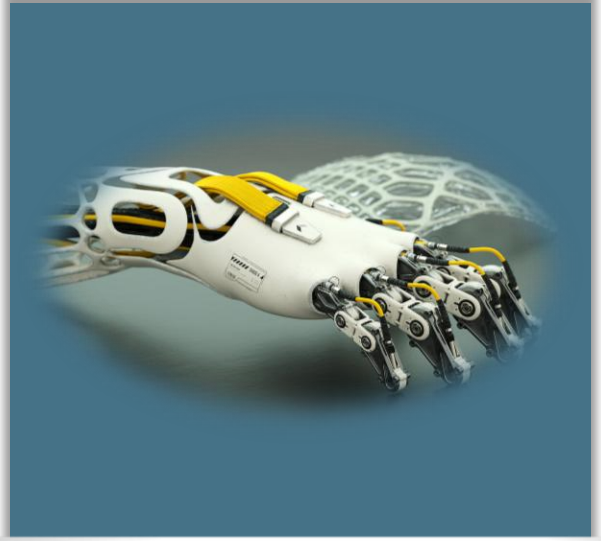
机器学习第三讲

授课人：王闻博

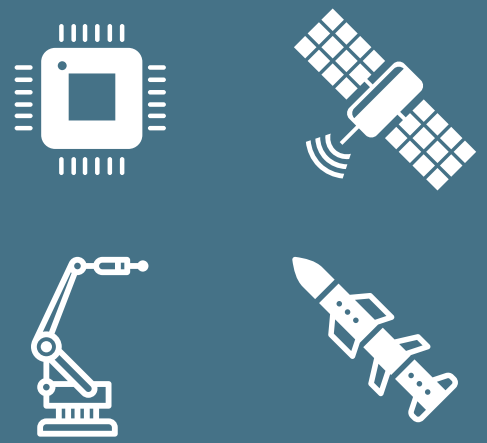
Email: wenbo_wang@kust.edu.cn

昆明理工大学 机电工程学院

2025年03月16-24日



统计机器学习中的回归问题



1. 线性回归中的最小二乘法
2. 贝叶斯线性回归
3. 线性回归中的正则化（Regularization）
4. 模型的质量指标及偏差-方差平衡问题

监督学习（Supervised Learning）：以线性回归为例

- 监督学习：
  - 从一个训练集（Training Data Set）中，获取N个样本（观测）值；
  - 训练目标是预测新样本值 x 对应的目标 t；
  - 训练集：N个样本$\{x_1,\ldots,x_n\}$及其对应的目标值 $\{t_1,\ldots,t_n\}$；
  - 确定性方法视角：构造一个函数模型$y(x)$，使得 $y(x)$ 是 t的预测值。
  - 概率性方法视角：对预测的条件分布 $p(t|x)$ 进行建模。
- 线性回归
  - 回归：构造一组给定非线性函数（称为基函数）的线性组合，$y(\mathbf{x},\mathbf{w})$，其中$\mathbf{w}$是线性组合的权重，而$\mathbf{x}$是训练集中的某个样本向量；
  - 回归任务的目标：使得$y(\mathbf{x},\mathbf{w})$的输出尽量拟合样本对应的目标值t，即$y(\mathbf{x},\mathbf{w})\approx t$。

最简单的线性回归模型

- **线性模型**

$$f(\mathbf{x}, \mathbf{w}) = w_1 x_1 + w_2 x_2 + \ldots + w_{M-1} x_{M-1} + b$$

  - 它是一个权重参数$\{w_1, w_2, \ldots, w_{M-1}, b\}$的线性函数；
  - 写成向量形式，为：$f(\mathbf{x}) = \mathbf{w}^{\mathrm{T}}\mathbf{x} + b$ 。
  - b是偏置量，大部分时间表为$b = w_0$；由此有$f(\mathbf{x}) = \mathbf{w}^{\mathrm{T}}\mathbf{x}$，其中令$x_0 = 1$，即统一为向量内积的形式。

- **线性基函数模型**

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

  - 其中，

$$\mathbf{w} = (w_0, \ldots, w_{M-1})^{\mathrm{T}} \quad \boldsymbol{\phi} = (\phi_0, \ldots, \phi_{M-1})^{\mathrm{T}} \quad \phi_0(\boldsymbol{x}) = 1$$

常见的基函数（单变量情况）

- 多项式（ Polynomial ）函数（左）

$$\phi_j(x) = x^j$$

- Gaussian函数（中）

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

- Sigmoid函数（右）

$$\phi_j(x) = \sigma\left(\frac{x-\mu_j}{s}\right) \quad \text{with } \sigma(a) = \frac{1}{1+e^{-a}}$$

- 其他：样条函数（Spline），傅里叶变换/小波变换的基函数等
  - 以傅里叶变换的基函数为例，选取M-1个频率的基函数，有

$$y(x) = x_0 + \sum_{j=1}^{M-1}(w_j \cos(\frac{2\pi jx}{T}) + v_j \sin(\frac{2\pi jx}{T}))$$

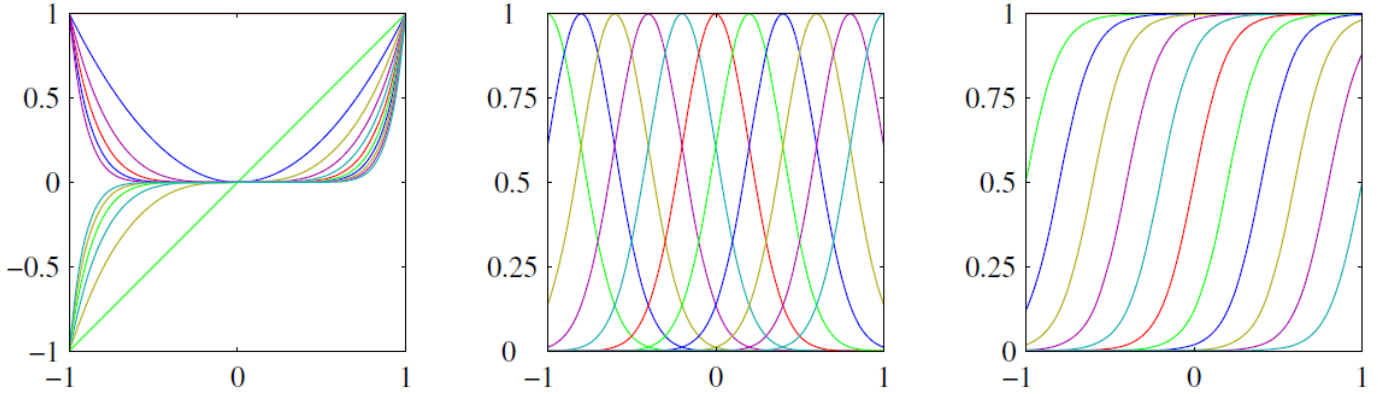


Figure 3.1 Examples of basis functions, showing polynomials on the left, Gaussians of the form (3.4) in the centre, and sigmoidal of the form (3.5) on the right.

图来源：Bishop《PRML》

训练过程：损失函数的设计

- **平方和误差（Sum-of-Square Error）**
  - 引入欧式距离相似性度量，设计损失函数：

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

  - 其中，N为训练集中的样本总量。

- **构造优化（最小化）问题：**

$$\mathbf{w}^* = \underset{\mathbf{w}}{\operatorname{argmin}}\frac{1}{2}\sum_{n=1}^{N}(f(\mathbf{x}_n,\mathbf{w}) - t_n)^2$$

$$= \underset{\mathbf{w}}{\operatorname{argmin}}\frac{1}{2}\sum_{n=1}^{N}\left(\mathbf{w}^{\mathrm{T}}\phi(\mathbf{x}_n) - t_n\right)^2$$

  - 以简单线性回归为例，有

$$\mathbf{w}^* = \underset{\mathbf{w}}{\operatorname{argmin}}\frac{1}{2}\sum_{n=1}^{N}\left(t_n - \mathbf{w}^{\mathrm{T}}\mathbf{x}_n\right)^2$$

基于SSE损失函数的优化问题闭合解

- **以单变量简单线性回归为例：最小二乘法**
  - 将$E_D(\mathbf{w})$对**w求梯度（偏导）**并令梯度为0，

$$\frac{\partial E_D}{\partial w} = \sum_{i=1}^{N} x_i(wx_i - t_i) = 0$$

  - 展开后得到

$$w\sum_{i=1}^{N} x_i^2 = \sum_{i=1}^{N} x_i t_i$$

  - 解得w的解析解：

$$w = \frac{\sum_{i=1}^{N} x_i t_i}{\sum_{i=1}^{N} x_i^2}$$

几何解释：简单最小二乘法就是试图找到一条直线，使得所有样本目标值到直线上的欧氏距离之和最小

简单线性回归拓展到输入为多元向量情况

- **多元线性回归：一个样本有d个属性（分量）描述**
  - 样本i：$x_i = [x_{i,1}, x_{i,2}, \dots, x_{id}]^T$，$t_i \in \mathbb{R}$
  - 模型参数：$\mathbf{w} = (w_0 \quad \cdots \quad w_d)^T$
- **数据集表示为张量（矩阵形式）**

$$X = \begin{pmatrix} 1 & x_{1,1} & \cdots & x_{1,d-1} & x_{1,d} \\ 1 & x_{2,1} & \cdots & x_{2,d-1} & x_{2,d} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 1 & x_{N,1} & \cdots & x_{N,d-1} & x_{N,d} \end{pmatrix} = \begin{pmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_m^T \end{pmatrix} \qquad \mathbf{t} = \begin{pmatrix} t_1 \\ t_2 \\ \vdots \\ t_N \end{pmatrix}$$

- **SSE误差最小化的矩阵表达式**

$$\mathbf{w}^* = \mathrm{argmin}_{\boldsymbol{w}} \, (\mathbf{t} - X\mathbf{w})^T (\mathbf{t} - X\mathbf{w})$$

多元向量的核函数线性回归的最小二乘解

- **最小二乘法求优化问题**：$\mathbf{w}^* = \text{argmin}_{\boldsymbol{w}} (\mathbf{t} - X\mathbf{w})^T(\mathbf{t} - X\mathbf{w})$
  - 令$E = (\mathbf{t} - X\mathbf{w})^T(\mathbf{t} - X\mathbf{w})$;
  - 求梯度并令其等于0：$\nabla_{\mathrm{w}} E = 2X^T(X\mathbf{w} - \mathbf{t}) = 0$;
  - 若X^TX满秩或正定则可以解得: $\mathbf{w}^* = (X^TX)^{-1}X^T\mathbf{t}$，　$(X^TX)^{-1}X^T$是X的左伪逆;
  - 若X^TX不满秩则可以解出多个$\mathbf{w}^*$。
- **拓展到核函数模型**：$\mathbf{t} = \mathbf{w}^{\mathrm{T}}\phi(\mathbf{x}_n)$ 则有

$$\boldsymbol{w}_{ML} = (\Phi^\top\Phi)^{-1}\Phi^\top\boldsymbol{t} \text{ with } \Phi_{M\times N} = [\phi_{mn}(\boldsymbol{x}_n)]$$

  - 一般情况下样本数远大于特征数，即$M \gg N$,Φ 由于是我们构造的基函数，则应列满秩，对应左伪逆;
  - 若Φ行满秩（此时样本数＜特征数）则使用右伪逆，（由于不合理）右伪逆不应出现在现实问题回归解中;
  - 若Φ秩亏损（$\text{rank}(X) < \min(M, N)$），应通过奇异值分解求广义伪逆（同样不合理）。

补充：矩阵和向量函数的微分（部分公式）

- **向量函数的微分（将自变量向量作为一个整体处理）**
  - 一般线性函数$\boldsymbol{f}(\mathbf{x})=\boldsymbol{a}^{\boldsymbol{T}}\mathbf{x}+\boldsymbol{b}$有：$\frac{\partial f(\mathbf{x})}{\partial \mathbf{x}}=\mathbf{a}$;
    - 或$\frac{\partial \mathbf{x}^{\boldsymbol{T}}\mathbf{a}}{\partial \mathbf{x}}=\frac{\partial \mathbf{a}^{\boldsymbol{T}}\mathbf{x}}{\partial \mathbf{x}}=\mathbf{a}$;
  - 二次型$\boldsymbol{f}(\mathbf{x})=\mathbf{x}^{\boldsymbol{T}}\mathrm{Ax}$（展开后对每个分量求导）：$\frac{\partial f(\mathbf{x})}{\partial \mathbf{x}}=\mathbf{2}\mathrm{Ax}$
    - （扩展）X是向量/矩阵：$\frac{\partial}{\partial \mathrm{X}}(\mathbf{Xb}+\mathbf{c})^{\boldsymbol{T}}\mathbf{D}(\mathbf{Xb}+\mathbf{c})=\left(\mathbf{D}+\mathbf{D}^{\boldsymbol{T}}\right)(\mathbf{Xb}+\mathbf{c})\,\mathrm{b}^{\boldsymbol{T}}$;
    - 假设W是对称矩阵：$\frac{\partial}{\partial \mathrm{x}}(\mathbf{y}-\mathbf{Ax})^{\boldsymbol{T}}\mathrm{W}(\mathbf{y}-\mathbf{Ax})=-\mathbf{2}\mathbf{A}^{\boldsymbol{T}}\mathrm{W}(\mathbf{y}-\mathbf{Ax})$;
  - 乘积法则$\boldsymbol{f}(\mathbf{x})=\mathbf{u}(\mathbf{x})^{\boldsymbol{T}}\mathbf{v}(\mathbf{x})$，$\mathbf{u}(\cdot)$、$\mathbf{v}(\cdot)$为向量函数

$$\frac{\partial f(\mathbf{x})}{\partial \mathbf{x}}=\left(\frac{\partial \mathbf{u}}{\partial \mathbf{x}}\right)^{\mathrm{T}}\mathbf{v}+\left(\frac{\partial \mathrm{v}}{\partial \mathbf{x}}\right)^{\mathrm{T}}\mathbf{u}$$

雅可比（Jacobian）阵

最小二乘法的几何解释

最小二乘法的几何意义在于：将数据拟合问题转化为高维空间中的投影问题

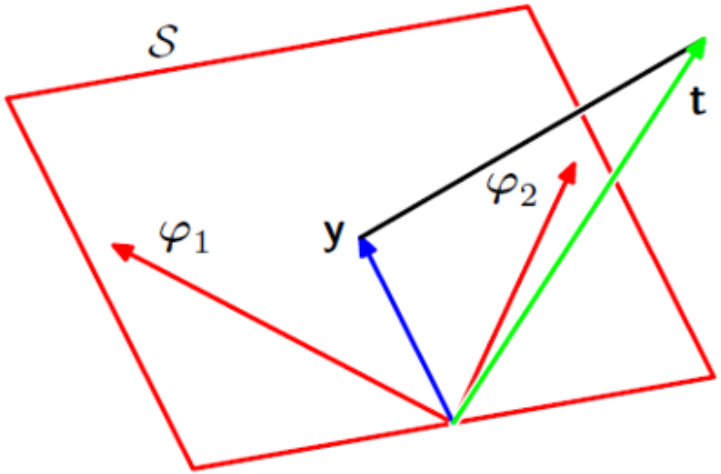
- 在以样本目标值$\{t_0, \ldots, t_N\}$为坐标轴的N维空间中，最小二乘回归通过将数据向量 **t**（绿色箭头）正交投影到由基函数矩阵Φ的列向量ϕ_j张成的子空间 ***S***（红色多边形）上来实现。
- 注意1：上述张成子空间中，基函数向量 ϕ_j 的第n个元素值是根据 第n个样本点对应元素求得的$\phi_j(x_n)$。
- 注意2：子空间 ***S***是基函数向量线性组合的集合，即模型可能的预测结果空间，是一个超平面。
- 最小二乘（优化过程）的本质：

  找到在子空间 ***S*** 中与数据向量 **t** 最接近的点 **y**（**t** 的正交投影）。

$$\Phi = \begin{pmatrix} 1 & \phi_1(x_{1,1}) & \cdots & \phi_{d-1}(x_{1,d-1}) & \phi_d(x_{1,d}) \\ 1 & \phi_1(x_{2,1}) & \cdots & \phi_{d-1}(x_{2,d-1}) & \phi_d(x_{2,d}) \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 1 & \phi_1(x_{N,1}) & \cdots & \phi_{d-1}(x_{N,d-1}) & \phi_d(x_{N,d}) \end{pmatrix}$$

样本排列

基函数列向量 ϕ_0 ϕ_1 ϕ_{d-1} ϕ_d



图来源：Bishop《PRML》

参数集更新的序列算法（Sequential Learning）

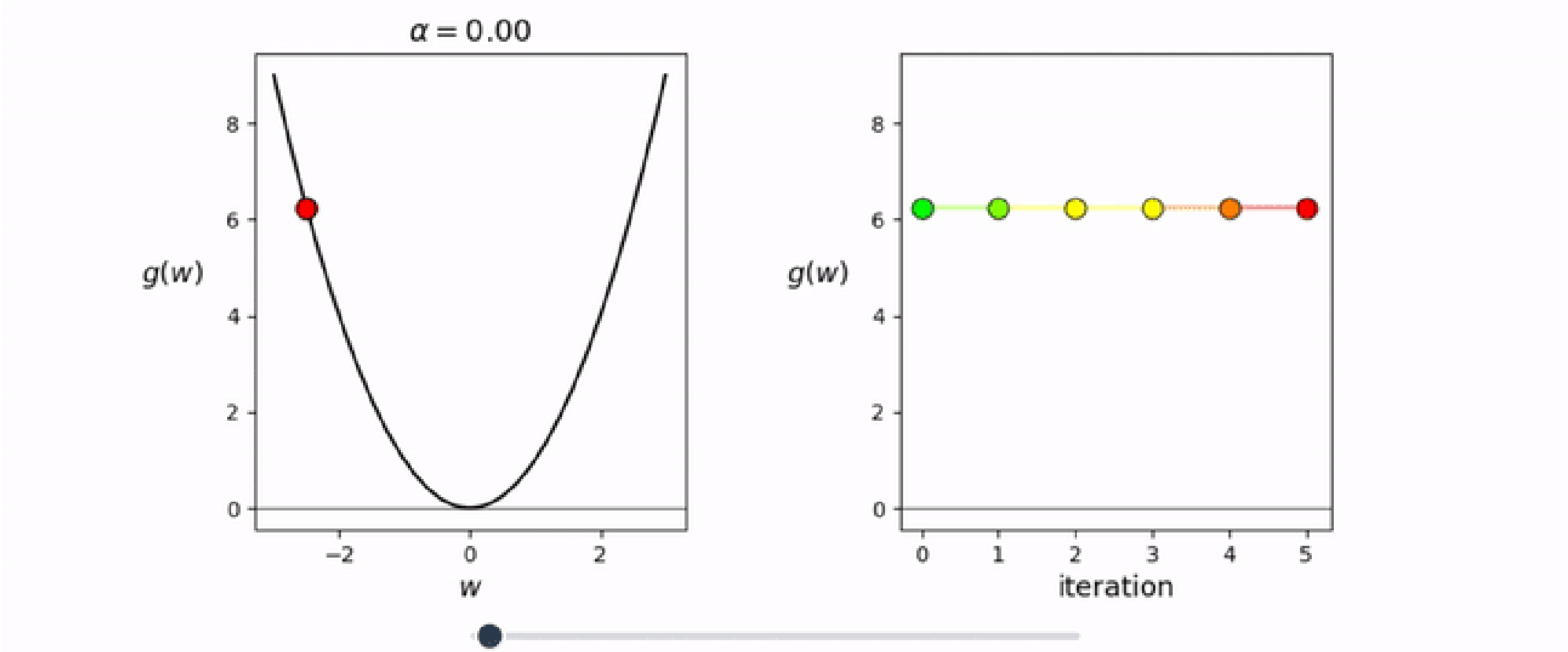
- **使用梯度下降法或随机梯度下降（Stochastic Gradient Descent）；**
  - 对第n个训练样本，构造平方误差：$E_n = \frac{1}{2}\left(\mathbf{w}^{\mathrm{T}}\phi(\mathbf{x}_n) - t_n\right)^2$；

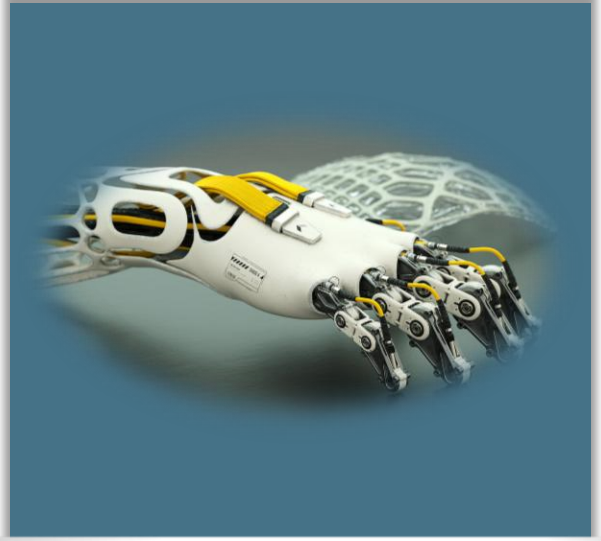
    则有：
    $$E_D(\boldsymbol{w}) = \frac{1}{2}\sum_{n=1}^{N}\left(t_n - \boldsymbol{w}^{\top}\boldsymbol{\phi}(\boldsymbol{x}_n)\right)^2 = \sum_n E_n$$
  - 使用梯度方法，对新加入的样本n更新参数$\mathbf{w}$：
    $$\mathbf{w}^{(\tau+1)} = \mathbf{w}^{(\tau)} - \eta\nabla E_n$$
  - 其中，η是参数更新的“学习率”（Learning Rate），τ是迭代数。
  - 更新总是从某个初始参数集$\mathbf{w}^0$开始的，对于一般基函数下的SSE损失函数，有：
    $$\boldsymbol{w}^{(\tau+1)} = \boldsymbol{w}^{(\tau)} + \eta\underbrace{\left(t_n - \boldsymbol{w}^{(\tau)\top}\boldsymbol{\phi}(\boldsymbol{x}_n)\right)\boldsymbol{\phi}(\boldsymbol{x}_n)}_{\nabla E_n}$$

梯度下降中的超参数影响

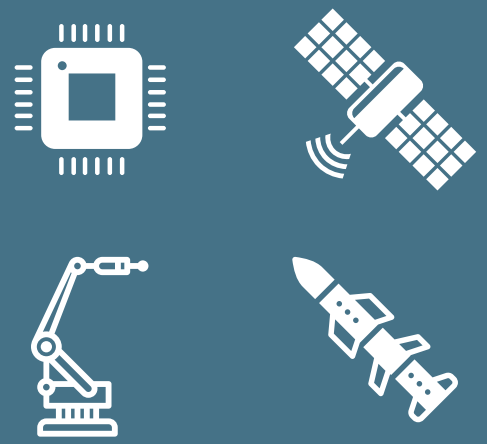
固定步长（此处以α表示）大小对优化结果的影响

- 迭代速度 vs. 目标函数质量





统计机器学习中的回归问题



1. 线性回归中的最小二乘法
2. 贝叶斯线性回归
3. 线性回归中的正则化（Regularization）
4. 模型的质量指标及偏差-方差平衡问题

极大似然法：考虑样本获取过程的随机性

- **考虑样本目标值仍由确定性函数决定，但采样过程引入多元变量加性Gaussian噪声：**

$$t = y(\mathbf{x}, \mathbf{w}) + \varepsilon$$

  - 此处ε是白噪声信号：**期望为零，协方差矩阵为对角阵，精度（方差的倒数）为β**的高斯随机变量；
  - 噪声采样ε为独立同分布（i.i.d.）。

- **由i.i.d.高斯分布，对整个目标集 t（列向量）而言，其在给定样本集x和参数集w下发生的条件概率可写为（注意此处有$y(\mathbf{x}, \mathbf{w}) = \mathbf{w}^{\mathrm{T}}\phi(\mathbf{x})$）：**

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

  - 在新输入x下，其对应目标值的期望值可根据条件概率$p(t|\mathbf{x})$表出

$$\mathbb{E}[t|\mathbf{x}] = \int t \cdot p(t|\mathbf{x})\mathrm{d}t = y(\mathbf{x}, \mathbf{w})$$

这是我们待定的预测函数

极大似然法（续）:构造不同的目标函数

- **似然函数（Likelihood Function）：**

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

（回忆）高斯分布：

$$\mathcal{N}(x|\mu,\sigma^2)=\frac{1}{(2\pi\sigma^2)^{1/2}}\exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right)$$

- **极大似然法（MLE）下的参数求取**
  - 似然函数取自然对数，有

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

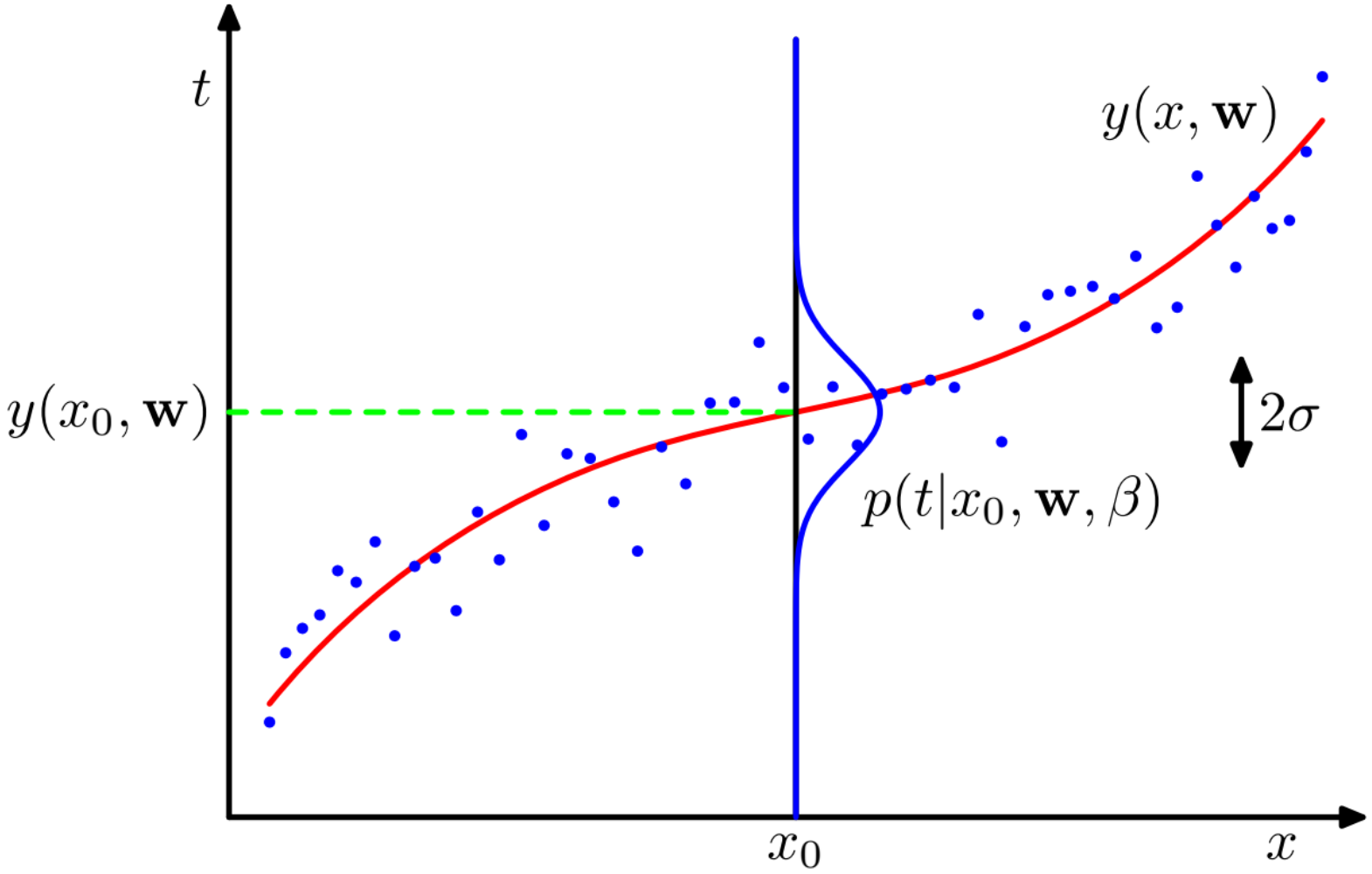
左式的前两项为常数。

  - 其中，E_D恰好是平方和误差（SSE）

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

总结：在高斯加性噪声的假设条件下，基于基函数线性组合的**极大似然法**求解等价于**最小化SSE损失函数**。

基于基函数的极大似然法构造线性回归示意图



来源：Bishop《PRML》

线性回归中输出为多变量的情况

- **目标值 t 由单变量变为多变量（K维向量）的情况时**
  - 使用同样形式的基函数$\phi(\mathbf{x})$；
  - 参数变量由向量$[\mathrm{w}]_{M\times 1}$拓展为2维张量（矩阵） $[\mathrm{W}]_{M\times K}$；
  - 预测函数形式： $y(\mathbf{x},\mathrm{W})=\mathrm{W}^{\mathrm{T}}\phi(\mathbf{x})$ 。
- **在高斯噪声假设下对单个样本目标值构造似然函数：**

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

- **基于全样本集，构造对数似然函数：**

$$\ln\left(p(\mathbf{T}|\mathbf{x},\mathrm{W},\beta)=\prod_{n=1}^{N}\mathcal{N}(\mathbf{t}|\boldsymbol{W}^{\mathrm{T}}\phi(\mathbf{x}_n),\beta^{-1}\mathbf{I})\right)=\frac{NK}{2}\ln\left(\frac{\beta}{2\pi}\right)-\frac{\beta}{2}\sum_{n=1}^{N}\left\|\mathbf{t}_n-W^{T}\phi(\mathbf{x})\right\|^{2}$$

线性回归中输出为多变量的情况（续）

- **面向参数W，有类似单目标值下的最小二乘表达结构（其中T扩展为M × K的矩阵）**

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

  - 对于每一个目标向量 $\mathbf{t}_k$，有：

$$\mathbf{w}_{\boldsymbol{k}} = (\boldsymbol{\Phi}^{\mathrm{T}}\,\Phi)^{-1}\Phi\boldsymbol{T}\mathbf{t}_k$$

  - 上式说明，矩阵W_{ML}中的所有列向量$\mathbf{w}_{\boldsymbol{k}}$共享同一个左伪逆矩阵$\Phi^{\dagger} = (\boldsymbol{\Phi}^{\mathrm{T}}\Phi)^{-1}\Phi\boldsymbol{T}$

  - 对比单变量情况下的最小二乘表达式

$$\boldsymbol{w}_{ML} = (\Phi^{\top}\Phi)^{-1}\Phi^{\top}\boldsymbol{t} \text{ with } \Phi_{M\times N} = [\phi_{mn}(\boldsymbol{x}_n)]$$

（单输出）线性回归的贝叶斯处理

- **考虑线性回归中使用基本最大似然估计方法的情况**

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

似然函数

- **对模型参数w引入高斯先验概率（Prior Probability）模型假设如下：**

$$p(\mathbf{w})=\mathcal{N}(\mathbf{w}|\mathbf{m}_0,\mathbf{S}_0)$$

这是一个以$\mathbf{m}_0$为期望，以$\mathbf{S}_0$为协方差的高斯分布

- **则后验概率根据样本集里的*N*个样本决定：**
  - 注意：由于引入的先验分布是高斯分布，则**后验分布依然是高斯分布**（证明略，高斯共轭）

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

  - 我们所关心的上述后验概率分布参数由样本集估计如下：

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

贝叶斯线性回归（接上页）

- **考虑各项同性的零均值先验高斯分布，并以α为其准确性参数（控制方差）**

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

$\mathbf{m}_0$ $\mathbf{S}_0$

各向同性，即参数维度间独立，协方差矩阵非对角元素为0

- **上述高斯分布对应参数为（从上页底结果可得）：**

$$\begin{aligned} \mathbf{m}_N &= \beta \mathbf{S}_N \boldsymbol{\Phi}^{\mathrm{T}} \mathbf{t} \\ \mathbf{S}_N^{-1} &= \alpha \mathbf{I} + \beta \boldsymbol{\Phi}^{\mathrm{T}} \boldsymbol{\Phi} \end{aligned}$$

这是一个以$\mathbf{m}_0$为期望，以$\mathbf{S}_0$为协方差的高斯分布

- **似然函数：**

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

- **根据Bayesian定理，更新先验分布为后验分布（信念）：**
  - $p(\mathbf{w}|\mathbf{t}, \mathbf{X}, \beta) \propto p(\mathbf{w})p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \beta)$

解释

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

模型证据（Model Evidence）/边际似然函数

模型概率/先验概率

边际似然衡量“给定模型时，获取当前数据的可能

贝叶斯线性回归

- **则对后验概率取对数，得（详细证明略）：**
  - （接上页）注意——后验概率的形式（先验模型 ×基于数据的似然）：

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

  - 是单样本似然函数取对数的和 （类似我们之前的似然数对数）+ 先验概率取对数（一个w的函数）

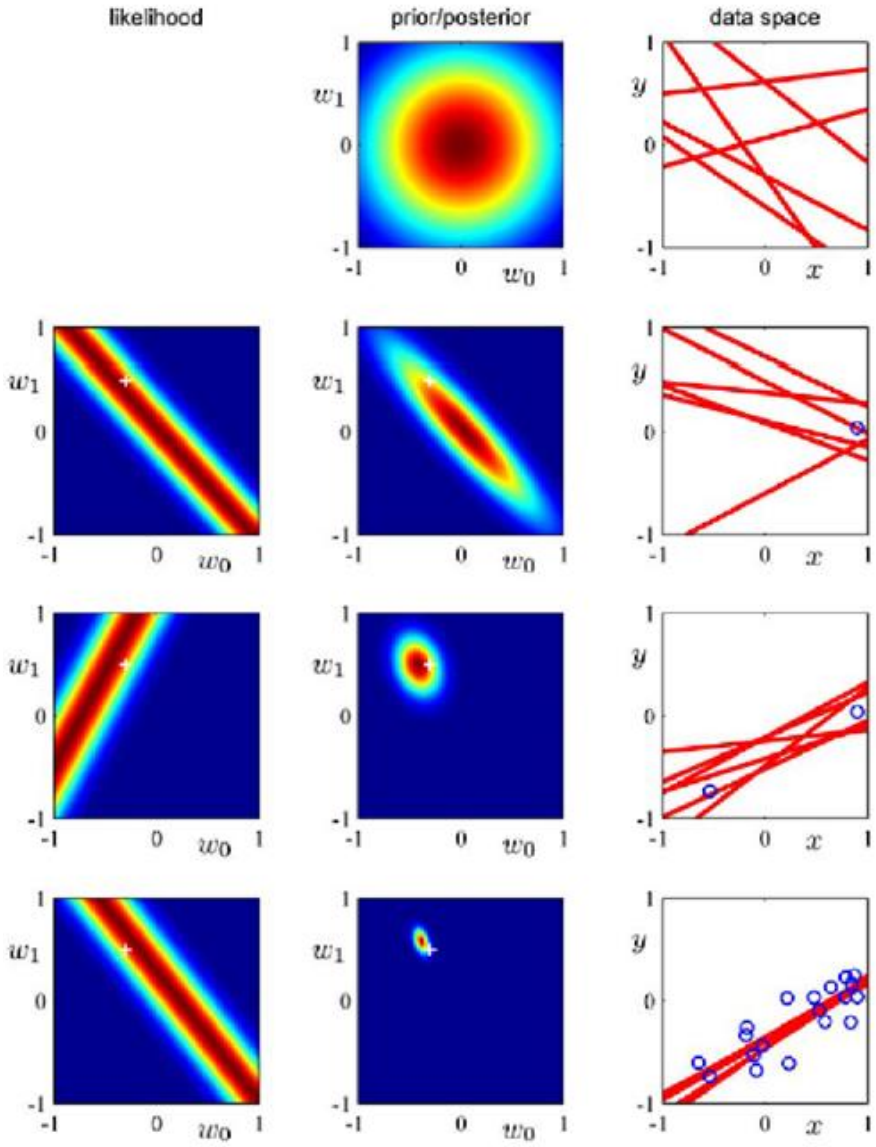
$$\ln p(\mathbf{w}|\mathbf{t}) = -\frac{\beta}{2}\sum_{n=1}^{N}\{t_n - \mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_n)\}^2 - \frac{\alpha}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w} + \text{const}$$

对上式相对于w求最大值的过程等价于对“SSE损失函数 + 一个二项式正则项“求最小值。

贝叶斯线性回归的可视化解释

- **考虑单输入下简单线性核的回归过程**
  - 线性模型 $y(x, \mathbf{w}) = w_0 + w_1 x$
  - 数据生成函数（真值）：

    $f(x, \mathbf{a}) = a_0 + a_1 x$ + Gaussian噪声
  - a_0=-0.3，a_1=0.5；
  - 右侧栏：红色线为模型预测线，蓝色点为样本点；
  - 中间栏：在w_0，w_1平面的先验或后验分布热力图；
  - 左侧栏：单样本似然函数$p(t|\mathbf{x}, \mathbf{w})$的热力分布图（由似然函数约束预测线必须接近每个采样）。



第一行：无样本值，**w**从先验分布取值得到的预测值

第二行：单样本点，左栏图十字代表实际的w值

第三行：两个样本点，左栏单样本似然函数乘第二行后验概率函数得到本行后验概率

第四行：20个样本点，左侧图代表第20个样本的似然值

图来源：Bishop《PRML》

（贝叶斯回归）预测值的概率分布

- **在回归问题中，我们构造w的目的最终是为了得到对“新输入x的对应目标值t”的预测**
  - 求取后验预测分布的情况（右端根据全概率公式）

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

    - 上式中，条件变量除训练集中的N维t向量外，为简便起见忽略了输入矩阵X。
  - 加性高斯噪声下的目标值生成条件概率（右端第一项）：

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

  - 后验概率（右端第二项）： $p(\mathbf{w}|\mathbf{t},\alpha,\beta)=\mathcal{N}(\mathbf{w}|\mathbf{m}_N,\mathbf{S}_N)$ 其中

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

预测值分布 → （上面两个分布相互独立）

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

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

- $\frac{1}{\beta}$：数据生成过程中的噪声方差
- $\boldsymbol{\phi}(\mathbf{x})^{\mathrm{T}}\mathbf{S}_N\boldsymbol{\phi}(\mathbf{x})$：与参数w相关的不确定性（参数分布的方差）

预测值分布函数的可视化解释

- **预测值分布函数**
  - 9个高斯核函数在不同训练集下的预测方差情况。
  - 绿色曲线对应于生成数据点的函数 sin(2πx)（生成过程中添加高斯噪声）。
  - 四个子图中，**蓝色圆圈**分别展示了包含 N=1、N=2、N=4 和 N=25 个数据点的数据集。
  - 在每个子图中，红色曲线表示对应高斯基函数预测分布的均值，红色阴影区域覆盖了均值两侧各一个标准差的范围。

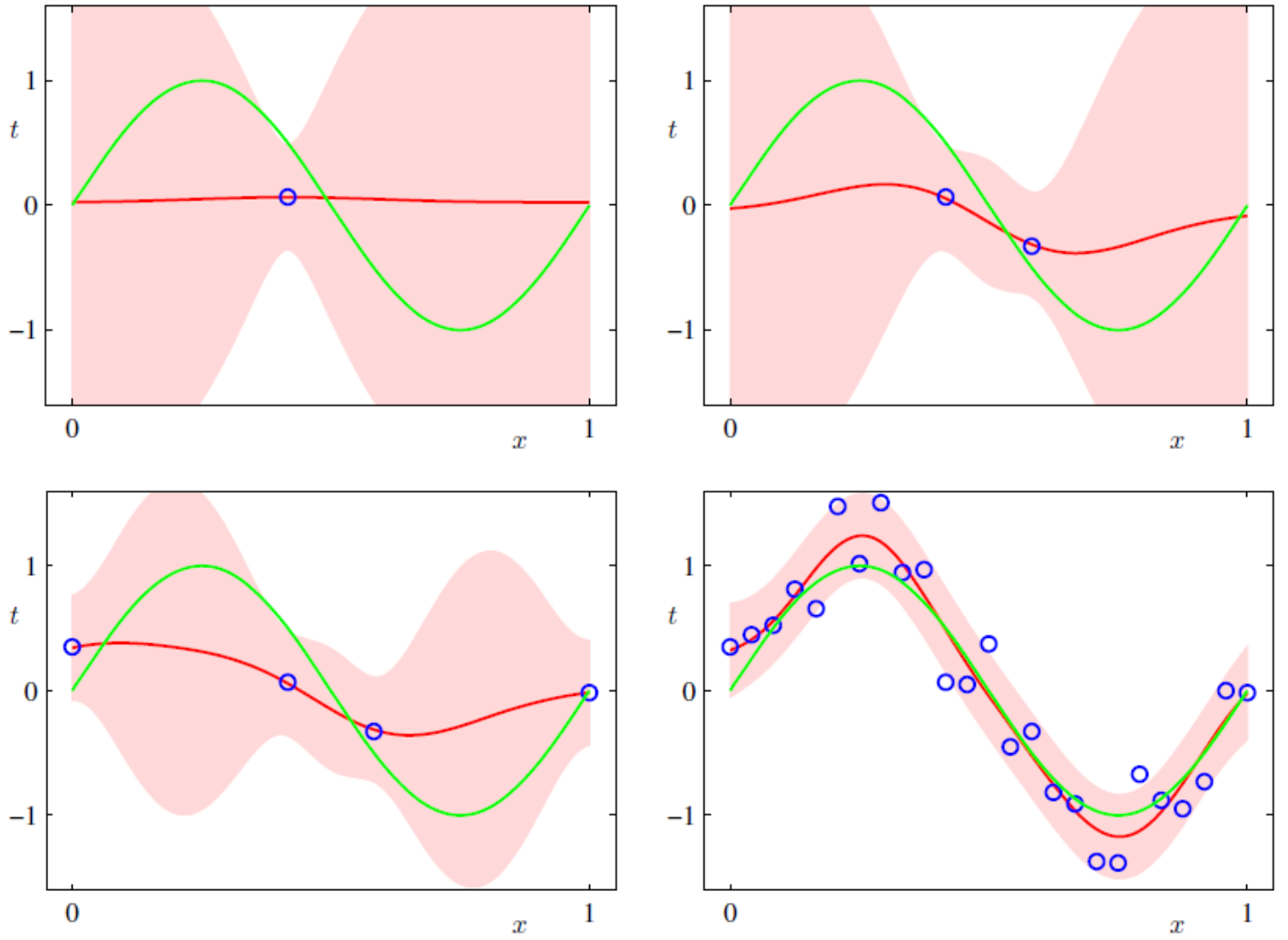


Figure 3.8 Examples of the predictive distribution (3.58) for a model consisting of 9 Gaussian basis functions of the form (3.4) using the synthetic sinusoidal data set of Section 1.1. See the text for a detailed discussion.

图来源：Bishop《PRML》

等价核（Equivalent Kernel）

- **将如下关于后验概率的预测均值参数取代w，代入预测方程$y(\mathbf{x},\mathbf{w})=\mathbf{w}^T\boldsymbol{\phi}(\mathbf{x})$：**

$$\begin{aligned}\mathbf{m}_N &= \beta\mathbf{S}_N\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{t}\\ \mathbf{S}_N^{-1} &= \alpha\mathbf{I}+\beta\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi}\end{aligned}$$

- 得预测期望为：$y(\mathbf{x},\mathbf{m}_N)=\mathbf{m}_N^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x})=\beta\boldsymbol{\phi}(\mathbf{x})^{\mathrm{T}}\mathbf{S}_N\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{t}=\sum_{n=1}^{N}\beta\boldsymbol{\phi}(\mathbf{x})^{\mathrm{T}}\mathbf{S}_N\boldsymbol{\phi}(\mathbf{x}_n)t_n$
- 上式可写为：

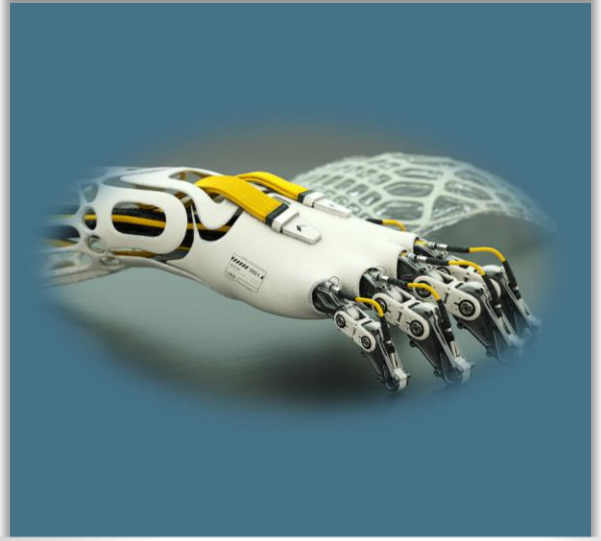
$$y(\mathbf{x},\mathbf{m}_N)=\sum_{n=1}^{N}k(\mathbf{x},\mathbf{x}_n)t_n$$

  - 其中

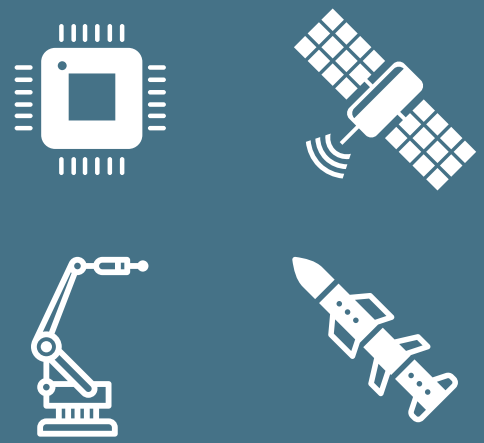
$$k(\mathbf{x},\mathbf{x}')=\beta\boldsymbol{\phi}(\mathbf{x})^{\mathrm{T}}\mathbf{S}_N\boldsymbol{\phi}(\mathbf{x}')$$

称为等价核（Equivalent Kernel）

- 此时，对任意输入的预测值输出由计算训练集各目标值的线性组合决定，权重为样本对应的等价核函数的输出。

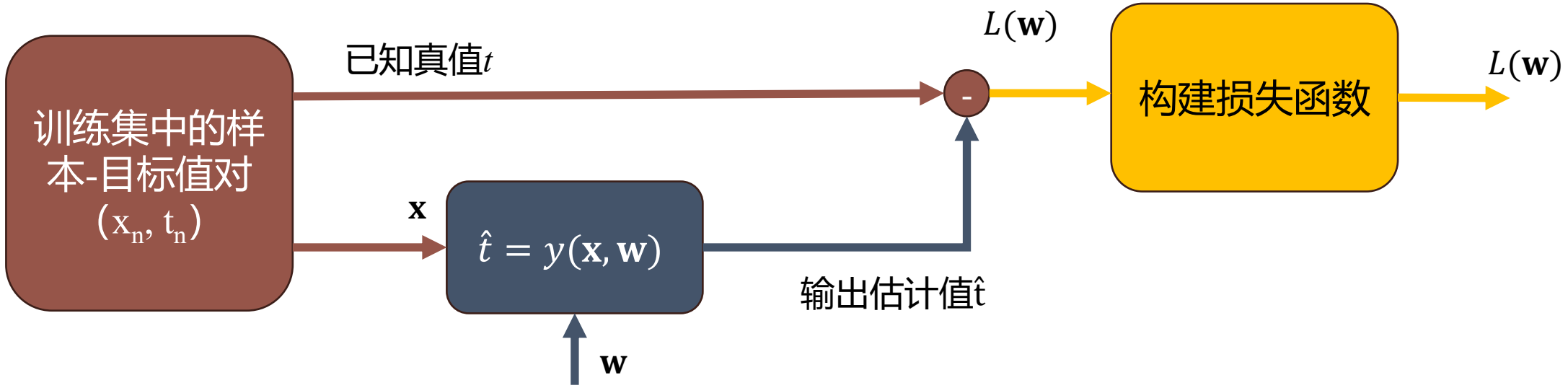


统计机器学习中的回归问题



1. 线性回归中的最小二乘法
2. 贝叶斯线性回归
3. 线性回归中的正则化（Regularization）
4. 模型的质量指标

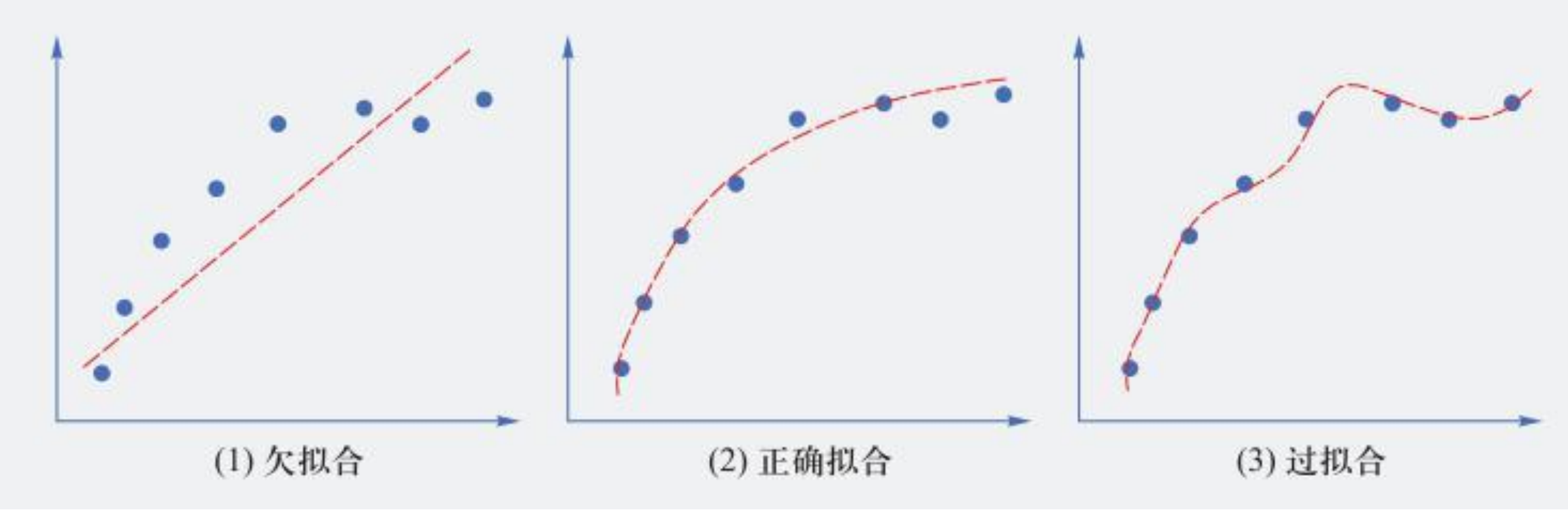
回顾：监督学习框架下的回归问题



- **求解回归问题的四个步骤**
  - 收集数据构建训练集；
  - 选取预测函数模型；
  - 确定损失函数；
  - 求取使损失函数最小化的参数集w。

在有限数据集上训练模型时发生的欠拟合和过拟合问题

- **模型泛化：期望模型在测试集上的表现和训练集上一样好。**
- **过拟合的发生条件**
  - 基函数数量大 + 训练数据有限 → 模型过拟合：在训练集上表现完美，但在未知数据上表现欠佳；
  - 同时，限制基函数数量会降低模型灵活性。



模型的偏差-方差分解与设计权衡

- **偏差（Bias）和方差（Variance）**
  - **偏差**：模型预测均值与真实目标值t的偏离；
  - **方差**：模型预测对训练数据敏感性的度量；
  - **噪声**：数据本身的不可约不确定性。
- **权衡原则：**
  - 高偏差 → 欠拟合（模型过于简单）；
  - 高方差 → 过拟合（模型过于复杂）。
- **处理方法：引入正则化（Regularization）操作**
  - 正则化：在损失函数中加入与模型参数相关的惩罚项

总损失 = 原始设计损失（如均方误差）+λ·正则化项

偏差-方差分解

- 目标值的条件分布（用$h(\mathbf{x})$表示）：

$$h(\mathbf{x}) = \mathbb{E}[t|\mathbf{x}] = \int t p(t|\mathbf{x})\,\mathrm{d}t$$

- 在Square Error误差函数情况下，单样本平方预测值的损失期望：

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

- 在给定数据集D下，单样本平方预测值的损失期望：

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

推导过程：

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

- 由此，总损失函数由以下项目构成：

$$\text{expected loss} = (\text{bias})^2 + \text{variance} + \text{noise}$$

- 其中：

$$\begin{aligned}(\text{bias})^2 &= \int \{\mathbb{E}_{\mathcal{D}}[y(\mathbf{x};\mathcal{D})] - h(\mathbf{x})\}^2 p(\mathbf{x})\,\mathrm{d}\mathbf{x}\\ \text{variance} &= \int \mathbb{E}_{\mathcal{D}}\left[\{y(\mathbf{x};\mathcal{D}) - \mathbb{E}_{\mathcal{D}}[y(\mathbf{x};\mathcal{D})]\}^2\right] p(\mathbf{x})\,\mathrm{d}\mathbf{x}\\ \text{noise} &= \int \{h(\mathbf{x}) - t\}^2 p(\mathbf{x}, t)\,\mathrm{d}\mathbf{x}\,\mathrm{d}t\end{aligned}$$

正则化的最小二乘法

- **在SSE损失函数基础上添加正则化项：**

$$\boldsymbol{E_D(\mathbf{w}) + \lambda E_w(\mathbf{w})}$$

  - 超参数$\lambda > 0$用来控制正则化项对总损失函数的贡献值。
  - 最简单的正则化项：权重参数的二范数（即sum-of-square）：

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

  - 总损失式可写为：

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

    - 此种正则项也称为“参数衰减（weight decay）”，因为在序列更新算法中（以线性核单输出预测函数为例）：

$$\mathbf{w}^{t+1} \leftarrow \mathbf{w}^t - \lambda\left(\mathbf{x}_n\left(\mathbf{x}_n^T\mathbf{w} - t\right) + \lambda\mathbf{w}\right)$$

在不受数据$\mathbf{x}_n$支持的情况下，更新总倾向使$\mathbf{w}$向零点方向衰减。

正则化的最小二乘法（续）

- **在SSE损失函数基础上添加l_2-Norm正则化项：**

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

- **最小二乘解：**

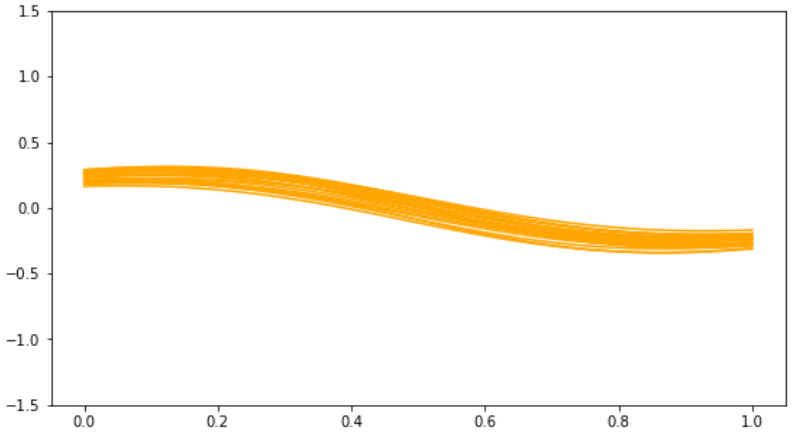
$$\mathbf{w} = \left(\lambda\mathbf{I} + \mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi}\right)^{-1}\mathbf{\Phi}^{\mathrm{T}}\mathbf{t}$$

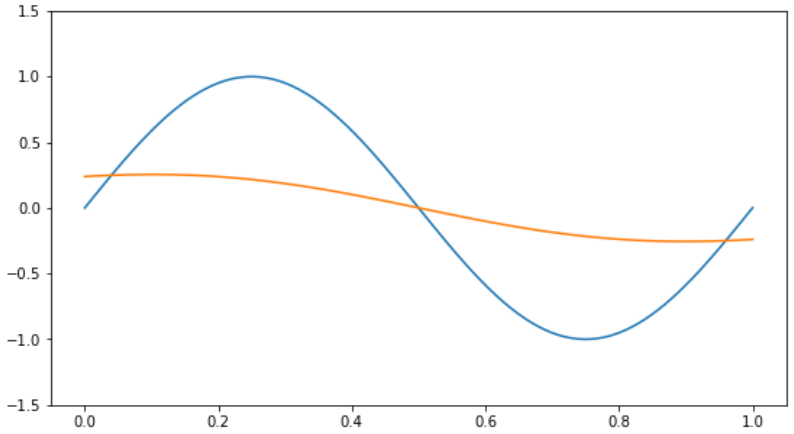
  - 正则化通过限制有效模的型复杂度，使得复杂模型能够在有限规模的数据集上进行训练而不会出现严重的过拟合（Overfitting）。
  - 注意：确定最优模型的复杂度的问题，在引入正则项后，则从选择合适数量的基函数转化为确定正则化系数 λ 的适宜值。

不同正则项控制系数λ下的预测函数表现

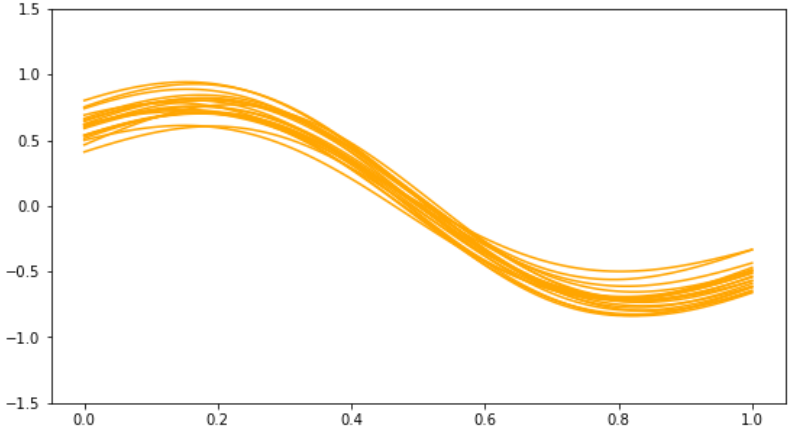
- **生成100个数据集，每个包含25个采样；基准数据生成函数$h(x) = sin(2\pi x)$；25个高斯基函数分量**

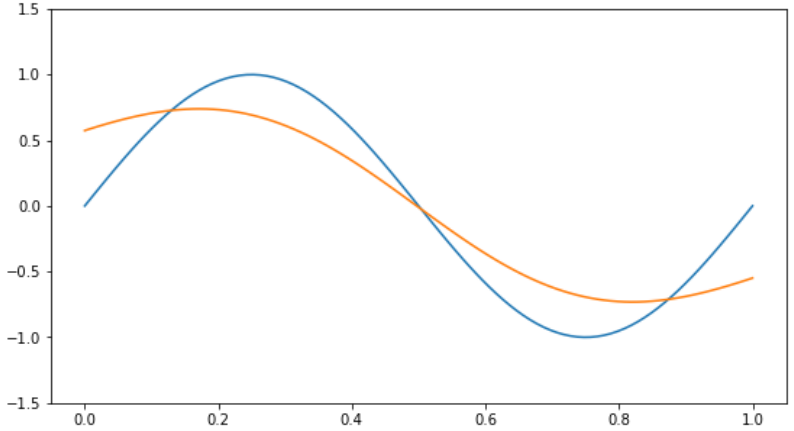
左侧：100次预测采样的结果





$\lambda = 100$;
黄色：预测值平均
蓝色：生成函数





$\lambda = 1$

一般的正则项构建方法

- **在SSE损失函数基础上添加正则化项：$E_D(\mathbf{w}) + \lambda E_W(\mathbf{w})$**
- **可以在l_2-Norm正则化项基础上构建一般的正则项表达式：**

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

  - 注意：q=2代表二项式（l_2-Norm）正则项；
  - q=1称为Lasso正则项（w_j=0处不再是连续可微函数）。当λ足够大时，若对应的基函数分量不起作用时（偏导接近0），w_j将被强制衰减到0，由此得到稀疏化模型，即对应的bias不再起作用。
  - 正则项的引入相当于基于SSE的损失函数搜索域被限制在由以下约束函数（Constraint）决定的子空间内：

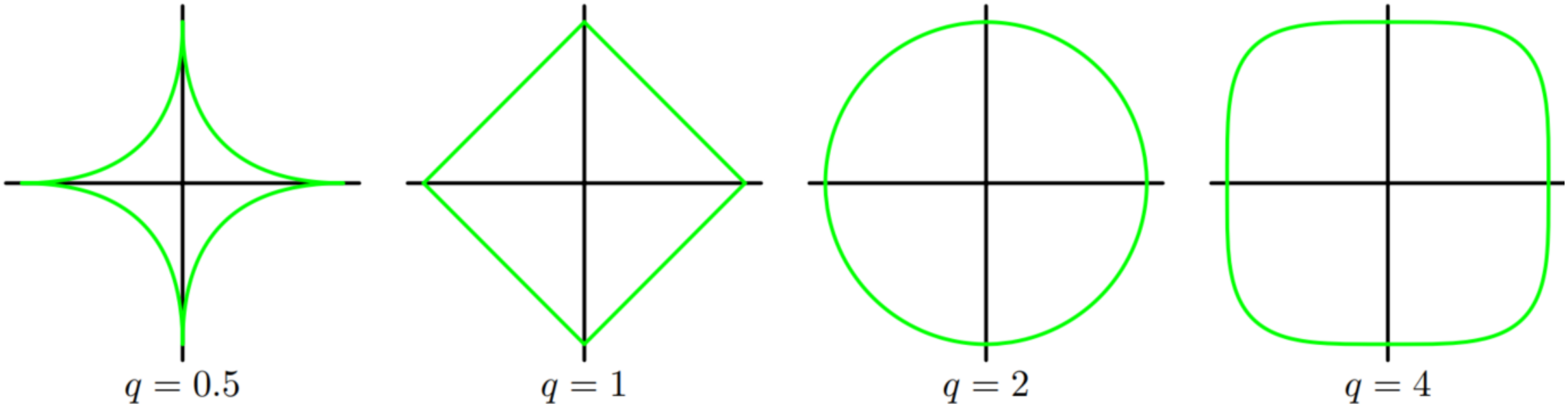
$$\sum_{j=1}^{M}|w_j|^q \leqslant \eta$$

则上式中，λ相当于拉格朗日乘子（Lagrange multiplier）

  - 注意：由于Lasso正则项不再连续可微，优化需使用次梯度下降（Sub-gradient）或坐标下降等特殊方法。

不同正则项结构系数q下的梯度等高包线（Contour）表现

- **q=0.5, q= 1, q= 2, q= 4:**

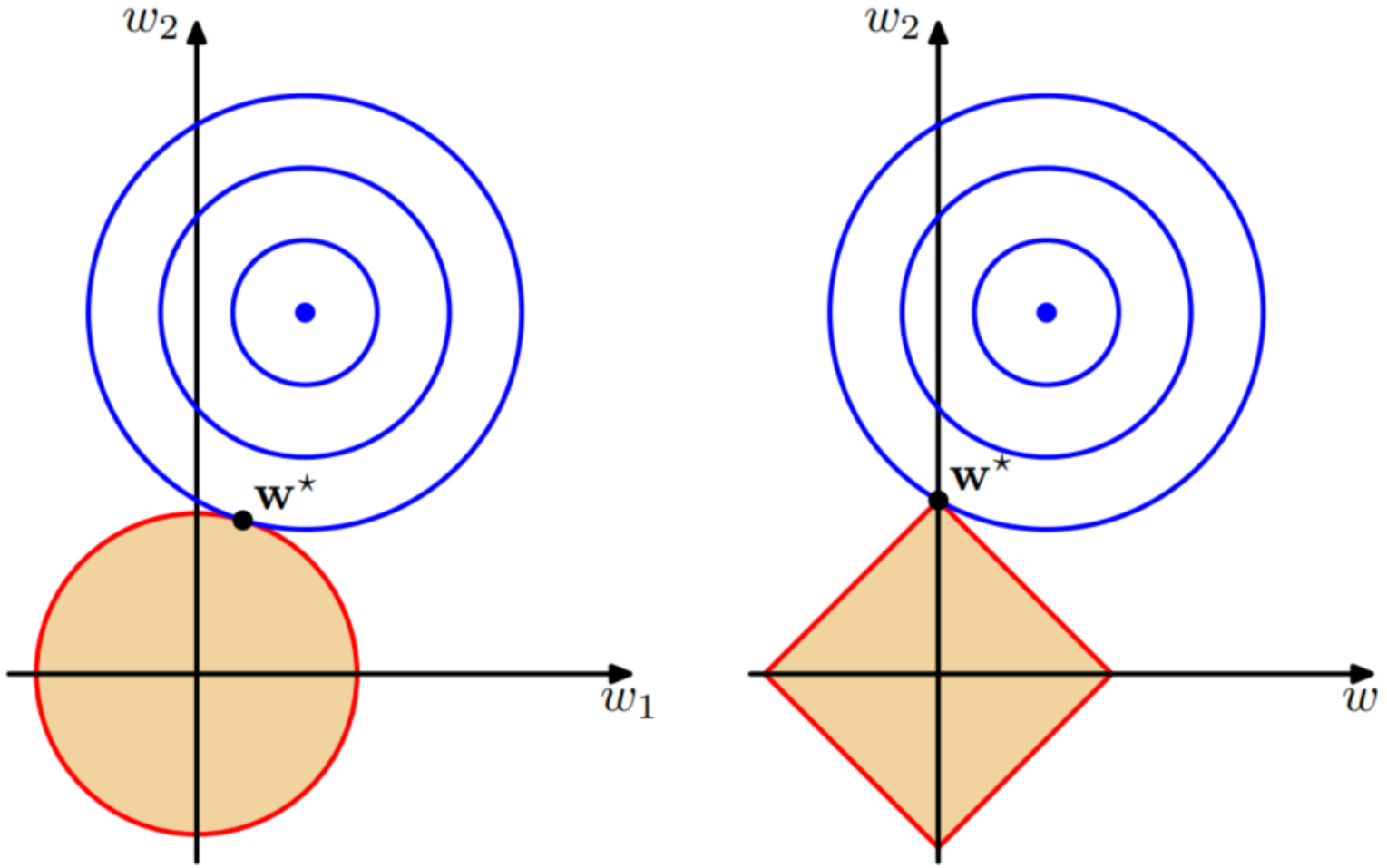


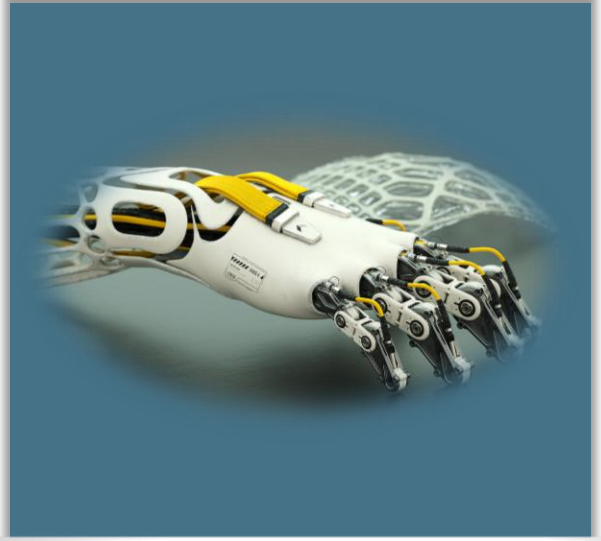
正则项作为硬约束条件对最优解搜索的影响

- **q=2（左） 对比 q=1（右）。**
- **约束条件:**

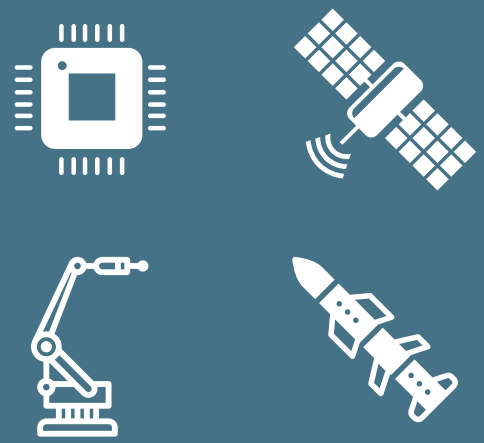
$$\sum_{j=1}^{M} |w_j|^q \leqslant \eta$$

  - 对于 $Lasso$ 正则项，有 $w_1 = 0$。





统计机器学习中的回归问题



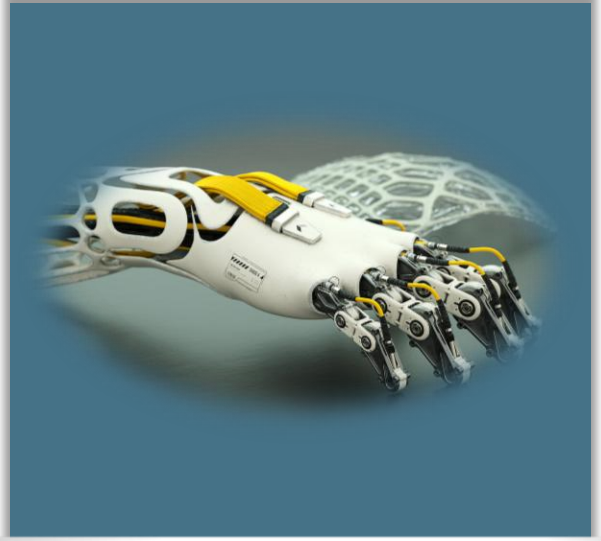
1. 线性回归中的最小二乘法
2. 贝叶斯线性回归
3. 线性回归中的正则化（Regularization）
4. 模型的质量指标

回归模型评估指标

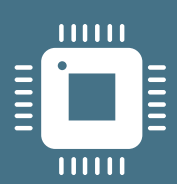
- **MSE：均方误差，Mean Square Error**
  - $\mathrm{MSE} = \frac{1}{N}\sum_{n=1}^{N}(y(\mathbf{x}_n) - t_n)^2$
  - 同样数据集下，MSE越小，模型效果越好；适用于梯度下降优化方法。
  - （变种）RMSE（Root Mean Square Error）： $\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{n=1}^{N}(y(x_n) - t_n)^2}$
- **MAPE：平均绝对百分比误差（Mean Absolute Percentage Error）：** $\mathrm{MAPE} = \frac{100}{N}\sum_{n=1}^{N}\left|\frac{t_n - y(\mathbf{x}_n)}{t_n}\right|$
- **MAE：平均绝对误差（Mean Absolute Error）：** $\mathrm{MSE} = \frac{1}{N}\sum_{n=1}^{N}|y(\mathbf{x}_n) - t_n|$
  - 对比：MAPE对零值敏感（不适用于目标变量含零值的数据）；
  - MAE对异常值不敏感（线性惩罚），鲁棒性强。

回归模型评估指标（续）

- **R-Square：决定系数**
  - $R^2 = \frac{SS_{res}}{SS_{tot}} = 1 - \frac{\sum_{n=1}^{N}(y(x_n)-t_n)^2}{\sum_{n=1}^{N}(t_n-E(t_n))^2}$
  - SS_{res}：预测误差平方和；
  - SS_{tot}：数据总方差（衡量原始数据的离散程度）；
  - 对异常值（outlier）不敏感，但随无关变量增加而虚高（需Adjusted R^2校正）。
- **Adjusted R^2：校正决定系数**
  - $R^2_{adj} = 1 - \frac{(n-1)(n-R^2)}{n-p-1}$
  - 其中*n*为样本数量，*p*为特征数量。
  - 校正目的：惩罚无关变量，即*p*（特征数）增加时降低评分。 p/n<0.05 时校正效果显著。



统计机器学习中的回归问题




讨论