回帰

モデルの振る舞い、複雑さを決定するうえで基底関数の数や形が重要となる。

→解くべき問題に応じて、モデルの複雑さを決定することが肝要

$$y(\mathbf{x}, \mathbf{w}) = w_0 + \sum_{j=1}^{M-1} w_j \phi_j(\mathbf{x})$$

- ・尤度関数の最大化では複雑なモデルになりすぎて過学習してしまう

→線形回帰モデルをベイズ的に収め扱うことにより

最尤推定の過学習の回避、訓練データのみでモデルの複雑さを決定する。

3.3.1 パラメータの分布 (線形回帰モデルのベイズ推論の流れ)

- モデルパラメータ \mathbf{w} の事前確率分布

尤度関数 $p(\mathbf{t}|\mathbf{w})$ は \mathbf{w} の二次の指数関数(gauss) \rightarrow 共役事前分布 $p(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_0, \mathbf{S}_0)$ (期待値 \mathbf{m}_0 , 共分散 \mathbf{S}_0)

$$p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \beta) = \prod_{n=1}^N \mathcal{N}(t_n|\mathbf{w}^T \phi(\mathbf{x}_n), \beta^{-1})$$

- 事後分布(\propto 尤度関数 \times 事前分布)

$$p(\mathbf{w}|\mathbf{t}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_N, \mathbf{S}_N) \quad \cdots(3.49) \quad \mathbf{m}_N = \mathbf{S}_N(\mathbf{S}_0^{-1}\mathbf{m}_0 + \beta\Phi^T\mathbf{t})$$

$$\mathbf{S}_N^{-1} = \mathbf{S}_0^{-1} + \beta\Phi^T\Phi$$

\rightarrow モード: \mathbf{m}_N 事後確率を最大化するパラメータ: $\mathbf{w}_{MAP} = \mathbf{m}_N$ (\because 事後分布もガウス分布)

\rightarrow 無限に広い事前分布 $\mathbf{S}_0 = \alpha^{-1}\mathbf{I}$ ($\alpha \rightarrow 0$) の時

$$\mathbf{m}_N = (\Phi^T\Phi)^{-1}\Phi^T\mathbf{t} = \mathbf{w}_{ML} \quad (\text{最尤推定値(3.15)})$$

- 逐次的にデータinput、任意時点の事後分布が次のデータにおける事前分布となる (次の事後分布も(3.49)の形)

- 今後は簡単のため、以下を考える

- 単一の精度パラメータ α \cdot 期待値0の等方ガウス分布

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

- 事後分布は(3.49)と同型 ($\mathbf{m}_N = \beta\mathbf{S}_N\Phi^T\mathbf{t}$ $\mathbf{S}_N^{-1} = \alpha\mathbf{I} + \beta\Phi^T\Phi$)

- 対数事後分布は \propto 尤度関数 \times 事前分布より、 $\ln p(\mathbf{w}|\mathbf{t}) = -\frac{\beta}{2} \sum_{n=1}^N \{t_n - \mathbf{w}^T \phi(\mathbf{x}_n)\}^2 - \frac{\alpha}{2} \mathbf{w}^T \mathbf{w} + \text{定数}$

線形基底関数モデル→ベイズ学習と事後分布の逐次的な更新方法

e.g.) 直線fitting : 線形モデル $y(x, \mathbf{w}) = \omega_0 + \omega_1 x$

- ・ 関数 : $y(x, a) = a_0 + a_1 x$ ($a_0 = -0.3, a_1 = 0.5$)
- ・ 訓練データの目標値 $t_n : f(x_n, a)$ にガウスノイズを加えたもの ($x_n \in U(x | -1, 1)$)
- ・ 目標 : パラメータ a_0, a_1 の復元、データサイズと推定値との関係
- ・ 精度パラメータ $\beta = 25, \alpha = 2.0$

e.g.) 他形式の事前分布 : ガウス事前分布の一般化

$$p(\mathbf{w}|\alpha) = \left[\frac{q}{2} \left(\frac{\alpha}{2} \right)^{1/q} \frac{1}{\Gamma(1/q)} \right]^M \exp \left(-\frac{\alpha}{2} \sum_{j=0}^{M-1} |w_j|^q \right)$$

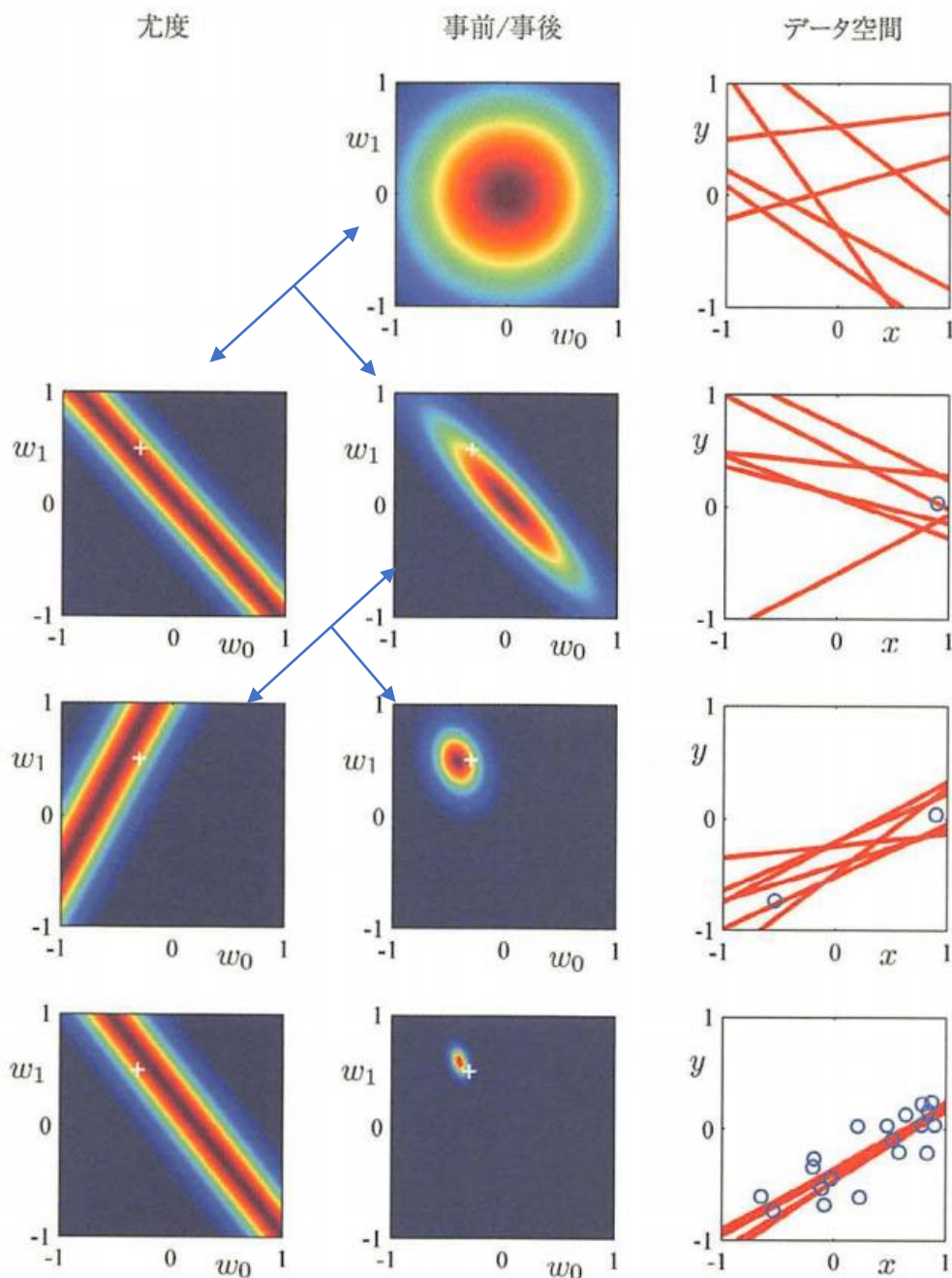
- ・ $q=2$ の時ガウス事前分布に一致

→ 尤度関数 $p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \beta) = \prod_{n=1}^N \mathcal{N}(t_n | \mathbf{w}^T \phi(\mathbf{x}_n), \beta^{-1})$ の共役事前分布

- ・ この時の 事後分布の \mathbf{w} に関する最大化 → 正則誤差関数(3,29)の最小化 に等しい

$$\frac{1}{2} \sum_{n=1}^N \{t_n - \mathbf{w}^T \phi(\mathbf{x}_n)\}^2 + \frac{\lambda}{2} \sum_{j=1}^M |w_j|^q$$

3.3.1 パラメータの分布 (線形回帰モデルのベイズ的取り扱い)



- 関数: $y(x, a) = a_0 + a_1 x$ ($a_0 = -0.3, a_1 = 0.5$)
 - 訓練データの目標値 $t_n: f(x_n, a)$ にガウスノイズを加えたもの ($x_n \in U(x| -1, 1)$)
 - 目標: パラメータ a_0, a_1 の復元、データサイズと推定値との関係
 - 精度パラメータ $\beta = 25, \alpha = 2.0$

- データ観測前: w 空間の事前分布とそこから抽出した6種の $y(x, w)$
- データ1つ観測後: 右図の○が1つ目のデータ点
 - 左図は尤度関数 $p(t|x, w)$
 - 白点が真の値 $a_0 = -0.3, a_1 = 0.5$
 - 事後分布 (中図 \propto 尤度関数(左図) \times 事前分布(上図))
 - 右図は事後分布からランダムに抽出した6種の $y(x, w)$ (データ点近く)
- 事後分布 = 3行目尤度関数 \times 直前の(2行目)事前分布
= 3行目 \times 2行目尤度関数 \times 1行目事前分布
→ 右図のデータ点と $y(x, w)$ (回帰関数のサンプル)を見るとかなり近くなっている
- 20点のデータを観測したとき
 - 左図はあくまで20点目のデータ点のみに関する尤度関数
 - 事後分布もかなり狭まり、真の w ($a_0 = -0.3, a_1 = 0.5$)に近い
→ ∞ データ点を用意すれば事後分布は δ 関数に収束

3.3.2 予測分布 (線形回帰モデルのベイズ的取り扱い)

実際問題パラメータ w よりも、入力値 x に対する目標値 t の予測が知りたい
→ 予測分布を評価する

予測分布: $p(t|\mathbf{t}, \alpha, \beta) = \int p(t|\mathbf{w}, \beta) p(\mathbf{w}|\mathbf{t}, \alpha, \beta) d\mathbf{w}$ \mathbf{t} は訓練データの目標値ベクトル

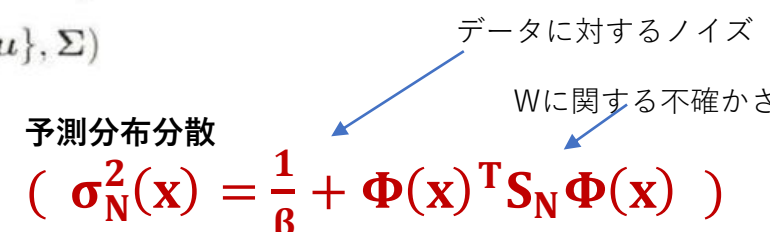
条件付き分布: $p(t|\mathbf{x}, \mathbf{w}, \beta) = \mathcal{N}(t|y(\mathbf{x}, \mathbf{w}), \beta^{-1})$

事後分布: $p(\mathbf{w}|\mathbf{t}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_N, \mathbf{S}_N)$

$$(2.115) \quad \begin{array}{ll} p(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1}) & \text{の時} \quad p(\mathbf{y}) = \mathcal{N}(\mathbf{y}|\mathbf{A}\boldsymbol{\mu} + \mathbf{b}, \mathbf{L}^{-1} + \mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^T) \\ p(\mathbf{y}|\mathbf{x}) = \mathcal{N}(\mathbf{y}|\mathbf{A}\mathbf{x} + \mathbf{b}, \mathbf{L}^{-1}) & p(\mathbf{x}|\mathbf{y}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\Sigma}\{\mathbf{A}^T\mathbf{L}(\mathbf{y} - \mathbf{b}) + \boldsymbol{\Lambda}\boldsymbol{\mu}\}, \boldsymbol{\Sigma}) \end{array}$$

よって

予測分布: $p(t|\mathbf{x}, \mathbf{t}, \alpha, \beta) = \mathcal{N}(t|\mathbf{m}_N^T \boldsymbol{\Phi}(\mathbf{x}), \sigma_N^2(\mathbf{x})) \rightarrow \left(\sigma_N^2(\mathbf{x}) = \frac{1}{\beta} + \boldsymbol{\Phi}(\mathbf{x})^T \mathbf{S}_N \boldsymbol{\Phi}(\mathbf{x}) \right)$



新しいデータが追加されると事後分布は狭まる $\rightarrow \therefore \sigma_{N+1}^2(\mathbf{x}) \leq \sigma_N^2(\mathbf{x})$

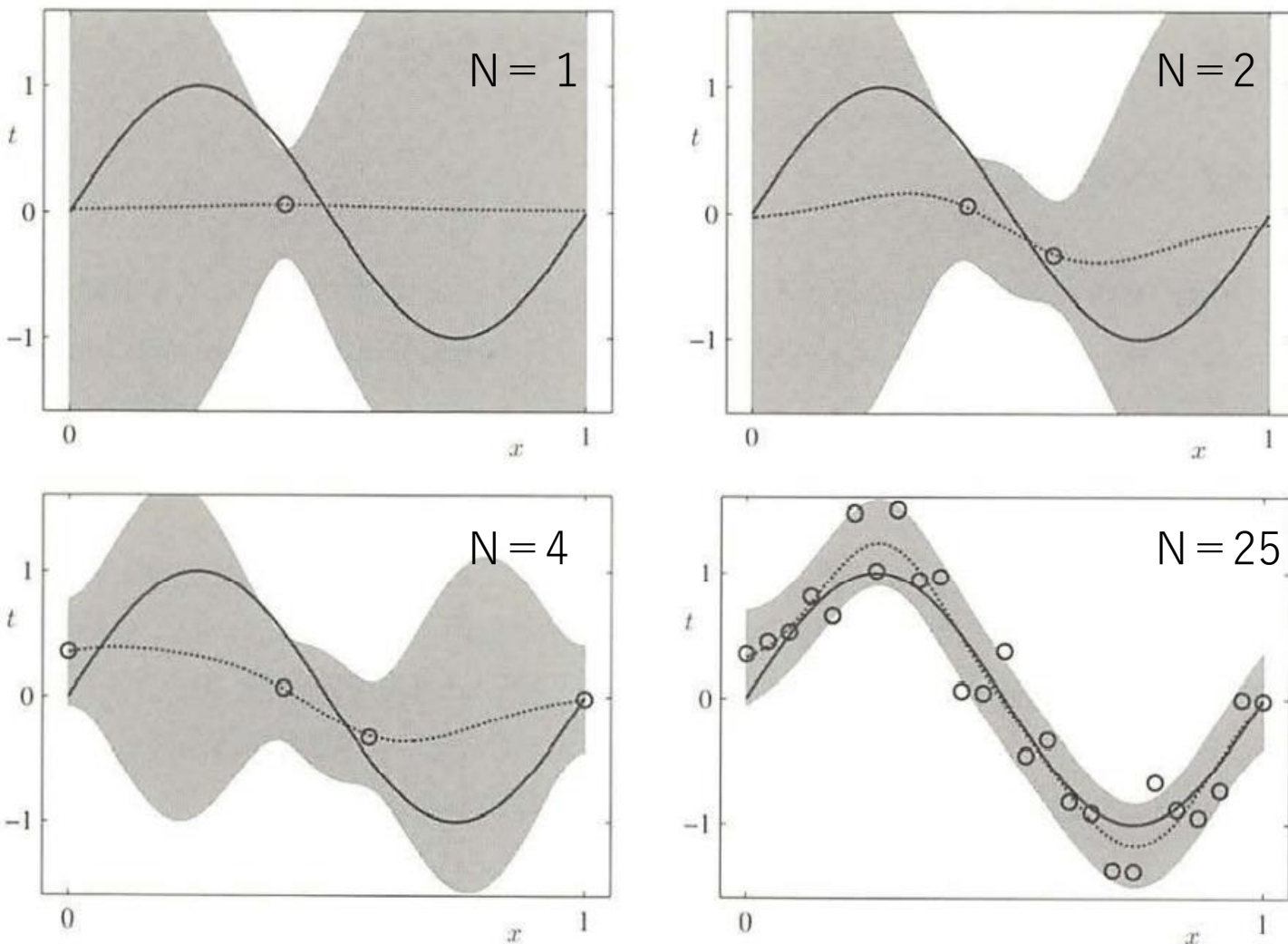
$\sigma_N^2(\mathbf{x})$ の第二項は $N \rightarrow \infty$ で0に収束するため、予測分布の分散は $N \rightarrow \infty$ でパラメータ β にのみ依存する

3.3.2 予測分布 (線形回帰モデルのベイズ的取り扱い)

ベイズ線形回帰モデルの予測分布

e.g.) $\sin(2\pi x)$ に対する予測分布、事後分布 (モデルはガウス関数9個 $\phi_j(x) = \exp\left\{-\frac{(x - \mu_j)^2}{2s^2}\right\}$)

図3.8



実線 : $\sin(2\pi x)$
点線 : ガウス予測分布の平均
灰色領域 : 予測分布における平均 \pm 標準誤差
○ : データ点

予測の不確かさはデータサイズに依存

各予測に対し

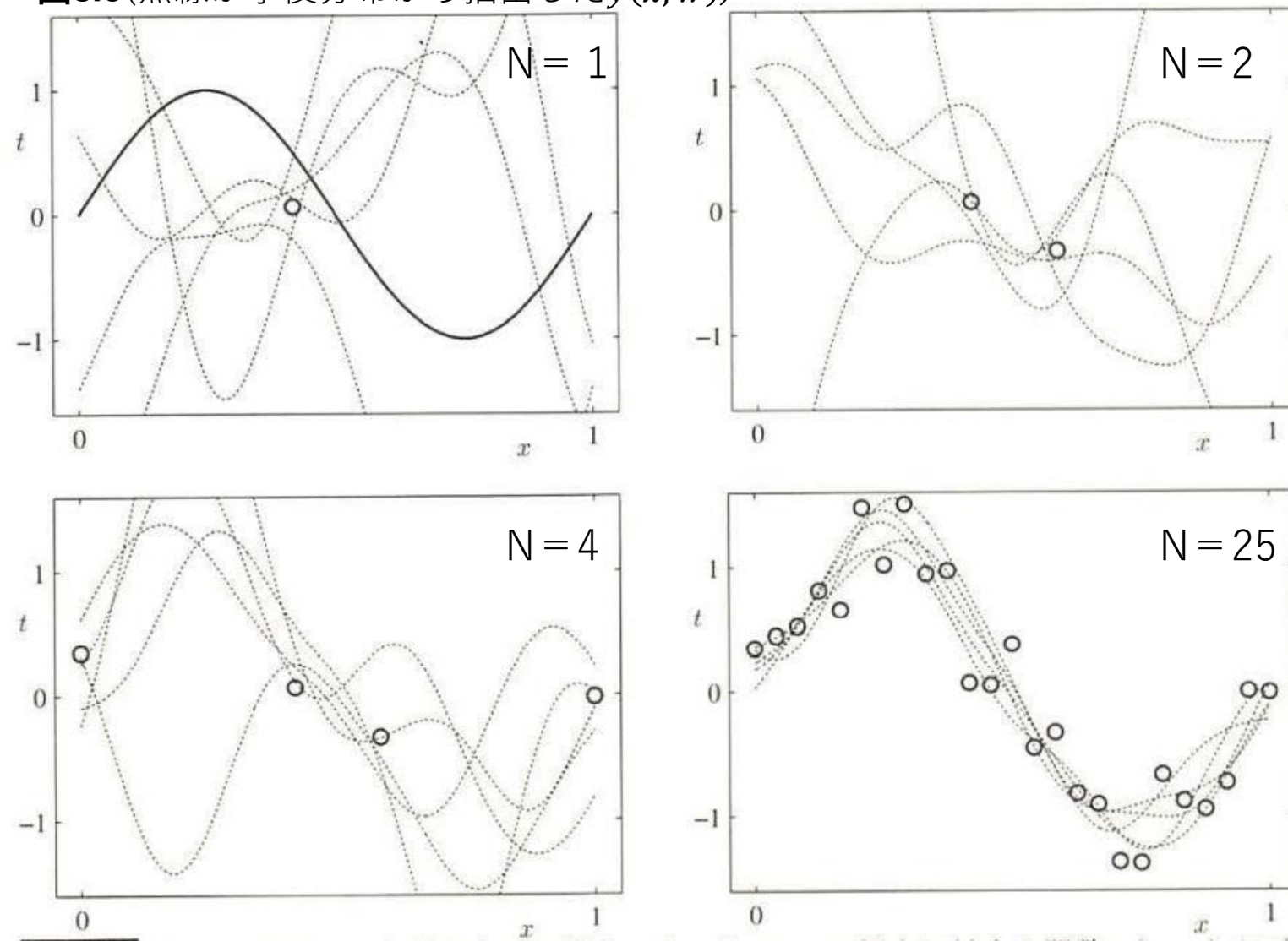
- ・ データ点近傍が最も不確かさが小さい
- ・ 不確かさは x に依存

3.3.2 予測分布 (線形回帰モデルのベイズ的取り扱い)

異なる x に対する予測値 $y(x, w)$ の共分散を調べたい

→ w の事後分布から w をいくつか抽出し対応する $y(x, w)$ をプロット

図3.9(点線が事後分布から抽出した $y(x, w)$)



局所的な基底関数の時 (ガウス関数)

→基底関数の中心から遠ざかるほど

$$\sigma_N^2(x) = \frac{1}{\beta} + \Phi(x)^T S_N \Phi(x)$$

予測分散の第2項の寄与が小さくなる

→基底関数の外側領域の補完 (外挿) の時
推定の信頼性が非常に高くなる
(望ましくはない)

→(6.4節)のガウス過程 (ベイズ的アプローチの1種)を使えば解決

3.3.3 等価カーネル(線形回帰モデルのベイズ的取り扱い)

カーネル法の導入 訓練データの目標値だけから予測

- ・ 線形基底関数モデルに対する事後分布の平均解 $m_N = \beta S_N \Phi^T t$
 $S_N^{-1} = S_0^{-1} + \beta \Phi^T \Phi$
 $y(\mathbf{x}, \mathbf{w}) = \mathbf{w}^T \Phi(\mathbf{x})$ より
- ・ 予測平均: $y(\mathbf{x}, m_N) = m_N^T \Phi(\mathbf{x}) = \beta \Phi(\mathbf{x})^T S_N \Phi^T t = \sum_{n=1}^N \beta \Phi(\mathbf{x})^T S_N \Phi^T(\mathbf{x}_n) t_n$

→ 点 \mathbf{x} での予測分布平均: $y(\mathbf{x}, m_N) = \sum_{n=1}^N k(\mathbf{x}, \mathbf{x}_n) t_n$ $k(\mathbf{x}, \mathbf{x}') = \beta \Phi(\mathbf{x})^T S_N \Phi^T(\mathbf{x}')$

この $k(\mathbf{x}, \mathbf{x}_n)$ を平滑化行列 (等価カーネル) と呼ぶ

→ 予測形式 y が**訓練データの目標値 t_n の線形結合**で表せる
この線形回帰 $y(\mathbf{x}, m_N)$ を線形平滑器と呼ぶ

※等価カーネル $k(\mathbf{x}, \mathbf{x}_n)$ は入力値 \mathbf{x}_n に依存

3.3.3 等価カーネル(線形回帰モデルのベイズ的取り扱い)

等価カーネル $k(\mathbf{x}, \mathbf{x}')$

- ・ $\mathbf{x}=\mathbf{x}'$ の周りに局在

ある \mathbf{x} での予測分布期待値 $y(\mathbf{x}, m_N)$

$$y(\mathbf{x}, m_N) = \sum_{n=1}^N k(\mathbf{x}, \mathbf{x}_n) t_n$$

- ・ 目標値の重み付き和で表せる。
- ・ 重みは \mathbf{x} 近傍のデータほど大きい

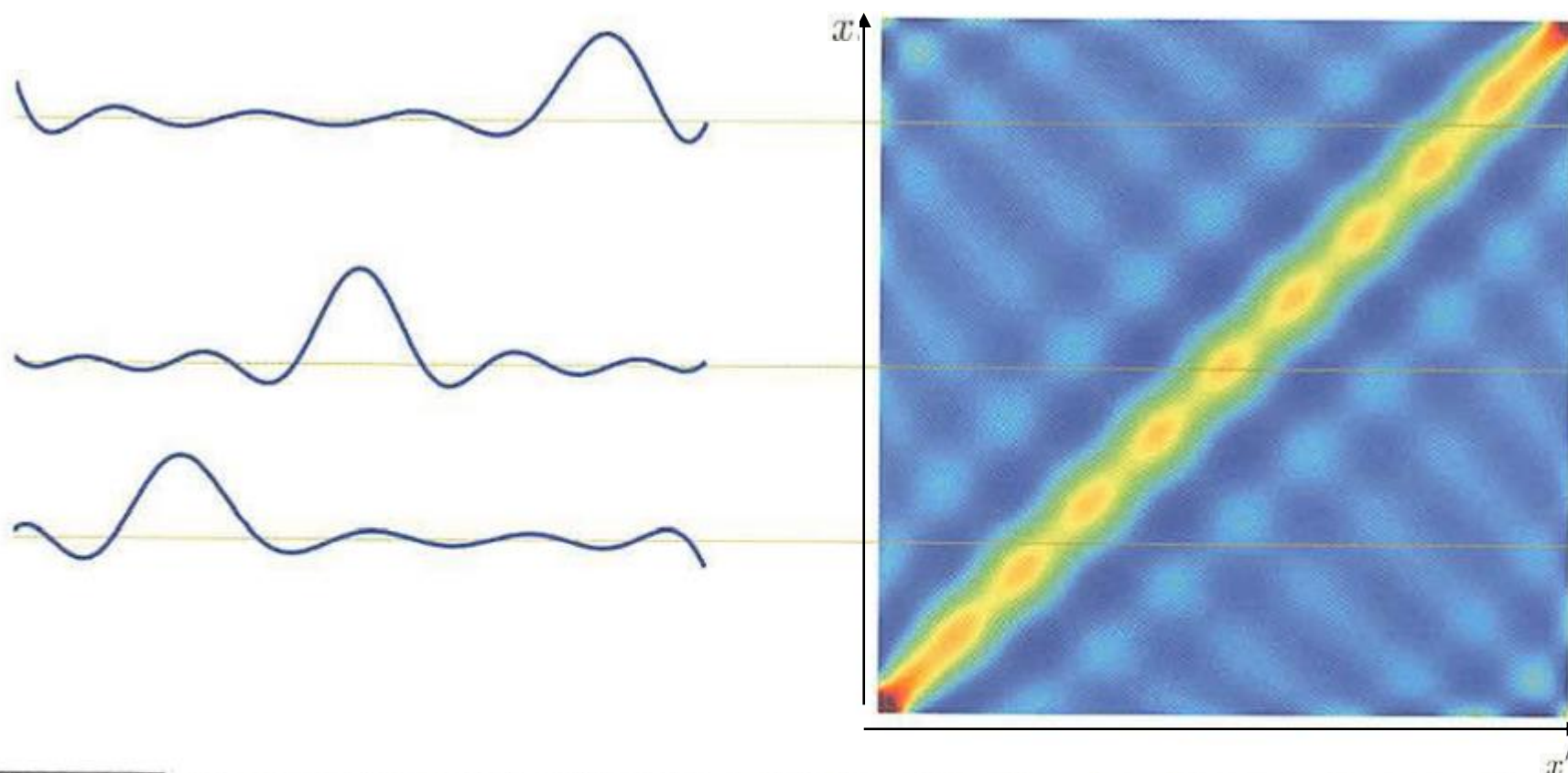


図 3.10 図 3.1 のガウス基底関数に対する等価カーネル $k(x, x')$. 右図の横軸は x' , 縦軸は x に対応している. 左側の 3 つのグラフは, この行列の 3 つの異なる x の値での切片である. このカーネルは, $(-1, 1)$ の区間の等間隔の 200 点の x の値からなるデータ集合を用いて生成した.

3.3.3 等価カーネル(線形回帰モデルのベイズ的取り扱い)

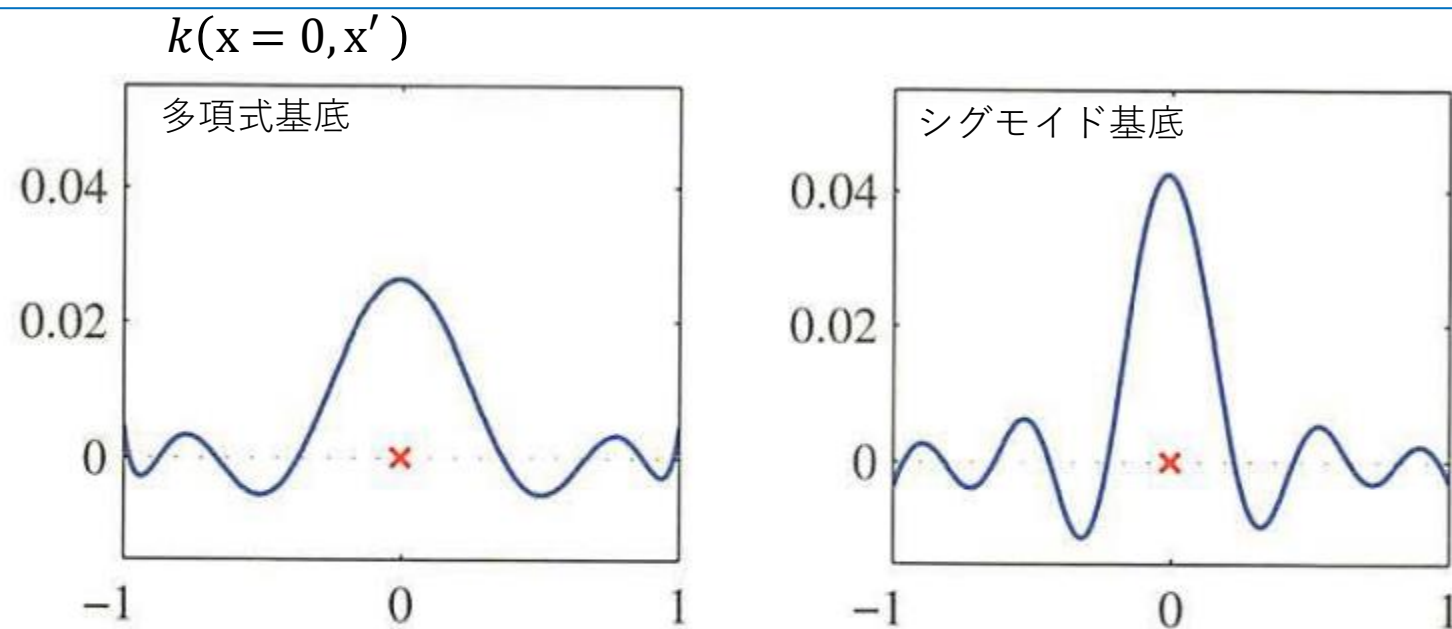


図 3.11 等価カーネル $k(x, x')$ の例. $x = 0$ に対して x' の関数としてプロットされている. 左は多項式基底関数, 右はシグモイド基底関数 (図 3.1 参照). 対応する基底関数は局所的でないにもかかわらず, 等価カーネルは x' に関して局所的な関数となっていることに注意.

等価カーネル $k(x, x')$

- $x=x'$ の周りに局在

ある x での予測分布期待値 $y(x, m_N)$

- 目標値の重み付き和で表せる。

- 重みは x 近傍のデータほど大きい

局所化の性質はガウス関数基底だけでない

- 多項式基底
- シグモイド基底

でも成り立つ

等価カーネルの役割

2点の y の相関： $y(\mathbf{x})$ と $y(\mathbf{x}')$ の共分散

$$\begin{aligned}\text{cov}[y(\mathbf{x}), y(\mathbf{x}')] &= \text{cov}[\phi(\mathbf{x})^T \mathbf{w}, \mathbf{w}^T \phi(\mathbf{x}')] \\ &= \phi(\mathbf{x})^T \mathbf{S}_N \phi(\mathbf{x}') = \beta^{-1} k(\mathbf{x}, \mathbf{x}')\end{aligned}$$

∴ $\mathbf{x}=\mathbf{x}'$ の近傍点での予測平均は互いに相関が大きく、より遠い点同士では相関が小さくなる

- ・基底関数の集合を固定化すれば等価カーネルが決まる。 $k(\mathbf{x}, \mathbf{x}') = \beta \phi(\mathbf{x})^T \mathbf{S}_N \phi(\mathbf{x}')$

→ **カーネル関数を用いた線形回帰問題の定式化**

- ・”基底関数の集合”ではなく、”局所的なカーネル”を直接定義することで
観測された訓練データ → カーネルを用いた予測 → 入力 \mathbf{x} に対する予測値

この定式化から6.4節で行う回帰に実用的な”**ガウス過程**”が得られる。

3.3.3 等価カーネル(線形回帰モデルのベイズ的取り扱い)

等価カーネル：訓練データの目標値 \mathbf{t}_n の和の重みづけに相当

- これらの重みの和はある仮定の下で1となり
$$\sum_{n=1}^N k(\mathbf{x}, \mathbf{x}_n) = 1$$

∴全ての n について $t_n = 1$ としたときの予測平均 $y^*(\mathbf{x})$ と等価であり、その時 $y^*(\mathbf{x})=1$

- 和が1という制約があったとしても
 - カーネル関数自体は、正・負どちらをとってもよい
 - 対応する予測器は訓練データの目標 \mathbf{t}_n の凸結合になるとは限らない

- カーネル関数が満たすべき重要な性質（非線形関数のベクトル $\Psi(\mathbf{x})$ の内積で表現）

→ $k(\mathbf{x}, \mathbf{z}) = \psi(\mathbf{x})^T \psi(\mathbf{z})$ (6章)

$$\psi(\mathbf{x}) = \beta^{1/2} \mathbf{S}_N^{1/2} \phi(\mathbf{x})$$

3.4 ベイズモデル比較

1章で過学習や交差確認(cv)による正則化パラメータ値の決定やモデルの選択法を論じた。

本節では**ベイズの立場からモデル選択**を考える。

→訓練データだけに基づいてモデル比較が行える

3.4 ベイズモデル比較(モデルが複数ある時どれが一番もっともらしいか)

最尤推定に関連した過学習は、モデルパラメータの値を点推定する代わりに周辺化（または和、積分）することにより回避できる。

・ ベイズモデル比較：モデル選択の不確かさを確率で表現（確率の加法乗法）

e.g) **L個のモデル $\{\mathcal{M}_i\}(i = 1 \sim L)$ の比較**

→ データ \mathcal{D} はL個のモデルのどれかに従って生成される

→ 事前確率分布 $p(\mathcal{M}_i)$ を用いて事後分布 $p(\mathcal{M}_i|\mathcal{D}) \propto p(\mathcal{M}_i)p(\mathcal{D}|\mathcal{M}_i)$ を評価する

※モデル \mathcal{M} は観測データ \mathcal{D} 上の確率分布

分布は目標値 t の集合上に定義、入力値 x は既知

今回は簡単のため、全ての $p(\mathcal{M}_i)$ は等確率

・ **周辺尤度（モデルエビデンス）： $p(\mathcal{D}|\mathcal{M}_i)$**

→ データ \mathcal{D} から見た時のモデル \mathcal{M}_i の好み（モデルでデータがどれくらい説明できるか）

・ **ベイズ因子**(2モデルにおけるエビデンス比)： $\frac{p(\mathcal{D}|\mathcal{M}_i)}{p(\mathcal{D}|\mathcal{M}_j)}$

・ 全体の予測分布(**混合分布**)： $p(t|x, \mathcal{D}) = \sum_{i=1}^{\{L\}} p(t|x, \mathcal{M}_i, \mathcal{D})p(\mathcal{M}_i|\mathcal{D})$

個々のモデル予測分布 $p(t|x, \mathcal{M}_i, \mathcal{D})$ と事後確率 $p(\mathcal{M}_i|\mathcal{D})$ の重みつき平均

3.4 ベイズモデル比較

モデル選択：一番もっともらしいモデルを選ぶ

- ・パラメータ w を持つモデルのモデルエビデンス： $p(\mathcal{D}|\mathcal{M}_i) = \int p(\mathcal{D}|w, \mathcal{M}_i)p(w|\mathcal{M}_i)dw$

※ベイズ定理によれば、パラメータ w の事後確率を計算するときにもエビデンスが登場する

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

パラメータを事前分布から適当にサンプリングしたときにデータ集合 \mathcal{D} が生成される確率

・近似によるモデルエビデンスの解釈

→近似条件：単一パラメータ w

→事後確率 $p(w|\mathcal{D}) \propto p(\mathcal{D}|w)p(w)$

：事後分布がモード w_{MAP} で幅 $\Delta w_{\text{posterior}}$ で尖っている（全体の積分が幅×最大値と近似）

：事前分布の幅 Δw_{prior}

→ $p(w) = \frac{1}{\Delta w_{\text{prior}}}$

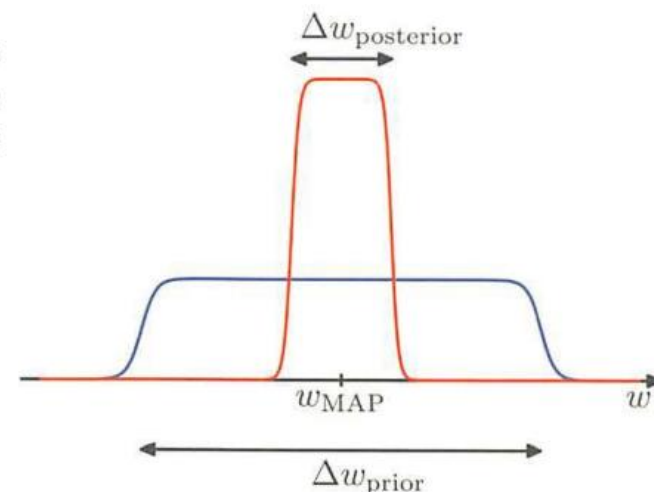
$$\rightarrow p(\mathcal{D}) = \int p(\mathcal{D}|w)p(w)dw \simeq p(\mathcal{D}|w_{\text{MAP}}) \frac{\Delta w_{\text{posterior}}}{\Delta w_{\text{prior}}}$$

対数

→

$$\ln p(\mathcal{D}) \simeq \ln p(\mathcal{D}|w_{\text{MAP}}) + \ln \left(\frac{\Delta w_{\text{posterior}}}{\Delta w_{\text{prior}}} \right)$$

図 3.12 パラメータの事後分布がモード w_{MAP} の近傍で鋭く尖っているとき、モデルエビデンスの大雑把な近似が得られる。



3.4 ベイズモデル比較

$$\text{近似したモデルエビデンス: } \ln p(\mathcal{D}) \simeq \ln p(\mathcal{D}|w_{\text{MAP}}) + \overset{\text{負}}{\ln} \left(\frac{\Delta w_{\text{posterior}}}{\Delta w_{\text{prior}}} \right)$$

事前分布が平坦な対数尤度
データへのフィッティング度 モデルの複雑さに対するペナルティ項

- ・ 第一項：一番もっともらしいパラメータ値によるデータへのフィッティング度
- ・ 第二項：モデルの複雑さにもとづいてペナルティを与えることに対応
 - $\Delta w_{\text{posterior}} < \Delta w_{\text{prior}}$ (事後の方が狭い) より **第二項は負 (ペナルティ)**
 $\Delta w_{\text{posterior}}$ が狭まるほど第二項はより小さく (負方向に大きく) なりペナルティが増す
∴ モデルがデータに強くフィットするとペナルティが強くなる (過学習の回避)

モデルパラメータがM個あり、全て同じパラメータ比 $\frac{\Delta w_{\text{posterior}}}{\Delta w_{\text{prior}}}$ を持つとき

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

- ・ 第一項はモデルの複雑さが増すことで大きくなる (よりデータにフィット)
- ・ 第二項 (ペナルティ) もパラメータ数Mに応じて強くなる

∴ エビデンスが最大となる最適なモデル複雑さ → 2 項のトレードオフ

3.4 ベイズモデル比較

モデル $\{\mathcal{M}_i\}$ からデータ集合を生成する際

- ・事前分布 $p(\mathbf{w})$ に従ってパラメータを選択
- ・その \mathbf{w} に対してデータを $p(\mathcal{D}|\mathbf{w})$ からサンプリング

1次多項式とか

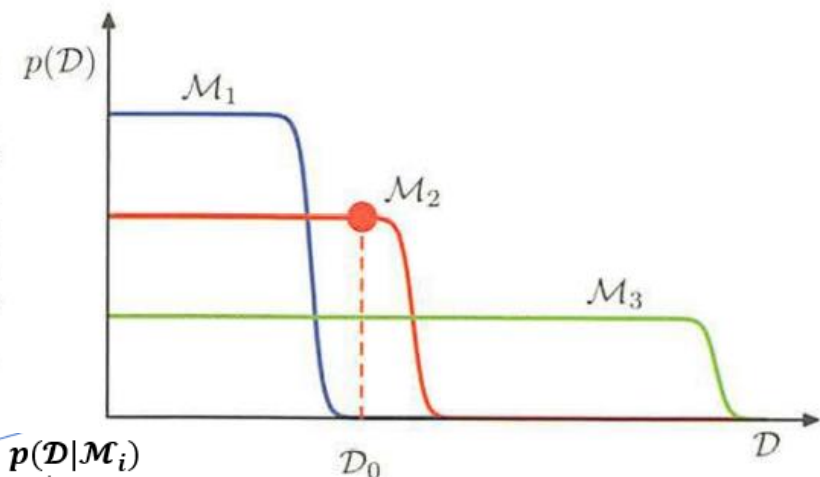
→単純モデル \mathcal{M}_1 は自由度が少ないため生成データ \mathcal{D} の多様性が少なく、 $p(\mathcal{D})$ は図軸の狭い範囲に集中

9次多項式とか

→複雑モデル \mathcal{M}_3 は多様性に富む： $p(\mathcal{D})$ は広がる

図 3.13 複雑さの異なる3つのモデルに対するデータの分布、 \mathcal{M}_1 が最も単純なモデルであり、 \mathcal{M}_3 が最も複雑なモデルである。分布は正規化されている。この例では、観測されたデータ \mathcal{D}_0 に対するエビデンスは、中間の複雑さを持つモデル \mathcal{M}_2 が最大である。

・周辺尤度 (モデルエビデンス) : $p(\mathcal{D}|\mathcal{M}_i)$
→データ \mathcal{D} から見た時のモデル \mathcal{M}_i の好み



あるデータ \mathcal{D}_0 に対してエビデンス $p(\mathcal{D}|\mathcal{M}_i)$ が最大となるモデルは中間の複雑さになることがある

(単純モデルはデータにフィットしにくいし、複雑モデルは予測分布が広くて特定のデータ \mathcal{D}_0 が割り当てられる確率が相対的に低い)

※前提仮定：ベイズモデル比較では想定するモデルの中にデータが生成される真の分布がある

e.g.)あるモデル \mathcal{M}_i と正しいモデル $\mathcal{M}_{correct}$ の比較

→ベイズ因子 $\frac{p(\mathcal{D}|\mathcal{M}_i)}{p(\mathcal{D}|\mathcal{M}_j)}$ のデータ集合の分布について平均すると

期待ベイズ因子： $\int p(\mathcal{D}|\mathcal{M}_{correct}) \frac{p(\mathcal{D}|\mathcal{M}_{correct})}{p(\mathcal{D}|\mathcal{M}_i)} d\mathcal{D}$ (カルバック-ライブラーダイバージェンス:相対エントロピー)

→相対エントロピーは2つの分布が等しいときに0(他は正) → ∴ 平均的には期待値は正しいモデルのベイズ因子の方が常に大きい

3.4 ベイズモデル比較(まとめ)

ベイズの枠組みの利点

- ・ 過学習が回避
- ・ 訓練データのみでのモデル比較

難点

- ・ 考えるモデル形の仮定が正しくないと間違える (モデルエビデンスは事前分布の特性に強く依存)
- ・ 変則事前分布に対してモデルエビデンスが定義できない ($\log p(\mathcal{D})$)
e.g.) ガウス事前分布における分散が無限大の分布とか

$$p(\mathcal{D}) = \int p(\mathcal{D}|w)p(w) dw \simeq p(\mathcal{D}|w_{\text{MAP}}) \frac{\Delta w_{\text{posterior}}}{\Delta w_{\text{prior} \rightarrow \infty}} \rightarrow 0$$

※ただし 2 モデルのエビデンスの比を先に計算して後に極限をとると解消される場合がある

∴実際に応用する場合はテスト用の独立データを用意して最終性能評価をするのが妥当

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いたい

→パラメータ w に加えて、ハイパーパラメータ α, β にも事前分布を導入→周辺化&予測?
→全部積分・周辺化するのは解析的にキツイ

- ∴①パラメータ w だけを積分して周辺尤度関数を得る
②周辺尤度関数を最大化するようハイパーパラメータの値を決める

2段階の近似

※統計学の文脈では「**経験ベイズ**」「**第二種の最尤推定**」「**一般化最尤推定**」

機械学習の文脈では「**エビデンス近似**」と呼ばれる

3.5 エビデンス近似(α, β, w に関して周辺化した予測分布)

とりあえず w, α, β の事前分布を導入 → 同時分布を w, α, β で周辺化して予測分布を考える

$$\text{予測分布: } p(t|\mathbf{t}) = \iiint p(t|w, \beta) p(w|\mathbf{t}, \alpha, \beta) p(\alpha, \beta|\mathbf{t}) dw d\alpha d\beta$$

$$\begin{aligned} \times \quad p(t|w, \beta) &= \mathcal{N}(t|y(x, w), \beta^{-1}) & p(w|\mathbf{t}, \alpha, \beta) &= \mathcal{N}(w|m_N, S_N) \\ & & m_N &= \beta S_N \Phi^T \mathbf{t} \\ & & S_N^{-1} &= \alpha I + \beta \Phi^T \Phi \end{aligned}$$

事後分布 $p(\alpha, \beta|\mathbf{t})$ がある値 $\hat{\alpha}$ と $\hat{\beta}$ で尖っている → α, β をその値で固定化する → w のみ周辺化で予測分布

$$\text{予測分布: } p(t|\mathbf{t}) \simeq p(t|\mathbf{t}, \hat{\alpha}, \hat{\beta}) = \int p(t|w, \hat{\beta}) p(w|\mathbf{t}, \hat{\alpha}, \hat{\beta}) dw$$

$$\alpha, \beta \text{ の事後分布: } p(\alpha, \beta|\mathbf{t}) \propto p(\mathbf{t}|\alpha, \beta) * p(\alpha, \beta)$$

周辺尤度関数

※(事前分布が平坦であれば)エビデンスの枠組みの中で、 $\hat{\alpha}$ と $\hat{\beta}$ の値は周辺尤度関数 $p(\mathbf{t}|\alpha, \beta)$ の最大化で得られる

今回考えること

①線形基底関数モデルに対して周辺尤度関数を求める

②それを最大化してハイパーパラメータ α, β の値を決定する ※比 α/β は正則化パラメータと同様の働きをする
最大化する方法

①エビデンス関数を解析的に評価(導関数を0にして α, β の再推定方程式を得る)

②EMアルゴリズム(9.3.4章、①も②も同じ解に収束)

3.5.1 エビデンス関数の評価 (線形基底関数モデルの周辺尤度関数の導出)

周辺尤度関数 : $p(\mathbf{t}|\alpha, \beta) = \int p(\mathbf{t}|\mathbf{w}, \beta)p(\mathbf{w}|\alpha)d\mathbf{w}$

$$\begin{aligned}\therefore p(\mathbf{t}|\alpha, \beta) &= \int \prod \mathcal{N}(t_n|\mathbf{w}^T\Phi(\mathbf{x}_n), \beta^{-1})\mathcal{N}(\mathbf{w}|0, \alpha^{-1}\mathbf{I})d\mathbf{w} \quad \text{Mはパラメータwの次元} \\ &= \int (\beta/2\pi)^{N/2} \exp(-\beta E_D(\mathbf{w})) (\alpha/2\pi)^{M/2} \exp(-\alpha/2\mathbf{w}^T\mathbf{w}) d\mathbf{w} \\ &= (\beta/2\pi)^{N/2} (\alpha/2\pi)^{M/2} \int \exp(-E(\mathbf{w})) d\mathbf{w}\end{aligned}$$

$$\ast E(\mathbf{w}) = \beta E_D + \alpha/2 \mathbf{w}^T\mathbf{w} = \beta/2 \|\mathbf{t} - \Phi\mathbf{w}\|^2 + \alpha/2 \mathbf{w}^T\mathbf{w}$$

→これは正則化最小二乗法で議論した誤差関数 $\frac{1}{2} \sum_{n=1}^N \{t_n - \mathbf{w}^T\phi(\mathbf{x}_n)\}^2 + \frac{\lambda}{2} \mathbf{w}^T\mathbf{w}$ の定数倍

$$\begin{aligned}\ln p(\mathbf{t}|\mathbf{w}, \beta) &= \sum_{n=1}^N \ln \mathcal{N}(t_n|\mathbf{w}^T\phi(\mathbf{x}_n), \beta^{-1}) \\ &= \frac{N}{2} \ln \beta - \frac{N}{2} \ln(2\pi) - \beta E_D(\mathbf{w}) \\ E_D(\mathbf{w}) &= \frac{1}{2} \sum_{n=1}^N \{t_n - \mathbf{w}^T\phi(\mathbf{x}_n)\}^2 \\ p(\mathbf{w}|\alpha) &= \mathcal{N}(\mathbf{w}|0, \alpha^{-1}\mathbf{I})\end{aligned}$$

$E(\mathbf{w})$ を \mathbf{w} に関して平方完成すると $E(\mathbf{w}) = E(\mathbf{m}_N) + \frac{1}{2}(\mathbf{w} - \mathbf{m}_N)^T \mathbf{A}(\mathbf{w} - \mathbf{m}_N)$

$$\ast E(\mathbf{m}_N) = \frac{\beta}{2} \|\mathbf{t} - \Phi\mathbf{m}_N\|^2 + \frac{\alpha}{2} \mathbf{m}_N^T \mathbf{m}_N, \quad \mathbf{A} = \alpha\mathbf{I} + \beta\Phi^T\Phi, \quad \mathbf{m}_N = \beta\mathbf{A}^{-1}\Phi^T\mathbf{t} = \beta\mathbf{S}_N\Phi^T\mathbf{t}$$

加えて \mathbf{A} は誤差関数の2回の導関数 $\mathbf{A} = \nabla\nabla E(\mathbf{w})$ となる (ヘッセ行列)

共分散 $\mathbf{S}_N^{-1} = \alpha\mathbf{I} + \beta\Phi^T\Phi$

事後分布の平均

更に周辺尤度関数を解析

$$\begin{aligned}\int \exp(-E(\mathbf{w})) d\mathbf{w} &= \exp(-E(\mathbf{m}_N)) \int \exp\{-\frac{1}{2}(\mathbf{w} - \mathbf{m}_N)^T \mathbf{A}(\mathbf{w} - \mathbf{m}_N)\} d\mathbf{w} \\ &= \exp(-E(\mathbf{m}_N)) (2\pi)^{M/2} |\mathbf{A}|^{-1/2}\end{aligned}$$

$$\therefore \text{対数周辺尤度関数 : } \ln p(\mathbf{t}|\alpha, \beta) = \frac{M}{2} \ln \alpha + \frac{N}{2} (\ln \beta - \ln 2\pi) - E(\mathbf{m}_N) - \frac{1}{2} \ln |\mathbf{A}| \quad \text{エビデンス関数!}$$

3.5.1 エビデンス関数の評価

1章でやった三角関数の多項式回帰を考えてみる

事前分布 $p(\mathbf{w} \mid \alpha) = \mathcal{N}(\mathbf{w} \mid \mathbf{0}, \alpha^{-1} \mathbf{I}) = \left(\frac{\alpha}{2\pi}\right)^{(M+1)/2} \exp\left\{-\frac{\alpha}{2} \mathbf{w}^T \mathbf{w}\right\}$

M: フィッティングする多項式の次元

結果

M=0: フィッティングが悪く、エビデンスも小さい

M=1: M=0よりフィッティングが向上し、エビデンスも大きく

M=2: M=1よりエビデンスが悪い、

残差の低下よりも、次数向上によるペナルティが大きい

※今回は真の関数が奇関数のため偶数次数の効力が低い

M=3: 最もエビデンスが高い

残差が最も小さくフィッティング能力が高い

M ≥ 4: Mに関してエビデンスが単調減少

→ 残差はM=3に比べ殆ど変わらないため

次数(複雑さ)向上によるペナルティが大きくなってる

∴ フィッティング能力と複雑さペナルティのトレードオフではM=3が最もよい

$$\text{対数周辺尤度関数: } \ln p(t \mid \alpha, \beta) = \frac{M}{2} \ln \alpha + \frac{N}{2} (\ln \beta - \ln 2\pi) - E(m_N) - \frac{1}{2} \ln |A| \quad \text{エビデンス関数}$$

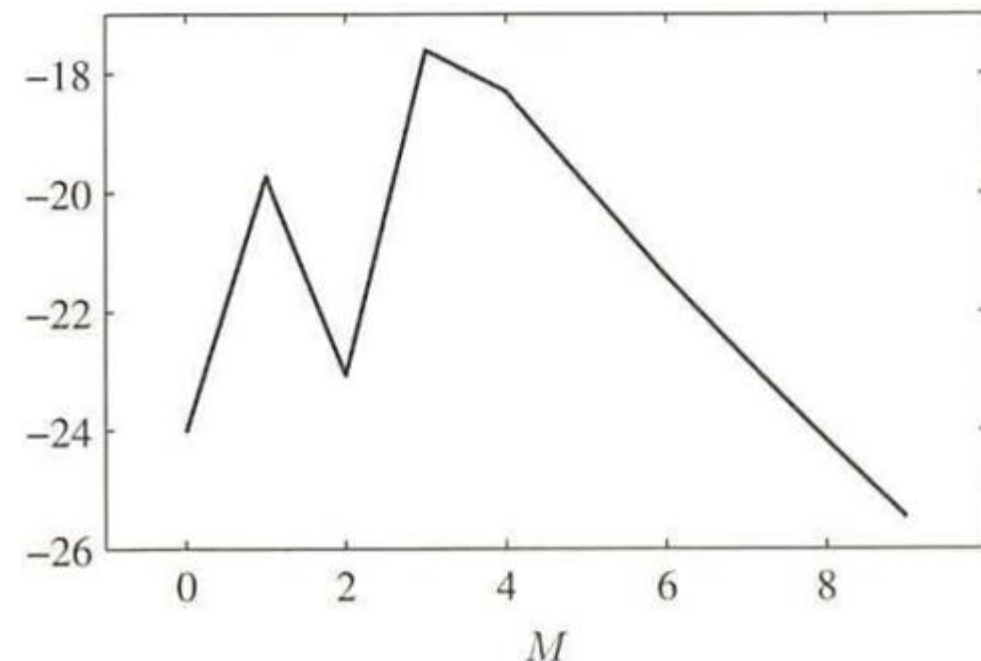
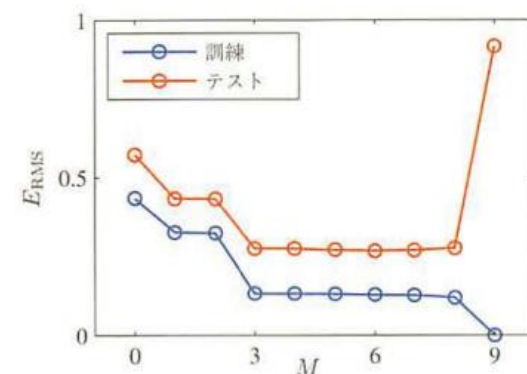


図 3.14 多項式回帰モデルにおける多項式の次元 M とモデルエビデンス (対数表示) との関係. エビデンスを最大にするモデルの次数は $M=3$ であることを示している.



3.5.2 エビデンス関数の最大化

動機：周辺尤度関数： $p(t|\alpha, \beta)$ の最大化を行いたい

$$\text{対数周辺尤度関数：} \ln p(t|\alpha, \beta) = \frac{M}{2} \ln \alpha + \frac{N}{2} (\ln \beta - \ln 2\pi) - E(m_N) - \frac{1}{2} \ln |A| \quad \text{エビデンス関数}$$

①エビデンス関数を解析的に評価(導関数を0にして α, β の再推定方程式を得る)

導入：固有ベクトル方程式： $(\beta \Phi^T \Phi) \mathbf{u}_i = \lambda_i \mathbf{u}_i$

※ $A = \alpha I + \beta \Phi^T \Phi$ より A は固有値 $\alpha + \lambda_i$ を持つ

∴ $\ln |A|$ の α に関する導関数を考えると

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

∴エビデンス関数の α に関する停留点は

$$0 = \frac{M}{2\alpha} - \frac{1}{2} \mathbf{m}_N^T \mathbf{m}_N - \frac{1}{2} \sum_i \frac{1}{\lambda_i + \alpha} \quad \text{を整理して、} \quad \alpha \mathbf{m}_N^T \mathbf{m}_N = M - \alpha \sum_i \frac{1}{\lambda_i + \alpha} = \sum_i \frac{\lambda_i + \alpha - \alpha}{\lambda_i + \alpha} = \sum_i \frac{\lambda_i}{\lambda_i + \alpha} = \gamma$$

$$E_W(m_N) = \frac{1}{2} \mathbf{m}_N^T \mathbf{m}_N$$

$$\therefore \alpha = \frac{\gamma}{\mathbf{m}_N^T \mathbf{m}_N} \quad \text{※ } \alpha \text{ の再推定方程式}$$

γ も \mathbf{m}_N も α に依存するため上式は α に陰に依存している→繰り返し手順により再推定

①適当な初期値 α を与える → ② $\mathbf{m}_N = \beta S_N \Phi^T t$, $\gamma = \sum_i \frac{\lambda_i}{\lambda_i + \alpha}$ を用いて α を再計算 →繰り返し……

→最終的に α の値は収束する。(この計算において $\Phi^T \Phi$ の値は変わらないため、最初に計算しておくだけでよい)

→以上の手順では訓練データのみで α の値が決定する (最尤推定のようなモデル複雑さの最適化用データを用意しなくてよい)

3.5.2 エビデンス関数の最大化

$$\text{対数周辺尤度関数: } \ln p(t|\alpha, \beta) = \frac{M}{2} \ln \alpha + \frac{N}{2} (\ln \beta - \ln 2\pi) - E(m_N) - \frac{1}{2} \ln |A| \quad \text{エビデンス関数}$$

同様に β に対しても対数周辺尤度関数（エビデンス関数）を最大化できる

①エビデンス関数を解析的に評価(導関数を0にして α, β の再推定方程式を得る)

導入：固有ベクトル方程式： $(\beta \Phi^T \Phi) \mathbf{u}_i = \lambda_i \mathbf{u}_i$

固有値 λ_i は β に比例するため $d\lambda_i/d\beta = \lambda_i/\beta$ となる。

$$\therefore \frac{d}{d\beta} \ln |A| = \frac{d}{d\beta} \sum_i \ln(\lambda_i + \alpha) = \frac{1}{\beta} \sum_i \frac{\lambda_i}{\lambda_i + \alpha} = \frac{\gamma}{\beta}$$

\therefore エビデンス関数の β に関する停留点は

$$0 = \frac{N}{2\beta} - \frac{\gamma}{2\beta} - \frac{1}{2} \sum_i \{t_n - \mathbf{m}_N^T \Phi(x_n)\}^2 \quad \text{※} E(m_N) = \frac{\beta}{2} \|t - \Phi m_N\|^2 + \frac{\alpha}{2} m_N^T m_N$$

これを解いて

$$\frac{1}{\beta} = \frac{1}{N - \gamma} \sum_i \{t_n - \mathbf{m}_N^T \Phi(x_n)\}^2$$

※ β の再推定方程式

右辺はやはり β に関する陰関数なので、適当な初期値 β 、 $\mathbf{m}_N = \beta S_N \Phi^T t$ 、 $\gamma = \sum_i \frac{\lambda_i}{\lambda_i + \alpha}$ を用いた再推定

→ β が収束して、 α 、 β とともに訓練データから決定できる。

3.5.3 有効パラメータ数

① $\alpha = \frac{\gamma}{m_N^T m_N}$

② $(\beta \Phi^T \Phi) \mathbf{u}_i = \lambda_i \mathbf{u}_i \quad (3.87)$

尤度関数の等高線は固有ベクトル \mathbf{u}_i の軸に沿った楕円になる。

右図において固有値 λ_i が楕円曲率を表す (曲率が小さいと等高線は伸びる)

②の左辺の $(\beta \Phi^T \Phi)$ は正定値行列なので固有値 λ_i は正
 $\therefore 0 \leq \sum_i \frac{\lambda_i}{\lambda_i + \alpha} = \gamma \leq M$ (Mはパラメータ \mathbf{w} の次元)

$\lambda_i \gg \alpha$ の時、対応するパラメータ w_i は最尤推定値に近く、 $\frac{\lambda_i}{\lambda_i + \alpha} \simeq 1$ となる。
その時のパラメータはデータによって強い制約を持つ→**well-determinedパラメータ**

$\lambda_i \ll \alpha$ の時、対応するパラメータ w_i は0に近く、 $\frac{\lambda_i}{\lambda_i + \alpha} \simeq 0$
→パラメータ変化に対する尤度関数の感度が悪い→データによらず、 \mathbf{w} は事前分布に従って小さく設定される

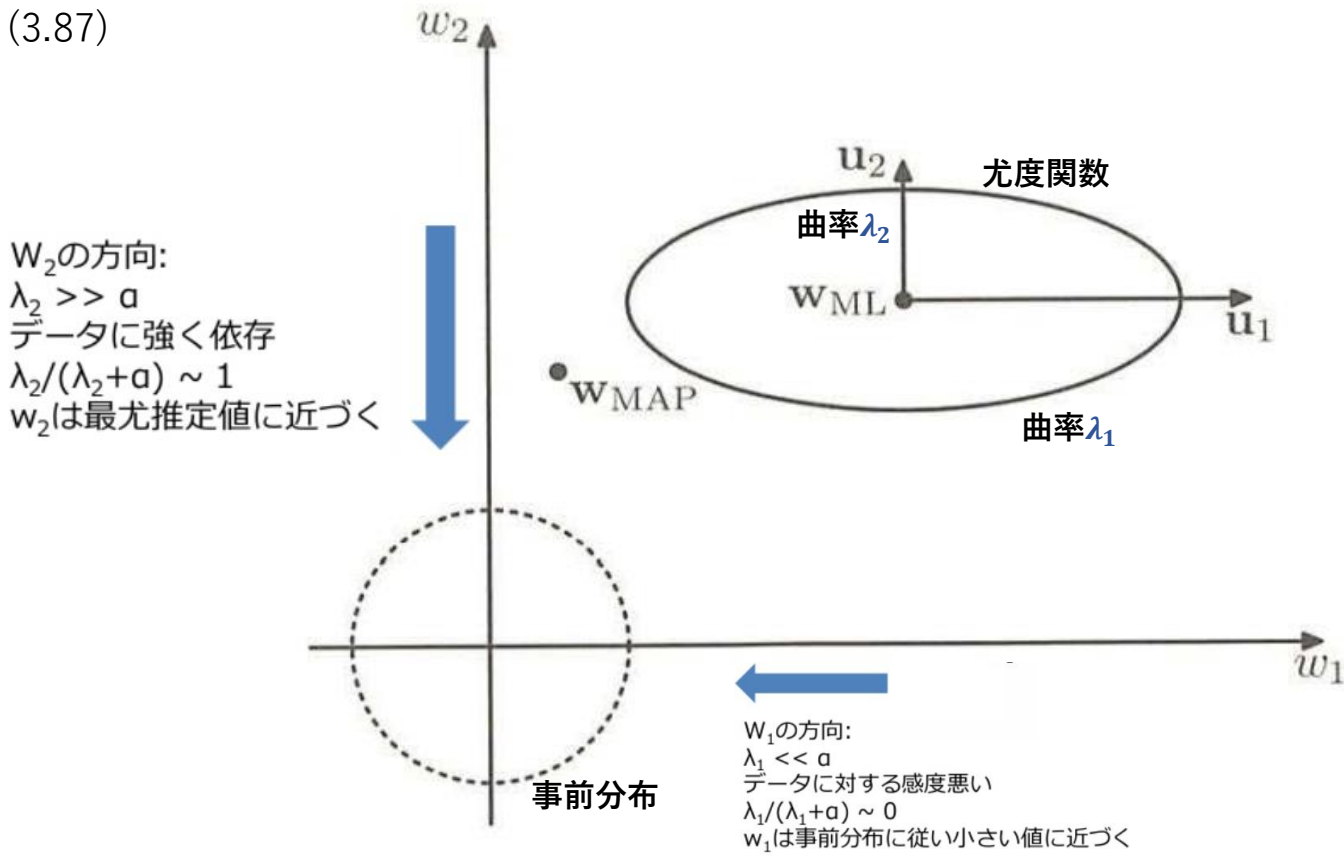


図 3.15 尤度関数（実線）と事前確率（破線）の等高線表示. パラメータ空間の各軸は、ヘッセ行列の固有ベクトル \mathbf{u}_i と重なるように回転してある. $\alpha = 0$ のとき、事後分布のモードは最尤解 \mathbf{w}_{ML} で与えられ、一方 $\alpha \neq 0$ のとき、モードは $\mathbf{w}_{MAP} = \mathbf{m}_N$ で与えられる. w_1 の方向に対しては、(3.87) で定義される固有値 λ_1 は α より小さいので、 $\lambda_1 / (\lambda_1 + \alpha)$ は 0 に近い. したがって、対応する w_1 のモードも 0 に近い値になる. 一方、 w_2 の方向に対しては、固有値 λ_2 は α より大きいので、 $\lambda_2 / (\lambda_2 + \alpha)$ は 1 に近い. したがって、 w_2 のモードは最尤推定値に近くなる.

3.5.3 有効パラメータ数

$$\alpha = \frac{\gamma}{m_N^T m_N}$$

$\therefore \gamma = \sum_i \frac{\lambda_i}{\lambda_i + \alpha}$ は well-determined パラメータの有効数を表している

$$(\beta \Phi^T \Phi) \mathbf{u}_i = \lambda_i \mathbf{u}_i$$

→ $\frac{1}{\beta} = \frac{1}{N-\gamma} \sum_i \{t_n - m_N^T \Phi(\mathbf{x}_n)\}^2$ と $\frac{1}{\beta_{ML}} = \frac{1}{N} \sum_i \{t_n - w_{ML}^T \Phi(\mathbf{x}_n)\}^2$ の比較
ベイズ推定で求めた β 3.1章の最尤推定量

→ 両者とも目標値とモデルによる推定値との差の二乗平均（分散：精度の逆数）

→ 分母の値で両者に違いがみられる

・ ここで1変数 x によるガウス分布分散は $\sigma_{ML}^2 = \frac{1}{N} \sum_{n=1}^N (x_n - \mu_{ML})^2$

→ 1章で見たように最尤アプローチは分布の分散が $(N-1)/N$ 倍過小評価されている (**バイアス**)

→ $\sigma_{MAP}^2 = \frac{1}{N-1} \sum_{n=1}^N (x_n - \mu_{ML})^2$ ベイズの結果の分母の $N-1$ は自由度の一つを平均のフィッティングと最尤推定のバイアスを取り除くことを考慮している。
分散の不偏推定量

・ 以上を線形回帰モデルで考えた場合

・ 目標分布の平均： M 個のパラメータを含む $\mathbf{w}^T \Phi(\mathbf{x})$ によって与えられる

→ しかし、全パラメータがデータによって調整されるわけではない（影響がないものもある）

\therefore 有効パラメータ数は $\gamma = \sum_i \frac{\lambda_i}{\lambda_i + \alpha}$ で与えられ、 $M - \gamma$ 個のパラメータは事前分布によって小さい値となる

\therefore ベイズ推定の分母の $N - \gamma$ は最尤推定の結果のバイアスを補正している。

3.5.3 有効パラメータ数

$$\text{対数周辺尤度関数: } \ln p(\mathbf{t}|\alpha, \beta) = \frac{M}{2} \ln \alpha + \frac{N}{2} (\ln \beta - \ln 2\pi) - E(\mathbf{m}_N) - \frac{1}{2} \ln |A| \quad \text{エビデンス関数}$$

e.g.)
1章の $\sin 2\pi x$ を9個のガウス基底関数モデルによって近似する

バイアスを含めて10個のパラメータ、 $\beta = 11.1$

図3.16

→エビデンス近似によるハイパーパラメータ α の決定

図3.17

→ハイパーパラメータ α はパラメータ $\{w_i\}$ の大きさを制御

$N \gg M$ (データ点がパラメータ数より十分大きい) とき
 $(\beta \Phi^T \Phi) \mathbf{u}_i = \lambda_i \mathbf{u}_i$ より全パラメータが well-determined

→ $\Phi^T \Phi$ はデータ点に関する和を陰に含んでいて
データサイズの増大に伴い固有値 λ_i も大きくなる
→ $\gamma \simeq M$ この時の α, β の再推定方程式

$$\alpha = \frac{\gamma}{\mathbf{m}_N^T \mathbf{m}_N} = \frac{M}{2E_W(\mathbf{m}_N)}, \quad \beta = \frac{N}{2E_D(\mathbf{m}_N)}$$

$$E_W(\mathbf{w}) = \frac{1}{2} \mathbf{w}^T \mathbf{w} \quad E_D(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^N \{t_n - \mathbf{w}^T \phi(\mathbf{x}_n)\}^2$$

上式の近似の利点はヘッセ行列の
固有値スペクトルの計算を介さないため比較的楽に計算できる

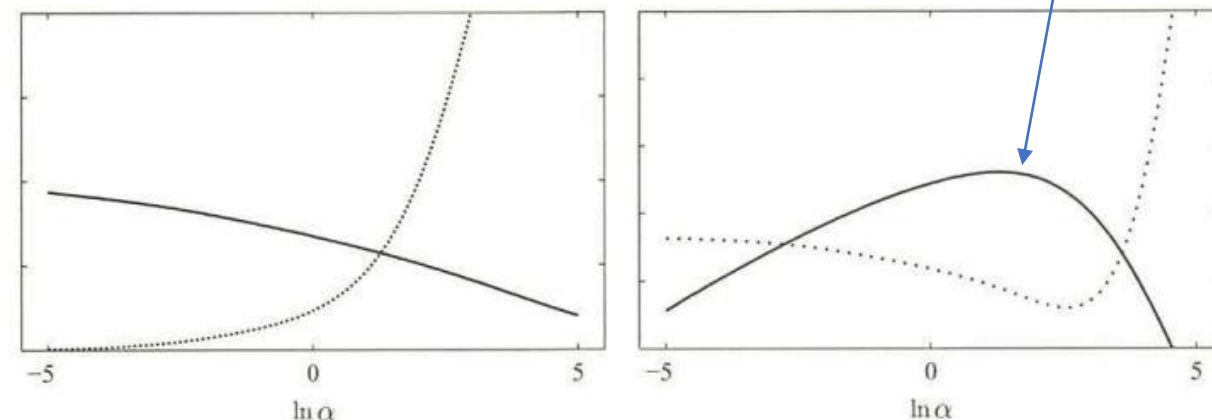
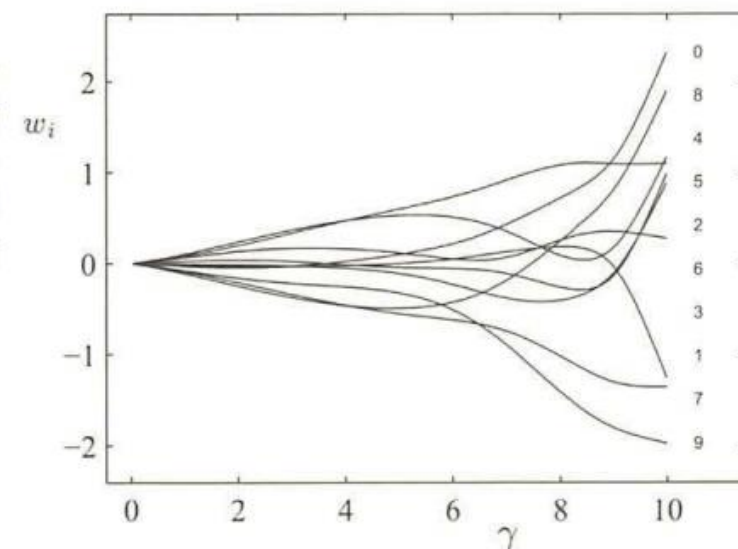


図 3.16 左のプロットは、三角関数の人工データ集合に対する γ (実線) と $2\alpha E_W(\mathbf{m}_N)$ (点線) を $\ln \alpha$ の関数として示している。エビデンスの手順によって得られる最適な α の値はこれらの2つの曲線の交点で定められる。右のプロットは、対応する対数エビデンス $\ln p(\mathbf{t}|\alpha, \beta)$ (実線) を $\ln \alpha$ の関数として示している。これより、対数エビデンスの最大値は左のプロットの交点と対応していることがわかる。また、これはテスト集合に対する誤差 (点線) を最小にする点に近いこともわかる。

図 3.17 ガウス基底関数モデルの10個のパラメータ w_i の値と有効パラメータ数 γ の関係。超パラメータ α を $0 \leq \alpha \leq \infty$ の範囲で変化させることにより、 γ は $0 \leq \gamma \leq M$ の範囲で変化する。



3.6 固定された基底関数の限界（3章のまとめ）

3章ではあらかじめ固定化した非線形基底関数の線形結合したモデルを扱ってきた

$$y(\mathbf{x}, \mathbf{w}) = w_0 + \sum_{j=1}^{M-1} w_j \phi_j(\mathbf{x})$$

3章の前提：モデルはパラメータ \mathbf{w} に対して線形である（仮定）

→利点

- ・ 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ぶことにより、入力値 \mathbf{x} から目標値 t への非線形変換をモデル化できる
(等価カーネル： $y(\mathbf{x}, \mathbf{m}_N) = \sum_{n=1}^N k(\mathbf{x}, \mathbf{x}_n) t_n$ $k(\mathbf{x}, \mathbf{x}') = \boldsymbol{\beta} \boldsymbol{\Phi}(\mathbf{x})^T \mathbf{S}_N \boldsymbol{\Phi}(\mathbf{x}')$)

- ・ 線形モデルには複数の致命的欠陥がある（後の章で議論）

e.g.)

訓練データ集合を観測する前に基底関数 $\boldsymbol{\Phi}_j(\mathbf{x})$ を固定する

→**次元の呪い**（入力空間の次元数 D に応じて、指数的に基底関数の数を増やさなければならない）

※実際のところ、現実のデータベクトルの本質的な次元は入力空間の次元より小さいことが多い

→入力変数同士が相関を持っていたりする

→今後は局所的な基底関数を用いた対処法（RBF、SVM、RVFなど）が紹介される。

同様に、尤度関数 (3.11) をノイズの精度パラメータ β に関しても最大化することができ、

$$\frac{1}{\beta_{\text{ML}}} = \frac{1}{N} \sum_{n=1}^N \{t_n - \mathbf{w}_{\text{ML}}^T \phi(\mathbf{x}_n)\}^2 \quad (3.21)$$

が得られる。これより、ノイズの精度の逆数は回帰関数周りでの目標値の残差分散で与えられることがわかる。