

回归的线性模型

李娜

监督学习

给定一个由 N 个观测值 $\{x_n\}$ 组成的数据集，其中 $n = 1, \dots, N$ ，以及对应的目标值 $\{t_n\}$ ，我们的目标是预测对于给定新的 x 值的情况下， t 的值。最简单的方法是，直接建立一个适当的函数 $y(x)$ ，对于新的输入 x ，这个函数能够直接给出对应的 t 的预测。更一般地，从一个概率的观点来看，我们的目标是对预测分布 $p(t | x)$ 建模，因为它表达了对于每个 x 值，我们对于 t 的值的

回归模型

简单回归模型

$$y(\boldsymbol{x}, \boldsymbol{w}) = w_0 + w_1 x_1 + \dots + w_D x_D$$

一般回归模型

$$y(\boldsymbol{x}, \boldsymbol{w}) = w_0 + \sum_{j=1}^{M-1} w_j \phi_j(\boldsymbol{x}) \implies y(\boldsymbol{x}, \boldsymbol{w}) = \sum_{j=0}^{M-1} w_j \phi_j(\boldsymbol{x}) = \boldsymbol{w}^T \boldsymbol{\phi}(\boldsymbol{x})$$

w_0 被称为偏置参数

其中 $\phi_j(\boldsymbol{x})$ 被称为基函数 (basis function)

常见基函数

多项式基函数 $\phi_j(x) = x^j$

高斯基函数 $\phi_j(x) = \exp \left\{ -\frac{(x - \mu_j)^2}{2s^2} \right\}$ [点击](#)

sigmoid基函数 $\phi_j(x) = \sigma \left(\frac{x - \mu_j}{s} \right)$

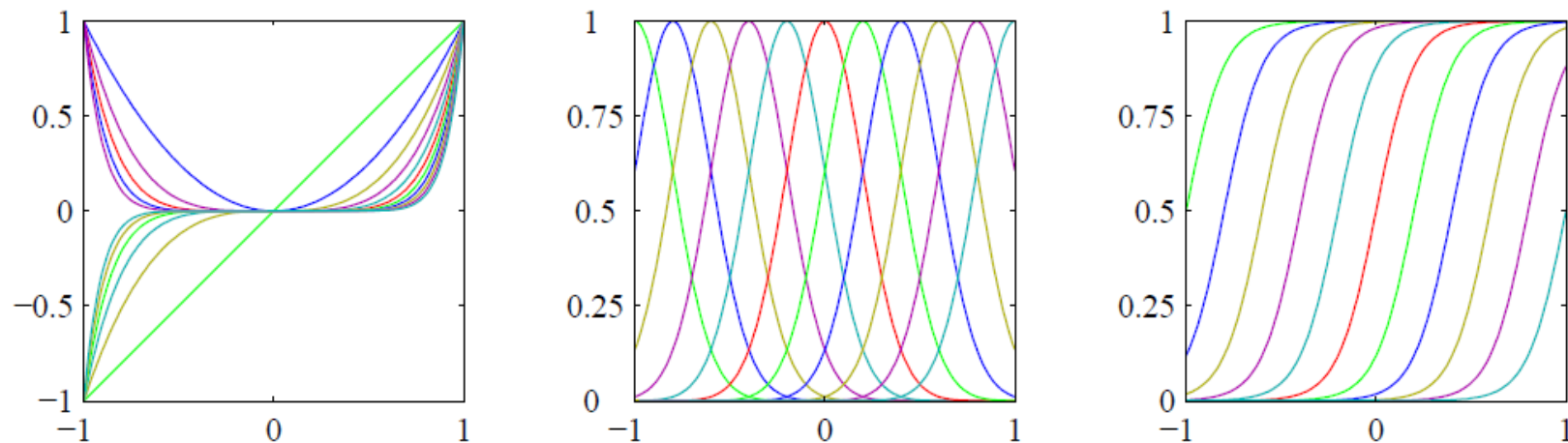


图 3.1: 基函数的例子, 左图是多项式基函数, 中图是形式为 (3.4) 的高斯基函数, 右图是形式为 (3.5) 的sigmoid基函数。

最大似然与最小平方

$t = y(\mathbf{x}, \mathbf{w}) + \epsilon$ 其中 ϵ 是一个零均值的高斯随机变量，精度（方差的倒数）为 β 。

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

$$\mathbb{E}[t \mid \mathbf{x}] = \int t p(t \mid \mathbf{x}) \, dt = y(\mathbf{x}, \mathbf{w})$$

考虑一个输入数据集 $\mathbf{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_N\}$ ，对应的目标值为 t_1, \dots, t_N 。我们把目标向量 $\{t_n\}$ 成一个列向量，记作 \mathbf{t} ：

$$p(\mathbf{t} \mid \mathbf{X}, \mathbf{w}, \beta) = \prod_{n=1}^N \mathcal{N}(t_n \mid \mathbf{w}^T \phi(\mathbf{x}_n), \beta^{-1}) \longrightarrow \begin{aligned} &\mathcal{N}(t_n \mid \mathbf{w}^T \phi(\mathbf{x}_n), \beta^{-1}) \\ &= \frac{\sqrt{\beta}}{\sqrt{2\pi}} \exp\left\{-\frac{\beta}{2}(t_n - \mathbf{w}^T \phi(\mathbf{x}_n))^2\right\} \end{aligned}$$

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

$$= \frac{N}{2} \ln \beta - \frac{N}{2} \ln(2\pi) - \beta E_D(\mathbf{w}) \longrightarrow E_D(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^N \{t_n - \mathbf{w}^T \phi(\mathbf{x}_n)\}^2$$

最大似然与最小平方

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

对 \mathbf{w} 求导

$$\nabla \ln p(\mathbf{t} \mid \mathbf{w}, \beta) = \beta \sum_{n=1}^N \{t_n - \mathbf{w}^T \phi(\mathbf{x}_n)\} \phi(\mathbf{x}_n)^T$$

$$0 = \sum_{n=1}^N t_n \phi(\mathbf{x}_n)^T - \mathbf{w}^T \left(\sum_{n=1}^N \phi(\mathbf{x}_n) \phi(\mathbf{x}_n)^T \right)$$

$$\mathbf{w}_{ML} = (\Phi^T \Phi)^{-1} \Phi^T \mathbf{t}$$

$\Phi^T \Phi$ 可逆吗？

$$\Phi = \begin{pmatrix} \phi_0(\mathbf{x}_1) & \phi_1(\mathbf{x}_1) & \cdots & \phi_{M-1}(\mathbf{x}_1) \\ \phi_0(\mathbf{x}_2) & \phi_1(\mathbf{x}_2) & \cdots & \phi_{M-1}(\mathbf{x}_2) \\ \vdots & \vdots & \ddots & \vdots \\ \phi_0(\mathbf{x}_N) & \phi_1(\mathbf{x}_N) & \cdots & \phi_{M-1}(\mathbf{x}_N) \end{pmatrix}$$

最大似然与最小平方

若显示的展现偏置参数 w_0

$$E_D(w) = \frac{1}{2} \sum_{n=1}^N \{t_n - w_0 - \sum_{j=1}^{M-1} w_j \phi_j(x_n)\}^2$$

对 w_0 求导

$$w_0 = \bar{t} - \sum_{j=1}^{M-1} w_j \bar{\phi}_j$$

偏置 w_0 补偿了目标值的平均值（在训练集上的）与基函数的值的平均值的加权求和之间的差。

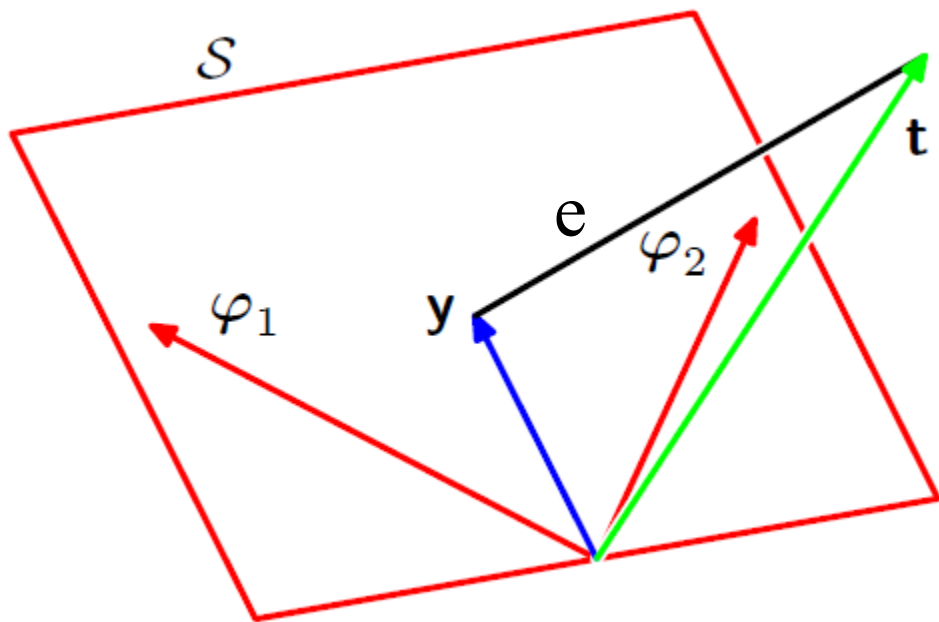
$$\bar{t} = \frac{1}{N} \sum_{n=1}^N t_n, \quad \bar{\phi}_j = \frac{1}{N} \sum_{n=1}^N \phi_j(x_n)$$

对 β 求导

$$\frac{1}{\beta_{ML}} = \frac{1}{N} \sum_{n=1}^N \{t_n - \mathbf{w}_{ML}^T \boldsymbol{\phi}(x_n)\}^2$$

噪声精度的倒数由目标值在回归函数周围的残留方差给出。

最小平方的几何描述



$$\varphi_1 x_1 + \varphi_2 x_2 = y = A\check{x}$$

$$e = t - y = t - A\check{x}$$

$$\varphi_1^T e = 0, \varphi_2^T e = 0 \Rightarrow A^T e = 0$$

$$\Rightarrow A^T (t - A\check{x}) = 0$$

$$\Rightarrow A^T (t - A\check{x}) = 0$$

$$\Rightarrow \check{x} = (A^T A)^{-1} A^T t$$

$\Phi^T \Phi$ 不可逆的情况

若其不可逆，则可以采取以下方法：

- (1) 移除冗余特征。
- (2) 特征太多时，删除一些特征，或者对于小样本，使用正则化。

若 A 可逆，则对任意 $x \neq 0$ ， $Ax \neq 0$ ，则

$$x^T (A^T A) x = (Ax)^T (Ax) > 0$$

故：

$A^T A$ 可逆

顺序学习

$$\left. \begin{aligned} \boldsymbol{w}^{(\tau+1)} &= \boldsymbol{w}^{(\tau)} - \eta \nabla E_n \\ E_D(\boldsymbol{w}) &= \frac{1}{2} \sum_{n=1}^N \{t_n - \boldsymbol{w}^T \phi(\boldsymbol{x}_n)\}^2 \end{aligned} \right\} \boldsymbol{w}^{(\tau+1)} = \boldsymbol{w}^{(\tau)} + \eta (t_n - \boldsymbol{w}^{(\tau)T} \phi_n) \phi_n$$

τ 表示迭代次数， η 是学习率参数

两种求参方式对比：梯度下降和正规方程

梯度下降 gradient descent	正规方程 normal equation
<p>缺点：</p> <ul style="list-style-type: none">● 需要选择学习率α● 需要多次迭代● 特征值范围相差太大，要特征缩放 <p>优点：</p> <ul style="list-style-type: none">● 当特征数n很大时，能够工作的很好	<p>优点：</p> <ul style="list-style-type: none">● 不需要选择学习率α● 不需要多次迭代● 不需要特征缩放(feature scaling) <p>缺点：</p> <ul style="list-style-type: none">● 当特征数n很大时，运算的很慢。因为求解逆矩阵的时间复杂度是$O(N^3)$
<p>何时选择梯度下降、正规方程：</p> <p>这个没有一个特定的标准，完全靠经验。给出NG的经验，$n < 10000$时，用正规方程。当$n \geq 10000$时，要考虑用梯度下降。</p> <p>一些更加复杂的算法只能选择用梯度下降</p>	

正则化最小平方

$$E_D(\mathbf{w}) + \lambda E_W(\mathbf{w})$$

$$E_W(\mathbf{w}) = \frac{1}{2} \mathbf{w}^T \mathbf{w}$$

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

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

对 \mathbf{w} 求导：

$$\mathbf{w} = (\lambda \mathbf{I} + \Phi^T \Phi)^{-1} \Phi^T \mathbf{t}$$

一般形式：

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

正则化最小平方

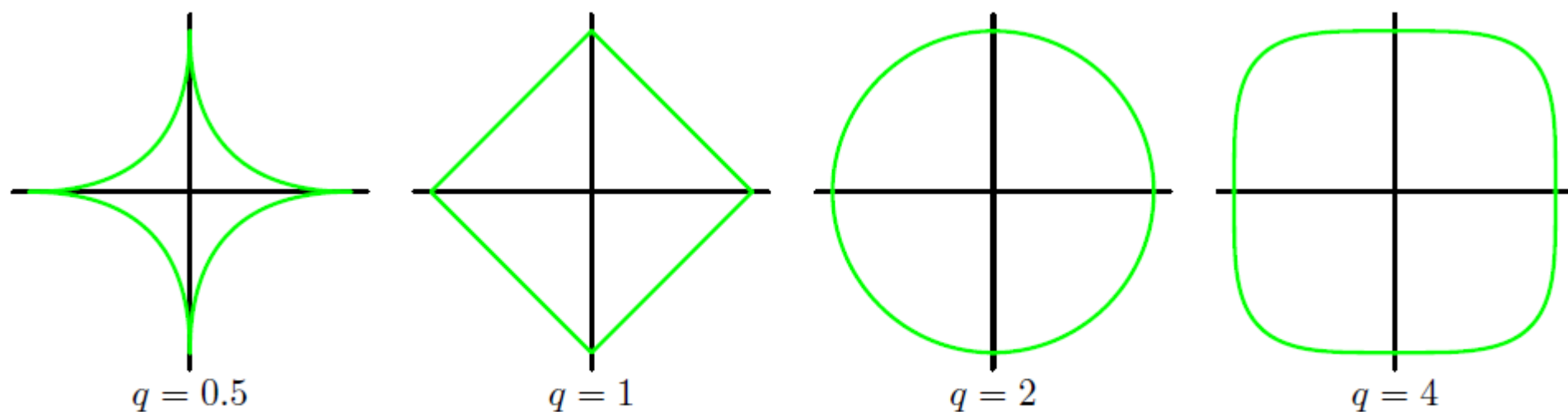


图 3.3: 对于不同的参数 q , 公式 (3.29) 中的正则化项的轮廓线。

在统计学的文献中, $q = 1$ 的情形被称为套索 (lasso) (Tibshirani, 1996)。它的性质为: 如果 λ 充分大, 那么某些系数 w_j 会变为零, 从而产生了一个稀疏 (sparse) 模型, 这个模型中对应的基函数不起作用。为了说明这一点, 我们首先注意到最小化公式 (3.19) 等价于在满足下面的限制的条件下最小化未正则化的平方和误差函数 (3.12)

$$\sum_{j=1}^M |w_j|^q \leq \eta \quad (3.30)$$

正则化最小平方

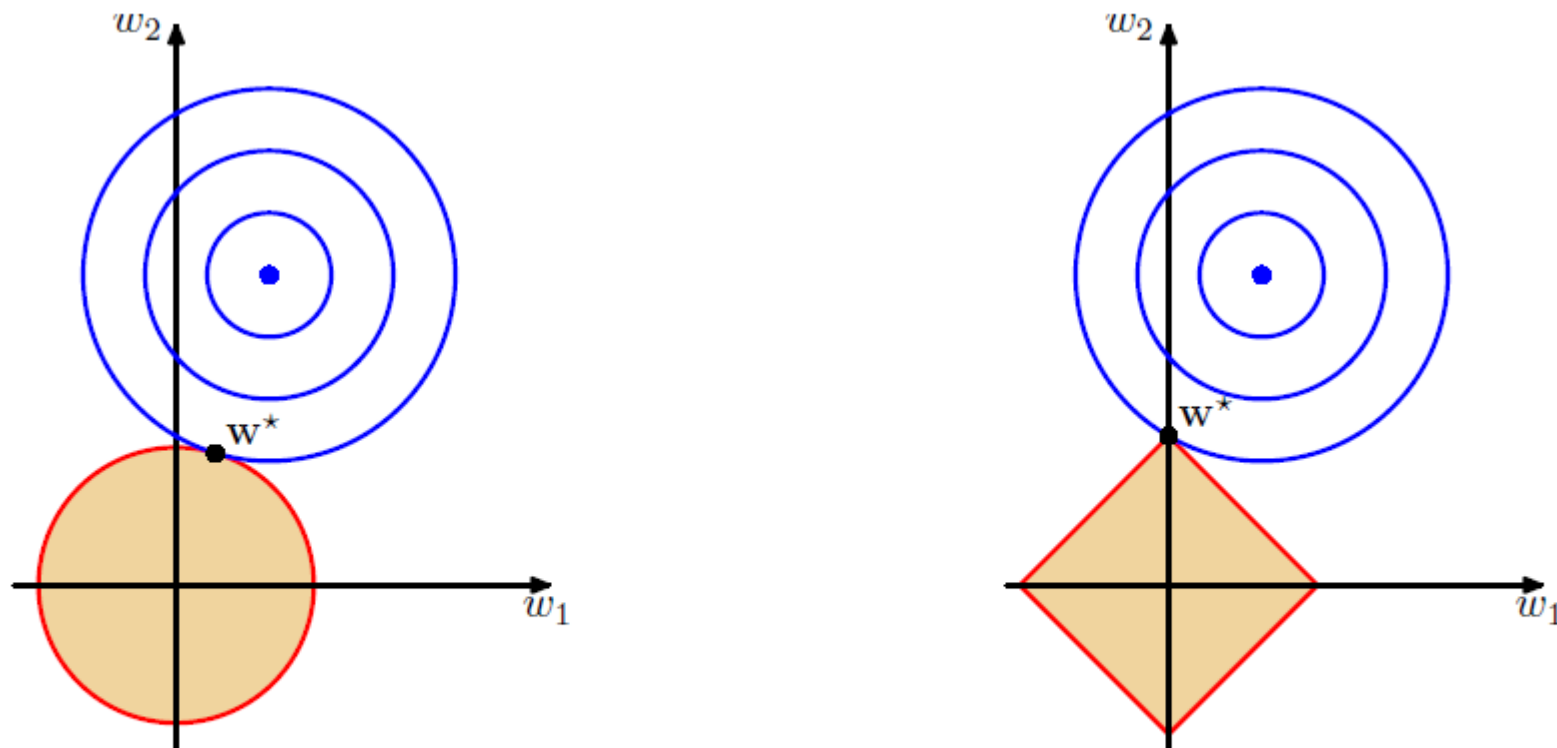


图 3.4: 未正则化的误差函数的轮廓线（蓝色）以及公式 (3.30) 给出的限制区域。左图是 $q = 2$ 的二次正则化项的限制区域，右图是 $q = 1$ 的套索正则化项的限制区域，其中参数向量 w 的值被记作 w^* 。套索正则化项给出了一个稀疏的解，其中 $w_1^* = 0$ 。

L1正则假设参数的先验分布是Laplace分布，可以保证模型的稀疏性，也就是某些参数等于0；
L2正则假设参数的先验分布是Gaussian分布，可以保证模型的稳定性，也就是参数的值不会太大或太小；
在实际使用中，如果特征是高维稀疏的，则使用L1正则；如果特征是低维稠密的，则使用L2正则。

正则化最小平方

正则化方法通过限制模型的复杂度，使得复杂的模型能够在有限大小的数据集上进行训练，而不会产生严重的过拟合。然而，这样做就使确定最优的模型复杂度的问题从确定合适的基函数数量的问题转移到了确定正则化系数 λ 的合适值的问题上。我们稍后在本章中还会回到这个模

多个输出

$$\underset{\text{K} \times 1}{y(x, w)} = \underset{\text{K} \times \text{M}}{\mathbf{W}^T} \underset{\text{M} \times 1}{\phi(x)}$$

$$p(\mathbf{t} \mid \mathbf{x}, \mathbf{W}, \beta) = \mathcal{N}(\mathbf{t} \mid \mathbf{W}^T \phi(\mathbf{x}), \beta^{-1} \mathbf{I})$$

如果我们有一组观测 t_1, \dots, t_N ，我们可以把这些观测组合为一个 $N \times K$ 的矩阵 \mathbf{T} ，使得矩阵的第 n 行为 \mathbf{t}_n^T 。类似地，我们可以把输入向量 $\mathbf{x}_1, \dots, \mathbf{x}_N$ 组合为矩阵 \mathbf{X} 。这样，对数似然函数为

$$\begin{aligned} \ln p(\underset{\text{N} \times \text{K}}{\mathbf{T}} \mid \underset{\text{N} \times \text{M}}{\mathbf{X}}, \underset{\text{M} \times \text{K}}{\mathbf{W}}, \beta) &= \sum_{n=1}^N \ln \mathcal{N}(\underset{\text{K} \times \text{M}}{\mathbf{t}_n} \mid \underset{\text{K} \times \text{M}}{\mathbf{W}^T} \underset{\text{M} \times 1}{\phi(\mathbf{x}_n)}, \beta^{-1} \mathbf{I}) \\ &= \frac{NK}{2} \ln \left(\frac{\beta}{2\pi} \right) - \frac{\beta}{2} \sum_{n=1}^N \|\mathbf{t}_n - \mathbf{W}^T \phi(\mathbf{x}_n)\|^2 \end{aligned}$$

$$\mathbf{W}_{ML} = (\Phi^T \Phi)^{-1} \Phi^T \mathbf{T}$$

似然函数与正则化

- 1、如果使用有限规模的数据集来训练复杂的模型，那么使用最大似然方法，或者等价地，使用最小平方法，会导致严重的**过拟合**问题。
- 2、通过限制基函数的数量来避免过拟合问题有一个负作用，即限制了模型描述数据中有趣且重要的规律的**灵活性**。
- 3、虽然引入正则化项可以控制具有多个参数的模型的过拟合问题，但是如何确定正则化系数的合适的值。同时关于权值 w 和正则化系数来最小化正则化的误差函数显然不是一个正确的方法，因为这样做会使得 $\lambda = 0$ ，从而产生非正则化的解。

回归问题的损失函数

期望损失 $\mathbb{E}[L] = \iint L(t, y(\mathbf{x})) p(\mathbf{x}, t) \, d\mathbf{x} \, dt$

$$\mathbb{E}[L] = \iint \{y(\mathbf{x}) - t\}^2 p(\mathbf{x}, t) \, d\mathbf{x} \, dt$$

选择 $y(\mathbf{x})$
优化 $E[L]$ $\frac{\delta \mathbb{E}[L]}{\delta y(\mathbf{x})} = 2 \int \{y(\mathbf{x}) - t\} p(\mathbf{x}, t) \, dt = 0$

回归函数 $y(\mathbf{x}) = \frac{\int t p(\mathbf{x}, t) \, dt}{p(\mathbf{x})} = \int t p(t \mid \mathbf{x}) \, dt = \mathbb{E}_t[t \mid \mathbf{x}]$

$$\begin{aligned} \{y(\mathbf{x}) - t\}^2 &= \{y(\mathbf{x}) - \mathbb{E}[t \mid \mathbf{x}] + \mathbb{E}[t \mid \mathbf{x}] - t\}^2 \\ &= \{y(\mathbf{x}) - \mathbb{E}[t \mid \mathbf{x}]\}^2 + 2\{y(\mathbf{x}) - \mathbb{E}[t \mid \mathbf{x}]\}\{\mathbb{E}[t \mid \mathbf{x}] - t\} \\ &\quad + \{\mathbb{E}[t \mid \mathbf{x}] - t\}^2 \end{aligned}$$

$$\mathbb{E}[L] = \int \{y(\mathbf{x}) - \mathbb{E}[t \mid \mathbf{x}]\}^2 p(\mathbf{x}) \, d\mathbf{x} + \iint \{\mathbb{E}[t \mid \mathbf{x}] - t\}^2 p(\mathbf{x}, t) \, d\mathbf{x} \, dt$$

偏置-方差分解

最优预测

$$h(\mathbf{x}) = \mathbb{E}[t \mid \mathbf{x}] = \int t p(t \mid \mathbf{x}) \, dt$$

期望损失函数

$$\mathbb{E}[L] = \iint \{y(\mathbf{x}) - t\}^2 p(\mathbf{x}, t) \, d\mathbf{x} \, dt \quad (1.5节)$$

平方损失函数期望

$$\mathbb{E}[L] = \int \underbrace{\{y(\mathbf{x}) - h(\mathbf{x})\}^2}_{\text{偏置}^2} p(\mathbf{x}) \, d\mathbf{x} + \iint \underbrace{\{h(\mathbf{x}) - t\}^2}_{\text{方差}} p(\mathbf{x}, t) \, d\mathbf{x} \, dt$$

$$\begin{aligned} & \{y(\mathbf{x}; \mathcal{D}) - \mathbb{E}_{\mathcal{D}}[y(\mathbf{x}; \mathcal{D})] + \mathbb{E}_{\mathcal{D}}[y(\mathbf{x}; \mathcal{D})] - h(\mathbf{x})\}^2 \\ &= \{y(\mathbf{x}; \mathcal{D}) - \mathbb{E}_{\mathcal{D}}[y(\mathbf{x}; \mathcal{D})]\}^2 + \{\mathbb{E}_{\mathcal{D}}[y(\mathbf{x}; \mathcal{D})] - h(\mathbf{x})\}^2 \\ & \quad + 2\{y(\mathbf{x}; \mathcal{D}) - \mathbb{E}_{\mathcal{D}}[y(\mathbf{x}; \mathcal{D})]\}\{\mathbb{E}_{\mathcal{D}}[y(\mathbf{x}; \mathcal{D})] - h(\mathbf{x})\} \end{aligned}$$

由数据本身的噪声造成的，
表示期望损失能够达到的
最小值。

$$\begin{aligned} & \mathbb{E}_{\mathcal{D}}[\{y(\mathbf{x}; \mathcal{D}) - h(\mathbf{x})\}^2] \\ &= \underbrace{\{\mathbb{E}_{\mathcal{D}}[y(\mathbf{x}; \mathcal{D})] - h(\mathbf{x})\}^2}_{(\text{偏置})^2} + \underbrace{\mathbb{E}_{\mathcal{D}}[\{y(\mathbf{x}; \mathcal{D}) - \mathbb{E}_{\mathcal{D}}[y(\mathbf{x}; \mathcal{D})]\}^2]}_{\text{方差}} \end{aligned}$$

表示所有数据集的平均预测与
预测的回归函数之间的差异

模型给出的解在平均值附近的波动
情况，即对特定数据集的敏感程度

偏置-方差分解

$$\text{期望损失} = \text{偏置}^2 + \text{方差} + \text{噪声}$$

$$\text{偏置}^2 = \int \{\mathbb{E}_{\mathcal{D}}[y(x; \mathcal{D})] - h(x)\}^2 p(x) \, dx$$

$$\text{方差} = \int \mathbb{E}_{\mathcal{D}}[\{y(x; \mathcal{D}) - \mathbb{E}_{\mathcal{D}}[y(x; \mathcal{D})]\}^2] p(x) \, dx$$

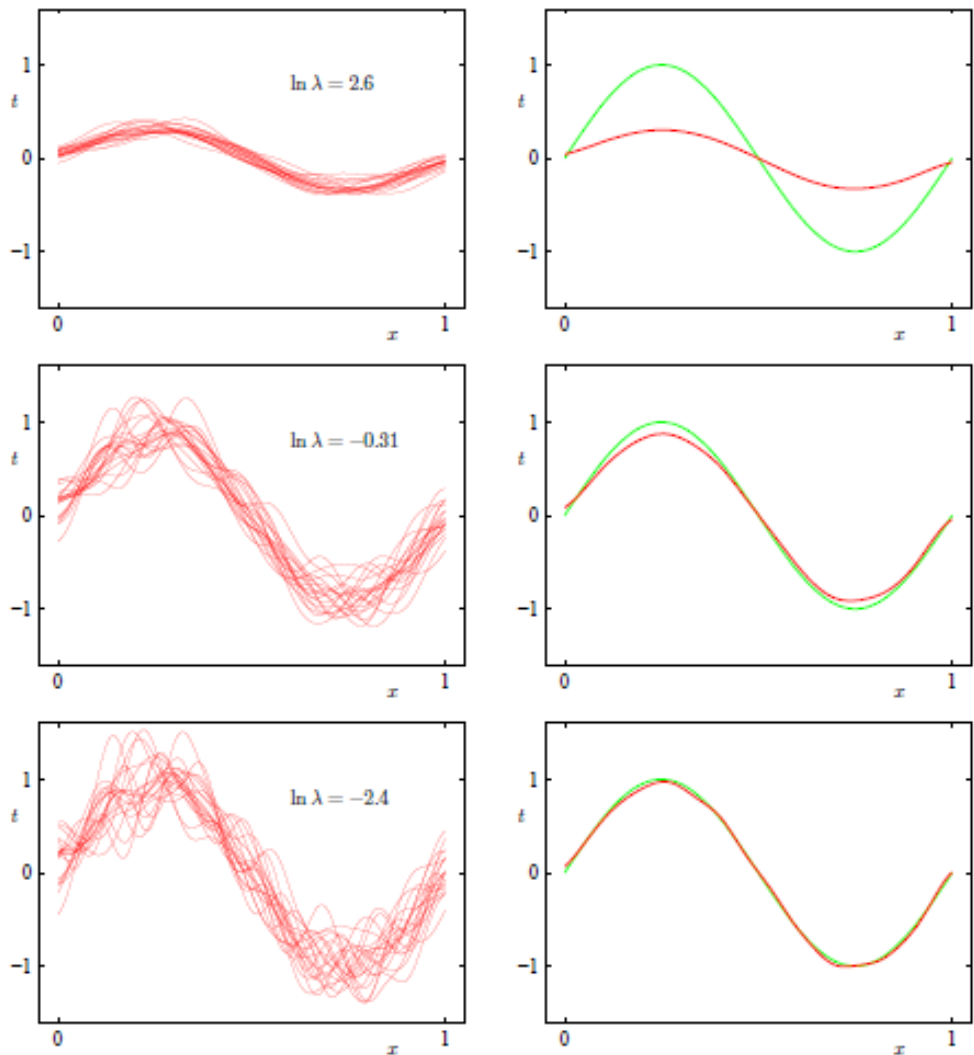
$$\text{噪声} = \iint \{h(x) - t\}^2 p(x, t) \, dx \, dt$$

偏置-方差分解

例子：采样于正弦曲线的100个数据集，每个数据集25个数据点，规则化参数 λ 采用不同的取值。

$$h(x) = \sin(2\pi x)$$

对于非常灵活的模型来说，偏置较小，方差较大。
对于相对固定的模型来说，偏置较大，方差较小。



偏差大
方差小

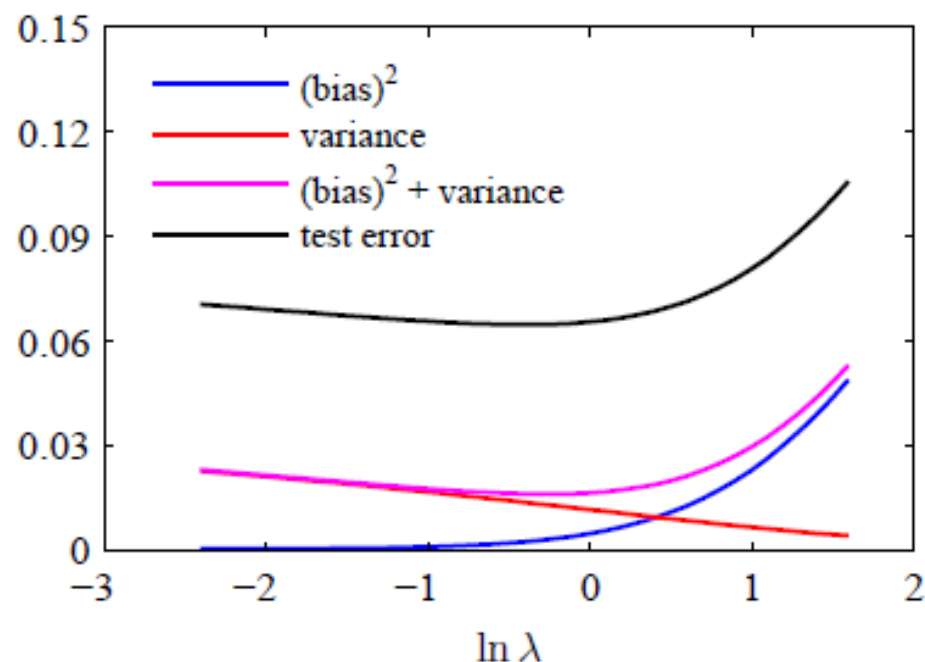
偏差小
方差大

偏置-方差分解

平均预测 $\bar{y}(x) = \frac{1}{L} \sum_{l=1}^L y^{(l)}(x)$

$$\text{偏置}^2 = \frac{1}{N} \sum_{n=1}^N \{\bar{y}(x_n) - h(x_n)\}^2$$

$$\text{方差} = \frac{1}{N} \sum_{n=1}^N \frac{1}{L} \sum_{l=1}^L \{y^{(l)}(x_n) - \bar{y}(x_n)\}^2$$



虽然偏置-方差分解能够从频率学家的角度对模型的复杂度提供一些有趣的认识，但是它的实用价值很有限。这是因为偏置-方差分解依赖于对所有的数据集求平均，而在实际应用中我们只有一个观测数据集。如果我们有大量的已知规模的独立的训练数据集，那么我们最好的方法是把它们组合成一个大的训练集，这显然会降低给定复杂度的模型的过拟合程度。

贝叶斯线性回归

这就产生了对于特定的应用确定合适的模型复杂度的问题。这个问题不能简单地通过最大化似然函数来确定，因为这总会产生过于复杂的模型和过拟合现象。独立的额外数据能够用来确定模型的复杂度，正如1.3节所说的那样，但是这需要较大的计算量，并且浪费了有价值的数据。因此我们转而考虑线性回归的贝叶斯方法，这会避免最大似然的过拟合问题，也会引出使用训练数据本身确定模型复杂度的自动化方法。与之前一样，为了简单起见，我们只考虑单一目标变量 t 的情形。对于多个目标变量情形的推广是很直接的，与3.1.5节的讨论很类似。

样本集合 D 中的样本都是从一个固定但是未知的概率密度函数 $p(x)$ 中独立抽取出来的，要求根据这些样本估计 x 的概率分布，记为 $p(x|D)$ ，并且使得 $p(x|D)$ 尽可能的接近 $p(x)$ ，这就是贝叶斯估计的核心问题。

条件高斯分布

多元高斯分布的一个重要性质是，如果两组变量是联合高斯分布，那么以一组变量为条件，另一组变量同样是高斯分布。类似地，任何一个变量的边缘分布也是高斯分布

$$\boldsymbol{x} = \begin{pmatrix} \boldsymbol{x}_a \\ \boldsymbol{x}_b \end{pmatrix} \quad \boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_a \\ \boldsymbol{\mu}_b \end{pmatrix} \quad \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{aa} & \boldsymbol{\Sigma}_{ab} \\ \boldsymbol{\Sigma}_{ba} & \boldsymbol{\Sigma}_{bb} \end{pmatrix} \quad \boldsymbol{\Lambda} = \begin{pmatrix} \boldsymbol{\Lambda}_{aa} & \boldsymbol{\Lambda}_{ab} \\ \boldsymbol{\Lambda}_{ba} & \boldsymbol{\Lambda}_{bb} \end{pmatrix}$$

多元高斯分布

$$\mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{\frac{D}{2}}} \frac{1}{|\boldsymbol{\Sigma}|^{\frac{1}{2}}} \exp \left\{ -\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}) \right\}$$

多元高斯分布
指数

$$\begin{aligned} -\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}) = & \\ & -\frac{1}{2}(\boldsymbol{x}_a - \boldsymbol{\mu}_a)^T \boldsymbol{\Lambda}_{aa}(\boldsymbol{x}_a - \boldsymbol{\mu}_a) - \frac{1}{2}(\boldsymbol{x}_a - \boldsymbol{\mu}_a)^T \boldsymbol{\Lambda}_{ab}(\boldsymbol{x}_b - \boldsymbol{\mu}_b) \\ & - \frac{1}{2}(\boldsymbol{x}_b - \boldsymbol{\mu}_b)^T \boldsymbol{\Lambda}_{ba}(\boldsymbol{x}_a - \boldsymbol{\mu}_a) - \frac{1}{2}(\boldsymbol{x}_b - \boldsymbol{\mu}_b)^T \boldsymbol{\Lambda}_{bb}(\boldsymbol{x}_b - \boldsymbol{\mu}_b) \end{aligned}$$

一般高斯分布
指数

$$-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}) = -\frac{1}{2}\boldsymbol{x}^T \boldsymbol{\Sigma}^{-1}\boldsymbol{x} + \boldsymbol{x}^T \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} + \text{常数}$$

条件高斯分布

x_a 的二阶项 $-\frac{1}{2}x_a^T \Lambda_{aa} x_a \longrightarrow \Sigma_{a|b} = \Lambda_{aa}^{-1}$

x_a 的常数项 $x_a^T \{ \Lambda_{aa} \mu_a - \Lambda_{ab}(x_b - \mu_b) \} \longrightarrow \mu_{a|b} = \Sigma_{a|b} \{ \Lambda_{aa} \mu_a - \Lambda_{ab}(x_b - \mu_b) \}$
 $= \mu_a - \Lambda_{aa}^{-1} \Lambda_{ab}(x_b - \mu_b)$

边缘高斯分布

$$p(x_a) = \mathcal{N}(x_a \mid \mu_a, \Sigma_{aa})$$

一般规则

给定 \boldsymbol{x} 的一个边缘高斯分布，以及在给定 \boldsymbol{x} 的条件下 \boldsymbol{y} 的条件高斯分布，形式为

$$p(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1}) \quad (2.113)$$

$$p(\boldsymbol{y} \mid \boldsymbol{x}) = \mathcal{N}(\boldsymbol{y} \mid \boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}, \boldsymbol{L}^{-1}) \quad (2.114)$$

\boldsymbol{y} 的边缘分布以及给定 \boldsymbol{y} 的条件下 \boldsymbol{x} 的条件分布为

$$p(\boldsymbol{y}) = \mathcal{N}(\boldsymbol{y} \mid \boldsymbol{A}\boldsymbol{\mu} + \boldsymbol{b}, \boldsymbol{L}^{-1} + \boldsymbol{A}\boldsymbol{\Lambda}^{-1}\boldsymbol{A}^T) \quad (2.115)$$

$$p(\boldsymbol{x} \mid \boldsymbol{y}) = \mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\Sigma}\{\boldsymbol{A}^T\boldsymbol{L}(\boldsymbol{y} - \boldsymbol{b}) + \boldsymbol{\Lambda}\boldsymbol{\mu}\}, \boldsymbol{\Sigma}) \quad (2.116)$$

其中

$$\boldsymbol{\Sigma} = (\boldsymbol{\Lambda} + \boldsymbol{A}^T\boldsymbol{L}\boldsymbol{A})^{-1} \quad (2.117)$$

贝叶斯线性回归-参数分布

$$p(\boldsymbol{w}) = \mathcal{N}(\boldsymbol{w} \mid \boldsymbol{m}_0, \boldsymbol{S}_0)$$

$$p(t|\boldsymbol{w}) = \mathcal{N}(\boldsymbol{w}^T \Phi(\boldsymbol{x}), \beta^{-1})$$

$$p(\boldsymbol{w} \mid \mathbf{t}) = \mathcal{N}(\boldsymbol{w} \mid \boldsymbol{m}_N, \boldsymbol{S}_N)$$

$$\begin{aligned} \boldsymbol{m}_0 &\rightarrow \boldsymbol{\mu} \\ \boldsymbol{S}_0 &\rightarrow \boldsymbol{\Lambda}^{-1} \\ \Phi(\boldsymbol{x}) &\rightarrow \boldsymbol{A} \\ \beta &\rightarrow L \end{aligned}$$

$$p(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1})$$

$$p(\boldsymbol{y} \mid \boldsymbol{x}) = \mathcal{N}(\boldsymbol{y} \mid \boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}, \boldsymbol{L}^{-1})$$

$$p(\boldsymbol{x} \mid \boldsymbol{y}) = \mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\Sigma}\{\boldsymbol{A}^T \boldsymbol{L}(\boldsymbol{y} - \boldsymbol{b}) + \boldsymbol{\Lambda}\boldsymbol{\mu}\}, \boldsymbol{\Sigma})$$

$$\boldsymbol{\Sigma} = (\boldsymbol{\Lambda} + \boldsymbol{A}^T \boldsymbol{L} \boldsymbol{A})^{-1}$$

$$\Rightarrow \begin{aligned} \boldsymbol{S}_N &= (\boldsymbol{S}_0^{-1} + \Phi^T \beta \Phi)^{-1} \\ \boldsymbol{m}_N &= \boldsymbol{S}_N \{\boldsymbol{S}_0^{-1} \boldsymbol{m}_0 + \Phi^T \beta \mathbf{t}\} \end{aligned}$$

$$\boldsymbol{S}_0 = \alpha^{-1} \boldsymbol{I}$$

$$\alpha \rightarrow 0$$

$$\Rightarrow \boldsymbol{m}_N = \boldsymbol{w}_{ML}$$

$$\boldsymbol{w}_{ML} = (\Phi^T \Phi)^{-1} \Phi^T \mathbf{t}$$

$$N = 0 \Rightarrow \text{后验等于先验}$$

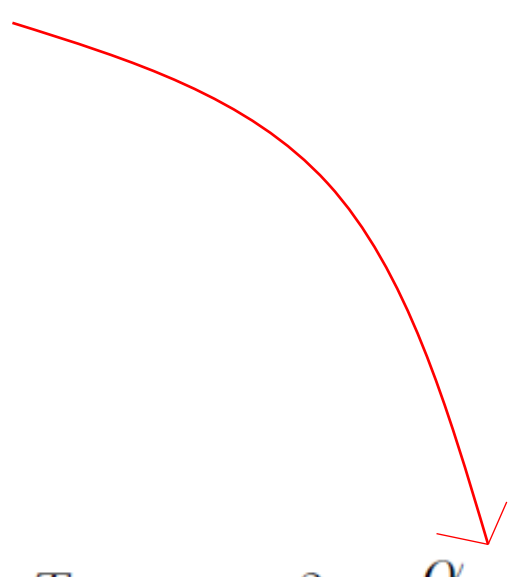
考虑0均值高斯分布

先验分布 $p(\mathbf{w} \mid \alpha) = \mathcal{N}(\mathbf{w} \mid \mathbf{0}, \alpha^{-1} \mathbf{I})$

$$\mathbf{m}_N = \beta \mathbf{S}_N \Phi^T \mathbf{t}$$

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

后验=先验*似然

$$\ln p(\mathbf{w} \mid \mathbf{t}) = -\frac{\beta}{2} \sum_{n=1}^N \{t_n - \mathbf{w}^T \phi(\mathbf{x}_n)\}^2 - \frac{\alpha}{2} \mathbf{w}^T \mathbf{w} + \text{常数}$$


后验分布关于 \mathbf{w} 的最大化等价于对平方和误差函数加上一个二次正则项进行最小化。

贝叶斯估计的增量学习过程

为了明确的表示样本集合 D 中有 n 个样本，这里采用记号： $D^n = \{x_1, x_2, \dots, x_n\}$ 。根据前一个公式，在 $n > 1$ 的情况下有：

$$p(D^n|\theta) = p(x_n|\theta)p(D^{n-1}|\theta)$$

可以很容易得到：

$$p(\theta|D^n) = \frac{p(\theta, D^n)}{p(D^n)} = \frac{p(D^n|\theta)p(\theta)}{\int p(D^n|\theta)p(\theta)d\theta} = \frac{p(x_n|\theta)p(D^{n-1}|\theta)p(\theta)}{\int p(x_n|\theta)p(D^{n-1}|\theta)p(\theta)d\theta} = \frac{\overset{\text{似然}}{p(x_n|\theta)}\overset{\text{先验}}{p(\theta|D^{n-1})}}{\int p(x_n|\theta)p(\theta|D^{n-1})d\theta}$$

当没有观测样本时，定义 $p(\theta|D^0) = p(\theta)$ ，为参数 θ 的初始估计。然后让样本集合依次进入上述公式，就可以得到一系列的概率密度函数： $p(\theta|D^0)$ 、 $p(\theta|D^1)$ 、 $p(\theta|D^2)$ 、...、 $p(\theta|D^n)$ ，这一过程称为参数估计贝叶斯递归法，也叫贝叶斯估计的增量学习。这是一个在线学习算法，它和随机梯度下降法有很多相似之处。

贝叶斯学习过程

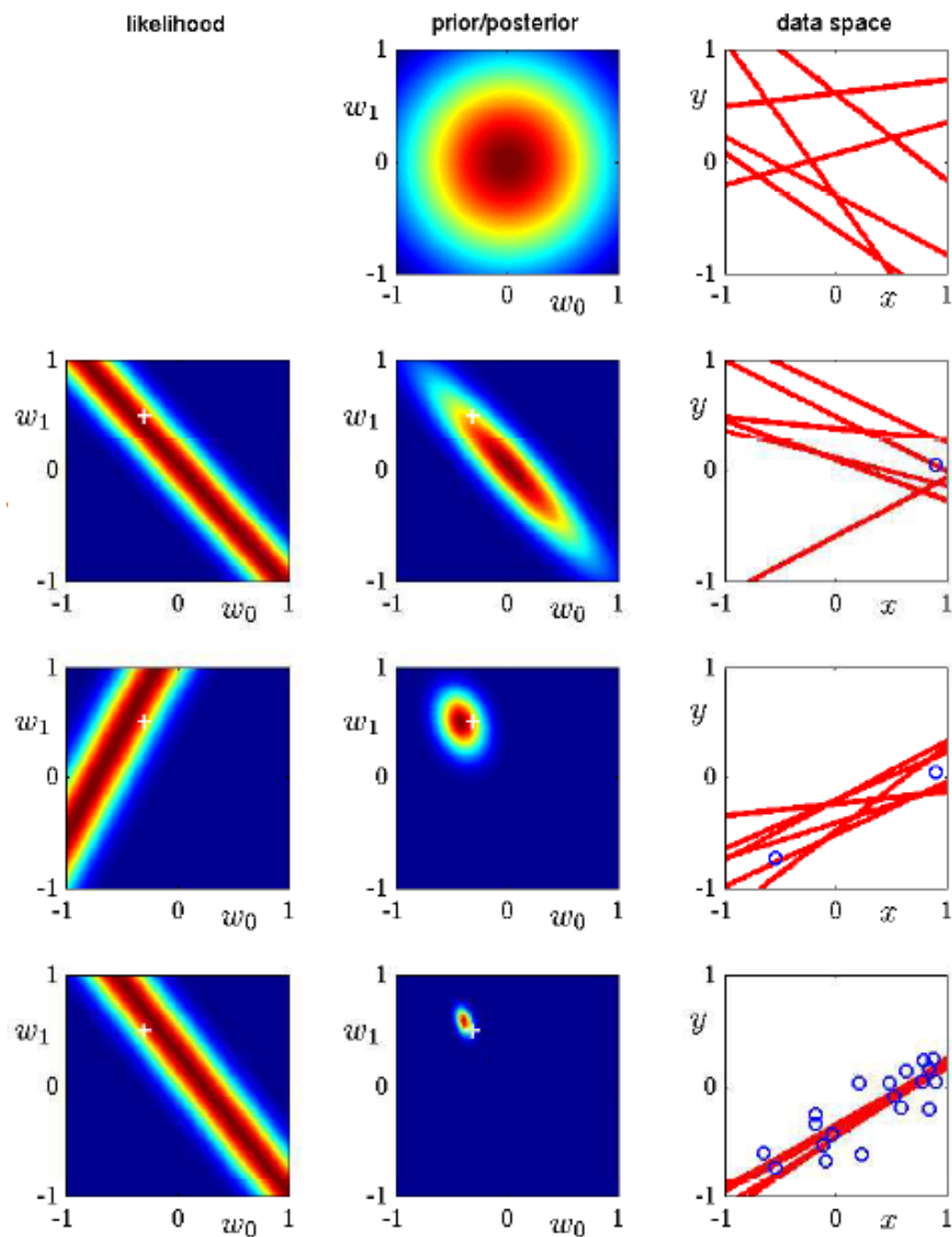
率分布的顺序更新过程。考虑一个单一输入变量 x ，一个单一目标变量 t ，以及一个形式为 $y(x, \mathbf{w}) = w_0 + w_1 x$ 的线性模型。由于这个模型只有两个可调节参数，因此我们可以直接在参数空间中画出先验分布和后验分布。我们从函数 $f(x, \mathbf{a}) = a_0 + a_1 x$ 中人工生成数据，其中 $a_0 = -0.3$ 且 $a_1 = 0.5$ 。生成数据的方法为：首先从均匀分布 $U(x | -1, 1)$ 中选择 x_n 的值，然后计算 $f(x_n, \mathbf{a})$ ，最后增加一个标准差为0.2的高斯噪声，得到目标变量 t_n 。我们的目标是从这样的数据中恢复 a_0 和 a_1 的值，并且我们想研究模型对于数据集规模的依赖关系。这里我们假设噪声方差是已知的，因此我们把精度参数设置为它的真实值 $\beta = (\frac{1}{0.2})^2 = 25$ 。类似地，我们把 α 固定

根据前面关于贝叶斯估计的增量学习可以很容易得到下面这个式子，这个就是贝叶斯学习过程：在前一个训练集合 D^{n-1} 的后验概率 $p(\theta | D^{n-1})$ 上，乘以新的测试样本点 x_n 的似然估计，得到新的集合 D^n 的后验概率 $p(\theta | D^n)$ ，这样，相当于 $p(\theta | D^{n-1})$ 成为了 $p(\theta | D^n)$ 的先验概率分布：

$$p(\theta | D^n) \propto p(x_n | \theta) p(\theta | D^{n-1})$$

贝叶斯学习过程

$$w_1 = \frac{1}{x}y - \frac{1}{x}w_0$$



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

预测分布

$$t = y(\boldsymbol{x}, \boldsymbol{w}) + \epsilon$$

$$p(t \mid \mathbf{t}, \alpha, \beta) = \int p(t \mid \boldsymbol{w}, \beta) p(\boldsymbol{w} \mid \mathbf{t}, \alpha, \beta) \, d\boldsymbol{w}$$

后验分布 $p(\boldsymbol{w} \mid \mathbf{t}) = \mathcal{N}(\boldsymbol{w} \mid \boldsymbol{m}_N, \boldsymbol{S}_N)$

条件概率 $p(t \mid \boldsymbol{x}, \boldsymbol{w}, \beta) = \mathcal{N}(t \mid \boldsymbol{w}^T \boldsymbol{\Phi}(\boldsymbol{x}), \beta^{-1})$

$$\left\{ \begin{array}{l} p(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1}) \\ p(\boldsymbol{y} \mid \boldsymbol{x}) = \mathcal{N}(\boldsymbol{y} \mid \boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}, \boldsymbol{L}^{-1}) \end{array} \right. \rightarrow p(\boldsymbol{y}) = \mathcal{N}(\boldsymbol{y} \mid \boldsymbol{A}\boldsymbol{\mu} + \boldsymbol{b}, \boldsymbol{L}^{-1} + \boldsymbol{A}\boldsymbol{\Lambda}^{-1}\boldsymbol{A}^T)$$

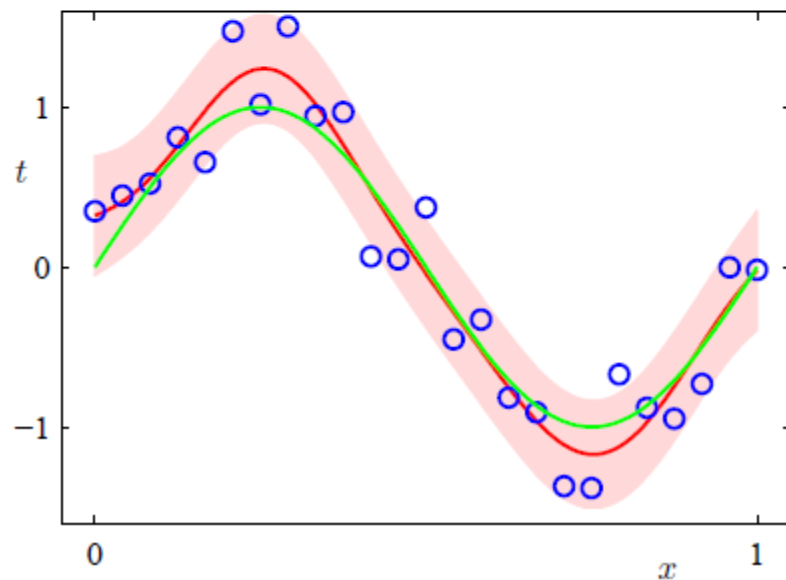
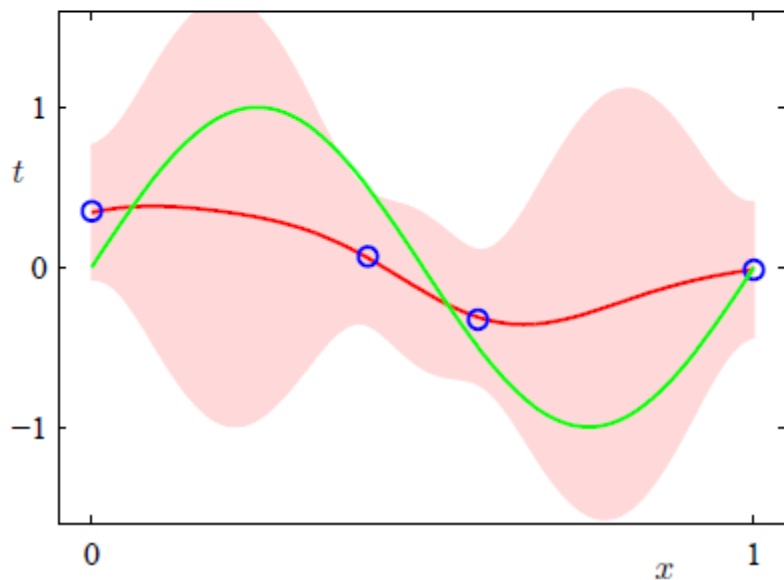
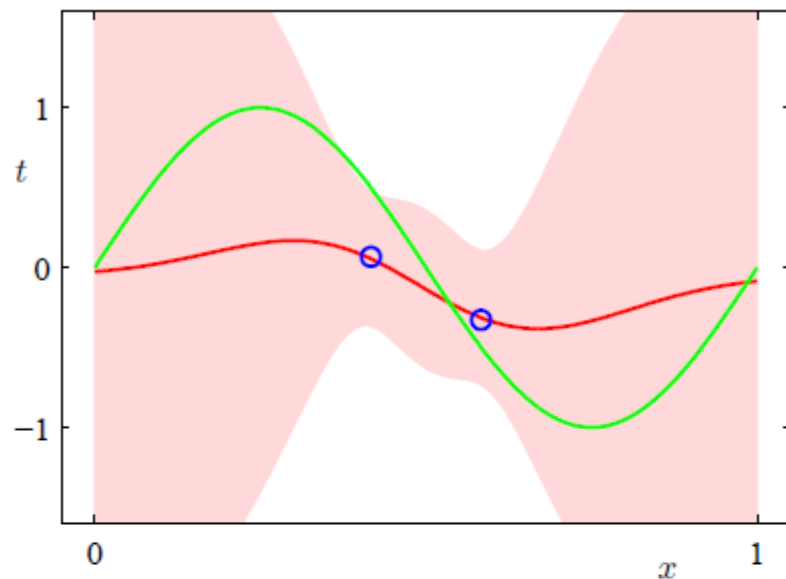
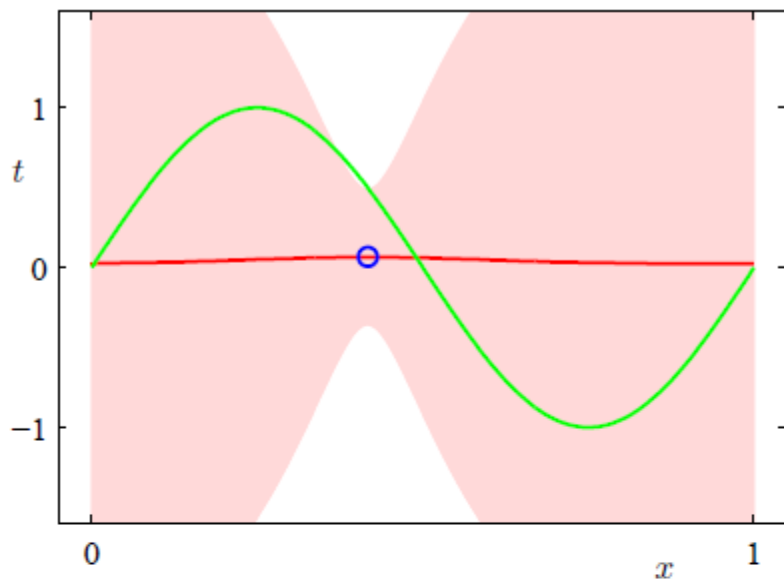
$$p(t \mid \boldsymbol{x}, \mathbf{t}, \alpha, \beta) = \mathcal{N}(t \mid \boldsymbol{m}_N^T \boldsymbol{\Phi}(\boldsymbol{x}), \sigma_N^2(\boldsymbol{x}))$$

$\sigma_N^2(\boldsymbol{x})$ 为

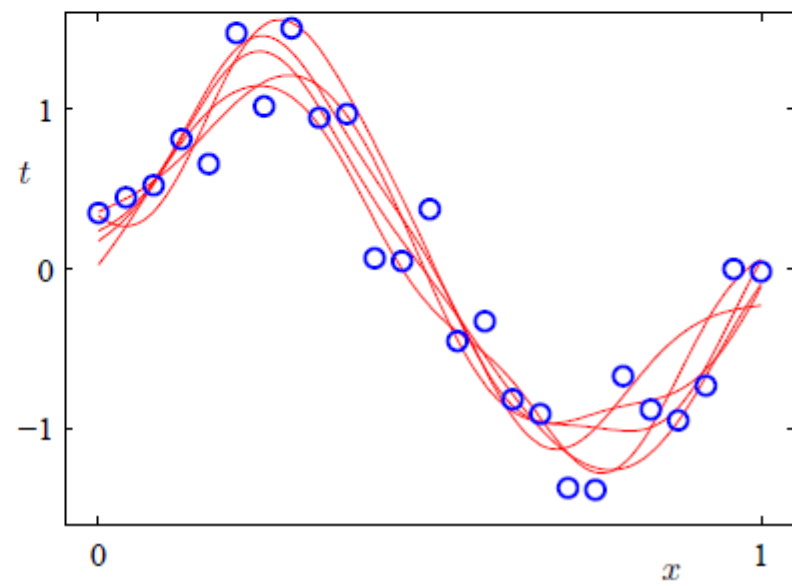
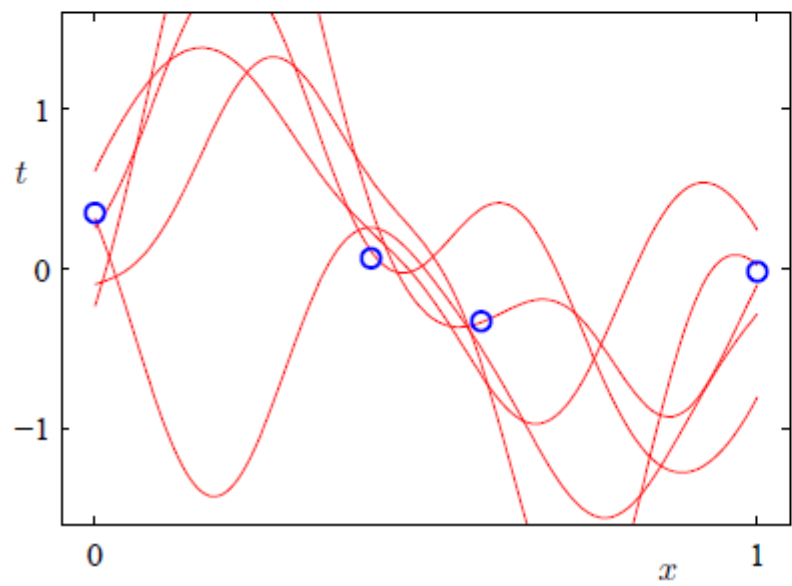
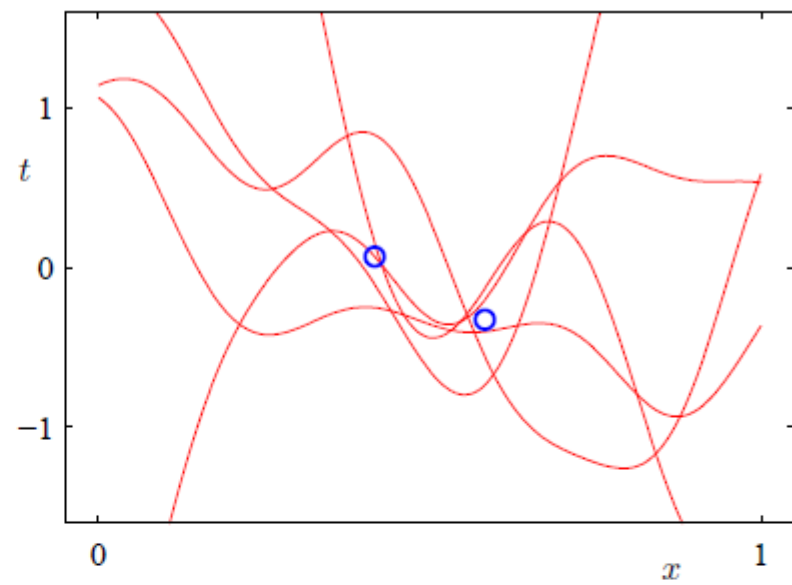
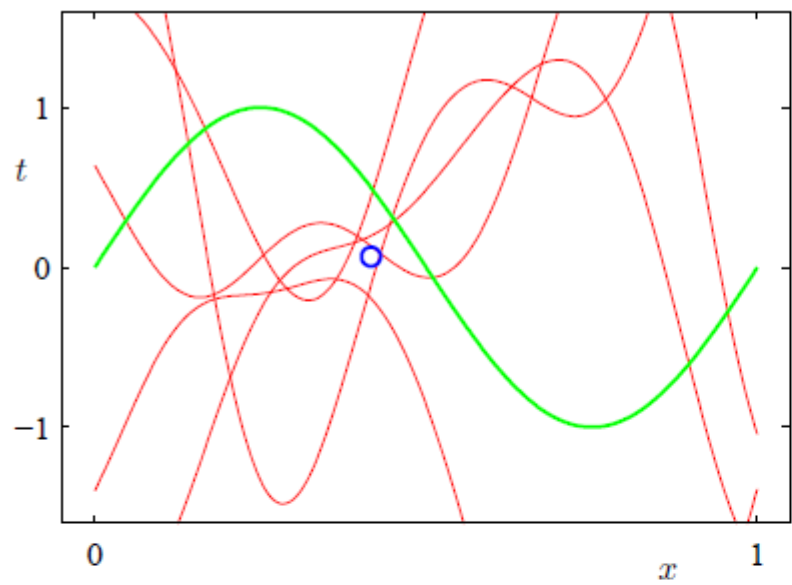
$$\sigma_N^2(\boldsymbol{x}) = \underbrace{\frac{1}{\beta}}_{\text{噪声}} + \underbrace{\boldsymbol{\Phi}(\boldsymbol{x})^T \boldsymbol{S}_N \boldsymbol{\Phi}(\boldsymbol{x})}_{\text{与参数 } \boldsymbol{w} \text{ 关联的不确定性}}$$

预测分布

- 红色曲线是预测分布的均值
 - 红色阴影区域是均值两侧的一个标准差范围的区域
-
- 预测的不确定性依赖于 x ，并且在数据点的邻域内最小
 - 不确定性程度随着观测到的数据点的增多而逐渐减小



预测分布



等价核

$$y(\boldsymbol{x}, \boldsymbol{w}) = \sum_{j=0}^{M-1} w_j \phi_j(\boldsymbol{x}) = \boldsymbol{w}^T \boldsymbol{\phi}(\boldsymbol{x})$$

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

$$p(\boldsymbol{w} \mid \mathbf{t}) = \mathcal{N}(\boldsymbol{w} \mid \boldsymbol{m}_N, \boldsymbol{S}_N)$$

预测均值

$$y(\boldsymbol{x}, \boldsymbol{m}_N) = \boldsymbol{m}_N^T \boldsymbol{\phi}(\boldsymbol{x}) = \beta \boldsymbol{\phi}(\boldsymbol{x})^T \boldsymbol{S}_N \boldsymbol{\Phi}^T \mathbf{t} = \sum_{n=1}^N \beta \boldsymbol{\phi}(\boldsymbol{x})^T \boldsymbol{S}_N \boldsymbol{\phi}(\boldsymbol{x}_n) t_n$$

$$y(\boldsymbol{x}, \boldsymbol{m}_N) = \sum_{n=1}^N k(\boldsymbol{x}, \boldsymbol{x}_n) t_n$$

$$k(\boldsymbol{x}, \boldsymbol{x}') = \beta \boldsymbol{\phi}(\boldsymbol{x})^T \boldsymbol{S}_N \boldsymbol{\phi}(\boldsymbol{x}')$$

等价核的局部性

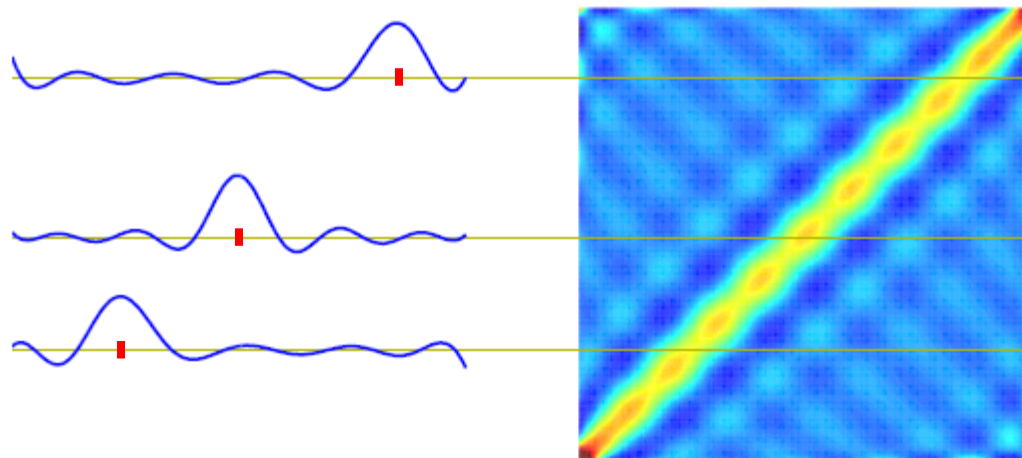


图 3.10: 图 3.1 中的高斯基函数的等价核 $k(x, x')$ ，图中给出了 x 关于 x' 的图像，以及通过这个矩阵的三个切片，对应于三个不同的 x 值。用来生成这个核的数据集由 x 的 200 个值组成， x 均匀地分布在区间 $(-1, 1)$ 中。

非局域基函数有
局域的等价核



图 3.11: $x = 0$ 时的等价核 $k(x, x')$ 的例子，图中给出了关于 x' 的函数图像。左图对应于多项式基函数，右图对应于 sigmoid 基函数，如图 3.1 所示。注意，这些是 x' 的局部函数，即使对应的基函数不是局部的。

等价核的局部性

$$p(\boldsymbol{w} \mid \mathbf{t}) = \mathcal{N}(\boldsymbol{w} \mid \boldsymbol{m}_N, \boldsymbol{S}_N)$$

$$k(\boldsymbol{x}, \boldsymbol{x}') = \beta \phi(\boldsymbol{x})^T \boldsymbol{S}_N \phi(\boldsymbol{x}')$$

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

根据等价核的形式，我们可以看到在附近的点处的预测均值相关性较高，而对于距离较远的点对，相关性就较低。

一般形式

$$k(\boldsymbol{x}, \boldsymbol{z}) = \psi(\boldsymbol{x})^T \psi(\boldsymbol{z})$$

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

可以不使用基函数，直接定义核函数来实现回归（或分类），即高斯过程（第6章）

等价核

对于所有的 \mathbf{x}_n ，有 $\sum_{n=1}^N k(\mathbf{x}, \mathbf{x}_n) = 1$

注1：上式可相当直觉地证明

注2：等价核可能为负

想理解这个式子，我们要知道非参数模型学习的过程：我们想对 x_{new} 进行预测得到 y_{new} ，并且在遇到 x_{new} 之前我们已经见到了 n 个样本 $\{(x_i, y_i) : i = 1, \dots, n\}$ ，那么 x_{new} 的预测值和这 n 个样本都有关系，关系的大小度量就是核 $k(x_{new}, x_i)$ 。所以我们可以把每个核看做一个权值，预测结果就是

$$y_{new} = \sum_{i=1}^n k(x_{new}, x_i) y_i$$

很明显这是一个加权求和，权值之和为1是很自然的。

3.4 贝叶斯模型比较

预测分布

$$p(t \mid \boldsymbol{x}, \mathcal{D}) = \sum_{i=1}^L p(t \mid \boldsymbol{x}, \mathcal{M}_i, \mathcal{D}) p(\mathcal{M}_i \mid \mathcal{D})$$

某一个模型的预测分布

权值

给定 \mathcal{D} ，估计
后验分布

$$p(\mathcal{M}_i \mid \mathcal{D}) \propto p(\mathcal{M}_i) p(\mathcal{D} \mid \mathcal{M}_i)$$

模型证据(边缘似然)，表达了数据展现出的不同模型的优先级

贝叶斯因子

$$\frac{p(\mathcal{D} \mid \mathcal{M}_i)}{p(\mathcal{D} \mid \mathcal{M}_j)}$$

模型选择：使用最可能的一个模型自己做预测

模型证据

$$p(\mathcal{D} \mid \mathcal{M}_i) = \int p(\mathcal{D} \mid \boldsymbol{w}, \mathcal{M}_i) p(\boldsymbol{w} \mid \mathcal{M}_i) \, d\boldsymbol{w}$$

从取样的角度来看，边缘似然函数可以被看成从一个模型中生成数据集 \mathcal{D} 的概率，这个模型的参数是从先验分布中随机取样的。

简单近似

假设后验分布在最大似然值 w_{MAP} 附近是一个尖峰，宽度为 $\Delta w_{\text{后验}}$

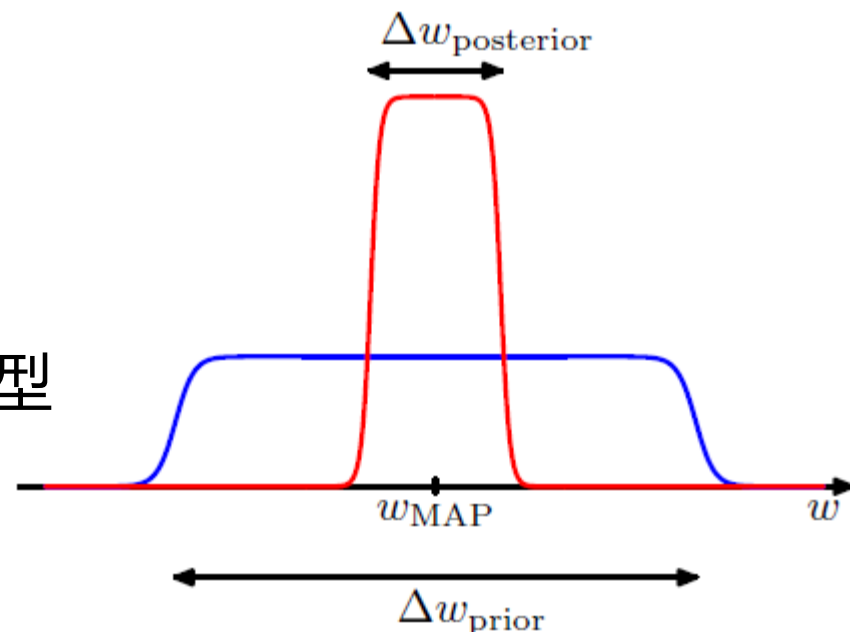
$$p(\mathcal{D}) = \int p(\mathcal{D} | w) p(w) dw \simeq p(\mathcal{D} | w_{MAP}) \frac{\Delta w_{\text{后验}}}{\Delta w_{\text{先验}}}$$

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

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

对数似然

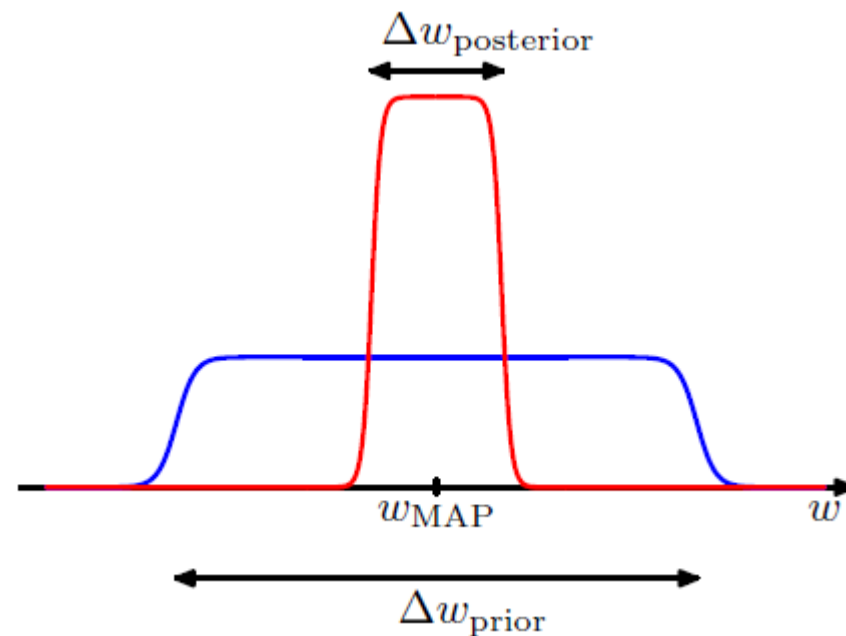
根据模型的复杂度来惩罚模型



对于一个有 M 个参数的模型，我们可以对每个参数进行类似的近似。假设所有的参数的 $\frac{\Delta w_{\text{后验}}}{\Delta w_{\text{先验}}}$ 都相同，我们有

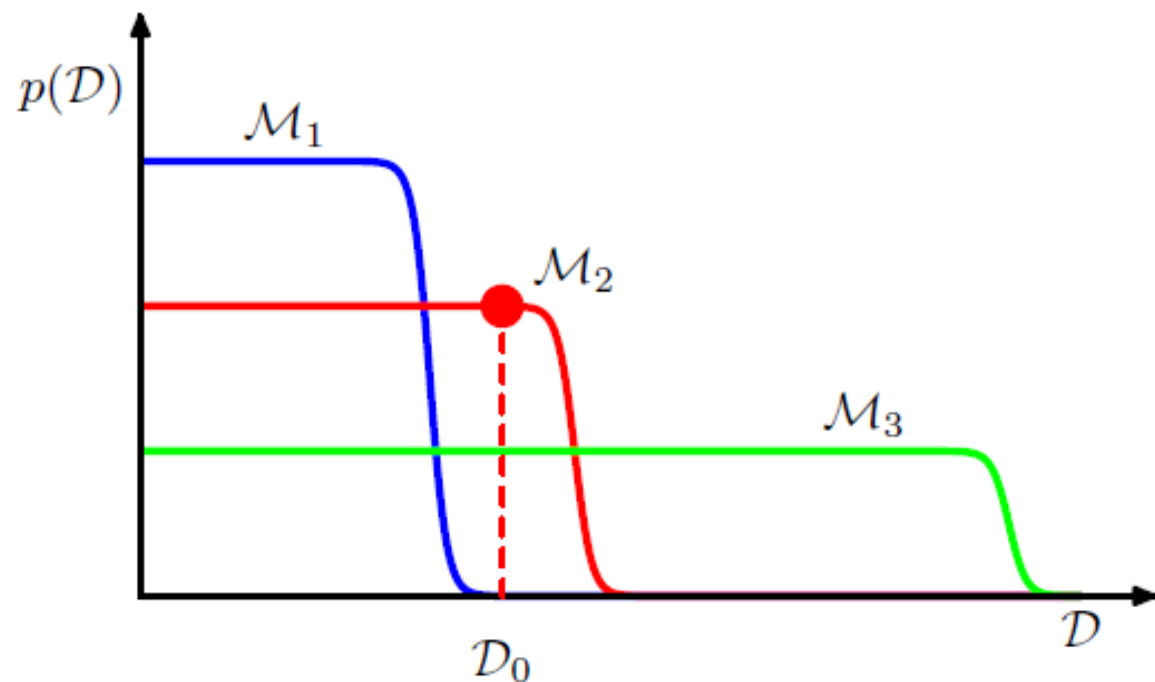
$$p(D) = \iiint p(D|w_1, w_2, \dots, w_n) p(w_1)p(w_2) \dots p(w_n)dw_1dw_2 \dots dw_n$$

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



举例：贝叶斯模型比较

$\mathcal{M}_1, \mathcal{M}_2$ 和 \mathcal{M}_3 ，复杂度依次增加



(横轴每个点对应一个具体的数据集)

简单的模型其值相对固定，而复杂的模型取值便相对比较多

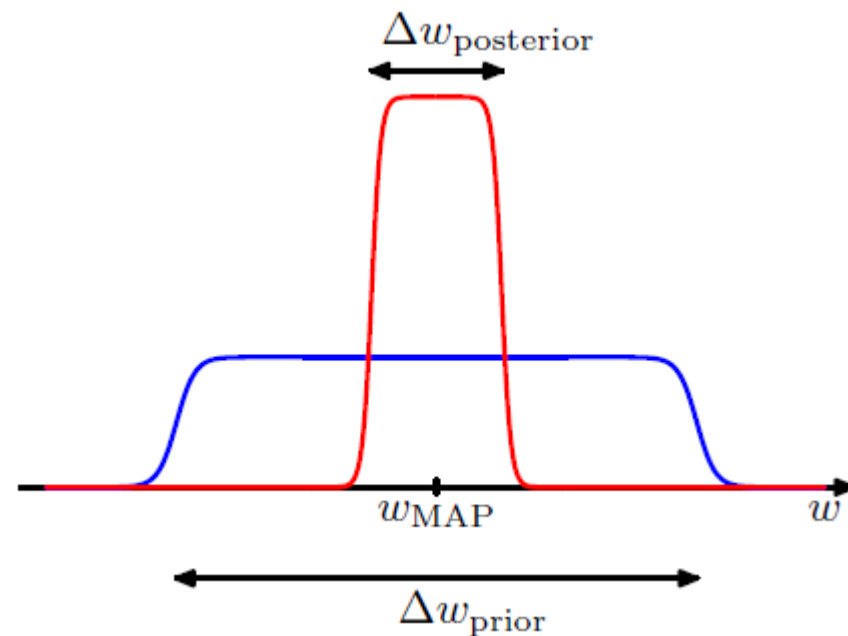
本质上说，简单的模型不能很好地拟合数据，而复杂的模型把它的预测概率散布于过多的可能的数据集当中，从而对它们当中的每一个赋予的概率都相对较小。

贝叶斯模型倾向于选择正确的模型

贝叶斯因子 $\frac{p(\mathcal{D}|\mathcal{M}_i)}{p(\mathcal{D}|\mathcal{M}_j)}$

期望贝叶斯因子 $\int p(\mathcal{D} | \mathcal{M}_1) \ln \frac{p(\mathcal{D} | \mathcal{M}_1)}{p(\mathcal{D} | \mathcal{M}_2)} d\mathcal{D}$ (KL散度)

真实的数据分布



3.5 证据近似

如果我们引入 α 和 β 上的超先验分布，那么预测分布可以通过对 w, α 和 β 求积分的方法得到，即

$$p(t \mid \mathbf{t}) = \iiint p(t \mid w, \beta) p(w \mid \mathbf{t}, \alpha, \beta) p(\alpha, \beta \mid \mathbf{t}) dw d\alpha d\beta \quad (3.74)$$

近似方法：我们首先对参数 w 求积分，得到边缘似然函数（marginal likelihood function），然后通过最大化边缘似然函数，确定超参数的值。

统计学中称为经验贝叶斯或者第二类最大似然

贝叶斯定理

$$p(\alpha, \beta \mid \mathbf{t}) \propto p(\mathbf{t} \mid \alpha, \beta) p(\alpha, \beta)$$

计算证据函数

边缘似然函数 $p(\mathbf{t} \mid \alpha, \beta)$ 是通过权值参数 \mathbf{w} 进行积分得到的，即

$$p(\mathbf{t} \mid \alpha, \beta) = \int p(\mathbf{t} \mid \mathbf{w}, \beta) p(\mathbf{w} \mid \alpha) d\mathbf{w}$$

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

$$= \frac{N}{2} \ln \beta - \frac{N}{2} \ln(2\pi) - \beta E_D(\mathbf{w})$$

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

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

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

$$E(\mathbf{w}) = \beta E_D(\mathbf{w}) + \alpha E_W(\mathbf{w})$$

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

推导：

$$p(\mathbf{t} \mid \alpha, \beta)$$

$$= \int \exp\left\{\frac{N}{2} \ln \frac{\beta}{2\pi} - \frac{\beta}{2} \sum_{i=1}^N \{t_n - \mathbf{w}^T \Phi\}^2\right\} \left(\frac{\alpha}{2\pi}\right)^{\frac{M}{2}} \exp\left\{-\frac{\alpha}{2} \mathbf{w}^2\right\} d\mathbf{w}$$

$$= \left(\frac{\beta}{2\pi}\right)^{\frac{N}{2}} \left(\frac{\alpha}{2\pi}\right)^{\frac{M}{2}} \int \exp\left\{-\frac{\beta}{2} \sum_{i=1}^N \{t_n - \mathbf{w}^T \Phi\}^2 - \frac{\alpha}{2} \mathbf{w}^2\right\} d\mathbf{w}$$

$$= \left(\frac{\beta}{2\pi}\right)^{\frac{N}{2}} \left(\frac{\alpha}{2\pi}\right)^{\frac{M}{2}} \int \exp\{-E(\mathbf{w})\} d\mathbf{w}$$

对 \mathbf{w} 配方:

现在对 \mathbf{w} 配平方, 可得

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

其中我们令

$$\mathbf{A} = \alpha \mathbf{I} + \beta \Phi^T \Phi = \nabla \nabla E(\mathbf{w})$$

以及

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

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

对w积分

$$\begin{aligned} & \int \exp\{-E(w)\} \, dw \\ &= \exp\{-E(m_N)\} \int \exp\left\{-\frac{1}{2}(w - m_N)^T A (w - m_N)\right\} \, dw \\ &= \exp\{-E(m_N)\} (2\pi)^{\frac{M}{2}} |A|^{-\frac{1}{2}} \end{aligned}$$

多元高斯分布

$$f = \frac{1}{(2\pi)^{\frac{M}{2}} |\Sigma|^{\frac{1}{2}}} \exp\left\{-\frac{(x-\mu)^T (\Sigma)^{-1} (x-\mu)}{2}\right\}$$

又因为

$$p(\mathbf{t} \mid \alpha, \beta) = \left(\frac{\beta}{2\pi}\right)^{\frac{N}{2}} \left(\frac{\alpha}{2\pi}\right)^{\frac{M}{2}} \int \exp\{-E(w)\} \, dw$$

=> 证据函数

$$\begin{aligned} \ln p(\mathbf{t} \mid \alpha, \beta) &= \ln\left\{\left(\frac{\beta}{2\pi}\right)^{\frac{N}{2}} \left(\frac{\alpha}{2\pi}\right)^{\frac{M}{2}} \exp\{-E(m_N)\} (2\pi)^{\frac{M}{2}} |A|^{-\frac{1}{2}}\right\} \\ &= \frac{N}{2} \ln \beta - \frac{N}{2} \ln(2\pi) + \frac{M}{2} \ln \alpha - \frac{M}{2} \ln 2\pi - E(m_N) + \frac{M}{2} \ln(2\pi) - \frac{1}{2} \ln |A| \\ &= \frac{N}{2} \ln \beta + \frac{M}{2} \ln \alpha - E(m_N) - \frac{1}{2} \ln |A| - \frac{N}{2} \ln(2\pi) \end{aligned}$$

最大化证据函数

证据函数 (边缘似然的对数)

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

$$\begin{aligned} (\beta \Phi^T \Phi) \mathbf{u}_i &= \lambda_i \mathbf{u}_i \\ \mathbf{A} &= \alpha \mathbf{I} + \beta \Phi^T \Phi \end{aligned} \quad \Rightarrow \quad \mathbf{A} \text{ 的特征值为 } \alpha + \lambda_i$$

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

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

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

$$\Rightarrow \alpha \mathbf{m}_N^T \mathbf{m}_N = M - \alpha \sum_i \frac{1}{\lambda_i + \alpha} = \gamma$$

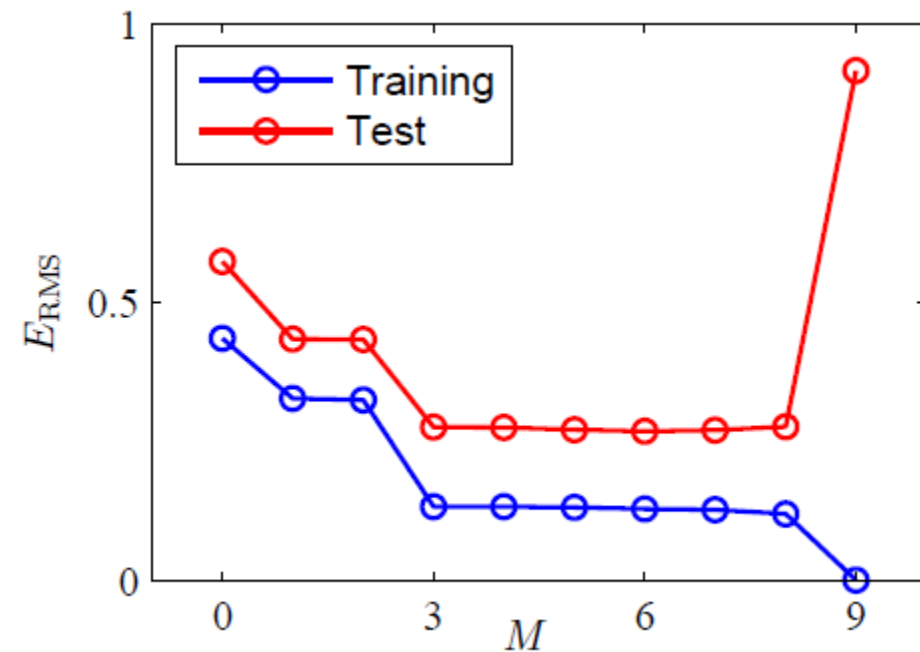
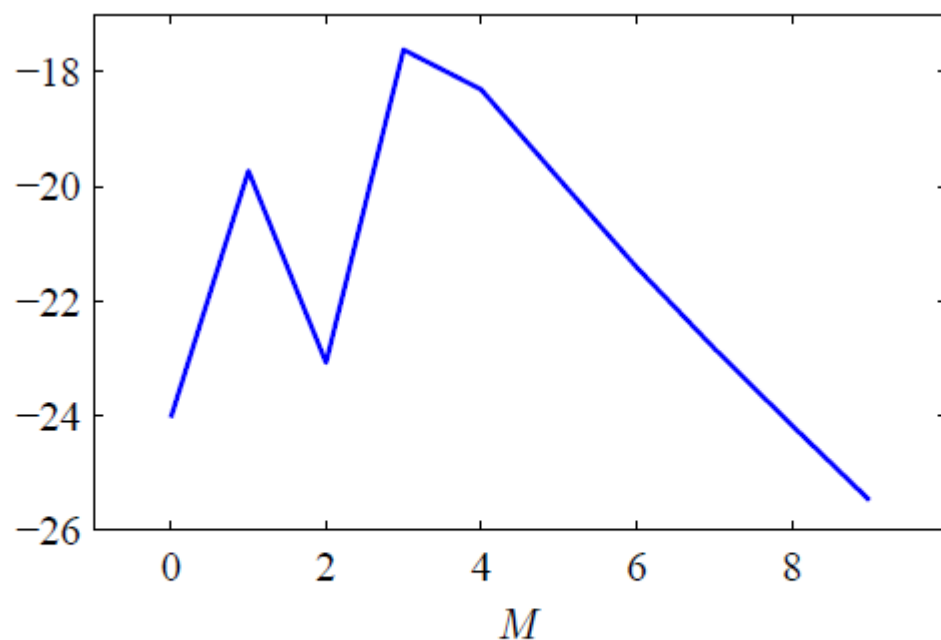
$$\Rightarrow \gamma = \sum_i \left(\frac{\lambda_i + \alpha}{\lambda_i + \alpha} - \frac{\alpha}{\lambda_i + \alpha} \right) = \sum_i \frac{\lambda_i}{\lambda_i + \alpha} \quad \alpha = \frac{\gamma}{\mathbf{m}_N^T \mathbf{m}_N}$$

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

$$0 = \frac{N}{2\beta} - \frac{1}{2} \sum_{n=1}^N \{t_n - \boldsymbol{m}_N^T \phi(\boldsymbol{x}_n)\}^2 - \frac{\gamma}{2\beta}$$

$$\frac{1}{\beta} = \frac{1}{N - \gamma} \sum_{n=1}^N \{t_n - \boldsymbol{m}_N^T \phi(\boldsymbol{x}_n)\}^2$$

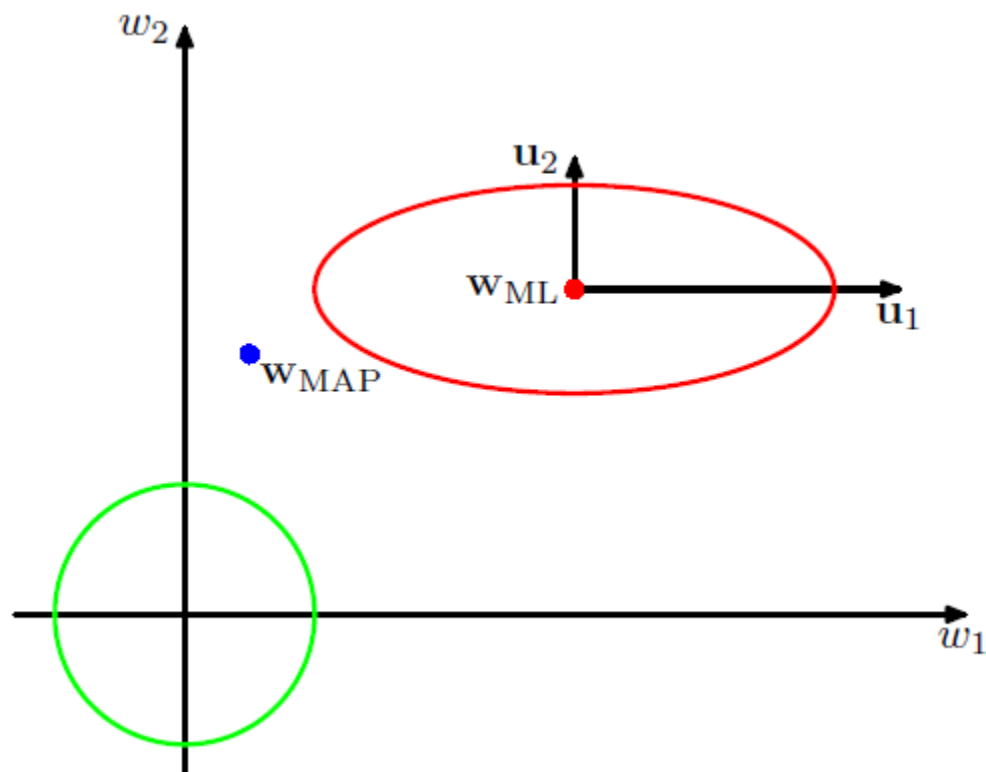
模型证据与多项式阶数的关系



根均方误差

多项式回归模型的模型对数证据与阶数 M 的关系图像，表明证据倾向于选择 $M = 3$ 的模型

参数的有效数量



特征值1小于2（因为较小的曲率对应着似然函数轮廓线较大的拉伸）

良好确定的参数：

$$\gamma = \sum_i \frac{\lambda_i}{\alpha + \lambda_i} \text{ 在 } 0 \sim M \text{ 之间}$$

中，特征值 λ_1 小于 λ_2 （因为较小的曲率对应着似然函数轮廓线较大的拉伸）。由于 $\beta \Phi^T \Phi$ 是一个正定矩阵，因此它的特征值为正数，从而比值 $\frac{\lambda_i}{\lambda_i + \alpha}$ 位于0和1之间。结果，由公式（3.91）定义的 γ 的取值范围为 $0 \leq \gamma \leq M$ 。对于 $\lambda_i \gg \alpha$ 的方向，对应的参数 w_i 将会与最大似然值接近，且比值 $\frac{\lambda_i}{\lambda_i + \alpha}$ 接近1。这样的参数被称为良好确定的（well determined），因为它们的值被数据紧紧

参数的有效数量

$$\frac{1}{\beta_{ML}} = \frac{1}{N} \sum_{n=1}^N \{t_n - \mathbf{w}_{ML}^T \phi(\mathbf{x}_n)\}^2 \quad \leftarrow \text{最大似然的结果}$$

$$\frac{1}{\beta} = \frac{1}{N - \gamma} \sum_{n=1}^N \{t_n - \mathbf{m}_N^T \phi(\mathbf{x}_n)\}^2 \quad \leftarrow \text{贝叶斯求参的结果}$$

$$s^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$$
$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$$

分母中的因子 $N - 1$ 反映了模型中的一个自由度被用于拟合均值的事实，它抵消了最大似然解的偏差。现在考虑线性回归模型的对应的结果。目标分布的均值现在由函数 $\mathbf{w}^T \phi(\mathbf{x})$ 给出，它包含了 M 个参数。但是，并不是所有的这些参数都按照数据进行了调解。由数据确定的有效参数的数量为 γ ，剩余的 $M - \gamma$ 个参数被先验概率分布设置为较小的值。这可以通过方差的贝叶斯结果中的因子 $N - \gamma$ 反映出来，因此修正了最大似然结果的偏差。

总结

- 1、主要内容为：线性基函数模型（最小平方问题、正则化等）、偏差-方差分解、贝叶斯方法、预测分布、贝叶斯比较、证据函数等。
- 2、线性模型仍然具有非常大的局限性。
- 3、基函数的数量随着输入空间的维度迅速增长，通常是指数方式的增长。