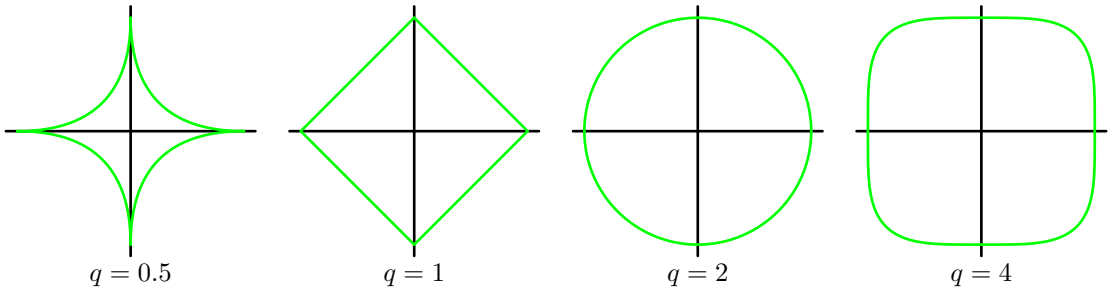

图 3.3: 对于不同的参数q，公式（3.29）中的正则化项的轮廓线。

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