第 13 章 过拟合的危害

13.1 什么是过拟合？

定义 13.1.1 (坏的泛化与过拟合)

当训练集大小 N 固定时，模型复杂度可用 VC 维 d_{VC} 表征。泛化能力由内外误差之差 $E_{\mathrm{out}} - E_{\mathrm{in}}$ 衡量，该差值应尽可能小。

1) 过拟合（Overfitting）：当将 d_{VC} 从最优值 d_{VC}^{*} 提高到极高维度（如 1126）时，

$$E_{\mathrm{in}} \downarrow, \quad E_{\mathrm{out}} \uparrow, \quad \Longrightarrow \quad E_{\mathrm{out}} - E_{\mathrm{in}} \text{ 显著增大.}$$

模型对训练数据拟合更好，但泛化能力恶化。

2) 欠拟合（Underfitting）：当将 d_{VC} 降至过低时，模型表达能力不足，难以捕捉数据规律，

$$E_{\mathrm{in}} \uparrow, \quad E_{\mathrm{out}} \uparrow .$$

内外误差均增大，表现出整体性能低下。

♣

命题 13.1.1 (过拟合的“驾驶”类比)

将机器学习过程类比为驾驶：

- 数据 $\leftrightarrow$ 观察到的路况；
- 目标函数 $\leftrightarrow$ 理想行驶路线；
- 模型容量（d_{VC}）$\leftrightarrow$ 车速。

则

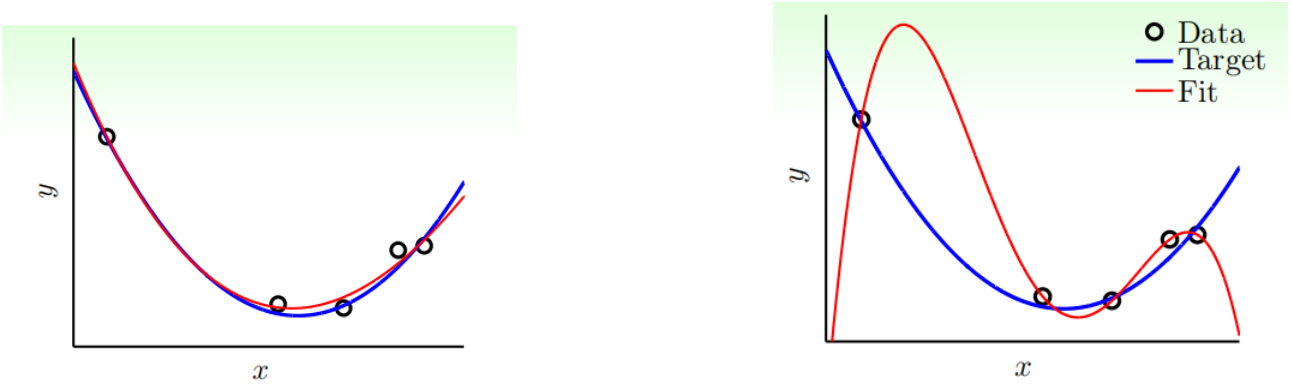
(a) 车速过高（d_{VC} 过大）$\Rightarrow$ 对路面微小起伏过度反应，即过拟合；

(b) 路面颠簸（数据噪声）$\Rightarrow$ 进一步放大失控风险；

(c) 观察窗口有限（样本量 N 小）$\Rightarrow$ 无法充分了解路况，同样导致偏离。

因此，噪声与数据规模共同决定了“失控”即过拟合发生的概率。

♠



(a) 好拟合图像　　(b) 过拟合图像

图 13.1.1: 机器学习拟合曲线图像

13.2 噪声与数据量的作用

命题 13.2.1 (过拟合的案例研究)

考虑两种不同的目标函数场景：

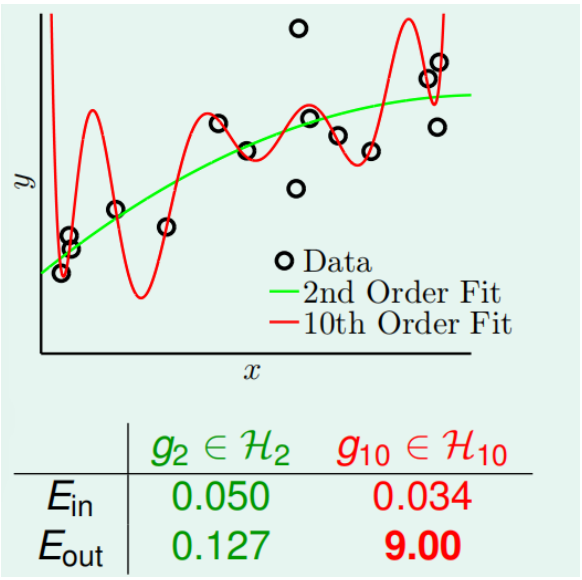
(a) 场景一：含噪声的 **10** 次目标函数　给定含噪声的 10 次目标函数，分别在 2 次多项式空间 $\mathcal{H}_2$ 中挑选假设 g_2，和在 10 次多项式空间 $\mathcal{H}_{10}$ 中挑选假设 g_{10}。得到如下结果：

	$g_2 \in \mathcal{H}_2$	$g_{10} \in \mathcal{H}_{10}$
E_{in}	0.050	0.034
E_{out}	0.127	9.00

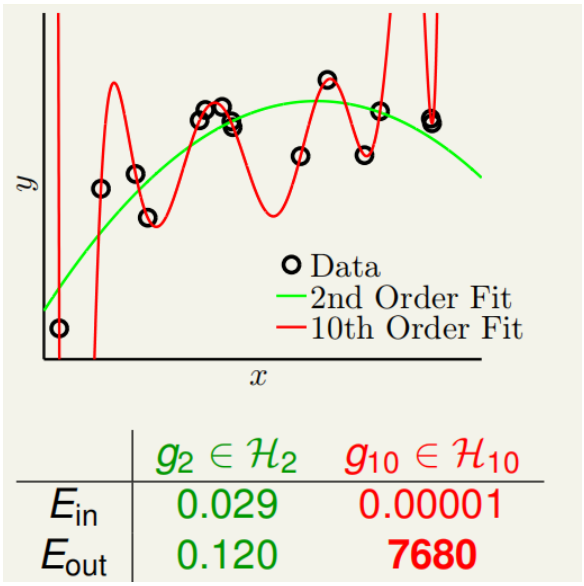
(b) 场景二：无噪声的 **50** 次目标函数　给定无噪声的 50 次目标函数，同样分别在 2 次多项式空间 $\mathcal{H}_2$ 中挑选假设 g_2，和在 10 次多项式空间 $\mathcal{H}_{10}$ 中挑选假设 g_{10}。得到如下结果：

	$g_2 \in \mathcal{H}_2$	$g_{10} \in \mathcal{H}_{10}$
E_{in}	0.029	0.00001
E_{out}	0.120	7680

在这两种场景中，从 g_2 到 g_{10} 均出现了过拟合现象，即模型在训练集上的误差 E_{in} 减小，但在测试集上的误差 E_{out} 显著增大。



(a) 十阶目标函数（有噪声）



(b) 五十阶目标函数（无噪声）

图 13.2.1: 二阶拟合与十阶拟合对比图

命题 13.2.2 (过拟合与模型选择的“讽刺”)

给定同一组数据，真实目标函数为十次多项式。

- 学习者 **Overfit**（过拟合者）：在十次多项式假设空间 $\mathcal{H}_{10}$ 中选择假设函数 g_{10}；
- 学习者 **Restrict**（受限者）：仅在二次多项式假设空间 $\mathcal{H}_2$ 中选择假设函数 g_2。

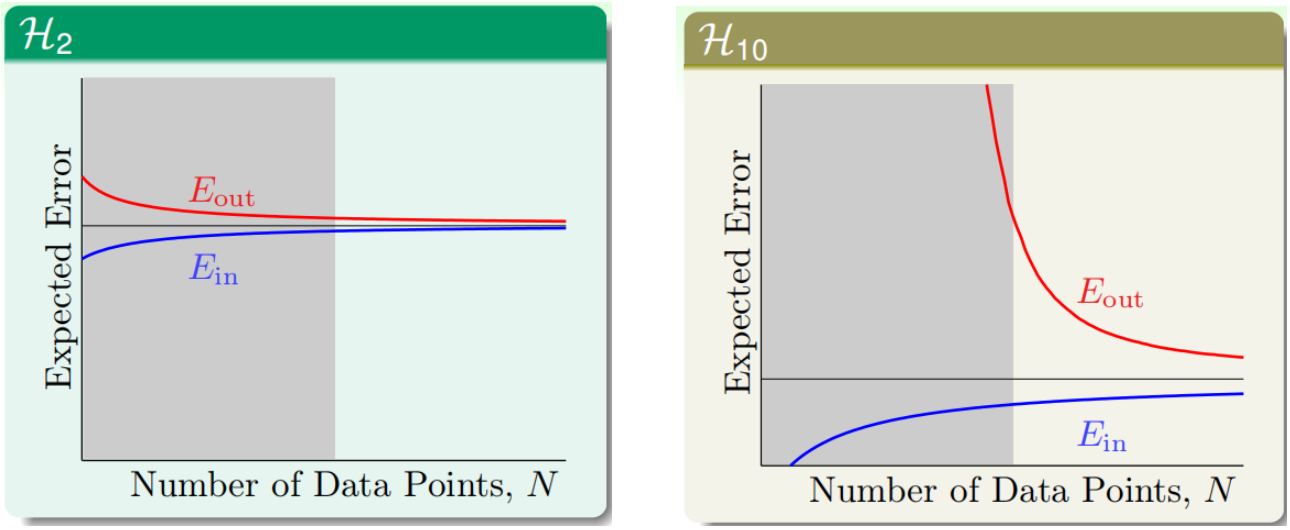
虽然两者都知道目标是十次多项式，学习者 Restrict 却主动放弃高阶拟合能力，结果是：

$$E_{\text{out}}(g_2) \ll E_{\text{out}}(g_{10}).$$

这一现象可以从学习曲线中得到进一步验证：

- 对于 $\mathcal{H}_{10}$，当数据点数量 N 趋近于无穷大时，预期的 E_{out} 较低，但在 N 较小时，泛化误差会大得多。
- 在数据点数量 N 较小的灰色区域，$\mathcal{H}_{10}$ 会出现过拟合现象，表现为 E_{in} 下降而 E_{out} 上升。

即使在无噪声的情况下，学习者 Restrict 仍然能够在 E_{out} 上取得优势，这是因为目标函数的复杂性本身就起到了类似噪声的作用。



(a) 假设空间 $\mathcal{H}_2$ 下期望误差与数据点数量关系图　(b) 假设空间 $\mathcal{H}_{10}$ 下期望误差与数据点数量关系图

图 13.2.2: 不同假设空间下期望误差与数据点数量关系图

13.3 确定性噪声

命题 13.3.1 (过拟合度量实验)

设目标函数 $f \in \mathcal{H}_{Q_f}$，其形式为 $f(x) = \sum_{q=0}^{Q_f} \alpha_q x^q$，其中 α_q 为系数。观测样本由以下方式生成：

$$y_n = f(x_n) + \varepsilon_n, \quad \varepsilon_n \sim \mathcal{N}(0, \sigma^2)$$

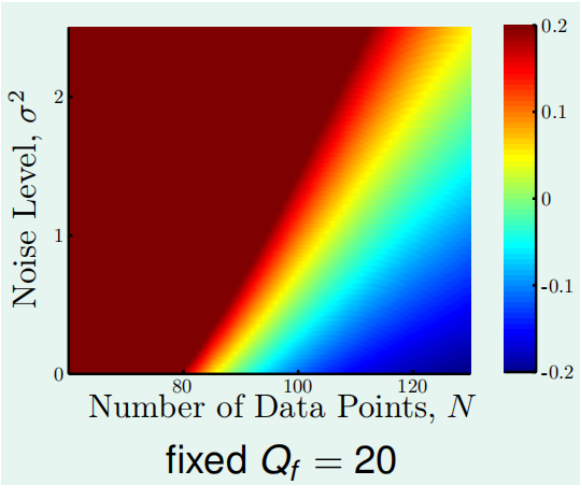
其中 x_n 在定义域上均匀分布，ε_n 为独立同分布的高斯噪声，噪声水平为 σ^2。

记 $g_2 \in \mathcal{H}_2$ 与 $g_{10} \in \mathcal{H}_{10}$ 分别为通过二次多项式拟合和十次多项式拟合得到的假设函数。过拟合程度可量化为：

$$\text{Overfit}(g_2, g_{10}) \triangleq E_{\text{out}}(g_{10}) - E_{\text{out}}(g_2)$$

并且恒有 $E_{\text{in}}(g_{10}) \leq E_{\text{in}}(g_2)$。

该过拟合度量值 $\text{Overfit}(g_2, g_{10})$ 随噪声水平 σ^2 的增大或样本量 N 的减小而单调上升。



(a) 固定 Q_f 时噪声水平与数据点数量关系热图　(b) 固定 σ^2 时目标复杂度与数据点数量关系热图

图 13.3.1: 目标复杂度或噪声水平与数据点数量关系热图

命题 13.3.2 (噪声与数据量对过拟合的影响)

设目标函数 $f \in \mathcal{H}_{Q_f}$，样本数为 N，则过拟合程度满足：

$$\text{Overfit}(g) \propto \frac{\sigma^2}{N} + \frac{\text{DetNoise}}{N}$$

其中 σ^2 为随机噪声强度，DetNoise 为确定性噪声 (deterministic noise)。

经验规律

- 固定 Q_f：数据量 N 减小或随机噪声 σ^2 增大 $\Rightarrow$ 过拟合程度上升；
- 固定 N：真实复杂度 Q_f 增大 $\Rightarrow$ 确定性噪声增大 $\Rightarrow$ 过拟合程度上升。

确定性噪声定义　若 $f \notin \mathcal{H}$，则存在：

$$\text{DetNoise}(x) = f(x) - h^*(x), \quad h^* = \arg\min_{h \in \mathcal{H}} E_{\text{out}}(h)$$

与随机噪声不同，确定性噪声具有以下特点：

- 由假设空间 $\mathcal{H}$ 决定；
- 对给定的 x 固定不变。

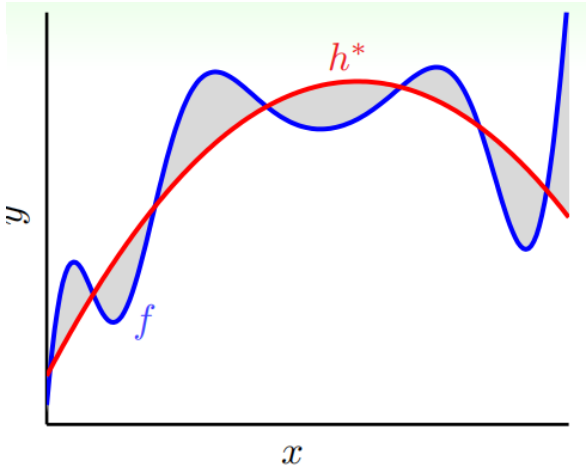


图 13.3.2: 确定性噪声图片

例题 13.1 选择题：确定性噪声的计算

考虑目标函数为 $\sin(1126x)$，$x \in [0, 2\pi]$，且 x 在该区间均匀采样。若用线性假设 $h(x) = w \cdot x$ 以平方误差近似目标函数，则每个 x 处的确定性噪声为：

1) $|\sin(1126x)|$
2) $|\sin(1126x) - x|$
3) $|\sin(1126x) + x|$
4) $|\sin(1126x) - 1126x|$

解答 正确选项为 $\boxed{1}$。确定性噪声是目标函数与最优近似函数的差异。对线性假设 $h(x) = w \cdot x$，通过尝试不同 w 值可知，当 $w = 0$ 时，$h^*(x) = 0$ 是最优假设（平方误差最小）。

因此，确定性噪声为 $|\sin(1126x) - h^*(x)| = |\sin(1126x)|$。 ■

13.4 应对过拟合

命题 13.4.1 (驾驶类比再探：防治过拟合的实用技术)

将机器学习过程类比为驾驶行为：

- 过拟合 $\leftrightarrow$ 发生车祸；
- 过度 **VC** 维（d_{VC}） $\leftrightarrow$ 驾驶速度过快；
- 噪声 $\leftrightarrow$ 颠簸的路面；
- 数据量有限（N） $\leftrightarrow$ 对路况的观察有限。

防治过拟合的对应策略如下：

机器学习	驾驶策略
从简单模型起步	低速驾驶
数据清洗/剪枝	使用更准确的路况信息
数据提示（hinting）	利用更多的路况信息
正则化	踩刹车
验证	监控仪表盘

这些一一对应的实用技术，都是防止“失控”——过拟合——的有效手段。 ♠

定义 13.4.1 (数据清洗 / 数据剪枝（Data Cleaning / Data Pruning）)

数据清洗（Data Cleaning）是从数据集、数据库表或记录集中检测、纠正或移除损坏、不准确、不完整、不相关或重复数据的过程，旨在将“脏数据”转换为满足质量要求的高质量可用数据。其核心任务包括：

- 一致性检查：发现并修正逻辑冲突、越界值；
- 缺失值处理：插补、删除或标记无效值；
- 异常值检测与修正：识别并替换错误或离群样本；
- 重复记录去重：合并或删除冗余条目；
- 数据类型与格式标准化：确保变量类型、单位和命名一致。

数据剪枝（Data Pruning）是数据清洗的一个子过程，专指在不损失关键信息的前提下，删除冗余、噪声或低质量样本，从而降低模型复杂度并减少过拟合风险。
清洗后的数据应满足准确性、完整性、一致性、唯一性和有效性等质量维度，为后续分析与建模奠定可靠基础。

定义 13.4.2 (数据提示（Data Hinting）)

数据提示（Data Hinting）是一种在不收集额外真实样本的前提下，通过“伪造”或“生成”虚拟样本（virtual examples）来扩充训练集规模的数据增强技术。其核心思想是：在保持标签不变的情况下，对已有数据进行合理的几何或语义变换，从而人为地引入更多有效信息，降低过拟合风险。

常见做法包括：

- 图像：平移、旋转、缩放、翻转、亮度/对比度调整；
- 文本：同义词替换、随机删除/插入词语、回译；
- 语音：加噪、变速、音高偏移；
- 数值特征：小幅随机扰动、线性插值。

注意事项

- 变换强度应受约束，避免引入错误标签；
- 对生成样本进行一致性检查，确保其符合真实数据分布。

数据提示与常规数据增强（data augmentation）在术语上常互换使用，但“hinting”更强调通过人工“提示”模型学习不变性或鲁棒性，而非单纯增加样本量。

13.5 总结

笔记 [过拟合的危害]

- 什么是过拟合?: E_{in} 降低，E_{out} 反而升高。
- 噪声与数据量的作用：过拟合“极易”发生！
- 确定性噪声：假设集 $\mathcal{H}$ 无法捕捉的部分表现为噪声。
- 应对过拟合：数据清洗/裁剪/提示等方法，以及其他策略。