

机器学习理论基础

从风险分解到局部复杂度与核方法

陆一平 北京大学北京国际数学研究中心

时间 2026 年 10 月 12 日，周一，18:40 至 20:30

地点 四元厅 225

本讲的问题。怎样用有限数据学到一个未知函数？我们从风险分解出发，先看平滑函数的逼近为什么受到维数限制，再把集中不等式逐步升级为函数类上的泛化控制。最后利用平方损失的曲率与核的几何，解释非参估计的收敛速率。证明按“为什么需要这一步、关键变换是什么、如何严格完成”的顺序展开。

阅读说明：全文保留完整证明。首次阅读可暂略 chaining 的技术细节、RKHS 局部谱计算和附录，先掌握每节的证明直觉与主定理。所需预备知识为条件期望、线性代数、基本微积分及初等概率不等式。

目录与符号

1 学习问题与风险分解	3
2 逼近与维数灾难	5
3 泛化：从一个函数的集中到整个函数类的控制	10
4 局部复杂度：让曲率参与统计误差分析	21
5 非参统计：把附近的数据平均起来	28
6 RKHS：从岭回归的对偶到函数空间几何	32
7 从核的谱衰减回到局部复杂度（课后）	41
8 把整讲串起来：分辨率、噪声与几何	44
9 讨论题与解题提示	45
A 从小凸起到非参统计下界（课后）	47
B 高斯集中不等式的完整证明（课后）	49
C 只给出半正定核，怎样恢复函数空间	50
参考文献与原课程资料	51

符号与约定

符号	含义
P, P_n	总体平均算子与训练样本平均算子。
$R(f), R_n(f)$	总体风险与经验风险；损失顺序统一为 $\ell(y, a)$ 。
f^*	平方损失下的真实回归函数 $\mathbb{E}[Y X]$ 。
$\mathcal{F}, \mathcal{H}$	候选函数类与再生核 Hilbert 空间。
$\ g\ _{L^2(P_X)}^2$	新输入上的总体平方误差。
$\ g\ _n^2$	给定设计点上的 $n^{-1} \sum_i g(x_i)^2$ 。
$N(T, d, \varepsilon)$	在指定度量 d 下覆盖 T 所需的最少球数。
$\widehat{\mathfrak{N}}_S$	本文统一使用含绝对值的经验 Rademacher 复杂度。
$w_n(r)$	半径 r 内的高斯复杂度，定义中不含噪声强度 σ 。
$\hat{\mu}_j$	经验核矩阵 K/n 的特征值。

默认 $n \geq 1, \delta \in (0, 1)$ ，尺度常数取正值；退化情形可直接处理。函数类与所用上确界均假设可测。 $a_n \lesssim b_n$ 表示存在与 n 无关的常数使 $a_n \leq Cb_n$ ， $a_n \asymp b_n$ 表示双向成立；常数是否依赖维数或平滑度会在相应结论中说明。各节重新声明其假设，特别是有界噪声与高斯噪声、固定设计与随机设计不会混用。

1 学习问题与风险分解

1.1 真正要估计的是新样本上的风险

设训练数据 $S = (Z_1, \dots, Z_n)$ 独立同分布, $Z_i = (X_i, Y_i) \sim P$ 。学习算法看到 S 后, 输出预测函数 $\hat{f} = \hat{f}_S$ 。对预先给定的函数 f , 定义总体风险与经验风险

$$R(f) = P\ell_f = \mathbb{E}[\ell(Y, f(X))], \quad R_n(f) = P_n\ell_f = \frac{1}{n} \sum_{i=1}^n \ell(Y_i, f(X_i)).$$

这里 $\ell_f(x, y) = \ell(y, f(x))$ 。在 $R(\hat{f}_S)$ 中, 期望针对一对独立于训练集的新样本 (X, Y) ; 训练集仍是随机的。因此

$$R(\hat{f}_S) = \mathbb{E}[\ell(Y, \hat{f}_S(X)) | S]$$

通常是关于 S 的随机变量。若再对训练集平均, 应明确写成 $\mathbb{E}_S R(\hat{f}_S)$ 。

证明直觉 训练误差是在做过的题目上打分; 风险是在尚未见过的题目上打分。它们之间的困难不只是用平均值近似期望, 而是答题规则本身也由这些题目选出。因此, 大数定律对一个固定函数成立, 并没有直接解决学习问题。

记 $R^* = \inf_f R(f)$ 为 Bayes risk。模型类 $\mathcal{F}$ 的逼近误差定义为

$$A(\mathcal{F}) = \inf_{f \in \mathcal{F}} R(f) - R^*.$$

只要 $\hat{f} \in \mathcal{F}$, 便有精确的恒等式

$$R(\hat{f}) - R^* = \underbrace{A(\mathcal{F})}_{\text{逼近误差}} + \underbrace{R(\hat{f}) - \inf_{f \in \mathcal{F}} R(f)}_{\text{估计误差}}. \quad (1.1)$$

两项均非负。逼近误差问的是模型类是否包含足够好的函数; 估计误差问的是有限数据能否从模型类中选出好函数。若模型类随 n 变化, 逼近误差也会通过 $\mathcal{F}_n$ 间接依赖 n 。

1.2 平方损失为何特别适合讲理论

假设 $\mathbb{E}Y^2 < \infty$, 取 $\ell(y, a) = (y - a)^2$, 并令

$$f^*(x) = \mathbb{E}[Y | X = x], \quad Y = f^*(X) + \xi, \quad \mathbb{E}[\xi | X] = 0.$$

命题 1.1 (平方风险就是函数的平方距离)。对任意 $f \in L^2(P_X)$,

$$R(f) = \mathbb{E}\xi^2 + \|f - f^*\|_{L^2(P_X)}^2, \quad R(f) - R^* = \|f - f^*\|_{L^2(P_X)}^2. \quad (1.2)$$

若 $f = \hat{f}_S$, 上述等式对每个满足可积性条件的训练集仍成立。

证明直觉 条件期望 $f^*(X)$ 是能从 X 中预测出的部分, ξ 是剩余噪声。任何只依赖 X 的预测偏差, 都与条件均值为零的噪声正交。平方损失把这种正交关系直接变成勾股恒等式。

证明. 展开平方得到

$$(Y - f(X))^2 = \xi^2 + (f(X) - f^*(X))^2 - 2\xi(f(X) - f^*(X)).$$

最后一项的期望为零, 因为

$$\mathbb{E}[\xi(f - f^*)(X)] = \mathbb{E}[(f - f^*)(X)\mathbb{E}[\xi | X]] = 0.$$

对学习得到的函数先固定 S , 然后对独立测试样本作同样计算即可。□

因此全文会区分 L^2 误差 $\|\hat{f} - f^*\|_{L^2(P_X)}$ 与平方风险 $\|\hat{f} - f^*\|_{L^2(P_X)}^2$ 。若对训练集也取平均, 均方根误差定义为 $\text{RMSE}(\hat{f}) = \{\mathbb{E}_S \|\hat{f} - f^*\|_{L^2(P_X)}^2\}^{1/2}$ 。例如前者为 $n^{-\beta/(2\beta+d)}$ 时, 后者才是 $n^{-2\beta/(2\beta+d)}$ 。

1.3 从风险分解到可以证明的上界

设算法只求近似经验最优解:

$$R_n(\tilde{f}) \leq \inf_{f \in \mathcal{F}} R_n(f) + \varepsilon_{\text{opt}}, \quad \varepsilon_{\text{opt}} \geq 0.$$

其中 ε_{opt} 是对经验最优值的非负优化误差。

命题 1.2 (逼近、统计与优化的统一上界)。令 $\Delta_n(\mathcal{F}) = \sup_{f \in \mathcal{F}} |(P - P_n)\ell_f|$ 。则

$$R(\tilde{f}) - R^* \leq A(\mathcal{F}) + 2\Delta_n(\mathcal{F}) + \varepsilon_{\text{opt}}. \quad (1.3)$$

若总体最优比较函数 $f_{\mathcal{F}}^* \in \arg \min_{f \in \mathcal{F}} R(f)$ 存在, 还有更精细的式子

$$R(\tilde{f}) - R(f_{\mathcal{F}}^*) \leq (P - P_n)(\ell_{\tilde{f}} - \ell_{f_{\mathcal{F}}^*}) + \varepsilon_{\text{opt}}. \quad (1.4)$$

证明直觉 算法只能比较训练集上的表现, 理论要比较真实表现。所以先把 $\tilde{f}$ 的风险搬到经验风险, 再利用经验最优性与比较函数交换, 最后搬回总体风险。每次搬运产生一项统计误差, 这就是常数 2 的来源。但若保留两次搬运的差, 就得到 excess loss; 局部复杂度正是从这里开始。

证明. 对任意 $f \in \mathcal{F}$,

$$\begin{aligned} R(\tilde{f}) - R(f) &= (R - R_n)(\tilde{f}) + \{R_n(\tilde{f}) - R_n(f)\} + (R_n - R)(f) \\ &\leq (P - P_n)(\ell_{\tilde{f}} - \ell_f) + \varepsilon_{\text{opt}} \\ &\leq 2\Delta_n(\mathcal{F}) + \varepsilon_{\text{opt}}. \end{aligned}$$

对 f 的总体风险取下确界, 便得到 (1.3); 这个证明不要求下确界被达到。若 $f_{\mathcal{F}}^*$ 存在, 将它代入第一行并保留 excess loss 即可。□

说明 1.3 (三个术语不能互换)。 $R(\hat{f}) - R_n(\hat{f})$ 是泛化间隙, 可以有正有负。 $R(\hat{f}) - \inf_{\mathcal{F}} R$ 是估计误差, 必定非负。 $R_n(\hat{f}) - \inf_{\mathcal{F}} R_n$ 才是这里的优化误差。特别地, 即便精确 ERM 满足 $R_n(\hat{f}) - R_n(f_{\mathcal{F}}^*) \leq 0$, 这个比较项也不应被叫作非负优化误差。

1.4 偏差与方差分解是另一条恒等式

令 $\bar{f}(x) = \mathbb{E}_S \hat{f}_S(x)$ 。在允许交换积分与期望的可积条件下，

$$\mathbb{E}_S \left\| \hat{f} - f^* \right\|_{L^2(P_X)}^2 = \left\| \bar{f} - f^* \right\|_{L^2(P_X)}^2 + \mathbb{E}_S \left\| \hat{f} - \bar{f} \right\|_{L^2(P_X)}^2. \quad (1.5)$$

证明仍是展开平方，利用 $\mathbb{E}_S(\hat{f} - \bar{f}) = 0$ 消去交叉项。这里的 bias 取决于估计器，一般不等于模型类的逼近误差 $A(\mathcal{F})$ 。正则化也会产生偏差，即便 $f^* \in \mathcal{F}$ 。原课程关于风险分解与非参回归的两讲是本节的起点[2, 4]。

2 逼近与维数灾难

2.1 立方体的角区：体积离开中央内切球

先沿着原 PPT 的图像理解高维空间[3, 第 46 页]。立方体 $C_d = [0, 1]^d$ 有 2^d 个顶点：每个坐标都可以独立选择 0 或 1。记中心为 $c = (1/2, \dots, 1/2)$ ，中央内切球为 $B_d = \{x : \|x - c\|_2 \leq 1/2\}$ 。二维图中，圆外剩下四个朝向顶点的角区；高维时，这个图像需要用体积来解释。

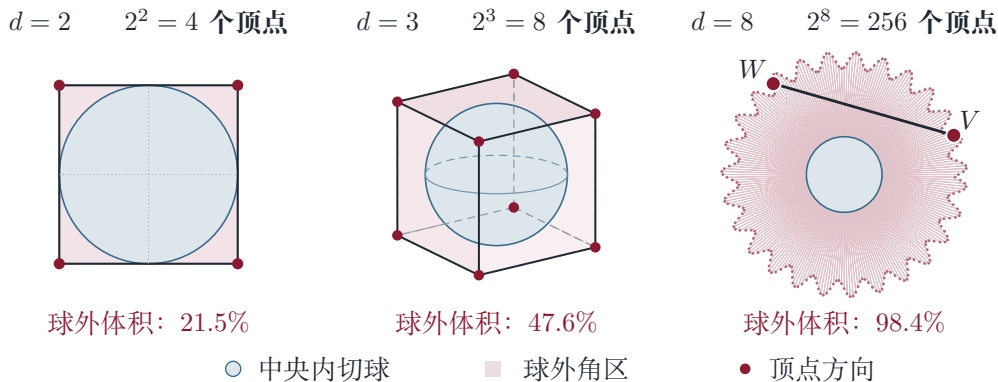


图 1: 按原课第 46 页改绘：维数增加时，顶点数为 2^d ，中央内切球的体积占比迅速下降。右图以 256 个方向编码八维顶点，两个标记顶点恰有四个坐标不同；它是角区的示意图，平面线段的长度与夹角不表示高维距离。百分比使用实际的 d 维体积计算。

沿每个坐标的中点把立方体分开，得到由顶点 $v \in \{0, 1\}^d$ 标记的 2^d 个子立方体 Q_v 。忽略体积为零的分界面，定义

$$A_v = Q_v \setminus B_d$$

为朝向 v 的球外角区。这里的“角区”指这一明确划分；高维时不要求这些区域彼此不连通。

命题 2.1 (几乎全部体积落在球外角区)。中央球占立方体的体积比例为

$$r_d = \frac{\pi^{d/2}}{2^d \Gamma(d/2 + 1)} \leq \left(\frac{\pi}{4}\right)^{\lfloor d/2 \rfloor} \rightarrow 0.$$

每个角区的体积为 $(1 - r_d)/2^d$ 。若 X, Y 是两个独立的均匀立方体随机点，则

$$\mathbb{P}(X, Y \text{ 都在球外, 且属于不同角区}) = (1 - r_d)^2 (1 - 2^{-d}) \rightarrow 1. \quad (2.1)$$

证明直觉 要落在中央球内，总的平方距离不能超过 $1/4$ 。因而无论取哪两个坐标，它们组成的平面坐标也必须落在对应的内切圆里。每增加一对独立坐标，就多一道概率为 $\pi/4 < 1$ 的必要条件；这些条件相乘，使中央球的体积比例迅速减小。

证明. 单位欧氏球的体积为 $\pi^{d/2}/\Gamma(d/2 + 1)$ ，半径缩为 $1/2$ 后体积乘以 2^{-d} ，得到 r_d 的公式。下面直接证明其趋于零，不需要估计 Gamma 函数。把坐标配成 $\lfloor d/2 \rfloor$ 对。若 $X \in B_d$ ，则每对 (X_{2j-1}, X_{2j}) 都满足

$$(X_{2j-1} - 1/2)^2 + (X_{2j} - 1/2)^2 \leq 1/4.$$

这个圆在单位正方形内的面积是 $\pi/4$ ；不同坐标对独立，故 $r_d \leq (\pi/4)^{\lfloor d/2 \rfloor}$ 。关于各坐标中点的反射对称性使全部角区等体积，得到 $(1 - r_d)/2^d$ 。最后，对 $2^d(2^d - 1)$ 个不同的有序角区对求和，就得到(2.1)。□

“球外”并不表示靠近某个顶点的小邻域。例如在每个顶点放一个各坐标宽度为 $\rho \in (0, 1/2)$ 的小方块，这些小方块的总体积为 $2^d \rho^d = (2\rho)^d \rightarrow 0$ 。式(2.1)也还没有给出点间距离下界：不同角区只保证至少一个坐标跨过中点。要理解典型距离，需要进一步看有多少坐标贡献了差异。

2.2 两个随机点：坐标差异累积成常数距离

为了在不同维数下比较同一尺度，以下使用归一化欧氏距离

$$D_d(x, y) = \frac{\|x - y\|_2}{\sqrt{d}}.$$

立方体在这个距离下的直径始终是 1。先看最简单的顶点模型：令 V, W 独立均匀取自 $\{0, 1\}^d$ ，并记

$$H = \sum_{i=1}^d \mathbf{1}_{\{V_i \neq W_i\}}, \quad H \sim \text{Bin}(d, 1/2), \quad D_d(V, W)^2 = \frac{H}{d}.$$

每个坐标不同的概率为 $1/2$ ，所以典型情形是约 $d/2$ 个坐标不同；每个不同坐标对平方距离贡献 $1/d$ ，加起来就是约 $1/2$ 。反过来，两个相邻顶点虽然不同，却只有一个坐标不同，距离为 $1/\sqrt{d}$ 。这说明关键在于大量坐标共同贡献差异。

定理 2.2 (乘积分布中两个随机点的距离集中)。固定一个不随维数改变、支持在 $[0, 1]$ 上的概率分布 ν ，设其方差为 $\sigma^2 > 0$ 。令 X, Y 独立服从 $\nu^{\otimes d}$ ，并记 $\mu = 2\sigma^2$ 。则 $\mathbb{E}D_d(X, Y)^2 = \mu$ ，且对任意 $t > 0$ ，

$$\mathbb{P}(|D_d(X, Y)^2 - \mu| \geq t) \leq \min \left\{ \frac{1}{4dt^2}, 2e^{-2dt^2} \right\}. \quad (2.2)$$

对距离本身，对任意 $\varepsilon > 0$ 有

$$\mathbb{P}(|D_d(X, Y) - \sqrt{\mu}| \geq \varepsilon) \leq 2e^{-2d\mu\varepsilon^2}. \quad (2.3)$$

特别地，当 $d \rightarrow \infty$ 时， $D_d(X, Y) \rightarrow \sqrt{\mu}$ ，其中收敛为依概率收敛。

证明直觉 平方距离是 d 个独立、取值有界的坐标贡献的平均。每一项的平均贡献相同，独立性又使不同项的协方差消失，所以平均后的方差缩小到 $1/d$ 。这个方差计算已经足以说明距离趋于常数；后面的集中不等式会把失败概率进一步压到关于 d 的指数小量。

证明. 令 $Z_i = (X_i - Y_i)^2$. 这些变量独立同分布, 取值于 $[0, 1]$, 且 X_i, Y_i 独立并具有相同均值, 所以

$$\mathbb{E}Z_i = \text{Var}(X_i - Y_i) = 2\sigma^2 = \mu.$$

又有 $\text{Var}(Z_i) \leq \mathbb{E}(Z_i - 1/2)^2 \leq 1/4$, 因此

$$\mathbb{E}D_d(X, Y)^2 = \mu, \quad \text{Var}(D_d(X, Y)^2) = \frac{1}{d^2} \sum_{i=1}^d \text{Var}(Z_i) \leq \frac{1}{4d}.$$

由逐点不等式 $t^2 \mathbf{1}_{\{|D_d^2 - \mu| \geq t\}} \leq (D_d^2 - \mu)^2$, 取期望得到(2.2)中的方差界。这就是 Chebyshev 方法, 此处已经给出了它所需的全部推导。

指数界则将 $U_i = Z_i \in [0, 1]$ 代入后文独立证明的 Hoeffding 不等式 (定理 3.6), 对总和取阈值 dt , 得到 $2\exp\{-2(dt)^2/d\} = 2e^{-2dt^2}$. 读者可以先用方差界理解集中现象, 再在第 3 节学过集中不等式后返回这一步。最后, 由

$$|D_d - \sqrt{\mu}| = \frac{|D_d^2 - \mu|}{D_d + \sqrt{\mu}} \leq \frac{|D_d^2 - \mu|}{\sqrt{\mu}}$$

在平方距离的尾界中取 $t = \sqrt{\mu}\varepsilon$, 便得到距离本身的概率界, 以及所述依概率收敛。□

顶点模型对应 $\nu = \frac{1}{2}\delta_0 + \frac{1}{2}\delta_1$, 其方差为 $1/4$; 连续均匀立方体对应 $\nu = \text{Unif}[0, 1]$, 其方差为 $1/12$. 因此, 两种模型的极限分别为

$$D_d(V, W) \longrightarrow \frac{1}{\sqrt{2}}, \quad D_d(X, Y) \longrightarrow \frac{1}{\sqrt{6}} \quad (\text{连续均匀立方体}).$$

“趋于常数”指的是归一化距离。通常的欧氏距离仍随 $\sqrt{d}$ 增长, 在这两种模型中分别集中在 $\sqrt{d}/2$ 与 $\sqrt{d}/6$ 附近的相对小范围内。

与局部覆盖的联系 仍在定理 2.2 的假设下, 令测试点 X 与训练点 $X^{(1)}, \dots, X^{(N)}$ 相互独立, 都服从 $\nu^{\otimes d}$. 若固定 $0 \leq r < \sqrt{\mu}$, 单个训练点落在测试点的 r 邻域内的概率指数小。对每一对使用定理 3.6 证明中的单侧形式, 再作并集界, 便有

$$\mathbb{P}\left(\min_{1 \leq j \leq N} D_d(X, X^{(j)}) \leq r\right) \leq N \exp\{-2d(\mu - r^2)^2\}. \quad (2.4)$$

具体地, 距离至多为 r 要求坐标平方差的平均值从 μ 向下偏离至少 $\mu - r^2$, 单对的概率因而至多为 $e^{-2d(\mu - r^2)^2}$. 并集界将这些概率相加, 不要求不同的事件独立。若要找到一个这样的近邻的概率至少为固定 $p \in (0, 1)$, 必需 $N \geq p e^{2d(\mu - r^2)^2}$. 这是随机局部覆盖的必要样本量; 它对一般学习算法是否构成困难, 还取决于目标函数类的结构。下面从网格计数出发, 把局部距离要求转成平滑函数的逼近复杂度。

2.3 先把困难变成一个网格计数问题

取 $\mathcal{X} = [0, 1]^d$. 先假设 $0 < \beta \leq 1$, 且目标满足 Hölder 条件

$$|f(x) - f(y)| \leq L \|x - y\|_\infty^\beta. \quad (2.5)$$

把每个坐标轴分成 q 份, 得到 $M = q^d$ 个边长 $h = 1/q$ 的小方块。在每个方块 Q_j 内取一点 c_j , 用常数 $f(c_j)$ 代替整个方块内的函数。

命题 2.3 (分片常数逼近)。设 $\mathcal{F}_M$ 是上述网格上的分片常数函数类。则

$$\inf_{g \in \mathcal{F}_M} \|f - g\|_\infty \leq Lh^\beta = LM^{-\beta/d}, \quad \inf_{g \in \mathcal{F}_M} \|f - g\|_{L^2(P_X)}^2 \leq L^2 M^{-2\beta/d}.$$

第二个结论对任意支持在 $[0, 1]^d$ 上的概率分布 P_X 成立。

证明直觉 平滑性说，在一个小方块里，函数来不及变化很多。维数灾难说，要让每个方向都足够近，小方块的总数必须相乘。提高平滑度改善每块的误差；提高维数增加要照顾的方块数。

证明. 对 $x \in Q_j$, 有 $\|x - c_j\|_\infty \leq h$, 因此 $|f(x) - f(c_j)| \leq Lh^\beta$ 。对 x 取上确界得到第一个式子。平方后对概率分布积分得到第二个式子。□

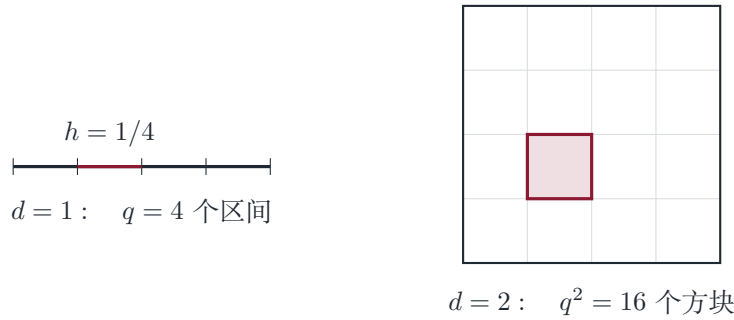


图 2: 同样的坐标分辨率 h , 在 d 维需要 h^{-d} 个区域。每个坐标的选择相乘, 把前面 2^d 个角区的计数推广到任意分辨率。

如果要控制的是函数误差 $\|f - g\|_\infty \leq \varepsilon$, 取 $M \gtrsim (L/\varepsilon)^{d/\beta}$ 即可由这个构造保证该精度。若要控制的是平方风险 $A(\mathcal{F}_M) \leq \varepsilon$, 同一构造给出的充分分辨率为 $M \gtrsim (L^2/\varepsilon)^{d/(2\beta)}$ 。先说清误差标准, 指数才有意义。

2.4 更高平滑度为什么可以换成更高阶多项式

令 $s = m + \alpha$, 其中 $m \geq 0$ 为整数、 $0 < \alpha \leq 1$ 。假设 $f \in C^{m, \alpha}([0, 1]^d)$, 即 m 阶偏导数为 α 次 Hölder, 相关范数由 L 控制。在每个方块使用 m 次 Taylor 多项式, 就有

$$\|f - g_h\|_\infty \leq C_{d,s} Lh^s, \quad N = \binom{d+m}{m} q^d \quad (2.6)$$

个系数, 因此在固定 d, s 时平方逼近误差为 $O(N^{-2s/d})$ 。

证明直觉 常数模型只记录函数的高度；一次模型还记录斜率；二次模型继续记录曲率。Taylor 展开把在小区域内可预测的变化全部扣除, 剩余误差才由最高阶导数的平滑性决定。

误差阶的推导. $m = 0$ 时就是命题 2.3。对 $m \geq 1$, 固定方块内 c, x , 令 $v = x - c$, $u(t) = f(c + tv)$ 。一维 Taylor 公式可改写为

$$u(1) - \sum_{k=0}^m \frac{u^{(k)}(0)}{k!} = \frac{1}{(m-1)!} \int_0^1 (1-t)^{m-1} \{u^{(m)}(t) - u^{(m)}(0)\} dt.$$

链式法则与 Hölder 条件给出 $|u^{(m)}(t) - u^{(m)}(0)| \leq C_{d,m} L \|v\|_\infty^{m+\alpha} t^\alpha$ 。因 $\|v\|_\infty \leq h$ ，积分得到 $C_{d,s} L h^s$ 。总次数不超过 m 的 d 元单项式数目是 $\binom{d+m}{m}$ ，乘以方块数即得 N 。□

这里的常数不能被当作与 d 无关； $\binom{d+m}{m}$ 也应在讨论维数依赖时保留。网格构造给出了一个上界。若想证明任意方法都无法避免维数依赖，需要再给函数类的复杂度下界。

2.5 许多互不干扰的小凸起给出复杂度下界

记 $\mathcal{H}^\beta(L)$ 为满足 (2.5) 且 $\|f\|_\infty \leq L$ 的函数类，仍令 $0 < \beta \leq 1$ 。在 $M = q^d$ 个网格方块中央，各放一个支撑宽度约为 $h/2$ 、高度为 ah^β 的小凸起 b_j ，其中 $a > 0$ 足够小。其具体构造和统计下界将在附录 A 完整给出。

对每个 $\theta \in \{0, 1\}^M$ ，定义 $f_\theta = \sum_{j=1}^M \theta_j b_j$ 。可以选择仅依赖 d, β, L 的 a ，使所有这些函数都属于 $\mathcal{H}^\beta(L)$ 。当 $\theta \neq \theta'$ 时，至少有一个凸起不同，因此

$$\|f_\theta - f_{\theta'}\|_\infty \geq ah^\beta.$$

所以任何半径小于 $ah^\beta/2$ 的一致范数球，至多包含其中一个函数。具体取 $q = \lfloor (a/(4\varepsilon))^{1/\beta} \rfloor$ 。对充分小的 ε ，有 $q \geq \frac{1}{2}(a/(4\varepsilon))^{1/\beta}$ ，且任意两个函数相距至少 4ε 。由此得到

$$\log N(\mathcal{H}^\beta(L), \|\cdot\|_\infty, \varepsilon) \geq c_{d,\beta,L} \varepsilon^{-d/\beta} \quad (2.7)$$

对所有足够小的 ε 成立。这里 $N(\cdot, \cdot, \varepsilon)$ 是覆盖数，正式定义见后文。

证明直觉 每个小区域可以独立决定“有没有一个凸起”。平滑性限制凸起高度，却没有把相距较远区域的选择绑在一起。因此，在分辨率 h 下，整个函数类携带约 h^{-d} 个独立的二元选择。覆盖数把这种需要分辨的自由度，转成一个严格的计数量。

说明 2.4 (下界到底约束了什么)。 式 (2.7) 是函数类的度量熵下界，也能解释有限精度编码需要的位数。它本身并不意味着任何使用 N 个实参数的方法都必须满足同样的参数下界：无限精度实数可以编码任意多的离散信息。要推参数数目下界，还需要限制参数精度、参数范围及解码映射的定量连续模，或模型类本身的覆盖数。

2.6 把逼近困难和统计困难放进同一个式子

网格模型有 M 个局部均值需要估计。若每块约有 n/M 个观测，每个局部平均的噪声方差约为 $\sigma^2 M/n$ 。因此后面的核平滑证明将严格实现下面的平衡：

$$\text{平方风险} \lesssim \underbrace{L^2 M^{-2\beta/d}}_{\text{局部逼近}} + \underbrace{\frac{\sigma^2 M}{n}}_{\text{估计局部均值}}. \quad (2.8)$$

令两项同阶，得到

$$M \asymp n^{d/(2\beta+d)}, \quad \mathbb{E} \left\| \hat{f} - f^* \right\|_{L^2(P_X)}^2 \lesssim n^{-2\beta/(2\beta+d)}$$

其中暂把 d, β, L, σ 看作固定常数。此处是预告推导结构，不是从参数个数直接推出一个普适泛化定理。

维数灾难的结论针对一般平滑函数类。如果目标只依赖少数坐标、具有可加结构或较小的字典表示范数，函数类会显著缩小。Maurey 方法与 RKHS 理论将展示：对某些结构化函数类，控制表示的范数与几何，比单纯数参数更有信息。这延续了原课程第一讲关于维数灾难的讨论[2]。

3 泛化：从一个函数的集中到整个函数类的控制

设 $S = (Z_1, \dots, Z_n)$ 是来自分布 P 的独立同分布样本。对实值函数 g ，记

$$Pg = \mathbb{E}[g(Z)], \quad P_n g = \frac{1}{n} \sum_{i=1}^n g(Z_i).$$

在学习问题中，我们把预测函数类 $\mathcal{F}$ 转成损失函数类 $\mathcal{G} = \{\ell_f : f \in \mathcal{F}\}$ ，其中 $\ell_f(x, y) = \ell(y, f(x))$ 。若无另说， $\mathcal{G}$ 非空且在取样前确定。本节要控制的核心量是

$$\Delta_S(\mathcal{G}) = \sup_{g \in \mathcal{G}} |(P_n - P)g|. \quad (3.1)$$

它控制的是同一批数据上所有候选模型的评估误差，因此也控制看过数据之后才选出的模型。下面默认函数类及其上确界可测；例如，存在一个可数子类，使所用上确界可由该子类逼近。集中不等式部分沿着原课程第七讲 PPT 的思路，从 Markov 与 Chebyshev 出发，逐步提升矩阶，再用指数矩和 Chernoff 方法处理独立和 [6, 第 6 至 15 页][5]。随后进入统一收敛与 Rademacher complexity [7, 8]。下面补齐各步之间的动机，证明中的记号和常数均在此独立给出。

3.1 Markov 与 Chebyshev：用已知矩限制尾部概率

对于固定的 g ， $P_n g - Pg$ 是均值为零的随机波动。我们希望在给定样本量 n 时，就能回答它超过 t 的概率有多大。最简单的办法，是把尾部事件的指示函数压在一个已知期望的随机变量下面。

命题 3.1 (Markov 与 Chebyshev 不等式)。若 $X \geq 0$ 且 $\mathbb{E}X < \infty$ ，则对所有 $t > 0$ ，

$$\mathbb{P}(X \geq t) \leq \frac{\mathbb{E}X}{t}.$$

若 X 有有限方差 $\sigma^2 = \mathbb{E}(X - \mathbb{E}X)^2$ ，则对所有 $t > 0$ ，

$$\mathbb{P}(|X - \mathbb{E}X| \geq t) \leq \frac{\sigma^2}{t^2}.$$

证明直觉 在事件 $\{X \geq t\}$ 上，每次至少贡献 t 的大小。既然总平均只有 $\mathbb{E}X$ ，这个事件的概率便不能超过 $\mathbb{E}X/t$ 。Chebyshev 使用完全相同的想法，只是把偏差先平方：这样既消除了正负号，又能调用已知的方差。

证明。 由逐点不等式 $t\mathbf{1}_{\{X \geq t\}} \leq X$ ，取期望就得到 Markov 界。再对非负变量 $(X - \mathbb{E}X)^2$ 应用它，并把阈值取为 t^2 ，即得 Chebyshev 界。 $\square$

例如，若 $U_1, \dots, U_n$ 独立同分布，均值为 μ 、方差为 σ^2 ，则独立性使不同项的协方差消失，因而

$$\text{Var}\left(\frac{1}{n} \sum_{i=1}^n U_i\right) = \frac{\sigma^2}{n}, \quad \mathbb{P}\left(\left|\frac{1}{n} \sum_{i=1}^n U_i - \mu\right| \geq t\right) \leq \frac{\sigma^2}{nt^2}.$$

固定失败概率 $\delta \in (0, 1)$ 时，阈值是 $\sigma/\sqrt{n\delta}$ 。这里已经看见 $n^{-1/2}$ 的平均尺度；需要改进的是对很小失败概率的依赖。稍后同时控制很多函数时， $\delta^{-1/2}$ 会很昂贵，而指数尾可以把它换成 $\sqrt{\log(1/\delta)}$ 。这种改进需要比有限方差更强的信息。

3.2 提升矩阶：先选变换，再让矩阶适应偏差

从一阶矩到二阶矩的过程可以继续。对任意 $p > 0$ ，只要 $\mathbb{E}|W|^p < \infty$ ，将 Markov 用于 $|W|^p$ 就有

$$\mathbb{P}(|W| \geq t) = \mathbb{P}(|W|^p \geq t^p) \leq \frac{\mathbb{E}|W|^p}{t^p}, \quad t > 0. \quad (3.2)$$

提高幂次会更强地惩罚大偏差，但分子里的高阶矩也会增长。因此，真正有用的信息是**矩随阶数怎样增长**。固定一个有限的 p ，式(3.2)仍只是 t^{-p} 的多项式尾；若能同时控制一族矩，就可以针对每个阈值 t 选择合适的 p 。

例 3.2 (高斯变量：优化矩阶可以恢复高斯型尾)。设 $G \sim N(0, 1)$ 。对整数 $k \geq 1$,

$$\mathbb{E}G^{2k} = (2k-1)(2k-3)\cdots 1 \leq (2k)^k.$$

于是 $\mathbb{P}(|G| \geq t) \leq (2k/t^2)^k$ 。当 $t \geq 2$ 时，选择 $k = \lfloor t^2/4 \rfloor$ ，便得到

$$\mathbb{P}(|G| \geq t) \leq 2^{-k} \leq \exp\left(-\frac{\log 2}{8} t^2\right).$$

这里要观察的是指数形式，常数还可以进一步改进。

证明直觉 若永远只用第四阶矩，尾界就永远只有 t^{-4} 。高斯变量的高阶矩增长足够慢，允许我们让 k 随 t^2 增长，同时把括号中的比值维持在小于 1 的常数。一个固定小于 1 的数取约 t^2 次幂，就变成了高斯型衰减。

证明. 记标准高斯密度为 ϕ 。由 $\phi'(x) = -x\phi(x)$ 分部积分，且边界项为零，

$$\mathbb{E}G^{2k} = - \int_{\mathbb{R}} x^{2k-1} \phi'(x) dx = (2k-1) \mathbb{E}G^{2k-2}.$$

从 $\mathbb{E}G^0 = 1$ 递推可得偶数矩公式；每个因子至多是 $2k$ ，所以得到所述矩上界。选择 $k = \lfloor t^2/4 \rfloor$ 后， $2k/t^2 \leq 1/2$ ，且 $t \geq 2$ 保证 $k \geq t^2/8$ ，代入(3.2)即得结论。□

这补足了 PPT 中从一阶、二阶矩走向指数矩的一步：指数尾并非由 Markov 或 Chebyshev 自动给出，而是来自额外的高阶矩控制，以及随阈值优化变换的做法。

3.3 MGF 与 Chernoff：重新加权全部矩，再优化参数

指数函数提供了一族更方便的变换。随机变量 X 的矩母函数 (moment generating function, MGF) 与对数矩母函数记为

$$M_X(\lambda) = \mathbb{E}e^{\lambda X}, \quad K_X(\lambda) = \log M_X(\lambda),$$

其中 K_X 只在 $M_X(\lambda) < \infty$ 处使用。若存在 $a > 0$ 使 $\mathbb{E}e^{a|X|} < \infty$ ，则当 $|\lambda| < a$ 时，

$$M_X(\lambda) = \sum_{k=0}^{\infty} \frac{\lambda^k \mathbb{E}X^k}{k!}.$$

对中心化变量 W ，零阶项是 1，一阶项为零，二阶项是 $\lambda^2 \text{Var}(W)/2$ 。参数 λ 通过权重 $\lambda^k/k!$ 重新分配各阶矩的作用。这正是原 PPT 上从多项式变换过渡到指数变换的动机：先把各阶矩汇总成一个函数，再选择合适的 λ 。

命题 3.3 (Chernoff 方法)。 设 X 的矩母函数在至少一个正参数处有限。对每个 $t > 0$,

$$\mathbb{P}(X \geq t) \leq \inf_{\substack{\lambda > 0 \\ M_X(\lambda) < \infty}} \exp(K_X(\lambda) - \lambda t). \quad (3.3)$$

若 X 在某个负参数处也有有限的 MGF, 则对 $-X$ 使用同一结论可以控制左尾。

证明直觉 增大 λ 会使阈值对应的因子 $e^{-\lambda t}$ 更小, 但还必须同时考虑随 λ 变化的指数加权代价 $M_X(\lambda)$ 。Chernoff 方法选择使二者乘积最小的参数。它与刚才选择矩阶 p 的策略相同: 都把 Markov 用于一个可调节的变换, 再针对所关心的阈值优化这个变换。

证明. 对任意允许的 $\lambda > 0$, 指数函数严格递增, 故

$$\mathbb{P}(X \geq t) = \mathbb{P}(e^{\lambda X} \geq e^{\lambda t}) \leq e^{-\lambda t} \mathbb{E} e^{\lambda X} = e^{K_X(\lambda) - \lambda t}.$$

该界对每个这样的 λ 都成立, 取下确界即可。 □

以 $G \sim N(0, 1)$ 为例, 配方给出

$$M_G(\lambda) = \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} e^{\lambda x - x^2/2} dx = e^{\lambda^2/2}.$$

于是 Chernoff 指数为 $\lambda^2/2 - \lambda t$, 最优参数为 $\lambda = t$ 。左右两尾合起来得到 $\mathbb{P}(|G| \geq t) \leq 2e^{-t^2/2}$ 。这与优化偶数矩得到相同的尾部形状, 而且计算更直接、常数更好。

说明 3.4 (提升矩阶与使用指数矩都需要条件)。 上面的级数交换可由支配收敛证明, 因为各部分和的绝对值都不超过 $e^{|\lambda||X|} \leq e^{a|X|}$ 。只有某个有限阶矩时, 不能继续调用更高阶矩; 即使所有阶矩都有限, 也不能直接假定正指数矩有限。例如 $X = e^G$ 满足 $\mathbb{E} X^p = e^{p^2/2} < \infty$, $p > 0$, 但对任意 $\lambda > 0$, $\mathbb{E} e^{\lambda X} = \infty$: 相应积分的指数 $\lambda e^x - x^2/2$ 在 $x \rightarrow \infty$ 时趋于正无穷。因此, 后面每次使用 MGF 都会给出能控制它的具体假设。

3.4 独立和: 指数变换把平均效应变成乘法

矩方法也能展开 $\mathbb{E}(\sum_i W_i)^p$, 但高阶混合项很快变得繁杂。指数函数则直接满足 $e^{\lambda \sum_i W_i} = \prod_i e^{\lambda W_i}$ 。若 W_i 相互独立, 且各个指数矩有限, 就有

$$M_{\sum_i W_i}(\lambda) = \prod_i M_{W_i}(\lambda), \quad K_{\sum_i W_i}(\lambda) = \sum_i K_{W_i}(\lambda). \quad (3.4)$$

取对数后, 独立和的尾部控制因此归结为各项指数矩上界的相加。

定义 3.5 (次高斯随机变量)。 设 $\nu \geq 0$ 。若 $\mathbb{E} W = 0$, 且对所有 $\lambda \in \mathbb{R}$,

$$\mathbb{E} e^{\lambda W} \leq e^{\nu^2 \lambda^2/2},$$

则称 W 是方差代理为 ν^2 的次高斯变量。这里 ν^2 是上界中的参数, 未必等于真实方差; 比较零附近的二阶项可知 $\text{Var}(W) \leq \nu^2$ 。

证明直觉 次高斯条件表示：用来做 Chernoff 估计的指数矩，至多与方差为 ν^2 的中心高斯变量一样大。独立性使各项的方差代理相加，平均又把总和缩小 n 倍，于是平均值的方差代理缩小到原来的 $1/n$ 。

具体地，设独立中心化变量 W_i 分别具有方差代理 ν_i^2 ，并令 $V = \sum_i \nu_i^2 > 0$ 。由(3.4)， $K_{\sum_i W_i}(\lambda) \leq V\lambda^2/2$ 。对总和取阈值 nt ，并在 Chernoff 界中取 $\lambda = nt/V$ ，得到

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^n W_i\right| \geq t\right) \leq 2\exp\left(-\frac{n^2 t^2}{2V}\right). \quad (3.5)$$

特别地，若每个 $\nu_i^2 \leq \nu^2$ ，则以至少 $1 - \delta$ 的概率，

$$\left|\frac{1}{n}\sum_{i=1}^n W_i\right| \leq \nu\sqrt{\frac{2\log(2/\delta)}{n}}.$$

固定 δ 仍是 $n^{-1/2}$ 尺度，但现在把失败概率压得很小只需付出对数代价。这里独立性负责乘法，次高斯假设负责控制每个因子；独立性本身并不能保证指数矩存在。若用同一个尺度 ν^2 控制 M 个固定候选的方差与次高斯代理，分别作并集界得到的阈值为 $\nu\sqrt{M/(n\delta)}$ 与 $\nu\sqrt{2\log(2M/\delta)/n}$ 。这预示了指数尾为何特别适合统一控制多个候选。

3.5 Hoeffding：有界性如何给出二次指数矩界

次高斯条件把所需的 MGF 控制写得很清楚。接下来的问题是：除了高斯变量，哪些常见变量也满足它？Hoeffding 引理说明，只要取值落在有限区间，中心化之后就有方差代理 $(b-a)^2/4$ 的次高斯控制。

定理 3.6 (Hoeffding 不等式)。设 $U_1, \dots, U_n$ 相互独立，且 $a_i \leq U_i \leq b_i$ 几乎处处成立。令 $V_0 = \sum_i (b_i - a_i)^2 > 0$ 。对 $t > 0$ ，

$$\mathbb{P}\left(\left|\sum_{i=1}^n (U_i - \mathbb{E}U_i)\right| \geq t\right) \leq 2\exp\left(-\frac{2t^2}{V_0}\right).$$

特别地，若 $g(Z) \in [0, B]$ ，则

$$\mathbb{P}(|P_n g - P g| \geq t) \leq 2e^{-2nt^2/B^2}. \quad (3.6)$$

证明直觉 先研究一个变量的对数指数矩 $K(\lambda)$ 。其二阶导数恰好是某个重新加权分布下的方差；重新加权不会改变变量的取值区间，所以二阶导数始终有界。把这个曲率界积分两次，便得到一个二次函数控制 K 。最后利用独立性把这些二次上界相加，再选择最合适的 λ 。

证明. 先令 $U \in [a, b]$ ，并定义 $K(\lambda) = \log \mathbb{E}e^{\lambda(U - \mathbb{E}U)}$ 。在密度与 $e^{\lambda U}$ 成比例的新分布 P_λ 下，直接求导得到

$$K''(\lambda) = \text{Var}_{P_\lambda}(U) \leq \mathbb{E}_{P_\lambda}\left(U - \frac{a+b}{2}\right)^2 \leq \frac{(b-a)^2}{4}.$$

又因为 $K(0) = K'(0) = 0$ ，带积分余项的 Taylor 公式给出，对所有实数 λ ，

$$K(\lambda) = \lambda^2 \int_0^1 (1-s)K''(s\lambda) ds \leq \frac{\lambda^2(b-a)^2}{8}.$$

因此, $T = \sum_i (U_i - \mathbb{E}U_i)$ 满足 $\mathbb{E}e^{\lambda T} \leq e^{\lambda^2 V_0/8}$ 。对任意 $\lambda > 0$, Markov 不等式给出

$$\mathbb{P}(T \geq t) \leq e^{-\lambda t} \mathbb{E}e^{\lambda T} \leq \exp\left(-\lambda t + \frac{\lambda^2 V_0}{8}\right).$$

取 $\lambda = 4t/V_0$, 得到单侧界 e^{-2t^2/V_0} 。对 $-T$ 重复一次并相加, 即得双侧界。令 $U_i = g(Z_i)$, 再把总和的阈值取成 nt , 得到(3.6)。□

3.6 Bernstein: 保留方差, 才能看见较快的统计速率

Hoeffding 只看取值区间。如果某个随机变量通常很接近零, 仅有很小概率达到区间端点, 它的实际波动会远小于区间长度所暗示的尺度。Bernstein 不等式把真实方差留在界中。

定理 3.7 (有界变量的 Bernstein 不等式)。设 $\xi_1, \dots, \xi_n$ 相互独立, $\mathbb{E}\xi_i = 0$, $|\xi_i| \leq b$, 其中 $b > 0$, 并记 $v = \sum_i \mathbb{E}\xi_i^2$ 。对任意 $t > 0$,

$$\mathbb{P}\left(\pm \sum_{i=1}^n \xi_i \geq t\right) \leq \exp\left(-\frac{t^2}{2(v + bt/3)}\right). \quad (3.7)$$

对任意 $u > 0$, 还可以得到如下单侧阈值形式:

$$\mathbb{P}\left(\pm \sum_{i=1}^n \xi_i \geq \sqrt{2vu} + \frac{bu}{3}\right) \leq e^{-u}. \quad (3.8)$$

因此, 若 $v \leq n\sigma^2$, 则以至少 $1 - 2e^{-u}$ 的概率,

$$\left|\frac{1}{n} \sum_{i=1}^n \xi_i\right| \leq \sqrt{\frac{2\sigma^2 u}{n}} + \frac{bu}{3n}. \quad (3.9)$$

证明直觉 指数函数的级数中, 一阶项由于中心化而消失, 二阶项正是方差。有界性使更高阶矩都能用二阶矩乘上 b 的幂控制。于是整个级数可以求和: 小偏差由方差控制, 大偏差还需支付有界性的代价。后面局部化分析的关键, 是把这里的 σ^2 再与未知的超额风险联系起来。

证明. 若 $v = 0$, 所有变量均为零, 结论显然。以下设 $v > 0$ 。有界性保证指数级数可以逐项取期望。对 $0 \leq \lambda < 3/b$, 使用 $|\mathbb{E}\xi_i^k| \leq b^{k-2}\mathbb{E}\xi_i^2$ 和 $k! \geq 2 \cdot 3^{k-2}$, 有

$$\begin{aligned} \mathbb{E}e^{\lambda \xi_i} &\leq 1 + \frac{\lambda^2 \mathbb{E}\xi_i^2}{2} \sum_{k=2}^{\infty} \left(\frac{\lambda b}{3}\right)^{k-2} \\ &\leq \exp\left(\frac{\lambda^2 \mathbb{E}\xi_i^2}{2(1 - \lambda b/3)}\right). \end{aligned}$$

独立性因而给出

$$\log \mathbb{E}e^{\lambda \sum_i \xi_i} \leq \frac{v\lambda^2}{2(1 - \lambda b/3)}.$$

对阈值 $t > 0$, 取 $\lambda = t/(v + bt/3)$, 直接得到(3.7)。为取得(3.8)中的阈值常数, 记 $c = b/3$, $a = \sqrt{2u/v}$, 取 $\lambda = a/(1 + ca)$ 。此时 $1 - c\lambda = 1/(1 + ca)$; 将阈值 $t = va + cu = \sqrt{2vu} + cu$ 代入 Chernoff 界, 指数恰好为

$$-\lambda t + \frac{v\lambda^2}{2(1 - c\lambda)} = -u.$$

这证明了正侧界; 对 $-\xi_i$ 应用同一论证, 得到负侧界。□

这里两项的意义不同： $\sqrt{\sigma^2 u/n}$ 是平均后留下的随机波动， $bu/(3n)$ 则来自尾部控制。若进一步有 $\sigma^2 \lesssim r$ ，我们会得到 $\sqrt{ru/n} + u/n$ ，而不再是一个与 r 无关的 $n^{-1/2}$ 上界。

3.7 为什么固定函数的集中还不够

例 3.8 (选择发生在观察数据之后)。令 X 在 $[0, 1]$ 上均匀分布，并考虑所有有限集合的指示函数 $g_A(x) = \mathbf{1}_{\{x \in A\}}$ 。对每个预先固定的有限集 A ， $Pg_A = 0$ ，且 $P_n g_A = 0$ 几乎必然成立。然而看过数据后可以选择 $A_S = \{X_1, \dots, X_n\}$ ；这时 $P_n g_{A_S} = 1$ ，但 $Pg_{A_S} = 0$ 。每个固定候选的评估都正确，无法推出从全部候选中选出的那个也正确。

学习算法输出 $\hat{f}(S)$ ，恰好具有这种数据依赖性。因此我们希望先得到一个事件，在这个事件上所有 $f \in \mathcal{F}$ 都被同时控制。风险分解中的泛化项便归结为(3.1)；对于经验风险近似最小化器，类内超额风险不超过 $2\Delta_S(\mathcal{G})$ 加优化误差。

命题 3.9 (有限函数类的统一界)。若确定性函数类 $\mathcal{G}$ 含有 M 个函数，且每个函数取值于 $[0, B]$ ，则以至少 $1 - \delta$ 的概率，

$$\Delta_S(\mathcal{G}) \leq B \sqrt{\frac{\log(2M/\delta)}{2n}}.$$

证明直觉 一个函数出错的概率指数小，同时考虑 M 个函数最多把失败概率乘以 M 。反解指数后，模型数量只以 $\log M$ 进入统计误差。这也解释了为什么应当数可以区分的预测行为，而不能只看有多少实数参数。

证明. 对每个函数应用(3.6)，再作并集界：

$$\mathbb{P}(\Delta_S(\mathcal{G}) > t) \leq \sum_{g \in \mathcal{G}} \mathbb{P}(|(P_n - P)g| > t) \leq 2Me^{-2nt^2/B^2}.$$

令右端等于 δ 即可。 □

3.8 Covering number: 把无限类变成有限分辨率的类

定义 3.10 (覆盖数与度量熵)。给定度量空间 (T, d) ，若有限集合 $\{t_1, \dots, t_M\} \subseteq T$ 满足每个 $t \in T$ 都有某个 j 使 $d(t, t_j) \leq \varepsilon$ ，则称其为 ε 网。最小的 M 记为 $N(T, d, \varepsilon)$ ， $\log N(T, d, \varepsilon)$ 称为度量熵。

这里度量是结论的一部分。对函数类，先用确定性的全局度量 $d_\infty(g, h) = \sup_z |g(z) - h(z)|$ 。它同时控制训练分布上的平均和这次训练样本上的平均。

定理 3.11 (确定性覆盖的统一界)。设 $\mathcal{G}$ 中所有函数取值于 $[0, B]$ 。对预先确定的 $\varepsilon > 0$ ，若 $N_\varepsilon = N(\mathcal{G}, d_\infty, \varepsilon) < \infty$ ，则以至少 $1 - \delta$ 的概率，

$$\Delta_S(\mathcal{G}) \leq 2\varepsilon + B \sqrt{\frac{\log(2N_\varepsilon/\delta)}{2n}}. \quad (3.10)$$

证明直觉 只在有限张网格上作概率估计，再靠连续性填满网格之间的空隙。网越细，填空误差越小，但需要同时控制的点越多。 ε 因而是一个分辨率选择：粗糙化误差与统计误差需要平衡。

证明. 固定一个大小为 N_ε 的网 $\mathcal{G}_\varepsilon$. 对每个 g 选择 $\pi(g) \in \mathcal{G}_\varepsilon$, 使 $d_\infty(g, \pi(g)) \leq \varepsilon$. 于是

$$\begin{aligned} |(P_n - P)g| &\leq |(P_n - P)\pi(g)| + P_n |g - \pi(g)| + P |g - \pi(g)| \\ &\leq \max_{h \in \mathcal{G}_\varepsilon} |(P_n - P)h| + 2\varepsilon. \end{aligned}$$

对这个预先固定的网应用命题 3.9 即可。 $\square$

可以在看数据之前选择使 (3.10) 右端较小的 ε . 例如若 $\log N_\varepsilon \lesssim D \log(C/\varepsilon)$, 取逐渐缩小的分辨率可得到大约 $\sqrt{D \log n/n}$ 的上界. 若损失对预测值是 L Lipschitz 的, 则预测函数的 ε 网提供损失函数的 $L\varepsilon$ 网. **不能把只在训练样本上接近的两个函数直接代入本定理**: 样本上的接近只控制 $P_n |g - h|$, 并未控制 $P |g - h|$. 后面将用条件于样本的随机符号过程处理这种度量.

例 3.12 (有限维球的覆盖数). $\mathbb{R}^D$ 中任意范数球 AB 满足 $N(AB, \|\cdot\|, \varepsilon) \leq (1 + 2A/\varepsilon)^D$. 证明只需取最大 ε 分离集: 以其各点为中心的 $\varepsilon/2$ 球互不相交, 且都位于 $(A + \varepsilon/2)B$ 中. 比较体积得到点数上界; 最大性又保证这些中心构成覆盖. 这里真正起作用的是所选度量下的几何; 把参数距离换成函数距离时, 相应的 Lipschitz 常数也必须计入.

3.9 Rademacher complexity: 函数类能拟合多少随机符号

覆盖数先把函数类离散化. 另一条路线是直接测量函数类对纯噪声的拟合能力. 取独立符号 $\sigma_i \in \{-1, 1\}$, 每个符号的两种取值概率相同, 且与样本独立.

定义 3.13 (本讲采用带绝对值的约定). 定义经验 Rademacher complexity 及其总体平均为

$$\hat{\mathfrak{R}}_S(\mathcal{G}) = \mathbb{E}_\sigma \sup_{g \in \mathcal{G}} \left| \frac{1}{n} \sum_{i=1}^n \sigma_i g(Z_i) \right|, \quad \mathfrak{R}_n(\mathcal{G}) = \mathbb{E}_S \hat{\mathfrak{R}}_S(\mathcal{G}).$$

这里先让函数适应随机符号, 再取符号的平均; 如果交换这两个次序, 每个固定函数与独立随机符号的平均相关性都是零.

一些教材去掉上式中的绝对值, 记相应复杂度为 $\hat{\mathfrak{R}}_S^+$. 当函数类关于原点对称时, 两种定义相等; 当 $0 \in \mathcal{G}$ 时, $\hat{\mathfrak{R}}_S^+ \leq \hat{\mathfrak{R}}_S \leq 2\hat{\mathfrak{R}}_S^+$. 下面所有常数均对应定义 3.13, 不混用两种约定.

定理 3.14 (对称化). 若相应期望有限, 则

$$\mathbb{E}_S \Delta_S(\mathcal{G}) \leq 2\mathfrak{R}_n(\mathcal{G}). \quad (3.11)$$

证明直觉 未知总体均值很难直接处理, 先把它写成另一批独立样本平均的期望. 两批样本地位完全相同, 所以每对样本谁在前、谁在后都可以随机交换. 这次交换产生的符号正是 Rademacher 变量. 最后把两批样本拆开, 就只剩下函数类对随机符号的拟合能力.

证明. 引入与 S 独立同分布的幽灵样本 $S' = (Z'_1, \dots, Z'_n)$, 记其经验平均为 P'_n . 对固定 S , 由 $Pg = \mathbb{E}_{S'} P'_n g$ 和凸性,

$$\sup_g |P_n g - P g| \leq \mathbb{E}_{S'} \sup_g |P_n g - P'_n g|.$$

现在对 S 也取平均。每一对 (Z_i, Z'_i) 的联合分布在交换两个分量后不变，不同的样本对又相互独立。因此整个以 g 为指标的过程在随机交换后分布不变：

$$\begin{aligned}\mathbb{E}_{S, S'} \sup_g \left| \frac{1}{n} \sum_i [g(Z_i) - g(Z'_i)] \right| &= \mathbb{E}_{S, S', \sigma} \sup_g \left| \frac{1}{n} \sum_i \sigma_i [g(Z_i) - g(Z'_i)] \right| \\ &\leq \mathbb{E}_{S, \sigma} \sup_g \left| \frac{1}{n} \sum_i \sigma_i g(Z_i) \right| + \mathbb{E}_{S', \sigma} \sup_g \left| \frac{1}{n} \sum_i \sigma_i g(Z'_i) \right| \\ &= 2\mathfrak{R}_n(\mathcal{G}).\end{aligned}$$

这里引入符号的等式是在对两批样本取平均后成立，不能把已经固定的两批数值当成仍然具有交换对称性。□

引理 3.15 (有界差分不等式)。若 $H(Z_1, \dots, Z_n)$ 在只替换第 i 个坐标时最多改变 c_i ，则

$$\mathbb{P}(H - \mathbb{E}H \geq t) \leq \exp\left(-\frac{2t^2}{\sum_i c_i^2}\right).$$

同样的界适用于 $\mathbb{E}H - H$ 。

证明直觉 一次一个地揭示样本。每次揭示造成的条件期望变化均值为零，而且取值区间长度至多为 c_i 。所以 Hoeffding 的指数矩论证可逐步条件化使用。

证明. 令 $M_i = \mathbb{E}[H \mid Z_1, \dots, Z_i]$, $D_i = M_i - M_{i-1}$ 。给定前 $i-1$ 个坐标后， D_i 条件均值为零，条件取值区间长度至多为 c_i 。Hoeffding 证明中的单变量界给出 $\mathbb{E}[e^{\lambda D_i} \mid Z_1, \dots, Z_{i-1}] \leq e^{\lambda^2 c_i^2 / 8}$ 。迭代条件期望得到 $\mathbb{E}e^{\lambda(H - \mathbb{E}H)} \leq e^{\lambda^2 \sum_i c_i^2 / 8}$ ，再优化 Chernoff 参数即可。□

定理 3.16 (Rademacher 泛化界)。若 $\mathcal{G}$ 中的函数均取值于 $[0, B]$ ，则以至少 $1 - \delta$ 的概率，

$$\Delta_S(\mathcal{G}) \leq 2\mathfrak{R}_n(\mathcal{G}) + B\sqrt{\frac{\log(1/\delta)}{2n}}.$$

另一个同样以至少 $1 - \delta$ 概率成立、只需经验复杂度的版本是

$$\Delta_S(\mathcal{G}) \leq 2\widehat{\mathfrak{R}}_S(\mathcal{G}) + 3B\sqrt{\frac{\log(2/\delta)}{2n}}. \quad (3.12)$$

证明直觉 对称化已经控制了平均的统一偏差。替换一个样本只能让统一偏差改变 B/n ，所以这个量集中在自己的均值附近。经验复杂度也有同样的稳定性，因而还可以用它替换未知的总体平均复杂度。

证明. 替换一个坐标时， Δ_S 的变化至多为 B/n 。引理 3.15 与 (3.11) 直接给出第一个结论。同理， $\widehat{\mathfrak{R}}_S$ 的变化也至多为 B/n 。分别对 Δ_S 的上偏差和 $\widehat{\mathfrak{R}}_S$ 的下偏差应用该引理，各分配失败概率 $\delta/2$ 。在两个事件的交集上，

$$\Delta_S \leq 2\mathfrak{R}_n + Bt_\delta, \quad \mathfrak{R}_n \leq \widehat{\mathfrak{R}}_S + Bt_\delta, \quad t_\delta = \sqrt{\frac{\log(2/\delta)}{2n}}.$$

代入即得系数 3。注意 $\sum_i (B/n)^2 = B^2/n$ ，所以集中指数正比于 n ，不是 n^2 。□

3.10 Contraction: 把损失的复杂度还原成预测的复杂度

定理 3.17 (带绝对值约定下的收缩原理)。固定样本, 令 $T \subset \mathbb{R}^n$ 为预测值向量的集合。若每个 $\phi_i: \mathbb{R} \rightarrow \mathbb{R}$ 都是 L Lipschitz 的, 且 $\phi_i(0) = 0$, 则

$$\mathbb{E}_\sigma \sup_{t \in T} \left| \frac{1}{n} \sum_i \sigma_i \phi_i(t_i) \right| \leq 2L \mathbb{E}_\sigma \sup_{t \in T} \left| \frac{1}{n} \sum_i \sigma_i t_i \right|. \quad (3.13)$$

证明直觉 随机符号在一个坐标上只做两件事: 奖励较大的取值, 或奖励较小的取值。取两种符号的平均后, 这个坐标所能带来的额外收益只取决于两个候选值的距离。Lipschitz 变换把距离至多放大 L 倍。逐个坐标替换, 就能把复杂的损失函数剥掉。

证明. 先证明没有绝对值的比较。给定其他符号后, 把剩余各坐标的贡献记为 A_t 。对一个 L Lipschitz 函数 ϕ ,

$$\begin{aligned} & \frac{1}{2} \sup_t (A_t + \phi(t_k)) + \frac{1}{2} \sup_t (A_t - \phi(t_k)) \\ &= \frac{1}{2} \sup_{s,t} \{A_s + A_t + \phi(s_k) - \phi(t_k)\} \\ &\leq \frac{1}{2} \sup_{s,t} \{A_s + A_t + L|s_k - t_k|\} \\ &= \frac{1}{2} \sup_s (A_s + Ls_k) + \frac{1}{2} \sup_t (A_t - Lt_k). \end{aligned}$$

最后一个等式利用了交换 s, t 不改变 $A_s + A_t$ 。依次对所有坐标应用这个不等式, 得到无绝对值的比较, 常数为 L 。

再令 $T_0 = T \cup \{0\}$ 。由于 $\phi_i(0) = 0$, $\sup_{t \in T_0} \sum_i \sigma_i \phi_i(t_i)$ 和其负号版本均非负。绝对值上确界不超过这两个上确界之和; 由符号对称性, 两者期望相同。对每一项使用刚才的无绝对值比较, 使得

$$\mathbb{E} \sup_{t \in T} \left| \sum_i \sigma_i \phi_i(t_i) \right| \leq 2L \mathbb{E} \sup_{t \in T_0} \sum_i \sigma_i t_i \leq 2L \mathbb{E} \sup_{t \in T} \left| \sum_i \sigma_i t_i \right|.$$

除以 n 即得结论。一般集合可先取有限子集, 再用可测的极限逼近。 $\square$

应用于损失时, 应该取 $\phi_i(t) = \ell(Y_i, t) - \ell(Y_i, 0)$, 而不是默认 $\ell(Y_i, 0) = 0$ 。例如, 当 $|Y_i|, |f(X_i)| \leq M$ 时, 平方损失在 $[-M, M]$ 上是 $4M$ Lipschitz 的; 在这个区间使用收缩即可。去掉损失基线后, 基线自身的经验偏差仍需单独保留, 或者直接研究两个模型的损失差。这一区别在采用绝对值复杂度时尤其需要注意。

3.11 两个可以直接算出的例子: 范数比参数数量更有信息

例 3.18 (ℓ_2 约束与 Hilbert 特征)。设 $f_w(x) = \langle w, \varphi(x) \rangle$, $\|w\| \leq A$, 其中内积和范数来自某个实 Hilbert 空间, 且 $\|\varphi(X_i)\| \leq R$ 。由对偶性、Jensen 不等式和随机符号的独立性,

$$\begin{aligned} \hat{\mathcal{R}}_S(\mathcal{F}) &= \frac{A}{n} \mathbb{E}_\sigma \left\| \sum_i \sigma_i \varphi(X_i) \right\| \\ &\leq \frac{A}{n} \sqrt{\mathbb{E}_\sigma \left\| \sum_i \sigma_i \varphi(X_i) \right\|^2} = \frac{A}{n} \sqrt{\sum_i \|\varphi(X_i)\|^2} \leq \frac{AR}{\sqrt{n}}. \end{aligned}$$

关键一步是展开平方后, $i \neq j$ 的项含有 $\mathbb{E}[\sigma_i \sigma_j] = 0$, 所以交叉项全部消失。结论可用于无限维 Hilbert 空间; 但若 R 或表示目标函数所需的 A 随维数增长, 维数依赖仍会通过它们出现。

例 3.19 (ℓ_1 约束与有限字典)。设 $f_w(x) = \sum_{j=1}^D w_j \varphi_j(x)$, $\|w\|_1 \leq A$, 且 $|\varphi_j(X_i)| \leq R$ 。 ℓ_1 与 ℓ_∞ 的对偶性给出

$$\hat{\mathfrak{R}}_S(\mathcal{F}) = \frac{A}{n} \mathbb{E}_\sigma \max_{1 \leq j \leq D} \left| \sum_i \sigma_i \varphi_j(X_i) \right| \leq AR \sqrt{\frac{2 \log(2D)}{n}}.$$

为看清最后一步, 记 $V_j = \sum_i \sigma_i \varphi_j(X_i)$ 。对任何 $\lambda > 0$, 单变量的 Hoeffding 指数矩界给出 $\mathbb{E} e^{\pm \lambda V_j} \leq e^{\lambda^2 n R^2 / 2}$, 所以

$$\mathbb{E} \max_j |V_j| \leq \frac{1}{\lambda} \log \sum_{j=1}^D (\mathbb{E} e^{\lambda V_j} + \mathbb{E} e^{-\lambda V_j}) \leq \frac{\log(2D)}{\lambda} + \frac{\lambda n R^2}{2}.$$

优化 λ 即得结论。这里 D 个坐标可以高度相关, 并不需要坐标之间独立; 独立的只是人为引入的 σ_i 。

3.12 Maurey 的 empirical method: 用随机稀疏化构造覆盖

前面把学习问题变成了几何问题: 某个范数球在函数距离下到底有多大? Maurey 的方法进一步把几何问题变成均值估计问题。课程推荐讲义中针对 ℓ_1 球的构造[12], 可以写成如下 Hilbert 空间形式。

定理 3.20 (有限字典的随机稀疏逼近与覆盖)。设 $\psi_1, \dots, \psi_M$ 属于实 Hilbert 空间 H , $\|\psi_j\|_H \leq R$, $A, R > 0$ 。定义

$$\mathcal{A}_A = \left\{ \sum_{j=1}^M a_j \psi_j : \sum_{j=1}^M |a_j| \leq A \right\}.$$

对每个 $f \in \mathcal{A}_A$ 和整数 $m \geq 1$, 存在 $v_1, \dots, v_m \in \{0, \pm \psi_1, \dots, \pm \psi_M\}$, 使

$$\left\| f - \frac{A}{m} \sum_{k=1}^m v_k \right\|_H^2 \leq \frac{A^2 R^2}{m}. \quad (3.14)$$

进而对 $\varepsilon > 0$,

$$\log N(\mathcal{A}_A, \|\cdot\|_H, \varepsilon) \leq \left\lceil \frac{A^2 R^2}{\varepsilon^2} \right\rceil \log(2M + 1). \quad (3.15)$$

证明直觉 把系数的绝对值看成概率, 把系数的符号放进被抽取的字典元素中。整个函数因此成为一个随机字典元素的期望。用 m 次抽样的平均代替这个期望, 平方误差便像均值估计一样按 $1/m$ 衰减。随机化最关键的作用是让中心化后的交叉项期望为零, 而不是保证每次随机抽样都逼近得好。

证明. 固定表示 $f = \sum_j a_j \psi_j$ 。先抽取指标 $J \in \{0, 1, \dots, M\}$, 使

$$\mathbb{P}(J = j) = \frac{|a_j|}{A} \quad (j \geq 1), \quad \mathbb{P}(J = 0) = 1 - \frac{\sum_j |a_j|}{A}.$$

若 $J = 0$, 令 $U = 0$; 否则令 $U = A \text{sign}(a_J) \psi_J$ 。于是 $\mathbb{E}U = f$, 且 $\mathbb{E}\|U\|_H^2 \leq A^2 R^2$ 。取独立同分布的 $U_1, \dots, U_m$, 记 $\bar{U} = m^{-1} \sum_k U_k$ 。展开平方可得

$$\begin{aligned} \mathbb{E}\|\bar{U} - f\|_H^2 &= \frac{1}{m^2} \sum_{k,l} \mathbb{E}\langle U_k - f, U_l - f \rangle_H \\ &= \frac{1}{m} \mathbb{E}\|U - f\|_H^2 = \frac{1}{m} (\mathbb{E}\|U\|_H^2 - \|f\|_H^2) \leq \frac{A^2 R^2}{m}. \end{aligned}$$

其中 $k \neq l$ 时, 独立性与中心化给出 $\mathbb{E}\langle U_k - f, U_l - f \rangle_H = \langle \mathbb{E}(U_k - f), \mathbb{E}(U_l - f) \rangle_H = 0$ 。因此至少存在一个抽样结果达到不大于期望的误差, 证明了(3.14)。

现在把所有可能的 $A m^{-1} \sum_k v_k$ 收集成集合 $\mathcal{C}_m$ 。它的点数至多为 $(2M+1)^m$, 且每个点仍属于 $\mathcal{A}_A$ 。虽然每个 f 所需的抽样概率不同, $\mathcal{C}_m$ 本身并不依赖于 f 。取 $m = \lceil A^2 R^2 / \varepsilon^2 \rceil$, (3.14)说明 $\mathcal{C}_m$ 是整个 $\mathcal{A}_A$ 的 ε 网, 因而得到(3.15)。□

说明 3.21 (这个结论怎样使用)。 可以取 $H = L^2(P_X)$, 得到总体均方逼近; 也可以条件于样本取 $H = \mathbb{R}^n$, 内积为 $n^{-1} \sum_i u_i v_i$, 得到经验 L^2 覆盖。这些都是平均平方意义的逼近, 并未证明逐点一致逼近。字典数只通过 $\log M$ 出现, 但 A, R 可能依赖输入维数, 因此还不能据此宣称一般光滑函数类的维数灾难已经消失。本定理是存在性与覆盖数结论; 随机采样需要知道给定表示的系数, 并不自动提供从训练数据找到最优稀疏表示的算法。

3.13 从经验覆盖到 chaining: 相近的候选不必重复支付完整代价

现在可以严谨地使用样本相关的度量

$$d_S(g, h) = \sqrt{\frac{1}{n} \sum_{i=1}^n [g(Z_i) - h(Z_i)]^2}.$$

固定 S 后, 网的选择也固定, 剩下的随机性只有独立符号。若 $\sup_g d_S(g, 0) \leq Q$, 对一个经验 ε 网应用 Cauchy Schwarz 不等式和前面有限向量最大值的指数矩论证, 得到

$$\hat{\mathfrak{N}}_S(\mathcal{G}) \leq \varepsilon + Q \sqrt{\frac{2 \log\{2N(\mathcal{G}, d_S, \varepsilon)\}}{n}}. \quad (3.16)$$

具体地, $n^{-1} |\sum_i \sigma_i(g - \pi(g))(Z_i)| \leq d_S(g, \pi(g))$, 而每个网点的符号和的指数矩由其经验平方范数控制。本论证没有对未知的 $P(g - \pi(g))$ 作任何替换, 因而允许网依赖样本。

单一分辨率仍可能浪费信息: 两个非常接近的函数被当成了两项全新的竞争者。Chaining 的想法是先作粗定位, 再逐层修正; 每层只为小增量付费。假设 $0 \in \mathcal{G}$, 令 $\varepsilon_j = 2^{-j} Q$, 取相应的网 $\mathcal{G}_j$, 并记 $N_j = |\mathcal{G}_j|$ 。令 $\pi_0(g) = 0, N_0 = 1$, 将 g 写成

$$g = (g - \pi_J(g)) + \sum_{j=1}^J (\pi_j(g) - \pi_{j-1}(g)).$$

第 j 层至多出现 $N_j N_{j-1}$ 个增量, 且每个实际出现的增量的 d_S 范数至多为 $\varepsilon_j + \varepsilon_{j-1} = 3\varepsilon_j$ 。对每一层再次应用有限向量的最大值界, 得到完整的有限层结论

$$\hat{\mathfrak{N}}_S(\mathcal{G}) \leq \varepsilon_J + \sum_{j=1}^J 3\varepsilon_j \sqrt{\frac{2 \log(2N_j N_{j-1})}{n}}. \quad (3.17)$$

把分层求和与积分比较, 就通向 Dudley 的熵积分。本讲需要记住的证明机制是: 先用粗网承担大尺度变化, 再用细网描述范数很小的增量; 不能把每层增量都按整个函数类的半径估计。这也说明, 覆盖数、随机符号与随机稀疏化是可以相互衔接的三种工具, 而不是互不相关的三个术语 [7, 8, 15]。

4 局部复杂度：让曲率参与统计误差分析

课堂定位。课堂从均值估计出发：先比较全局复杂度界与显式解，再讲平方损失的基本恒等式、有限类的快速率，以及星形缩放的证明图景，最后用局部分析回到同一个均值问题。临界半径的完整概率论证与 RKHS 谱计算可在课后阅读。本节延续原课程 Lecture 11 的局部化思路[9]，将它写成假设明确、可以独立使用的有限样本结论。固定设计部分也可与陈育东的讲义对照[13]。

4.1 先看均值估计：同一个估计器为何出现两种速率

考虑最简单的统计任务。观测 $Y_1, \dots, Y_n$ 独立同分布，且 $Y_i \in [-1, 1]$ 几乎处处。记

$$m = \mathbb{E}Y, \quad \tau^2 = \text{Var}(Y), \quad \bar{Y} = \frac{1}{n} \sum_{i=1}^n Y_i.$$

用一个常数 $\theta \in [-1, 1]$ 预测 Y ，并采用平方损失

$$\ell_\theta(y) = (y - \theta)^2, \quad R(\theta) = \mathbb{E}\ell_\theta(Y), \quad R_n(\theta) = \frac{1}{n} \sum_{i=1}^n \ell_\theta(Y_i).$$

本小节以 P 和 P_n 分别表示对 Y 的总体平均与经验平均。将 $Y - \theta = (Y - m) - (\theta - m)$ 展开，并利用 $\mathbb{E}(Y - m) = 0$ ，得到

$$R(\theta) = \tau^2 + (\theta - m)^2, \quad R(\theta) - R(m) = (\theta - m)^2. \quad (4.1)$$

因此总体最优参数就是 m 。我们研究同一个经验风险最小化器

$$\hat{\theta} \in \arg \min_{\theta \in [-1, 1]} R_n(\theta).$$

参数误差与风险误差差一个平方。以下用 r 表示参数误差半径，用 $q = r^2$ 表示超额风险的尺度。

第一种分析：直接使用全局复杂度。先完全照搬上一节的统一泛化界。完整平方损失类为 $\mathcal{L} = \{\ell_\theta : \theta \in [-1, 1]\}$ ，每个损失都取值于 $[0, 4]$ 。固定观测样本，取独立 Rademacher 符号 ε_i 。按本讲带绝对值的约定，展开平方后有

$$\begin{aligned} \hat{\mathfrak{R}}_S(\mathcal{L}) &= \mathbb{E}_\varepsilon \sup_{|\theta| \leq 1} \left| \frac{1}{n} \sum_{i=1}^n \varepsilon_i (\theta^2 - 2\theta Y_i + Y_i^2) \right| \\ &\leq \frac{1}{n} \mathbb{E}_\varepsilon \left| \sum_i \varepsilon_i \right| + \frac{2}{n} \mathbb{E}_\varepsilon \left| \sum_i \varepsilon_i Y_i \right| + \frac{1}{n} \mathbb{E}_\varepsilon \left| \sum_i \varepsilon_i Y_i^2 \right| \\ &\leq \frac{\sqrt{n} + 2\sqrt{\sum_i Y_i^2} + \sqrt{\sum_i Y_i^4}}{n} \leq \frac{4}{\sqrt{n}}. \end{aligned} \quad (4.2)$$

倒数第二步只用了 $\mathbb{E}_\varepsilon |\sum_i \varepsilon_i b_i| \leq \sqrt{\sum_i b_i^2}$ ：展开平方时，符号的交叉项期望为零。上界对每个样本都成立，故总体平均复杂度同样至多为 $4/\sqrt{n}$ 。

取任意 $\delta \in (0, 1)$ 。将损失上界 $B = 4$ 代入定理 3.16 的第一个版本，以至少 $1 - \delta$ 的概率有

$$\sup_{|\theta| \leq 1} |R_n(\theta) - R(\theta)| \leq \frac{8}{\sqrt{n}} + 4\sqrt{\frac{\log(1/\delta)}{2n}}.$$

ERM 的类内超额风险至多为上述统一偏差的两倍，因而

$$(\hat{\theta} - m)^2 \leq \frac{C_\delta}{\sqrt{n}}, \quad |\hat{\theta} - m| \leq \sqrt{C_\delta} n^{-1/4}, \quad C_\delta = 16 + 8\sqrt{\frac{\log(1/\delta)}{2}}. \quad (4.3)$$

固定置信度时，这条全局分析只给出参数误差 $O(n^{-1/4})$ 。到这里每一步都是正确的；我们还需要判断这个上界是否反映了估计器本身的精度。

第二种分析：先把 ERM 显式求出来。对同一个经验目标作平方配方：

$$R_n(\theta) = (\theta - \bar{Y})^2 + \frac{1}{n} \sum_{i=1}^n (Y_i - \bar{Y})^2.$$

因为 $\bar{Y} \in [-1, 1]$ ，唯一最小点就是

$$\hat{\theta} = \bar{Y}. \quad (4.4)$$

对区间长度为 2 的独立观测使用 Hoeffding 不等式，得到

$$\mathbb{P}\{|\bar{Y} - m| \geq t\} \leq 2 \exp\{-nt^2/2\}, \quad t > 0.$$

记

$$a_n(\delta) = \sqrt{\frac{2 \log(2/\delta)}{n}}. \quad (4.5)$$

于是以至少 $1 - \delta$ 的概率，

$$|\hat{\theta} - m| \leq a_n(\delta), \quad R(\hat{\theta}) - R(m) \leq a_n(\delta)^2. \quad (4.6)$$

参数误差现在是 $O(n^{-1/2})$ ，超额风险是 $O(n^{-1})$ 。更直接地，独立性给出精确的均方误差

$$\mathbb{E}(\hat{\theta} - m)^2 = \frac{\tau^2}{n}. \quad (4.7)$$

若分布固定且 $\tau^2 > 0$ ，均方根误差恰好为 $\tau/\sqrt{n}$ 。这说明先前的 $n^{-1/4}$ 是泛化分析给出的松上界。

差别出在哪里：二次代价与随机倾斜。令 $a = \theta - m$ 。对比 θ 与 m 的损失时，公共项 $(Y - m)^2$ 消失，只留下

$$\mathcal{L}_a(Y) := \ell_{m+a}(Y) - \ell_m(Y) = a^2 - 2a(Y - m). \quad (4.8)$$

因此总体与经验的超额风险分别为

$$P\mathcal{L}_a = \underbrace{a^2}_{\text{二次代价}}, \quad P_n\mathcal{L}_a = a^2 - \underbrace{2a(\bar{Y} - m)}_{\text{随机线性倾斜}}. \quad (4.9)$$

证明直觉 偏离最优点的距离为 r 时，确定性的风险代价是 r^2 。随机线性倾斜的大小却是 r 乘上均值波动，因而只有 $r/\sqrt{n}$ 的量级。全局界用一个不随 r 缩小的统一数去控制随机扰动，于是只剩下 r^2 与 $n^{-1/2}$ 的比较。局部分析应当保留扰动中的距离因子，让 r^2 与 $r/\sqrt{n}$ 比较。快率来自二次增长与局部波动的联合使用。

图中 $\varepsilon_n \asymp n^{-1/2}$ 表示固定置信度时的量级，省略了常数。两个交点对应同一个估计器的两种分析上界。如果把整个损失乘以一个正常数，二次代价与随机倾斜也会同时乘上这个常数，估计器和关于 n 的指数都不变；曲率数值变大本本身不产生更快的统计速率。接下来先建立一般函数类的局部分分析方法，再回到这个均值问题，仅从 ERM 性质恢复 $n^{-1/2}$ 的参数精度。

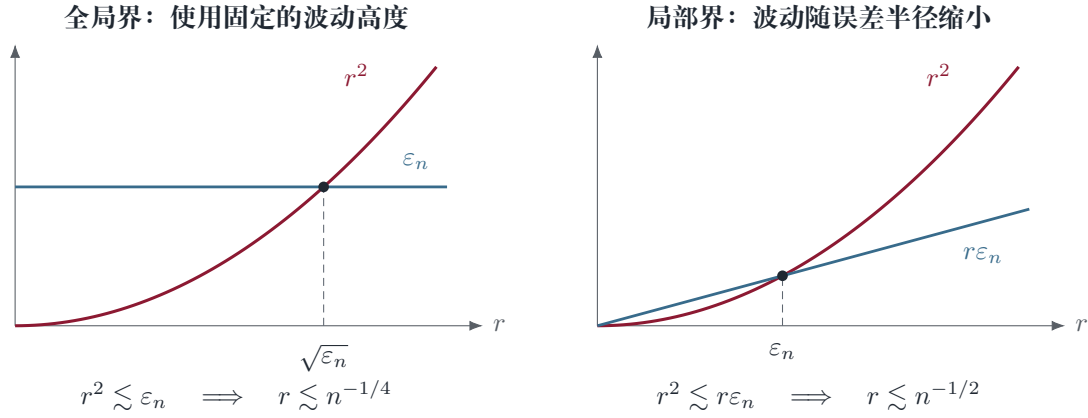


图 3: 平方损失的曲率如何改变可证明的误差半径。这里 $\varepsilon_n \asymp n^{-1/2}$ ，省略固定置信度与常数。红色曲线为风险代价，蓝色曲线为相应的随机波动界。全局分析把波动固定在整个参数区间的尺度；局部分析保留其随 r 缩小的线性因子，因而能在更小的半径处排除错误候选。图中高度仅为量级示意。

4.2 平方损失同时给出曲率和较小的方差

现在把均值问题扩展为随机设计回归。设 $Z_i = (X_i, Y_i)$ 独立同分布，以 $Ph = \mathbb{E}[h(X, Y)]$ 、 $P_n h = n^{-1} \sum_{i=1}^n h(X_i, Y_i)$ 表示总体平均与经验平均。记

$$f^*(x) = \mathbb{E}[Y | X = x], \quad \xi = Y - f^*(X), \quad g = f - f^*.$$

本小节假设 $f^* \in \mathcal{F}$ ，即真实回归函数属于候选类。对平方损失定义超额损失函数

$$\mathcal{L}_f(X, Y) = (Y - f(X))^2 - (Y - f^*(X))^2.$$

展开平方得到整个局部分分析的起点：

$$\mathcal{L}_f = g^2 - 2\xi g, \quad P\mathcal{L}_f = Pg^2 =: q_f. \quad (4.10)$$

第二个等号来自 $\mathbb{E}[\xi | X] = 0$ 。这里 q_f 是超额风险，而 $\mathcal{L}_f(X, Y)$ 本身可以为负。

证明直觉 离真实函数的距离扩大一倍，确定性的代价 Pg^2 扩大四倍。相应的噪声项对 g 是线性的，只扩大两倍。证明应保留这两个不同的缩放规律。若一开始就用整个函数类上的统一绝对偏差替换它们，噪声会被固定成一个与距离无关的数，平方损失的曲率就没有参与最后的比较。

命题 4.1 (超额损失的 Bernstein 方差条件)。若 $|Y| \leq B$ ，且每个 $f \in \mathcal{F}$ 都满足 $|f(X)| \leq B$ 几乎处处，则 $|f^*(X)| \leq B$ ，并且

$$\text{Var}(\mathcal{L}_f) \leq P\mathcal{L}_f^2 \leq 16B^2 q_f, \quad |\mathcal{L}_f - q_f| \leq 8B^2. \quad (4.11)$$

证明. 使用平方差分解比逐项展开四次方更直接：

$$\mathcal{L}_f = (f - f^*)(f + f^* - 2Y), \quad |\mathcal{L}_f| \leq 4B|f - f^*|.$$

平方后取期望，便有 $P\mathcal{L}_f^2 \leq 16B^2 P(f - f^*)^2$ 。此外，两项平方损失都在 $[0, 4B^2]$ 内，所以 $|\mathcal{L}_f| \leq 4B^2$ 。又因 $0 \leq q_f \leq 4B^2$ ，中心化后的绝对值不超过 $8B^2$ 。□

这个条件揭示了快速率的来源：风险越小，超额损失的方差也越小。通常称 $\text{Var}(\mathcal{L}_f) \leq CP\mathcal{L}_f$ 为 Bernstein 条件。它是模型的性质，与作为概率工具的 Bernstein 不等式是两个不同的概念。

定理 4.2 (有限函数类的快速率). 在上述有界回归假设下, 设 $|\mathcal{F}| = M < \infty$ 、 $f^* \in \mathcal{F}$, 并令

$$\hat{f} \in \arg \min_{f \in \mathcal{F}} P_n(Y - f(X))^2.$$

对任意 $\delta \in (0, 1)$, 以至少 $1 - \delta$ 的概率,

$$P(\hat{f} - f^*)^2 \leq \frac{40B^2 \log(M/\delta)}{n}. \quad (4.12)$$

证明直觉 考虑一个风险为 q 的候选函数。它若想在训练样本上胜过 f^* , 就必须依靠一个大小至少为 q 的负偏差。其方差只有 $O(B^2 q)$, 因此这样的伪装概率是 $\exp\{-cnq/B^2\}$ 。让这个概率小到足以同时排除 M 个候选函数, 只需要 nq/B^2 大于 $\log M$ 。

证明. 固定一个 $f \in \mathcal{F}$, 且 $q_f > 0$ 。把定理 3.7 用于独立中心化变量 $q_f - \mathcal{L}_f(Z_i)$ 。由命题 4.1, 其方差至多为 $16B^2 q_f$, 绝对值至多为 $8B^2$ 。于是

$$\begin{aligned} \mathbb{P}\{P_n \mathcal{L}_f \leq 0\} &\leq \exp\left\{-\frac{nq_f^2}{2\{16B^2 q_f + (8B^2)q_f/3\}}\right\} \\ &\leq \exp\left\{-\frac{nq_f}{40B^2}\right\}. \end{aligned}$$

令 $a = 40B^2 \log(M/\delta)/n$ 。对所有满足 $q_f > a$ 的候选函数取并集, 得到

$$\mathbb{P}\{\exists f \in \mathcal{F} : q_f > a, P_n \mathcal{L}_f \leq 0\} \leq M \exp\left\{-\frac{na}{40B^2}\right\} = \delta.$$

而经验风险最小化与 $f^* \in \mathcal{F}$ 保证 $P_n \mathcal{L}_{\hat{f}} \leq 0$ 。因此在上述坏事件之外, $q_{\hat{f}} \leq a$ 。□

说明 4.3 (为什么这里是 n^{-1}). 全局有界损失的 Hoeffding 分析通常给出 $B^2 \sqrt{\log(M/\delta)/n}$ 。这里在风险为 q 的位置, 波动的主要量级变成 $B\sqrt{q \log(M/\delta)/n}$ 。把它与确定性代价 q 比较, 就得到 $q \asymp B^2 \log(M/\delta)/n$ 。这一推理依赖方差条件与 $f^* \in \mathcal{F}$; 对任意损失、任意模型失配情形, 不能直接沿用这个快速率。

4.3 固定设计: 把随机误差写成一个线性过程

下面改用一个便于完整展示局部化技术的模型。设计点 $x_1, \dots, x_n$ 固定, 观测为

$$Y_i = f^*(x_i) + \sigma Z_i, \quad Z_i \stackrel{\text{iid}}{\sim} N(0, 1), \quad \sigma > 0. \quad (4.13)$$

本模型不再假设响应 Y_i 有界。仍假设 $f^* \in \mathcal{F}$, 并令 $\hat{f}$ 为平方损失的 ERM。写成函数值向量后, 误差集合为

$$\mathcal{V} = \{(f(x_i) - f^*(x_i))_{i=1}^n : f \in \mathcal{F}\} \subseteq \mathbb{R}^n.$$

假设 $\mathcal{V}$ 闭且关于原点星形: $v \in \mathcal{V}$ 、 $a \in [0, 1]$ 蕴含 $av \in \mathcal{V}$ 。闭性保证有限维最小二乘解存在, 也使下文的局部上确界可测。若 $\mathcal{F}$ 为凸集且包含 f^* , 则星形条件自动成立。

记

$$\langle u, v \rangle_n = \frac{1}{n} \sum_{i=1}^n u_i v_i, \quad \|v\|_n^2 = \frac{1}{n} \sum_{i=1}^n v_i^2, \quad \hat{v} = (\hat{f}(x_i) - f^*(x_i))_{i=1}^n.$$

比较 $\hat{f}$ 与 f^* 的经验损失并展开平方, 直接得到

$$\underbrace{\|\hat{v}\|_n^2}_{\text{二次代价}} \leq \underbrace{2\sigma \langle Z, \hat{v} \rangle_n}_{\text{线性噪声收益}}. \quad (4.14)$$

这称为最小二乘的基本不等式。它不涉及任何概率界, 是 ERM 定义的确定性结果。

引理 4.4 (一个球面就能排除所有远处的候选点)。设 $\mathcal{V}$ 关于原点星形, $\hat{v} \in \mathcal{V}$ 满足基本不等式 (4.14)。若某个 $r > 0$ 满足

$$\sigma \sup_{v \in \mathcal{V}: \|v\|_n = r} |\langle Z, v \rangle_n| \leq \frac{r^2}{4}, \quad (4.15)$$

则 $\|\hat{v}\|_n \leq r$ 。若该球面与 $\mathcal{V}$ 无交点, 结论同样成立。

证明直觉 局部化并不预先假设估计器已经接近真实函数。证明从一个远处的候选点出发, 把它缩回固定球面。只要这个球面上噪声收益已经小于二次代价, 射线上更远的位置就更不可能满足 ERM 的基本不等式。这是一个排除论证, 因此没有循环推理。

证明. 假设 $t = \|\hat{v}\|_n > r$ 。将这个候选点沿射线缩回球面:

$$u = \frac{r}{t} \hat{v} \in \mathcal{V}, \quad \|u\|_n = r.$$

星形条件保证缩放后仍然可行。线性噪声项满足

$$\sigma |\langle Z, \hat{v} \rangle_n| = \frac{t}{r} \sigma |\langle Z, u \rangle_n| \leq \frac{tr}{4}.$$

基本不等式便要求 $t^2 \leq tr/2$, 从而 $t \leq r/2$, 与 $t > r$ 矛盾。如果球面没有可行点, 星形性本身就排除了半径大于 r 的可行点。□

4.4 临界半径: 复杂度与曲率在哪里相交

定义 4.5 (局部高斯复杂度)。给定 $r > 0$, 定义

$$T_r(Z) = \sup_{v \in \mathcal{V}: \|v\|_n \leq r} |\langle Z, v \rangle_n|, \quad w_n(r) = \mathbb{E}_Z T_r(Z).$$

这里的期望只对标准高斯向量 Z 取。实际噪声过程的期望是 $\sigma w_n(r)$ 。

引理 4.6 (星形缩放控制复杂度的增长)。对 $0 < r \leq s$, 有

$$\frac{w_n(s)}{s} \leq \frac{w_n(r)}{r}. \quad (4.16)$$

证明. 对每个 $v \in \mathcal{V} \cap \{\|v\|_n \leq s\}$, 星形性保证 $(r/s)v \in \mathcal{V} \cap \{\|v\|_n \leq r\}$ 。因此逐个 Z 都有 $(r/s)T_s(Z) \leq T_r(Z)$ 。取期望即得结论。□

引理 4.7 (高斯 Lipschitz 集中)。若 $Z \sim N(0, I_n)$, 且 $H: \mathbb{R}^n \rightarrow \mathbb{R}$ 满足 $|H(z) - H(z')| \leq L\|z - z'\|_2$, 其中 $L > 0$, 则对任意 $a > 0$,

$$\mathbb{P}\{H(Z) \geq \mathbb{E}H(Z) + L\sqrt{2a}\} \leq e^{-a}.$$

其完整证明见附录 B。

定理 4.8 (临界半径控制固定设计预测误差)。 在模型 (4.13) 与上述闭性、星形假设下, 取任意确定的 $r_* > 0$, 使得

$$\sigma w_n(r_*) \leq \frac{r_*^2}{8}. \quad (4.17)$$

则对任意 $\delta \in (0, 1)$, 以至少 $1 - \delta$ 的概率,

$$\frac{1}{n} \sum_{i=1}^n (\hat{f}(x_i) - f^*(x_i))^2 \leq r_*^2 + \frac{128\sigma^2 \log(1/\delta)}{n}. \quad (4.18)$$

这里所有概率都只来自观测噪声。

证明. 第一步: 固定半径的集中。 对任意 $z, z' \in \mathbb{R}^n$, 由 Cauchy 不等式,

$$|T_r(z) - T_r(z')| \leq \sup_{\|v\|_n \leq r} |\langle z - z', v \rangle_n| \leq \frac{r}{\sqrt{n}} \|z - z'\|_2.$$

所以引理 4.7 给出

$$\sigma T_r(Z) \leq \sigma w_n(r) + \sigma r \sqrt{\frac{2 \log(1/\delta)}{n}} \quad (4.19)$$

的概率至少为 $1 - \delta$ 。

第二步: 选一个确定的检验半径。 令

$$r = \max \left\{ r_*, 8\sigma \sqrt{\frac{2 \log(1/\delta)}{n}} \right\}.$$

这个半径由设计、模型与置信度确定, 不依赖已经实现的噪声。由引理 4.6 和临界半径条件,

$$\sigma w_n(r) \leq \frac{r}{r_*} \sigma w_n(r_*) \leq \frac{r r_*}{8} \leq \frac{r^2}{8}.$$

同时 $\sigma r \sqrt{2 \log(1/\delta)/n} \leq r^2/8$ 。故在事件 (4.19) 上, $\sigma T_r(Z) \leq r^2/4$ 。

第三步: 排除半径之外的 ERM。 球内上确界控制了球面上确界, 引理 4.4 因此给出 $\|\hat{v}\|_n \leq r$ 。最后利用 $r^2 \leq r_*^2 + 128\sigma^2 \log(1/\delta)/n$ 。□

说明 4.9 (固定点到底是哪两个量相等)。 可以把满足 (4.17) 的较小半径称为临界半径。本定理使用任意满足条件的半径, 从而无须讨论下确界是否取得。这样的半径总是存在: Cauchy 不等式给出 $w_n(r) \leq r \mathbb{E} \|Z\|_2 / \sqrt{n} \leq r$, 所以 $r = 8\sigma$ 已经可行。核心平衡是

$\text{平方误差 } r^2 \asymp \text{噪声强度 } \sigma \times \text{局部复杂度 } w_n(r).$

半径 r 与风险 r^2 不应混用; 同样, $w_n(r)$ 的定义中没有包含 σ 。

例 4.10 (线性子空间给出参数个数除以样本量)。 若 $\mathcal{V}$ 是一个 $p \geq 1$ 维线性子空间, 用 $\Pi_{\mathcal{V}}$ 表示欧氏正交投影, 则

$$T_r(Z) = \frac{r}{\sqrt{n}} \|\Pi_{\mathcal{V}} Z\|_2, \quad w_n(r) \leq r \sqrt{\frac{p}{n}}.$$

后一式由 Jensen 不等式和 $\mathbb{E} \|\Pi_{\mathcal{V}} Z\|_2^2 = p$ 得到。取 $r_* = 8\sigma \sqrt{p/n}$, 定理给出

$$\|\hat{v}\|_n^2 \leq \frac{64\sigma^2 p + 128\sigma^2 \log(1/\delta)}{n}.$$

局部复杂度中的 r 与左边的 r^2 平衡, 自然恢复了最小二乘的 p/n 量级。 $p = 0$ 时可行误差只有零向量, 结论直接成立。

4.5 回到均值估计：局部复杂度恢复正确速率

现在回到第 4.1 小节的同一个有界均值模型： Y_i 独立同分布于 $[-1, 1]$ ， $m = \mathbb{E}Y$ ，参数集合为 $[-1, 1]$ 。这一遍从 ERM 的基本不等式出发完成证明。

先把局部复杂度真正算出来。参数误差 $a = \theta - m$ 位于区间 $I_m = [-1 - m, 1 - m]$ 。对半径 $r > 0$ ，考虑常数误差函数的局部类

$$\mathcal{G}_r = \{g_a : y \mapsto a : a \in I_m, |a| \leq r\}.$$

其经验 Rademacher 复杂度满足

$$\widehat{\mathfrak{R}}_S(\mathcal{G}_r) = \mathbb{E}_\varepsilon \sup_{\substack{a \in I_m \\ |a| \leq r}} \left| \frac{a}{n} \sum_i \varepsilon_i \right| \leq \frac{r}{n} \mathbb{E}_\varepsilon \left| \sum_i \varepsilon_i \right| \leq \frac{r}{\sqrt{n}}. \quad (4.20)$$

局部复杂度随参数半径线性缩小，正好保留了开头全局分析丢掉的距离因子。

也可以直接计算超额损失的局部类 $\mathcal{L}_r = \{\mathcal{L}_a : a \in I_m, |a| \leq r\}$ 。由 $\mathcal{L}_a(Y) = a^2 - 2a(Y - m)$,

$$\begin{aligned} \widehat{\mathfrak{R}}_S(\mathcal{L}_r) &\leq \frac{r^2}{n} \mathbb{E}_\varepsilon \left| \sum_i \varepsilon_i \right| + \frac{2r}{n} \mathbb{E}_\varepsilon \left| \sum_i \varepsilon_i (Y_i - m) \right| \\ &\leq \frac{r^2}{\sqrt{n}} + \frac{2r}{n} \sqrt{\sum_i (Y_i - m)^2} \leq \frac{r^2 + 4r}{\sqrt{n}}. \end{aligned} \quad (4.21)$$

再对样本取期望，还可以用 $\tau^2 = \text{Var}(Y)$ 写成 $\mathfrak{R}_n(\mathcal{L}_r) \leq (r^2 + 2\tau r)/\sqrt{n}$ 。不过， a^2 是关于观测的常数，在实际经验过程 $(P - P_n)\mathcal{L}_a$ 中完全消失。若直接中心化， $\mathcal{L}_a - P\mathcal{L}_a = -2a(Y - m)$ ，同一计算给出

$$\mathfrak{R}_n(\{\mathcal{L}_a - P\mathcal{L}_a : a \in I_m, |a| \leq r\}) \leq \frac{2\tau r}{\sqrt{n}}, \quad \tau \leq 1.$$

因此，本例能比一般的对称化界更直接地控制局部波动。

命题 4.11 (从局部波动恢复均值估计的速率)。对任意 $\delta \in (0, 1)$ ，令 $a_n(\delta)$ 如式 (4.5)。以至少 $1 - \delta$ 的概率，平方损失的 ERM 满足

$$|\widehat{\theta} - m| \leq 2a_n(\delta), \quad R(\widehat{\theta}) - R(m) \leq 4a_n(\delta)^2 = \frac{8 \log(2/\delta)}{n}. \quad (4.22)$$

证明直觉 只需控制一次均值噪声。同一个噪声事件就控制所有参数 a 、所有半径 r 的局部波动。然后把 ERM 的实际误差代入这个已经同时成立的不等式，让二次代价与线性噪声收益直接比较。这样既保留了局部复杂度的缩放，也避免了先假设估计器已在某个小球里的循环推理。

证明. 由 Hoeffding 不等式，事件

$$\mathcal{E}_\delta = \{|\bar{Y} - m| \leq a_n(\delta)\}$$

的概率至少为 $1 - \delta$ 。由式 (4.8)，对每个 $a \in I_m$ 都有精确恒等式

$$(P - P_n)\mathcal{L}_a = 2a(\bar{Y} - m).$$

所以在同一个事件 $\mathcal{E}_\delta$ 上, 同时对所有 a 与所有 $r > 0$,

$$|(P - P_n)\mathcal{L}_a| \leq 2|a|a_n(\delta), \quad \sup_{\substack{a \in I_m \\ |a| \leq r}} |(P - P_n)\mathcal{L}_a| \leq 2ra_n(\delta). \quad (4.23)$$

这不是把随机半径代入只对固定半径成立的概率界; 一个事件已经同时控制了全部半径。

记 $\hat{a} = \hat{\theta} - m$ 。由于 m 是可行参数, ERM 性质给出 $P_n\mathcal{L}_{\hat{a}} \leq 0$ 。因此在 $\mathcal{E}_\delta$ 上,

$$\hat{a}^2 = P\mathcal{L}_{\hat{a}} = P_n\mathcal{L}_{\hat{a}} + (P - P_n)\mathcal{L}_{\hat{a}} \leq 2|\hat{a}|a_n(\delta).$$

若 $\hat{a} = 0$, 结论直接成立; 否则除以 $|\hat{a}|$, 得到 $|\hat{a}| \leq 2a_n(\delta)$ 。平方即得风险界。 $\square$

这个证明只用了局部波动与 ERM 性质。所得参数界与显式解 (4.6) 相差常数 2, 关于 n 的速率已经完全匹配。用参数半径 r 与风险尺度 $q = r^2$ 写出平衡, 便是

$$r^2 \asymp \frac{r}{\sqrt{n}} \implies r \asymp n^{-1/2}, \quad q \asymp \frac{\sqrt{q}}{\sqrt{n}} \implies q \asymp n^{-1}.$$

这里仍省略固定置信度下的常数。全局分析、显式求解与局部分析始终研究同一个估计器; 局部分析恢复的是这个估计器原本就具有的精度。

直接代入高斯临界半径定理。作为对一般定理的单独核对, 这一段改用 $Y_i = m + \sigma Z_i$ 、 $Z_i \stackrel{\text{iid}}{\sim} N(0, 1)$ 、 $\sigma > 0$, 并令参数集合为 $\mathbb{R}$ 。此处采用高斯噪声模型, 不使用前面有界响应的假设。误差向量集合为 $\mathcal{V} = \{a\mathbf{1} : a \in \mathbb{R}\}$, 它是闭线性子空间。由 $n^{-1} \sum_i Z_i \sim N(0, 1/n)$,

$$w_n(r) = r\mathbb{E} \left| \frac{1}{n} \sum_i Z_i \right| = r\sqrt{\frac{2}{\pi n}}.$$

取 $r_* = 8\sigma\sqrt{2/(\pi n)}$, 恰好满足 $\sigma w_n(r_*) = r_*^2/8$ 。定理 4.8 因而给出

$$(\hat{\theta} - m)^2 \leq \frac{128\sigma^2}{n} \left\{ \frac{1}{\pi} + \log(1/\delta) \right\}$$

的概率至少为 $1 - \delta$ 。这正是一般临界半径理论在一个常数方向上的完整实例。

5 非参统计：把附近的数据平均起来

前面的复杂度分析问：一个函数类能在多大程度上追随随机噪声？现在把问题落实到一个具体算法。若附近的输入有相近的回归函数值, 就可以用附近的观测回答当前位置的问题。于是需要决定：附近究竟有多近, 附近究竟有多少数据？本节对应原课程第三讲的局部平均与核平滑内容 [4], 下面给出包含随机分母和边界的完整证明。

5.1 局部平均的两个尺度

设训练样本 $(X_i, Y_i)_{i=1}^n$ 独立同分布, 且

$$Y_i = f^*(X_i) + \xi_i, \quad \mathbb{E}[\xi_i | X_i] = 0, \quad \mathbb{E}[\xi_i^2 | X_i] \leq \sigma^2. \quad (5.1)$$

这里的条件零均值很关键：它说明在每一个输入位置, 噪声平均以后都不会留下系统误差。单有 $\mathbb{E}\xi_i = 0$ 不能保证这一点。

为看清证明，先取最简单的方盒核。对 $x \in [0, 1]^d$ 和带宽 $0 < h \leq 1$ ，定义

$$A_h(x) = \{z \in [0, 1]^d : \|z - x\|_\infty \leq h\}, \quad N_h(x) = \sum_{i=1}^n \mathbf{1}\{X_i \in A_h(x)\}.$$

局部平均估计量为

$$\hat{f}_h(x) = \begin{cases} \frac{1}{N_h(x)} \sum_{i: X_i \in A_h(x)} Y_i, & N_h(x) > 0, \\ 0, & N_h(x) = 0. \end{cases} \quad (5.2)$$

最后一行使算法在每个样本上都有定义。它产生的误差将单独由空邻域概率控制。

证明直觉 固定训练输入以后，算法做的事情只剩下平均。如果平均了 N 个条件独立的噪声，方差至多为 σ^2/N 。但参与平均的函数值来自距离 x 至多为 h 的位置，它们与 $f^*(x)$ 之间有平滑性误差。因此证明应当先固定输入，分别处理函数变化和噪声，再对输入的随机性取期望。

5.2 随机分母的一个小技巧

引理 5.1 (二项计数的倒数)。若 $N \sim \text{Bin}(n, p)$ ，其中 $p > 0$ ，则

$$\mathbb{E} \left[\frac{\mathbf{1}\{N > 0\}}{N \vee 1} \right] \leq \frac{2}{(n+1)p}, \quad \mathbb{P}(N = 0) = (1-p)^n \leq e^{-np}.$$

证明直觉 这里不需要证明 N 在每个样本上都接近 np 。我们真正要控制的是倒数的期望，所以直接寻找适合倒数的恒等式。证明中需要哪个随机量，就针对哪个随机量设计工具，通常比先证明一个更强的集中事件更短。

证明。不能直接把 $\mathbb{E}(1/N)$ 换成 $1/\mathbb{E}N$ ，而且 N 可能等于零。对整数 $k \geq 1$ ，用

$$\frac{1}{k} \leq \frac{2}{k+1}$$

把分母平移一格。这个平移的好处是二项式系数恰好可以求和：

$$\begin{aligned} \mathbb{E} \frac{1}{N+1} &= \sum_{k=0}^n \frac{1}{k+1} \binom{n}{k} p^k (1-p)^{n-k} \\ &= \frac{1}{(n+1)p} \sum_{k=0}^n \binom{n+1}{k+1} p^{k+1} (1-p)^{n-k} \\ &= \frac{1 - (1-p)^{n+1}}{(n+1)p} \leq \frac{1}{(n+1)p}. \end{aligned}$$

因此第一式成立。第二式由 $1-p \leq e^{-p}$ 得到。 □

5.3 核平滑的完整误差界

定理 5.2 (方盒核的偏差与方差)。在模型 (5.1) 下, 假设 X 的分布在 $[0, 1]^d$ 上有密度 q , 且 $q(x) \geq q_0 > 0$ 几乎处处成立。设 $0 < \beta \leq 1$, 并且

$$|f^*(x) - f^*(z)| \leq L \|x - z\|_\infty^\beta, \quad \sup_x |f^*(x)| \leq M.$$

对每一个 $x \in [0, 1]^d$, 记 $p_h(x) = \mathbb{P}\{X \in A_h(x)\}$ 。则

$$\mathbb{E}[(\hat{f}_h(x) - f^*(x))^2] \leq L^2 h^{2\beta} + \frac{2\sigma^2}{(n+1)p_h(x)} + M^2 e^{-np_h(x)}. \quad (5.3)$$

特别地, $p_h(x) \geq q_0 h^d$, 所以

$$\mathbb{E} \left\| \hat{f}_h - f^* \right\|_{L^2(P_X)}^2 \leq L^2 h^{2\beta} + \frac{2\sigma^2}{(n+1)q_0 h^d} + M^2 e^{-nq_0 h^d}. \quad (5.4)$$

上式中的期望对整个训练样本取, $L^2(P_X)$ 积分对应独立的新输入。

证明. 第一步: 固定输入, 分开函数变化与噪声。 固定 x , 并条件于 $X_{1:n} = (X_1, \dots, X_n)$ 。在事件 $N_h(x) > 0$ 上写成

$$\hat{f}_h(x) - f^*(x) = \underbrace{\frac{1}{N_h(x)} \sum_{i: X_i \in A_h(x)} [f^*(X_i) - f^*(x)]}_{b_h(x; X_{1:n})} + \underbrace{\frac{1}{N_h(x)} \sum_{i: X_i \in A_h(x)} \xi_i}_{\eta_h(x)}.$$

第一项在固定输入后是确定量, 而且

$$|b_h(x; X_{1:n})| \leq L h^\beta.$$

样本对的独立性使噪声在条件于全部输入后仍独立。由条件零均值, 交叉项的条件期望为零, 并且

$$\mathbb{E}[\eta_h(x)^2 | X_{1:n}] = \frac{1}{N_h(x)^2} \sum_{i: X_i \in A_h(x)} \mathbb{E}[\xi_i^2 | X_i] \leq \frac{\sigma^2}{N_h(x)}.$$

这一步直接解释了平方偏差和方差分别来自哪里。

第二步: 把空邻域也放回去。 在 $N_h(x) = 0$ 时, 算法输出零, 平方误差就是 $f^*(x)^2$ 。因此

$$\begin{aligned} \mathbb{E}[(\hat{f}_h(x) - f^*(x))^2] &\leq L^2 h^{2\beta} + \sigma^2 \mathbb{E} \left[\frac{\mathbf{1}\{N_h(x) > 0\}}{N_h(x) \vee 1} \right] \\ &\quad + M^2 \mathbb{P}\{N_h(x) = 0\}. \end{aligned}$$

现在 $N_h(x) \sim \text{Bin}(n, p_h(x))$, 应用引理 5.1 即得 (5.3)。

第三步: 检查边界, 再积分。 即使 x 位于立方体的边界, 每个坐标方向上 $[x_j - h, x_j + h] \cap [0, 1]$ 的长度仍至少为 h , 故

$$p_h(x) = \int_{A_h(x)} q(z) dz \geq q_0 \text{vol}(A_h(x)) \geq q_0 h^d.$$

将它代入逐点界, 再对 $x \sim P_X$ 积分, 由非负函数的积分交换定理得到 (5.4)。□

说明 5.3. 证明中的 $L^2 h^{2\beta}$ 控制的是非空邻域上的条件平滑偏差。随机设计带来的波动已经包含在这个逐样本有效的控制中；空邻域造成的误差另列一项。这样就无需把随机分母错误地当成确定数，也无需把全部随机性都归入观测噪声。

推论 5.4 (带宽选择与维数灾难). 在定理 5.2 的条件下，取

$$h = n^{-1/(2\beta+d)},$$

则存在仅依赖于 $L, M, q_0, \sigma, \beta, d$ 的常数 C ，使

$$\mathbb{E} \left\| \hat{f}_h - f^* \right\|_{L^2(P_X)}^2 \leq C n^{-2\beta/(2\beta+d)}.$$

证明直觉 带宽 h 把空间分成大约 h^{-d} 个可分辨区域，每个区域平均获得约 nh^d 个样本。因此方差可写成

$$\frac{1}{nh^d} = \frac{\text{可分辨区域数}}{\text{总样本数}}.$$

这与有限维回归的维数除以样本数是同一种计数。平滑性则决定每个区域用常数近似要付多少偏差。若用 $m \asymp h^{-d}$ 表示模型分辨率，同一个平衡变成

$$\underbrace{m^{-2\beta/d}}_{\text{平方逼近误差}} + \underbrace{m/n}_{\text{统计误差}}.$$

维数同时出现在逼近的困难和获得足够邻居的困难之中。

证明. 两个主要误差项为 $h^{2\beta}$ 和 $1/(nh^d)$ 。令它们同阶，得到 $h^{2\beta+d} = 1/n$ 。此时 $nh^d = n^{2\beta/(2\beta+d)}$ ，空邻域项指数衰减，也被所述多项式界控制。□

这里的 $n^{-2\beta/(2\beta+d)}$ 是平方误差的速率，均方根误差是 $n^{-\beta/(2\beta+d)}$ 。因而达到均方根精度 ε 的充分样本量阶为 $\varepsilon^{-(2+d/\beta)}$ 。本节证明的是具体估计量的上界；要把它称为极小极大最优速率，还需要对给定噪声模型和函数类证明匹配下界。

5.4 一般核与更高阶平滑性

一般的局部平均把硬截断换成权重。给定非负平滑核 K ，可写

$$w_i(x) = \frac{K((x - X_i)/h)}{\sum_{j=1}^n K((x - X_j)/h)}, \quad \hat{f}_h(x) = \sum_{i=1}^n w_i(x) Y_i,$$

其中分母为零时仍需约定输出。若权重非负、总和为一，且仅使用距离 x 至多为 h 的输入，条件偏差仍至多为 Lh^β 。条件噪声方差则满足

$$\text{Var}(\hat{f}_h(x) | X_{1:n}) \leq \sigma^2 \sum_{i=1}^n w_i(x)^2.$$

所以 $1/\sum_i w_i(x)^2$ 可以看作加权平均的有效样本量。方盒核的有效样本量恰好是邻居数。

说明 5.5 (为什么高阶光滑性需要高阶算法). 本节的完整定理只假设 $0 < \beta \leq 1$ 。对通常意义下 $\beta > 1$ 的 Hölder 类，函数可以有非零一阶导数，不能继续假设 $|f(x) - f(z)| \leq L \|x - z\|^\beta$ ；这样的差值条件在连通区域上会迫使函数为常数。

利用高阶光滑性时，应在邻域内拟合局部多项式，并通过矩条件消除低阶 Taylor 项。边界附近邻域不对称，也需要相应处理。一个非负核平均不会仅因真函数更加光滑就自动获得 h^β 偏差。高阶方法可以延续相同的偏差与方差平衡，但这是需要另行证明的结论。

6 RKHS：从岭回归的对偶到函数空间几何

核平滑告诉我们：可以把新位置的预测写成观测标签的加权平均。下一步自然要问，权重除了依赖输入距离，还能否由一组更适合任务的特征来决定？当特征很多，甚至有无限多个时，是否仍然能计算、仍然能控制学习的复杂度？

原课程第一份 RKHS 笔记的前 3 页先从非线性特征讲到岭回归的对偶与预测，第 4 页再把数据点解释为取值泛函；第二份笔记第 2 至 3 页则从导数能量反向推导核的 Green 函数解释 [10, 11]。本节按这个动机顺序组织。我们先得到一个有用的算法，再问应当用什么函数空间解释和推广它。

本节先回答三个问题。怎样让线性模型表达非线性关系？怎样在不展开全部特征的情况下计算预测？怎样把“参数范数小”改写成只涉及函数本身的约束？RKHS 的定义将从这三个问题中自然出现。课堂上可以抓住 (6.3)、(6.4) 和表示定理的正交分解；完整的核空间构造与 Green 函数推导可留作课后细读。

6.1 先有非线性问题，再选择线性特征

假设二维分类问题的边界是一条以原点为圆心、半径为 ρ 的圆：圆内为正类，圆外为负类。直接用 $w^\top x + b$ 无法表达这条边界，但令

$$\phi(x) = (x_1, x_2, x_1^2 + x_2^2),$$

边界就变成特征空间中的一个平面 $\phi_3(x) = \rho^2$ 。这里“线性”指模型对参数线性，输入可以先经过非线性变换。回归同样如此：取 $\phi(x) = (1, x, x^2)$ ，模型 $w^\top \phi(x)$ 就是二次多项式。

这种方法保留了线性回归的计算结构，但出现了一个新问题：更丰富的特征意味着更长的参数向量。如果我们希望同时使用很多阶的多项式，甚至无限多个特征，逐个存储和更新参数就不方便了。

证明直觉 训练与预测真正需要的，可能只是特征之间的内积。先在有限维情形把这个事实推导出来，再让特征空间变成无限维，就能看清楚核方法解决的具体问题。

6.2 重访岭回归：为什么只需要特征之间的内积

设 $\phi(x) \in \mathbb{R}^p$ ，用 $\Phi \in \mathbb{R}^{n \times p}$ 表示特征矩阵，第 i 行为 $\phi(x_i)^\top$ 。考虑

$$\hat{w} = \operatorname{argmin}_{w \in \mathbb{R}^p} \left\{ \frac{1}{2n} \|\Phi w - y\|_2^2 + \frac{\lambda}{2} \|w\|_2^2 \right\}, \quad \lambda > 0. \quad (6.1)$$

两个 $1/2$ 只为简化求导，最小化解与本讲后面的平方损失约定相同。参数空间中的正规方程给出

$$\hat{w} = (\Phi^\top \Phi + n\lambda I_p)^{-1} \Phi^\top y.$$

这是一个 p 维线性系统。下面换到由 n 个训练样本索引的坐标。

命题 6.1 (岭回归的对偶表示与核预测)。令

$$k(x, z) = \langle \phi(x), \phi(z) \rangle, \quad K = \Phi \Phi^\top, \quad \mathbf{k}_X(x) = (k(x_1, x), \dots, k(x_n, x))^\top.$$

则

$$a = (K + n\lambda I_n)^{-1}y, \quad \hat{w} = \Phi^\top a, \quad (6.2)$$

并且对任意新输入 x ,

$$\hat{f}_\lambda(x) = \langle \hat{w}, \phi(x) \rangle = \mathbf{k}_X(x)^\top (K + n\lambda I_n)^{-1}y. \quad (6.3)$$

因而只要能计算 $k(x, z)$, 就能计算训练矩阵和全部预测值。

证明直觉 把拟合值 $z = \Phi w$ 单独作为变量。这样, 损失只看 z , 正则项只看 w , 两者通过约束相连。约束的每一个乘子对应一个训练样本; 对 w 求导后, 最优参数自动落在这些样本特征的线性张成中。对偶表示因此是从最优性条件推出来的。

证明. 将 (6.1) 写成带约束问题

$$\min_{w, z} \left\{ \frac{1}{2n} \|z - y\|_2^2 + \frac{\lambda}{2} \|w\|_2^2 \right\} \quad \text{使得} \quad z = \Phi w.$$

为了让后面的系数直接对应核展开, 采用缩放后的 Lagrange 乘子 λa :

$$\mathcal{L}(w, z, a) = \frac{1}{2n} \|z - y\|_2^2 + \frac{\lambda}{2} \|w\|_2^2 + \lambda a^\top (z - \Phi w).$$

对两个原变量求导, 得到

$$\nabla_w \mathcal{L} = \lambda w - \lambda \Phi^\top a = 0, \quad \nabla_z \mathcal{L} = \frac{z - y}{n} + \lambda a = 0.$$

因此 $w = \Phi^\top a$, $z = y - n\lambda a$ 。代回约束 $z = \Phi w$, 得到

$$(K + n\lambda I_n)a = y.$$

由于 $K \succeq 0$ 且 $\lambda > 0$, 这个系统有唯一解。得到的 (w, z) 可行且满足全部驻点条件; 目标严格凸, 故它给出唯一最小化解。最后

$$\hat{f}_\lambda(x) = \left\langle \sum_{i=1}^n a_i \phi(x_i), \phi(x) \right\rangle = \sum_{i=1}^n a_i k(x_i, x),$$

即为 (6.3)。消去 w, z 后, 对偶目标本身也可以写成

$$\max_{a \in \mathbb{R}^n} \left\{ \lambda a^\top y - \frac{\lambda}{2} a^\top K a - \frac{n\lambda^2}{2} \|a\|_2^2 \right\},$$

其一阶条件正是上面的线性系统。 □

驻点关系还说明 $a_i = (Y_i - \hat{f}_\lambda(x_i))/(n\lambda)$: 每个样本的系数就是经过缩放的拟合残差, 残差通过该样本的特征方向影响预测。

为什么这一步重要? 参数坐标中的矩阵是 $p \times p$, 样本坐标中的矩阵是 $n \times n$ 。例如对 $\phi(x) = (1, x, x^2)$, 核为 $k(x, z) = 1 + xz + x^2z^2$ 。只要内积能够直接计算, 预测公式里就不需要显式恢复全部参数。后面的 Gaussian 核将展示同一件事如何容纳无限多个特征。当然, n 很大时, 存储和求解核矩阵本身仍有计算成本。

与上一节核平滑的联系。令 $\omega_\lambda(x) = (K + n\lambda I_n)^{-1} \mathbf{k}_X(x)$, 则

$$\hat{f}_\lambda(x) = \sum_{i=1}^n \omega_{\lambda,i}(x) Y_i.$$

在核、输入和 λ 固定时, 估计仍然对标签 y 线性; 它对 x 可以高度非线性。与局部平均相比, 这组权重由全部训练点之间的特征几何共同决定, 一般可以有负值, 也不必和为一。

6.3 从特征空间到函数空间: 再生性质从哪里来

现在允许 $\phi(x)$ 取值于一个可能无限维的实 Hilbert 空间 $\mathcal{G}$, 仍写

$$f_w(x) = \langle w, \phi(x) \rangle_{\mathcal{G}}, \quad k(x, z) = \langle \phi(x), \phi(z) \rangle_{\mathcal{G}}.$$

参数范数 $\|w\|_{\mathcal{G}}$ 很自然, 但我们希望给出函数本身的复杂度。这里须先处理一个细节: 不同参数可能表示同一个函数。

令

$$\mathcal{G}_0 = \overline{\text{span}\{\phi(x) : x \in \mathcal{X}\}}.$$

若 $w = w_0 + w_\perp$ 是相对于 $\mathcal{G}_0$ 的正交分解, 则 $f_w = f_{w_0}$, 且 $\|w_0\| \leq \|w\|$ 。所以我们总可以只保留 $\mathcal{G}_0$ 中的参数。在这个子空间上, $w \mapsto f_w$ 是单射: 如果 $f_w(x) = 0$ 对所有 x 成立, 则 w 与一个稠密生成集正交, 只能是零。

于是可以定义

$$\mathcal{H} = \{f_w : w \in \mathcal{G}_0\}, \quad \langle f_w, f_v \rangle_{\mathcal{H}} = \langle w, v \rangle_{\mathcal{G}}.$$

这个定义把 $\mathcal{G}_0$ 的内积与完备性等距搬到函数空间上。它也解释了函数范数的含义:

$$\|f\|_{\mathcal{H}} = \min\{\|w\|_{\mathcal{G}} : f_w = f\}.$$

接下来观察: $\phi(x)$ 本身也是 $\mathcal{G}_0$ 中的一个参数, 它表示的函数正是

$$f_{\phi(x)}(z) = \langle \phi(x), \phi(z) \rangle_{\mathcal{G}} = k(x, z).$$

将这个函数记为 $k_x = k(x, \cdot)$ 。由搬过来的内积, 立即得到

$$\langle f_w, k_x \rangle_{\mathcal{H}} = \langle w, \phi(x) \rangle_{\mathcal{G}} = f_w(x). \quad (6.4)$$

证明直觉 再生性质就是“用特征做预测”在函数坐标中的写法。原先输入 x 对应参数空间中的向量 $\phi(x)$; 现在它对应函数空间中的向量 k_x 。与这个向量取内积, 就读取函数在 x 的值。训练数据因此指定了函数空间中的一组可观察方向。

6.4 为什么要要求点值连续, 核又能否任意选择

学习算法要在训练点读取 $f(x_i)$, 而复杂度约束控制的是函数范数。两者必须相容: 范数很小的扰动, 在固定输入处也应当很小。Hilbert 结构本身并不保证这一点。

例如取内部点 $x \in (0, 1)$, 令

$$g_h(t) = \max\{1 - |t - x|/h, 0\}, \quad 0 < h < \min\{x, 1 - x\}.$$

这是越来越窄的连续尖峰，满足

$$g_h(x) = 1, \quad \|g_h\|_{L^2[0,1]}^2 = \frac{2h}{3} \rightarrow 0.$$

所以即使先只考虑连续函数， L^2 范数也不能连续控制取值；在完整 L^2 等价类上，单点取值更没有唯一含义。选择 RKHS，就是让我们用来约束复杂度的范数能够控制学习实际读取的量。

另一方面，如果 $k(x, z)$ 真的是特征内积，那么任意 $c \in \mathbb{R}^m$ 都必须满足

$$\sum_{i,j=1}^m c_i c_j k(x_i, x_j) = \left\| \sum_{i=1}^m c_i \phi(x_i) \right\|_{\mathcal{G}}^2 \geq 0.$$

核的半正定性由内积几何强制给出。我们现在可以把这两个观察写成定义。

6.5 为什么一个核能够定义一个函数空间

定义 6.2 (正定核与再生核 Hilbert 空间)。实值对称函数 $k: \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ 称为正定核，如果对任意有限点集 $x_1, \dots, x_m$ 和 $a \in \mathbb{R}^m$,

$$\sum_{i,j=1}^m a_i a_j k(x_i, x_j) \geq 0.$$

这里允许等号成立，即 Gram 矩阵只需半正定。一个由 $\mathcal{X}$ 上实值函数构成的 Hilbert 空间 $\mathcal{H}$ 称为再生核 Hilbert 空间，简称 RKHS，如果每个取值映射 $f \mapsto f(x)$ 都是连续线性泛函。其表示元记为 k_x ，满足

$$f(x) = \langle f, k_x \rangle_{\mathcal{H}}, \quad k(x, z) = \langle k_x, k_z \rangle_{\mathcal{H}}.$$

对抽象定义中的连续取值泛函，Riesz 表示定理保证唯一的 k_x 存在；上面的特征模型则把它直接构造了出来。更进一步，即使最初只给出一个半正定核，也能恢复整套函数空间几何。

这里的“连续”指 $f \mapsto f(x)$ 相对于函数空间范数连续。若还希望函数值随输入 x 连续变化，则要再结合核关于输入的正则性；这是两个不同层面的性质。

命题 C.1 证明，半正定性也足以保证对应的 RKHS 存在。它的自然特征就是核截面 $x \mapsto k_x$ ，并满足

$$\|k_x\|_{\mathcal{H}}^2 = k(x, x), \quad |f(x)| \leq \|f\|_{\mathcal{H}} \sqrt{k(x, x)}.$$

若 $\sup_x k(x, x) \leq \kappa^2$ ，便有 $\|f\|_{\infty} \leq \kappa \|f\|_{\mathcal{H}}$ 。从有限核组合、消去零范数表示到完备化的严格构造放在附录 C；下面继续看这套几何怎样指导模型设计。

6.6 三个可以直接算清楚的例子

例 6.3 (线性核)。在 $\mathcal{X} = \mathbb{R}^p$ 上取 $k(x, z) = x^\top z$ 。其 RKHS 是所有线性函数 $f_w(x) = w^\top x$ ，内积为 $\langle f_w, f_v \rangle_{\mathcal{H}} = w^\top v$ 。因此核方法包含普通的欧氏几何，并没有改变岭回归的基本原理。

例 6.4 (Gaussian 核与无限维特征)。给定 $a > 0$ ，令

$$k(x, z) = \exp\left(-\frac{\|x - z\|_2^2}{2a^2}\right).$$

对多重指标 $\alpha = (\alpha_1, \dots, \alpha_d)$ ，定义

$$\phi_\alpha(x) = e^{-\|x\|_2^2/(2a^2)} \frac{x^\alpha}{a^{|\alpha|} \sqrt{\alpha!}}.$$

指数函数的幂级数给出

$$k(x, z) = \sum_{\alpha \in \{0,1,2,\dots\}^d} \phi_\alpha(x) \phi_\alpha(z), \quad \sum_{\alpha} \phi_\alpha(x)^2 = 1.$$

所以这确实是 ℓ^2 特征的内积，正定性随之成立。算法可以直接计算核值，从而使用这些无限多个特征，而不必逐个存储它们。

例 6.5 (从导数能量反向推导 Brownian 核)。 现在从一个统计建模要求出发：希望拟合曲线不要无代价地剧烈振荡，因此用导数能量作为复杂度约束。考虑

$$\mathcal{H} = \{f : [0, 1] \rightarrow \mathbb{R} : f \text{ 绝对连续}, f(0) = 0, f' \in L^2[0, 1]\}, \quad \langle f, g \rangle_{\mathcal{H}} = \int_0^1 f'(t) g'(t) dt.$$

条件 $f(0) = 0$ 使导数能量成为范数：常数函数不会留下非零的零范数元素。映射 $v \mapsto [x \mapsto \int_0^x v(t) dt]$ 将 $L^2[0, 1]$ 等距映到 $\mathcal{H}$ ，所以空间完备。

先问核应该满足什么，再把它求出来。对固定的 x ，我们希望找到 $k_x \in \mathcal{H}$ ，使

$$\int_0^1 f'(t) k'_x(t) dt = f(x) = \int_0^x f'(t) dt \quad \text{对所有 } f \in \mathcal{H} \text{ 成立.}$$

f' 可以是任意 L^2 函数，因此唯一可能的导数是 $k'_x(t) = \mathbf{1}\{t < x\}$ ，几乎处处成立。再用 $k_x(0) = 0$ 积分，便得到

$$k_x(t) = \int_0^t \mathbf{1}\{u < x\} du = \min\{x, t\}, \quad k(x, z) = \min\{x, z\}. \quad (6.5)$$

所以这个核是从我们想使用的正则项推导出来的。代回上式，同时完成了再生性质的证明。进一步，

$$|f(x) - f(z)| \leq \left(\int_{x \wedge z}^{x \vee z} |f'(t)|^2 dt \right)^{1/2} \sqrt{|x - z|} \leq \|f\|_{\mathcal{H}} \sqrt{|x - z|}.$$

范数约束在这里直接控制了函数的振荡。这个例子把函数分析中的导数正则性与统计中的复杂度约束联系起来。

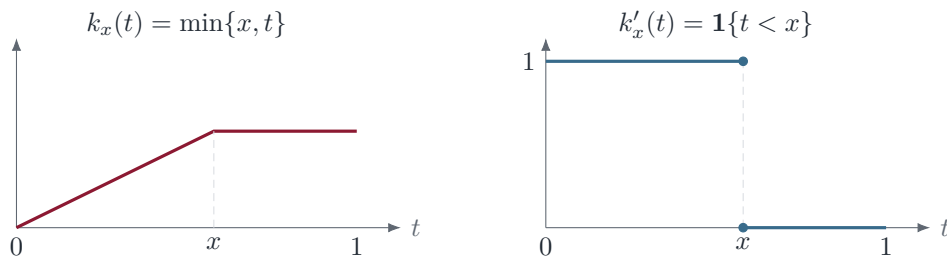


图 4: 导数能量下的取值方向。与 k'_x 取 L^2 内积，只累计 $[0, x]$ 上的导数，恰好还原 $f(x)$ 。折点处导数的具体赋值不影响积分。

说明 6.6 (为什么原课把核称为 Green 函数)。 再生关系在这里就是弱方程

$$\int_0^1 k'_x(t) v'(t) dt = v(x), \quad v \in \mathcal{H}.$$

当 $0 < x < 1$ 时，它对应于

$$-k''_x = \delta_x, \quad k_x(0) = 0, \quad k'_x(1) = 0.$$

第二个导数在分布意义下理解：\$k'_x\$ 在 \$x\$ 处从 1 跳到 0，于是 \$-k''_x\$ 是单位点源。\$\delta_x\$ 表示取值泛函，不能当成一个 \$L^2\$ 函数。右端的自然边界条件则使分部积分的边界项消失。

这给出另一条选择核的路线：先规定正则化能量，再寻找能量对应算子对一个点源的响应。以 \$A \succ 0\$ 时的离散二次能量 \$f^\top A f\$ 类比，读取第 \$i\$ 个分量的表示元满足 \$A k_i = e_i\$；连续情形中，\$A^{-1} e_i\$ 就变成相应的 Green 函数。

边界条件属于模型的一部分。上述弱方程对 \$x \in [0, 1]\$ 都成立，而写出的齐次右端条件只针对内部点。当 \$x = 1\$ 时 \$k_1(t) = t\$，取值来自右端通量。若把函数空间改为两端都取零，则核相应变为 \$\min\{x, z\} - xz\$。

说明 6.7 (平滑性为什么要与维数一起考虑)。 原课从 Green 函数的奇性引出高维中的平滑性要求。一个可以直接验证的版本是在周期区域 \$\mathbb{T}^d\$ 上，给三角多项式定义 Sobolev 范数

$$\|f\|_{H^s}^2 = \sum_{m \in \mathbb{Z}^d} (1 + 4\pi^2 \|m\|_2^2)^s |\hat{f}(m)|^2.$$

对 Fourier 展开应用 Cauchy 与 Schwarz 不等式，得到

$$|f(x)| \leq \|f\|_{H^s} \left[\sum_{m \in \mathbb{Z}^d} (1 + 4\pi^2 \|m\|_2^2)^{-s} \right]^{1/2}.$$

频率壳层 \$2^j \leq \|m\|_2 < 2^{j+1}\$ 中有数量级为 \$2^{jd}\$ 的格点，每项权重的数量级为 \$2^{-2js}\$，所以这一层的贡献为 \$2^{j(d-2s)}\$。方括号中的级数恰在 \$s > d/2\$ 时收敛，此时完备化有连续的点值泛函。反过来，固定 \$x\$，令 \$a_m = (1 + 4\pi^2 \|m\|_2^2)^{-s}\$，并取 \$f_N(t) = \sum_{\|m\|_2 \leq N} a_m e^{2\pi i m \cdot (t-x)}\$。对称的频率配对使 \$f_N\$ 实值，且

$$\frac{|f_N(x)|}{\|f_N\|_{H^s}} = \left[\sum_{\|m\|_2 \leq N} a_m \right]^{1/2} \rightarrow \infty \quad \text{当 } s \leq d/2.$$

这说明临界情形一般也不能控制点值。维数提高后，同一频率尺度内存在更多方向，需要更强的高频惩罚才能把它们的贡献加起来。这里讨论的是这族 Sobolev 空间的阈值，并非所有 RKHS 都必须由经典导数阶数来刻画。

6.7 表示定理：训练数据只看得到有限个方向

定理 6.8 (表示定理)。 固定训练输入 \$x_1, \dots, x_n\$。考虑

$$J(f) = \Psi(f(x_1), \dots, f(x_n)) + \Omega(\|f\|_{\mathcal{H}}), \quad f \in \mathcal{H},$$

其中 \$\Omega\$ 非递减。记 \$S = \text{span}\{k_{x_1}, \dots, k_{x_n}\}\$。对每个 \$f \in \mathcal{H}\$，其正交投影 \$f_S\$ 都满足 \$J(f_S) \leq J(f)\$。因此，若最小值存在，就有一个最小化解形如

$$\hat{f}(\cdot) = \sum_{i=1}^n \alpha_i k(x_i, \cdot).$$

若 \$\Omega\$ 严格递增，则每个最小化解都在 \$S\$ 中。

证明直觉 训练数据只能通过 $f(x_i)$ 观察函数。与所有 k_{x_i} 正交的方向对数据完全不可见，却仍消耗函数范数。最优解自然舍去这些方向。无限维问题之所以化为有限维问题，原因就在这个可见与不可见的正交分解。

证明. 有限维子空间 S 是闭的，将任意函数正交分解成 $f = f_S + f_\perp$ 。因为 $k_{x_i} \in S$ ，再生性质给出

$$f_\perp(x_i) = \langle f_\perp, k_{x_i} \rangle_{\mathcal{H}} = 0.$$

于是 f 与 f_S 在全部训练点取值相同，损失 Ψ 完全相同。另一方面，勾股定理给出

$$\|f\|_{\mathcal{H}}^2 = \|f_S\|_{\mathcal{H}}^2 + \|f_\perp\|_{\mathcal{H}}^2.$$

删除 $f_\perp$ 不会增大正则项；若 Ω 严格递增且 $f_\perp \neq 0$ ，正则项还会严格减小。 $\square$

6.8 核岭回归：一个矩阵求解与一组谱收缩

核岭回归取平方损失和二次正则项：

$$\hat{f}_\lambda = \operatorname{argmin}_{f \in \mathcal{H}} \left\{ \frac{1}{n} \sum_{i=1}^n (Y_i - f(x_i))^2 + \lambda \|f\|_{\mathcal{H}}^2 \right\}, \quad \lambda > 0. \quad (6.6)$$

令 $K_{ij} = k(x_i, x_j)$, $y = (Y_1, \dots, Y_n)^\top$ 。表示定理将目标变为

$$\frac{1}{n} \|y - K\alpha\|_2^2 + \lambda \alpha^\top K \alpha.$$

取

$$\alpha_\lambda = (K + n\lambda I)^{-1} y, \quad \hat{f}_\lambda(x) = \sum_{i=1}^n \alpha_{\lambda,i} k(x_i, x). \quad (6.7)$$

这个系数满足 $(K + n\lambda I)\alpha_\lambda = y$ ，代入即验证目标的一阶条件。原函数目标由 $\lambda \|f\|_{\mathcal{H}}^2$ 保证严格凸，因此所得函数是唯一最小化解。即使 K 奇异， $K + n\lambda I$ 也正定可逆；此时同一个函数的系数表示可以不唯一，式 (6.7) 给出其中一个确定选择。

记 K/n 的特征分解为

$$\frac{K}{n} = U \operatorname{diag}(\hat{\mu}_1, \dots, \hat{\mu}_n) U^\top, \quad \hat{\mu}_j \geq 0.$$

训练点上的拟合向量为

$$\hat{m} = S_\lambda y, \quad S_\lambda = K(K + n\lambda I)^{-1} = U \operatorname{diag} \left(\frac{\hat{\mu}_j}{\hat{\mu}_j + \lambda} \right) U^\top. \quad (6.8)$$

因此 λ 规定了保留数据方向的尺度：当 $\hat{\mu}_j \gg \lambda$ 时，方向 j 几乎完整保留；当 $\hat{\mu}_j \ll \lambda$ 时，该方向被强烈压缩。核值编码的几何信息，就这样进入了学习算法。

6.9 固定设计下的精确偏差方差分解

下面条件于训练输入，把它们记为确定点 $x_1, \dots, x_n$ 。令 $m^* = (f^*(x_1), \dots, f^*(x_n))^T$ ，并写 $y = m^* + \xi$ 。假设 $\mathbb{E}\xi = 0$ 、 $\text{Cov}(\xi) = \Sigma \preceq \sigma^2 I$ 。例如，条件独立且条件方差至多为 σ^2 的噪声满足此条件。

定理 6.9 (核岭回归的经验预测误差)。对任意真均值向量 m^* ，均有精确分解

$$\frac{1}{n} \mathbb{E}_\xi \|\hat{m} - m^*\|_2^2 = \frac{1}{n} \|(I - S_\lambda)m^*\|_2^2 + \frac{1}{n} \text{tr}(S_\lambda \Sigma S_\lambda). \quad (6.9)$$

若进一步 $f^* \in \mathcal{H}$ 且 $\|f^*\|_{\mathcal{H}} \leq R$ ，则

$$\mathbb{E}_\xi \|\hat{f}_\lambda - f^*\|_n^2 \leq \frac{\lambda R^2}{4} + \frac{\sigma^2}{n} \sum_{j=1}^n \left(\frac{\hat{\mu}_j}{\hat{\mu}_j + \lambda} \right)^2 \quad (6.10)$$

$$\leq \frac{\lambda R^2}{4} + \frac{\sigma^2}{n} \hat{D}(\lambda), \quad \hat{D}(\lambda) = \sum_{j=1}^n \frac{\hat{\mu}_j}{\hat{\mu}_j + \lambda}. \quad (6.11)$$

这里 $\|g\|_n^2 = n^{-1} \sum_i g(x_i)^2$ 。当 $\Sigma = \sigma^2 I$ 时，(6.10) 中的方差项为精确等式。

证明直觉 一个小特征值方向会被正则化强烈压缩，但 RKHS 球也只允许真函数在这个方向上携带少量信号，见椭圆约束 (6.12)。偏差因此可控。另一方面，大特征值方向保留得多，也就保留了更多噪声。有效维数 $\hat{D}(\lambda)$ 将每个方向按保留程度计数，取代了粗略的参数个数。

带宽 h 控制空间中的分辨率，正则化 λ 控制核谱中的分辨率。两者都通过压缩难以可靠估计的细节来降低方差，并通过平滑性或范数约束保证所损失的信号不大。

证明. 第一步：线性平滑器直接给出分解。由 (6.8)，

$$\hat{m} - m^* = (S_\lambda - I)m^* + S_\lambda \xi.$$

平方后对噪声取期望，交叉项因 $\mathbb{E}\xi = 0$ 消失，随机项的平方范数期望等于 $\text{tr}(S_\lambda \Sigma S_\lambda)$ 。这证明了 (6.9)。由协方差上界，

$$\text{tr}(S_\lambda \Sigma S_\lambda) \leq \sigma^2 \text{tr}(S_\lambda^2) = \sigma^2 \sum_j \left(\frac{\hat{\mu}_j}{\hat{\mu}_j + \lambda} \right)^2.$$

第二步：函数范数限制每个谱方向的信号能量。令 $\tilde{m}_j = (U^\top m^*)_j$ 。对每个 $\hat{\mu}_j > 0$ ，定义

$$e_j = \frac{1}{\sqrt{n\hat{\mu}_j}} \sum_{i=1}^n U_{ij} k_{x_i} \in \mathcal{H}.$$

由 $U^\top K U = n \text{diag}(\hat{\mu}_j)$ ，这些 e_j 两两正交且范数为一，同时

$$\langle f^*, e_j \rangle_{\mathcal{H}} = \frac{\tilde{m}_j}{\sqrt{n\hat{\mu}_j}}.$$

投影的勾股定理于是给出

$$\sum_{j: \hat{\mu}_j > 0} \frac{\tilde{m}_j^2}{n\hat{\mu}_j} \leq \|f^*\|_{\mathcal{H}}^2 \leq R^2. \quad (6.12)$$

若 $\hat{\mu}_j = 0$ ，对应的核截面组合范数为零，再生性质说明 $\tilde{m}_j = 0$ ，所以无需对零特征值作除法。

第三步：只需一个一元不等式。在谱坐标中，平方偏差为

$$\begin{aligned} \frac{1}{n} \|(I - S_\lambda)m^*\|_2^2 &= \sum_{j: \hat{\mu}_j > 0} \frac{\lambda^2 \hat{\mu}_j}{(\hat{\mu}_j + \lambda)^2} \frac{\tilde{m}_j^2}{n \hat{\mu}_j} \\ &\leq \frac{\lambda}{4} \sum_{j: \hat{\mu}_j > 0} \frac{\tilde{m}_j^2}{n \hat{\mu}_j} \leq \frac{\lambda R^2}{4}. \end{aligned}$$

这里用了 $\lambda^2 \mu / (\mu + \lambda)^2 \leq \lambda/4$ ，等价于 $(\mu - \lambda)^2 \geq 0$ 。合并偏差与方差，再用 $0 \leq s \leq 1 \Rightarrow s^2 \leq s$ ，即得 (6.11)。□

说明 6.10 (经验预测与新样本预测)。定理 6.9 控制的是固定训练输入上的真回归函数误差，不是对带噪声标签的训练损失，也不是独立新输入上的总体误差。随机设计下， K/n 和 $\hat{D}(\lambda)$ 都依赖于训练输入。把这个界转移为 $L^2(P_X)$ 风险，需要额外比较经验几何与总体几何。不能直接把经验特征值换成总体积分算子的特征值，而不证明这一步。

6.10 从全局复杂度到局部谱信息

最后把 RKHS 与前面的泛化工具接起来。对半径为 R 的球 $\mathcal{F}_R = \{f \in \mathcal{H} : \|f\|_{\mathcal{H}} \leq R\}$ ，固定样本上的 Rademacher 复杂度满足

$$\begin{aligned} \hat{\mathfrak{R}}_S(\mathcal{F}_R) &= \frac{R}{n} \mathbb{E}_\varepsilon \left\| \sum_{i=1}^n \varepsilon_i k_{x_i} \right\|_{\mathcal{H}} \\ &\leq \frac{R}{n} \sqrt{\mathbb{E}_\varepsilon \left\| \sum_{i=1}^n \varepsilon_i k_{x_i} \right\|_{\mathcal{H}}^2} = \frac{R}{n} \sqrt{\sum_{i=1}^n k(x_i, x_i)}. \end{aligned}$$

第一步是 Hilbert 空间中的对偶范数；第二步是 Cauchy 与 Schwarz 不等式；最后利用独立随机符号消去全部交叉项。如果 $k(x, x) \leq \kappa^2$ ，就得到

$$\hat{\mathfrak{R}}_S(\mathcal{F}_R) \leq \frac{R\kappa}{\sqrt{n}}.$$

无限维并不自动意味着复杂度无穷大：这里起作用的是范数预算与核对角线。不过， R 和核的选择本身可能依赖维数，因此不能据此宣布所有高维问题都没有维数灾难。

这个全局界只使用 $\sum_i k(x_i, x_i)$ ，没有区分各个谱方向。定理 6.9 中的有效维数，以及第 4 节的局部复杂度，保留了更多谱信息。局部分析再加上一个条件：函数差已经接近真函数。RKHS 范数球与这个小的预测误差球相交，便形成截断椭球；它的宽度由 $\sum_j \min\{r^2, R^2 \hat{\mu}_j\}$ 控制，具体证明见命题 7.1。这说明泛化工具、非参平滑和核方法可以围绕同一个问题组织：在当前误差尺度下，还有多少可以被噪声扰动的方向？

例 6.11 (Brownian 核给出分段线性估计)。将例 6.5 代入表示定理，核岭回归解为

$$\hat{f}_\lambda(x) = \sum_{i=1}^n \alpha_{\lambda, i} \min\{x_i, x\}.$$

它在相邻训练输入之间是线性的，在训练输入处可以改变斜率； $\hat{f}_\lambda(0) = 0$ ，在最大训练输入右侧为常数。正则项则是 $\lambda \int_0^1 |f'(t)|^2 dt$ 。所以一个无限维的导数能量最小化问题，最终得到由数据位置决定节点的分段线性函数。这是表示定理可以直接画出来的实例。

7 从核的谱衰减回到局部复杂度 (课后)

RKHS 的定义与核方法见第 6 节。这里仅使用再生性质 $g(x) = \langle g, k(x, \cdot) \rangle_{\mathcal{H}}$ 。令核矩阵为 $K = (k(x_i, x_j))_{i,j=1}^n$ ，并按降序排列 K/n 的特征值：

$$\hat{\mu}_1 \geq \cdots \geq \hat{\mu}_n \geq 0.$$

对 RKHS 球中的函数值向量定义

$$\mathcal{V}_R = \{(g(x_i))_{i=1}^n : \|g\|_{\mathcal{H}} \leq R\}.$$

命题 7.1 (局部椭圆复杂度)。 对任意 $r, R > 0$,

$$\mathbb{E}_Z \sup_{v \in \mathcal{V}_R : \|v\|_n \leq r} |\langle Z, v \rangle_n| \leq \left\{ \frac{2}{n} \sum_{j=1}^n \min(r^2, R^2 \hat{\mu}_j) \right\}^{1/2}. \quad (7.1)$$

把 Z_i 换成独立 Rademacher 符号，右边同样成立。

证明直觉 沿第 j 个经验特征方向，可动的幅度有两个限制。局部经验范数给出幅度 r ，RKHS 范数给出幅度 $R\sqrt{\hat{\mu}_j}$ 。因此真正起作用的是两者中较小的一个。谱很小的方向可以存在很多个，但每个方向对局部噪声收益的贡献都很小。

证明. 令 $K/n = Q \text{diag}(\hat{\mu}_j) Q^T$ ，其中 Q 正交。对 $v = (g(x_i))_{i=1}^n$ ，设 $\theta = Q^T v / \sqrt{n}$ 。将 g 正交投影到 $\text{span}\{k(x_i, \cdot) : 1 \leq i \leq n\}$ ，函数值不变而 RKHS 范数不增加。在这个有限维子空间中展开即可得到

$$\|v\|_n^2 = \sum_j \theta_j^2, \quad \sum_{j: \hat{\mu}_j > 0} \frac{\theta_j^2}{\hat{\mu}_j} \leq \|g\|_{\mathcal{H}}^2,$$

且 $\hat{\mu}_j = 0$ 时必有 $\theta_j = 0$ 。等价地，第二项是最小范数插值函数的平方范数 $v^T K^\dagger v$ 。

只对正特征值坐标计算，令 $a_j^2 = \min(r^2, R^2 \hat{\mu}_j)$ 。两个约束一起蕴含

$$\begin{aligned} \sum_j \frac{\theta_j^2}{a_j^2} &= \sum_j \max \left\{ \frac{\theta_j^2}{r^2}, \frac{\theta_j^2}{R^2 \hat{\mu}_j} \right\} \\ &\leq \sum_j \frac{\theta_j^2}{r^2} + \sum_j \frac{\theta_j^2}{R^2 \hat{\mu}_j} \leq 2. \end{aligned}$$

写 $\eta = Q^T Z$ ，于是

$$|\langle Z, v \rangle_n| = \frac{1}{\sqrt{n}} \left| \sum_j \eta_j \theta_j \right| \leq \sqrt{\frac{2}{n}} \left(\sum_j a_j^2 \eta_j^2 \right)^{1/2}.$$

先对可行 v 取上确界，再由 Jensen 不等式取期望，利用 $\mathbb{E} \eta_j^2 = 1$ 即得结论。Rademacher 情形下 $Q^T Z$ 的坐标未必独立，但仍有 $\mathbb{E} \eta_j^2 = 1$ ，同一个证明完全适用。□

定义尺度 r 下的经验有效维数

$$d_{\text{eff}}(r) = \sum_{j=1}^n \min \left\{ 1, \frac{R^2 \hat{\mu}_j}{r^2} \right\}.$$

上面的结果可以写成 $w_n(r) \leq r\sqrt{2d_{\text{eff}}(r)/n}$ 。于是临界半径条件由下式保证：

$$r^2 \geq \frac{128\sigma^2}{n} d_{\text{eff}}(r). \quad (7.2)$$

这与线性子空间的 $\sigma^2 p/n$ 是同一种平衡，区别在于现在有效的方向数会随估计尺度改变。

推论 7.2 (经验谱衰减给出的非参数快速率)。在固定设计高斯回归中，令 $\mathcal{F} = \{f \in \mathcal{H} : \|f\|_{\mathcal{H}} \leq A\}$ ，其中 $A > 0$ ，并假设 $f^* \in \mathcal{F}$ 。设对所有 $1 \leq j \leq n$ 有

$$\hat{\mu}_j \leq C_0 j^{-q}, \quad C_0 > 0, \quad q > 1.$$

记 $R = 2A$ 、 $D = R\sqrt{C_0}$ ，且 $c_q = \sqrt{2q/(q-1)}$ 。则以至少 $1 - \delta$ 的概率，约束最小二乘解满足

$$\|\hat{f} - f^*\|_n^2 \leq (8c_q)^{\frac{2q}{q+1}} \left(\frac{\sigma^2}{n}\right)^{\frac{q}{q+1}} D^{\frac{2}{q+1}} + \frac{128\sigma^2 \log(1/\delta)}{n}. \quad (7.3)$$

证明. 有限样本取值像是有限维闭椭球，因此满足定理所需的闭性。原函数类是凸集，移位后的误差类关于原点星形，并且 $\|f - f^*\|_{\mathcal{H}} \leq 2A = R$ 。因此可用 $\mathcal{V}_R$ 的局部复杂度控制误差类。利用积分表示与计数界，有

$$\begin{aligned} \sum_{j=1}^n \min(r^2, R^2 \hat{\mu}_j) &\leq \sum_{j \geq 1} \min(r^2, D^2 j^{-q}) \\ &= \int_0^{r^2} \#\{j \geq 1 : D^2 j^{-q} \geq t\} dt \\ &\leq \int_0^{r^2} (D^2/t)^{1/q} dt = \frac{q}{q-1} D^{2/q} r^{2-2/q}. \end{aligned}$$

所以 $w_n(r) \leq c_q n^{-1/2} D^{1/q} r^{1-1/q}$ 。取

$$r_* = \left(\frac{8c_q \sigma D^{1/q}}{\sqrt{n}} \right)^{q/(q+1)},$$

便满足 $\sigma w_n(r_*) \leq r_*^2/8$ 。应用定理 4.8 即可。□

当 $q = 2s/d > 1$ ，即 $s > d/2$ 时，上式的主要样本量因子为

$$\boxed{n^{-\frac{2s}{2s+d}}.}$$

这里严格使用的是经验核矩阵的谱衰减假设。若已知的只是总体核积分算子的谱衰减，还需要额外分析经验谱；不能在公式中直接替换。条件 $q > 1$ 保证上述无穷级数与积分估计成立。对 $q \leq 1$ ，有限样本的式 (7.1) 仍然成立，但这个推论的无穷级数证明不适用。

说明 7.3 (经验预测误差与总体预测误差)。定理 4.8 与推论 7.2 控制的是设计点上的 $n^{-1} \sum_i (\hat{f}(x_i) - f^*(x_i))^2$ 。它既不是带噪声的训练损失，也不是新样本上的总体风险。若 X_i 随机，且条件于设计点之后噪声仍为独立高斯，定理可条件使用；但要进一步控制 $\mathbb{E}_X[(\hat{f}(X) - f^*(X))^2]$ ，还需研究局部类上的经验二次型与总体二次型之间的关系。一般的随机设计局部 Rademacher 理论正是在处理这些额外步骤[14]。

与核平滑的联系。回看推论 5.4, 带宽 h 的局部平均给出 $h^{2\beta} + 1/(nh^d)$ 这两个量级。令 $m \asymp h^{-d}$, 它们就变成 $m^{-2\beta/d} + m/n$: 保留更多空间分辨率, 逼近误差变小, 统计波动变大。这里的局部谱分析从预测半径 r 出发, 通过 $d_{\text{eff}}(r)$ 计算在这个尺度上仍然可见的方向数。两种计算都把光滑性转化为有效维数, 再让统计成本与所需精度平衡。不过, 核平滑的结论使用随机设计下的总体均方误差, 本节结论使用固定设计下的经验预测误差; 平滑度条件也分别是该推论的 $0 < \beta \leq 1$ 与本节的 $s > d/2$ 。这些条件应连同速率一起阅读。

可复用的证明结构。先从估计器的定义提取一个保留曲率的基本不等式, 再确定误差函数真正位于哪个局部集合。计算这个集合上的线性噪声复杂度后, 用一个确定的半径建立集中事件, 最后通过缩放或分层排除更远的候选函数。有限类证明用方差随风险减小来完成同一件事, RKHS 证明则让有效维数随尺度改变。

8 把整讲串起来：分辨率、噪声与几何

8.1 核平滑、核岭回归与局部复杂度的共同结构

核平滑选择一个空间分辨率 h 。邻域越小，平滑偏差越小，但能平均的观测数也越少。核岭回归选择一个谱分辨率 λ 。保留更多特征方向可以减少偏差，也会让更多噪声进入预测。局部复杂度则从目标误差 r 出发，反过来问：在这个误差尺度上，还有多少方向可被噪声利用？

命题 8.1 (两种有效维数的直接联系)。给定半径 $R > 0$ 、误差尺度 $r > 0$ ，令 $\lambda = r^2/R^2$ 。对同一组非负经验特征值，定义

$$d_{\text{eff}}(r) = \sum_j \min\{1, \hat{\mu}_j/\lambda\}, \quad \hat{D}(\lambda) = \sum_j \frac{\hat{\mu}_j}{\hat{\mu}_j + \lambda}.$$

则

$$\frac{1}{2}d_{\text{eff}}(r) \leq \hat{D}(\lambda) \leq d_{\text{eff}}(r). \quad (8.1)$$

证明直觉 一个定义把每个方向按 1 截断计数，另一个用平滑的分数计数。只要分辨率对应，它们看见的是同一批方向，区别只有固定常数。这解释了为什么正则化的偏差方差平衡与约束估计的临界半径常常给出同样的速率。

证明. 对 $u \geq 0$ ，分别考虑 $u \leq 1$ 与 $u \geq 1$ ，可得

$$\frac{1}{2} \min\{1, u\} \leq \frac{u}{1+u} \leq \min\{1, u\}.$$

令 $u = \hat{\mu}_j/\lambda$ 并求和即可。 $\square$

例如在对应的 RKHS 球内，核岭回归的上界具有 $\lambda R^2 + \sigma^2 \hat{D}(\lambda)/n$ 的结构。代入 $\lambda = r^2/R^2$ ，两项同阶正是

$$r^2 \asymp \frac{\sigma^2 d_{\text{eff}}(r)}{n}.$$

这与局部复杂度的固定点是一致的尺度关系。这里比较的是同一设计与同经验谱下的几何；两种估计器的定理仍各自保留前文给出的假设。

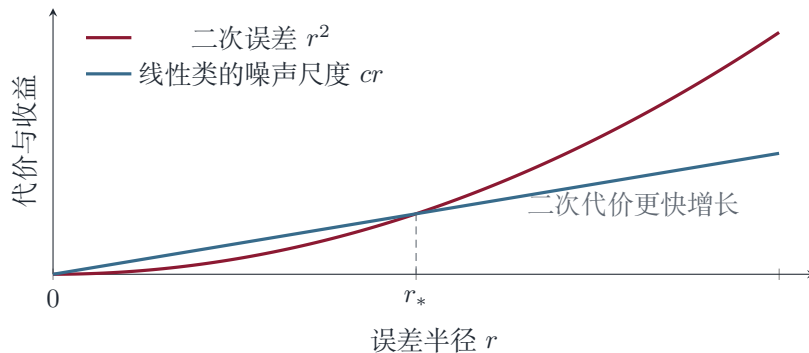


图 5: 局部化的示意图。图中以线性类的 $w_n(r) \propto r$ 说明缩放机制，常数经归一化。一般函数类的曲线由其局部几何决定；严格定理使用带常数余量的临界半径条件。

8.2 同一个误差标准下比较结论

下表省略固定常数及相应置信度项，所有“平方误差”均相对真实回归函数，而不是相对带噪声训练标签。

模型与方法	控制的量	本文证明的主要量级
有限类、一般有界损失	类内总体超额风险	$\sqrt{\log M/n}$ ；统一偏差分析。
有限类、有界平方回归	总体平方风险	$\log M/n$ ；要求 $f^* \in \mathcal{F}$ 与相应方差条件。
p 维子空间最小二乘	设计点上的平方误差	$\sigma^2 p/n$ ；固定设计、高斯噪声。
Hölder 核平滑	期望总体平方风险	$n^{-2\beta/(2\beta+d)}$ ； $0 < \beta \leq 1$ 。
核岭回归	期望经验平方误差	$\lambda R^2 + \sigma^2 \hat{D}(\lambda)/n$ 。
RKHS 约束最小二乘	经验平方误差	$n^{-q/(q+1)}$ ；经验谱 $\hat{\mu}_j \lesssim j^{-q}$ ， $q > 1$ 。

相同的幂次并不意味着相同的结论。判断一个学习定理时，至少同时读清楚：模型假设、噪声假设、误差范数、期望或高概率形式，以及隐藏常数的依赖。对核平滑，附录 A 还在高斯回归模型下证明了匹配下界。对上述经验谱结论，本文只给出相应上界。

8.3 可以带到别的问题里的证明技巧

技巧	什么时候用，为什么有效
加减同一个比较量	算法优化的是经验目标，结论却关心总体目标；用比较函数把两者接上。
先条件化再平方	权重或设计是随机的；先固定它们，中心化噪声便能消去交叉项。
指数矩与参数优化	独立求和；指数把和变成乘积，最后优化一个实参数。
有限分辨率与补空隙	函数类无限；先控制有限网，再用所选度量控制剩余误差。
幽灵样本与交换对称	总体均值未知；用独立副本替代，再随机交换样本对。
随机稀疏化	对象是凸组合或均值；抽样平均给出一个可数的近似，独立性消去交叉项。
曲率与星形缩放	误差代价是二次而噪声收益是线性；控制一个局部球面即可排除远处解。
正交投影	数据只依赖有限个线性泛函；删除数据不可见却消耗范数的方向。
难区分的局部变化	需要下界；把估计问题还原成许多二元检验，再累加错误。

9 讨论题与解题提示

讨论题 1：ERM 的比较项。为什么 $R_n(\hat{f}) - R_n(f_{\mathcal{F}}^*) \leq 0$ ，而 $R(\hat{f}) - R(f_{\mathcal{F}}^*) \geq 0$ ？这两个不等式分别来自什么最优性？若只求到 ε_{opt} 精度，哪一项会改变？

提示。第一个比较的是经验目标，第二个比较的是总体目标。在命题 1.2 的第一行保留三项，再逐项使用对应的定义。

讨论题 2：维数究竟让样本量增加多少？固定 $\beta = 1$ 。分别写出 $d = 1, 2, 10$ 时的最优平方风险指数，以及达到均方根误差 ε 时的样本量指数。

解。平方风险分别为 $n^{-2/3}$ 、 $n^{-1/2}$ 、 $n^{-1/6}$ ；对应的样本量阶分别为 ε^{-3} 、 ε^{-4} 、 ε^{-12} 。这些比较固定了函数类、噪声和其他常数，不能直接当作实际数据量预测。

讨论题 3：为什么范数条件能改变维数依赖？ 假设输入特征满足 $\|x_i\|_\infty \leq 1$ 。比较 $\|w\|_1 \leq A$ 与 $\|w\|_2 \leq A$ 两个线性预测类的复杂度。

提示。 前者可得 $A\sqrt{2\log(2D)/n}$ ；后者从 $\|x_i\|_2 \leq \sqrt{D}$ 只能直接得到 $A\sqrt{D/n}$ 。两类函数并不相同，且它们的最佳表示范数也可能不同。因此不能只看一个界中有没有显式的 D 。

讨论题 4：线性最小二乘的精确常数。 固定设计，假设真均值向量 m^* 属于一个 p 维线性子空间 $V \subseteq \mathbb{R}^n$ ， $y = m^* + \sigma Z$ ， $Z \sim N(0, I_n)$ 。用投影直接求 $\mathbb{E} \|\hat{m} - m^*\|_2^2 / n$ 。

解。 最小二乘解是 $\hat{m} = \Pi_V y$ ，故误差为 $\sigma \Pi_V Z$ 。于是

$$\frac{1}{n} \mathbb{E} \|\hat{m} - m^*\|_2^2 = \frac{\sigma^2}{n} \text{tr}(\Pi_V) = \frac{\sigma^2 p}{n}.$$

通用的局部复杂度论证恢复相同量级，显式投影则利用更多结构给出精确常数。

讨论题 5：为什么普通 L^2 不是 RKHS？ 取 $x_0 \in (0, 1)$ ，令 $f_m(x) = (1 - m|x - x_0|)_+$ 。证明 $\|f_m\|_{L^2} \rightarrow 0$ ，但 $f_m(x_0) = 1$ 。

解。 对足够大的 m ， $\|f_m\|_{L^2}^2 = 2/(3m)$ 。所以即便只看连续函数，点值也不是关于 L^2 范数的连续泛函。完整 L^2 空间中的元素还是几乎处处等价类，点值本身也没有天然定义。RKHS 的再生性质恰好提供了所缺少的连续性。

讨论题 6：Maurey 方法到底逃开了什么？ 为何随机平均可以给出 $m^{-1/2}$ 的 Hilbert 范数误差，却不推翻一般 Hölder 函数类的维数灾难？

提示。 依次检查目标是否有受控范数的字典表示、原子的 Hilbert 范数是否受控，以及所控制的是 L^2 还是一致范数。定理改变了函数类及其几何假设；表示范数也可能携带维数依赖。

A 从小凸起到非参统计下界（课后）

本附录为维数灾难补上一个统计意义上的下界。证明只需高斯密度的计算、一个二元检验不等式，以及把整体函数拆成许多局部选择的想法。我们完整处理 $0 < \beta \leq 1$ ，与正文箱形核平滑的上界使用相同的平滑函数类。

A.1 先证明一个二元检验工具

对两个概率分布 Q_0, Q_1 ，记

$$\mathrm{TV}(Q_0, Q_1) = \sup_A |Q_0(A) - Q_1(A)|, \quad \mathrm{KL}(Q_0 \| Q_1) = \int \log \frac{dQ_0}{dQ_1} dQ_0.$$

若 Q_0 不绝对连续于 Q_1 ，约定 KL 为 $+\infty$ 。

引理 A.1 (两个世界很相近时，判别必定会出错)。任何根据观测输出 $\hat{\theta} \in \{0, 1\}$ 的检验，都满足

$$Q_0(\hat{\theta} = 1) + Q_1(\hat{\theta} = 0) \geq 1 - \mathrm{TV}(Q_0, Q_1) \geq 1 - \sqrt{\frac{\mathrm{KL}(Q_0 \| Q_1)}{2}}.$$

证明. 记 $A = \{\hat{\theta} = 1\}$ ，两种错误概率之和为 $Q_0(A) + 1 - Q_1(A) \geq 1 - \mathrm{TV}(Q_0, Q_1)$ 。第二步也可从头证明：将密度比分别在 A 与 A^c 上用 Jensen 不等式平均，得到

$$\mathrm{KL}(Q_0 \| Q_1) \geq p \log(p/q) + (1-p) \log\{(1-p)/(1-q)\},$$

其中 $p = Q_0(A), q = Q_1(A)$ 。把右边看成 p 的函数，在 $p = q$ 处函数值和一阶导数均为零，二阶导数为 $1/[p(1-p)] \geq 4$ ，故右边至少为 $2(p-q)^2$ 。对 A 取上确界即得第二步；端点情形取极限即可。随机化检验可以把额外独立随机数包含在观测中，结论不变。□

A.2 把一个未知函数变成许多未知比特

令 $\phi(u) = (1 - 4\|u\|_\infty)_+$ ，其中 $a_+ = \max(a, 0)$ 。它支撑在 $[-1/4, 1/4]^d$ 上， $\phi(0) = 1$ ，且

$$|\phi(u) - \phi(v)| \leq \min\{1, 4\|u - v\|_\infty\} \leq 4^\beta \|u - v\|_\infty^\beta.$$

令 $c_\phi = \int_{\mathbb{R}^d} \phi(u)^2 du \in (0, \infty)$ 。把 $[0, 1]^d$ 分成 $M = q^d$ 个边长 $h = 1/q$ 的方块，中心为 c_j 。取 $0 < a \leq L/4$ ，定义

$$b_j(x) = ah^\beta \phi((x - c_j)/h), \quad f_\theta(x) = \sum_{j=1}^M \theta_j b_j(x), \quad \theta \in \{0, 1\}^M.$$

凸起非负且支撑互不相交，所以 $f_\theta(x) = \max_j \{\theta_j b_j(x)\}$ 。每个 b_j 的 Hölder 常数至多为 $a4^\beta \leq L$ ；有限个具有相同 Hölder 常数的函数，其最大值仍具有该常数。因此

$$|f_\theta(x) - f_\theta(y)| \leq L\|x - y\|_\infty^\beta, \quad \|f_\theta\|_\infty \leq L.$$

因此所有 f_θ 都落在正文定义的 $\mathcal{H}^\beta(L)$ 内。每个凸起的平方能量恰为

$$v_h := \|b_j\|_{L^2([0,1]^d)}^2 = a^2 c_\phi h^{2\beta+d}. \quad (\text{A.1})$$

证明直觉 每个小区域藏着一个“有凸起还是没有凸起”的比特。凸起越窄，看到它的样本越少；凸起越低，即便看到样本也越难穿过噪声判断。我们把宽度选到恰好难以判别的尺度，再把所有局部错误累加起来。

定理 A.2 (Hölder 回归的平方风险下界)。固定 $d \geq 1$, $0 < \beta \leq 1$, $L > 0$, $\sigma > 0$ 。考虑随机设计高斯回归

$$X_i \stackrel{\text{iid}}{\sim} \text{Unif}([0, 1]^d), \quad Y_i = f(X_i) + \xi_i, \quad \xi_i \stackrel{\text{iid}}{\sim} N(0, \sigma^2),$$

其中噪声独立于设计。存在只依赖 d, β, L, σ 的 $c > 0$ 与 n_0 ，使所有 $n \geq n_0$ 满足

$$\inf_{\hat{f}} \sup_{f \in \mathcal{H}^\beta(L)} \mathbb{E}_f \left\| \hat{f} - f \right\|_{L^2([0, 1]^d)}^2 \geq cn^{-2\beta/(2\beta+d)}. \quad (\text{A.2})$$

下确界遍历所有可测估计器，允许它们随机化。

证明. 若某个候选分布下的期望平方误差为无穷，下界显然成立；其余情形在 L^2 中作以下比较。随机化估计器的独立随机种子可并入观测，下面的 KL 不变。

第一步：相邻两个函数产生的样本分布有多近？ 用 Q_θ 表示 f_θ 对应的全部训练数据分布。若 θ 与 θ' 只在第 j 个比特不同，高斯均值平移的 KL 公式给出

$$\text{KL}(Q_\theta \| Q_{\theta'}) = \frac{n}{2\sigma^2} \|f_\theta - f_{\theta'}\|_{L^2}^2 = \frac{nv_h}{2\sigma^2}. \quad (\text{A.3})$$

这个公式来自条件于 $X = x$ 后两个方差同为 σ^2 的高斯密度：其对数密度比的期望为两均值之差的平方除以 $2\sigma^2$ ；再对 X 积分，并利用独立观测的对数似然比相加。

选择

$$q = \left\lceil \left(\frac{4na^2c_\phi}{\sigma^2} \right)^{1/(2\beta+d)} \right\rceil, \quad h = q^{-1}.$$

当 n 足够大时，这使得 $\text{KL}(Q_\theta \| Q_{\theta'}) \leq 1/8$ ，从而 $\text{TV}(Q_\theta, Q_{\theta'}) \leq 1/4$ 。

第二步：任何函数估计器都能诱导一个逐块检验。 给定 $\hat{f}$ ，在第 j 个方块 Q_j 上比较它离 0 与 b_j 哪个更近，据此输出 $\hat{\theta}_j$ 。若判错，三角不等式保证

$$\left\| \hat{f} - \theta_j b_j \right\|_{L^2(Q_j)}^2 \geq \frac{1}{4} \|b_j\|_{L^2(Q_j)}^2 = \frac{v_h}{4}.$$

例如 $\theta_j = 0$ 却选择了 b_j ，则 $\left\| \hat{f} - b_j \right\|_{L^2(Q_j)} \leq \left\| \hat{f} \right\|_{L^2(Q_j)}$ ，因此 $\|b_j\|_{L^2(Q_j)} \leq 2 \left\| \hat{f} \right\|_{L^2(Q_j)}$ 。另一个方向同理。于是

$$\mathbb{E}_\theta \left\| \hat{f} - f_\theta \right\|_{L^2}^2 \geq \frac{v_h}{4} \sum_{j=1}^M Q_\theta(\hat{\theta}_j \neq \theta_j).$$

第三步：平均每个比特的两种可能。 对所有 θ 均匀平均。固定其余 $M-1$ 个比特后，把第 j 个比特为 0 和 1 的两个世界配成一对。引理 A.1 说明，这一对的两种错误概率之和至少为 $3/4$ ，所以均匀平均后的错误概率至少为 $3/8$ 。因此

$$\sup_{\theta} \mathbb{E}_\theta \left\| \hat{f} - f_\theta \right\|_{L^2}^2 \geq \frac{3Mv_h}{32} = \frac{3a^2c_\phi}{32} h^{2\beta}.$$

当 n 足够大时，取整只改变常数， $h \asymp n^{-1/(2\beta+d)}$ ，得到结论。 $\square$

为什么与核平滑得到同一个速率？ 核平滑的上界在比较 $h^{2\beta}$ 与 $\sigma^2/(nh^d)$ 。下界把局部假设之间的可区分性选在

$$\frac{nh^d \cdot h^{2\beta}}{\sigma^2} \asymp 1.$$

左边就是“局部样本数乘以单个样本的信号强度”。上界与下界其实在观察同一个临界尺度。因此，在这里明确给定的函数类与观测模型下，正文的核平滑上界达到最优的样本量指数。这给原课程非参回归中的最优速率补充了一条完整的检验证明[4]。

B 高斯集中不等式的完整证明（课后）

本段补上引理 4.7 所需的概率工具。只需在首次阅读时掌握结论及其 Lipschitz 常数，下面的半群计算可以在熟悉 Gaussian integration by parts 后细读。

引理 4.7 的证明。记标准高斯分布为 γ ，并定义

$$P_t h(x) = \mathbb{E}[h(e^{-t}x + \sqrt{1-e^{-2t}}Z)].$$

先对光滑、各阶导数有界、严格正且有正下界的有界函数 h 计算。高斯分布在 P_t 下不变，且 $\nabla P_t h = e^{-t} P_t \nabla h$ 。由 Gaussian integration by parts，生成元 $\mathcal{A} = \Delta - x \cdot \nabla$ 满足 $\mathbb{E}_\gamma[u \mathcal{A}v] = -\mathbb{E}_\gamma \langle \nabla u, \nabla v \rangle$ 。因此

$$\frac{d}{dt} \mathbb{E}_\gamma[(P_t h) \log P_t h] = -\mathbb{E}_\gamma \frac{\|\nabla P_t h\|_2^2}{P_t h}.$$

把 t 从 0 积分到 ∞ ，左边的极限给出

$$\text{Ent}_\gamma(h) := \mathbb{E}_\gamma[h \log h] - (\mathbb{E}_\gamma h) \log(\mathbb{E}_\gamma h) = \int_0^\infty \mathbb{E}_\gamma \frac{\|\nabla P_t h\|_2^2}{P_t h} dt.$$

带权 Cauchy 不等式保证

$$\|P_t \nabla h\|_2^2 \leq (P_t h) P_t \left(\frac{\|\nabla h\|_2^2}{h} \right).$$

利用不变性并积分 e^{-2t} ，得到

$$\text{Ent}_\gamma(h) \leq \frac{1}{2} \mathbb{E}_\gamma \frac{\|\nabla h\|_2^2}{h}. \quad (\text{B.1})$$

这就是高斯对数 Sobolev 不等式的这个形式。

对光滑的 L Lipschitz 函数 H 取 $h = e^{\lambda H}$ 。先截断 H 再作光滑逼近即可使用上式；由于 Lipschitz 函数至多线性增长，高斯可积性允许去掉截断。记 $\psi(\lambda) = \log \mathbb{E} e^{\lambda(H - \mathbb{E}H)}$ 。式 (B.1) 给出

$$\lambda \psi'(\lambda) - \psi(\lambda) \leq \frac{\lambda^2 L^2}{2}.$$

将其除以 λ^2 并从 0 到 $\lambda > 0$ 积分，利用 $\lim_{\lambda \downarrow 0} \psi(\lambda)/\lambda = 0$ ，得到

$$\log \mathbb{E} e^{\lambda(H - \mathbb{E}H)} \leq \frac{\lambda^2 L^2}{2}.$$

最后用指数 Markov 不等式，并取 $\lambda = u/L^2$ ，得 $\mathbb{P}\{H - \mathbb{E}H \geq u\} \leq e^{-u^2/(2L^2)}$ 。令 $u = L\sqrt{2a}$ 即得结论。一般 Lipschitz 函数通过保持 Lipschitz 常数的光滑逼近得到同一结论。□

C 只给出半正定核，怎样恢复函数空间

证明直觉 既然 k_x 应当是“读取 x 处函数值”的方向，那么这些方向的内积只能取为 $\langle k_x, k_z \rangle_{\mathcal{H}} = k(x, z)$ 。接下来先允许有限线性组合，合并不产生函数变化的冗余表示，再补上极限。这就是下面构造的每一步分别在做的事情。

命题 C.1 (从核构造空间)。每个正定核 k 都对应一个 RKHS，其核截面为 $k_x(\cdot) = k(x, \cdot)$ ，并满足

$$|f(x)| \leq \|f\|_{\mathcal{H}} \sqrt{k(x, x)}, \quad \|k_x\|_{\mathcal{H}}^2 = k(x, x). \quad (\text{C.1})$$

证明. 第一步：先在有限组合上规定内积。以形式符号 e_x 为生成元，考虑有限和 $u = \sum_i a_i e_{x_i}$ 。规定双线性形式

$$B\left(\sum_i a_i e_{x_i}, \sum_j b_j e_{z_j}\right) = \sum_{i,j} a_i b_j k(x_i, z_j).$$

正定性给出 $B(u, u) \geq 0$ 。由 $B(u + tv, u + tv) \geq 0$ 对所有 $t \in \mathbb{R}$ 成立，得到 Cauchy 与 Schwarz 不等式 $|B(u, v)|^2 \leq B(u, u)B(v, v)$ 。

第二步：合并零范数的表示。令 $\mathcal{N} = \{u : B(u, u) = 0\}$ 。上述不等式说明 $u \in \mathcal{N}$ 时， $B(u, v) = 0$ 对每个 v 都成立，所以对 $\mathcal{N}$ 取商以后， B 成为真正的内积。商类 $[u]$ 对应函数

$$f_u(x) = B(u, e_x) = \sum_i a_i k(x_i, x).$$

零范数表示不会改变任何函数值。反过来，若 $f_u(x)$ 处处为零，则 $B(u, u) = \sum_i a_i f_u(x_i) = 0$ ，所以两个商类也不会代表同一个非零差函数。

第三步：允许极限，并保留点值。对上述内积空间取完备化，得到 Hilbert 空间 $\mathcal{H}$ 。有限组合上已有

$$|f_u(x)| = |B(u, e_x)| \leq \sqrt{B(u, u)} \sqrt{k(x, x)}.$$

因此每个取值映射都连续延拓到完备化，且 $f(x) = \langle f, k_x \rangle_{\mathcal{H}}$ 。如果一个极限元素在每一点都取零，它就与所有 k_x 正交，因而与这些截面的稠密线性张成正交，只能是零元素。这保证完备化仍可确实看作函数空间。最后取 $f = k_x$ ，即得 (C.1)。□

证明同时说明再生性质为何意味着取值连续：函数在 $\mathcal{H}$ 范数中变化很小，则每一个固定位置的函数值也变化很小。若 $\sup_x k(x, x) \leq \kappa^2$ ，还得到统一的界 $\|f\|_{\infty} \leq \kappa \|f\|_{\mathcal{H}}$ 。

参考文献与原课程资料

- [1] 陆一平. *Statistical Learning, IEMS 402*, 美国西北大学, 2025 年冬季课程主页.
<https://2prime.github.io/teaching/2025-Statistical-Learning>.
- [2] Yiping Lu. *Lecture 1 and 2: Overview*. IEMS 402 课程讲义.
https://2prime.github.io/files/IEMS_402_Note_lec1.pdf.
- [3] Yiping Lu. *Lecture 1 and 2: What is Machine Learning?*. IEMS 402 课程幻灯片, 第 46 页的高维几何图及批注.
<https://2prime.github.io/files/IEMS402/402lec1and2.pdf>.
课堂标注版: <https://2prime.github.io/files/IEMS402/402lec1and2fill.pdf>.
- [4] Yiping Lu. *Lecture 3: Nonparametric Regression and Density Estimation*. IEMS 402 课程讲义.
https://2prime.github.io/files/IEMS_402_Note_lec3.pdf.
- [5] Yiping Lu. *Lecture 7: Concentration Inequalities*. IEMS 402 课程讲义.
https://2prime.github.io/files/IEMS402/IEMS_402_Note_lec7.pdf.
- [6] Yiping Lu. *Lecture 7: Concentration Inequalities*. IEMS 402 课程幻灯片, 特别参见第 6 至 15 页.
<https://2prime.github.io/files/IEMS402/402lec7.pdf>.
课堂标注版: <https://2prime.github.io/files/IEMS402/402lec7fill.pdf>.
- [7] Yiping Lu. *Lecture 8: Generalization and Covering Numbers*. IEMS 402 课程讲义.
https://2prime.github.io/files/IEMS402/IEMS_402_Note_lec8.pdf.
- [8] Yiping Lu. *Lecture 9: Rademacher Complexity and Chaining*. IEMS 402 课程讲义.
https://2prime.github.io/files/IEMS402/IEMS_402_Note_lec9.pdf.
- [9] Yiping Lu. *Lecture 11: Localized Complexity*. IEMS 402 课程幻灯片.
<https://2prime.github.io/files/IEMS402/402lec11.pdf>.
- [10] Yiping Lu. *Reproducing Kernel Hilbert Space, Part 1*. IEMS 402 手写讲义.
<https://2prime.github.io/files/IEMS402/rkhs1.pdf>.
- [11] Yiping Lu. *Reproducing Kernel Hilbert Space, Part 2*. IEMS 402 手写讲义.
<https://2prime.github.io/files/IEMS402/rkhs2.pdf>.
- [12] Eric Price. *Lecture 13: Maurey's Empirical Method*. Sublinear Algorithms, University of Texas at Austin, 2016.
<https://www.cs.utexas.edu/~ecprice/courses/sublinear/fa16/notes/lec13.pdf>.
- [13] Yudong Chen. *Lectures 16 and 17: Nonparametric Least Squares*. CS 839, Probability and Learning in High Dimension, University of Wisconsin Madison, 2022.
https://pages.cs.wisc.edu/~yudongchen/cs839_sp22/16-17_nonparametric_least_squares.pdf.
- [14] Peter L. Bartlett, Olivier Bousquet and Shahar Mendelson. Local Rademacher complexities. *The Annals of Statistics*, 33(4), 1497 至 1537, 2005.
<https://doi.org/10.1214/009053605000000282>.
- [15] Francis Bach. *Learning Theory from First Principles*. MIT Press, 2024. 原课程教材; 可对照风险分解、统一收敛、局部平均与核方法相关章节.
<https://www.di.ens.fr/~fbach/ltfp/>.

资料说明: 本讲以原课程资料为起点重新组织内容与证明, 统一风险、范数、样本量和经验谱的记号。部分参考条目按本次讨论班主题标注, 便于与课程主页定位。文中所有有限样本结论均给出所使用的假设; 课堂记录中的简写不作为省略假设的依据。