3 回归的线性模型

目前为止，本书的关注点是无监督学习，包括诸如概率密度估计和数据聚类等话题。我们现在开始讨论有监督学习，首先讨论的是回归问题。回归问题的目标是在给定D维输入（input）变量$\boldsymbol{x}$的情况下，预测一个或者多个连续目标（target）变量t的值。在第1章中，我们已经遇到了回归问题的一个例子：多项式曲线拟合问题。多项式是被称为线性回归模型的一大类函数的一个具体的例子。线性回归模型有着可调节的参数，具有线性函数的性质，将会成为本章的关注点。线性回归模型的最简单的形式也是输入变量的线性函数。但是，通过将一组输入变量的非线性函数进行线性组合，我们可以获得一类更加有用的函数，被称为基函数（basis function）。这样的模型是参数的线性函数，这使得其具有一些简单的分析性质，同时关于输入变量是非线性的。

给定一个由N个观测值$\{\boldsymbol{x}_n\}$组成的数据集，其中$n = 1, \ldots, N$，以及对应的目标值$\{t_n\}$，我们的目标是预测对于给定新的$\boldsymbol{x}$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$y(\boldsymbol{x})$，对于新的输入$\boldsymbol{x}$，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$p(t \mid \boldsymbol{x})$建模，因为它表达了对于每个$\boldsymbol{x}$值，我们对于t的值的不确定性。从这个条件概率分布中，对于任意的$\boldsymbol{x}$的新值，我们可以对t进行预测，这种方法等同于最小化一个恰当选择的损失函数的期望值。正如在1.5.5节讨论的那样，对于实值变量来说，损失函数的一个通常的选择是平方误差损失，这种情况下最优解由t的条件期望给出。

虽然线性模型对于模式识别的实际应用来说有很大的局限性，特别是对于涉及到高维输入空间的问题来说更是如此，但是他们有很好的分析性质，并且组成了后续章节中将要讨论的更加复杂的模型的基础。

3.1 线性基函数模型

回归问题的最简单模型是输入变量的线性组合

$$y(\boldsymbol{x}, \boldsymbol{w}) = w_0 + w_1 x_1 + \ldots + w_D x_D \tag{3.1}$$

其中$\boldsymbol{x} = (x_1, \ldots, x_D)^T$。这通常被简单地称为线性回归（linear regression）。这个模型的关键性质是它是参数$w_0, \ldots, w_D$的一个线性函数。但是，它也是输入变量x_i的一个线性函数，这给模型带来的极大的局限性。因此我们这样扩展模型的类别：将输入变量的固定的非线性函数进行线性组合，形式为

$$y(\boldsymbol{x}, \boldsymbol{w}) = w_0 + \sum_{j=1}^{M-1} w_j \phi_j(\boldsymbol{x}) \tag{3.2}$$

其中$\phi_j(\boldsymbol{x})$被称为基函数（basis function）。通过把下标j的最大值记作$M - 1$，这个模型中的参数总数为M。

参数w_0使得数据中可以存在任意固定的偏置，这个值通常被称为偏置参数（bias parameter）。注意不要把这里的“偏置”与统计学中的“偏置”弄混淆。通常，定义一个额外的虚“基函数”$\phi_0(\boldsymbol{x}) = 1$是很方便的，这时

$$y(\boldsymbol{x}, \boldsymbol{w}) = \sum_{j=0}^{M-1} w_j \phi_j(\boldsymbol{x}) = \boldsymbol{w}^T \boldsymbol{\phi}(\boldsymbol{x}) \tag{3.3}$$

其中$\boldsymbol{w} = (w_0, \ldots, w_{M-1})^T$且$\boldsymbol{\phi} = (\phi_0, \ldots, \phi_{M-1})^T$。在许多模式识别的实际应用中，我们会对原始的数据变量进行某种固定形式的预处理或者特征抽取。如果原始变量由向量$\boldsymbol{x}$组成，那么特征可以用基函数$\{\phi_j(\boldsymbol{x})\}$来表示。

通过使用非线性基函数，我们能够让函数$y(\boldsymbol{x}, \boldsymbol{w})$成为输入向量$\boldsymbol{x}$的一个非线性函数。但是，形如（3.2）的函数被称为线性模型，因为这个函数是$\boldsymbol{w}$的线性函数。正是这种关于参数的线性极大地简化了对于这列模型的分析。然而，这也造成了一些巨大的局限性，正如我们在3.6节讨论的那样。

第1章中讨论的多项式拟合的例子是这个模型的一个特例，那里有一个输入变量x，基函数是x的幂指数的形式，即$\phi_j(x) = x^j$。多项式基函数的一个局限性是它们是输入变量的全局函

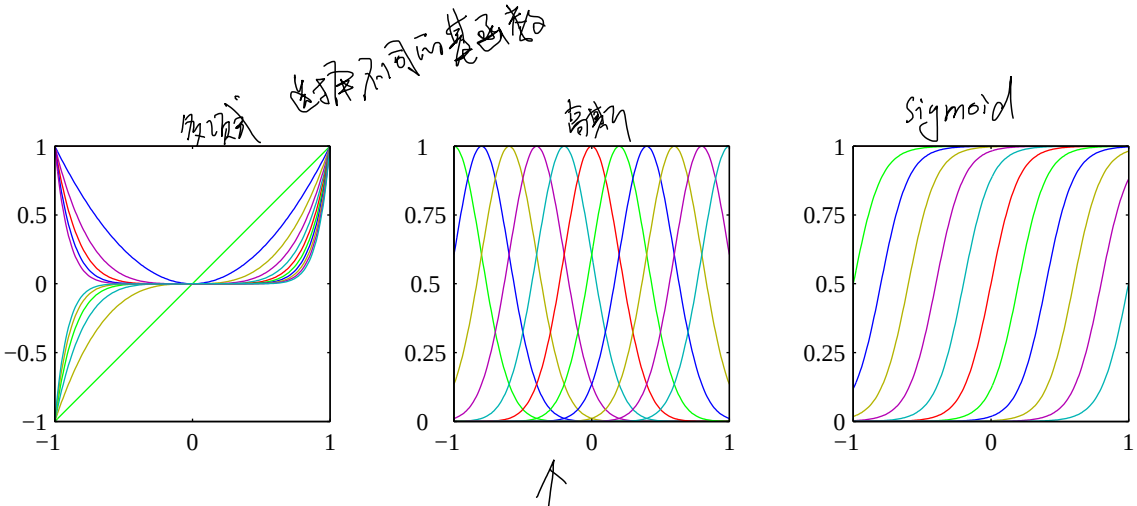

图 3.1: 基函数的例子，左图是多项式基函数，中图是形式为（3.4）的高斯基函数，右图是形式为（3.5）的sigmoid基函数。

数，因此对于输入空间一个区域的改变将会影响所有其他的区域。这个问题可以这样解决：把输入空间切分成若干个区域，然后对于每个区域用不同的多项式函数拟合。这样的函数叫做样条函数（spline function）（Hastie et al., 2001）。

对于基函数，有许多其他的选择，例如

$$\phi_j(x) = \exp\left\{-\frac{(x-\mu_j)^2}{2s^2}\right\} \tag{3.4}$$

其中μ_j控制了基函数在输入空间中的位置，参数s控制了基函数的空间大小。这种基函数通常被称为“高斯”基函数，但是应该注意它们未必一定是一个概率表达式。特别地，归一化系数不重要，因为这些基函数会与一个调节参数w_j相乘。

另一种选择是sigmoid基函数，形式为

$$\phi_j(x) = \sigma\left(\frac{x-\mu_j}{s}\right) \tag{3.5}$$

其中$\sigma(a)$是logistic sigmoid函数，定义为

$$\sigma_a = \frac{1}{1+\exp(-a)} \tag{3.6}$$

等价地，我们可以使用tanh函数，因为它和logistic sigmoid函数的关系为$\tanh(a) = 2\sigma(2a) - 1$，因此logistic sigmoid函数的一般的线性组合等价于tanh函数的一般的线性组合。图3.1说明了基函数的不同选择情况。

基函数的另一种可能的选择是傅里叶基函数，它可以用正弦函数展开。每个基函数表示一个具体的频率，它在空间中有无限的延伸。相反，限制在输入空间中的有限区域的基函数要由不同空间频率的一系列频谱组成。在许多信号处理的应用中，一个吸引了研究者兴趣的问题是考虑同时在空间和频率受限的基函数。这种研究产生了一类被称为小波（wavelet）的函数。为了简化应用，这些基函数被定义为相互正交的。当应用中的输入值位于正规的晶格中时，应用小波最合适。这种应用包括时间序列中的连续的时间点，以及图像中的像素。关于小波的有用的教科书包括Ogden（1997），Mallat （1999）和Vidakovic（1999）。

但是，本章中的大部分讨论都与基函数的选择无关。因此对于我们的大部分讨论，我们不会具体化基函数的特定形式，除非我们为了数值说明。事实上，我们的大部分讨论将会同等地适用于基函数向量$\boldsymbol{\phi}(\boldsymbol{x})$的形式为$\boldsymbol{\phi}(\boldsymbol{x}) = \boldsymbol{x}$的情形。此外，为了保持记号的简洁，我们把注意力集中于单一目标变量t的情形。但是在3.1.5节里，我们将会简短地考虑必要的修改，来处理多个目标变量的情形。

3.1.1 最大似然与最小平方

在第1章，我们通过最小化平方和误差函数，用多项式函数拟合数据集。我们也证明了，这种误差函数可以看成高斯噪声模型的假设下的最大似然解。现在让我们回到这种讨论中，更加详细地考虑最小平方的方法以及它与最大似然方法的关系。

与之前一样，我们假设目标变量t由确定的函数$y(\boldsymbol{x}, \boldsymbol{w})$给出，这个函数被附加了高斯噪声，即

$$t = y(\boldsymbol{x}, \boldsymbol{w}) + \epsilon \tag{3.7}$$

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。因此我们有

$$p(t \mid \boldsymbol{x}, \boldsymbol{w}, \beta) = \mathcal{N}(t \mid y(\boldsymbol{x}, \boldsymbol{w}), \beta^{-1}) \tag{3.8}$$

回忆一下，如果我们假设一个平方损失函数，那么对于$\boldsymbol{x}$的一个新值，最优的预测由目标变量的条件均值给出。在公式（3.8）给出的高斯条件分布的情况下，条件均值可以简单地写成

$$\mathbb{E}[t \mid \boldsymbol{x}] = \int t p(t \mid \boldsymbol{x}) \, \mathrm{d}t = y(\boldsymbol{x}, \boldsymbol{w}) \tag{3.9}$$

注意高斯噪声的假设表明，给定$\boldsymbol{x}$的条件下，t的条件分布是单峰的，这对于一些实际应用来说是不合适的。第14.5.1节将扩展到条件高斯分布的混合，那种情况下可以描述多峰的条件分布。

现在考虑一个输入数据集$\boldsymbol{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$，对应的目标值为$t_1, \ldots, t_N$。我们把目标向量$\{t_n\}$组成一个列向量，记作$\mathbf{t}$。这个变量的字体与多元目标值的一次观测（记作$\boldsymbol{t}$）不同。假设这些数据点是独立地从分布（3.8）中抽取的，那么我们可以得到下面的似然函数的表达式，它是可调节参数$\boldsymbol{w}$和β的函数，形式为

$$p(\mathbf{t} \mid \boldsymbol{X}, \boldsymbol{w}, \beta) = \prod_{n=1}^{N} \mathcal{N}(t_n \mid \boldsymbol{w}^T \boldsymbol{\phi}(\boldsymbol{x}_n), \beta^{-1}) \tag{3.10}$$

其中我们使用了公式（3.3）。注意，在有监督学习问题中（例如回归问题和分类问题），我们不是在寻找模型来对输入变量的概率分布建模。因此$\boldsymbol{x}$总会出现在条件变量的位置上。因此从现在开始，为了保持记号的简洁性，我们在诸如$p(\mathbf{t} \mid \boldsymbol{x}, \boldsymbol{w}, \beta)$这类的表达式中不显式地写出$\boldsymbol{x}$。取对数似然函数的对数，使用一元高斯分布的标准形式（2.146），我们有

$$\begin{aligned} \ln p(\mathbf{t} \mid \boldsymbol{w}, \beta) &= \sum_{n=1}^{N} \ln \mathcal{N}(t_n \mid \boldsymbol{w}^T \boldsymbol{\phi}(\boldsymbol{x}_n), \beta^{-1}) \\ &= \frac{N}{2} \ln \beta - \frac{N}{2} \ln(2\pi) - \beta E_D(\boldsymbol{w}) \end{aligned} \tag{3.11}$$

其中平方和误差函数的定义为

$$E_D(\boldsymbol{w}) = \frac{1}{2} \sum_{n=1}^{N} \{t_n - \boldsymbol{w}^T \boldsymbol{\phi}(\boldsymbol{x}_n)\}^2 \tag{3.12}$$

写出了似然函数，我们可以使用最大似然的方法确定$\boldsymbol{w}$和β。首先关于$\boldsymbol{w}$求最大值。正如我们已经在1.2.5节中已经看到的那样，我们看到在条件高斯噪声分布的情况下，线性模型的似然函数的最大化等价于平方和误差函数的最小化。平方和误差函数由$E_D(\boldsymbol{w})$给出。公式（3.11）给出的对数似然函数的梯度为

$$\nabla \ln p(\mathbf{t} \mid \boldsymbol{w}, \beta) = \beta \sum_{n=1}^{N} \{t_n - \boldsymbol{w}^T \boldsymbol{\phi}(\boldsymbol{x}_n)\} \boldsymbol{\phi}(\boldsymbol{x}_n)^T \tag{3.13}$$

令这个梯度等于零，可得

$$0 = \sum_{n=1}^{N} t_n \boldsymbol{\phi}(\boldsymbol{x}_n)^T - \boldsymbol{w}^T \left(\sum_{n=1}^{N} \boldsymbol{\phi}(\boldsymbol{x}_n) \boldsymbol{\phi}(\boldsymbol{x}_n)^T \right) \tag{3.14}$$

求解$\boldsymbol{w}$，我们有

$$\boldsymbol{w}_{ML} = (\boldsymbol{\Phi}^T \boldsymbol{\Phi})^{-1} \boldsymbol{\Phi}^T \mathbf{t} \tag{3.15}$$

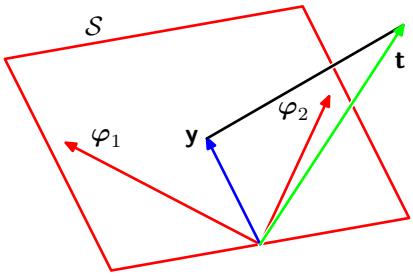

图 3.2: 最小平方解的几何表示，在一个N维空间中，坐标轴是$t_1, \ldots, t_N$的值。最小平方回归函数可以通过下面的方式得到：寻找数据向量$\mathbf{t}$在由基函数$\phi_j(\boldsymbol{x})$张成的子空间上的正交投影，其中每个基函数都可以看成一个长度为N的向量$\boldsymbol{\varphi}_j$，它的元素为$\phi_j(\boldsymbol{x}_n)$。

这被称为最小平方问题的规范方程（normal equation）。这里$\boldsymbol{\Phi}$是一个$N \times M$的矩阵，被称为设计矩阵（design matrix），它的元素为$\Phi_{nj} = \phi_j(\boldsymbol{x}_n)$，即

$$\boldsymbol{\Phi} = \begin{pmatrix} \phi_0(\boldsymbol{x}_1) & \phi_1(\boldsymbol{x}_1) & \cdots & \phi_{M-1}(\boldsymbol{x}_1) \\ \phi_0(\boldsymbol{x}_2) & \phi_1(\boldsymbol{x}_2) & \cdots & \phi_{M-1}(\boldsymbol{x}_2) \\ \vdots & \vdots & \ddots & \vdots \\ \phi_0(\boldsymbol{x}_N) & \phi_1(\boldsymbol{x}_N) & \cdots & \phi_{M-1}(\boldsymbol{x}_N) \end{pmatrix} \tag{3.16}$$

量

$$\boldsymbol{\Phi}^\dagger \equiv (\boldsymbol{\Phi}^T\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^T \tag{3.17}$$

被称为矩阵$\boldsymbol{\Phi}$的Moore-Penrose伪逆矩阵（pseudo-inverse matrix）（Rao and Mitra, 1971; Golub and Van Loan, 1996）。它可以被看成逆矩阵的概念对于非方阵的矩阵的推广。实际上，如果$\boldsymbol{\Phi}$是方阵且可逆，那么使用性质$(\boldsymbol{AB})^{-1} = \boldsymbol{B}^{-1}\boldsymbol{A}^{-1}$，我们可以看到$\boldsymbol{\Phi}^\dagger \equiv \boldsymbol{\Phi}^{-1}$。

现在，我们可以更加深刻地认识偏置参数w_0。如果我们显式地写出偏置参数，那么误差函数（3.12）变为

$$E_D(\boldsymbol{w}) = \frac{1}{2}\sum_{n=1}^{N}\{t_n - w_0 - \sum_{j=1}^{M-1} w_j\phi_j(\boldsymbol{x}_n)\}^2 \tag{3.18}$$

令关于w_0的导数等于零，解出w_0，可得

$$w_0 = \bar{t} - \sum_{j=1}^{M-1} w_j\bar{\phi}_j \tag{3.19}$$

其中我们已经定义了

$$\bar{t} = \frac{1}{N}\sum_{n=1}^{N} t_n, \quad \bar{\phi}_j = \frac{1}{N}\sum_{n=1}^{N}\phi_j(\boldsymbol{x}_n) \tag{3.20}$$

因此偏置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。

我们也可以关于噪声精度参数β最大化似然函数（3.11），结果为

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

因此我们看到噪声精度的倒数由目标值在回归函数周围的残留方差（residual variance）给出。

3.1.2 最小平方的几何描述

现在，考虑最小平方解的几何描述有助于理解这种方法。我们考虑一个N维空间，它的坐标轴由t_n给出，即$\mathbf{t} = (t_1, \ldots, t_N)^T$是这个空间中的一个向量。每个在$N$个数据点处估计的基函数$\phi_j(\boldsymbol{x}_n)$也可以表示为这个空间中的一个向量，记作$\boldsymbol{\varphi}_j$，如图3.2所示。注意，$\boldsymbol{\varphi}_j$对应于$\boldsymbol{\Phi}$的第$j$列，而$\boldsymbol{\phi}(\boldsymbol{x}_n)$对应于$\boldsymbol{\Phi}$的第$i$行。如果基函数的数量$M$小于数据点的数量$N$，那么$M$个向量$\boldsymbol{\varphi}_j$将会张成一个$M$维的子空间$\mathcal{S}$。我们定义$\mathbf{y}$是一个$N$维向量，它的第$n$个元素为$y(\boldsymbol{x}_n, \boldsymbol{w})$，其中$n = 1, \ldots, N$。由于$\mathbf{y}$是向量$\boldsymbol{\varphi}_j$的任意线性组合，因此它可以位于$M$维子空间的任何位置。这样，平方和误差函数（3.12）就等于$\mathbf{y}$和$\mathbf{t}$之间的平方欧氏距离（只相差一个因子$\frac{1}{2}$）。因此，$\boldsymbol{w}$的最小平方解对应于位于子空间$\mathcal{S}$的与$\mathbf{t}$最近的$\mathbf{y}$的选择。直观来看，根据图3.2，我们猜想这个解对应于$\mathbf{t}$在子空间$\mathcal{S}$上的正交投影。事实上确实是这样，并且很容易证明。注意到$\mathbf{y}$是由$\boldsymbol{\Phi}\boldsymbol{w}_{ML}$给出的，然后证明它的表达式为正交投影即可。

在实际应用中，当$\boldsymbol{\Phi}^T\boldsymbol{\Phi}$接近奇异矩阵时，直接求解规范方程会导致数值计算上的困难。特别地，当两个或者更多的基向量$\boldsymbol{\varphi j}$共线或者接近共线时，最终的参数值会相当大。这样的退化在处理真实数据集的时候并不罕见。这种数值计算上的困难可以通过奇异值分解（singular value decomposition）或者简称SVD的方法解决（Press et al., 1992; Bishop and Nabney, 2008）。注意，正则项的添加确保了矩阵是非奇异的，即使在退化的情况下也是如此。

3.1.3 顺序学习

最大似然解（3.15）的求解过程涉及到一次处理整个数据集。这种批处理技术对于大规模数据集来说计算量相当大。正如我们在第1章讨论的那样，如果数据集充分大，那么使用顺序算法（也被称为在线算法）可能更有价值。顺序算法中，每次只考虑一个数据点，模型的参数在每观测到一个数据点之后进行更新。顺序学习也适用于实时的应用。在实时应用中，数据观测以一个连续的流的方式持续到达，我们必须在观测到所有数据之前就做出预测。

我们可以获得一个顺序学习的算法通过考虑随机梯度下降（stochastic gradient descent）也被称为顺序梯度下降（sequential gradient descent）的方法。如果误差函数由数据点的和组成$E = \sum_n E_n$，那么在观测到模式n之后，随机梯度下降算法使用下式更新参数向量$\boldsymbol{w}$

$$\boldsymbol{w}^{(\tau+1)} = \boldsymbol{w}^{(\tau)} - \eta \nabla E_n \tag{3.22}$$

其中τ表示迭代次数，η是学习率参数。我们稍后会讨论η的选择问题。$\boldsymbol{w}$被初始化为某个起始向量$\boldsymbol{w}^{(0)}$。对于平方和误差函数（3.12）的情形，我们有

$$\boldsymbol{w}^{(\tau+1)} = \boldsymbol{w}^{(\tau)} + \eta(t_n - \boldsymbol{w}^{(\tau)T}\boldsymbol{\phi}_n)\boldsymbol{\phi}_n \tag{3.23}$$

其中$\boldsymbol{\phi}_n = \boldsymbol{\phi}(\boldsymbol{x}_n)$。这被称为最小均方（least-mean-squares）或者LMS算法。η的值需要仔细选择，确保算法收敛（Bishop and Nabney, 2008）。

3.1.4 正则化最小平方

在1.1节，我们介绍了为误差函数添加正则化项的思想来控制过拟合，因此需要最小化的总的误差函数的形式为

$$E_D(\boldsymbol{w}) + \lambda E_W(\boldsymbol{w}) \tag{3.24}$$

其中λ是正则化系数，控制数据相关的误差$E_D(\boldsymbol{w})$和正则化项$E_W(\boldsymbol{w})$的相对重要性。正则化项的一个最简单的形式为权向量的各个元素的平方和

$$E_W(\boldsymbol{w}) = \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} \tag{3.25}$$

如果我们考虑平方和误差函数

$$E_D(\boldsymbol{w}) = \frac{1}{2}\sum_{n=1}^{N}\{t_n - \boldsymbol{w}^T\boldsymbol{\phi}(\boldsymbol{x}_n)\}^2 \tag{3.26}$$

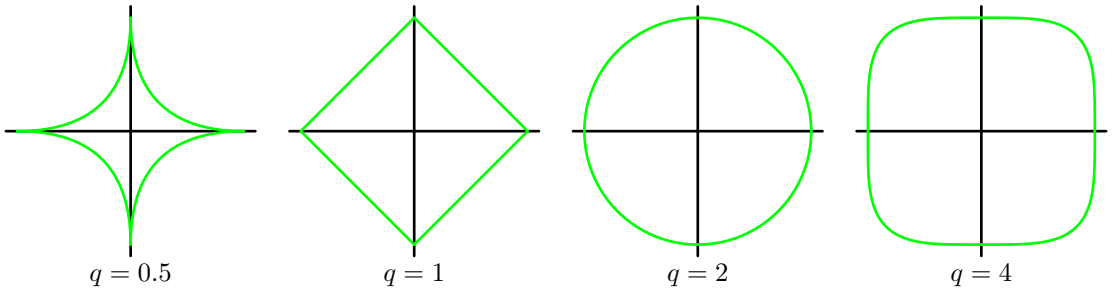

图 3.3: 对于不同的参数q，公式（3.29）中的正则化项的轮廓线。

那么总误差函数就变成了

$$\frac{1}{2}\sum_{n=1}^{N}\{t_n - \boldsymbol{w}^T\boldsymbol{\phi}(\boldsymbol{x}_n)\}^2 + \frac{\lambda}{2}\boldsymbol{w}^T\boldsymbol{w} \tag{3.27}$$

这种对于正则化项的选择方法在机器学习的文献中被称为权值衰减（weight decay）。这是因为在顺序学习算法中，它倾向于让权值向零的方向衰减，除非有数据支持。在统计学中，它提供了一个参数收缩（parameter shrinkage）方法的例子，因为这种方法把参数的值向零的方向收缩。这种方法的优点在于，误差函数是$\boldsymbol{w}$的二次函数，因此精确的最小值具有解析解。具体来说，令公式（3.27）关于$\boldsymbol{w}$的梯度等于零，解出$\boldsymbol{w}$，我们有

$$\boldsymbol{w} = (\lambda\boldsymbol{I} + \boldsymbol{\Phi}^T\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^T\mathbf{t} \tag{3.28}$$

这是最小平方解（3.15）的一个简单的扩展。

有时使用一个更加一般的正则化项，这时正则化的误差函数的形式为

$$\frac{1}{2}\sum_{n=1}^{N}\{t_n - \boldsymbol{w}^T\boldsymbol{\Phi}(\boldsymbol{x}_n)\}^2 + \frac{\lambda}{2}\sum_{j=1}^{M}|w_j|^q \tag{3.29}$$

其中$q = 2$对应于二次正则化项（3.27）。图3.3给出了不同q值下的正则化函数的轮廓线。

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）

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

参数η要选择一个合适的值。这样，这两种方法通过拉格朗日乘数法被联系到了一起。稀疏性的来源可以从图3.4中看出来。图3.4给出了在限制条件（3.30）下误差函数的最小值。随着λ的增大，越来越多的参数会变为零。

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型复杂度的问题上。

对于本章的其余部分，我们将把注意力放在二次正则化项（3.27）上，因为它在实际应用中很重要，并且数学计算上比较容易。

3.1.5 多个输出

目前为止，我们已经考虑了单一目标变量t的情形。在某些应用中，我们可能想预测$K > 1$个目标变量。我们把这些目标变量聚集起来，记作目标向量$\boldsymbol{t}$。这个问题可以这样解决：对于$\boldsymbol{t}$的

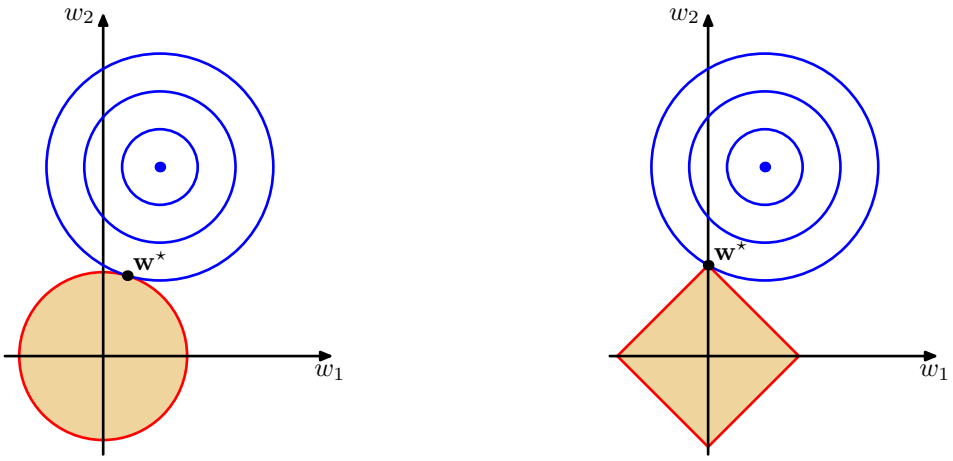

图 3.4: 未正则化的误差函数的轮廓线（蓝色）以及公式（3.30）给出的限制区域。左图是$q = 2$的二次正则化项的限制区域，右图是$q = 1$的套索正则化项的限制区域，其中参数向量$\boldsymbol{w}$的值被记作$\boldsymbol{w}^*$。套索正则化项给出了一个稀疏的解，其中$w_1^* = 0$。

每个分量，引入一个不同的基函数集合，从而变成了多个独立的回归问题。但是，一个更有趣的并且更常用的方法是对目标向量的所有分量使用一组相同的基函数来建模，即

$$\boldsymbol{y}(\boldsymbol{x}, \boldsymbol{w}) = \boldsymbol{W}^T \boldsymbol{\phi}(\boldsymbol{x}) \tag{3.31}$$

其中$\boldsymbol{y}$是一个K维列向量，$\boldsymbol{W}$是一个$M \times K$的参数矩阵，$\boldsymbol{\phi}(\boldsymbol{x})$是一个$M$为列向量，每个元素为$\phi_j(\boldsymbol{x})$，并且与之前一样，$\phi_0(\boldsymbol{x}) = 1$。假设我们令目标向量的条件概率分布是一个各向同性的高斯分布，形式为

$$p(\boldsymbol{t} \mid \boldsymbol{x}, \boldsymbol{W}, \beta) = \mathcal{N}(\boldsymbol{t} \mid \boldsymbol{W}^T \boldsymbol{\phi}(\boldsymbol{x}), \beta^{-1} \boldsymbol{I}) \tag{3.32}$$

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

$$\begin{aligned} \ln p(\boldsymbol{T} \mid \boldsymbol{X}, \boldsymbol{W}, \beta) &= \sum_{n=1}^{N} \ln \mathcal{N}(\boldsymbol{t}_n \mid \boldsymbol{W}^T \boldsymbol{\phi}(\boldsymbol{x}_n), \beta^{-1} \boldsymbol{I}) \\ &= \frac{NK}{2} \ln \left(\frac{\beta}{2\pi} \right) - \frac{\beta}{2} \sum_{n=1}^{N} \| \boldsymbol{t}_n - \boldsymbol{W}^T \boldsymbol{\phi}(\boldsymbol{x}_n) \|^2 \end{aligned} \tag{3.33}$$

与之前一样，我们可以关于$\boldsymbol{W}$最大化这个函数，可得

$$\boldsymbol{W}_{ML} = (\boldsymbol{\Phi}^T \boldsymbol{\Phi})^{-1} \boldsymbol{\Phi}^T \boldsymbol{T} \tag{3.34}$$

如果我们对于每个目标变量t_k考察这个结果，那么我们有

$$\boldsymbol{w}_k = (\boldsymbol{\Phi}^T \boldsymbol{\Phi})^{-1} \boldsymbol{\Phi}^T \mathbf{t}_k = \boldsymbol{\Phi}^\dagger \mathbf{t}_k \tag{3.35}$$

这里，$\mathbf{t}_k$是一个N维列向量，元素为t_{nk}其中$n = 1, \ldots, N$。因此不同目标变量的回归问题在这里被分解开，并且我们只需要计算一个伪逆矩阵$\boldsymbol{\Phi}^\dagger$，这个矩阵是被所有向量$\boldsymbol{w}_k$所共享的。

推广到具有任意协方差矩阵的一般的高斯噪声分布是很直接的。与之前一样，这个问题可以被分解为K个独立的回归问题。这种结果毫不令人惊讶，因为参数$\boldsymbol{W}$只定义了高斯噪声分布的均值，并且我们从2.3.4节中知道多元高斯分布均值的最大似然解与协方差无关。从现在开始，为了简单起见，我们值考虑单一目标变量t的情形。

3.2 偏置-方差分解

目前为止，我们对于回归的线性模型的讨论中，我们假定了基函数的形式和数量都是固定的。正如我们在第1章中看到的那样，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，使用最小平方方法，会导致严重的过拟合问题。然而，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。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，但是这就产生了一个问题：如何确定正则化系数λ的合适的值。同时关于权值$\boldsymbol{w}$和正则化系数λ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。

正如我们在之前的章节中看到的那样，过拟合现象确实是最大似然方法的一个不好的性质。但是当我们在使用贝叶斯方法对参数进行求和或者积分时，过拟合现象不会出现。本章中，我们会稍微深入地从贝叶斯观点讨论模型的复杂度。但是，在进行这样的讨论之前，从频率学家的观点考虑一下模型的复杂度问题是很有指导意义的。这种频率学家的观点被称为偏置-方差折中（bias-variance trade-off）。虽然我们将在线性基函数模型中介绍这个概念，因为这样介绍可以使用简单的例子来说明一些基本的思想，但是实际上这种讨论有着更加普遍的适用性。

在1.5.5节，当我们讨论回归问题的决策论时，我们考虑了不同的损失函数。一旦我们知道了条件概率分布$p(t \mid \boldsymbol{x})$，每一种损失函数都能够给出对应的最优预测结果。使用最多的一个选择是平方损失函数，此时最优的预测由条件期望（记作$h(\boldsymbol{x})$）给出，即

$$h(\boldsymbol{x}) = \mathbb{E}[t \mid \boldsymbol{x}] = \int t p(t \mid \boldsymbol{x}) \, \mathrm{d}t \tag{3.36}$$

现在，有必要区分决策论中出现的平方损失函数以及模型参数的最大似然估计中出现的平方和误差函数。我们可以使用比最小平方更复杂的方法，例如正则化或者纯粹的贝叶斯方法，来确定条件概率分布$p(t \mid \boldsymbol{x})$。为了进行预测，这些方法都可以与平方损失函数相结合。

我们在1.5.5节证明了平方损失函数的期望可以写成

$$\mathbb{E}[L] = \int \{y(\boldsymbol{x}) - h(\boldsymbol{x})\}^2 p(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x} + \iint \{h(\boldsymbol{x}) - t\}^2 p(\boldsymbol{x}, t) \, \mathrm{d}\boldsymbol{x} \, \mathrm{d}t \tag{3.37}$$

回忆一下，与$y(\boldsymbol{x})$无关的第二项，是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。第一项与我们对函数$y(\boldsymbol{x})$的选择有关，我们要找一个$y(\boldsymbol{x})$的解，使得这一项最小。由于它是非负的，因此我们希望能够让这一项的最小值等于零。如果我们有无限多的数据（以及无限多的计算资源），那么原则上我们能够以任意的精度寻找回归函数$h(\boldsymbol{x})$，这会给出$y(\boldsymbol{x})$的最优解。然而，在实际应用中，我们的数据集$\mathcal{D}$只有有限的N个数据点，从而我们不能够精确地知道回归函数$h(\boldsymbol{x})$。

如果我们使用由参数向量$\boldsymbol{w}$控制的函数$y(\boldsymbol{x}, \boldsymbol{w})$对$h(\boldsymbol{x})$建模，那么从贝叶斯的观点来看，我们模型的不确定性是通过$\boldsymbol{w}$的后验概率分布来表示的。但是，频率学家的方法涉及到根据数据集$\mathcal{D}$对$\boldsymbol{w}$进行点估计，然后试着通过下面的思想实验来表示估计的不确定性。假设我们有许多数据集，每个数据集的大小为N，并且每个数据集都独立地从分布$p(t, \boldsymbol{x})$中抽取。对于任意给定的数据集$\mathcal{D}$，我们可以运行我们的学习算法，得到一个预测函数$y(\boldsymbol{x}; \mathcal{D})$。不同的数据集会给出不同的函数，从而给出不同的平方损失的值。这样，特定的学习算法的表现就可以通过取各个数据集上的表现的平均值来进行评估。

考虑公式（3.37）的第一项的被积函数，对于一个特定的数据集$\mathcal{D}$，它的形式为

$$\{y(\boldsymbol{x}; \mathcal{D}) - h(\boldsymbol{x})\}^2 \tag{3.38}$$

由于这个量与特定的数据集$\mathcal{D}$相关，因此我们对所有的数据集取平均。如果我们在括号内加上然后减去$\mathbb{E}_{\mathcal{D}}[y(\boldsymbol{x}; \mathcal{D})]$，然后展开，我们有

$$\begin{aligned}
&\{y(\boldsymbol{x};\mathcal{D}) - \mathbb{E}_{\mathcal{D}}[y(\boldsymbol{x}; \mathcal{D})] + \mathbb{E}_{\mathcal{D}}[y(\boldsymbol{x}; \mathcal{D})] - h(\boldsymbol{x})\}^2 \\
&\quad = \{y(\boldsymbol{x}; \mathcal{D}) - \mathbb{E}_{\mathcal{D}}[y(\boldsymbol{x}; \mathcal{D})]\}^2 + \{\mathbb{E}_{\mathcal{D}}[y(\boldsymbol{x}; \mathcal{D})] - h(\boldsymbol{x})\}^2 \\
&\qquad + 2\{y(\boldsymbol{x}; \mathcal{D}) - \mathbb{E}_{\mathcal{D}}[y(\boldsymbol{x}; \mathcal{D})]\}\{\mathbb{E}_{\mathcal{D}}[y(\boldsymbol{x}; \mathcal{D})] - h(\boldsymbol{x})\}
\end{aligned} \tag{3.39}$$

我们现在关于$\mathcal{D}$求期望，然后注意到最后一项等于零，可得

$$\begin{aligned}&\mathbb{E}_{\mathcal{D}}[\{y(\boldsymbol{x};\mathcal{D})-h(\boldsymbol{x})\}^2]\\&=\underbrace{\{\mathbb{E}_{\mathcal{D}}[y(\boldsymbol{x};\mathcal{D})]-h(\boldsymbol{x})\}^2}_{(\text{偏置})^2}+\underbrace{\mathbb{E}_{\mathcal{D}}[\{y(\boldsymbol{x};\mathcal{D})-\mathbb{E}_{\mathcal{D}}[y(\boldsymbol{x};\mathcal{D})]\}^2]}_{\text{方差}}\end{aligned}\tag{3.40}$$

我们看到，$y(\boldsymbol{x};\mathcal{D})$与回归函数$h(\boldsymbol{x})$的差的平方的期望可以表示为两项的和。第一项，被称为平方偏置（bias），表示所有数据集的平均预测与预期的回归函数之间的差异。第二项，被称为方差（variance），度量了对于单独的数据集，模型所给出的解在平均值附近波动的情况，因此也就度量了函数$y(\boldsymbol{x};\mathcal{D})$对于特定的数据集的选择的敏感程度。稍后我们会考虑一个简单的例子，来直观地说明这些概念。

目前为止，我们已经考虑了单一输入变量$\boldsymbol{x}$的情形。如果我们把这个展开式带回到公式（3.37）中，那么我们就得到了下面的对于期望平方损失的分解

$$\text{期望损失}=\text{偏置}^2+\text{方差}+\text{噪声}\tag{3.41}$$

其中

$$\text{偏置}^2=\int\{\mathbb{E}_{\mathcal{D}}[y(\boldsymbol{x};\mathcal{D})]-h(\boldsymbol{x})\}^2p(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}\tag{3.42}$$

$$\text{方差}=\int\mathbb{E}_{\mathcal{D}}[\{y(\boldsymbol{x};\mathcal{D})-\mathbb{E}_{\mathcal{D}}[y(\boldsymbol{x};\mathcal{D})]\}^2]p(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}\tag{3.43}$$

$$\text{噪声}=\iint\{h(\boldsymbol{x})-t\}^2p(\boldsymbol{x},t)\,\mathrm{d}\boldsymbol{x}\,\mathrm{d}t\tag{3.44}$$

现在，偏置和方差指的是积分后的量。

我们的目标是最小化期望损失，它可以分解为（平方）偏置、方差和一个常数噪声项的和。正如我们将看到的那样，在偏置和方差之间有一个折中。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。有着最优预测能力的模型时在偏置和方差之间取得最优的平衡的模型。这里通过第1章讨论过的正弦数据集来说明。我们产生了100个数据集合，每个集合都包含$N=25$个数据点，都是独立地从正弦曲线$h(x)=\sin(2\pi x)$抽取的。数据集的编号为$l=1,\ldots,L$，其中$L=100$，并且对于每个数据集$\mathcal{D}^{(l)}$，我们通过最小化正则化的误差函数（3.27）拟合了一个带有24个高斯基函数的模型，然后给出了预测函数$y^{(l)}(x)$，如图3.5所示。第一行对应着较大的正则化系数λ，这样的模型的方差很小（因为左侧图中的红色曲线看起来很相似），但是偏置很大（因为右侧图中的两条曲线看起来相当不同）。相反，在最后一行，正则化系数λ很小，这样模型的方差较大（因为左侧图中的红色曲线变化性相当大），但是偏置很小（因为平均拟合的结果与原始正弦曲线十分吻合）。注意，把$M=25$这种复杂模型的多个解进行平均，会产生对于回归函数非常好的拟合，这表明求平均是一个很好的步骤。事实上，将多个解加权平均是贝叶斯方法的核心，虽然这种求平均针对的是参数的后验分布，而不是针对多个数据集。

对于这个例子，我们也可以定量地考察偏置-方差折中。平均预测由下式求出

$$\bar{y}(x)=\frac{1}{L}\sum_{l=1}^{L}y^{(l)}(x)\tag{3.45}$$

并且积分后的平方偏置以及积分后的方差为

$$\text{偏置}^2=\frac{1}{N}\sum_{n=1}^{N}\{\bar{y}(x_n)-h(x_n)\}^2\tag{3.46}$$

$$\text{方差}=\frac{1}{N}\sum_{n=1}^{N}\frac{1}{L}\sum_{l=1}^{L}\{y^{(l)}(x_n)-\bar{y}(x_n)\}^2\tag{3.47}$$

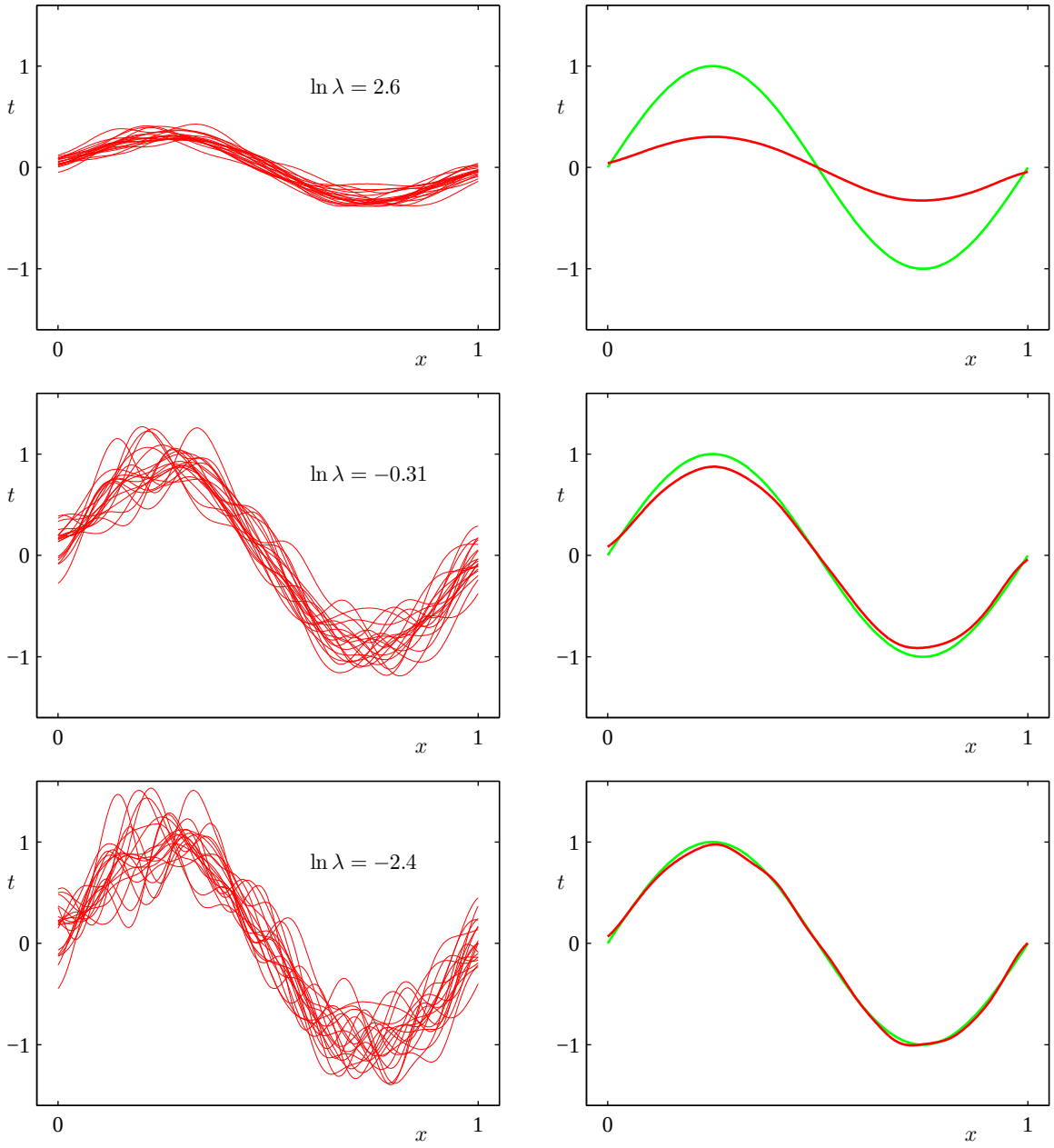

图 3.5: 模型复杂度对于偏置和方差的依赖的说明。模型的复杂度由正则化参数λ控制，数据集是第1章中的正弦数据。有$L = 100$个数据集，每个数据集有$N = 25$个数据点，每个模型有24个高斯基函数，从而参数的总数为$M = 25$（包括偏置参数）。左侧一列给出了对于不同的$\ln \lambda$值，根据数据集拟合模型的结果。为了清晰起见，我们只给出了100个拟合模型中的20个。右侧一列给出了对应的100个拟合的均值（红色）以及用于生成数据集的正弦函数（绿色）。

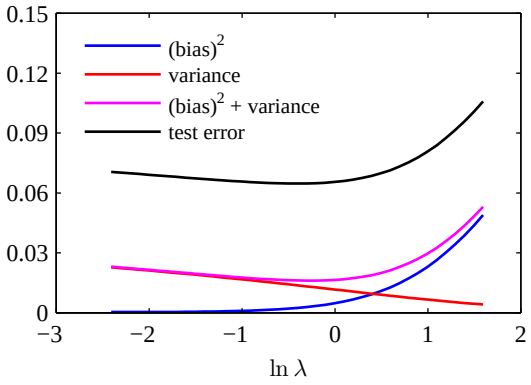

图 3.6: 平方偏置和方差的图像，以及它们的加和，对应于图3.5给出的结果。同样给出的还有大小为1000个数据点的测试数据的平均测试误差。(偏置)^2 + 方差的最小值出现在$\ln \lambda = -0.31$的位置，它接近于在测试数据上取得最小误差的位置。

其中由概率分布$p(x)$加权的x的积分由来自那个概率分布的有限数据点的加和来近似。图3.6给出了这些量以及它们的求和关于$\ln \lambda$的函数图像。我们看到，小的λ使得模型对于各个数据集里的噪声的拟合效果非常好，导致了较大的方差。相反，大的λ把权值参数拉向零，导致了较大的偏置。

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。

由于有这么多局限性，因此我们在下一节里将讨论线性基函数模型的贝叶斯观点。它不仅提供了对于过拟合现象的深刻认识，还提出了解决模型复杂度问题的实用的技术。

3.3 贝叶斯线性回归

在我们讨论使用最大似然方法设置线性回归模型的参数时，我们已经看到由基函数的数量控制的模型的复杂度需要根据数据集的规模进行调整。为对数似然函数增加一个正则化项意味着模型的复杂度可以通过正则化系数的值进行控制，虽然基函数的数量和形式的选择仍然对于确定模型的整体行为十分重要。

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。

3.3.1 参数分布

关于线性拟合的贝叶斯方法的讨论，我们首先引入模型参数$\boldsymbol{w}$的先验概率分布。现在这个阶段，我们把噪声精度参数β当做已知常数。首先，我们注意到，由公式（3.10）定义的似然函数$p(\mathbf{t} \mid \boldsymbol{w})$是$\boldsymbol{w}$的二次函数的指数形式。于是对应的共轭先验是高斯分布，形式为

$$p(\boldsymbol{w}) = \mathcal{N}(\boldsymbol{w} \mid \boldsymbol{m}_0, \boldsymbol{S}_0) \tag{3.48}$$

均值为$\boldsymbol{m}_0$，协方差为$\boldsymbol{S}_0$。

接下来我们计算后验分布，它正比于似然函数与先验分布的乘积。由于共轭高斯先验分布的选择，后验分布也将是高斯分布。我们可以对指数项进行配平方，然后使用归一化的高斯分

布的标准结果找到归一化系数，这样就计算出了后验分布的形式。但是，我们在推导公式（2.116）已经进行了必要的工作，这让我们能够直接写出后验概率分布的形式

$$p(\boldsymbol{w} \mid \mathbf{t}) = \mathcal{N}(\boldsymbol{w} \mid \boldsymbol{m}_N, \boldsymbol{S}_N) \quad \text{后验分布，} \tag{3.49}$$

其中

$$\boldsymbol{m}_N = \boldsymbol{S}_N(\boldsymbol{S}_0^{-1}\boldsymbol{m}_0 + \beta\boldsymbol{\Phi}^T\mathbf{t}) \tag{3.50}$$

$$\boldsymbol{S}_N^{-1} = \boldsymbol{S}_0^{-1} + \beta\boldsymbol{\Phi}^T\boldsymbol{\Phi} \tag{3.51}$$

注意，由于后验分布是高斯分布，它的众数恰好与它的均值相同。因此最大后验权向量的结果就是$\boldsymbol{w}_{MAP} = \boldsymbol{m}_N$。如果我们考虑一个无限宽的先验$\boldsymbol{S}_0 = \alpha^{-1}\boldsymbol{I}$，其中$\alpha \to 0$，那么后验概率分布的均值$\boldsymbol{m}_N$就变成了由公式（3.15）给出的最大似然值$\boldsymbol{w}_{ML}$。类似地，如果$N = 0$，那么后验概率分布就变成了先验分布。此外，如果数据点是顺序到达的，那么任何一个阶段的后验概率分布都可以看成后续数据点的先验。此时新的后验分布再次由公式（3.49）给出。

对于本章的剩余部分，为了简化起见，我们将考虑高斯先验的一个特定的形式。具体来说，我们考虑零均值各向同性高斯分布。这个分布由一个精度参数α控制，即

$$p(\boldsymbol{w} \mid \alpha) = \mathcal{N}(\boldsymbol{w} \mid \mathbf{0}, \alpha^{-1}\boldsymbol{I}) \tag{3.52}$$

对应的$\boldsymbol{w}$的后验概率分布由公式（3.49）给出，其中

$$\boldsymbol{m}_N = \beta\boldsymbol{S}_N\boldsymbol{\Phi}^T\mathbf{t} \tag{3.53}$$

$$\boldsymbol{S}_N^{-1} = \alpha\boldsymbol{I} + \beta\boldsymbol{\Phi}^T\boldsymbol{\Phi} \tag{3.54}$$

后验概率分布的对数由对数似然函数与先验的对数求和的方式得到。它是$\boldsymbol{w}$的函数，形式为

$$\ln p(\boldsymbol{w} \mid \mathbf{t}) = -\frac{\beta}{2}\sum_{n=1}^{N}\{t_n - \boldsymbol{w}^T\boldsymbol{\phi}(\boldsymbol{x}_n)\}^2 - \frac{\alpha}{2}\boldsymbol{w}^T\boldsymbol{w} + \text{常数} \tag{3.55}$$

于是，后验分布关于$\boldsymbol{w}$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。正则项对应于公式（3.27），其中$\lambda = \frac{\alpha}{\beta}$。

我们可以使用直线拟合的简单的例子来说明线性基函数的贝叶斯学习过程，以及后验概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$y(x, \boldsymbol{w}) = w_0 + w_1 x$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$f(x, \boldsymbol{a}) = a_0 + a_1 x$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$\mathrm{U}(x \mid -1, 1)$中选择$x_n$的值，然后计算$f(x_n, \boldsymbol{a})$，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$\alpha$固定为2.0。我们稍后会简短地讨论从训练数据中确定α和β的值的策略。图3.7给出了当数据集的规模增加时贝叶斯学习的结果，还展示了贝叶斯学习的顺序本质，即当新数据点被观测到的时候，当前的后验分布变成了先验分布。花时间仔细研究一下这幅图是很值得的，因为它说明了贝叶斯推断的一些重要的概念。这张图的第一行对应于观测到任何数据点之前的情况，给出了$\boldsymbol{w}$空间的先验概率分布的图像，以及函数$y(x, \boldsymbol{w})$的六个样本，这六个样本的$\boldsymbol{w}$都是从先验概率分布中抽取的。在第二行，我们看到了观测到一个数据点之后的情形。数据点的位置(x, t)由右侧一列中的蓝色圆圈表示。左侧一列是对于这个数据点的似然函数$p(t, \boldsymbol{w})$关于$\boldsymbol{w}$的函数图像。注意，似然函数提供了一个温和的限制，即直线必须穿过数据点附近的位置，其中附近位置的范围由噪声精度β确定。为了进行对比，用来生成数据集的真实参数值$a_0 = -0.3$以及$a_1 = 0.5$在图3.7的左侧一列被标记为白色十字。如果我们把这个似然函数与第一行的先验概率相乘，然后归一化，我们就得到了第二行中间的图给出的后验概率分布。从这个后验概率分布中抽取$\boldsymbol{w}$的样本，对应的回归函数$y(x, \boldsymbol{w})$被画在了右侧一列的途中。注意，这些样本直线全部穿过数据点的附近位置。这张图的第三行展示了观测到第二个数据点的效果。与之前一样，这个数据点由右侧一列的蓝色圆圈表示。第二个数据点自身对应的似然函数在左侧一列的图中给出。如果我们

把这个似然函数与第二行的后验概率分布相乘，我们就得到了第三行中间一列的图给出的后验概率分布。注意，这个后验概率分布与我们将原始的先验分布结合两个数据点的似然函数得到的后验概率分布完全相同。现在，后验概率分布被两个数据点影响。由于两个点足够定义一条直线，因此目前已经得到了相对较好的后验概率分布。从这个后验分布中抽取的样本产生了第三列中红色的函数，我们看到这些函数同时穿过了两个数据点的附近。第四行展示了观测到20个数据点的效果。左侧的图展示了第20个数据点自身的似然函数，中间的图展示了融合了20次观测信息的后验概率分布。注意与第三行相比，这个后验概率分布变得更加尖锐。在无穷多个数据点的极限情况下，后验概率分布会变成一个Delta函数。这个函数的中心是用白色十字标记出的真实参数值。

也可以考虑参数的其他形式的先验分布。例如，我们可以推广高斯先验分布，得到

$$p(\boldsymbol{w} \mid \alpha) = \left[\frac{q}{2}\left(\frac{\alpha}{2}\right)^{\frac{1}{q}} \frac{1}{\Gamma(\frac{1}{q})}\right]^{M} \exp\left(-\frac{\alpha}{2}\sum_{j=0}^{M-1} |w_j|^q\right) \tag{3.56}$$

其中$q = 2$的情形对应于高斯分布，并且只有在这种情形下的先验分布才是公式（3.10）给出的似然函数的共轭先验。找到$\boldsymbol{w}$的后验概率分布的最大值对应于找到正则化误差函数（3.29）的最小值。在高斯先验的情况下，后验概率分布的众数等于均值，但是如果$q \neq 2$，这个性质就不成立了。

3.3.2 预测分布

在实际应用中，我们通常感兴趣的不是$\boldsymbol{w}$本身的值，而是对于新的$\boldsymbol{x}$值预测出t的值。这需要我们计算出预测分布（predictive distribution），定义为

$$p(t \mid \mathbf{t}, \alpha, \beta) = \int p(t \mid \boldsymbol{w}, \beta) p(\boldsymbol{w} \mid \mathbf{t}, \alpha, \beta)\ \mathrm{d}\boldsymbol{w} \tag{3.57}$$

其中$\mathbf{t}$是训练数据的目标变量的值组成的向量。并且，为了简化记号，我们在右侧省略了条件概率中出现的输入向量。目标变量的条件概率分布$p(t \mid \boldsymbol{w}, \boldsymbol{w}, \beta)$由公式（3.8）给出，后验分布由公式（3.49）给出。我们看到公式（3.57）涉及到两个高斯分布的卷积，因此使用2.3.3节的公式（2.115）的结果，我们看到预测分布的形式为

$$p(t \mid \boldsymbol{x}, \mathbf{t}, \alpha, \beta) = \mathcal{N}(t \mid \boldsymbol{m}_N^T \boldsymbol{\phi}(\boldsymbol{x}), \sigma_N^2(\boldsymbol{x})) \tag{3.58}$$

其中预测分布的方差$\sigma_N^2(\boldsymbol{x})$为

$$\sigma_N^2(\boldsymbol{x}) = \frac{1}{\beta} + \boldsymbol{\phi}(\boldsymbol{x})^T \boldsymbol{S}_N \boldsymbol{\phi}(\boldsymbol{x}) \tag{3.59}$$

公式（3.59）的第一项表示数据中的噪声，而第二项反映了与参数$\boldsymbol{w}$关联的不确定性。由于噪声和$\boldsymbol{w}$的分布是相互独立的高斯分布，因此它们的值是可以相加的。注意，当额外的数据点被观测到的时候，后验概率分布会变窄。从而可以证明出$\sigma_{N+1}^2(\boldsymbol{x}) \leq \sigma_N^2(\boldsymbol{x})$（Qazaz et al., 1997）。在极限$N \to \infty$的情况下，公式（3.59）的第二项趋于零，从而预测分布的方差只与参数β控制的具有可加性的噪声有关。

为了说明贝叶斯线性回归模型的预测分布，让我们回到第1.1节人工生成的正弦数据集。在图3.8中，我们调整一个由高斯基函数线性组合的模型，使其适应于不同规模的数据集，然后观察对应的后验概率分布。这里，绿色曲线对应着产生数据点的函数$\sin(2\pi x)$（带有附加的高斯噪声）。大小为$N = 1, N = 2, N = 4$和$N = 25$的数据集在四幅图中用蓝色圆圈表示。对于每幅图，红色曲线是对应的高斯预测分布的均值，红色阴影区域是均值两侧的一个标准差范围的区域。注意，预测的不确定性依赖于x，并且在数据点的邻域内最小。还要注意，不确定性的程度随着观测到的数据点的增多而逐渐减小。

图3.8中的图像只给出了每个点处的预测方差与x的函数关系。为了更加深刻地认识对于不同的x值的预测之间的协方差，我们可以从$\boldsymbol{w}$的后验概率分布中抽取样本，然后画出对应的函数$y(x, \boldsymbol{w})$，如图3.9所示。

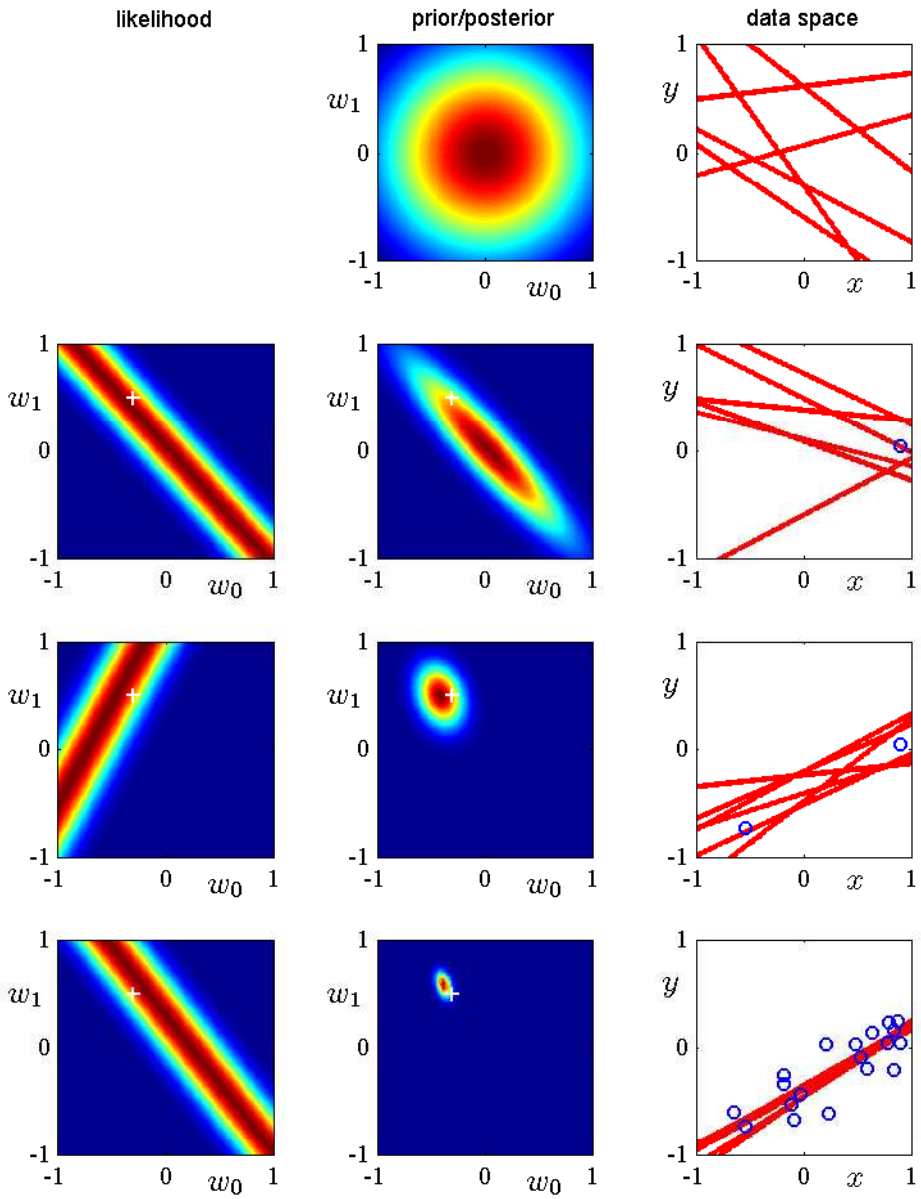

图 3.7: 顺序贝叶斯学习的例子。模型是一个简单的线性模型，形式为$y(x, \boldsymbol{w}) = w_0 + w_1 x$。本图的详细描述见正文。

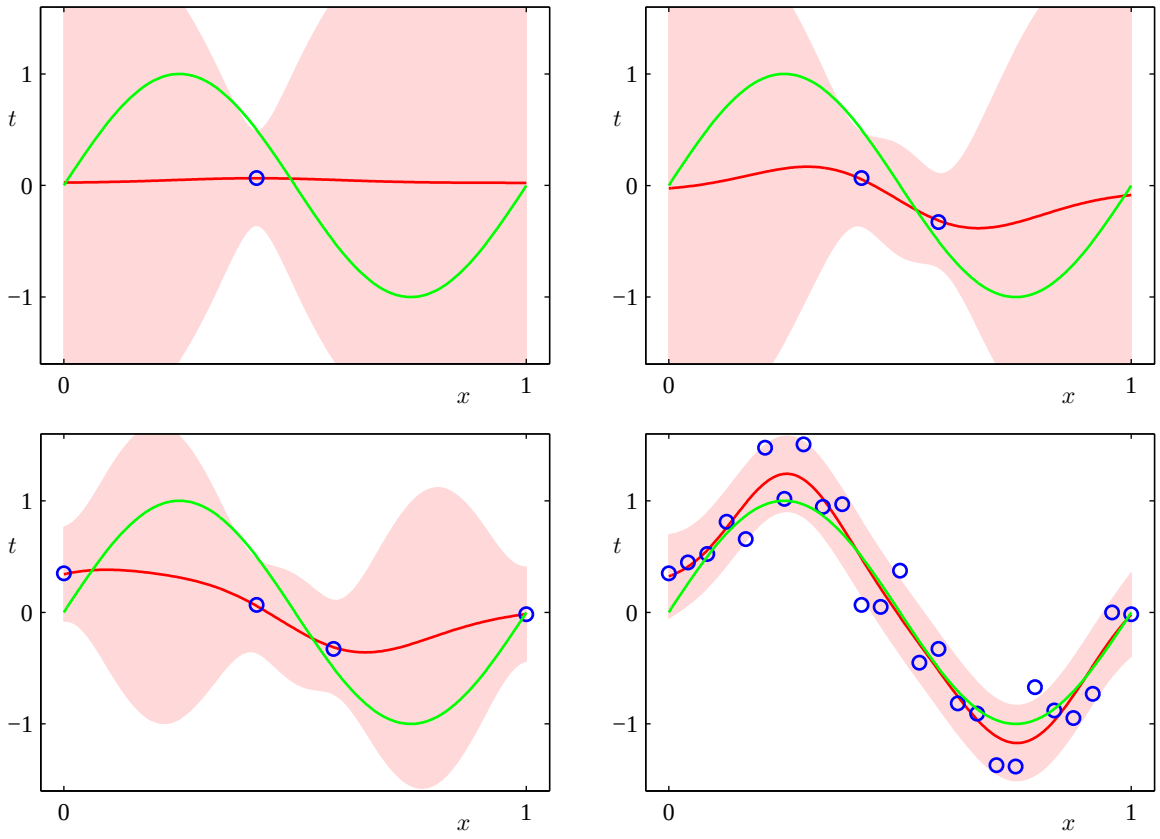

图 3.8: 包含9个高斯基函数（3.4）的模型的预测分布（3.58），使用了1.1节的人工生成的正弦数据集。详细的讨论见正文。

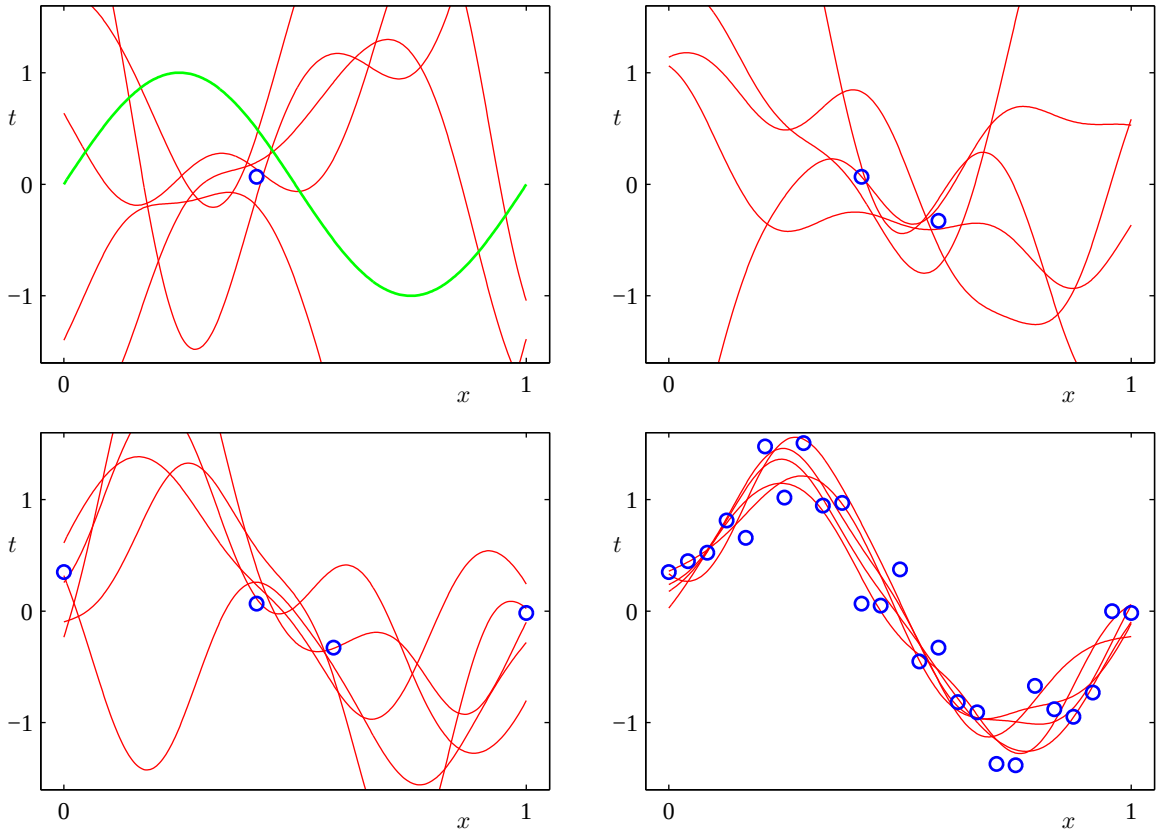

图 3.9: 函数$y(x, \boldsymbol{w})$的图像，使用了服从$\boldsymbol{w}$上的后验概率分布的样本，对应于图3.8。

如果我们使用局部的基函数（例如高斯基函数），那么在距离基函数中心比较远的区域，公式（3.59）给出的预测方差的第二项的贡献将会趋于零，只剩下噪声的贡献β^{-1}。因此，当对基函数所在的区域之外的区域进行外插的时候，模型对于它做出的预测会变得相当确定，这通常不是我们想要的结果。通过使用被称为高斯过程的另一种贝叶斯回归方法，这个问题可以被避免。

注意，如果$\boldsymbol{w}$和β都被当成未知的，那么根据2.3.6节的讨论，我们可以引入一个由高斯-Gamma分布定义的共轭先验分布$p(\boldsymbol{w}, \beta)$（Denison et al., 2002）。在这种情况下，预测分布是一个学生t分布。

3.3.3 等价核

公式（3.53）给出的线性基函数模型的后验均值解有一个有趣的解释，这个解释为核方法（包括高斯过程）提供了舞台。如果我们把公式（3.53）代入表达式（3.3），我们看到预测均值可以写成下面的形式

$$y(\boldsymbol{x}, \boldsymbol{m}_N) = \boldsymbol{m}_N^T \boldsymbol{\phi}(\boldsymbol{x}) = \beta \boldsymbol{\phi}(\boldsymbol{x})^T \boldsymbol{S}_N \boldsymbol{\Phi}^T \mathbf{t} = \sum_{n=1}^{N} \beta \boldsymbol{\phi}(\boldsymbol{x})^T \boldsymbol{S}_N \boldsymbol{\phi}(\boldsymbol{x}_n) t_n \tag{3.60}$$

其中$\boldsymbol{S}_N$由公式（3.51）定义。因此在点$\boldsymbol{x}$处的预测均值由训练集目标变量t_n的线性组合给出，即

$$y(\boldsymbol{x}, \boldsymbol{m}_N) = \sum_{n=1}^{N} k(\boldsymbol{x}, \boldsymbol{x}_n) t_n \tag{3.61}$$

其中，函数

$$k(\boldsymbol{x}, \boldsymbol{x}') = \beta \boldsymbol{\phi}(\boldsymbol{x})^T \boldsymbol{S}_N \boldsymbol{\phi}(\boldsymbol{x}') \tag{3.62}$$

合，然后对新的$\boldsymbol{x}$值做预测。可以证明这些权值的和等于1，即

$$\sum_{n=1}^{N} k(\boldsymbol{x}, \boldsymbol{x}_n) = 1 \tag{3.64}$$

对于所有的$\boldsymbol{x}$值都成立。这个直观上令人兴奋的结果可以很容易地用非形式化的方式证明出来。我们注意到，这个加和等价于对于所有的n都有$t_n = 1$的目标数据集的预测均值$\widehat{y}(\boldsymbol{x})$。假设基函数是线性独立的，且数据点的数量多于基函数的数量，并且其中一个基函数是常量（对应于偏置参数），那么很明显我们可以精确地拟合训练数据，因此预测均值就是简单的$\widehat{y}(\boldsymbol{x}) = 1$，这样我们就可以得到共识（3.64）。注意，核函数可以为负也可以为正，因此它虽然满足加和限制，但是对应的预测未必是训练集的目标值的凸组合。

最后，我们注意到，公式（3.62）给出的等价核满足一般的核函数共有的一个重要性质，即它可以表示为非线性函数的向量$\boldsymbol{\psi}(\boldsymbol{x})$的内积的形式，即

$$k(\boldsymbol{x}, \boldsymbol{z}) = \boldsymbol{\psi}(\boldsymbol{x})^T \boldsymbol{\psi}(\boldsymbol{z}) \tag{3.65}$$

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

3.4 贝叶斯模型比较

在第1章中，我们强调了过拟合的问题，也介绍了通过使用交叉验证的方法，来设置正则化参数的值，或者从多个模型中选择合适的一个。这里，我们从贝叶斯的角度考虑模型选择的问题。在本节中，我们的讨论是非常一般的。之后在3.5节，我们将会看到这些想法是如何应用到线性回归的正则化参数确定的问题中的。

正如我们将看到的那样，与最大似然估计相关联的过拟合问题可以通过对模型的参数进行求和或者积分的方式（而不是进行点估计）来避免。这样，模型可以直接在训练数据上进行比较，而不需要验证集。这使得所有的数据都能够被用于训练，并且避免了交叉验证当中每个模型要运行多次训练过程的问题。它也让多个复杂度参数可以同时在训练过程中被确定。例如，在第7章，我们会介绍相关向量机（relevance vector machine），这是一个贝叶斯模型，它对于每个训练数据点都有一个复杂度参数。

模型比较的贝叶斯观点仅仅涉及到使用概率来表示模型选择的不确定性，以及恰当地使用概率的加和规则和乘积规则。假设我们想比较L个模型$\{\mathcal{M}_i\}$，其中$i = 1, \ldots, L$。这里，一个模型指的是观测数据$\mathcal{D}$上的概率分布。在多项式曲线拟合的问题中，概率分布被定义在目标值$\mathbf{t}$上，而输入值$\boldsymbol{X}$被假定为已知的。其他类型的模型定义了$\boldsymbol{X}$和$\mathbf{t}$上的联合分布。我们会假设数据是由这些模型中的一个生成的，但是我们不知道究竟是哪一个。我们的不确定性通过先验概率分布$p(\mathcal{M}_i)$表示。给定一个训练数据集$\mathcal{D}$，我们想估计后验分布

$$p(\mathcal{M}_i \mid \mathcal{D}) \propto p(\mathcal{M}_i) p(\mathcal{D} \mid \mathcal{M}_i) \tag{3.66}$$

先验分布让我们能够表达不同模型之间的优先级。让我们简单地假设所有的模型都有相同的先验概率。比较有意思的一项是模型证据（model evidence）$p(\mathcal{D} \mid \mathcal{M}_i)$，它表达了数据展现出的不同模型的优先级，我们稍后会稍微详细地考察这一项。模型证据有时也被称为边缘似然（marginal likelihood），因为它可以被看做在模型空间中的似然函数，在这个空间中参数已经被求和或者积分。两个模型的模型证据的比值$\frac{p(\mathcal{D}|\mathcal{M}_i)}{p(\mathcal{D}|\mathcal{M}_j)}$被称为贝叶斯因子（Bayes factor）（Kass and Raftery, 1995）。

一旦我们知道了模型上的后验概率分布，那么根据概率的加和规则与乘积规则，预测分布为

$$p(t \mid \boldsymbol{x}, \mathcal{D}) = \sum_{i=1}^{L} p(t \mid \boldsymbol{x}, \mathcal{M}_i, \mathcal{D}) p(\mathcal{M}_i \mid \mathcal{D}) \tag{3.67}$$

这是混合分布（mixture distribution）的一个例子。这个公式中，整体的预测分布由下面的方式获得：对各个模型的预测分布$p(t \mid \boldsymbol{x}, \mathcal{M}_i, \mathcal{D})$求加权平均，权值为这些模型的后验概

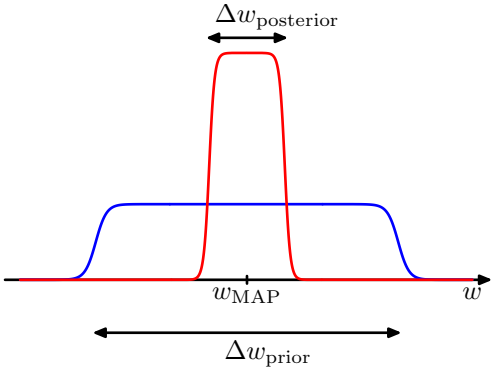

图 3.12: 我们可以粗略地近似模型证据，如果我们假设参数上的后验概率分布在众数w_{MAP}附近有一个尖峰。

率$p(\mathcal{M}_i \mid \mathcal{D})$。例如，如果我们有两个模型，这两个模型的后验概率相等。一个模型预测了$t = a$附近的一个很窄的分布，而另一个模型预测了$t = b$附近的一个很窄的分布，这样整体的预测分布是一个双峰的概率分布，峰值位于$t = a$和$t = b$处，而不是在$t = \frac{a+b}{2}$处的一个单一的模型。

对于模型求平均的一个简单的近似是使用最可能的一个模型自己做预测。这被称为模型选择（model selection）。

对于一个由参数$\boldsymbol{w}$控制的模型，根据概率的加和规则和乘积规则，模型证据为

$$p(\mathcal{D} \mid \mathcal{M}_i) = \int p(\mathcal{D} \mid \boldsymbol{w}, \mathcal{M}_i) p(\boldsymbol{w} \mid \mathcal{M}_i) \, \mathrm{d}\boldsymbol{w} \tag{3.68}$$

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。还有一件有趣的事情是，我们注意到模型证据恰好就是在估计参数的后验分布时出现在贝叶斯定理的分母中的归一化项，因为

$$p(\boldsymbol{w} \mid \mathcal{D}, \mathcal{M}_i) = \frac{p(\mathcal{D} \mid \boldsymbol{w}, \mathcal{M}_i) p(\boldsymbol{w} \mid \mathcal{M}_i)}{p(\mathcal{D} \mid \mathcal{M}_i)} \tag{3.69}$$

通过对参数的积分进行一个简单的近似，我们可以更加深刻地认识模型证据。首先考虑模型有一个参数w的情形。这个参数的后验概率正比于$p(\mathcal{D} \mid w)p(w)$，其中为了简化记号，我们省略了它对于模型$\mathcal{M}_i$的依赖。如果我们假设后验分布在最大似然值$w_{MAP}$附近是一个尖峰，宽度为$\Delta w_{\text{后验}}$，那么我们可以用被积函数的值乘以尖峰的宽度来近似这个积分。如果我们进一步假设先验分布是平的，宽度为$\Delta w_{\text{先验}}$，即$p(w) = \frac{1}{\Delta w_{\text{先验}}}$，那么我们有

$$p(\mathcal{D}) = \int p(\mathcal{D} \mid w) p(w) \, \mathrm{d}w \simeq p(\mathcal{D} \mid w_{MAP}) \frac{\Delta w_{\text{后验}}}{\Delta w_{\text{先验}}} \tag{3.70}$$

取对数可得

$$\ln p(\mathcal{D}) \simeq \ln p(\mathcal{D} \mid w_{MAP}) + \ln \left(\frac{\Delta w_{\text{后验}}}{\Delta w_{\text{先验}}} \right) \tag{3.71}$$

图3.12说明了这个近似。第一项表示拟合由最可能参数给出的数据。对于平的先验分布来说，这对应于对数似然。第二项用于根据模型的复杂度来惩罚模型。由于$\Delta w_{\text{后验}} < \Delta w_{\text{先验}}$，因此这一项为负，并且随着$\frac{\Delta w_{\text{后验}}}{\Delta w_{\text{先验}}}$的减小，它的绝对值会增加。因此，如果参数精确地调整为后验分布的数据，那么惩罚项会很大。

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

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

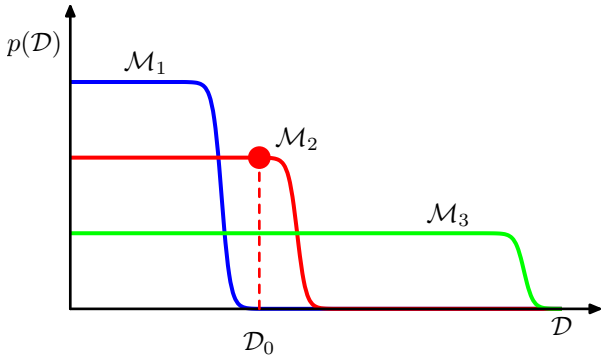

图 3.13: 对于三个具有不同复杂度的模型，数据集的概率分布的图形表示，其中$\mathcal{M}_1$是最简单的，$\mathcal{M}_3$是最复杂的。注意，概率分布是归一化的。在这个例子中，对于特定的观测数据集$\mathcal{D}_0$，具有中间复杂度的模型$\mathcal{M}_2$具有最大的模型证据。

因此，在这种非常简单的近似下，复杂度惩罚项的大小随着模型中可调节参数M的数量线性增加。随着我们增加模型的复杂度，第一项通常会增大，因为一个更加复杂的模型能够更好地拟合数据，而第二项会减小，因为它依赖于M。由最大模型证据确定的最优的模型复杂度需要在这两个相互竞争的项之间进行折中。我们后面会介绍这种近似的一个更加精炼的版本，那个版本依赖于后验概率分布的高斯近似。

通过图3.13，我们可以进一步深入认识贝叶斯模型比较，并且理解边缘似然是如何倾向于选择中等复杂度的模型的。这里，横轴是可能的数据集构成的空间的一个一维表示，因此轴上的每个点都对应着一个具体的数据集。我们现在考虑三个模型$\mathcal{M}_1, \mathcal{M}_2$和$\mathcal{M}_3$，复杂度依次增加。假设我们让这三个模型自动产生样本数据集，然后观察生成的数据集的分布。任意给定的模型都能够生成一系列不同的数据集，这是因为模型的参数由先验概率分布控制，对于任意一种参数的选择，在目标变量上都可能有随机的噪声。为了从具体的模型中生成一个特定的数据集，我们首先从先验分布$p(\boldsymbol{w})$中选择参数的值，然后对于这些参数的值，我们按照概率$p(\mathcal{D} \mid \boldsymbol{w})$对数据进行采样。一个简单的模型（例如，基于一阶多项式的模型）几乎没有变化性，因此生成的数据集彼此之间都十分相似。于是它的分布$p(\mathcal{D})$就被限制在横轴的一个相对小的区域。相反，一个复杂的模型（例如九阶多项式）可以生成变化性相当大的数据集，因此它的分布$p(\mathcal{D})$遍布了数据集空间的一个相当大的区域。由于概率分布$p(\mathcal{D} \mid \mathcal{M}_i)$是归一化的，因此我们看到特定的数据集$\mathcal{D}_0$对中等复杂度的模型有最高的模型证据。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。

贝叶斯模型比较框架中隐含的一个假设是，生成数据的真实的概率分布包含在考虑的模型集合当中。如果这个假设确实成立，那么我们我们可以证明，平均来看，贝叶斯模型比较会倾向于选择出正确的模型。为了证明这一点，考虑两个模型$\mathcal{M}_1$和$\mathcal{M}_2$，其中真实的概率分布对应于模型$\mathcal{M}_1$。对于给定的有限数据集，确实有可能出现错误的模型反而使贝叶斯因子较大的事情。但是，如果我们把贝叶斯因子在数据集分布上进行平均，那么我们可以得到期望贝叶斯因子

$$\int p(\mathcal{D} \mid \mathcal{M}_1) \ln \frac{p(\mathcal{D} \mid \mathcal{M}_1)}{p(\mathcal{D} \mid \mathcal{M}_2)} \, \mathrm{d}\mathcal{D} \tag{3.73}$$

上式是关于数据的真实分布求的平均值。这是Kullback-Leibler散度的一个例子，满足下面的性质：如果两个分布相等，则Kullback-Leibler散度等于零，否则恒为正。因此平均来讲，贝叶斯因子总会倾向于选择正确的模型。

我们已经看到，贝叶斯框架避免了过拟合的问题，并且使得模型能够基于训练数据自身进行对比。但是，与模式识别中任何其他的方法一样，贝叶斯方法需要对模型的形式作出假设，并且如果这些假设不合理，那么结果就会出错。特别地，我们从图3.12可以看出，模型证据对先验分布的很多方面都很敏感，例如在低概率处的行为等等。实际上，如果先验分布是反常的，那么模型证据无法定义，因为反常的先验分布有着任意的缩放因子（换句话说，归一化系数无法定义，因为分布根本无法被归一化）。如果我们考虑一个正常的先验分布，然后取一个适当的极限来获得一个反常的先验（例如高斯先验中，我们令方差为无穷大），那么模型证据就会趋

于零，这可以从公式（3.70）和图3.12中看出来。但是这种情况下也可能通过首先考虑两个模型的证据比值，然后取极限的方式来得到一个有意义的答案。

因此，在实际应用中，一种明智的做法是，保留一个独立的测试数据集，这个数据集用来评估最终系统的整体表现。

3.5 证据近似

在处理线性基函数模型的纯粹的贝叶斯方法中，我们会引入超参数α和β的先验分布，然后通过对超参数以及参数$\boldsymbol{w}$求积分的方式做预测。但是，虽然我们可以解析地求出对$\boldsymbol{w}$的积分或者求出对超参数的积分，但是对所有这些变量完整地求积分是没有解析解的。这里我们讨论一种近似方法。这种方法中，我们首先对参数$\boldsymbol{w}$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。这个框架在统计学的文献中被称为经验贝叶斯（empirical Bayes）（Bernardo and Smith, 1994; Gelman et al., 2004），或者被称为第二类最大似然（type 2 maximum likelihood）（Berger, 1985），或者被称为推广的最大似然（generalized maximum likelihood）。在机器学习的文献中，这种方法也被称为证据近似（evidence approximation）（Gull, 1989; MacKay, 1992a）。

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$\boldsymbol{w}, \alpha$和β求积分的方法得到，即

$$p(t \mid \mathbf{t}) = \iiint p(t \mid \boldsymbol{w}, \beta) p(\boldsymbol{w} \mid \mathbf{t}, \alpha, \beta) p(\alpha, \beta \mid \mathbf{t}) \, \mathrm{d}\boldsymbol{w} \, \mathrm{d}\alpha \, \mathrm{d}\beta \tag{3.74}$$

其中$p(t \mid \boldsymbol{w}, \beta)$由公式（3.8）给出，$p(\boldsymbol{w} \mid \mathbf{t}, \alpha, \beta)$由公式（3.49），其中$\boldsymbol{m}_N$和$\boldsymbol{S}_N$分别由公式（3.53）和公式（3.54）定义。这里，为了让记号简洁，我们省略了对于输入变量$\boldsymbol{x}$的依赖关系。如果后验分布$p(\alpha, \beta \mid \mathbf{t})$在$\widehat{\alpha}$和$\widehat{\beta}$附近有尖峰，那么预测分布可以通过对$\boldsymbol{w}$积分的方式简单地得到，其中$\alpha$和$\beta$被固定为$\widehat{\alpha}$和$\widehat{\beta}$，即

$$p(t \mid \mathbf{t}) \simeq p(t \mid \mathbf{t}, \widehat{\alpha}, \widehat{\beta}) = \int p(t \mid \boldsymbol{w}, \widehat{\beta}) p(\boldsymbol{w} \mid \mathbf{t}, \widehat{\alpha}, \widehat{\beta}) \, \mathrm{d}\boldsymbol{w} \tag{3.75}$$

根据贝叶斯定理，α和β的后验分布为

$$p(\alpha, \beta \mid \mathbf{t}) \propto p(\mathbf{t} \mid \alpha, \beta) p(\alpha, \beta) \tag{3.76}$$

如果先验分布相对比较平，那么在证据框架中，$\widehat{\alpha}$和$\widehat{\beta}$可以通过最大化边缘似然函数$p(\mathbf{t} \mid \alpha, \beta)$来获得。我们接下来会计算线性基函数模型的边缘似然函数，然后找到它的最大值。这将使我们能够从训练数据本身确定这些超参数的值，而不需要交叉验证。回忆一下比值$\frac{\alpha}{\beta}$类似于正则化参数。

此外，值得注意的一点是，如果我们定义α和β上的共轭（Gamma）先验分布，那么对公式（3.74）中的这些超参数求积分可以解析地计算出来，得到$\boldsymbol{w}$上的学生t分布（见第2.3.7节）。虽然得到的$\boldsymbol{w}$上的积分不再有解析解，但是我们可以认为对这个积分求近似会给证据框架提供了另一种实用的方法（Buntine and Weigend, 1991）。其中，可以使用拉普拉斯近似方法（见第4.4节）对这个积分求近似。拉普拉斯近似方法的基础是以后验概率分布的众数为中心的局部高斯近似方法。然而，作为$\boldsymbol{w}$的函数的被积函数的众数通常很不准确，因此拉普拉斯近似方法不能描述概率质量中的大部分信息。这就导致最终的结果要比最大化证据的方法给出的结果更差（MacKay, 1999）。

回到证据框架中，我们注意到有两种方法可以用来最大化对数证据。我们可以解析地计算证据函数，然后令它的导数等于零，得到了对于α和β的重新估计方程（将在3.5.2节讨论）。另一种方法是，我们使用一种被称为期望最大化（EM）算法的方法，这个算法将在9.3.4节讨论，那里我们还会证明这两种方法会收敛到同一个解。

3.5.1 计算证据函数

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

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

一种计算这个积分的方法是再次使用公式（2.115）给出的线性-高斯模型的条件概率分布的结果。这里，我们使用另一种方法计算这个积分，即通过对指数项配平方，然后使用高斯分布的归一化系数的基本形式。

根据公式（3.11）、公式（3.12）和公式（3.52），我们可以把证据函数写成下面的形式

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

其中M是$\boldsymbol{w}$的维数，并且，我们定义了

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

我们看到，如果忽略一些比例常数，公式（3.79）等于正则化的平方和误差函数（3.27）。我们现在对$\boldsymbol{w}$配平方，可得

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

其中我们令

$$\boldsymbol{A} = \alpha\boldsymbol{I} + \beta\boldsymbol{\Phi}^T\boldsymbol{\Phi} \tag{3.81}$$

以及

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

注意$\boldsymbol{A}$对应于误差函数的二阶导数

$$\boldsymbol{A} = \nabla\nabla E(\boldsymbol{w}) \tag{3.83}$$

被称为Hessian矩阵。这里我们也定义了$\boldsymbol{m}_N$为

$$\boldsymbol{m}_N = \beta\boldsymbol{A}^{-1}\boldsymbol{\Phi}^T\mathbf{t} \tag{3.84}$$

使用公式（3.54），我们看到$\boldsymbol{A} = \boldsymbol{S}_N^{-1}$，因此公式（3.84）等价于之前的定义（3.53），从而它表示后验概率分布的均值。

通过比较多元高斯分布的归一化系数，关于$\boldsymbol{w}$的积分现在可以很容易地计算出来了，即

$$\begin{aligned} &\int \exp\{-E(\boldsymbol{w})\} \, \mathrm{d}\boldsymbol{w} \\ &= \exp\{-E(\boldsymbol{m}_N)\} \int \exp\left\{-\frac{1}{2}(\boldsymbol{w} - \boldsymbol{m}_N)^T\boldsymbol{A}(\boldsymbol{w} - \boldsymbol{m}_N)\right\} \mathrm{d}\boldsymbol{w} \\ &= \exp\{-E(\boldsymbol{m}_N)\}(2\pi)^{\frac{M}{2}}|\boldsymbol{A}|^{-\frac{1}{2}} \end{aligned} \tag{3.85}$$

使用公式（3.78），我们可以把边缘似然函数的对数写成下面的形式

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

这就是证据函数的表达式。

回到多项式回归问题，我们可以画出模型证据与多项式阶数之间的关系，如图3.14所示。这里，我们已经假定先验分布的形式为公式（1.65），参数α的值固定为$\alpha = 5 \times 10^{-3}$。这个图像的形式非常有指导意义。我们回头看图1.4，我们看到$M = 0$的多项式对数据的拟合效果非常差，结果模型证据的值也相对较小。$M = 1$的多项式对于数据的拟合效果有了显著的提升，因此模型证据变大了。但是，对于$M = 2$的多项式，拟合效果又变得很差，因为产生数据的正弦函数是奇函数，因此在多项式展开中没有偶次项。事实上，图1.5给出的数据残差从$M = 1$到$M = 2$只有微小的减小。由于复杂的模型有着更大的复杂度惩罚项，因此

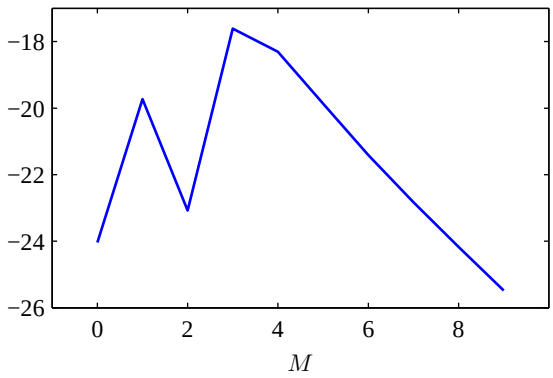

图 3.14: 多项式回归模型的模型对数证据与阶数M的关系图像，表明证据倾向于选择$M = 3$的模型。

从$M = 1$到$M = 2$，模型证据实际上减小了。当$M = 3$时，我们对于数据的拟合效果有了很大的提升，如图1.4所示，因此模型证据再次增大，给出了多项式拟合的最高的模型证据。进一步增加M的值，只能少量地提升拟合的效果，但是模型的复杂度却越来越复杂，这导致整体的模型证据会下降。再次看图1.5，我们看到泛化错误在$M = 3$到$M = 8$之间几乎为常数，因此单独基于这幅图很难对模型做出选择。然而，模型证据的值明显地倾向于选择$M = 3$的模型，因为这是能很好地解释观测数据的最简单的模型。

3.5.2 最大化证据函数

让我们首先考虑$p(\mathbf{t} \mid \alpha, \beta)$关于$\alpha$的最大化。首先定义下面的特征向量方程

$$(\beta \boldsymbol{\Phi}^T \boldsymbol{\Phi})\boldsymbol{u}_i = \lambda_i \boldsymbol{u}_i \tag{3.87}$$

根据公式（3.81），可知$\boldsymbol{A}$的特征值为$\alpha + \lambda_i$。现在考虑公式（3.86）中涉及到$\ln|\boldsymbol{A}|$的项关于α的导数

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

因此函数（3.86）关于α的驻点满足

$$0 = \frac{M}{2\alpha} - \frac{1}{2}\boldsymbol{m}_N^T\boldsymbol{m}_N - \frac{1}{2}\sum_i \frac{1}{\lambda_i + \alpha} \tag{3.89}$$

两侧乘以2α，整理，可得

$$\alpha\boldsymbol{m}_N^T\boldsymbol{m}_N = M - \alpha\sum_i \frac{1}{\lambda_i + \alpha} = \gamma \tag{3.90}$$

由于i的求和式中一共有M项，因此γ可以写成

$$\gamma = \sum_i \frac{\lambda_i}{\alpha + \lambda_i} \tag{3.91}$$

γ的意义稍后会讨论。根据方程（3.90），我们看到最大化边缘似然函数的α满足

$$\alpha = \frac{\gamma}{\boldsymbol{m}_N^T\boldsymbol{m}_N} \tag{3.92}$$

注意，这是α的一个隐式解，不仅因为γ与α相关，还因为后验概率本身的众数$\boldsymbol{m}_N$也与α的选择有关。因此我们使用迭代的方法求解。首先我们选择一个α的初始值，使用这个初始值找到$\boldsymbol{m}_N$（由公式（3.53）求得），利用公式（3.91）计算γ。之后这些值被公式（3.92）用来重新

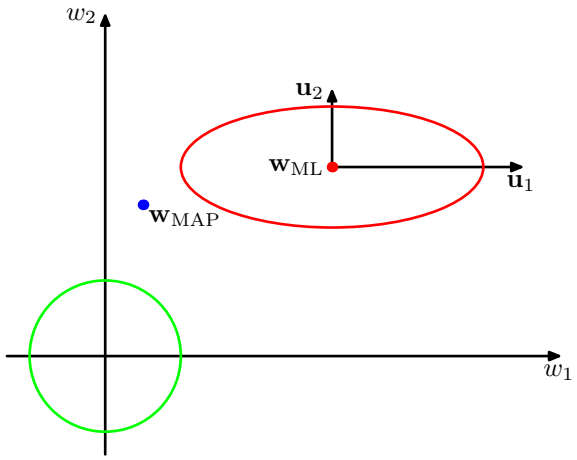

图 3.15: 似然函数的轮廓线（红色）和先验概率分布（绿色），其中参数空间中的坐标轴被旋转，与Hessian矩阵的特征向量$\boldsymbol{u}_i$对齐。对于$\alpha = 0$，后验概率分布的众数由最大似然解$\boldsymbol{w}_{ML}$给出，而对于非零的α，众数位于$\boldsymbol{w}_{MAP} = \boldsymbol{m}_N$的位置。在方向$w_1$上，由公式（3.87）定义的特征值$\lambda_1$与$\alpha$相比较小，因此$\lambda_1/(\lambda_1 + \alpha)$接近零，对应的$w_1$的MAP值也接近零。相反，在$w_2$的方向上，特征值$\lambda_2$与$\alpha$相比较大，因此$\lambda_2/(\lambda_2 + \alpha)$接近1，$w_2$的MAP值接近于最大似然值。

估计α。这个过程不断进行，直到收敛。注意，由于矩阵$\boldsymbol{\Phi}^T\boldsymbol{\Phi}$是固定的，因此我们可以在最开始的时候计算一次特征值，然后接下来只需乘以β就可以得到λ_i的值。

应该强调的是，α的值是纯粹通过观察训练集确定的。与最大似然方法不同，最优化模型复杂度不需要独立的数据集。

我们可以类似地关于β最大化对数边缘似然函数（3.86）。为了完成这一点，我们注意到公式（3.87）定义的特征值λ_i正比于β，因此$\frac{\mathrm{d}}{\mathrm{d}\beta} = \frac{\lambda_i}{\beta}$。于是

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

边缘似然函数的驻点因此满足

$$0 = \frac{N}{2\beta} - \frac{1}{2}\sum_{n=1}^{N}\{t_n - \boldsymbol{m}_N^T\boldsymbol{\phi}(\boldsymbol{x}_n)\}^2 - \frac{\gamma}{2\beta} \tag{3.94}$$

整理，我们可以得到

$$\frac{1}{\beta} = \frac{1}{N - \gamma}\sum_{n=1}^{N}\{t_n - \boldsymbol{m}_N^T\boldsymbol{\phi}(\boldsymbol{x}_n)\}^2 \tag{3.95}$$

与之前一样，这是β的一个隐式解，可以通过迭代的方法解出。首先选择β的一个初始值，然后使用这个初始值计算$\boldsymbol{m}_N$和γ，然后使用公式（3.95）重新估计β的值，重复直到收敛。如果α和β的值都要从数据中确定，那么他们的值可以在每次更新γ之后一起重新估计。

3.5.3 参数的有效数量

公式（3.92）给出的结果有一个十分优雅的意义（MacKay, 1992a），它提供给我们关于α的贝叶斯解的更深刻的认识。考虑似然函数的轮廓线以及先验概率分布，如图3.15所示。这里，我们隐式地把参数空间的坐标轴进行了旋转变换，使其与公式（3.87）定义的特征向量对齐。这样，似然函数的轮廓线就变成了轴对齐的椭圆。特征值λ_i度量了似然函数的曲率，因此在图3.15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$\beta\boldsymbol{\Phi}^T\boldsymbol{\Phi}$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$\gamma$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

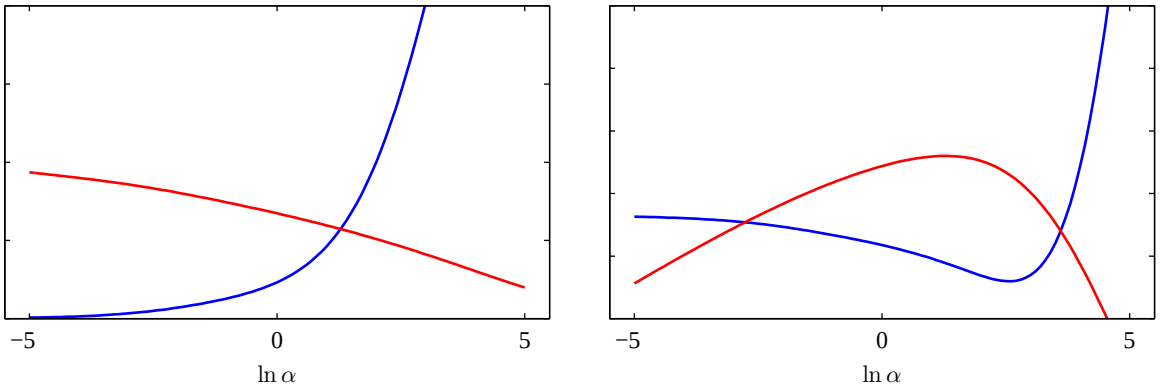

图 3.16: 左图给出了γ与$\ln\alpha$的关系（红色曲线）以及$2\alpha E_W(\boldsymbol{m}_N)$与$\ln\alpha$的关系（蓝色曲线），数据集为正弦数据集。这两条曲线的交点定义了α的最优解，由模型证据的步骤给出。右图给出了对应的对数证据$\ln p(\mathbf{t}\mid\alpha,\beta)$关于$\ln\alpha$的图像（红色曲线），说明了峰值与左图中曲线的交点恰好重合。同样给出的时测试集误差（蓝色曲线），说明模型证据最大值的位置接近于具有最好泛化能力的点。

地限制着。相反，对于$\lambda_i \ll \alpha$的方向，对应的参数w_i将会接近0，比值$\frac{\lambda_i}{\lambda_i+\alpha}$也会接近0。这些方向上，似然函数对于参数的值相对不敏感，因此参数被先验概率设置为较小的值。公式（3.91）定义的γ因此度量了良好确定的参数的有效总数。

我们可以更深刻地研究一下用于重新估计β的公式（3.95）。让我们把β和公式（3.21）给出的对应的最大似然结果进行比较。这两个公式都把方差（精度的倒数）表示为目标值和模型预测值的差的平方的平均值。但是，它们的区别在于，最大似然结果的分母是数据点的数量N，而贝叶斯结果的分母是$N-\gamma$。根据公式（1.56），我们看到单一变量x的高斯分布的方差的最大似然估计为

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

这个估计是有偏的，因为均值的最大似然解μ_{ML}拟合了数据中的一些噪声。从效果上来看，这占用了模型的一个自由度。对应的无偏的估计由公式（1.59）给出，形式为

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

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$\boldsymbol{w}^T\boldsymbol{\phi}(\boldsymbol{x})$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。

我们可以说明使用1.1节的正弦数据超参数的有效框架，以及由9个基函数组成的高斯基函数模型，因此模型中的参数的总数为$M=10$，这里包含了偏置。这里为了说明的简洁性，我们已经把β设置成了真实值11.1，然后使用证据框架来确定α，如图3.16所示。

我们也可以看到参数α是如何控制参数$\{w_i\}$的大小的。图3.17给出了独立的参数关于有效参数数量γ的函数图像。

如果我们考虑极限情况$N \gg M$，数据点的数量大于参数的数量，那么根据公式（3.87），所有的参数都可以根据数据良好确定。因为$\boldsymbol{\Phi}^T\boldsymbol{\Phi}$涉及到数据点的隐式求和，因此特征值$\lambda_i$随着数据集规模的增加而增大。在这种情况下，$\gamma=M$，并且$\alpha$和$\beta$的重新估计方程变为了

$$\alpha = \frac{M}{2E_W(\boldsymbol{m}_N)} \tag{3.98}$$

$$\beta = \frac{N}{2E_D(\boldsymbol{m}_N)} \tag{3.99}$$

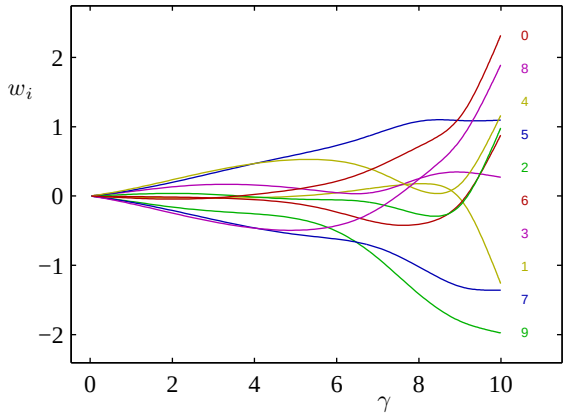

图 3.17: 高斯基函数模型中的10个参数w_i与参数有效数量γ的关系，其中超参数的变化范围为$0 \leq \alpha \leq \infty$，使得$\gamma$的变化范围为$0 \leq \gamma \leq M$。

其中E_W和E_D分别由公式（3.25）和公式（3.26）定义。这些结果可以用作完整的重新估计公式的简化计算的近似，因为它们不需要计算Hessian矩阵的一系列特征值。

3.6 固定基函数的局限性

在本章中，我们已经关注了由固定的非线性基函数的线性组合组成的模型。我们已经看到，对于参数的线性性质的假设产生了一系列有用的性质，包括最小平方问题的解析解，以及容易计算的贝叶斯方法。此外，对于一个合适的基函数的选择，我们可以建立输入向量到目标值之间的任意非线性映射。在下一章中，我们会研究类似的用于分类的模型。

因此，似乎这样的模型建立的解决模式识别问题的通用框架。不幸的是，线性模型有一些重要的局限性，这使得我们在后续的章节中要转而关注更加复杂的模型，例如支持向量机和神经网络。

困难的产生主要是因为我们假设了基函数在观测到任何数据之前就被固定了下来，而这正是1.4节讨论的维度灾难问题的一个表现形式。结果，基函数的数量随着输入空间的维度D迅速增长，通常是指数方式的增长。

幸运的是，真实数据集有两个性质，可以帮助我们缓解这个问题。第一，数据向量$\{\boldsymbol{x}_n\}$通常位于一个非线性流形内部。由于输入变量之间的相关性，这个流形本身的维度小于输入空间的维度。我们将在第12章中讨论手写数字识别时给出一个例子来说明这一点。如果我们使用局部基函数，那么我们可以让基函数只分布在输入空间中包含数据的区域。这种方法被用在径向基函数网络中，也被用在支持向量机和相关向量机当中。神经网络模型使用可调节的基函数，这些基函数有着sigmoid非线性的性质。神经网络可以通过调节参数，使得在输入空间的区域中基函数会按照数据流形发生变化。第二，目标变量可能只依赖于数据流形中的少量可能的方向。利用这个性质，神经网络可以通过选择输入空间中基函数产生响应的方向。

3.7 练习

（3.1）（*）证明，双曲正切函数与公式（3.6）定义的logistic sigmoid函数的关系为

$$\tanh(a) = 2\sigma(2a) - 1 \tag{3.100}$$

这也就能够证明，logistic sigmoid函数的一个一般的线性组合

$$y(x, \boldsymbol{w}) = w_0 + \sum_{j=1}^{M} w_j \sigma\left(\frac{x - \mu_j}{s}\right) \tag{3.101}$$

等价于一个双曲正切函数的线性组合

$$y(x, \boldsymbol{u}) = u_0 + \sum_{j=1}^{M} u_j \tanh\left(\frac{x - \mu_j}{2s}\right) \tag{3.102}$$

寻找一个表达式，将新的参数$\{u_0, \ldots, u_M\}$与原始的参数$\{w_0, \ldots, w_M\}$关联起来。

（3.2）（**）证明矩阵

$$\boldsymbol{\Phi}(\boldsymbol{\Phi}^T\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^T \tag{3.103}$$

会把任意的向量$\boldsymbol{v}$投影到由$\boldsymbol{\Phi}$的列张成的空间上。使用这个结果证明最小平方解（3.15）对应于向量$\mathbf{t}$在流形$\mathcal{S}$上的一个正交投影，如图3.2所示。

（3.3）（*）考虑一个数据集，其中每个数据点t_n都与一个权因子$r_n > 0$相关联，从而平方和误差函数变为了

$$E_D(\boldsymbol{w}) = \frac{1}{2}\sum_{n=1}^{N} r_n\{t_n - \boldsymbol{w}^T\boldsymbol{\phi}(\boldsymbol{x}_n)\}^2 \tag{3.104}$$

找到最小化这个误差函数的解$\boldsymbol{w}^*$的表达式。说出这种加权的平方和误差函数的两个意义，分别根据（1）数据对噪声方差的依赖性（2）复制的数据点。

（3.4）（*）考虑一个线性模型

$$y(\boldsymbol{x}, \boldsymbol{w}) = w_0 + \sum_{i=1}^{D} w_i x_i \tag{3.105}$$

以及平方和误差函数

$$E_D(\boldsymbol{w}) = \frac{1}{2}\sum_{n=1}^{N}\{y(\boldsymbol{x}_n, \boldsymbol{w}) - t_n\}^2 \tag{3.106}$$

现在假设服从均值为零方差为σ^2的高斯分布的噪声ϵ_i被独立地加到每个输入变量x_i上。通过使用$\mathbb{E}[\epsilon_i] = 0$和$\mathbb{E}[\epsilon_i\epsilon_j] = \delta_{ij}\sigma^2$，证明，对在噪声分布上做平均的$E_D$进行最小化，等价于对附加权值衰减的正则化项的无噪声输入变量的平方和误差函数进行最小化，其中偏置参数w_0从正则化项中被省略掉。

（3.5）（*）使用附录E中讨论的拉格朗日乘数法，证明最小化正则化的误差函数（3.29）等价于在限制条件（3.30）下最小化未正则化的平方和误差函数（3.12）。讨论参数η和λ的关系。

（3.6）（*）考虑多元目标变量$\boldsymbol{t}$的线性基函数回归模型，其中$\boldsymbol{t}$服从高斯分布，形式为

$$p(\boldsymbol{t} \mid \boldsymbol{W}, \boldsymbol{\Sigma}) = \mathcal{N}(\boldsymbol{t} \mid \boldsymbol{y}(\boldsymbol{x}, \boldsymbol{W}), \boldsymbol{\Sigma}) \tag{3.107}$$

其中

$$\boldsymbol{y}(\boldsymbol{x}, \boldsymbol{W}) = \boldsymbol{W}^T\boldsymbol{\phi}(\boldsymbol{x}) \tag{3.108}$$

训练数据集由基向量输入$\boldsymbol{\phi}(\boldsymbol{x}_n)$和对应的目标向量$\boldsymbol{t}_n$组成，其中$n = 1, \ldots, N$。证明参数矩阵$\boldsymbol{W}$的最大似然解$\boldsymbol{W}_{ML}$具有这样的性质：每一列由形如（3.15）的表达式给出，它是各向同性的噪声分布的解。注意，这个最大似然解与协方差矩阵$\boldsymbol{\Sigma}$无关。证明，$\boldsymbol{\Sigma}$的最大似然解为

$$\boldsymbol{\Sigma} = \frac{1}{N}\sum_{n=1}^{N}\left(\boldsymbol{t}_n - \boldsymbol{W}_{ML}^T\boldsymbol{\phi}(\boldsymbol{x}_n)\right)\left(\boldsymbol{t}_n - \boldsymbol{W}_{ML}^T\boldsymbol{\phi}(\boldsymbol{x}_n)\right)^T \tag{3.109}$$

（3.7）（*）通过使用配平方的方法，证明公式（3.49）给出的线性基函数模型中的参数$\boldsymbol{w}$的后验概率分布的结果，其中$\boldsymbol{m}_N$和$\boldsymbol{S}_N$分别由公式（3.50）和公式（3.51）定义。

（3.8）（**）考虑3.1节的线性基函数模型。假设我们已经观测到了N个数据点，从而$\boldsymbol{w}$的后验概率分布由公式（3.49）给出。这个后验概率可以被当成下一次观测的先验概率。通过考虑一个额外的数据点$(\boldsymbol{x}_{N+1}, t_{N+1})$，使用为指数项配平方的方法，证明最终的后验概率分布仍然由公式（3.49）给出，但是$\boldsymbol{S}_N$被替换为了$\boldsymbol{S}_{N+1}$，$\boldsymbol{m}_N$被替换为了$\boldsymbol{S}_{N+1}$。

（3.9）（**）重复上一个练习，但这次不是用手配平方，而是使用公式（2.116）给出的线性高斯模型的一般结果。

（3.10）（**）使用公式（2.115）给出的结果，计算公式（3.57）的积分，证明贝叶斯线性回归模型的预测分布由公式（3.58）给出，其中与输入相关的变量由公式（3.59）给出。

（3.11）（**）我们已经看到，随着数据集规模的增加，模型参数的后验概率分布的不确定性会降低。使用矩阵恒等式（附录C）

$$(\boldsymbol{M}+\boldsymbol{v}\boldsymbol{v}^T)^{-1}=\boldsymbol{M}^{-1}-\frac{(\boldsymbol{M}^{-1}\boldsymbol{v})(\boldsymbol{v}^T\boldsymbol{M}^{-1})}{1+\boldsymbol{v}^T\boldsymbol{M}^{-1}\boldsymbol{v}} \tag{3.110}$$

证明公式（3.59）给出的线性回归函数的不确定性$\sigma_N^2(\boldsymbol{x})$满足

$$\sigma_{N+1}^2(\boldsymbol{x})\leq\sigma_N^2(\boldsymbol{x}) \tag{3.111}$$

（3.12）（**）我们在2.3.6节看到，具有未知均值和未知精度（方差倒数）的高斯分布的共轭先验是正态-Gamma分布。这个性质对于线性回归模型的条件高斯分布$p(t\mid\boldsymbol{x},\boldsymbol{w},\beta)$也成立。如果我们考虑似然函数（3.10），那么$\boldsymbol{w}$和$\beta$的共轭先验为

$$p(\boldsymbol{w},\beta)=\mathcal{N}(\boldsymbol{w}\mid\boldsymbol{m}_0,\beta^{-1}\boldsymbol{S}_0)\,\mathrm{Gam}(\beta\mid a_0,b_0) \tag{3.112}$$

证明对应的后验概率分布具有相同的函数形式，即

$$p(\boldsymbol{w},\beta\mid\mathbf{t})=\mathcal{N}(\boldsymbol{w}\mid\boldsymbol{m}_N,\beta^{-1}\boldsymbol{S}_N)\,\mathrm{Gam}(\beta\mid a_N,b_N) \tag{3.113}$$

并且找出后验概率参数$\boldsymbol{m}_N,\boldsymbol{S}_N,a_N$和$b_N$的表达式。

（3.13）（**）证明练习3.12中讨论的模型的预测分布$p(t\mid\boldsymbol{x},\mathbf{t})$是学生t分布，形式为

$$p(t\mid\boldsymbol{x},\mathbf{t})=\mathrm{St}(t\mid\mu,\lambda,\nu) \tag{3.114}$$

并求出μ,λ和ν的表达式。

（3.14）（**）本练习中，我们仔细研究公式（3.62）定义的等价核的性质，其中$\boldsymbol{S}_N$由公式（3.54）定义。假设基函数$\phi_j(\boldsymbol{x})$是线性独立的，且观测数据点的数量N大于基函数的数量M。此外，令某一个基函数为常数，例如$\phi_0(\boldsymbol{x})=1$。通过对这些基函数进行恰当的线性变换，我们可以建立一个新的基的集合$\psi_j(\boldsymbol{x})$。这个新的基的集合能够张成同样的空间，但是基是单位正交的，即

$$\sum_{n=1}^{N}\psi_j(\boldsymbol{x}_n)\psi_k(\boldsymbol{x}_n)=I_{jk} \tag{3.115}$$

其中，如果$j=k$，则I_{jk}为1，否则为0。并且，我们取$\psi_0(\boldsymbol{x})=1$。证明对于$\alpha=0$，等价核可以写成$k(\boldsymbol{x},\boldsymbol{x}')=\boldsymbol{\psi}(\boldsymbol{x})^T\boldsymbol{\psi}(\boldsymbol{x}')$，其中$\boldsymbol{\psi}=(\psi_0,\ldots,\psi_{M-1})^T$。使用这个结果证明，核满足下面的加和限制

$$\sum_{n=1}^{N}k(\boldsymbol{x},\boldsymbol{x}_n)=1 \tag{3.116}$$

（3.15）（*）考虑回归的线性基函数模型，其中参数α和β通过模型证据框架来设定。证明由公式（3.82）定义的函数$E(\boldsymbol{m}_N)$满足关系$2E(\boldsymbol{m}_N)=N$。

（3.16）（**）使用公式（2.115）直接计算积分（3.77），推导线性回归模型的对数证据函数的结果（3.86）。

（3.17）（*）证明贝叶斯线性回归模型的证据函数可以写成公式（3.78）的形式，其中$E(\boldsymbol{w})$由公式（3.79）定义。

（3.18）（**）通过关于$\boldsymbol{w}$配平方，证明贝叶斯线性回归的误差函数（3.79）可以写成公式（3.80）的形式。

（3.19）（**）证明贝叶斯线性回归模型中，对$\boldsymbol{w}$积分会得到结果（3.85）。从而也就证明了对数边缘似然函数由公式（3.86）给出。

（3.20）（**）证明，对于对数边缘似然函数（3.86）关于α进行最大化的步骤会产生出重估计方程（3.92）。

（3.21）（**）另一种推导模型证据框架中最优的α值的结果（3.92）的方法是使用恒等式

$$\frac{\mathrm{d}}{\mathrm{d}\alpha}\ln|\boldsymbol{A}| = \mathrm{Tr}\left(\boldsymbol{A}^{-1}\frac{\mathrm{d}}{\mathrm{d}\alpha}\boldsymbol{A}\right) \tag{3.117}$$

通过考虑实对称矩阵$\boldsymbol{A}$的特征值展开式，然后使用由$\boldsymbol{A}$的特征值表示的行列式和迹的标准结果（附录C），证明这个恒等式。然后使用公式（3.117），从公式（3.86）开始，推导公式（3.92）。

（3.22）（**）证明，对于对数边缘似然函数（3.86）关于β进行最大化的步骤会产生出重估计方程（3.95）。

（3.23）（**）证明练习3.12描述的模型的数据的边缘概率分布（即模型证据）为

$$p(\mathbf{t}) = \frac{1}{(2\pi)^{\frac{N}{2}}}\frac{b_0^{a_0}}{b_N^{a_N}}\frac{\Gamma(a_N)}{\Gamma(a_0)}\frac{|\boldsymbol{S}_N|^{\frac{1}{2}}}{|\boldsymbol{S}_0|^{\frac{1}{2}}} \tag{3.118}$$

首先关于$\boldsymbol{w}$求积分，然后关于β求积分即可。

（3.24）（**）重复上一个练习，但是这次使用贝叶斯定理

$$p(\mathbf{t}) = \frac{p(\mathbf{t}\mid\boldsymbol{w},\beta)p(\boldsymbol{w},\beta)}{p(\boldsymbol{w},\beta\mid\mathbf{t})} \tag{3.119}$$

然后将先验概率分布、后验概率分布以及似然函数代入上面的表达式，推导出公式（3.118）的结果。