第 14 章 正则化

14.1 正则化假设集

命题 14.1.1 (正则化)

设 Q 阶多项式特征映射为:

$$\phi_Q(x) = \left(1,\, x,\, x^2, \ldots, x^Q\right) \in \mathbb{R}^{Q+1}$$

对应假设集为:

$$\mathcal{H}_Q = \left\{h(x) = w^\top \phi_Q(x) \mid w \in \mathbb{R}^{Q+1}\right\}$$

约束 从 $\mathcal{H}_{10}$ 退到 $\mathcal{H}_2$ 等价于在 $\mathcal{H}_{10}$ 中施加线性约束:

$$w_3 = w_4 = \cdots = w_{10} = 0$$

带约束的回归

$$\mathcal{H}_{10}\ \text{回归}:\quad \min_{w\in\mathbb{R}^{11}} E_{\text{in}}(w)$$

$$\mathcal{H}_{2}\ \text{回归}:\quad \min_{\substack{w\in\mathbb{R}^{11}\\ w_3=\cdots=w_{10}=0}} E_{\text{in}}(w)$$

注 为何不直接令 $w \in \mathbb{R}^3$？因为显式约束形式为后续更灵活的正则化（如 L_2、L_1 惩罚）奠定统一框架。

命题 14.1.2 (从硬约束到软约束的正则化)

设多项式特征映射后的权重维度为 $d = 10$，记

$$\phi(x) = \left(1, x, x^2, \ldots, x^{10}\right), \quad w = (w_0, w_1, \ldots, w_{10}) \in \mathbb{R}^{11}$$

(a) 稀疏硬约束:

$$\mathcal{H}_2^{\text{hard}} = \left\{w \in \mathbb{R}^{11} \;\middle|\; w_3 = w_4 = \cdots = w_{10} = 0\right\}$$

优化问题为:

$$\min_{w\in\mathbb{R}^{11}} E_{\text{in}}(w) \quad \text{s.t. } w_3 = \cdots = w_{10} = 0$$

(b) 宽松稀疏约束:

$$\mathcal{H}_2^{\text{sparse}} = \left\{w \in \mathbb{R}^{11} \mid \|w\|_0 \le 3\right\}$$

其中 $\|w\|_0$ 为非零权重个数。该约束更灵活: $\mathcal{H}_2^{\text{hard}} \subset \mathcal{H}_2^{\text{sparse}} \subset \mathcal{H}_{10}$，但求解 $\min E_{\text{in}}$ 在 $\mathcal{H}_2^{\text{sparse}}$ 下为 NP-hard。

(c) 软约束（L_2 正则化）:

$$\mathcal{H}(C) = \left\{w \in \mathbb{R}^{11} \mid \|w\|_2^2 \le C\right\}, \quad C \ge 0$$

优化问题为:

$$\min_{w\in\mathbb{R}^{11}} E_{\text{in}}(w) \quad \text{s.t. } \|w\|_2^2 \le C$$

该集合对 C 连续嵌套：

$$\mathcal{H}(0) \subset \mathcal{H}(1.126) \subset \cdots \subset \mathcal{H}(1126) \subset \cdots \subset \mathcal{H}(\infty) = \mathcal{H}_{10}$$

结论：软约束 $\mathcal{H}(C)$ 既保留模型容量，又通过连续调节 C 实现平滑正则化，从而避免 NP-hard 求解，成为实际中最常用的正则化框架。

例题 14.1 选择题：正则化假设集的成员判定

对于 $Q \geq 1$，下列哪个假设（权重向量 $\boldsymbol{w} \in \mathbb{R}^{Q+1}$）不属于正则化假设集 $\mathcal{H}(1)$？

1) $\boldsymbol{w}^T = [0, 0, \ldots, 0]$
2) $\boldsymbol{w}^T = [1, 0, \ldots, 0]$
3) $\boldsymbol{w}^T = [1, 1, \ldots, 1]$
4) $\boldsymbol{w}^T = \left[\sqrt{\frac{1}{Q+1}}, \sqrt{\frac{1}{Q+1}}, \ldots, \sqrt{\frac{1}{Q+1}}\right]$

解答 正确选项为 $\boxed{3}$。正则化假设集 $\mathcal{H}(1)$ 要求权重向量的 L_2 范数满足 $\|\boldsymbol{w}\|_2 \leq 1$。

- 选项 1：$\|\boldsymbol{w}\|_2 = 0 \leq 1$，满足条件。
- 选项 2：$\|\boldsymbol{w}\|_2 = 1 \leq 1$，满足条件。
- 选项 3：$\|\boldsymbol{w}\|_2 = \sqrt{Q+1} > 1$（因 $Q \geq 1$），不满足条件。
- 选项 4：$\|\boldsymbol{w}\|_2 = 1 \leq 1$，满足条件。 ■

14.2 权重衰减正则化

命题 14.2.1 (正则化回归的矩阵形式与拉格朗日乘子法)

矩阵形式给定特征矩阵 $Z \in \mathbb{R}^{N \times (Q+1)}$ 和标签向量 $y \in \mathbb{R}^N$，正则化回归问题可表示为：

$$\min_{w \in \mathbb{R}^{Q+1}} E_{\text{in}}(w) = \frac{1}{N}(Zw - y)^\top (Zw - y) \quad \text{s.t.} \quad w^\top w \leq C$$

其中 $w^\top w \leq C$ 表示可行域为半径 $\sqrt{C}$ 的超球面。

拉格朗日乘子法构造拉格朗日函数：

$$\mathcal{L}(w, \lambda) = \frac{1}{N}(Zw - y)^\top (Zw - y) + \lambda(w^\top w - C), \quad \lambda \geq 0$$

最优解 w_{REG} 满足 KKT 条件：

$$\nabla_w \mathcal{L}(w, \lambda) = \frac{2}{N} Z^\top (Zw - y) + 2\lambda w = 0$$

解得：

$$w_{\text{REG}} = (Z^\top Z + N\lambda I)^{-1} Z^\top y$$

且互补松弛条件给出：

$$\lambda(w_{\text{REG}}^\top w_{\text{REG}} - C) = 0$$

几何解读 在 $w^\top w = C$ 边界上，负梯度 $-\nabla E_{\text{in}}(w_{\text{REG}})$ 与法向量 w_{REG} 平行，保证在不违反约束的前提下无法再降低 E_{in}。

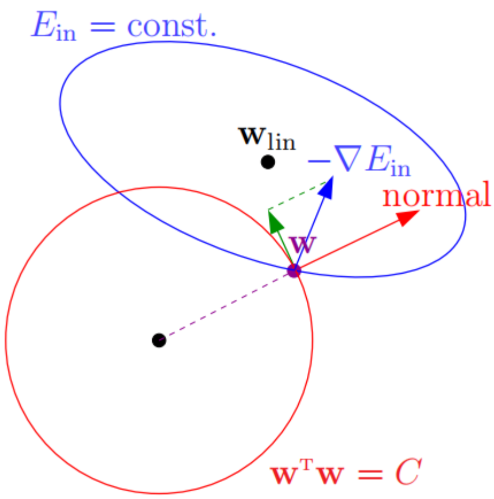


图 14.2.1: 带约束的样本内误差最小化几何解释图

命题 14.2.2 (增广误差与岭回归的等价性)

增广误差的定义设正则化参数 $\lambda > 0$ 已知（可由先验知识或交叉验证确定），定义增广误差：

$$E_{\text{aug}}(w) = \underbrace{\frac{1}{N}\|Zw - y\|_2^2}_{E_{\text{in}}(w)} + \underbrace{\frac{\lambda}{N}\|w\|_2^2}_{\text{正则化项}}$$

其中 $\|\cdot\|_2$ 表示欧几里得范数。

无约束优化解　对增广误差求导并令导数为零：

$$\nabla_w E_{\text{aug}}(w) = \frac{2}{N}(Z^\top Zw - Z^\top y) + \frac{2\lambda}{N}w = 0$$

解得：

$$w_{\text{REG}} = \left(Z^\top Z + \lambda I\right)^{-1} Z^\top y$$

该式即为统计学中的 岭回归（**ridge regression**）。

与约束优化的等价性　最小化无约束增广误差 $E_{\text{aug}}(w)$ 等价于求解以下带约束优化问题：

$$\min_w \frac{1}{N}\|Zw - y\|_2^2 \quad \text{s.t.} \quad \|w\|_2^2 \le C$$

其中 C 是与 λ 相关的常数，二者通过拉格朗日对偶关系一一对应。

命题 14.2.3 (正则化效果)

固定特征映射与数据集，仅改变正则化参数 λ：

- $\lambda = 0$：无正则化，易过拟合；
- $\lambda = 0.0001$：轻微正则，拟合与泛化平衡；
- $\lambda = 0.01$：适中正则，进一步抑制过拟合；
- $\lambda = 1$：强正则，可能欠拟合。

权重衰减（**weight-decay**）　把 $\frac{\lambda}{N}\|w\|^2$ 加进损失函数，等价于

$$\text{更大的}\lambda \iff \text{偏好更短的}w \iff \text{有效缩小约束半径}C.$$

结论：任意非线性变换 + 线性模型 + 权重衰减，即可稳健地控制复杂度与泛化性能。

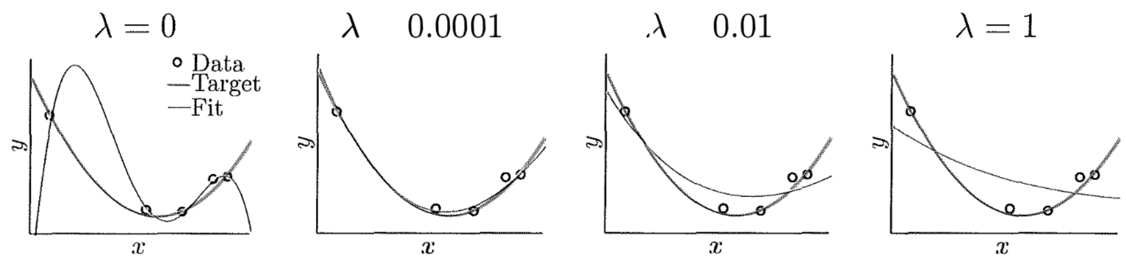


图 14.2.2: 不同的正则化参数值的权重衰减

定义 14.2.1 (多项式特征映射：从朴素到归一化)

给定输入变量 $x \in [-1,1]$，常用的两种多项式特征构造方式：

1) 朴素多项式映射

$$\phi_{\text{naive}}(x) = \left(1,\, x,\, x^2, \ldots, x^Q\right).$$

当 $x \in [-1,1]$ 时，$|x^q|$ 随阶数 q 增大而迅速减小，导致需要极大的权重 w_q 才能产生显著影响。

2) 归一化多项式映射（**Legendre** 正交基）采用勒让德多项式 $\{L_q(x)\}_{q=0}^{Q}$ 作为正交归一基函数：

$$\phi_{\text{norm}}(x) = \left(L_0(x),\, L_1(x), \ldots, L_Q(x)\right),$$

其中

$$\begin{aligned}
L_0(x) &= 1, \\
L_1(x) &= x, \\
L_2(x) &= \tfrac{1}{2}(3x^2 - 1), \\
L_3(x) &= \tfrac{1}{2}(5x^3 - 3x), \\
L_4(x) &= \tfrac{1}{8}(35x^4 - 30x^2 + 3), \\
L_5(x) &= \tfrac{1}{8}(63x^5 - 70x^3 + 15x), \ \ldots
\end{aligned}$$

由于 $\{L_q(x)\}$ 在区间 $[-1,1]$ 上正交归一，数值稳定性高，可显著抑制所需权重 w_q 的幅度，从而缓解数值病态与过拟合。♣

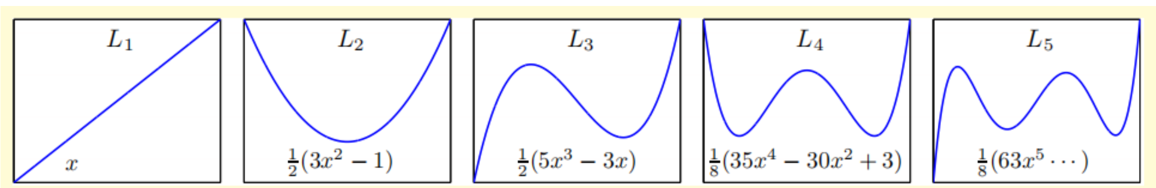


图 14.2.3: Legendre 正交基

例题 14.2 选择题：正则化权重与普通权重的等价条件

在什么情况下正则化权重 $\boldsymbol{w}_{\text{REG}}$ 等于普通最小二乘权重 $\boldsymbol{w}_{\text{LIN}}$？

1) $\lambda = 0$
2) $C = \infty$

3) $C \geq \|\boldsymbol{w}_{\mathrm{LIN}}\|^2$

4) 以上所有情况

解答 正确选项为 $\boxed{4}$。正则化权重 $\boldsymbol{w}_{\mathrm{REG}}$ 由目标函数 $\sum(y_n - \boldsymbol{w}^T\boldsymbol{x}_n)^2 + \lambda\|\boldsymbol{w}\|^2$ 最小化得到，普通权重 $\boldsymbol{w}_{\mathrm{LIN}}$ 是 $\lambda = 0$ 时的特例。

- 选项 1：$\lambda = 0$ 时，正则化退化为普通最小二乘，$\boldsymbol{w}_{\mathrm{REG}} = \boldsymbol{w}_{\mathrm{LIN}}$。
- 选项 2：$C = \infty$ 时，正则化约束被忽略，等价于 $\lambda = 0$。
- 选项 3：$C \geq \|\boldsymbol{w}_{\mathrm{LIN}}\|^2$ 时，正则化约束不限制 $\boldsymbol{w}_{\mathrm{LIN}}$。

所有选项均满足条件。 ■

14.3 正则化与 VC 理论

命题 14.3.1 (正则化与 VC 理论的统一视角)

1. 受约束优化与 **VC** 界考虑权重范数约束下的经验风险最小化问题：

$$\min_w E_{\mathrm{in}}(w) \quad \text{s.t.} \quad w^\top w \leq C$$

根据 VC 理论，其泛化误差满足：

$$E_{\mathrm{out}}(w) \leq E_{\mathrm{in}}(w) + \Omega\big(\mathcal{H}(C)\big)$$

其中 $\mathcal{H}(C) = \{h \mid h(x) = w^\top\phi(x),\ w^\top w \leq C\}$，且约束半径 C 与某个拉格朗日乘子 λ 一一对应。

2. 增广误差的再阐释将约束转化为无约束正则化问题：

$$\min_w E_{\mathrm{aug}}(w) = E_{\mathrm{in}}(w) + \frac{\lambda}{N} w^\top w$$

此时：

- 正则项 $w^\top w$：度量单个假设的复杂度；
- $\Omega(\mathcal{H})$：度量整个假设集的复杂度。

若 $\frac{\lambda}{N} w^\top w$ 能良好地近似 $\Omega(\mathcal{H})$，则 E_{aug} 比 E_{in} 更接近 E_{out}，从而在理论上享有整个假设集 $\mathcal{H}$ 的灵活性，同时实践中利用正则化避免过拟合。

3. 有效 **VC** 维

- 名义 VC 维：$d_{\mathrm{VC}}(\mathcal{H}) = d + 1$，因优化时" 理论上" 遍历所有 w；
- 有效 VC 维：$d_{\mathrm{EFF}}(\mathcal{H}, \lambda)$，对应实际受正则化约束的假设子集 $\mathcal{H}(C)$；
- 当 λ 较大时，$d_{\mathrm{EFF}}(\mathcal{H}, \lambda) \ll d_{\mathrm{VC}}(\mathcal{H})$，从而解释正则化如何在高名义复杂度下保持低有效复杂度。

结论：

$$\text{正则化} = \text{用较小的} d_{\mathrm{EFF}} \text{ 控制较大的} d_{\mathrm{VC}}$$

♠

14.4 通用正则化项

定义 14.4.1 (通用正则项 $\Omega(w)$ 的设计准则)

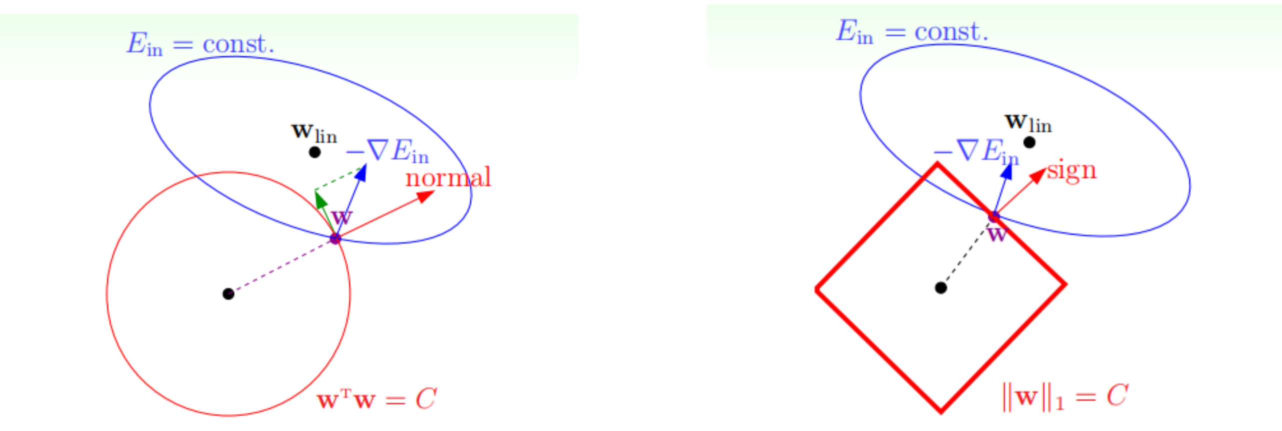
正则项 $\Omega(w)$ 的选择应兼顾以下三类动机，并可与误差函数 E_{err} 共同构成增广误差

$$E_{\text{aug}}(w) = E_{\text{err}}(w) + \lambda\,\Omega(w).$$

1. 目标依赖型（target-dependent）
   若对真实目标函数的先验已知，可构造指向该先验的约束方向。例如：
   - 对称正则：当目标函数具有奇偶对称性时，强制对应奇/偶系数 w_q 为零；
   - 平滑正则：鼓励高频系数 w_q 趋零，以逼近更平滑或更简单的真实函数。
2. 合理性型（plausible）
   对随机或确定性噪声均呈现非光滑特性时，选用
   - 稀疏正则 (L_1)：$\Omega(w) = \sum_q |w_q|$，诱导稀疏解；
   - 权重衰减 (L_2)：$\Omega(w) = \sum_q w_q^2$，抑制大权重。
3. 友好型（friendly）
   正则项本身需易于优化：
   - L_2 正则：凸、可微、闭式解；
   - L_1 正则：虽非光滑，但凸且可高效求解（如坐标下降、近端梯度）。

结论：

正则项 $\leftrightarrow$ 误差度量：均遵循“目标依赖 / 合理 / 友好”的相同设计哲学。♣



(a) L_2 正则化　　(b) L_1 正则化

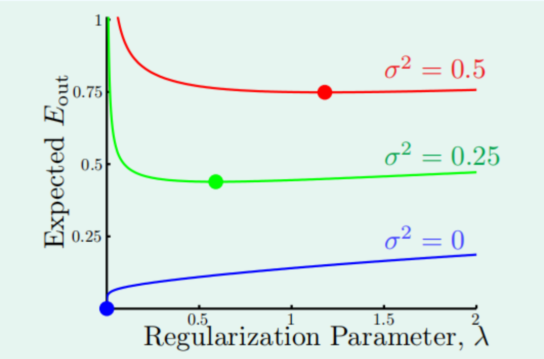
图 14.4.1: 正则化图像

命题 14.4.1 (最优正则化参数 λ 的选择)

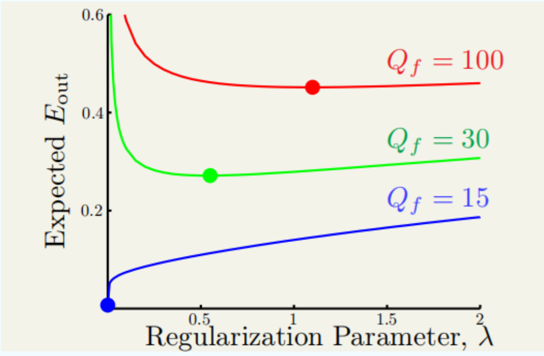
给定真实目标复杂度 Q_f（多项式阶数）与两类噪声：

- 随机噪声（stochastic noise）；
- 确定性噪声（deterministic noise）。

数值实验结果（固定 N）如图所示：



(a) 不同噪声水平下期望样本外误差与正则化参数关系图



(b) 不同目标复杂度下期望样本外误差与正则化参数关系图

图14.4：正则化图像

结论

- 噪声越大 $\Leftrightarrow$ 需要更大的 λ；
- 噪声未知时，必须通过交叉验证等方法选择恰当 λ，否则易过拟合或欠拟合。

14.5 总结

笔记 [正则化]

- 正则化假设集：在原始假设空间 $\mathcal{H}$ 上施加约束。
- 权重衰减正则化：在目标函数中增加 $\lambda\Omega(w)$，得到增广误差 $E_{\text{aug}}(w) = E_{\text{in}}(w) + \lambda\Omega(w)$。
- 正则化与 VC 理论：正则化通过减少有效参数降低复杂度 C，从而减小有效 VC 维 d_{eff}。
- 通用正则化项：可依据“目标相关”“合理”或“算法友好”等原则构造。