统计建模

作者：_____

2026 年 3 月 16 日

摘要

本文整理统计模型的核心框架，围绕条件分布建模、估计与推断、模型诊断与选择展开，并补充两类常用计算工具：自助法（Bootstrap）与期望最大化算法（EM）。

关键词：统计建模；线性模型；广义线性模型；Bootstrap；EM；缺失数据

1 引言

本书的目的不是罗列与展示各种统计模型，而是帮助读者建立一种*推断秩序*：在一个永远无法被完整观测的世界里，我们如何用有限的、带偏的、并且受机制筛选的数据，形成可辩护的知识？在这个意义上，统计建模并非“从数据里提取真相”，而是一种**受约束的认知实践**——它要求我们在证据不足时明确不确定性与知识空缺，在假设介入时说明这些假设带来的识别代价与外推范围，并始终记得：结论只在其前提与适用条件之内才成立。

这种实践有一条不可倒置的次序：先说清**你究竟要回答什么**——研究问题与对应的 estimand；再交代**你为何能回答它**——数据从何而来、对照如何成立、识别依赖哪些关键假设；最后才讨论**你如何把答案算出来并说明可信度**——选择何种模型与算法、如何评估、如何做推断并量化不确定性。因此，本书把统计建模理解为一套**相互衔接的认知实践**：

- **对象层面：**你在问什么？estimand 如何定义？目标人群与适用边界是什么？
- **识别与证据层面：**数据通过什么机制产生并进入样本？哪些比较构成有效对照？主要威胁（混杂、选择、测量误差等）是什么？
- **估计与不确定性层面：**在既定识别路径下如何估计？不确定性如何量化与检验？有限样本与模型错设的风险如何诊断与报告？

这条主线把识别（*能不能回答*）、方法（*怎么回答*）与保障（*在什么条件下、能做到多好*）连成一个整体。统计建模获得的每一个结论都涉及：它基于哪些数据与比较而来，关键依赖哪些假设，以及结论在什么范围内成立、在什么情况下可能失效。

1.1 对因果问题的回答

统计建模面对的是往往是带有因果含义的问题——*做与不做会怎样*——我们必须把讨论从“拟合一个模型”前移到更根本的两件事：**先定义要回答的因果问题，再说明这份数据为何能识别或者说回答这个问题**。只有当这两步站住之后，模型、算法与推断才有明确的指向：**在有限样本下如何计算、如何评估、以及如何量化不确定性？**

因果问题不只是“有没有效”，还包含了对对象与边界的选择：对谁有效（总体、亚组、边缘人群）？什么处理版本（一次性、持续性、强度不同、组合干预）？在哪个时间尺度上有效（短期、长期、动态反馈）？一旦这些没有说清，所谓“估计结果”往往只是某个便利定义下的数值，而不是你原本关心的政策或科学问题。因此，“问题先行”的最低要求是：先把要推断的因果量写成一个明确对象，再决定用什么数据与模型去探究它。

同一个 estimand 放到不同的数据生成机制下，推断含义会发生根本变化。举例来说：

- **随机试验：**随机化在设计层面制造可比性，使组间差异可以被解释为处理与对照的差异。
- **自然实验：**制度规则或者外界环境提供近似随机的变异；关键在于论证这种变异为何可被视为“接近随机”。
- **观测研究：**处理往往由行为与制度决定；要把比较解释为因果效应，必须依赖明确的识别假设（例如关于混杂、选择的约束），并用数据与背景知识说明这些假设为何合理。

因此，**统计建模的核心不是选择某种估计器，而是建立一套证据结构：**你要说明比较从何而来，主要偏差机制如何被控制，以及结论依赖哪些不可检验但可辩护的假设。当建模过程缺乏论证，再复杂的算法也只能给出相关性的拟合输出；当建模的路径清晰时，简单的估计方式反而更容易被理解、检验与信任。

1.2 模型的角色：推断的语言，而非因果的替身

在“问题先行”的框架下，统计模型不应被理解为“只要拟合得好就等于理解了世界”。模型是一种推断语言：它把研究问题中需要被估计的对象与数据之间的关系，组织成可计算、可交流、可检验的形式，从而让“问题—数据—结论”的推理过程可被复现与审查。

模型的特点：把复杂性变成可处理的结构。现实的数据生成过程往往高度复杂且不可完全观测；模型的作用是做出结构化的简化，用有限的参数、函数类或约束来刻画我们愿意表达的关系。线性回归用线性与加性给出最易解释的近似；广义线性模型用链接函数与指数族扩展到非正态结局；半参数模型在保留关键目标量可解释性的同时放松部分分布假设；非参数与现代学习方法则扩大函数空间，以更弱的形式假设换取对复杂规律的拟合能力。无论复杂与否，这些方法的共同点是：**模型不是“还原世界”，而是“选择一种结构来表达世界的某一部分”。**

模型能提供的价值：表达、估计、诊断。

- **表达：**把研究对象写成明确的数学形式（参数、函数、条件期望、风险函数等），便于沟通、比较与复现。
- **估计：**在样本有限、目标复杂的情况下，模型给出可实施的估计策略（闭式解、优化问题、正则化、算法流程）。
- **诊断：**模型把一部分假设暴露在数据面前，使我们可以进行检查（残差结构、拟合与校准、敏感性、外推/稳定性等），从而知道结论在哪些维度上脆弱。

模型的局限：拟合不等于识别，解释不等于因果。模型最常见的误用，是把“拟合得更好”误当作“回答得更对”，或者把“条件关联”自动升级为“干预效应”。在因果问题上，参数的因果含义不由模型本身赋予，而取决于它是否嵌入一个可辩护的识别框架：

- **对照从何而来：**随机化、外生规则/冲击，还是依赖可交换性/重叠等假设构造的对照？

- **主要偏差机制如何处理：**混杂、选择、同时性、测量误差、干预外溢等是否被控制或界定在可接受范围内?
- **估计对象是否一致：**你声称的 estimand 是否与数据实际支持的对象一致，是否因选择/删失/外推而被“悄悄替换”?

如果这些问题没有被回答，再灵活的算法也只能给出“在这份数据上相关性如何变化”的描述。当识别路径清楚时，简单透明的模型往往更有说服力，因为它让证据呈现得更直接、假设更容易被看见与质疑；当识别不足时，把希望寄托在“更复杂的模型”上通常只会把问题隐藏得更深，而不是把结论变得更可信。

1.3 三类常用的统计方法：模型评估、Bootstrap、EM 方法

模型评估帮助你评价模型表现并进行模型选择，**Bootstrap** 让你在复杂统计量下仍能量化抽样波动，**EM** 则让你在缺失与潜变量场景中把“看不见”的结构纳入推断。它们是统计建模中三类特别实用的方法，分别对应建模流程中的三个核心环节：

1. **模型评估（model assessment）：**回答“这个模型在这份数据机制下够不够用”。评估不仅仅是报一个 R^2 或 AUC，更包括残差诊断、拟合优度与校准、样本外验证、交叉验证与敏感性检查等。它的目标是把模型从“一个可以拟合的公式”变成“一个在关键方面不违背数据、且能支撑你的结论”的工具：该捕捉的结构有没有捕捉到？关键假设（线性、独立、同方差、链接函数、平行趋势等）是否与数据明显冲突？若模型只是方便的近似，结论对近似误差是否稳健?
2. **重抽样（bootstrap）：**回答“在不依赖解析近似时，不确定性有多大”。当统计量的抽样分布难以解析推导，或常规渐近近似在有限样本下不可靠时，bootstrap 用数据自身近似重复抽样过程，从而估计标准误、置信区间与偏差修正。它把“不确定性”从一页公式推导转化为一套可执行的计算流程，也提醒我们：推断的可靠性同样依赖于抽样/重抽样机制是否与数据生成机制相容。
3. **缺失与潜变量结构下的迭代极大化（EM）：**回答“当世界的一部分不可见时，如何仍然最大化似然并完成估计”。在含潜变量、隐状态或缺失数据的模型里，似然往往不可直接最大化。EM 的思想是把“不可见”当作缺失部分，在 E 步计算其条件期望（或充分统计的期望），在 M 步更新参数，从而把难优化的问题拆成一串可计算的子问题。它提供了一条将机制假设（缺失机制、测量过程、隐状态演化）转化为可实施估计器的通用路径。

1.4 理论分析：经验风险最小化与 Minimax 框架

在高维、非参数与现代学习问题中，我们关心的不止是“能不能估计”，更关心在有限样本下能做到多好：误差如何随样本量 n、模型/函数类复杂度与噪声水平变化，以及为了达到某种精度需要付出什么代价（结构假设、正则化强度与计算资源）。经验风险最小化（ERM）提供了统一的表述：给定损失 ℓ 与函数类 $\mathcal{F}$，用经验风险 $\hat{R}_n(f) = \frac{1}{n}\sum_{i=1}^{n} \ell(f(X_i), Y_i)$ 近似真实风险 $R(f) = \mathbb{E}[\ell(f(X), Y)]$，并取

$$\hat{f} \in \arg\min_{f\in\mathcal{F}} \hat{R}_n(f) \quad \text{或} \quad \hat{f} \in \arg\min_{f\in\mathcal{F}} \{\hat{R}_n(f) + \lambda\,\Omega(f)\}.$$

理论分析的关键是：控制 $\sup_{f\in\mathcal{F}}|\hat{R}_n(f)-R(f)|$，从而得到风险上界，其规模由函数类复杂度（VC 维、Rademacher 复杂度、覆盖数/熵、局部化复杂度等）决定。最终误差通常可写成近似误差 + 估计误差的权衡：$\mathcal{F}$ 越大表达能力越强、近似误差可能更小，但估计误差随复杂度上升而增大；正则化与算法选择的作用，就是在给定 n 下实现这两者的可控折中。

结构性难度的标尺：minimax、下界与最优性（简写） ERM 主要回答“某类方法在给定函数类与正则化下可以做到多好”，而 minimax 则把视角提升到任务本身：在给定模型/分布族 $\mathcal{P}$（或函数类 $\mathcal{F}$）与损失度量下，

$$\inf_{\hat{\theta}}\sup_{P\in\mathcal{P}}\mathbb{E}_P\big[d(\hat{\theta},\theta(P))\big] \quad 或 \quad \inf_{\hat{f}}\sup_{f\in\mathcal{F}}\mathbb{E}_f\big[\mathcal{L}(\hat{f},f)\big]$$

刻画了**任何估计器在最坏情况下最多只能达到的误差水平**。某些误差率来自信息论层面的不可区分性，并非“算法还不够聪明”。因此 minimax 一方面给出误差率如何随 n、维度、平滑度、稀疏度等结构参数变化，从而界定问题难度；另一方面提供判断最优性的客观标尺：若某个 ERM/正则化/网络估计器的上界与下界在常数或对数因子意义下匹配，则可称其（近似）minimax 最优；若差距很大，则说明方法或分析尚未利用关键结构。更重要的是，下界常反向揭示“必须引入怎样的结构约束（平滑/稀疏/低秩/架构偏置）才能避免退化”，从而把算法选择与建模假设放回同一套可比较的理论框架中。

用这种框架看待现代方法（核方法、稀疏学习、神经网络等），我们关心的不只是“它能拟合”，而是：它隐含了什么结构约束？复杂度如何被控制？在什么函数类上能达到怎样的误差率？这些误差率是否接近信息论下界？当这些问题被放进同一套 ERM–minimax 语言里，现代学习方法的经验现象就更容易被解释为“结构、复杂度与样本量”三者之间的可计算权衡，而不只是调参技巧的堆叠。

1.5 比模型更重要的是判断力

我们希望读者最终获得的不是模型与技巧的清单，而是一种可迁移的判断能力：能够提出可检验、可计算的主张，并清楚说明其前提、代价与适用边界。

2 问题先行：设计决定回答能力

统计学家和数据科学家在统计建模时，往往需要面对一些非常具体的因果问题：**如果我对某个单位施加一个干预，结果会不会因此改变、改变多少？**推行政策、给药、开展筛查、优化流程——我们想知道的不是“谁和谁一起变化”，而是“做与不做的差异”。但这类问题的困难在于：它指向的比较对象并不直接存在于数据中。最理想的比较是同一个人走两条路——一条路接受处理，一条路不接受处理，然后把两种结果一比，效应就出来了。现实却只允许一条路发生，另一条路永远不可见。这就是反事实 (Counterfactural) 难题：**你缺的不是一个更复杂的模型，而是一种可信的对照生成机制**。换句话说，因果推断的第一道门槛，不是“拟合得够不够好”，而是“比较对象从哪里来”。在观察数据里，我们看到的是已经发生的那条路：谁接受了处理、在什么背景下接受、随后出现了什么结果；但我们想比较的是同一批人在同一时间面对两种可能世界的差异。这个差异本身并不在数据表里，它必须由某种机制“补出来”。

因此，接下来讨论因果问题时，最容易犯的错误是把任务误读成“从相关到因果只差一个更灵活的模型”。事实上，从相关走向因果，缺的通常是一块更硬的基座：要么有一个机制让处理

组与对照组在处理前可比（如随机化或近似随机的制度变化）；要么你能用足够强的信息与假设把“谁会接受处理”解释清楚，从而在可观测维度上把对照“对齐”。一旦这块基座不稳，模型越复杂，往往只是把“不可比带来的偏差”包装得更精致。也正因为如此，我们需要按一个更可靠的顺序来推进：先把想比较的“效果”定义清楚，再检查你手上的数据是否真的提供了一个可信的反事实来源，最后才谈同一份数据究竟能回答什么、回答不了什么。

第一步：把问题问清楚——你要的“效果”到底是什么？ 同一句“处理有效吗”，背后可能是不同的问题：对谁有效（总体还是某类人群）？处理的具体版本是什么（强度、时长）？效果在哪个时间尺度上（短期、长期、动态反馈）？在什么制度与环境下成立？你的数据覆盖什么人群、处理什么版本、观察什么窗口，决定了后面进一步谈“能不能回答”的边界。

第二步：回答能力从哪里来——你有没有一个可信的反事实来源？ 既然反事实看不到，能不能回答因果问题，取决于你能否用研究设计提供一个可信的对照。常见的来源大致有三类：

- **随机对照实验 (Randomized Controlled Trial, RCT)：**用抽签决定“谁接受处理”。随机化的价值不在于它让模型更漂亮，而在于它（在理想情况下）让两组在处理前可比，从而把“差异”更自然地解释为“处理造成的变化”。当然，RCT 也会失效：不依从、失访、污染、口径变化等都会让你最终比较的对象被悄悄替换。
- **自然实验：**研究者没有亲手抽签，但现实制度偶尔提供“近似随机”的变化源——阈值规则、名额抽签、分期上线、边界、意外冲击等。自然实验的关键不是回归形式，而是把分配机制讲清楚：**为什么在你关心的范围内，这个变化不像是人们自己挑出来的**；并用证据支持“对照确实像”。
- **观测研究：**处理往往由个体或机构自选。此时回答能力高度依赖：你是否真的拥有解释“谁会接受处理、为什么接受处理”的信息。如果关键混杂因素没被测到、测得不准，或选择机制强烈依赖不可观测动机与资源，那么再精细的回归也只能在假设上行走——它可以把差异算得更精确，却不能把不可比的对照变得可比。

第三步：同一份数据，能回答什么、回答不了什么。 最常见的研究错位是：**数据不差，但它没有能力回答你想问的那个因果问题**。一项随机对照实验得到的数据也许能可靠回答“被随机分到处理组”的平均效应，却未必能直接回答“真正接受处理者”的效应；一项政策前后对比可能混入共同冲击与测量口径变化；一项大规模观测数据库可能预测极佳，却对“干预是否有效”缺乏可信对照。这里的关键结论是：

设计决定你能识别什么；模型只是在可识别的范围内，把它算出来。

本章路线：用三个真实案例来审问设计的可信度。 接下来我们不急着堆公式，而是围绕同一个动作：**审问一项研究设计有没有回答能力，以及可信度来自哪里**。我们将依次讨论三个实际案例，分别代表随机对照实验、自然实验与观测研究。对每个案例，我们都用同一套问题去拆解：

1. **因果问题是什么？**（效果定义、目标人群、时间尺度）
2. **对照从哪里来？**（随机化/制度规则/观测调整）
3. **这份设计真正识别的是什么？**（比较对象是否被“悄悄替换”）

4. **最可能的失效点在哪里？**（不依从、失访、污染、共同冲击、样本选择、测量差异）

5. **我们能做哪些可信度检查？**（平衡性、趋势检查、不同时间窗口回归对比、敏感性分析等）

读完这三类案例，你应该形成一种稳定的判断顺序：看到一个“回归结果很显著”的因果结论，先不急着问它用了什么模型，而是先问——**这份设计凭什么能回答它提出的因果问题？如果不能，缺的那一块到底是什么？**

2.1 案例一：Salk 脊髓灰质炎疫苗现场试验（随机对照与观察性对照对比）

为什么这是“设计优先于模型”的经典案例 Salk 疫苗现场试验之所以常被统计学教材用作开篇案例，不是因为需要复杂模型，而是因为它在同一场公共卫生行动中并排出现了两种不同的研究设计：

- **安慰剂随机对照区（Randomized Controlled Trial, RCT）**：二年级儿童被随机分配到“疫苗/安慰剂”；
- **观察性对照区（Observed Control, OC）**：二年级自愿接种，一、三年级被当作“对照”（非随机）。

正如 Freedman et al.[9, Ch. 1] 强调的：*数据的含义首先由数据如何生成决定*。因此，两种设计即便做同样的统计计算，也对应不同的可解释性边界。

设计结构与关键数据 下面两张表给出两类研究区的“分组人数（暴露/处理规模）”与三类报告结果（瘫痪 Paralytic / 非瘫痪 NonParalytic / 误报 FalseReports），便于在同一尺度下对比设计差异。[1]

表 1: 安慰剂对照区（随机）：分组人数与报告病例（Paralytic / NonParalytic / FalseReports）。

分组	人数	瘫痪	非瘫痪	误报
疫苗（随机）	200,745	33	24	25
安慰剂（随机）	201,229	115	27	20
未接种（同区其他儿童）	338,778	121	36	25
未完成接种	8,484	1	1	0
合计	749,236	270	88	70

表 2: 观察性对照区（非随机）：分组人数与报告病例（Paralytic / NonParalytic / FalseReports）。

分组	人数	瘫痪	非瘫痪	误报
疫苗（二年级自愿接种）	221,998	38	18	20
对照（一、三年级）	725,173	330	61	48
二年级未接种	123,605	43	11	12
未完成接种	9,904	4	0	0
合计	1,080,680	415	90	80

[1] 许多汇总表的“报告病例数”是三类报告之和，并不等同于瘫痪病例。

同一统计操作，在两种设计下有不同含义 先用最简单的统计操作：比较**瘫痪发生率**（每 10 万人中的瘫痪病例数）：

- **RCT（可直接作因果解释）**：疫苗组 33/200,745 ≈ 16.4，安慰剂组 115/201,229 ≈ 57.1（每 10 万）。随机化提供可比性，因此“疫苗 vs 安慰剂”的差异可解释为疫苗的因果效应。
- **OC(不能直接作因果解释)**:二年级接种者 38/221,998 ≈ 17.1,一、三年级对照 330/725,173 ≈ 45.5（每 10 万）。这里的差异同时混入年龄/年级差异、自愿接种的选择偏差、以及地区暴露强度等因素；除非额外提出强假设，否则不能把它当作“接种的因果效应”。

设计先于模型 在 RCT 中，设计（随机化）把“可比性”写进数据生成机制，所以简单率差就有明确因果含义；在 OC 中，对照来自年级划分与自愿选择，即便使用回归或分层，也是在弥补设计缺口，并且结论依赖不可检验的假设。换言之[9]Ch. 1：

> 统计计算可以相同，但因果含义由设计（数据机制）决定。

2.2 案例二：新泽西最低工资研究——政策冲击下的自然实验

1992 年 4 月 1 日，新泽西州（NJ）将最低工资从每小时 $4.25 提高至 $5.05，而相邻宾夕法尼亚州（PA）最低工资保持不变。[4] 研究问题为：最低工资上调是否会导致低工资行业（如快餐业）减少雇用？

自然实验设计的核心逻辑 此研究的自然实验性质来源于政策冲击的外生性：政策仅在 NJ 生效，PA 保持不变；研究样本主要来自两州交界附近的快餐店。对单个店铺而言，其所在州在短期内是既定事实，因而政策变化可被视作外部冲击，而非店铺自主选择的待遇安排。这确保了对照组的可信性——即 PA 的快餐店可以作为 NJ 在“未实施政策”情况下的潜在对照。

比较策略：前后对照与差中差（Difference-in-Differences, DiD） 作者在政策实施前后分别调查快餐店的用工规模（全职当量，Full-Time Equivalent, FTE），形成两次横截面数据。分析步骤如下：

1. 计算 NJ 和 PA 各自政策前后的 FTE 平均值；
2. 取 NJ 与 PA 的变化量之差，得到差中差估计：

$$\mathrm{DiD} = (\bar{Y}_{\text{NJ, after}} - \bar{Y}_{\text{NJ, before}}) - (\bar{Y}_{\text{PA, after}} - \bar{Y}_{\text{PA, before}}).$$

表 3: 最低工资上调前后：NJ 与 PA 快餐店平均就业（FTE/店）及变化量。[4]

	PA	NJ	NJ−PA
政策前平均就业	23.33	20.44	−2.89
政策后平均就业	21.17	21.03	−0.14
变化量（后 − 前）	−2.28	+0.47	+2.75

表 3显示：PA 的就业下降反映共同趋势，而 NJ 相对 PA 的“额外变化”（DiD）为正，暗示最低工资上调并未减少 NJ 的快餐店就业。

设计优先于模型 差中差的核心假设是平行趋势（parallel trends）：若无政策变化，NJ 与 PA 快餐店就业趋势应大体相同。这一假设通过地理邻近、行业相似和前后对比得到支撑，而非依赖回归模型自动保证。研究者首先明确“对照为什么可信”，然后才运用统计方法计算政策效应。这个案例直观地说明了本章核心观点：**先把“比较为什么公平”讲清楚，再谈模型如何计算效应**。

现实提醒：数据与可比性的重要性 自然实验的有效性依赖于对照选择及数据质量。后续研究显示，不同数据源可能导致结果差异，并讨论了测量误差与样本代表性问题[5,13]。由此可见，透明的比较口径和合理的对照设计是因果解释的前提，而复杂回归模型无法替代良好的研究设计。

2.3 案例三：Yule 与贫困：观察性研究里的混杂与回归的边界

Yule（1899）研究英格兰的一个政策争论：*地方政府更偏向“院外救济”，会不会反而让贫困者变多？*这项研究常被看作回归在社会科学中的早期应用之一，但也正好展示了观察性研究最难的地方：**你没有抽签分组，很多事情会搅在一起**。

当时英国救济大致有两种方式。**院内救济**：进救济院（poor-house）；**院外救济**：不进院，在院外给补助（out-relief）。不同地区的当局会更偏向其中一种做法。Yule 想问的是：*这种偏好会不会“制造贫困”——让贫困率上升？*

Yule 怎么做：用回归来表达“其他条件不变” Yule 收集了多个行政区在不同十年（1871–1881、1881–1891）的数据，用各变量在十年间的**百分比变化**来做比较，并提出回归方程[8,26]：

$$\Delta\text{Paup} = a + b\,\Delta\text{Out} + c\,\Delta\text{Old} + d\,\Delta\text{Pop} + \text{error}.$$

这里 Δ 表示“百分比变化（percentage change over time）”；Paup 是贫困者比例（percentage of paupers），Out 是院外救济比 N/D（N 为院外救济人数，D 为救济院内人数），Old 是 65 岁以上人口所占比例，Pop 是人口规模。Yule 的直觉是：*在扣除人口增长与老龄化的变化后，如果 $b > 0$，就更像是“院外救济倾向”推动了贫困上升。*

先看看原始数据长什么样。在做任何回归之前，可以先看一眼同一时期（1881 相对 1871）的原始比值数据（表 4）。它展示了不同地区在贫困率、院外救济比、老年比例与人口规模上的变化幅度——你会直观地看到：各区差异很大，而且这些变量可能纠缠在一起（例如人口变化、贫困变化、救济结构变化往往同时发生）。

Yule 得到了什么：系数显著，但前后不一致. Yule 在两个十年窗口对都市联合区（metropolitan unions）的拟合结果：

$$\Delta\text{Paup} = 13.19 + 0.755\,\Delta\text{Out} - 0.022\,\Delta\text{Old} - 0.322\,\Delta\text{Pop} + \text{error},$$

$$\Delta\text{Paup} = 1.36 + 0.324\,\Delta\text{Out} + 1.37\,\Delta\text{Old} - 0.369\,\Delta\text{Pop} + \text{error}.$$

这里 Δ 表示“十年间的百分比变化（percentage change over time）”。

可以看到，ΔOut 的系数在两个时期都为正，Yule 因而倾向于把它解读为“院外救济导致贫困上升”。但同一个变量在不同时期的系数大小差异明显；ΔOld 的符号甚至从负变正。这至少提醒我们：这些系数可能更多是对特定时期数据的拟合摘要，而不一定是稳定的“政策规律”。[8]

为什么回归在这里不够：你缺的不是“控制变量”，而是“公平对照”。Yule 的回归最多说明**条件相关**——在控制若干变量后，贫困变化与院外救济变化正相关。[8] 但这离“院外救济导致贫困”还差很远，因为在现实里，政策和贫困往往是一起变化的，背后还有很多看不见的力量：

- **漏掉的重要因素：**地区经济结构、失业冲击、迁移与城市化、地方治理水平等，既可能影响贫困，也可能影响政策选择；
- **反过来影响：**贫困上升可能促使当局扩大院外救济（政策在回应贫困，而不是制造贫困）；
- **统计口径变化：**登记制度、统计方法、救济标准变化，可能让数字“看起来变了”；
- **地区互相影响：**一个地方的政策会把人“吸过来”或“推过去”。邻区政策一变，贫困者可能搬家、流动、转去别处申请救济，于是你看到的“本区贫困”和“本区救济”不再只是本区政策的结果，而是周边政策一起作用的结果。

表 4: 贫困率、院外救济比、老年人口比例与人口规模（1881 相对 1871 的比值 ×100）。英格兰都市联合区（Metropolitan Unions）。Yule (1899, Table XIX)

地区（Union）	贫困率（Paup）	院外救济比（Out）	老年比例（Old）	人口（Pop）
Kensington	27	5	104	136
Paddington	47	12	115	111
Fulham	31	21	85	174
Chelsea	64	21	81	124
St. George’s	46	18	113	96
Westminster	52	27	105	91
Marylebone	81	36	100	97
St. John, Hampstead	61	39	103	141
St. Pancras	61	35	101	107
Islington	59	35	101	132
Hackney	33	22	91	150
St. Giles’	76	30	103	85
Strand	64	27	97	81
Holborn	79	33	95	93
City	79	64	113	68
Shoreditch	52	21	108	100
Bethnal Green	46	19	102	106
Whitechapel	35	6	93	93
St. George’s East	37	6	98	98
Stepney	34	10	87	101
Mile End	43	15	102	113
Poplar	37	20	102	135
St. Saviour’s	52	22	100	111
St. Olave’s	57	32	102	110
Lambeth	57	38	99	122
Wandsworth	23	18	91	168
Camberwell	30	14	83	168
Greenwich	55	37	94	131
Lewisham	41	24	100	142
Woolwich	76	20	119	110
Croydon	38	29	101	142
West Ham	38	49	86	203

一个特别直观的陷阱：把“治理水平”当作看不见的背景噪声 在 Yule 这种观察性数据里，有一个很容易被忽略、但特别致命的混杂来源：治理能力（或行政效率）本身。更会治理的地区，可能一方面更能压低贫困（比如就业更稳定、救助更有序、登记更规范），另一方面也更可能采用不同的救济结构（未必依赖院外救济，或者同样的救济被更有效地管理）。于是你在数据里看到的“院外救济多 → 贫困高”，很可能并不是“院外救济造成贫困”，而是“治理能力差”的地

区两边都更糟：贫困更高，救济结构也更混乱。关键麻烦在于：治理能力这种东西往往很难被准确量化，就算你在回归里塞进几个“看起来相关”的控制变量，也未必能把它真正扣掉。于是模型可以把相关算得很精细，但并不能保证你得到的是因果。

3 因果语言

为什么需要“因果语言”：从随机实验到观测研究 在理想情况下，回答因果问题的金标准是随机对照实验（RCT）：研究者用随机分配制造可信对照，从而把“做与不做”的比较变成一个有明确因果含义的量。但现实里，随机化常常不可行——出于伦理、成本、制度约束或实施难度，我们更常面对的是观测数据：处理并非抽签产生，而是由选择、制度与环境共同决定。因此，我们需要一套**因果语言**，把“我们要回答什么”与“我们凭什么能回答”说清楚。

本章的因果语言有两条互补的线索。第一条是**潜在结果（potential outcomes）**：它直接给出因果问题的对象（estimand）；第二条是**图与结构语言**（路径模型、DAG、联立方程、工具变量、选择机制），它把“为什么能/为什么不能”的识别依据表达为结构，并帮助我们追踪偏差从哪里进入。

3.1 潜在结果：把“因果效应”定义成一个可讨论的对象

设处理 $T \in \{0,1\}$，对每个单位 i 定义两种潜在结果 $Y_i(1), Y_i(0)$，分别表示“接受处理”和“未接受处理”时的结局。个体层面的因果效应是 $\tau_i := Y_i(1) - Y_i(0)$，但我们永远无法同时观测二者；现实中只能看到

$$Y_i = T_i Y_i(1) + (1 - T_i) Y_i(0).$$

这就是反事实难题，也是“因果”与“相关”分家的根源。

因此，实证研究通常关心的是总体层面的**因果量（estimand）**，例如

$$\mathrm{ATE} := \mathbb{E}[Y(1) - Y(0)]$$

以及更细粒度的异质性效应

$$\mathrm{CATE}(x) := \mathbb{E}[Y(1) - Y(0) \mid X = x].$$

3.1.1 随机实验与潜在结果：为什么它是金标准

在随机对照实验中，处理 T 由研究者随机分配。用潜在结果语言，这对应一个极其强的性质：

$$T \perp \big(Y(1), Y(0)\big),$$

即处理与两种潜在结果独立（至少在期望意义上成立）。它意味着处理组与对照组在潜在结果意义下是可比的，于是

$$\mathbb{E}[Y \mid T = 1] - \mathbb{E}[Y \mid T = 0] = \mathbb{E}[Y(1)] - \mathbb{E}[Y(0)] = \mathrm{ATE}.$$

换句话说，**随机化把识别问题提前解决了：组间差就是因果效应**。模型在这里的角色主要是提高计算规范性与推断效率，而不是赋予因果含义。

3.1.2 从随机实验走向观测研究：用假设与结构替代随机化

在观测数据中，处理变量 T 往往不是随机分配的。例如，更健康的个体可能更倾向接受筛查，更富裕的个体可能获得更优治疗，企业是否采用某项管理制度可能与治理能力相关，地区是否实施政策可能与当地产业结构相关。形式上，通常存在

$$T \not\perp (Y(1), Y(0)),$$

即处理的分配与潜在结果相关。在这种情况下，朴素比较

$$\mathbb{E}[Y \mid T = 1] - \mathbb{E}[Y \mid T = 0]$$

不仅反映处理效应，还混入系统性差异，因此不等于因果效应。

为了在观测数据中识别因果效应，需要明确提出某种**替代随机化的条件**。常见做法是引入协变量 X 并假设**无混杂（unconfoundedness）**：

$$(Y(1), Y(0)) \perp T \mid X,$$

即在给定协变量 X 后，处理分配相当于随机。同时，还需满足**正性（positivity）**条件：

$$0 < P(T = 1 \mid X) < 1 \quad \text{对所有可能的} X \text{成立},$$

保证每个协变量组合下既有处理组，也有对照组可供比较。

在满足无混杂与正性条件的前提下，可以利用观测数据中的条件比较来识别平均处理效应（ATE）或条件处理效应（CATE）。具体地，由无混杂性可得：

$$\mathbb{E}[Y(t) \mid X] = \mathbb{E}[Y \mid T = t, X], \quad t = 0, 1,$$

于是条件平均处理效应（CATE）为

$$\tau(X) := \mathbb{E}[Y(1) - Y(0) \mid X] = \mathbb{E}[Y \mid T = 1, X] - \mathbb{E}[Y \mid T = 0, X].$$

进一步利用迭代期望（law of iterated expectations），整体平均处理效应（ATE）可以写作：

$$\text{ATE} = \mathbb{E}[Y(1) - Y(0)] = \mathbb{E}_X[\tau(X)] = \mathbb{E}_X\Big[\mathbb{E}[Y \mid T = 1, X] - \mathbb{E}[Y \mid T = 0, X]\Big].$$

这一公式说明，只要无混杂与正性条件成立，就可以通过对协变量条件比较的加权平均，在观测数据中识别因果效应。

前文公式显示，只要无混杂与正性条件成立，就能从观测数据识别因果效应。但在实际研究中，这些条件并非回归分析自动保证：它们可能依赖于研究设计（如抽样、匹配、随访、测量）；也可能依赖于对机制的理解（哪些变量是混杂、哪些是碰撞点、哪些是处理后变量）；或者依赖外生变化来源（工具变量、制度阈值、自然冲击）。因此，引入图与结构化语言，可以将这些潜在偏差机制（混杂、同时性、选择等）显式表示，使它们不再是模糊背景噪声，而是可分析、可推导、可验证的识别前提。这为后续章节中因果识别与结构化建模提供了背景与动机。

3.2 路径模型：把回归系统画成图，以及它的限度

路径模型起源于 Wright 在遗传学中的工作。[22-23] 它的核心是：用一张有向图来表达一组回归方程（或结构方程），并用“扰动项”表示未显式纳入模型的因素。图中通常约定箭头从解释变量指向被解释变量，每条箭头对应一个回归系数（*path coefficient*）。

从图到方程：路径图是一组结构化回归 典型路径图对应结构方程系统，例如

$$Y = \beta_X X + \beta_V V + \delta_Y, \qquad W = \gamma_X X + \gamma_V V + \delta_W,$$

其中 δ_Y, δ_W 表示未被显式建模的影响（扰动、残差、遗漏因素的综合）。在形式上，这样的写法暗示了一种“生成顺序”，即假设某些变量按先后顺序影响其他变量。然而在观察性数据中，同一组协方差结构往往可以由多种互相矛盾的箭头方向解释。因此，更稳妥的理解是：路径图本质上是一种**结构化假设的表达**，而非数据自动生成的因果事实。它的作用在于把研究者对变量之间依赖关系的信念明确化，并检验这些假设是否与观测到的相关性或协方差模式相容[8]。换言之，路径模型提供了一种机制假设的可视化与定量表示，但其因果解释依赖于假设的合理性，而非仅凭数据本身。

3.2.1 标准化与路径系数：为何常从相关矩阵出发

社会科学路径分析常把变量标准化（均值 0、方差 1），以便将系数解释为“标准化效应”。在标准化条件下，路径系数可直接由相关矩阵与回归公式给出。比如

$$Y = aX + bV + \delta$$

若 X, V, Y 均已标准化，则 (a, b) 是对 Y 的多元回归系数，而 $\mathrm{Var}(\delta)$ 则由回归的残差方差公式得到。这使路径分析在操作上常把“样本相关矩阵 + 若干线性关系假设”作为输入，强调对协方差结构的拟合，而不是逐条方程的预测性能。

3.2.2 示例：Blau–Duncan 地位获得模型（路径图与识别边界）

Blau–Duncan (1967) 的地位获得（status attainment）研究，是社会学中经典的路径模型案例：用父亲教育、父亲职业等变量解释儿子教育、第一份工作与成年职业地位，并将这些关系绘成一张“从出身到成就”的箭头链。[1,8]

这种表示法之所以看起来像因果机制，关键在于它把社会学叙事拆解为若干直接路径，每条路径对应一条回归系数：父母教育和职业影响子女教育，教育影响第一份工作，第一份工作影响成年地位。读者容易将箭头理解为“作用方向”，将路径系数理解为“效应大小”。

箭头的因果与非因果解读

- **父亲教育 → 儿子教育**：若无遗漏变量且控制其他混杂，箭头可解读为教育的因果效应；否则仅表示条件相关性。
- **父亲职业 → 儿子教育**：理论假定直接影响，但在观测数据中可能混入父亲教育、文化资本等未测量因素。

- **儿子教育 → 第一份工作**：教育提升技能，可影响职业起点；同时可能反映家庭背景和社会网络的选择效应。
- **第一份工作 → 成年职业地位**：工作经验、职位积累理论上影响后续职业；在观测数据中也可能受教育与未观察因素影响。
- **儿子教育 → 成年职业地位**：直接和间接效应叠加，但可能同时受隐藏变量干扰。
- **父亲教育/职业 → 第一份工作**：理论上的直接路径，但在观察性数据中可能混入背景因素。

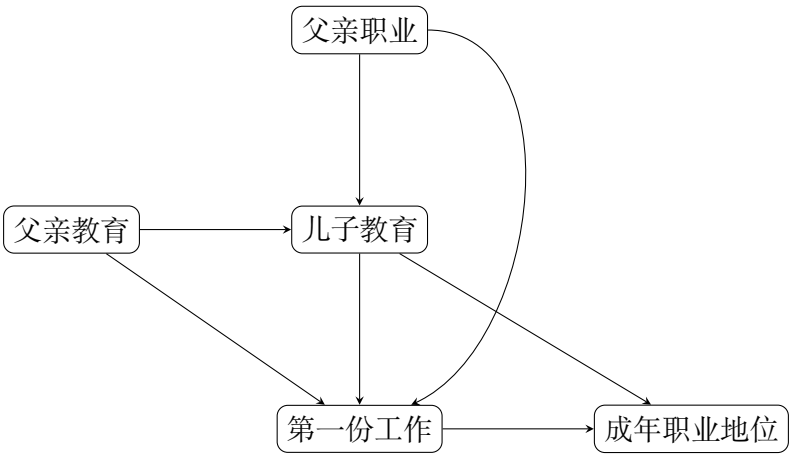

图 1: Blau–Duncan 地位获得模型路径图（示意）

识别限制与路径系数解释 在观察性数据中，路径系数仅拟合协方差结构，并不自动承载因果解释。要让系数具有因果意义，需要额外的*识别依据*：

- 说明处理如何产生、对照如何形成、哪些混杂被排除；
- 依赖研究设计（如随机化）、外生冲击（自然实验、工具变量）或可信的结构假设；
- 辅以假设诊断与稳健性检验。

因此，更合理的解读流程是：先给出可辩护的识别故事（设计与假设）⇒ 再用路径模型/回归作为计算与总结工具，而不是“画箭头 + 跑回归 ⇒ 自动因果”。这也说明了 Blau–Duncan 路径图的主要价值在于**明确理论结构与经验相关的分解方式**，而非直接提供干预效应证据。[8]

3.2.3 示例：Blau–Duncan 的地位获得模型（为何看起来像因果）

Blau–Duncan 的地位获得 (status attainment) 研究常被当作路径模型的代表：用父亲教育、父亲职业等变量解释儿子教育、第一份工作与成年职业地位，并把这些关系画成一张“从出身到成就”的箭头链。[1,8]

这种表示法之所以*看起来*很像因果机制，关键在于它把一个社会学叙事（例如“家庭背景影响教育，教育影响第一份工作，第一份工作影响成年地位”）拆成若干条“直接路径”，并把每条路径对应到一条回归系数。读者很自然会把“有箭头”理解成“有作用”，把“路径系数”理解成“效应大小”。

但在观察性数据里，这种直觉往往站不住脚：同一条箭头（例如父亲职业 → 儿子成年地位）的系数，既可能反映某种干预意义上的作用，也可能只是共同原因造成的相关。举例说，家庭的文化资本、居住地机会结构、父母的社会网络、学校质量等因素，即使没有被测量进数据，也可能同时影响儿子的教育与职业轨迹；它们会以“遗漏变量”的形式进入多条方程的扰动项，并在多条路径上制造系统性相关。在这种情况下，路径图仍然可以*很好地拟合协方差结构*，但拟合得好并不意味着箭头方向被识别，也不意味着系数可解释为干预效应。[8] 因此，Blau–Duncan 式路径图更可靠的价值，是把理论结构与经验相关的分解方式写清楚，而不是把它当成因果机制的证据本身。

3.2.4 从路径到相关：协方差结构与可检验性（路径模型在统计上“检验”的是什么）

路径模型不仅是一张图，它还隐含了一组对协方差结构的限制：*一旦你指定了哪些箭头存在、各条路径系数是多少，以及扰动项之间是否允许相关，模型就会给出每对变量的协方差/相关应当是什么样。*[8]

直观规则可以这样记：

- 若 $X \to Y$，且 $Y \to Z$，那么 X 对 Z 的关联会通过 Y 传递；
- 传递强度沿路径相乘：$X \to Y$ 的系数乘上 $Y \to Z$ 的系数，贡献到 $\mathrm{Cov}(X, Z)$（或相关）里；
- 若存在多条平行路径（例如 $X \to Y \to Z$ 与 $X \to W \to Z$），各条路径的贡献相加；
- 若某些扰动项相关（例如 $\mathrm{Cov}(\delta_Y, \delta_Z) \neq 0$），这会额外在相应变量之间“直接注入”相关。

看一个最小的例子：设 $Y = aX + \delta_Y, Z = bY + \delta_Z$，并假定 X 与扰动独立、且 $\mathrm{Cov}(\delta_Y, \delta_Z) = 0$。则

$$\mathrm{Cov}(X, Z) = \mathrm{Cov}\big(X,\, bY + \delta_Z\big) = b\,\mathrm{Cov}(X, Y) = b\,\mathrm{Cov}\big(X,\, aX + \delta_Y\big) = ab\,\mathrm{Var}(X).$$

如果再标准化使 $\mathrm{Var}(X) = 1$，那么相关就变成 $(X, Z) = ab$：这正是“沿路径系数相乘”的含义。若同时存在另一条路径或扰动相关，上式会再多出相应的加项。

因此，路径模型所谓“可检验”，主要检验的是：*这些协方差约束是否捕捉了数据中的相关结构*（例如模型隐含某些零协方差/零相关或特定的相关分解，但样本相关矩阵明显违背它）。需要强调的是：即使这些约束与数据相容，得到的也仍然是“这套线性协方差结构假设没有被数据拒绝”，而不是“箭头被证明为因果”。[8]

3.2.5 为何容易被误读为因果：路径模型的脆弱前提

路径图把回归系统“机制化”之后，很容易诱导读者把回归系数当作因果效应。[8] 在观察性研究中，下列问题会系统性破坏这种解释：

- **遗漏变量与共同原因**：未观测因素同时影响多条路径，回归系数偏误；
- **反向因果与同时性**：变量互相影响时，单向箭头未必对应可识别的生成机制；
- **测量误差**：解释变量的测量误差会扭曲系数与路径分解；
- **函数形式错设**：线性/加性错设时，路径分解的“效应叙事”会误导解释。

因此，路径模型更可靠的角色，是作为假设的记号系统：它把研究者关于变量间依赖关系的主张（哪些边存在、哪些边不存在、扰动项是否相关）压缩成一组结构方程，并据此导出一组对相关/协方差结构的限制[8]。在这个层面上，路径模型能做的主要是两件事：第一，把“理论故事”写成可讨论、可批评的形式；第二，检查该故事是否至少不与数据的协方差结构明显冲突（例如某些隐含的零相关/条件独立是否被数据否定）。

但仅凭这些并不能推出“箭头就是因果”。原因在于：在观察性数据中，同一套协方差结构往往可以由不同的因果机制产生，而遗漏变量、反向因果、共同扰动等都可能在不改变拟合优度的情况下改变箭头的因果解释。换言之，路径模型在统计上拟合得好，最多说明“这组线性协方差约束与数据相容”，却不足以证明“系数等于干预效应”。

要让路径系数承载因果含义，必须在模型之外再补上识别依据：说明处理如何产生、对照如何形成、哪些混杂被排除，以及在什么假设下回归/方程系数可被解释为某个因果 estimand。这类依据通常来自研究设计与外部信息——例如随机分配带来的处理不依赖潜在结果，自然实验提供的近似随机变异，或工具变量提供的外生冲击与排除限制——并辅以对关键假设的诊断与稳健性检验。[8] 因此，更合理的写法不是“画出箭头 + 跑回归 ⇒ 因果证明”，而是“先给出可辩护的识别故事（设计与假设）⇒ 再用路径模型/回归作为计算与总结工具”。

3.3 DAG 视角：控制变量何时有效、何时有害

DAG（directed acyclic graph，定向无环图）通过图形的方式将因果机制呈现出来，其中每一条箭头代表一个可能的直接因果作用，图中也隐含着可推导出的条件独立关系。[14] 与路径模型相比，DAG 的核心优势不仅仅是“更加像因果机制”，而是它将一个常常被忽视的结构性问题推向了显著位置：你控制了哪些变量，实际上就等同于你在做哪一种条件比较，而这种比较是否能识别你想要的干预效应，取决于结构的设定。

这种结构化的因果推断方法，极大地帮助我们理解哪些变量的控制是有效的，哪些控制反而会引入偏差。DAG 突破了传统回归分析中依赖假设的局限，通过图形化的方式明确了变量之间的因果关系和相关性路径。

DAG 的背景与应用 DAG 在因果推断中扮演着至关重要的角色，它不仅可以帮助我们明确哪些变量是混杂变量(confounders)，哪些是碰撞点(colliders)，哪些是处理后变量(post-treatment)，还能够揭示在何种条件下可以合理控制哪些变量来实现准确的因果推断。

例如，DAG 在流行病学研究中被广泛应用，尤其是在公共卫生和社会科学领域，研究人员可以使用 DAG 来理清暴露因素、治疗、结果之间的因果关系，并帮助确定在分析数据时，哪些变量需要控制，哪些应该避免控制。它还被用于经济学、教育学和心理学等领域，特别是那些涉及多个潜在影响因素和复杂因果关系的主题时，DAG 能够清晰地展示变量之间的因果链条。

三类关键结构：混杂、碰撞点与处理后变量 在因果推断中，DAG 使我们能够清晰地区分不同类型的变量结构。以下是三种常见的且容易导致推断偏差的结构：

- **混杂变量（confounder）**：$T \leftarrow C \rightarrow Y$。混杂变量指的是那些同时影响处理 T 和结果 Y 的变量。若我们不控制混杂变量 C，那么即使 T 与 Y 之间有关系，这种关系也可能是由 C 引起的，导致我们无法准确识别 T 对 Y 的因果效应。
- **碰撞点（collider）**：$T \rightarrow M \leftarrow Y$。在碰撞点结构中，$M$ 既受 T 影响，也受 Y 影响。如果我们对 M 进行条件化（控制 M），反而可能“打开”原本被阻断的关联通路，引入选择偏差，从而导致错误的因果推断。

- **处理后变量（post-treatment）**：例如 $T \to M \to Y$ 或 $T \to M$。处理后变量是指那些在处理 T 后，作为中介影响 Y 的变量。控制这些变量会改变我们估计的因果效应：总效应可能会被拆解为直接效应和间接效应，同时也可能引入额外的偏差，尤其是在存在未观测混杂的情况下。

这三类结构的共同含义是：控制变量并非越多越好，错误的控制可能会带来更大的问题。控制变量应当依据因果结构进行，而非随意添加。

后门准则与可交换性：一张“该控制谁”的清单 DAG 提供了一个非常实用的识别规则——后门准则（*backdoor criterion*），它帮助我们选择哪些变量应当控制，哪些变量不应当控制。具体而言，若存在一组协变量 S，使得

- S 不包含 T 的后代；
- 给定 S 后，所有从 T 到 Y 的后门路径（即不通过处理变量 T 的路径）都被阻断，

则在 S 条件下，处理 T 与潜在结果 Y 是可交换的，即调整 S 后，可以通过计算调整后的差异来识别因果效应。[14]

这一规则明确了：我们在做因果推断时，目标不是无差别地控制所有变量，而是要根据因果图识别出“需要控制”而不会引入偏差的变量集合。这也意味着控制变量的选择需要结构性地判断，而非依赖经验或直觉。

DAG 与回归：回归只是计算工具，识别在回归之外 回归分析本质上是通过条件比较来计算变量之间的关系。然而，正如 DAG 所提醒的，回归系数能否被解释为因果效应的估计，取决于是否存在满足识别条件的调整集 S。如果存在未控制的混杂变量，或者错误地调整了碰撞点或处理后变量，回归系数即使看起来很显著，也不能作为因果效应来解释。[8,14]

DAG 通过图形化表达，提前为我们指出哪些路径需要阻断，哪些变量需要控制，从而确保回归分析是针对真实的因果效应而不是仅仅是相关性的计算。

DAG 的局限性与其他方法的结合 虽然 DAG 提供了非常清晰的因果路径和控制变量的指南，但它本身并不是万能的。DAG 依赖于我们正确地构建因果模型，如果模型设定有误（例如遗漏了重要的混杂变量），DAG 的推断仍然可能是错误的。此外，DAG 也无法直接处理随机误差或模型不完全的问题。因此，DAG 更常见的应用是与其他因果推断方法相结合，例如随机实验、工具变量、差分法等，这些方法可以帮助验证 DAG 所推导的因果路径是否成立。

在实际研究中，研究人员需要结合设计和数据来检查 DAG 模型的有效性。例如，通过实验设计（如随机对照试验），我们可以验证 DAG 假设是否符合实际。通过外生冲击（如自然实验），可以进一步验证控制变量的选择是否恰当。

小结：从“把回归画成机制”到“用结构讨论识别” 路径模型擅长把理论结构写成方程与协方差约束，但它并不自动解决因果识别。DAG 则把识别问题前置，明确了哪些路径该阻断、哪些节点不该条件化、哪些变量集合能支撑可交换性。从路径模型走向 DAG，并不是从“相关”跳到“因果”的魔法，而是从“回归系数叙事”转向“设计与识别语言”的训练。理解 DAG 的真正价值在于它帮助我们系统地考虑因果推断中的各类问题，确保在数据分析中做出合理的因果解释。

3.4 联立方程：为什么会出现内生性

在许多经济学与社会科学问题中，研究者观测到的解释变量并不是“先给定、后影响结果”的外部输入，而是系统内部的**均衡结果**：它与结果变量在同一机制下被共同决定（joint determination）。当我们用单方程回归去刻画其中一条结构关系时，就很容易出现**内生性**——不是因为线性回归“不够复杂”，而是因为数据来自一个同时决定的系统。

外生性与内生性的回归定义 考虑线性回归（为简化省略截距）

$$Y = \beta X + \varepsilon.$$

若满足**外生性**（exogeneity）

$$\mathbb{E}(\varepsilon \mid X) = 0 \quad (\text{常用的弱形式为 } \mathrm{Cov}(X, \varepsilon) = 0),$$

则 OLS 在大样本下是一致的，常规标准误与显著性检验也有正确的解释。

内生性（endogeneity）指外生性不成立：

$$\mathbb{E}(\varepsilon \mid X) \neq 0 \quad (\text{或 } \mathrm{Cov}(X, \varepsilon) \neq 0).$$

在这种情形下，OLS 往往产生**偏误**或**不一致**：即便样本量很大，估计也会系统性偏离目标结构参数。内生性的典型来源包括同时性（simultaneity）、遗漏变量（omitted variables）、反向因果（reverse causality）与测量误差等；本节聚焦其中最原型的一类：**联立决定导致的同时性**。

3.4.1 同时性（simultaneity）偏差：供需模型作为原型

设 Q 为交易量、P 为价格。用两条**结构方程**描述需求与供给：

$$\text{需求：} \quad Q = \alpha - \beta P + u, \qquad \text{供给：} \quad Q = \gamma + \delta P - v,$$

其中 $\beta > 0, \delta > 0$；u 是需求侧冲击（偏好上升等，使给定价格下需求增加），v 是供给侧冲击（成本上升等，使给定价格下供给减少）。市场中观测到的 (P, Q) 来自**均衡**：同一时刻必须满足“需求量 = 供给量”。

均衡解：P, Q 是系统的内生结果 均衡条件给出

$$\alpha - \beta P + u \;=\; \gamma + \delta P - v,$$

从而

$$P \;=\; \frac{\alpha - \gamma}{\beta + \delta} + \frac{u + v}{\beta + \delta}, \qquad Q \;=\; \frac{\alpha\delta + \beta\gamma}{\beta + \delta} + \frac{\delta u - \beta v}{\beta + \delta}.$$

这里要强调的不是“价格不受外生因素影响”，恰恰相反：**外生冲击 u, v 是输入，而 P, Q 是由均衡机制在系统内部解出来的输出**。因此价格会随当期的需求与供给冲击共同变化。

比较静态：为什么两侧冲击都会推高价格 由均衡价格式直接得到

$$\frac{\partial P}{\partial u} = \frac{1}{\beta + \delta} > 0, \qquad \frac{\partial P}{\partial v} = \frac{1}{\beta + \delta} > 0.$$

含义是：u 上升使需求曲线右移，在原价格处出现超额需求，价格上升直到再次均衡；v 上升使供给曲线左移，在原价格处出现供给不足，价格同样上升直到再次均衡。换言之，**在均衡数据中，价格的波动同时夹杂了需求与供给两种来源**。

同时性偏差：为何单方程 OLS 无法识别结构斜率 如果我们把观测到的均衡样本直接拿来估“需求斜率”，做回归

$$Q = a + bP + e,$$

希望 $b \approx -\beta$，问题在于：在真实需求结构式中误差项就是 u，但均衡解告诉我们

$$P = \frac{\alpha - \gamma}{\beta + \delta} + \frac{u + v}{\beta + \delta},$$

因此 P 中包含 u 的成分，通常

$$\mathrm{Cov}(P, u) \neq 0.$$

同理，在供给式中 P 也与 v 相关。于是单方程回归的解释变量 P 与误差项相关，外生性条件被破坏，OLS 斜率就不再对应某一条结构曲线的斜率。你在数据里看到的是：*均衡点随着 u, v 的随机波动在平面上移动形成的一团云*，OLS 拟合的是这团云的相关斜率——它由需求斜率、供给斜率以及两类冲击的相对波动共同决定，而不是纯粹的 $-\beta$ 或 δ。这正是联立方程环境下最经典的同时性（内生性）问题。

3.4.2 识别（identification）：参数是否能由数据唯一决定

识别讨论的是：在给定一组模型假设下，目标参数是否能由*总体分布*唯一确定。更形式化地说，若存在两组不同的结构参数（连同误差分布）能生成完全相同的观测数据分布，则这些参数在该信息集合下**不可识别**（not identified）。不可识别不是“估计方法不够聪明”，而是**数据所包含的信息不足以区分不同的结构解释**。

为什么仅靠均衡点的相关性通常不够 在供需系统里，我们观测到的 (P, Q) 来自均衡解；均衡点会随着需求冲击 u 与供给冲击 v 的变化而移动，从而在平面上形成一团点云。关键在于：点云的相关斜率并不只由需求斜率 $-\beta$ 或供给斜率 δ 决定，还取决于两类冲击的相对波动大小及其在均衡中的传导方式。因此，仅凭 (P, Q) 的相关结构通常不足以把需求与供给的结构参数拆开。

额外识别信息的核心目标 要识别结构参数，需要额外信息来制造一种**单侧外生移动**：*让均衡点主要因为某一侧曲线的外生位移而变化，而另一侧保持不动（除了沿曲线的常规反应）*。这类信息最经典的形式是**排除限制**，从而引出工具变量。

排除限制与工具变量：用供给移位识别需求斜率

考虑带有供给移位变量 Z 的结构系统

$$Q^d = \alpha - \beta P + u, \qquad Q^s = \gamma + \delta P - v + \pi Z,$$

并令均衡满足 $Q^d = Q^s = Q$。观测数据为 (Q, P, Z) 的总体分布。

命题 3.1 (供给移位变量识别需求斜率)**.** *在上述模型下，若满足*

1. **排除/外生性（对需求误差）：** $\mathbb{E}[u \mid Z] = 0$（若变量中心化则等价于 $\mathrm{Cov}(Z, u) = 0$）；
2. **相关性（*relevance*）：** $\mathrm{Cov}(Z, P) \neq 0$，

则需求斜率参数 β 由观测分布 (Q, P, Z) **唯一确定**，*即 β 在该信息集合下可识别，并且具有识别式*

$$\beta \;=\; -\frac{\mathrm{Cov}(Z, Q)}{\mathrm{Cov}(Z, P)}.$$

证明. 由需求结构方程 $Q = \alpha - \beta P + u$ 得

$$u = Q - \alpha + \beta P.$$

对两边取与 Z 的协方差：

$$\mathrm{Cov}(Z, u) = \mathrm{Cov}(Z, Q - \alpha + \beta P) = \mathrm{Cov}(Z, Q) + \beta \mathrm{Cov}(Z, P),$$

其中 $\mathrm{Cov}(Z, \alpha) = 0$ 因为 α 为常数。由外生性条件 $\mathbb{E}[u \mid Z] = 0$ 可得 $\mathrm{Cov}(Z, u) = 0$，于是

$$0 = \mathrm{Cov}(Z, Q) + \beta \mathrm{Cov}(Z, P).$$

若再由相关性条件 $\mathrm{Cov}(Z, P) \neq 0$，即可唯一解出

$$\beta = -\frac{\mathrm{Cov}(Z, Q)}{\mathrm{Cov}(Z, P)}.$$

右侧完全由观测变量 (Q, P, Z) 的总体分布决定，因此 β 被该分布唯一确定，即 β 可识别。 □

几点说明

- 命题中外生性条件是相对于需求方程的误差 u 而言的；它并不要求 $\mathrm{Cov}(Z, v) = 0$。直觉上，Z 的作用正是提供供给侧位移，使价格发生外生变化；识别需求所需的是这种变化不要与需求侧未观测冲击 u 同源。
- 对称地，若存在只影响需求而不直接影响供给的移位变量 Z_d，并满足 $\mathbb{E}[v \mid Z_d] = 0$ 与 $\mathrm{Cov}(Z_d, P) \neq 0$，则可识别供给斜率 δ。

3.5 工具变量（IV）与两阶段最小二乘（2SLS）

本节讨论一类经典的“用设计补模型”的方法：**工具变量**（instrumental variables, IV）。它解决的问题很明确：解释变量 X 与误差项相关（内生性），导致 OLS 不能解释为 $X \to Y$ 的因

果效应。IV 的核心思想是：不再使用 X 的全部变动，而只使用其中可被视为外生的一部分变动来识别因果效应。

3.5.1 问题：为什么 OLS 会错

考虑结构方程（可含控制变量 W；为简化先写无截距形式）

$$Y = \beta X + \varepsilon, \qquad \mathrm{Cov}(X, \varepsilon) \neq 0.$$

这里 $\mathrm{Cov}(X, \varepsilon) \neq 0$ 可能来自同时性、遗漏变量、反向因果或测量误差等。此时 OLS 估计的斜率一般不是 β，因为回归把 X 的“内生成分”也当成了可用于识别的变化来源。

3.5.2 工具变量：把 X 分解为“外生部分 + 内生部分”

工具变量 Z 的作用可以用一句话概括：

Z 提供一种外部推动，使 X 发生变化；而这推动与 ε 无关，并且不直接影响 Y。

形式化地，设我们有外生控制变量 W，则 Z 需要满足两条识别条件（都是相对于结构误差 ε 而言的）：

(1) 相关性（relevance） 在控制 W 后，Z 必须能解释 X：

$$\mathrm{Cov}(Z, X \mid W) \neq 0.$$

否则 Z 推不动 X，就无法提供关于 β 的信息。

(2) 外生性与排除限制（exogeneity & exclusion） Z 必须与结构误差无关，并且对 Y 的影响只能通过 X 发生。最常用的表述是矩条件

$$\mathbb{E}[Z\,\varepsilon \mid W] = 0 \quad (\text{等价地，中心化后} \mathrm{Cov}(Z, \varepsilon \mid W) = 0).$$

“只能通过 X”包含两层含义：没有直接效应（$Z \not\to Y$ 直通路径），也不通过与 ε 同源的遗漏因素间接影响 Y。

识别公式（单工具、单内生变量的总体结论） 下面把“协方差比值”写成一个严格的识别命题，并给出证明。

命题 3.2 (单工具识别公式)**.** 设结构方程为

$$Y = \beta X + \varepsilon,$$

其中 X 可能内生（允许 $\mathrm{Cov}(X, \varepsilon) \neq 0$）。给定外生控制变量 W 与工具变量 Z，若满足：

1. **排除/外生性（*IV* 外生性）**：
$$\mathbb{E}[Z\,\varepsilon \mid W] = 0;$$

2. **相关性（*relevance*）**：

$$\mathrm{Cov}(Z, X \mid W) \neq 0;$$

则结构参数 β 在该信息集合下可识别，并由观测分布唯一确定为

$$\beta = \frac{\mathrm{Cov}(Z, Y \mid W)}{\mathrm{Cov}(Z, X \mid W)}.$$

证明. 从结构方程出发，有恒等式

$$Y = \beta X + \varepsilon.$$

在给定 W 的条件下，两边同乘以 Z 并取条件期望：

$$\mathbb{E}[ZY \mid W] = \mathbb{E}[Z(\beta X + \varepsilon) \mid W] = \beta\,\mathbb{E}[ZX \mid W] + \mathbb{E}[Z\varepsilon \mid W].$$

由外生性假设 $\mathbb{E}[Z\varepsilon \mid W] = 0$，得到

$$\mathbb{E}[ZY \mid W] = \beta\,\mathbb{E}[ZX \mid W].$$

现在对上式两边减去 $\mathbb{E}[Z \mid W]\mathbb{E}[Y \mid W]$ 与 $\mathbb{E}[Z \mid W]\mathbb{E}[X \mid W]$，可得

$$\mathrm{Cov}(Z, Y \mid W) = \beta\,\mathrm{Cov}(Z, X \mid W).$$

（这是因为 $\mathrm{Cov}(Z, Y \mid W) = \mathbb{E}[ZY \mid W] - \mathbb{E}[Z \mid W]\mathbb{E}[Y \mid W]$，同理对 X 也成立。）

若进一步由相关性假设 $\mathrm{Cov}(Z, X \mid W) \neq 0$，则可在几乎处处（或在总体意义下）除以该量并唯一解出

$$\beta = \frac{\mathrm{Cov}(Z, Y \mid W)}{\mathrm{Cov}(Z, X \mid W)}.$$

右侧完全由观测变量 (Y, X, Z, W) 的总体分布决定，因此 β 被该分布唯一确定，即 β 可识别。 □

解释：命题说明 IV 的“识别信息”来自矩条件 $\mathbb{E}[Z\varepsilon \mid W] = 0$。它并不要求 X 外生；相反，它用 Z 抽取出 X 中与 ε 不同源的那部分变动，从而恢复结构效应 β。）

3.5.3 结构方程、第一阶段与约化式：IV 从哪里来

上一小节给出了识别公式

$$\beta = \frac{\mathrm{Cov}(Z, Y \mid W)}{\mathrm{Cov}(Z, X \mid W)},$$

它告诉我们：β 只利用“由 Z 诱发的 X 变动”来识别。把这一逻辑写成三条方程，会更直观。

结构方程（structural equation）：我们想要的因果关系

$$Y = \beta X + \Lambda W + \varepsilon.$$

这里 X 可能内生（$\mathrm{Cov}(X, \varepsilon \mid W) \neq 0$），OLS 不能直接解释为因果效应。

第一阶段（first stage）：Z 是否真的推动 X

$$X = \pi Z + \Gamma W + \eta.$$

π 衡量工具的强度。没有第一阶段（$\pi = 0$），就没有可用的外生变动来源。

约化式（reduced form）：Z 对 Y 的总影响 把上面两式合在一起（把 X 的表达代回结构方程），得到

$$Y = (\beta\pi)Z + (\beta\Gamma + \Lambda)W + (\beta\eta + \varepsilon).$$

因此在控制 W 后，Z 对 Y 的系数就是

$$\rho = \beta\pi, \qquad \Rightarrow \qquad \beta = \frac{\rho}{\pi}.$$

这就是 IV 的“来源”：**用 Z 对 Y 的总作用（约化式）除以 Z 对 X 的作用（第一阶段），得到 $X \to Y$ 的结构效应**。它与协方差比值形式完全等价。

3.5.4 两阶段最小二乘（2SLS）：计算方式与直觉

当存在控制变量（以及可能多个工具）时，最常用的 IV 实现是两阶段最小二乘（2SLS）。它的核心可以概括为：用工具变量张成的空间，从 X 中抽取“可被视为外生”的那部分变动，再用这部分变动识别 β。

设定与较弱的外生性（矩条件） 结构方程写为

$$Y = \beta X + \Lambda W + \varepsilon, \qquad \varepsilon := Y - \beta X - \Lambda W.$$

IV 的外生性在本节采用**较弱的矩条件**表述：

$$\mathbb{E}[Z\,\varepsilon \mid W] = 0, \tag{Exog}$$

并配合相关性（第一阶段不为零）$\mathrm{Cov}(Z, X \mid W) \neq 0$。

第一阶段：得到 X 的外生成分 用 (Z, W) 回归 X，得到拟合值

$$\hat{X} \;=\; (X \mid Z, W),$$

即 X 在由 (Z, W) 张成空间中的投影。直觉上，$\hat{X}$ 只保留了“能被工具与外生控制解释的那部分 X 变动”。

第二阶段：用 $\hat{X}$ 估计结构效应 用 $(\hat{X}, W)$ 回归 Y：

$$Y = \beta\hat{X} + \Lambda W + \text{残差},$$

得到 $\hat{\beta}_{\mathrm{2SLS}}$。

为什么“两阶段回归”会实现外生性矩条件？（把等价关系说透） 要把两者连起来，只需抓住一个事实：**第二阶段用的是 $\hat{X} = (X \mid Z, W)$，它本身是 (Z, W) 的线性组合**。于是“用 $(\hat{X}, W)$ 做 OLS”的一阶条件，会自动转化成“残差与 (Z, W) 正交”，也就对应了外生性矩条件。

第一步：把两阶段写成“投影 + 一次 OLS” 第一阶段得到

$$\hat{X} = (X \mid Z, W) = aZ + b'W$$

（在总体或样本中，投影都落在 $\mathrm{span}(Z, W)$ 这个线性空间里）。

第二阶段做 OLS：

$$(\hat{\beta}, \hat{\Lambda}) = \arg\min_{\beta, \Lambda} \mathbb{E}\big[(Y - \beta\hat{X} - \Lambda W)^2\big].$$

令第二阶段残差为

$$r := Y - \beta\hat{X} - \Lambda W.$$

第二步：第二阶段的 OLS 一阶条件 $\Rightarrow$ 残差与回归量正交 OLS 的一阶条件是（对 $\hat{X}$ 与 W 分别求导）：

$$\mathbb{E}[\hat{X}\, r] = 0, \qquad \mathbb{E}[W\, r] = 0. \tag{FOC}$$

这只是“最小二乘的正交性”：残差与回归量正交。

第三步：因为 $\hat{X}$ 在工具空间里，所以 (FOC) 等价于残差与工具正交关键：$\hat{X} \in \mathrm{span}(Z, W)$，因此存在常数 (a, b) 使 $\hat{X} = aZ + b'W$。把它代入 $\mathbb{E}[\hat{X}\, r] = 0$：

$$0 = \mathbb{E}[(aZ + b'W)r] = a\,\mathbb{E}[Zr] + b'\mathbb{E}[Wr].$$

而第二个一阶条件已经给出 $\mathbb{E}[Wr] = 0$，于是推出

$$\mathbb{E}[Z\, r] = 0.$$

结合 $r = Y - \beta\hat{X} - \Lambda W$，得到

$$\mathbb{E}\big[Z\,(Y - \beta\hat{X} - \Lambda W)\big] = 0. \tag{Z-orth}$$

这就是“第二阶段 OLS $\Rightarrow$ 残差与工具正交”的严格推导。

第四步：把 $\hat{X}$ 换回 X（为什么这对应 IV 的矩条件） 注意 (Z-orth) 中出现的是 $\hat{X}$。而 IV 的识别矩条件写成

$$\mathbb{E}[Z\,(Y - \beta X - \Lambda W)] = 0. \tag{IV-MC}$$

两者为什么一致？因为第一阶段的定义 $\hat{X} = (X \mid Z, W)$ 等价于

$$X = \hat{X} + u, \qquad \text{且} \quad \mathbb{E}[Z\, u] = 0,\ \mathbb{E}[W\, u] = 0,$$

即“投影残差 u 与工具空间正交”。于是

$$\mathbb{E}[Z\,(Y - \beta X - \Lambda W)] = \mathbb{E}[Z\,(Y - \beta\hat{X} - \Lambda W)] - \beta\,\mathbb{E}[Z\, u].$$

由 (Z-orth) 得第一项为 0，而由投影正交性得 $\mathbb{E}[Z\,u]=0$，因此

$$\mathbb{E}[Z\,(Y-\beta X-\Lambda W)]=0.$$

这就把“两阶段回归”严格地连接到了“工具正交的矩条件”。

一句话总结： *2SLS 之所以等价于矩条件，是因为它等价于先把 X 投影到工具空间，再做一次 OLS；而 OLS 的一阶条件必然让残差与该空间正交，从而（连同投影残差的正交性）推出 Z 与结构残差正交的 IV 矩条件。*

3.6 选择偏倚（Selection Bias）：谁进入了数据，决定了你能学到什么

3.6.1 概念：选择机制改变了可观测世界

选择偏倚（selection bias）指的是：样本进入数据集的机制并非对目标总体（target population）的随机抽样，而是与研究中的结果、处理或协变量有关，从而使基于观测样本得到的关系不能代表目标总体中的关系。更形式化地说，设 $S\in\{0,1\}$ 表示个体是否被纳入样本（$S=1$ 表示被观测到）。若纳入概率

$$P(S=1\mid Y,X,T)$$

依赖于 Y,X,T 中的某些量，则即使你在 $S=1$ 的样本里做了“标准的”统计分析，所对应的推断对象也可能不再是目标总体中的规律。

关键提醒（推断对象会被“悄悄替换”）： 一旦你只在 $S=1$ 的子样本上分析，你回答的往往是“*在被观测到的人群里 Y 与 X,T 的关系是什么*”，而不是“*在目标总体里 Y 与 X,T 的关系是什么*”。选择偏倚因此不仅意味着“估计量有偏”，更意味着 **estimand 已被数据生成机制改变**。

3.6.2 选择偏倚与内生性：为何条件外生性会被破坏

为突出机制，先从最简单的回归出发。设目标总体中存在结构关系

$$Y=\beta X+\varepsilon,\qquad \mathbb{E}(\varepsilon\mid X)=0,$$

也就是说，在总体里 X 是外生的。如果我们只在被观测到的子样本 $S=1$ 上回归 Y 对 X，则 OLS 的关键条件变成了

$$\mathbb{E}(\varepsilon\mid X,S=1)=0.$$

但只要选择机制让 S 同时与 X 和 ε（或与 Y）有关，上式一般不成立：

$$\mathbb{E}(\varepsilon\mid X,S=1)\neq 0,$$

于是 X 在子样本里变得“内生”，回归系数会系统偏离目标总体的 β。

3.6.3 一个最小例子：条件化 $S=1$ 如何诱发 X 与误差相关

考虑一个极简选择规则：只记录“结果足够大”的个体，

$$S=1\{Y>\text{cut}\}.$$

在总体中仍假设

$$Y = \beta X + \varepsilon, \qquad \varepsilon \perp X, \quad \mathbb{E}(\varepsilon) = 0.$$

给定 $X = x$ 时，$S = 1$ 等价于

$$\varepsilon > \text{cut} - \beta x.$$

因此在 $S = 1$ 的子样本中，ε 的条件分布是一个被截断的分布，其条件均值一般不为零，并且随 x 改变：

$$\mathbb{E}(\varepsilon \mid X = x, S = 1) = \mathbb{E}(\varepsilon \mid \varepsilon > \text{cut} - \beta x),$$

右边依赖于阈值 $\text{cut} - \beta x$，从而依赖于 x。这就意味着：虽然在总体里 $\varepsilon \perp X$，但在条件 $S = 1$ 下 X 与 ε 被“选择机制”人为关联起来，外生性被破坏。

3.6.4 三种常见选择形态：缺失、截断与条件化中间事件

实践里“进入样本”并不只有一种含义，至少有三类常见形态：

1. **缺失/未响应（missing/attrition）：** 个体属于目标总体，变量在概念上存在，但由于失访、拒访或记录不全而未被观测到（$S = 0$）。
2. **截断/只在子总体中定义（truncation）：** 某些变量只对特定子总体有观测或有定义（如工资通常只对就业者可见），因此样本天然限制在 $S = 1$ 的子总体上。
3. **条件化某个中间事件（conditioning on an intermediate event）：** 例如只在“住院者”“被检测者”“被平台曝光者”中比较处理与结局；此时 S 往往既受处理也受未观测风险（或偏好）影响，是引入偏倚的高危情形。

这三类都可以写成“只看 $S = 1$ 的数据”，但其识别含义不同：缺失更像“没看见”，截断更像“变量只在一类人身上定义”，而条件化中间事件常直接对应碰撞器偏倚。

3.6.5 DAG 视角：选择常等价于对碰撞器条件化

用因果图语言，选择偏倚常可理解为对碰撞器（collider）条件化。考虑结构

$$X \to S \leftarrow U \to Y,$$

其中 U 表示影响结局的未观测因素（也可以理解为误差项来源）。在总体中 X 与 U 可以独立；但一旦我们只分析 $S = 1$ 的样本，就等价于对碰撞器 S 条件化，从而诱发 X 与 U 的相关，最终表现为 X 与 Y 的“被选择出来的关联”。

常见误区：“多控制一些变量更稳健”并不总对。如果你控制（或隐含条件化）的是由结果、由中间事件、或由选择过程决定的变量（或其强代理），就可能打开本来关闭的路径，反而引入偏差。

3.6.6 识别的核心分水岭：选择是否在协变量下可忽略

从识别角度看，关键问题是：在给定可观测信息后，选择是否仍与结果的不可观测部分相关。用“缺失机制”的语言可以把直觉分成三层（此处仅作识别层级的描述）：

- **（近似）完全随机：** $S \perp (Y, X, T)$。此时在 $S = 1$ 上分析通常仍无偏（但效率下降）。

- **在协变量下可忽略：** $S \perp Y \mid (X,T)$。此时可用 $P(S=1 \mid X,T)$ 做再平衡，尝试恢复目标总体量。
- **不可忽略选择：** $S \not\perp Y \mid (X,T)$。此时仅靠加权一般不够，需要额外外生信息或更强结构假设。

因此，“能不能修正选择偏倚”首先是一个**识别问题**：你是否愿意并且能够为选择机制提供足够的、可辩护的外生信息。

3.6.7 三个例子：同一机制在不同学科里的面孔

1. **只在就业者中研究工资决定因素：**工资只对就业者可见，而就业与能力、家庭条件、健康等共同决定。在就业者样本上回归“工资对教育”，容易把“谁更可能就业”的筛选混入误差，形成经典的样本选择问题。
2. **只在住院者中研究治疗与死亡：**是否住院受病情严重程度与就医可得性共同影响；住院后接受何种治疗也与严重程度相关。在住院样本里比较治疗方案，常把“被住院/被治疗”的选择机制误当成治疗效应。
3. **平台数据与算法推荐：**你能看到的点击/购买建立在“被展示”之上，而展示概率依赖用户偏好与内容特征。如果忽略展示这一选择环节，直接用曝光后的数据解释偏好或因果效应，往往会把推荐机制当成用户偏好。

3.6.8 应对思路：先定 estimand，再讨论能否修正

选择偏倚没有通用的自动修复方式，但可以按“先把问题说清楚、再谈修正”的顺序推进。

1. **先明确 estimand（你要推断谁）：**你关心的是目标总体中的效应/关系，还是“被观测者 $(S=1)$”子总体中的效应/关系？若数据天然只覆盖 $S=1$，至少应明示你的推断对象，并说明从 $S=1$ 外推到目标总体需要哪些额外假设。
2. **设计层面的修正：**在数据收集阶段改为更接近随机的抽样/随访，或通过设计减少 S 对 Y 与 T 的依赖；这类修正通常比事后建模更可信。
3. **加权与再平衡（选择在协变量下可忽略时）：**若可以相信 $S \perp Y \mid (X,T)$，并且存在足够重叠（positivity），可估计 $\pi(X,T)=P(S=1 \mid X,T)$ 并使用逆概率加权

   $$w_i = \frac{1}{\hat{\pi}(X_i,T_i)}$$

   将样本“还原”为目标总体。需要强调：当 $\pi(X,T)$ 在某些区域接近 0 时，权重会爆炸，推断将依赖少数样本点而极不稳定。
4. **选择模型与外生信息（不可忽略选择时）：**当 $S \not\perp Y \mid (X,T)$，需要更强的外生信息或结构假设，例如显式的选择方程模型，或存在只影响 S 而不直接影响 Y 的变量（可视作“选择的工具”）。
5. **敏感性分析：**当选择机制无法可信识别时，应报告在不同不可观测选择假设下结论如何变化，把不可检验部分转化为可讨论的不确定性区间或情景比较。

最低限度的实践诊断（建议报告）：（i）画出从目标总体到分析样本的纳入流程（每一步都对应一个选择节点）；（ii）若能获得未纳入者的部分信息，比较 $S=1$ 与 $S=0$ 在关键 X 上的差异；（iii）若使用 IPW，检查权重分布是否极端（结论是否由少数高权重样本“撑起来”）。

3.6.9 小结：把“谁被看见”从背景噪声变成识别前提

选择偏倚强调：识别不仅需要结果方程，还需要样本进入机制。总体中成立的外生性条件，在条件 $S=1$ 的可观测世界里可能立即失效；更常见的是，你的推断对象在不知不觉中从目标总体变成了被选择出来的子总体。因此，一旦数据并非随机进入，就必须显式写出（或至少系统讨论）选择过程；否则回归形式再标准，推断含义也可能已经改变。

4 线性回归模型

4.1 线性回归

为了讨论后续的路径模型、联立方程与识别问题，我们需要一套最小的回归语言来连接“数据机制”与“可计算的统计量”。

条件均值与线性近似 设 Y 为响应变量，$X\in\mathbb{R}^p$ 为协变量向量。回归首先关心条件均值函数

$$\eta(x)=\mathbb{E}(Y\mid X=x).$$

线性回归把 $\eta(x)$ 限制（或近似）在一个线性族中：存在参数 $\alpha\in\mathbb{R}$ 与 $\beta\in\mathbb{R}^p$ 使

$$\mathbb{E}(Y\mid X)=\alpha+X^\top\beta.$$

等价地，可以写成

$$Y=\alpha+X^\top\beta+\varepsilon,\qquad \mathbb{E}(\varepsilon\mid X)=0,$$

其中 ε 表示 Y 相对于其条件均值的偏离（包含未建模因素与随机波动）。需要强调的是：这里的“线性”是对条件均值的建模选择；β 是否具有因果含义并非由线性形式自动保证，而取决于 X 的外生性、混杂控制以及数据生成机制是否支持这种解释。

外生性：何时可以把系数当作稳定关系的参数 要让回归系数对应于条件均值中的系统性部分，一个核心要求是**外生性**（正交条件）：

$$\mathbb{E}(\varepsilon\mid X)=0.$$

它的直观含义是：在给定 X 后，剩余项 ε 不再携带与 X 系统相关的信息；因此，X 的变化可以被用来刻画 $\mathbb{E}(Y\mid X)$ 的变化。外生性成立时，OLS 在大样本下可以一致估计该条件均值的线性近似系数，并且在额外误差结构假设下得到标准误与推断公式。一旦外生性不成立（例如遗漏变量、同时性、反向因果、测量误差、选择进入样本等导致 $\mathbb{E}(\varepsilon\mid X)\neq 0$），OLS 往往出现系统性偏误，而且这种偏误通常不会随着样本量增大而消失——估计量可能收敛到与目标无关的“伪真值”。这也是后续讨论识别、工具变量与各种外生变化设计时反复强调的核心动机。

最小二乘（OLS）作为投影：正规方程、闭式解与几何直觉 给定样本 $\{(x_i, y_i)\}_{i=1}^n$，令设计矩阵与响应向量为

$$\boldsymbol{X} = \begin{pmatrix} x_1^\top \\ \vdots \\ x_n^\top \end{pmatrix} \in \mathbb{R}^{n\times p}, \qquad \boldsymbol{y} = (y_1, \dots, y_n)^\top \in \mathbb{R}^n,$$

并用常数列并入设计矩阵：$\tilde{\boldsymbol{X}} = [\mathbf{1}, \boldsymbol{X}] \in \mathbb{R}^{n\times(p+1)}$，参数向量 $\tilde{\boldsymbol{\beta}} = (\alpha, \beta^\top)^\top \in \mathbb{R}^{p+1}$。最小二乘估计量定义为

$$\hat{\boldsymbol{\beta}} = \arg\min_{\tilde{\boldsymbol{\beta}}} \ \|\boldsymbol{y} - \tilde{\boldsymbol{X}}\tilde{\boldsymbol{\beta}}\|_2^2 = \arg\min_{\alpha,\beta} \sum_{i=1}^n (y_i - \alpha - x_i^\top \beta)^2.$$

对目标函数求导并令其为零，得到**正规方程**（normal equations）：

$$\tilde{\boldsymbol{X}}^\top(\boldsymbol{y} - \tilde{\boldsymbol{X}}\hat{\boldsymbol{\beta}}) = \mathbf{0},$$

等价地说，残差 $\hat{\boldsymbol{\varepsilon}} = \boldsymbol{y} - \tilde{\boldsymbol{X}}\hat{\boldsymbol{\beta}}$ 与 $\tilde{\boldsymbol{X}}$ 的每一列都正交。几何上，拟合值

$$\hat{\boldsymbol{y}} = \tilde{\boldsymbol{X}}\hat{\boldsymbol{\beta}}$$

是把 $\boldsymbol{y}$ 正交投影到列空间 $\mathrm{col}(\tilde{\boldsymbol{X}})$ 的结果。

定理 4.1 (OLS 的闭式解（满列秩条件下）)**.** *若 $\tilde{\boldsymbol{X}}^\top\tilde{\boldsymbol{X}}$ 可逆，则*

$$\hat{\boldsymbol{\beta}} = (\tilde{\boldsymbol{X}}^\top\tilde{\boldsymbol{X}})^{-1}\tilde{\boldsymbol{X}}^\top\boldsymbol{y}.$$

何时不存在闭式解？ 当 $\tilde{\boldsymbol{X}}$ 不满列秩（如 p 很大或存在完全共线性）时，$\tilde{\boldsymbol{X}}^\top\tilde{\boldsymbol{X}}$ 不可逆，最小二乘解可能不唯一。此时可用广义逆给出一组最小二乘解，或引入正则化（如岭回归）获得稳定解；这一点也与后续的可解释性与共线性讨论直接相关（见 §4.6）。

投影算子（帽子矩阵） 在满列秩下，定义帽子矩阵与残差算子

$$H = \tilde{\boldsymbol{X}}(\tilde{\boldsymbol{X}}^\top\tilde{\boldsymbol{X}})^{-1}\tilde{\boldsymbol{X}}^\top, \qquad M = I_n - H,$$

则 $\hat{\boldsymbol{y}} = H\boldsymbol{y}$、$\hat{\boldsymbol{\varepsilon}} = M\boldsymbol{y}$。它们满足 $H^2 = H$、$H^\top = H$（对称幂等），以及 $M^2 = M$、$M^\top = M$。。

4.2 OLS 的方差、估计误差与解释

沿用前文记号：$\tilde{\boldsymbol{X}} = [\mathbf{1}, \boldsymbol{X}] \in \mathbb{R}^{n\times(p+1)}$，$\tilde{\boldsymbol{\beta}} = (\alpha, \beta^\top)^\top \in \mathbb{R}^{p+1}$，模型写作

$$\boldsymbol{y} = \tilde{\boldsymbol{X}}\tilde{\boldsymbol{\beta}} + \boldsymbol{\varepsilon}.$$

在外生性（A）$\mathbb{E}(\boldsymbol{\varepsilon} \mid \tilde{\boldsymbol{X}}) = \mathbf{0}$ 成立时，

$$\mathbb{E}(\hat{\boldsymbol{\beta}} \mid \tilde{\boldsymbol{X}}) = \tilde{\boldsymbol{\beta}}.$$

在进一步假设（B）同方差不相关 $\mathrm{Var}(\boldsymbol{\varepsilon} \mid \tilde{\boldsymbol{X}}) = \sigma^2 I_n$ 时，

$$\mathrm{Var}(\hat{\boldsymbol{\beta}} \mid \tilde{\boldsymbol{X}}) = \sigma^2(\tilde{\boldsymbol{X}}^\top\tilde{\boldsymbol{X}})^{-1}.$$

若再假设（C）条件正态 $\boldsymbol{\varepsilon} \mid \tilde{\boldsymbol{X}} \sim N(0, \sigma^2 I_n)$，则

$$\hat{\boldsymbol{\beta}} \mid \tilde{\boldsymbol{X}} \sim N\big(\tilde{\boldsymbol{\beta}},\ \sigma^2 (\tilde{\boldsymbol{X}}^\top \tilde{\boldsymbol{X}})^{-1}\big),$$

并且与若干平方和统计量一起导出 t、χ^2 与 F 的精确有限样本分布（见 §4.4）。

残差平方和与自由度 记帽子矩阵与残差算子为

$$H = \tilde{\boldsymbol{X}}(\tilde{\boldsymbol{X}}^\top \tilde{\boldsymbol{X}})^{-1} \tilde{\boldsymbol{X}}^\top, \qquad M = I_n - H,$$

则 $\hat{\boldsymbol{y}} = H\boldsymbol{y}$，$\hat{\boldsymbol{\varepsilon}} = M\boldsymbol{y}$。残差平方和（residual sum of squares, RSS）为

$$\mathrm{RSS} = \|\hat{\boldsymbol{\varepsilon}}\|_2^2 = \hat{\boldsymbol{\varepsilon}}^\top \hat{\boldsymbol{\varepsilon}} = \boldsymbol{y}^\top M \boldsymbol{y}.$$

在同方差不相关假设下常用

$$\hat{\sigma}^2 = \frac{\mathrm{RSS}}{n - (p+1)}$$

估计 σ^2，其中 $n - (p+1)$ 是残差自由度（因为参数维度为 $p+1$，包含截距）。

4.3 Gauss–Markov：为什么说 OLS 是 BLUE

在经典线性模型（A）+（B）下，OLS 具有一个“效率最优”性质。

线性无偏估计器 考虑所有形如 $\tilde{\boldsymbol{\beta}}^* = A\boldsymbol{y}$ 的估计器，其中 A 是只依赖于 $\tilde{\boldsymbol{X}}$ 的确定矩阵。若对任意 $\tilde{\boldsymbol{\beta}}$ 都有 $\mathbb{E}(\tilde{\boldsymbol{\beta}}^* \mid \tilde{\boldsymbol{X}}) = \tilde{\boldsymbol{\beta}}$，则称其为线性无偏估计器。无偏性条件等价于

$$A\tilde{\boldsymbol{X}} = I_{p+1}.$$

BLUE（Best Linear Unbiased Estimator） 在（A）+（B）下，OLS 估计量

$$\hat{\boldsymbol{\beta}} = (\tilde{\boldsymbol{X}}^\top \tilde{\boldsymbol{X}})^{-1} \tilde{\boldsymbol{X}}^\top \boldsymbol{y}$$

在所有线性无偏估计器中具有最小协方差矩阵：对任意线性无偏 $\tilde{\boldsymbol{\beta}}^*$，

$$\mathrm{Var}(\tilde{\boldsymbol{\beta}}^* \mid \tilde{\boldsymbol{X}}) - \mathrm{Var}(\hat{\boldsymbol{\beta}} \mid \tilde{\boldsymbol{X}})$$

为半正定矩阵。这就是 Gauss–Markov 定理（OLS is BLUE）。

一句话理解 在同方差且不相关的条件下，OLS 可以理解为：在满足无偏性的所有线性解里，选取方差最小的那一个。注意这种“最优性”是有条件的：若误差异方差或相关，OLS 仍可能无偏（取决于外生性），但不再是 BLUE。

4.4 正态理论与 t/χ^2/F：推断公式从何而来（补充）

在（A）+（B）的基础上，很多教材进一步假设条件正态性（C）：

$$\boldsymbol{\varepsilon} \mid \tilde{\boldsymbol{X}} \sim N(0, \sigma^2 I_n).$$

这一假设并非 OLS 一致性所必需，但它使得若干统计量具有精确的有限样本分布，从而得到 t、χ^2 与 F 检验的标准形式。

从正态到 t：单个系数的区间与检验 在正态假设下，

$$\hat{\boldsymbol{\beta}} \mid \tilde{\boldsymbol{X}} \sim N\big(\tilde{\boldsymbol{\beta}},\ \sigma^2(\tilde{\boldsymbol{X}}^\top \tilde{\boldsymbol{X}})^{-1}\big).$$

记第 j 个分量为 $\hat{\boldsymbol{\beta}}_j$，并令 $s_j^2 = \big[(\tilde{\boldsymbol{X}}^\top \tilde{\boldsymbol{X}})^{-1}\big]_{jj}$，　　$\widehat{\mathrm{se}}(\hat{\boldsymbol{\beta}}_j) = \hat{\sigma}\sqrt{s_j^2}$，其中

$$\hat{\sigma}^2 = \frac{\mathrm{RSS}}{n-(p+1)}, \qquad \mathrm{RSS} = \|\hat{\boldsymbol{\varepsilon}}\|_2^2.$$

则在原假设 $H_0 : \tilde{\boldsymbol{\beta}}_j = \tilde{\boldsymbol{\beta}}_{j,0}$ 下，t 统计量 $t_j = \frac{\hat{\boldsymbol{\beta}}_j - \tilde{\boldsymbol{\beta}}_{j,0}}{\hat{\sigma}\sqrt{s_j^2}}$ 满足

$$t_j \ \sim\ t_{\,n-(p+1)}.$$

这解释了 t 分布针对的对象：它用于对**单个线性系数（或单个线性对比）**做检验与置信区间，核心原因是我们用同一份数据用 $\hat{\sigma}$ 代替了未知的 σ。

为何出现 χ^2：残差平方和的分布 在正态假设下，残差向量是对正态噪声做正交投影后的结果，因而

$$\frac{\mathrm{RSS}}{\sigma^2} = \frac{\hat{\boldsymbol{\varepsilon}}^\top \hat{\boldsymbol{\varepsilon}}}{\sigma^2} \ \sim\ \chi^2_{n-(p+1)}.$$

并且在经典线性模型下，$\hat{\boldsymbol{\beta}}$ 与 RSS 条件独立（给定 $\tilde{\boldsymbol{X}}$），这也是 t 分布推导的关键一步：正态的分子与独立的 χ^2 分母之比给出 t。

为何出现 F：多个线性约束的联合检验 考虑 q 个线性约束

$$H_0 :\ R\tilde{\boldsymbol{\beta}} = r, \qquad R \in \mathbb{R}^{q\times(p+1)}.$$

记无约束模型残差平方和为 RSS_0，约束模型残差平方和为 RSS_1（约束下拟合更差，故 $\mathrm{RSS}_1 \geq \mathrm{RSS}_0$）。则在 H_0 下，F 统计量

$$F = \frac{(\mathrm{RSS}_1 - \mathrm{RSS}_0)/q}{\mathrm{RSS}_0/\{n-(p+1)\}}$$

满足

$$F \ \sim\ F_{q,\ n-(p+1)}.$$

因此，F 检验对应的是**多个系数（或多个线性对比）的联合检验**；而当 $q = 1$ 时，F 检验与对应的 t 检验等价（$F = t^2$）。

4.5 误差异方差或相关：GLS、FGLS 与稳健推断

在很多应用中，误差并不满足 $\mathrm{Var}(\boldsymbol{\varepsilon} \mid \tilde{\boldsymbol{X}}) = \sigma^2 I_n$：它可能异方差，或在时间/空间/群组内相关。一种经典修正是将误差协方差写成

$$\mathrm{Var}(\boldsymbol{\varepsilon} \mid \tilde{\boldsymbol{X}}) = \Sigma,$$

其中 Σ 为对称正定矩阵（允许非对角相关与不等方差）。

GLS 估计（Σ 已知） 若 Σ 已知，则广义最小二乘（GLS）可写成加权最小二乘问题：

$$\hat{\boldsymbol{\beta}}_{\mathrm{GLS}} = \arg\min_{\tilde{\boldsymbol{\beta}}} (\boldsymbol{y} - \tilde{\boldsymbol{X}}\tilde{\boldsymbol{\beta}})^\top \Sigma^{-1} (\boldsymbol{y} - \tilde{\boldsymbol{X}}\tilde{\boldsymbol{\beta}}),$$

其闭式解为

$$\hat{\boldsymbol{\beta}}_{\mathrm{GLS}} = (\tilde{\boldsymbol{X}}^\top \Sigma^{-1} \tilde{\boldsymbol{X}})^{-1} \tilde{\boldsymbol{X}}^\top \Sigma^{-1} \boldsymbol{y}.$$

从“变换”的角度看，令 $C^\top C = \Sigma^{-1}$（例如 Cholesky 分解），则 $C\boldsymbol{y} = C\tilde{\boldsymbol{X}}\tilde{\boldsymbol{\beta}} + C\boldsymbol{\varepsilon}$ 的误差协方差为单位阵；在变换后的模型上做 OLS，即得到 GLS。

FGLS 与实践困难（Σ 未知） 现实中 Σ 通常未知，需要先估计得到 $\hat{\Sigma}$，再代入形成可行 GLS（feasible GLS, FGLS）。关键难点在于：$\hat{\Sigma}$ 的建模与估计本身也会带来误差，而 GLS 的表现对 Σ 的设定较敏感；在小样本中，FGLS 甚至可能恶化不确定性评估。

4.6 共线性与中心化

共线性是什么 当设计矩阵 $\boldsymbol{X}$ 的列几乎线性相关时，称为**（近）共线性**（(near) multicollinearity）。它不一定导致偏误，但会放大方差：$(\tilde{\boldsymbol{X}}^\top \tilde{\boldsymbol{X}})^{-1}$ 的某些方向变大，从而使系数不稳定、置信区间变宽。

中心化与尺度 当模型中包含交互项或多项式项时，中心化（对变量减去均值）常能降低数值共线性并改善解释口径；标准化（除以标准差）则便于比较不同量纲变量的相对贡献，但会改变系数单位解释，需在写作中明确。

4.7 线性回归模型的正则化

当特征维度 p 与样本量 n 同量级，甚至出现 $p \gg n$ 时，经典最小二乘会遭遇两类问题：

- **方差爆炸与不稳定：** 共线性（或近共线）使得 $\hat{\beta}_{\mathrm{OLS}}$ 对噪声极其敏感；
- **可识别性：** 当 $p > n$ 时，$X^\top X$ 不可逆，OLS 不再唯一，训练误差可以被“拟合到几乎为零”，泛化却很差。

正则化的核心思想是：*允许一点偏差（bias），换取显著的方差下降（variance reduction）与更可控的泛化误差*。在回归中最常见的两种正则化是 L_2（Ridge）与 L_1（Lasso），以及二者折中的 Elastic Net。

4.7.1 Ridge：偏差–方差权衡与收缩

定义与闭式解 Ridge 回归通过 L_2 惩罚收缩系数：

$$\hat{\beta}_{\text{ridge}}(\lambda) = \arg\min_{\beta} \frac{1}{2n}\|y - X\beta\|_2^2 + \lambda\|\beta\|_2^2, \qquad \lambda \geq 0.$$

若 X 已中心化（y 也中心化）且不含截距项，则

$$\hat{\beta}_{\text{ridge}} = (X^\top X + 2n\lambda I)^{-1} X^\top y.$$

当 $p > n$ 时，$X^\top X$ 不可逆，但加上 λI 后变为可逆，从而得到唯一解。

偏差–方差直觉（用特征值看） 设 $X^\top X = VDV^\top$（D 为特征值对角阵），则

$$\hat{\beta}_{\text{ridge}} = V(D + 2n\lambda I)^{-1} V^\top X^\top y,$$

其效果相当于在每个主方向上把系数按因子 $\frac{d_j}{d_j + 2n\lambda}$ 进行收缩：小特征值（不稳定方向）会被强力压缩，从而显著降低方差；代价是引入偏差。

与贝叶斯与惩罚的对应 Ridge 等价于在高斯噪声线性模型下对 β 施加零均值各向同性高斯先验：$\beta \sim \mathcal{N}(0, \tau^2 I)$（参数对应关系与 λ 成比例），因此可以视为“把系数拉回到 0 附近”的一种结构化保守。

实践要点

- **标准化：** Ridge 的惩罚对尺度敏感，通常对每列 x_j 做零均值单位方差标准化（截距单独处理）。
- **选 λ：** 用验证集或（嵌套）交叉验证选 λ；常用“最小误差”或“1-SE 规则”取更稳健的较大 λ。
- **可解释性：** Ridge 不产生精确零系数（通常），更适合“许多变量都可能有小贡献”的场景，以及强共线条件下的稳定预测。

4.7.2 Lasso：稀疏性与变量选择

定义 Lasso 用 L_1 惩罚诱导稀疏：

$$\hat{\beta}_{\text{lasso}}(\lambda) = \arg\min_{\beta} \frac{1}{2n}\|y - X\beta\|_2^2 + \lambda\|\beta\|_1, \qquad \|\beta\|_1 = \sum_{j=1}^{p} |\beta_j|.$$

为何 L_1 会产生“精确为 0”？（几何直觉） L_1 约束集合是一个带尖角的高维菱形（cross-polytope）；最小二乘的等高线往往首先触到尖角处，从而使某些坐标精确为 0。这就是 Lasso 同时实现“正则化 + 变量选择”的原因。

软阈值与近端算子（最核心的一步） 在正交设计（或坐标下降的单坐标子问题）中，解具有软阈值形式：

$$\hat{\beta}_j \leftarrow S(z_j, \lambda) = \text{sign}(z_j) \cdot (|z_j| - \lambda)_+,$$

其中 $S(\cdot)$ 是 soft-thresholding。它形象地说明：小信号直接被压成 0，大信号被向 0 收缩。

算法：坐标下降与路径 Lasso 的工业级实现通常用坐标下降（coordinate descent），配合一串 λ 从大到小的路径（warm start）。经典的“整条路径”思路还有 LARS（Least Angle Regression）方法，能高效给出分段线性路径。

何时 Lasso “选得准”？（非常简略的理论提示） 在高维理论中，Lasso 的好性质通常依赖：

- 真系数足够稀疏（$s = \|\beta^\star\|_0$ 小）；
- 设计矩阵满足某种“受限特征值/兼容性条件”（避免强相关导致不可识别）；
- 变量选择的一致性还需要更强的条件（如 irrepresentable condition）。

直观上：**强相关会让“选变量”这件事变得不稳定**——Lasso 可能在一组相关变量里随机挑一个；这不是算法缺陷，而是信息本身不足导致的不可区分。

实践要点

- **标准化几乎是必须的：**否则惩罚会不公平地偏向尺度小的变量。
- **选择结果的不稳定：**当特征高度相关时，选出的集合可能对数据扰动敏感；可以用重复抽样/稳定性选择（stability selection）等方式诊断。
- **“选模后推断”要谨慎：**Lasso 先看数据做选择，再做系数检验会导致显著性膨胀；若需要推断，考虑去偏（debiased/desparsified）Lasso 或选择后推断框架（见文献）。

4.7.3 Elastic Net 与实践细节：标准化、路径算法与可解释性

Elastic Net：介于 Ridge 与 Lasso Elastic Net 用 $L_1 + L_2$ 的组合惩罚：

$$\hat{\beta}_{\text{enet}}(\lambda, \alpha) = \arg\min_{\beta} \ \frac{1}{2n}\|y - X\beta\|_2^2 + \lambda\Big(\alpha\|\beta\|_1 + \frac{1-\alpha}{2}\|\beta\|_2^2\Big),$$

其中 $\alpha \in [0, 1]$：

- $\alpha = 1$ 退化为 Lasso；
- $\alpha = 0$ 退化为 Ridge（差一个常数缩放）。

为什么它在相关特征下更好？（grouping effect） 在一组高度相关的变量中：

- Lasso 倾向于“选一个、丢其余”（更稀疏，但不稳定）；
- Ridge 倾向于“大家一起缩”（稳定但不稀疏）；
- Elastic Net 常表现为“成组进入/退出”，既稀疏又更稳定，尤其适合强相关的高维特征（例如基因、文本、影像特征）。

实现细节 1：标准化与截距 几乎所有惩罚回归都默认：

- 对 X 列标准化：$x_j \leftarrow (x_j - \bar{x}_j)/s_j$；
- 截距不惩罚；y 也常中心化以简化处理。

否则“惩罚哪个变量”会被量纲主导而不是信息主导。

实现细节 2：路径算法与 warm start 工业实现（如 `glmnet`）通常固定一组 α（或少量候选），对 λ 生成一条递减网格：

$$\lambda_{\max} \to \cdots \to \lambda_{\min},$$

用上一点的解作为下一点初值（warm start），配合坐标下降快速收敛。这也是为什么“全路径 + CV”在实践中很高效。

实现细节 3：如何选 (λ, α)？

- 若只调 λ：常见做法先固定 α（如 0.5 或 0.1），对 λ 做 CV。
- 若同时调二者：用网格搜索（少量 α 候选）+ 内层 CV 选 (λ, α)，外层用嵌套 CV 估计最终泛化误差，避免调参乐观偏差。
- 经验上：相关性越强、越希望稳定选择，α 可取更小（更接近 Ridge）。

可解释性：三件事要分开讲清楚

1. **预测性能：**用测试集/外层 CV 指标报告；
2. **变量选择：**Lasso/EN 给出的“非零集合”是带选择机制的结果，需讨论稳定性；
3. **系数大小解释：**惩罚带来系统性收缩偏差，不能直接当作无偏效应估计；若需要效应推断，请明确采用适当的后处理（去偏、选择后推断、或独立确认数据）。

小结：什么时候用谁？

- **Ridge：**多数变量可能有小贡献、强共线、追求稳定预测；
- **Lasso：**相信信号稀疏、想做变量筛选（但要警惕相关特征下不稳定）；
- **Elastic Net：**既想稀疏又想在相关特征下更稳定，常是高维预测的默认首选之一。

5 广义线性模型与似然

广义线性模型（Generalized Linear Model, GLM）把“线性预测子”与“非正态的响应变量分布”系统地粘合在一起：它既保留了线性回归里清晰的结构（线性预测子、可解释的系数），又允许响应变量来自更广的分布族（如二项、泊松、Gamma 等），并通过**链接函数 (link function)** 把均值映射到线性结构上。本节将 GLM 放回到**似然 (likelihood)** 的统一框架下：从指数族形式出发推导对数似然、得分函数与信息矩阵，解释为什么估计可以写成 IRLS，以及 deviance、Wald/LR/Score 等推断工具从何而来。

5.1 指数族与链接函数：均值–方差关系

从线性回归到 GLM：我们到底在推广什么？ 在线性回归中，

$$Y_i = x_i^\top \beta + \varepsilon_i, \qquad \mathbb{E}[\varepsilon_i \mid x_i] = 0, \qquad \mathrm{Var}(\varepsilon_i \mid x_i) = \sigma^2,$$

于是

$$\mu_i := \mathbb{E}[Y_i \mid x_i] = x_i^\top \beta, \qquad \mathrm{Var}(Y_i \mid x_i) = \sigma^2$$

（方差与均值无关）。但很多数据里方差会随均值变化：计数数据、二元数据、正偏连续数据等。GLM 的核心是两件事：

1. 用一个合适的**指数族分布**来刻画 $Y \mid X$ 的条件分布；
2. 用一个**链接函数**把条件均值 μ_i 与线性预测子 $\eta_i = x_i^\top \beta$ 联系起来。

指数族：统一的分布写法 单参数指数族常写为

$$f(y;\theta,\phi) = \exp\left(\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right), \tag{EF}$$

其中 θ 是自然参数、ϕ 是离散参数，$b(\theta)$ 控制均值–方差结构。

指数族的恒等式：均值与方差 由 (EF) 推得

$$\mu := \mathbb{E}[Y \mid X] = b'(\theta), \qquad \mathrm{Var}(Y \mid X) = a(\phi)\, b''(\theta). \tag{MV}$$

常写为

$$\mathrm{Var}(Y \mid X) = \phi\, V(\mu), \tag{VV}$$

其中 $V(\mu)$ 为方差函数。

链接函数：把均值嵌入线性结构 GLM 的系统部分是

$$\eta_i = x_i^\top \beta, \qquad g(\mu_i) = \eta_i, \qquad \mu_i = \mathbb{E}[Y_i \mid x_i]. \tag{Link}$$

链接函数 $g(\cdot)$ 负责把均值空间映到实数轴（如 $(0,1) \to \mathbb{R}$、$(0,\infty) \to \mathbb{R}$）。

规范链接（canonical link）与自然参数 若

$$\eta_i = \theta_i, \tag{Can}$$

则称为规范链接。它把“统计分布的自然参数”与“线性预测子”对齐，使得似然的代数形式最干净，也让后面的得分、信息矩阵与 IRLS 推导更直接。

5.2 GLM 的对数似然、得分与信息矩阵

从指数族到对数似然 在 GLM 中，条件独立假设给出联合对数似然

$$\ell(\beta,\phi)=\sum_{i=1}^{n}\log f(y_i;\theta_i,\phi)=\sum_{i=1}^{n}\left\{\frac{y_i\theta_i-b(\theta_i)}{a(\phi)}+c(y_i,\phi)\right\}, \tag{LL}$$

其中 θ_i 通过 μ_i、链接函数与线性预测子与 β 相连：

$$\eta_i=x_i^\top\beta,\qquad \mu_i=g^{-1}(\eta_i),\qquad \mu_i=b'(\theta_i).$$

因此 β 影响 ℓ 的路径是 $\beta\to\eta\to\mu\to\theta\to\ell$。

得分函数：加权残差正交 对 β 求导可得到统一形式（这是 GLM 推断的核心）：

$$U(\beta)=\frac{\partial\ell(\beta,\phi)}{\partial\beta}=\sum_{i=1}^{n}x_i\,\frac{y_i-\mu_i}{\operatorname{Var}(Y_i\mid x_i)}\,\frac{d\mu_i}{d\eta_i}. \tag{Score}$$

因此 MLE 的一阶条件 $U(\hat\beta)=0$ 表示：在权重 $\frac{1}{\operatorname{Var}(Y_i|x_i)}\frac{d\mu_i}{d\eta_i}$ 下，残差 $(y_i-\mu_i)$ 与每个回归方向 x_i 正交。

Fisher 信息与（近似）方差 在正则条件下，Fisher 信息矩阵为

$$I(\beta)=\mathbb{E}\Big[U(\beta)U(\beta)^\top\Big]=-\mathbb{E}\left[\frac{\partial^2\ell(\beta,\phi)}{\partial\beta\partial\beta^\top}\right],$$

其样本版本（或在收敛点的工作近似）可写成

$$I_n(\beta)\approx X^\top WX, \tag{Info}$$

其中 $W=\operatorname{diag}(w_i)$，而

$$w_i=\frac{\left(\frac{d\mu_i}{d\eta_i}\right)^2}{\operatorname{Var}(Y_i\mid x_i)}=\frac{\left(\frac{d\mu_i}{d\eta_i}\right)^2}{\phi\,V(\mu_i)}. \tag{W}$$

于是渐近方差（或标准误的工作近似）为

$$\widehat{\operatorname{Var}}(\hat\beta)\approx(X^\top\hat WX)^{-1},\qquad \text{若需估计离散参数，则}\widehat{\operatorname{Var}}(\hat\beta)\approx\hat\phi\,(X^\top\hat WX)^{-1}.$$

IRLS：从 Fisher scoring 到“迭代加权最小二乘” 用 Fisher scoring 更新

$$\beta^{(t+1)}=\beta^{(t)}+I_n(\beta^{(t)})^{-1}U(\beta^{(t)}),$$

代入 (Score)–(W) 可化为一次加权最小二乘问题：

$$\beta^{(t+1)}=\arg\min_{\beta}\sum_{i=1}^{n}w_i^{(t)}\big(z_i^{(t)}-x_i^\top\beta\big)^2, \tag{IRLS}$$

其中工作响应与工作权重为

$$z_i^{(t)} = \eta_i^{(t)} + \frac{y_i - \mu_i^{(t)}}{(d\mu_i/d\eta_i)^{(t)}}, \qquad w_i^{(t)} = \frac{\big((d\mu_i/d\eta_i)^{(t)}\big)^2}{\mathrm{Var}(Y_i \mid x_i)^{(t)}}.$$

这解释了：**GLM 的估计就是在对数似然的局部二次近似下，不断解“加权线性回归”**。

5.3 Logistic 回归：似然视角下的系数、检验与陷阱

模型与对数似然 Bernoulli 情形 $Y_i \in \{0,1\}$：

$$Y_i \mid x_i \sim \mathrm{Bernoulli}(p_i), \qquad \eta_i = x_i^\top \beta = \log\frac{p_i}{1-p_i}, \qquad p_i = \frac{e^{\eta_i}}{1+e^{\eta_i}}.$$

对数似然为

$$\ell(\beta) = \sum_{i=1}^{n} \{y_i \log p_i + (1-y_i)\log(1-p_i)\}.$$

得分与权重落到熟悉形式：

$$U(\beta) = \sum_{i=1}^{n} x_i(y_i - p_i), \qquad W = \mathrm{diag}\big(p_i(1-p_i)\big) \quad (\phi = 1).$$

因此 $\widehat{\mathrm{Var}}(\hat\beta) \approx (X^\top \hat W X)^{-1}$。

推断：Wald / LR / Score 在 logit 中的具体实现 对单个系数检验 $H_0 : \beta_j = 0$，

- Wald：$\hat\beta_j / \mathrm{se}(\hat\beta_j)$；
- LR：$2(\ell_{\text{full}} - \ell_{\text{reduced}}) \approx \chi^2_1$；
- Score：在约束模型处计算 $U(\tilde\beta)$ 与 $I(\tilde\beta)$。

在稀有事件、强共线或近分离时，LR/Score 往往比 Wald 更稳。

完全分离（separation）：似然为何“不封顶” 若存在某个线性组合把 $Y=1$ 与 $Y=0$ 完全分开，则可令 $\|\beta\| \to \infty$ 使对数似然持续上升，从而 MLE 不存在。这是典型的**非正则**似然问题。处理办法常包括：惩罚似然（ridge/lasso）、Firth 校正、或重构特征。

5.4 Poisson 回归：似然、offset 与过度离散

模型与对数似然

$$Y_i \mid x_i \sim \mathrm{Poisson}(\mu_i), \qquad \eta_i = x_i^\top \beta = \log \mu_i, \qquad \mu_i = e^{x_i^\top \beta}.$$

对数似然（忽略与 β 无关常数）为

$$\ell(\beta) = \sum_{i=1}^{n} \{y_i \eta_i - \mu_i\}.$$

得分与权重为

$$U(\beta) = \sum_{i=1}^{n} x_i(y_i - \mu_i), \qquad W = \mathrm{diag}(\mu_i) \quad (\phi = 1).$$

offset：把暴露量并入似然 若观测窗口/暴露量为 t_i，令

$$\log \mu_i = x_i^\top \beta + \log t_i,$$

等价于对率 μ_i/t_i 建模，且 $\log t_i$ 的系数固定为 1（不参与估计）。

过度离散：从似然到准似然/稳健方差 泊松强制 $\mathrm{Var}(Y \mid X) = \mu$，但实际常有 $\mathrm{Var}(Y \mid X) > \mu$。若均值模型仍可信，则 $\hat\beta$ 往往仍可视为“均值参数”的合理估计，但需要用 $\hat\phi > 1$ 或 sandwich/聚类稳健方差修正标准误。

5.5 Deviance：把“似然差”变成 GLM 的统一度量

偏差（deviance）与饱和模型 定义

$$D = 2\{\ell(\text{saturated}) - \ell(\text{fitted})\}.$$

它是“当前模型”相对“饱和模型”的似然差距：越小表示越接近可达到的最优拟合。

嵌套模型比较：偏差差就是 LR 若 $\mathcal{M}_0 \subset \mathcal{M}_1$，

$$D(\mathcal{M}_0) - D(\mathcal{M}_1) = 2\{\ell(\hat\theta_1) - \ell(\hat\theta_0)\} = LR \quad \overset{a}{\sim} \quad \chi^2_{\Delta\mathrm{df}}.$$

因此 GLM 的“逐步加变量/整体检验”通常直接比较 deviance。

Pearson 统计量与离散参数的线索 另一常用量是

$$X^2 = \sum_{i=1}^{n} \frac{(y_i - \hat\mu_i)^2}{\widehat{\mathrm{Var}}(Y_i \mid x_i)}.$$

对 Poisson 等模型，若 D/df 或 X^2/df 显著大于 1，往往提示过度离散或模型失配，从而需要准似然/稳健方差或更换分布族。

本节小结：从似然看 GLM

- 指数族给出对数似然的统一表达 (LL)，并锁定均值–方差关系；
- 得分方程 (Score) 把 MLE 解释为“加权残差正交”，信息矩阵 (Info) 给出标准误；
- Fisher scoring 的二阶近似导出 IRLS，使 GLM 估计可视作迭代加权最小二乘；
- deviance 把似然差变成可比较的拟合优度度量，嵌套比较时 deviance 差即 LR 检验。

6 非参数回归与平滑

本节从“把回归函数当作对象来估计”的角度介绍非参数回归。与 GLM 等参数模型不同，非参数方法不把 $\mathbb{E}[Y \mid X = x]$ 限制为有限维形式（如 $g(\mu) = x^\top \beta$），而是直接估计未知函数 $m(x) := \mathbb{E}[Y \mid X = x]$（以及更一般的条件分布特征，如分位数）。它的代价与魅力在同一处：为了让函数估计可行且稳定，我们必须引入**平滑/正则化**；而所有关键结论（预测误差、推断、收敛率）都围绕**偏差–方差权衡**展开。

6.1 为何需要非参数回归：回归函数与 L_2 风险

6.1.1 从预测最优性到回归函数：$\eta(x) = \mathbb{E}(Y \mid X = x)$

平方损失下的最优预测器 设我们用某个函数 f 来预测 Y，并以平方损失衡量预测误差：

$$R(f) := \mathbb{E}\big[(Y - f(X))^2\big].$$

经典结论是：使 $R(f)$ 最小的函数（贝叶斯最优预测器）是条件期望

$$\eta(x) := \mathbb{E}[Y \mid X = x].$$

证明只需对条件期望做“配方”：

$$\mathbb{E}\big[(Y - f(X))^2\big] = \mathbb{E}\Big[\mathbb{E}\big[(Y - f(X))^2 \mid X\big]\Big] = \mathbb{E}\Big[\mathrm{Var}(Y \mid X) + \big(\eta(X) - f(X)\big)^2\Big],$$

因此最小化等价于令 $f(X) = \eta(X)$。

不可约误差与可学习部分 上式给出一个重要分解：

$$R(f) = \mathbb{E}[\mathrm{Var}(Y \mid X)] + \mathbb{E}\big[(f(X) - \eta(X))^2\big]. \tag{Risk}$$

第一项是**不可约噪声**（无论怎样建模都无法消去），第二项是**可学习误差**（逼近回归函数的误差）。非参数回归的目标就是：在尽量少做结构假设的前提下，让第二项尽可能小。

6.1.2 L_2 风险、L_2 误差与应用场景

L_2 误差 如果我们关心的是“整体上”估计函数的精度，常用

$$\|\hat{\eta} - \eta\|^2_{L_2(P_X)} := \mathbb{E}\big[(\hat{\eta}(X) - \eta(X))^2\big]$$

来衡量（这正是 (Risk) 第二项）。

什么时候 L_2 风险是合适的？

- **预测导向：**关心平均预测误差（regression / forecasting）。
- **作为后续估计的中间量：**例如在因果推断/半参数估计中，需要估计倾向得分或结果回归，往往只要求这些 nuisance 函数在某种 L_2 意义下收敛即可。

- **平滑信号：**当 $m(\cdot)$ 被认为具有某种光滑性（例如连续、可导），L_2 风险与平滑正则化自然匹配。

6.1.3 参数化与非参数化：假设强弱、可解释性与误差来源

参数化模型：用结构换效率 例如线性回归假设 $\eta(x) = x^\top \beta$。若假设正确，参数估计可达 $n^{-1/2}$ 的“快”收敛率，并且系数可解释。但若结构误设，误差会出现**模型偏差**（misspecification bias）。

非参数模型：用灵活换稳健 非参数不强行规定 $\eta(\cdot)$ 的形式，减少了结构误设风险；但估计一个函数比估计有限维参数更难，收敛率通常更慢，且多维下会遭遇维数灾难。因此非参数回归的核心技术是：**用平滑/正则化把“函数学习”变成可控问题**。

6.2 问题设定与“为什么需要平滑”

回归设定 考虑

$$Y_i = m(X_i) + \varepsilon_i, \qquad \mathbb{E}[\varepsilon_i \mid X_i] = 0, \qquad i = 1, \dots, n,$$

其中 $X_i \in \mathbb{R}^d$，$m(\cdot)$ 未知。我们的目标是估计 $m(x)$ 并理解不确定性。

为什么不能“逐点拟合”？ 如果允许在每个 x 处自由设定 $m(x)$，训练误差可做到极小甚至插值，但对新点的方差会非常大。非参数回归的基本原则是**局部性/平滑性**：相近的 x 应该共享信息，且 $m(\cdot)$ 不应在微小邻域内剧烈振荡。

两种等价视角：局部平均 vs 正则化 多数非参数方法都可看作下列两类思想的实例（经常等价）：

- **局部平均 (local averaging)**：在 x 附近对 Y 做加权平均或局部拟合（核回归、k-NN、局部多项式）。
- **正则化 (regularization)**：在全局拟合中惩罚函数粗糙度（样条、惩罚样条、平滑样条）。

6.3 局部平均方法：核回归与 k-NN 的统一视角

统一写法 许多局部平均估计都可写成

$$\hat{m}(x) = \sum_{i=1}^{n} w_i(x) Y_i, \qquad w_i(x) \geq 0, \qquad \sum_{i=1}^{n} w_i(x) = 1. \tag{LA}$$

其中权重 $w_i(x)$ 随 x 变化，并把更近的点赋予更大权重。在这一视角下，“平滑”就是：**每个 x 使用多少邻域、邻域内如何加权**。

核回归（Nadaraya–Watson） 一维情形中，核回归对应

$$w_i(x) = \frac{K\left(\frac{X_i - x}{h}\right)}{\sum_{j=1}^{n} K\left(\frac{X_j - x}{h}\right)}, \qquad \hat{m}_h(x) = \frac{\sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right) Y_i}{\sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right)}, \tag{NW}$$

K 为核函数，h 为带宽（平滑参数）。

局部多项式（提示） 核回归可视为“局部常数”拟合。把邻域内拟合提升为局部线性或更高阶多项式，可显著缓解边界偏差，并自然给出导数估计；这一点在实践中很重要（后文将再次出现）。

6.4 偏差–方差分解：为何平滑能降低方差却引入偏差

局部平均的方差直觉 由 (LA)，若 $\mathrm{Var}(\varepsilon_i \mid X_i) = \sigma^2$ 且误差独立，则

$$\mathrm{Var}\big(\hat{m}(x) \mid X_{1:n}\big) = \mathrm{Var}\Big(\sum_{i=1}^{n} w_i(x)\varepsilon_i \;\Big|\; X_{1:n}\Big) \approx \sigma^2 \sum_{i=1}^{n} w_i(x)^2. \tag{Var}$$

权重越“分散”（更多点参与平均），$\sum w_i^2$ 越小，方差越小。

偏差来自“把邻域内的真函数当成常数/低阶” 写 $Y_i = m(X_i) + \varepsilon_i$，则

$$\mathbb{E}[\hat{m}(x) \mid X_{1:n}] - m(x) = \sum_{i=1}^{n} w_i(x)\big(m(X_i) - m(x)\big). \tag{Bias}$$

当邻域变大，$m(X_i) - m(x)$ 的系统性差异累计，偏差上升；当邻域变小，偏差下降但权重变“尖”，方差上升。

MSE 分解 因此在点 x 处

$$\mathrm{MSE}\big(\hat{m}(x)\big) = \mathbb{E}\big[(\hat{m}(x) - m(x))^2\big] \approx \mathrm{Bias}^2\big(\hat{m}(x)\big) + \mathrm{Var}\big(\hat{m}(x)\big). \tag{MSE}$$

6.5 平滑参数的角色：带宽 h、近邻数 k、基函数维数 m

三种“同构”的复杂度控制 虽然方法不同，但平滑参数的作用完全同构：

- 核回归：带宽 h（邻域半径）；
- k-NN：近邻数 k（邻域内样本量）；
- 基函数/样条：基函数维数 m（可拟合的波动自由度）与惩罚强度 λ。

它们都在回答同一问题：**局部要用多少信息、函数允许有多“弯”**。

补充：样条/惩罚样条的正则化表述 以惩罚样条为例，选择基函数 $B_1, \ldots, B_m$，写

$$m(x) \approx \sum_{j=1}^{m} \beta_j B_j(x) = B(x)^\top \beta,$$

并通过

$$\hat{\beta} = \arg\min_{\beta} \sum_{i=1}^{n} \big(Y_i - B(X_i)^\top \beta\big)^2 + \lambda\, \beta^\top P \beta \tag{PS}$$

控制粗糙度。这里 m 增大提高表达能力，λ 增大提高平滑程度；两者共同决定有效复杂度（可用有效自由度 edf 来描述）。

6.6 维数灾难与结构性假设：稀疏、可加性、低维表示

维数灾难的根源 当 $X \in \mathbb{R}^d$，局部方法需要在 d 维空间“填满邻域”才能稳定平均；但体积随维数爆炸，导致有效样本极其稀疏，从而收敛率迅速变慢。这不是算法问题，而是信息论意义上的困难：**弱假设下多维函数太难估**。

结构性假设：把“有效维度”降下来 实践中常用三类结构来缓解维数灾难：

- **稀疏性：**只有少数变量真正影响 $m(x)$（高维但低有效维）。
- **可加性：** $m(x) = \sum_{j=1}^{d} m_j(x_j)$（引出下一章 GAM）。
- **低维表示/单指标：** $m(x) = h(x^\top \beta)$ 或存在低维表示 $Z = \varphi(X)$。

6.7 自适应的目标：数据驱动选择与分布无关的表述

数据驱动选择：CV/GCV 的统一动机 平滑参数（h, k, m, λ）通常通过数据驱动方式选择，例如交叉验证：

$$\hat{\theta} = \arg\min_{\theta \in \Theta} \ \mathrm{CV}(\theta),$$

其中 θ 代表任一种平滑参数。其目标是逼近最小化泛化误差的选择。

当平滑器是线性的（如核回归/样条常见形式 $\hat{\mathbf{y}} = S_\theta \mathbf{y}$）时，还可用广义交叉验证（GCV）近似：

$$\mathrm{GCV}(\theta) = \frac{\|\mathbf{y} - \hat{\mathbf{y}}\|^2 / n}{(1 - \mathrm{tr}(S_\theta)/n)^2}. \tag{GCV}$$

分布无关的表述：风险、光滑度类与最优阶 从理论上，非参数的“最好能做到多好”常用风险来刻画，例如

$$\mathbb{E}\big[(\hat{m}(x) - m(x))^2\big] \quad \text{或} \quad \mathbb{E} \int (\hat{m}(x) - m(x))^2 \, dx.$$

为了谈“最优阶”，需要把 $m(\cdot)$ 限定在某个光滑度类（如 Hölder/Sobolev 类）。“自适应”通常指：在未知光滑度的情况下，数据驱动选择仍能达到接近最优的误差阶，而且尽量不依赖误差分布的细节（例如只用有限矩条件）。

6.8 （补充）收敛率的直觉：一维可行，多维受限

一维的典型结论（给直觉） 若 $d = 1$ 且 m 足够光滑（例如二阶可导），许多局部平滑估计能达到

$$\hat{m}(x) - m(x) = O_p\Big(n^{-2/5}\Big),$$

对应最优带宽 $h \asymp n^{-1/5}$。这解释了经验事实：一维平滑很强大，但需要认真选 h。

多维的代价 若 d 增大，局部邻域内有效样本量规模近似为 nh^d；要让 $nh^d \to \infty$，h 不能太小，但 h 不小又会导致偏差。因此收敛率随 d 很快变慢，这正是维数灾难的定量体现。

6.9 （补充）推断：区间估计与偏差校正

为什么推断更难？ 非参数估计往往存在不可忽略的平滑偏差。如果直接用“点估计 ± 正态标准误”，覆盖率可能严重不足。

常见策略（给读者路线图）

- **欠平滑（undersmoothing）：**选更小的 h 让偏差相对方差更小，再做正态近似；
- **显式偏差校正：**估计主导偏差项并扣除；
- **自举/重抽样：**用 bootstrap 校准区间或构造置信带（尤其在复杂平滑器中常用）。

本节小结：非参数回归的主线

- 在平方损失下，最优预测器是回归函数 $\eta(x) = \mathbb{E}[Y \mid X = x]$，非参数回归就是直接学习它；
- 平滑/正则化是把“函数估计”变成可控问题的代价：它降低方差但引入偏差；
- 核回归与 k-NN 属于统一的局部平均框架，h 与 k 控制邻域规模；
- 多维下的主要障碍是维数灾难，因此需要稀疏、可加性或低维表示等结构性假设；
- 数据驱动的平滑选择追求自适应：在尽量少依赖分布细节的前提下达到接近最优的误差阶，并为推断提供可操作方案。

7 半参数模型与部分似然

半参数（semiparametric）模型介于参数模型与非参数模型之间：我们对感兴趣的部分保持有限维参数化（通常是回归系数 β），但允许分布中的某些成分保持为未知函数（无限维扰动项，nuisance function）。这种做法的动机很直接：**我们希望保留“可解释的效应参数”，同时尽量少做不可检验的分布假设**。

本节以生存分析中的 Cox 比例风险模型为主线，引入**部分似然（partial likelihood）**这一半参数推断的经典工具；并给出半参数推断的三个关键词：**识别（identification）**、**效率（efficiency）**与**稳健性（robustness）**。

7.1 半参数模型：参数 + 无限维扰动

一般形式 设观测 O_i 的分布族写成

$$\mathcal{P} = \{P_{\beta,\eta} : \ \beta \in \mathbb{R}^p, \ \eta \in \mathcal{H}\},$$

其中 β 是我们关心的有限维参数（target parameter），η 是无限维参数（nuisance parameter），$\mathcal{H}$ 是某个函数空间。在回归语境下，常见例子是：

- **半参数回归/单指标模型：** $\mathbb{E}[Y \mid X] = h(X^\top \beta)$，其中 $h(\cdot)$ 未知；
- **分位数回归：** $Q_\tau(Y \mid X) = X^\top \beta(\tau)$，不指定误差分布；
- **生存模型：** $\lambda(t \mid X) = \lambda_0(t) \exp(X^\top \beta)$，其中基线风险 $\lambda_0(\cdot)$ 未知。

为什么“只参数化一部分”是有意义的？ 参数模型（如 GLM）把 $Y \mid X$ 的分布锁死在一个有限维族中，推断效率高，但误设风险也高；非参数模型假设弱但维数灾难严重。半参数模型的策略是：**把不可检验、但又极易误设的部分留给 η，把最关心且可解释的部分留给 β**。

7.2 Cox 比例风险模型：半参数的标志性例子

时间到事件与删失 设 T 为事件时间，C 为删失时间，观测到

$$\tilde{T} = \min(T, C), \qquad \Delta = \mathbb{I}(T \le C),$$

并观测协变量 X。删失通常假设为“独立删失”（条件于 X）：

$$T \perp C \mid X.$$

比例风险假设 Cox 模型假设条件风险函数（hazard）满足

$$\lambda(t \mid X) = \lambda_0(t) \exp(X^\top \beta), \tag{Cox}$$

其中 β 为回归系数，$\lambda_0(t)$ 为未知的基线风险函数（无限维扰动）。注意：这里并不需要指定 $T \mid X$ 的完整分布，只要求风险比是指数线性的。

风险比解释 对两个个体 X 与 X'，

$$\frac{\lambda(t \mid X)}{\lambda(t \mid X')} = \exp\big((X - X')^\top \beta\big),$$

与时间 t 无关，这就是“比例风险”。

7.3 部分似然：消去基线风险

核心直觉：只用“谁先发生”这类信息 设发生事件的时刻为 $t_{(1)} < \cdots < t_{(m)}$（只计未删失的事件时刻），在每个 $t_{(k)}$ 时刻，定义风险集（risk set）

$$R(t_{(k)}) = \{i : \ \tilde{T}_i \ge t_{(k)}\},$$

即在 $t_{(k)}$ 之前还“活着且未被删失”的个体集合。

在 Cox 模型 (Cox) 下，条件于“在风险集里有人在 $t_{(k)}$ 发生事件”，事件发生在个体 i 的概率与 $\lambda_0(t)$ 的比例项会相消，得到

$$\Pr(\text{在} t_{(k)} \text{ 发生事件的是} i \mid R(t_{(k)})) = \frac{\exp(X_i^\top \beta)}{\sum_{j \in R(t_{(k)})} \exp(X_j^\top \beta)}. \tag{PL-piece}$$

把这些条件概率在所有事件时刻相乘，就得到**部分似然**：

$$L_p(\beta) = \prod_{k=1}^{m} \frac{\exp(X_{i_k}^\top \beta)}{\sum_{j \in R(t_{(k)})} \exp(X_j^\top \beta)}, \qquad \ell_p(\beta) = \log L_p(\beta), \tag{PL}$$

其中 i_k 是在 $t_{(k)}$ 发生事件的个体索引。

部分似然是什么、不是什🄵

- 它**不是**完整似然（因为我们没有对 $\lambda_0(\cdot)$ 做参数化，也没有用到所有关于时间间隔的密度信息）。
- 它**是**一个针对 β 的有效“似然型目标函数”：最大化 $\ell_p(\beta)$ 给出 $\hat{\beta}$，并且在正则条件下具有良好的渐近性质（下面会给出推断形式）。

并列的另一种理解：条件 logit 式 (PL-piece) 的结构与 logistic 回归非常相似：在每个事件时刻，把风险集 $R(t_{(k)})$ 看作一个“匹配集合”，事件个体是“被选中”的那个，选择概率是 softmax。因此 Cox 的部分似然常被看作一类**条件 logit**。

7.4 得分方程与推断：Wald/LR/Score 的半参数版本

得分与信息 对数部分似然为

$$\ell_p(\beta) = \sum_{k=1}^{m} \left\{ X_{i_k}^\top \beta - \log \sum_{j \in R(t_{(k)})} \exp(X_j^\top \beta) \right\}.$$

其得分函数

$$U_p(\beta) = \frac{\partial \ell_p(\beta)}{\partial \beta} = \sum_{k=1}^{m} \left\{ X_{i_k} - \bar{X}(\beta, t_{(k)}) \right\},$$

其中

$$\bar{X}(\beta, t) = \frac{\sum_{j \in R(t)} X_j \exp(X_j^\top \beta)}{\sum_{j \in R(t)} \exp(X_j^\top \beta)}$$

是风险集内的加权平均协变量。$\hat{\beta}$ 满足 $U_p(\hat{\beta}) = 0$。

信息矩阵（负 Hessian）可写成风险集内的加权协方差的和：

$$I_p(\beta) = -\frac{\partial^2 \ell_p(\beta)}{\partial \beta \partial \beta^\top} = \sum_{k=1}^{m} \mathrm{Var}_\beta\big(X \mid R(t_{(k)})\big).$$

渐近正态与标准误 在正则条件下

$$\sqrt{n}\,(\hat{\beta} - \beta_0) \Rightarrow N\big(0,\ \mathcal{I}(\beta_0)^{-1}\big),$$

其中 $\mathcal{I}(\beta_0)$ 是半参数意义下的有效信息（efficient information），实践中常用 $I_p(\hat{\beta})^{-1}$ 作为方差估计（再配合稳健版本，见下）。

Wald / LR / Score 检验 检验 $H_0: R\beta = r$ 时，可以直接复用参数模型里的三类检验形式，但把 ℓ 换成 ℓ_p、把信息换成 I_p：

- Wald：$(R\hat{\beta} - r)^\top [R\widehat{\mathrm{Var}}(\hat{\beta})R^\top]^{-1}(R\hat{\beta} - r)$；
- LR：$2\{\ell_p(\hat{\beta}) - \ell_p(\tilde{\beta})\} \approx \chi^2$；
- Score：$U_p(\tilde{\beta})^\top I_p(\tilde{\beta})^{-1} U_p(\tilde{\beta}) \approx \chi^2$。

7.5 稳健性与“sandwich”：当删失或相关性更复杂

为什么需要稳健方差？ 部分似然推断依赖一些结构性条件（如独立删失、模型形式正确、观测独立等）。当数据存在聚类（同一医院/家庭）、重复事件、或删失机制更复杂时，即便 $\hat{\beta}$ 仍可作为某种“平均效应参数”的估计，标准误也常需要用稳健形式修正。

sandwich 方差（概念层面） 可用“信息的逆 × 得分的外积 × 信息的逆”的形式：

$$\widehat{\mathrm{Var}}(\hat{\beta}) = A^{-1} B A^{-1},$$

其中 A 通常来自（部分）似然的二阶导数，B 来自逐观测贡献的得分外积（可做聚类求和）。这与 GLM 中的稳健/聚类稳健推断在形式上完全一致。

7.6 半参数效率：为什么说 Cox 很“省假设但不太亏效率”

有效影响函数与效率界（只给直觉） 半参数模型里，β 的最优渐近方差由**效率界（efficiency bound）**决定，对应某个**有效影响函数（efficient influence function）**。部分似然估计在 Cox 模型的经典条件下是**半参数有效（semiparametrically efficient）**的：即在不指定 $\lambda_0(\cdot)$ 的前提下，仍能达到对 β 可实现的最小渐近方差。这解释了 Cox 模型在实践中的地位：*用较弱的分布假设换来很强的可用性，同时推断效率并不差。*

7.7 更多半参数范式：分位数回归与单指标模型（概览）

分位数回归 分位数回归直接建模

$$Q_\tau(Y \mid X) = X^\top \beta(\tau),$$

不需要指定误差分布;目标函数通常写成“check”损失的最小化,推断可用渐近理论或 bootstrap。它把“均值模型”替换为“分位数模型”，在异方差与长尾情形更稳健。

单指标/半参数链接 考虑

$$\mathbb{E}[Y \mid X] = h(X^\top \beta),$$

其中 $h(\cdot)$ 未知。它介于线性回归（h 为恒等）与一般非参数回归之间：只通过一个指数 $X^\top \beta$ 进入，从而显著缓解维数灾难，并给出可解释的方向参数 β。

本节小结：半参数与部分似然的三句话

- 半参数模型用 β 捕捉可解释结构，用无限维扰动 η 吸收难以可靠参数化的部分；
- Cox 模型的部分似然通过风险集条件化消去基线风险，直接对 β 做似然型推断；
- Wald/LR/Score 与稳健（sandwich）方差在半参数框架下依然成立，并把“少假设”与“可推断”结合起来。

8 模型评估

本节讨论统计建模中非常核心但常被低估的一件事：如何可靠地评估预测性能，并在此基础上做模型选择。注意，这里的“好”主要指泛化误差（*generalization error*）小，即在未来/新样本上预测得准；它与“系数是否显著”“解释是否因果”是两条不同的链路。如果你把同一份数据既当作“提出模型的灵感来源”，又当作“确认结论的证据来源”，就会出现系统性的乐观偏差：你会高估模型在新数据上的表现，也会高估显著性与可重复性。

8.1 训练–验证–测试与数据泄漏

基本切分与目标 给定样本 $(x_i, y_i)_{i=1}^n$，预测建模的典型流程是三段式切分：

- **训练集（train）**：拟合模型参数（例如回归系数、树的分裂、神经网络权重等）。
- **验证集（validation）**：用于调参、选模型、做特征选择、选正则化强度等“选择性决策”。
- **测试集（test）**：只在所有选择做完之后使用一次，用于给出最终、近似无偏的泛化性能估计。

把它写成形式：对一组候选模型族 $\{\mathcal{M}_\lambda\}$（λ 表示复杂度/超参数），训练集得到 $\hat{f}_\lambda$，验证集选择

$$\hat{\lambda} \in \arg\min_{\lambda} \ \widehat{R}_{\mathrm{val}}(\hat{f}_\lambda),$$

最后在测试集上报告

$$\widehat{R}_{\mathrm{test}}(\hat{f}_{\hat{\lambda}}).$$

关键点是：**测试集必须在选择过程之外**，否则“最终成绩”会被选择过程污染。

数据泄漏（data leakage）：最常见的失败模式 数据泄漏指任何形式的“用到了不该用的信息”，导致评估偏乐观。常见类型包括：

- **预处理泄漏：**在全数据上做标准化/缺失值插补/特征筛选，然后再切分。正确做法是：所有会学习数据分布的步骤（如均值方差、插补模型、*PCA* 方向、目标编码等）都必须只在训练折/训练集上拟合，并把同一变换应用到验证/测试上。
- **特征选择泄漏：**用全数据（含测试）挑选变量或确定交互项，再去报告测试误差或显著性。
- **时间序列与分组泄漏：**未来信息“流回”过去。比如按随机打乱切分时间序列，会让模型在训练中看到未来时期的分布特征；又如同一患者/同一用户的多条记录被分到训练与测试两边，会让测试看起来异常容易。这类问题应使用时间切分（*rolling/blocked split*）或分组切分（*grouped split*）。
- **目标泄漏（target leakage）：**某些特征本质上由 y 产生或在预测时不可得，例如“住院天数”去预测“是否重症”（住院天数是结果的一部分）、或使用事后才会知道的字段。

经验法则：把流程当作函数 把整个建模流程看成一个从数据到预测器的映射

$$\mathcal{A}:\ \mathcal{D}_{\text{train}} \mapsto \hat{f},$$

其中 $\mathcal{A}$ 包含了预处理、特征工程、调参、模型选择等全部步骤。那么“泛化评估”应该评估的是 *算法 $\mathcal{A}$ 的性能*，而不是某一次固定参数的 $\hat{f}$。这也是为什么外层验证（嵌套 CV）会在严格意义上更可靠（见后文）。

8.2 数据窥探（data snooping）与显著性膨胀：为何需要预注册与验证集

在回归实践中，一个常见风险是：研究者在同一份数据上反复尝试不同的变量组合、变换方式或模型设定，并最终“挑出最显著”的结果进行报告。这类过程常被称为数据窥探（data snooping）或“多重尝试”。其统计后果是：即使所有零假设都成立，也会因为反复搜索而更容易出现看似“显著”的发现，导致名义显著性水平（例如 5%）被系统性地打破。

直觉 如果我们对同一数据做很多次检验，总会有一两次 p 值很小；当研究流程把“挑出来的最小 p 值”当作一次性检验的结果来解释时，就会严重低估虚假发现的概率。

一个定量化的提醒（Bonferroni 直觉） 假设你做了 m 个相互独立的检验，每个检验在零假设成立时有 α 的概率误报。那么“至少出现一次误报”的概率近似为

$$\mathbb{P}(\text{至少一次假阳性}) = 1-(1-\alpha)^m \approx 1-e^{-m\alpha}.$$

当 m 稍大，这个概率会迅速上升。例如 $\alpha = 0.05, m = 20$ 时，$1-0.95^{20} \approx 0.64$：即使都为真零，也很可能“发现显著”。

方法论对应 在建模流程中，缓解数据窥探的关键手段不是更复杂的回归，而是更清晰的流程分离：

- 将探索（提出候选模型）与确认（评估与推断）分开；
- 使用验证集/交叉验证来评估泛化误差；
- 在可能时进行预注册或至少完整记录“尝试过哪些模型”（可重复、可审计）；
- 对多重比较进行校正，或用重抽样/外层验证来刻画选择带来的不确定性；
- 对“选择之后的推断”保持克制：选择过程改变了统计量的分布。

与因果推断的呼应 你在上一节强调“设计优先于模型”。在这里同样成立：如果结论来自“反复尝试后的最优展示”，那么再漂亮的模型形式也无法挽回推断的可信度。预测评估靠的是*流程隔离*；推断（特别是因果）靠的是*数据机制与识别假设*。

8.3 交叉验证：K 折、嵌套 CV 与不确定性

K 折交叉验证（K-fold CV） 当样本量不大、单次切分不稳定时，常用 K 折交叉验证估计泛化误差。把数据分成 K 个互不重叠的折 $\mathcal{I}_1, \ldots, \mathcal{I}_K$，第 k 次用 $\mathcal{I}_k$ 做验证，其余做训练，得到

误差估计

$$\widehat{R}_{\mathrm{CV}} = \frac{1}{K}\sum_{k=1}^{K} \widehat{R}\big(\hat{f}^{(-k)}, \mathcal{I}_k\big),$$

其中 $\hat{f}^{(-k)}$ 表示在除第 k 折外的数据上训练得到的模型。

K 怎么选？

- $K = 5$ 或 10 是常用折中：计算量适中，方差通常也不至于太大。
- $K = n$ 是留一法（LOOCV），偏差小但方差可能较大，且对某些模型计算昂贵。

当你在 CV 里做了“选择” 如果你用同一层 CV 同时调参/选模型并报告性能，会产生轻微到明显的乐观偏差，因为你在多个候选中挑了“验证误差最小者”。这时推荐用嵌套交叉验证。

嵌套交叉验证（Nested CV） 嵌套 CV 将“选择”和“评估”分到两层：

- **外层（outer loop）**：用于评估最终流程的泛化误差（相当于多次模拟测试集）。
- **内层（inner loop）**：在外层训练部分内部，再做 CV 用于调参/选模型。

外层每一折都重新做一次“完整选择流程”，因此更接近评估算法 $\mathcal{A}$ 的真实性能。

不确定性：别只报一个数 泛化误差估计本身也有抽样误差，尤其当数据量不大或数据异质性强时。实践中可以：

- 报告外层折误差的均值与标准差（或中位数与四分位距）；
- 对“预测误差差异”使用配对比较（同一折上比较模型 A 与 B）；
- 用重抽样（bootstrap）或重复 CV（repeated CV）观察稳定性；
- 强调：这些不确定性描述主要针对预测性能，不是传统意义的参数置信区间。

数据结构决定切分方式

- **时间序列：**使用时间阻断/滚动验证，保证训练只用过去预测未来。
- **分组数据（用户/病人/学校）：**使用 group CV，保证同组不跨训练测试。
- **空间数据：**考虑空间阻断（spatial blocking），避免邻近泄漏。

切分方式不是“实现细节”，而是你在定义什么叫“未来样本”。

8.4 指标：MSE/MAE/R^2 与校准的基本观念

回归：MSE 与 MAE 常用损失函数/指标包括：

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2, \qquad \mathrm{RMSE} = \sqrt{\mathrm{MSE}}, \qquad \mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|.$$

- MSE 对大误差更敏感（平方惩罚），适合你特别在乎大偏差的场景；
- MAE 更稳健，对异常值不那么敏感；
- 指标选择应来自业务/科学目标：你究竟在惩罚什么？

R^2：解释度不是泛化度 在线性回归中

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}.$$

注意：

- R^2 是相对于“只预测均值”的改进；
- 训练集上的 R^2 会随模型复杂度上升而不降；
- 更关键的是：**应报告测试/外层 CV 上的** R^2，否则它更像拟合程度而非预测能力。

分类：对数损失、AUC 与阈值依赖指标 若输出的是概率 $\hat{p}_i = \mathbb{P}(Y = 1 \mid x_i)$，对数损失（log loss / cross-entropy）为

$$\text{LogLoss} = -\frac{1}{n} \sum_{i=1}^{n} \left[y_i \log \hat{p}_i + (1 - y_i) \log(1 - \hat{p}_i) \right].$$

它直接评价概率预测质量；而准确率（accuracy）、精确率/召回率等依赖阈值，易受类别不平衡影响。AUC 衡量排序能力，但不直接保证概率可用。

校准（calibration）：概率预测必须“像概率” 预测“概率”时，除了区分度（discrimination），还要看校准：

> 在所有预测为 0.7 的样本中，真实发生比例是否约为 0.7？

校准差会导致决策层面系统性偏差：你以为风险是 20%，其实长期是 40%。常见做法是画可靠性图（reliability diagram），或报告校准误差（例如 ECE）。（本书不展开实现细节，但建议读者把“校准”视为概率预测的基本卫生标准。）

指标与目标：把“评价”写清楚 很多争论来自评价目标不明确：

- 你是在做点预测还是区间预测？
- 错一次代价对称吗？是否更怕漏报/误报？
- 未来分布会变吗（distribution shift）？如果会，测试集是否代表未来？

指标不是万能尺；它是你目标函数的投影。

8.5 信息准则：AIC/BIC（可选）

当候选模型是参数化似然模型（例如正态线性回归、GLM 等），信息准则提供一种无需显式验证集的复杂度惩罚思路。设模型似然为 $L(\hat{\theta})$、参数维度为 k、样本量 n，则

$$\mathrm{AIC} = -2\log L(\hat{\theta}) + 2k, \qquad \mathrm{BIC} = -2\log L(\hat{\theta}) + k\log n.$$

直觉上：第一项奖励拟合，第二项惩罚复杂度；BIC 的惩罚随 n 增大而更强，更偏向选择更简洁的模型。

AIC/BIC 与 CV 的关系

- 在很多经典条件下，AIC 可被理解为对“期望预测误差”的近似修正；
- BIC 更接近于某些贝叶斯模型比较的近似（对模型维度惩罚更强）；
- 但它们依赖于似然假设与正则条件；当模型错设、数据结构复杂、或你使用的是非似然型学习器（树、boosting 等），CV 往往更通用。

提醒：信息准则不解决数据窥探 如果你在同一份数据上不断试探特征工程、非线性变换、交互项、筛变量，再用 AIC/BIC “挑最小者”，本质仍是“选择后报告”，仍会偏乐观；此时应回到验证集/嵌套 CV/预注册等流程性手段。

8.6 模型选择之后你还能做什么推断？

本节的主题是预测，但读者常会自然追问：选完模型后，能否对系数做显著性检验、置信区间、解释变量重要性？一个原则性的回答是：

> **如果模型是通过看数据而被选择出来的，那么把后续推断当作“预先指定模型下的常规推断”通常是不可信的。**

原因不在于“线性回归不好”，而在于：你观测到的统计量分布已经被选择机制改变。要么你把“选择”也纳入推断框架（选择后推断、重抽样刻画不确定性等），要么你把推断目标从“参数”转回“预测性能”。

本章小结

- 预测评估的核心是**隔离**：训练用于拟合，验证用于选择，测试用于最终评估。
- 数据泄漏与数据窥探是“流程层面”的错误，会系统性乐观。
- 交叉验证是评估泛化误差的常用工具；若涉及调参选模，嵌套 CV 更可靠。
- 指标必须与任务匹配；概率预测还必须检查校准。
- AIC/BIC 是似然模型下的复杂度惩罚工具，但不替代良好的验证流程。

9 Bootstrap：用重抽样做推断

Bootstrap 的出发点很朴素：我们关心统计量 $T_n = T(Z_1, \dots, Z_n)$ 的抽样分布，但真实总体分布 F（以及可能的依赖/抽样设计）未知；于是用数据本身诱导的近似世界来替代真实世界，在计算机上反复生成“伪样本”并重算统计量，从而得到标准误与置信区间。在实践中，Bootstrap 最常用也最关键的功能是：**把复杂估计量的（渐近）方差/极限协方差数值化**，特别是在解析方差推导困难、稳健/长程方差形式复杂、或推断对象是两步/加权/非线性估计量时。

9.1 非参数 Bootstrap

定义 9.1 (非参数 Bootstrap)**.** 给定样本 $Z_1, \dots, Z_n \overset{\text{i.i.d.}}{\sim} F$，经验分布

$$\hat{F}_n := \frac{1}{n} \sum_{i=1}^{n} \delta_{Z_i}.$$

非参数 Bootstrap 是指在条件于观测数据 $(Z_1, \dots, Z_n)$ 下，从 $\hat{F}_n$ 有放回抽样得到 $Z_1^*, \dots, Z_n^* \overset{\text{i.i.d.}}{\sim} \hat{F}_n$，并用重复计算得到的统计量分布来近似原统计量 $T_n = T(Z_1, \dots, Z_n)$ 的抽样分布。

核心思想：用“经验分布”替代“真分布”。 若 T_n 的抽样波动来自未知总体分布 F，那么把 F 换成可计算的 $\hat{F}_n$，就得到一个可模拟的近似世界；在这个世界里反复抽样并重算 T，得到条件分布

$$\mathcal{L}^*(T_n^*) := \mathcal{L}\big(T(Z_1^*, \dots, Z_n^*) \,\big|\, Z_1, \dots, Z_n\big),$$

用它去逼近真实抽样分布 $\mathcal{L}(T_n)$（或其适当标准化后的版本）。

算法（Monte Carlo 近似）。 给定复本数 B（如 $B = 1000, 2000$），对 $b = 1, \dots, B$：

1. 从 $\{Z_i\}_{i=1}^n$ 等概率有放回抽样得到 $Z_1^{*(b)}, \dots, Z_n^{*(b)}$；
2. 计算统计量复本
$$T^{*(b)} := T\big(Z_1^{*(b)}, \dots, Z_n^{*(b)}\big).$$

用经验分布 $\{T^{*(b)}\}_{b=1}^B$ 近似 $\mathcal{L}^*(T_n^*)$，进而近似 $\mathcal{L}(T_n)$。

与“参数化 Bootstrap”的对比（点到为止）。 **参数化 Bootstrap** 是先拟合一个参数模型 $F_{\hat{\theta}}$，再从 $F_{\hat{\theta}}$ 抽样。非参数 Bootstrap 则不假设参数模型，直接用 $\hat{F}_n$ 近似 F；优点是稳健、实现简单，代价是它对统计量的平滑性/正则性更敏感。

何时有效：一个常用的充分条件（直觉版）。 把目标写成标准化形式，例如

$$\sqrt{n}\big(T_n - \theta\big) \;\Rightarrow\; G,$$

并希望 Bootstrap 满足

$$\mathcal{L}^*\big(\sqrt{n}(T_n^* - T_n)\big) \;\Rightarrow^{P}\; G,$$

即条件分布在概率意义下收敛到同一极限。直觉上，若统计量是“足够平滑”的函数（例如许多 $\sqrt{n}$-渐近线性/影响函数型统计量），非参数 Bootstrap 往往有效；反之对于**不连续/非平滑/极**

值型统计量，Bootstrap 可能失效或需要修正（例如 m-out-of-n bootstrap、subsampling、平滑 bootstrap 等）。

常见失效情形（提醒即可）。

- **非光滑/非正则**：极值、阈值、某些模型选择后的统计量；
- **边界/弱识别/不规则**：参数在边界、识别弱、分布有原子点或密度不连续；
- **依赖或复杂抽样**：时间序列/聚类/调查设计下不能直接做逐点 i.i.d. 重抽样（见 §9.3）。

9.2 标准误与置信区间

Bootstrap 输出的核心：对（渐近）方差/极限协方差的数值化。 很多推断最终都可归结为

$$\sqrt{n}\big(T_n-\theta\big) \;\Rightarrow\; N(0,V) \quad\Longrightarrow\quad \mathrm{Var}(T_n)\approx \frac{1}{n}V,$$

其中 V 是由数据生成机制决定的**渐近方差**（或更一般的极限协方差）。当 V 的解析形式很复杂（非线性估计、两步估计、加权、稳健/长程方差、复杂抽样等）时，Bootstrap 的优势在于：它用重抽样在计算机上直接近似 V，从而给出标准误与区间所需的方差尺度，而不必手推一套易错的方差公式。

Bootstrap 标准误（standard error）。 令点估计为 T_n。用复本 $\{T^{*(b)}\}_{b=1}^{B}$ 的样本标准差估计 $\mathrm{sd}^*(T_n^*)\approx \mathrm{sd}(T_n)$：

$$\widehat{\mathrm{se}}_{\mathrm{boot}}(T_n) := \left\{\frac{1}{B-1}\sum_{b=1}^{B}\Big(T^{*(b)}-\bar T^*\Big)^2\right\}^{1/2}, \qquad \bar T^* := \frac{1}{B}\sum_{b=1}^{B}T^{*(b)}.$$

Bootstrap 估计偏差（可选）。

$$\widehat{\mathrm{bias}}_{\mathrm{boot}}(T_n) := \bar T^* - T_n.$$

偏差校正点估计可写为 $T_n^{\mathrm{bc}} := T_n - \widehat{\mathrm{bias}}_{\mathrm{boot}}$，但偏差校正未必总是更好（可能增大方差），此处不展开。

置信区间：从“分位数”到“反演”。 Bootstrap 给出的是 $\mathcal{L}^*(T_n^*)$ 或 $\mathcal{L}^*(T_n^*-T_n)$ 的近似，据此构造区间的思路有两类：

- **分位数法（percentile）**：直接用 T_n^* 的分位数；
- **反演法（basic / pivotal / studentized）**：先近似误差 $T_n-\theta$ 的分布，再反推出 θ 的区间。

Percentile 区间（最常用、最简单）。 记 q_α^* 为 $\{T^{*(b)}\}_{b=1}^{B}$ 的经验 α 分位数，则

$$\mathrm{CI}_{1-\alpha}^{\mathrm{perc}} = \Big[q_{\alpha/2}^*,\; q_{1-\alpha/2}^*\Big].$$

Basic 区间（反演/对称于 T_n）。 设 r^*_α 为 $\{T^{*(b)} - T_n\}_{b=1}^B$ 的经验 α 分位数，则

$$\mathrm{CI}^{\mathrm{basic}}_{1-\alpha} = \Big[T_n - r^*_{1-\alpha/2},\ T_n - r^*_{\alpha/2}\Big] = \Big[2T_n - q^*_{1-\alpha/2},\ 2T_n - q^*_{\alpha/2}\Big].$$

Studentized / bootstrap-t（精度更高但更贵）。 若能为每个复本计算标准误 $\widehat{\mathrm{se}}^{*(b)}$，可构造

$$Z^{*(b)} := \frac{T^{*(b)} - T_n}{\widehat{\mathrm{se}}^{*(b)}}.$$

取 z^*_α 为 $\{Z^{*(b)}\}$ 的经验 α 分位数，则

$$\mathrm{CI}^{t}_{1-\alpha} = \Big[T_n - z^*_{1-\alpha/2}\ \widehat{\mathrm{se}}_{\mathrm{boot}}(T_n),\ T_n - z^*_{\alpha/2}\ \widehat{\mathrm{se}}_{\mathrm{boot}}(T_n)\Big].$$

BCa（点到为止）。 BCa（bias-corrected and accelerated）通过偏差校正 z_0 与加速参数 a 修正 percentile 的分位点水平；小样本与偏斜情形下常更稳健，但公式与实现细节较多，此处略。

实践建议（简短版）。

- 做区间通常至少 $B \geq 1000$，更稳妥可取 2000 或更高。
- 先看 $\{T^{*(b)}\}$ 的分布形状；明显偏斜/离群时 percentile/basic/BCa 的差异会变大。
- 若数据存在依赖或复杂抽样结构，必须调整 resampling 机制（见下节），否则标准误与区间可能严重失真。

9.3 相关数据与复杂抽样：哪些时候必须调整 Bootstrap

Freedman 的核心提醒可以改写成一句操作性原则：

Bootstrap 不是“重抽样”本身，而是“复现你相信的抽样/生成机制”。

在大样本推断里，我们真正依赖的是某个极限方差/协方差 V（可能是长程方差、簇稳健方差、设计方差等）：

$$\sqrt{n}\big(T_n - \theta\big) \ \Rightarrow\ N(0, V).$$

Bootstrap 的任务就是在计算机上近似这个 V；因此一旦你用错了重抽样机制，你近似的就不再是“真实世界对应的 V”，而是某个*错误世界*的 V_{wrong}，典型后果是标准误偏小、区间覆盖率偏低。

什么时候必须调整：触发条件 ⇒ 对应处方

1. **聚类/分组数据（簇内相关、簇间近似独立）** ⇒ *cluster bootstrap*（以簇为单位重抽），或 *cluster wild bootstrap*（回归且异方差时常更合适）。
   典型场景：学校/村庄/医院/企业/家庭聚类；cluster randomization；多期追踪里“个体”为簇。

2. **时间序列/面板的时间依赖（自相关、波动聚集）** ⇒ *block bootstrap*（MBB/circular/stationary）或 *sieve/model-based bootstrap*（拟合 AR(p) 等再递推生成）。

要点：这里要保留的是相关结构与长程方差（long-run variance），逐点 i.i.d. 会把相关打碎。

3. **空间相关（邻近相关、区域聚集）** ⇒ *spatial block bootstrap*（按空间块/区域重抽）或按更高层级地理单元聚类重抽。
   要点：抽样单位应覆盖主要相关尺度，否则方差仍会被低估。

4. **回归中强异方差（固定设计，关心系数标准误）** ⇒ *wild bootstrap*；若同时存在聚类，则用 *cluster wild bootstrap*。
   直觉：残差 i.i.d. 重抽相当于把异方差“抹平”，而 wild bootstrap 试图保留条件方差结构。

5. **复杂抽样设计（survey：分层/整群/多阶段/PPS/权重）** ⇒ *design-based / weighted bootstrap*（按 PSU/抽样单位重抽，并保持权重与分层结构）。
   要点：此时 V 是设计方差，必须复现抽样设计而非把记录当作 SRS。

6. **缺失/删失/进入样本的机制是关键（MNAR、依赖 Y/X 的选择）** ⇒ Bootstrap 需同时复现“观测机制”（例如对缺失过程建模、或在重抽样时保持两阶段抽样/选择结构）。否则你得到的区间往往对应的是“条件于被观测者”的目标，而非原始总体目标。

7. **统计量非平滑或包含选择（模型选择、阈值、极值、断点/带宽选择）** ⇒ 常规 bootstrap 可能失效；需考虑 *m-out-of-n bootstrap*、*subsampling*、或直接转向专门的后选择推断/重抽样修正方案。

把问题“拆成三类随机性”，通常就能选对 bootstrap

1. **抽样随机性（sampling design）**：你如何从总体进入样本？（分层、整群、权重）
2. **依赖结构（dependence）**：样本内部如何相关？（时间、空间、簇、网络）
3. **误差结构（error structure）**：给定协变量后噪声如何变化？（异方差、相关、重尾）

逐点 i.i.d. 非参数 bootstrap 默认把三者都简化为“观测点独立可交换”。当你发现自己需要 cluster-robust、Newey–West、设计方差、或任何“稳健/长程”修正时，几乎总意味着：你在乎的正是一个非 *i.i.d.* 的极限方差 V，此时 bootstrap 也必须随之调整，否则就不是在估计同一个 V。

10 EM 算法：从 E+M 到变分、采样与对抗的统一图景

很多人第一次接触 EM（Expectation–Maximization）时，会把它记成一句口号：

$$\text{EM} = \text{E-step} + \text{M-step}.$$

但 EM 更深的本质是：**在含隐变量/缺失数据的似然优化中，构造一个对数似然的局部下界（lower bound），再在下界上做坐标上升（coordinate ascent）**。沿着这条主线，Hard EM、广义 EM（GEM）、变分 EM（VBEM）、Weiss–something（WS）式的“松弛 + 优化”，乃至 Gibbs 抽样，都可以放进同一个“下界/散度”框架里理解。

10.1 第一层：EM 就是 E + M

隐变量模型与不完全数据似然 设观测数据为 x，隐变量为 z，参数为 θ。联合分布写作

$$p_\theta(x,z) = p_\theta(z)\,p_\theta(x \mid z) \qquad \text{或更一般地} \qquad p_\theta(x,z) = p_\theta(x \mid z,\theta)\,p_\theta(z \mid \theta).$$

我们能直接写出**完全数据**（complete data）对数似然

$$\log p_\theta(x,z),$$

但真实世界只观测到 x，因此要最大化**不完全数据**（incomplete data）对数似然

$$\ell(\theta) = \log p_\theta(x) = \log \sum_z p_\theta(x,z) \quad (\text{离散}z), \qquad \text{或} \qquad \ell(\theta) = \log \int p_\theta(x,z)\,dz \quad (\text{连续}z).$$

困难来自：log 与求和/积分不交换，导致 $\log\sum$ / $\log\int$ 的直接优化通常没有闭式形式，而且往往是非凸的。

两个关键对象：后验 $p_\theta(z \mid x)$ 与证据下界（ELBO） EM 的出发点是：既然 z 不可见，那就用一个分布 $q(z)$ 来“填补”缺失信息。对任意 $q(z)$，有恒等分解

$$\log p_\theta(x) = \underbrace{\mathbb{E}_q[\log p_\theta(x,z)] - \mathbb{E}_q[\log q(z)]}_{\mathcal{L}(q,\theta)\ (\text{ELBO})} + \underbrace{KL\big(q(z) \,\|\, p_\theta(z \mid x)\big)}_{\geq 0}. \tag{1}$$

其中

$$\mathcal{L}(q,\theta) = \mathbb{E}_q[\log p_\theta(x,z)] + H(q)$$

称为**证据下界**（evidence lower bound, ELBO），$H(q) = -\mathbb{E}_q \log q$ 是熵。由于 KL 非负，(1) 立刻给出

$$\log p_\theta(x) \;\geq\; \mathcal{L}(q,\theta),$$

并且当且仅当 $q(z) = p_\theta(z \mid x)$ 时取等号。

EM 的两步：坐标上升（coordinate ascent）的视角 EM 可以理解为对 (1) 中的两变量 (q,θ) 做交替优化：

- **E-step（更新 q）：** 固定 $\theta = \theta^{(t)}$，最大化 $\mathcal{L}(q,\theta^{(t)})$ 关于 q。由 (1) 可知这等价于最小化 $(q \,\|\, p_{\theta^{(t)}}(z \mid x))$，因此最优解为

  $$q^{(t)}(z) = p_{\theta^{(t)}}(z \mid x).$$

  这一步把 ELBO 的“松弛”收紧到与当前模型后验一致（KL 变为 0）。

- **M-step（更新 θ）：** 固定 $q = q^{(t)}$，最大化 ELBO 关于 θ：

  $$\theta^{(t+1)} \in \arg\max_\theta \mathcal{L}(q^{(t)},\theta).$$

由于 $H(q^{(t)})$ 与 θ 无关，M-step 等价于最大化

$$Q(\theta \mid \theta^{(t)}) = \mathbb{E}_{z\sim q^{(t)}}[\log p_\theta(x,z)],$$

即“完全数据对数似然”的后验期望。

因此，EM 既可以看作“E + M”，也可以看作**对 ELBO 的坐标上升算法**。

为什么 EM 会单调提高不完全数据似然？（基本性质） EM 的核心保证是：若每步都做精确的 E-step 且 M-step 不下降，则

$$\ell(\theta^{(t+1)}) \;\geq\; \ell(\theta^{(t)}).$$

证明思路非常短：由 (1)，

$$\ell(\theta) = \log p_\theta(x) \;\geq\; \mathcal{L}(q,\theta).$$

E-step 取 $q^{(t)} = p_{\theta^{(t)}}(z \mid x)$ 使得等号成立：

$$\ell(\theta^{(t)}) = \mathcal{L}(q^{(t)},\theta^{(t)}).$$

而 M-step 使 ELBO 不下降：

$$\mathcal{L}(q^{(t)},\theta^{(t+1)}) \geq \mathcal{L}(q^{(t)},\theta^{(t)}).$$

再用 $\ell(\theta^{(t+1)}) \geq \mathcal{L}(q^{(t)},\theta^{(t+1)})$ 夹起来，即得单调性。

从 Jensen 不等式看 EM（经典教科书推导） 同一件事也可用 Jensen 表达。对任意分布 q，

$$\log p_\theta(x) = \log \sum_z q(z)\frac{p_\theta(x,z)}{q(z)} \;\geq\; \sum_z q(z)\log\frac{p_\theta(x,z)}{q(z)} = \mathcal{L}(q,\theta).$$

这正是 ELBO 下界。E-step 选取使下界最紧的 q（即后验），M-step 最大化该下界。

常见术语与直观

- **complete data / incomplete data：**是否把隐变量 z 也当作“观测到的”。
- **E-step：**不是“求一个数”，而是求一个分布（或其关键期望），即 $p_{\theta^{(t)}}(z \mid x)$。
- **M-step：**在“补全”后的世界里做极大似然；因为用的是期望，所以叫期望最大化。
- **软分配 vs 硬分配：**E-step 给的是后验权重（soft assignment）；若把后验退化为指派到最大后验的单点，就得到类似 k-means 的硬分配算法（但不再是标准 EM）。

多样本形式（更常用的写法） 若观测为 $x_{1:n}$，隐变量为 $z_{1:n}$ 且在模型下条件独立，则

$$Q(\theta \mid \theta^{(t)}) = \sum_{i=1}^{n} \mathbb{E}_{z_i\sim p_{\theta^{(t)}}(z_i\mid x_i)}\big[\log p_\theta(x_i,z_i)\big],$$

E-step 与 M-step 都可逐样本分解（这也是混合模型 EM 可计算的关键）。

重要提醒：EM 解决了什么、没解决什么？

- EM 保证 **似然单调不降**，但一般只保证收敛到**驻点/局部最优**（非凸问题常见）。
- M-step 有时无法闭式解，则可以用 **Generalized EM (GEM)**：只要让 Q 上升即可。
- E-step 若后验不可得，可用 **变分 EM**（限制 q 的族）或 **Monte Carlo EM**（采样近似期望）。

10.2 第二层：K-Means 是一种 Hard EM

从高斯混合到 Hard assignment 考虑 K-component 高斯混合模型（GMM）：

$$z_i \sim \mathrm{Cat}(\pi_1, \dots, \pi_K), \qquad x_i \mid (z_i = k) \sim N(\mu_k, \Sigma_k), \qquad i = 1, \dots, n.$$

完全数据对数似然为

$$\log p_\theta(x_{1:n}, z_{1:n}) = \sum_{i=1}^{n} \sum_{k=1}^{K} \mathbf{1}\{z_i = k\} \Big(\log \pi_k + \log \varphi(x_i; \mu_k, \Sigma_k) \Big),$$

其中 $\varphi(\cdot; \mu, \Sigma)$ 为高斯密度。

标准 EM 的 E-step 计算**软责任度**（posterior responsibilities）

$$r_{ik}^{(t)} := p_{\theta^{(t)}}(z_i = k \mid x_i) = \frac{\pi_k^{(t)} \varphi(x_i; \mu_k^{(t)}, \Sigma_k^{(t)})}{\sum_{\ell=1}^{K} \pi_\ell^{(t)} \varphi(x_i; \mu_\ell^{(t)}, \Sigma_\ell^{(t)})}, \qquad \sum_{k=1}^{K} r_{ik}^{(t)} = 1.$$

Hard EM 的思想是把后验“硬化”：用一个确定的簇标签替代 r_{ik} 的分布信息，

$$\hat{z}_i^{(t)} = \arg\max_k r_{ik}^{(t)}, \qquad \hat{r}_{ik}^{(t)} := \mathbf{1}\{\hat{z}_i^{(t)} = k\} \in \{0, 1\}.$$

然后在 M-step 中，把软权重 $r_{ik}^{(t)}$ 替换成硬权重 $\hat{r}_{ik}^{(t)}$，更新参数

$$\theta^{(t+1)} \in \arg\max_\theta \sum_{i=1}^{n} \sum_{k=1}^{K} \hat{r}_{ik}^{(t)} \Big(\log \pi_k + \log \varphi(x_i; \mu_k, \Sigma_k) \Big).$$

这一步就是“把每个样本只交给一个簇”之后的完全数据极大似然。

把模型进一步简化：退化到 K-Means 的平方误差目标 为了看清 K-Means 的结构，我们进一步施加 GMM 的典型约束：

$$\Sigma_k = \sigma^2 I_d \ \text{（同方差、球形）}, \qquad \pi_k = \frac{1}{K} \ \text{（等权，可选）}.$$

在该约束下，

$$\log \varphi(x_i; \mu_k, \sigma^2 I_d) = -\frac{d}{2} \log(2\pi\sigma^2) \ - \ \frac{1}{2\sigma^2} \|x_i - \mu_k\|_2^2.$$

代回 Hard EM 的 M-step 目标，去掉与 $(\mu_1,\dots,\mu_K)$ 无关的常数项后，最大化等价于最小化

$$\sum_{i=1}^{n}\sum_{k=1}^{K}\hat{r}_{ik}^{(t)}\|x_i-\mu_k\|_2^2.$$

因此在固定硬分配 $\hat{r}^{(t)}$ 时，

$$\mu_k^{(t+1)}=\arg\min_{\mu}\ \sum_{i=1}^{n}\hat{r}_{ik}^{(t)}\|x_i-\mu\|_2^2=\frac{\sum_{i=1}^{n}\hat{r}_{ik}^{(t)}x_i}{\sum_{i=1}^{n}\hat{r}_{ik}^{(t)}}\quad\text{(簇均值)}.$$

而在固定中心 $\mu^{(t)}$ 时，Hard E-step 的 $\arg\max_k r_{ik}^{(t)}$ 在上述约束下等价于“选最近中心”：

$$\hat{z}_i^{(t)}=\arg\min_{k\in\{1,\dots,K\}}\ \|x_i-\mu_k^{(t)}\|_2^2,\qquad \hat{r}_{ik}^{(t)}=\mathbf{1}\Big\{k=\arg\min_{\ell}\|x_i-\mu_\ell^{(t)}\|_2^2\Big\}.$$

把两步合在一起就得到 K-Means 的经典迭代：

$$\textbf{(Assign)}\ \hat{z}_i\leftarrow\arg\min_k\|x_i-\mu_k\|^2,\qquad \textbf{(Update)}\ \mu_k\leftarrow\frac{1}{|\{i:\hat{z}_i=k\}|}\sum_{i:\hat{z}_i=k}x_i.$$

目标函数与“单调下降” K-Means 的显式目标是

$$\min_{\mu_1,\dots,\mu_K}\ \min_{z_1,\dots,z_n\in\{1,\dots,K\}}\sum_{i=1}^{n}\|x_i-\mu_{z_i}\|_2^2,$$

也常写成

$$\min_{\mu}\ \sum_{i=1}^{n}\min_{k\le K}\|x_i-\mu_k\|_2^2.$$

两步迭代对应对该目标的交替最小化：

- 固定 μ，选择最近中心使目标不增；
- 固定 z，令 μ_k 为簇均值使目标不增（这是平方损失的最小二乘解）。

因此 K-Means 的目标值单调不增并在有限步内达到一个局部最优（或停在一个固定点）。

Hard EM 与标准 EM：差别在哪里？

- **信息保留：** 标准 EM 用 $r_{ik}\in[0,1]$ 保留“不确定性”，Hard EM 直接把后验压成 0/1；当簇有重叠、边界样本很多时，硬化会损失关键信息。
- **优化目标：** 标准 EM 是在最大化不完全数据对数似然（或其 ELBO）上做坐标上升；Hard EM 更接近最大化分类似然（classification likelihood）或在一个“离散化的下界”上做交替优化，一般不再保证 $\log p_\theta(x)$ 单调上升。
- **计算与稳定性：** Hard EM/K-Means 每步更便宜、实现更简单，但对初始化更敏感，更容易陷入差的局部解（比如空簇、塌缩到坏划分等）。

直觉总结：K-Means = 简化 GMM + 硬后验 + 交替最小二乘　K-Means 可以被视为：在球形同方差 GMM 里，把 E-step 的“软责任度”硬化成“最近中心”的指派，并在 M-step 上把参数更新退化为簇均值更新。它牺牲了概率模型的细腻性，换来了速度与可扩展性；代价则是更脆弱的局部最优与更强的初始化依赖。

10.3 第三层：MM (minorize–maximize) 视角下的 EM

EM 的“单调上升”到底来自哪里？ EM 最核心的性质是：在温和的可积性/可微性条件下，每次迭代都不会降低不完全数据对数似然

$$\ell(\theta) := \log p_\theta(x).$$

更一般地，EM 可以被看成一类 **MM（minorize–maximize）算法**的特例：在当前点构造一个下界（*minorizer*），然后最大化这个下界，从而保证目标函数单调上升。

MM 的一般模板　给定要最大化的目标 $F(\theta)$。若我们能构造一个 surrogate（代理函数）$G(\theta \mid \theta^{(t)})$ 满足

$$\text{(i)} \quad G(\theta \mid \theta^{(t)}) \le F(\theta) \ \ \forall \theta, \qquad \text{(ii)} \quad G(\theta^{(t)} \mid \theta^{(t)}) = F(\theta^{(t)}),$$

则称 G **minorize**（下界）F。此时令

$$\theta^{(t+1)} \in \arg\max_\theta \ G(\theta \mid \theta^{(t)})$$

即可得到单调性：

$$F(\theta^{(t+1)}) \ \ge \ G(\theta^{(t+1)} \mid \theta^{(t)}) \ \ge \ G(\theta^{(t)} \mid \theta^{(t)}) = F(\theta^{(t)}).$$

这就是 MM 的“上升链”。

把 EM 写成 MM：关键是 Jensen 不等式　对隐变量模型，

$$\ell(\theta) = \log p_\theta(x) = \log \int p_\theta(x, z)\, dz.$$

取任意分布 $q(z)$（满足 $q(z) > 0$ 的地方 $p_\theta(x, z)$ 可积），写成

$$\ell(\theta) = \log \int q(z) \frac{p_\theta(x, z)}{q(z)}\, dz \ \ge \ \int q(z) \log \frac{p_\theta(x, z)}{q(z)}\, dz,$$

其中不等号来自 Jensen：$\log \mathbb{E}_q[U] \ge \mathbb{E}_q[\log U]$。定义

$$\mathcal{L}(\theta, q) := \int q(z) \log p_\theta(x, z)\, dz - \int q(z) \log q(z)\, dz = \mathbb{E}_q[\log p_\theta(x, z)] + H(q),$$

则对任意 q 都有

$$\ell(\theta) \ge \mathcal{L}(\theta, q).$$

并且当 $q(z) = p_{\theta^{(t)}}(z \mid x)$ 时，该下界在 $\theta = \theta^{(t)}$ 处是“贴住”的：

$$\ell(\theta^{(t)}) = \mathcal{L}(\theta^{(t)}, q^{(t)}), \qquad q^{(t)}(z) := p_{\theta^{(t)}}(z \mid x).$$

因此

$$G(\theta \mid \theta^{(t)}) := \mathcal{L}(\theta, q^{(t)})$$

正是 $\ell(\theta)$ 在当前点 $\theta^{(t)}$ 处的一个 minorizer。

E-step/M-step = 构造下界/最大化下界

- **E-step（tighten 下界）**：在固定 $\theta = \theta^{(t)}$ 时，最大化 $\mathcal{L}(\theta^{(t)}, q)$ 的 q 解为

  $$q^{(t)}(z) = p_{\theta^{(t)}}(z \mid x).$$

  这一步把下界“拉紧”，使其在当前点与 ℓ 相切。

- **M-step（maximize 下界）**：固定 $q = q^{(t)}$ 后，更新

  $$\theta^{(t+1)} \in \arg\max_{\theta} \ \mathcal{L}(\theta, q^{(t)}) = \arg\max_{\theta} \ \mathbb{E}_{q^{(t)}}[\log p_\theta(x, z)],$$

  因为 $H(q^{(t)})$ 与 θ 无关。这就回到了标准的 $Q(\theta \mid \theta^{(t)})$ 形式。

用 KL 散度把单调性写得更“结构化” 把 $\ell(\theta)$ 与下界的差写成 KL：

$$\ell(\theta) - \mathcal{L}(\theta, q) = KL\big(q(z) \parallel p_\theta(z \mid x)\big) \ \geq \ 0.$$

因此

$$\ell(\theta) = \mathcal{L}(\theta, q) + KL\big(q \| p_\theta(\cdot \mid x)\big).$$

当 $q = q^{(t)} = p_{\theta^{(t)}}(\cdot \mid x)$ 时，

$$\ell(\theta) = \mathcal{L}(\theta, q^{(t)}) + KL\big(p_{\theta^{(t)}}(\cdot \mid x) \parallel p_\theta(\cdot \mid x)\big),$$

这清楚地说明了：

- M-step 最大化 $\mathcal{L}(\theta, q^{(t)})$ 会推动 $\ell(\theta)$ 上升；
- 若用 *generalized EM* 只做到让 $\mathcal{L}$ 上升（不必全局最大化），依然能保证 ℓ 不下降。

一句话总结：EM 是 MM 的一个“Jensen 下界”实例 EM 的 E-step 构造一个在当前点贴住的 Jensen 下界（minorizer），M-step 则最大化该下界。因此 EM 的经典单调性并非“巧合”，而是 MM 机制的直接产物。

10.4 第四层：EM 的一个特例是 VBEM

为什么需要 VB：E-step 不可解时怎么办？ 标准 EM 的核心是：在每一步用当前参数 $\theta^{(t)}$ 计算精确后验

$$p_{\theta^{(t)}}(z \mid x) = \frac{p_{\theta^{(t)}}(x, z)}{p_{\theta^{(t)}}(x)},$$

并令 $q^{(t)}(z) = p_{\theta^{(t)}}(z \mid x)$（E-step）。但在高维隐变量、复杂图结构（含环）、或深度生成模型中，后验往往不可解析、也难以在每次迭代中精确积分/求和或精确采样，从而“精确 E-step”不 tractable。

VBEM（Variational Bayes EM）通过把后验近似限制在一个可计算的变分族 $\mathcal{Q}$ 内，使 E-step 变成一个可解的优化问题。

ELBO：把对数似然拆成“下界 + KL” 对任意分布 $q(z)$，都有恒等式

$$\log p_\theta(x) = \mathcal{L}(q, \theta) \; + \big(q(z) \,\|\, p_\theta(z \mid x)\big),$$

其中

$$\mathcal{L}(q, \theta) := \mathbb{E}_q[\log p_\theta(x, z)] - \mathbb{E}_q[\log q(z)] = \mathbb{E}_q[\log p_\theta(x, z)] + H(q)$$

称为 **ELBO**（evidence lower bound），$H(q)$ 是熵。因此

$$\log p_\theta(x) \;\geq\; \mathcal{L}(q, \theta),$$

且等号成立当且仅当 $q(z) = p_\theta(z \mid x)$（即 KL 为 0）。

VBEM 的两步：在受限变分族内做“坐标上升” VBEM 选择一个可计算的变分族 $\mathcal{Q}$（常见为均值场分解）

$$\mathcal{Q} = \Big\{ q(z) = \prod_{j=1}^{J} q_j(z_j) \Big\},$$

并进行如下迭代：

$$\textbf{(VB-E step)} \quad q^{(t)} \in \arg\max_{q \in \mathcal{Q}} \; \mathcal{L}(q, \theta^{(t)}), \qquad \textbf{(M step)} \quad \theta^{(t+1)} \in \arg\max_{\theta} \; \mathcal{L}(q^{(t)}, \theta).$$

与标准 EM 的区别是：E-step 不再令 $q^{(t)}$ 等于真后验，而是在 $\mathcal{Q}$ 内找使 ELBO 最大的近似后验。因此一般有

$$\big(q^{(t)} \| p_{\theta^{(t)}}(z \mid x)\big) > 0,$$

下界不再贴住真对数似然，但每一步都可计算。

等价形式：VB-E step 是对真后验的 KL 投影 由“下界 + KL”分解可知，对固定的 θ，

$$q^\star(\theta) \in \arg\max_{q \in \mathcal{Q}} \mathcal{L}(q, \theta) \quad \Longleftrightarrow \quad q^\star(\theta) \in \arg\min_{q \in \mathcal{Q}} \big(q \| p_\theta(\cdot \mid x)\big).$$

也就是说，VBEM 的 E-step 在做：把不可解的真后验 $p_\theta(z \mid x)$ 投影到一个可解的近似族 $\mathcal{Q}$。

均值场的坐标更新公式（可作为“记忆模板”） 若 $\mathcal{Q}$ 采取均值场形式 $q(z)=\prod_{j=1}^{J} q_j(z_j)$，则在固定 θ 与其它因子 q_{-j} 时，最优的 q_j 满足

$$\log q_j^\star(z_j)=\mathbb{E}_{q_{-j}}\big[\log p_\theta(x,z)\big]+\text{const},\qquad q_j^\star(z_j)\ \propto\ \exp\Big(\mathbb{E}_{q_{-j}}[\log p_\theta(x,z)]\Big),$$

其中 $q_{-j}:=\prod_{\ell\neq j} q_\ell$。因此 VB-E step 往往通过反复更新各个 q_j（坐标上升）来实现。

单调性：上升的是 ELBO（不一定是真似然） 只要每一步都精确最大化对应的坐标块（q 或 θ），就有

$$\mathcal{L}(q^{(t+1)},\theta^{(t+1)})\ \geq\ \mathcal{L}(q^{(t)},\theta^{(t)}).$$

因此 VBEM 保证 ELBO 单调不降；但由于 KL 项会随 θ 变化，$\log p_\theta(x)$ 本身不一定严格单调。

一句话总结 VBEM = EM 的“受限 E-step”版本：用可计算的变分近似后验替代精确后验，从而可计算地最大化 ELBO。

10.5 第五层：EM 的另一个特例是 Wake–Sleep（WS）算法

WS 的全称与抽象定位 WS 的全称是 **Wake–Sleep algorithm**（唤醒–睡眠算法）。它最初出现在带隐变量的生成模型/识别模型（inference network）联合训练语境中，可以理解为一类非常典型的 EM/GEM 思想的工程化实现：当标准 EM 的 E-step（精确后验）不可解或代价太高时，引入一个可计算的**识别分布**（recognition / inference model）$q_\phi(z\mid x)$ 来近似后验，并交替更新

$$\text{(生成模型参数) }\theta\quad\text{与}\quad\text{(推断网络参数) }\phi,$$

在“代理目标上升”的意义下推进训练。

设定：生成模型 + 推断网络 考虑生成模型 $p_\theta(x,z)=p_\theta(z)\,p_\theta(x\mid z)$，以及推断网络 $q_\phi(z\mid x)$。目标直觉上是两件事同时做好：

- 让生成模型解释数据：$p_\theta(x)$ 大（类似 M-step 的精神）；
- 让推断网络逼近真后验：$q_\phi(z\mid x)\approx p_\theta(z\mid x)$（类似 E-step 的精神）。

Wake step：像 “M-step”，更新生成模型 Wake（唤醒）阶段把 $q_\phi(z\mid x)$ 当作当前的“近似后验”，用它来更新 θ：

$$\theta^{(t+1)}\approx\arg\max_\theta\ \mathbb{E}_{x\sim\text{data}}\ \mathbb{E}_{z\sim q_{\phi^{(t)}}(z|x)}\big[\log p_\theta(x,z)\big].$$

这与 EM 的 $Q(\theta\mid\theta^{(t)})=\mathbb{E}_{p_{\theta^{(t)}}(z|x)}[\log p_\theta(x,z)]$ 完全同构，只是把真后验换成了可计算的 q_ϕ，因此更准确地说属于 **GEM / variational GEM**。

Sleep step：像 “E-step”，更新推断网络 Sleep（睡眠）阶段反过来用当前生成模型产生“梦境样本”来训练推断网络，使其逼近后验。最常见的写法是从 $p_{\theta^{(t+1)}}(x,z)$ 采样，然后最小化

$$\phi^{(t+1)}\approx\arg\min_\phi\ \mathbb{E}_{(x,z)\sim p_{\theta^{(t+1)}}(x,z)}\big[-\log q_\phi(z\mid x)\big]=\arg\min_\phi\ \mathbb{E}_{x\sim p_{\theta^{(t+1)}}(x)}\big(p_{\theta^{(t+1)}}(z\mid x)\,\|\,q_\phi(z\mid x)\big).$$

注意这里出现的是 反向 KL：$(p\|q)$，这与 VB 常见的 $(q\|p)$ 方向相反，因此 WS 与 VBEM 虽然同属“近似 E-step”的家族，但其优化几何与偏好不同：$(p\|q)$ 更倾向“覆盖”真后验的支撑，而 $(q\|p)$ 更倾向“择一模式”（mode-seeking）。

WS 与 EM/VBEM 的关系（放到层级里）

- **EM**：E-step 用真后验 $p_\theta(z \mid x)$；
- **VBEM**：E-step 在变分族内极大化 ELBO（通常对应 $(q\|p)$）；
- **Wake–Sleep (WS)**：用推断网络 $q_\phi(z \mid x)$ 近似后验做 Wake 更新，并用模型生成的样本在 Sleep 阶段训练 q_ϕ（常对应 $(p\|q)$ 或其近似），整体更像 **GEM + 采样驱动的 E-like 更新**。

实用直觉　Wake–Sleep 的核心信息是：只要你用一个可计算的推断分布替代不可解的后验，并交替更新“生成模型参数”和“推断网络参数”，你就在 EM/GEM 的大框架下做事；差别只在于你选择了什么代理目标、用哪种 KL 方向、以及用采样还是解析期望来实现更新。

10.6 第六层：EM 的再一个特例是 Gibbs 抽样算法

从“优化 q”到“采样 z”：两条近似 E-step 的路线　在隐变量模型里，EM 的关键瓶颈往往是 E-step：需要计算 $p_\theta(z \mid x)$ 或其期望 $\mathbb{E}_{p_\theta(z|x)}[\log p_\theta(x,z)]$. VBEM 的路线是**显式引入一个近似分布 q 并确定性优化它**；而 MCMC（尤其是 Gibbs 抽样）的路线是**不显式写出 q，而是构造一个以真后验为平稳分布的马尔可夫链**，用时间平均来逼近后验期望。

Gibbs 抽样：用条件分布拼出后验　设 $z = (z_1, \dots, z_J)$。若我们能写出每个坐标的全条件分布

$$p_\theta(z_j \mid z_{-j}, x), \qquad z_{-j} := (z_1, \dots, z_{j-1}, z_{j+1}, \dots, z_J),$$

则 Gibbs 抽样在第 t 次迭代做

$$z_1^{(t+1)} \sim p_\theta(z_1 \mid z_{-1}^{(t)}, x), \quad z_2^{(t+1)} \sim p_\theta(z_2 \mid z_{-2}^{(t+1)}, x), \ \dots, \ z_J^{(t+1)} \sim p_\theta(z_J \mid z_{-J}^{(t+1)}, x).$$

在常见可积条件（不可约、非周期、遍历性）下，这条链的平稳分布就是 $p_\theta(z \mid x)$，并且对合适的函数 h 有马尔可夫链大数定律（MCMC-LLN）：

$$\frac{1}{T}\sum_{t=1}^{T} h\big(z^{(t)}\big) \ \xrightarrow[T\to\infty]{} \ \mathbb{E}_{p_\theta(z|x)}\big[h(z)\big].$$

因此 Gibbs 能为 E-step 中的后验期望提供一个可计算的近似。

MCEM：把 Gibbs 当作“随机版 E-step”　把 Gibbs 与 EM 真正连在一起的标准桥梁是 **Monte Carlo EM (MCEM)**：在第 t 次迭代，用 Gibbs 生成近似后验样本 $z^{(t,1)}, \dots, z^{(t,S)} \approx p_{\theta^{(t)}}(z \mid x)$, 用样本平均近似 Q 函数

$$Q(\theta \mid \theta^{(t)}) = \mathbb{E}_{p_{\theta^{(t)}}(z|x)}[\log p_\theta(x,z)] \ \approx \ \widehat{Q}_S(\theta \mid \theta^{(t)}) := \frac{1}{S}\sum_{s=1}^{S} \log p_\theta\big(x, z^{(t,s)}\big),$$

然后做近似 M-step：

$$\theta^{(t+1)} \in \arg\max_{\theta} \widehat{Q}_S(\theta \mid \theta^{(t)}).$$

当 $S \to \infty$ 且抽样误差可控时，MCEM 会在“噪声”下逼近 EM 的单调上升行为；但由于 Q 是随机近似，**单调性一般不再是逐步严格保证**（需要额外的采样精度安排与收敛论证）。

与 VBEM 的同与不同：都是“可行近似”，但方式不同 把 VBEM 与 Gibbs 放在同一层的原因是：二者都在做

$$\text{不可解的精确 E-step} \quad \longrightarrow \quad \text{可行的近似替代}.$$

但它们的“近似”本质不同：

- **VB（确定性近似）**：选一个受限族 $\mathcal{Q}$，显式优化 $q \in \mathcal{Q}$，通常对应最小化 $(q\|p)$（或等价最大化 ELBO），偏向“择一模式”，速度快、但有偏差（bias）。
- **Gibbs/MCMC（随机近似）**：不限制显式族，而用马尔可夫链让经验分布逼近 p；在理想极限下可以渐近无偏，但有**混合时间**与**自相关**带来的方差与计算代价。

自由能/散度视角：Gibbs 也在“下降某种量” 虽然 Gibbs 不显式最大化 ELBO，但可以从更抽象的“距离后验越来越近”的角度理解其推进：设马尔可夫算子为 $\mathcal{K}$，从任意初始分布 $q^{(0)}$ 出发，迭代得到 $q^{(t+1)} = q^{(t)}\mathcal{K}$。在合适条件下，$q^{(t)} \Rightarrow p_\theta(z \mid x)$，从而诸如全变差距离、KL（在需要额外正则条件时）或其他散度会随迭代趋近于 0。因此它同样体现“用可行机制逼近不可行后验”的 EM 精神：**把 E-step 的积分困难交给随机近似与时间平均**。

小结：Gibbs 在 EM 层级中的位置

- 若把 EM 看作“在后验下取期望 + 在参数上最大化”的交替优化，则 Gibbs 给出一条重要替代：**用采样近似后验期望**（MCEM）。
- 与 VBEM 相比，Gibbs 更接近“原则上可精确”的近似（$T \to \infty$），代价是计算与混合；VBEM 更快更稳，但引入结构性偏差。

10.7 第七层：WS 与 VAE/GAN 的联系（作为直觉）

先把 WS 说清楚：Wake–Sleep algorithm（唤醒–睡眠算法） 这里的“WS”指 **Wake–Sleep algorithm**（Hinton 等在 Helmholtz machine 语境下提出）。它考虑一对模型（两张网）：

- **生成模型（generative model）** $p_\theta(x, z) = p(z)\, p_\theta(x \mid z)$；
- **识别/推断模型（recognition / inference model）** $q_\phi(z \mid x)$，用于近似后验 $p_\theta(z \mid x)$（这就是后来所谓的 amortized inference）。

在真实后验不可解时，WS 的核心想法是：**不做精确 E-step**，而是**学习一个** $q_\phi(z \mid x)$ **来替代**，并用两个方向相反的训练阶段交替推动 θ 与 ϕ。

Wake step：像 VBEM 的 “M-step” ——用近似后验推动生成模型 在 “wake” 阶段，固定识别模型 q_ϕ，用它给出的近似后验（或其采样/权重）来更新生成模型参数 θ，使数据在生成模型下更 “可能”：

$$\max_\theta \ \mathbb{E}_{x\sim p_{\text{data}}}\mathbb{E}_{z\sim q_\phi(z|x)}\big[\log p_\theta(x,z)\big] = \max_\theta \ \mathbb{E}_{x\sim p_{\text{data}}}\mathbb{E}_{z\sim q_\phi(z|x)}\big[\log p(z) + \log p_\theta(x\mid z)\big]. \quad (2)$$

这一步可以看作 “用 q_ϕ 代替真实后验” 的 *VBEM/M-step* 风格：固定一个（近似）q，提升生成模型对数据的解释力。若 z 连续，实践中常用从 $q_\phi(z\mid x)$ 采样、或用重参数化技巧估计梯度；若 z 离散，也可用 REINFORCE/控制变量等得到无偏或低方差梯度估计。

Sleep step：像把 “E-step 学习化” ——让 q_ϕ 逼近后验（但在模型分布下） 在 “sleep” 阶段，固定生成模型 p_θ，从当前生成模型采样 “梦境数据”

$$(z,x)\sim p_\theta(z,x) = p(z)p_\theta(x\mid z),$$

并用这些样本训练识别网络去预测 z：

$$\min_\phi \ \mathbb{E}_{(z,x)\sim p_\theta(z,x)}\big[-\log q_\phi(z\mid x)\big]. \quad (3)$$

把 (3) 改写成 KL 形式：

$$\mathbb{E}_{(z,x)\sim p_\theta}\big[-\log q_\phi(z\mid x)\big] = \mathbb{E}_{x\sim p_\theta(x)}\Big[\mathbb{E}_{z\sim p_\theta(z|x)}[-\log q_\phi(z\mid x)]\Big] = \mathbb{E}_{x\sim p_\theta(x)}\Big[\big(p_\theta(z\mid x)\,\|\,q_\phi(z\mid x)\big)\Big] + \text{const},$$

因此 sleep step 等价于

$$\min_\phi \ \mathbb{E}_{x\sim p_\theta(x)}\big(p_\theta(z\mid x)\,\|\,q_\phi(z\mid x)\big). \quad (4)$$

关键差异（方向与分布两点都不同）：

- **KL 方向：** sleep 用的是 “reverse direction” 的 $(p\|q)$，而标准变分推断/ELBO 常见的是 $(q\|p)$；两者会导致不同的近似偏好（例如 mode-covering vs mode-seeking 的差异）。
- **期望分布：** sleep 的外层期望在 $x\sim p_\theta(x)$（模型分布）下，而不是 $x\sim p_{\text{data}}$（真实数据分布）下；当模型还没学好时，sleep 会用 “质量一般的梦境数据” 训练推断器，这也是 WS 可能不稳的来源之一。

WS 与 EM/VB 的对照：它在 “KL gap 地图” 上处于哪里？ 回忆变分恒等式（你在下一小节会系统写出）：

$$\log p_\theta(x) = \mathcal{L}(q,\theta) + \big(q(z)\,\|\,p_\theta(z\mid x)\big).$$

若我们选 $q(z) = q_\phi(z\mid x)$，则理想的 VBEM/VAE 会直接最大化统一的 ELBO：

$$\max_{\theta,\phi} \ \mathbb{E}_{x\sim p_{\text{data}}}\Big[\mathcal{L}\big(q_\phi(\cdot\mid x),\theta\big)\Big].$$

而 WS 不是在同一个统一目标上做严格坐标上升：

- wake step (2) 更像“给定 q_ϕ，提高 θ”的一块更新；
- sleep step (4) 则是在模型分布下，用 $(p_\theta \| q_\phi)$ 训练推断器。

所以你可以把 WS 读作：

> **用两个更容易实现的代理目标，分别近似“提升似然”与“拟合后验”这两件事；它捕捉了变分学习的精神，但并不等同于严格最大化 ELBO。**

与 VBEM/VAE 的关系：WS 是“摊销推断”的早期形式，但优化目标更松弛 把 VBEM 写成

$$\max_{q\in\mathcal{Q},\theta} \mathcal{L}(q,\theta), \qquad \mathcal{L}(q,\theta) = \mathbb{E}_q[\log p_\theta(x,z)] - \mathbb{E}_q[\log q(z)].$$

VAE 的关键是：用参数化族 $q_\phi(z \mid x)$ 并**在真实数据分布下**直接最大化 ELBO：

$$\max_{\theta,\phi} \mathbb{E}_{x\sim p_{\text{data}}}\Big[\mathbb{E}_{z\sim q_\phi(z|x)} \log p_\theta(x \mid z) - \big(q_\phi(z \mid x)\|p(z)\big)\Big]. \tag{5}$$

与此相比，WS 的同源之处在于：**都用推断网络把 E-step 学习化 (amortize)**，并与生成模型参数共同训练；差别在于：**VAE 有一个统一的、可微的 ELBO 目标 (5)；WS 则以 wake/sleep 两个代理阶段近似该理想**。因此可以把 WS 视为“VAE 之前的工程型变分学习路线”：思想接近，但目标函数与期望分布并不完全一致。

与 GAN 的类比：都是“引入额外网络 + 代理目标”使训练可行（仅作结构直觉） GAN 的判别器–生成器博弈把“分布匹配”写成对抗代理目标：

$$\min_\theta \max_\psi \mathbb{E}_{x\sim p_{\text{data}}}\big[\log D_\psi(x)\big] + \mathbb{E}_{z\sim p(z)}\big[\log(1 - D_\psi(G_\theta(z)))\big].$$

GAN 的判别器 D_ψ 是**辅助网络**：它不在最终生成模型的显式似然里出现，但提供训练信号；WS 的识别网络 q_ϕ 也是**辅助网络**：它不改变 $p_\theta(x,z)$ 的形式，但让不可解后验近似变得可训练。因此二者可以做如下“工程结构类比”：

> **原始目标难直接优化 ⇒ 引入辅助网络 ⇒ 优化一个可行代理目标并交替（或联合）更新。**

但必须强调：**这只是结构类比**——GAN 的核心不是 ELBO/KL gap；而 WS/VAE 的核心仍是“后验近似 + 似然/下界”的变分逻辑。

常见实践提醒：WS 何时好用、何时会“漂”？(给读者的直觉护栏)

- **推断网络会受“梦境分布”影响：**若 p_θ 还很差，sleep 训练信号偏离真实数据，q_ϕ 容易学到“对梦境有效、对真实无效”的映射；
- **KL 方向影响近似形态：**sleep 的 $(p_\theta\|q_\phi)$ 往往更鼓励 q_ϕ 覆盖 p_θ 的支持（mode-covering），这与变分常见的 $(q\|p)$（更 mode-seeking）在行为上可能不同；
- **wake/sleep 目标不统一：**因此缺少“单调提升同一个标量目标”的干净保证，这一点与标准 EM/VBEM 有本质差别。

一句话总结 **Wake–Sleep 是把“E-step”用推断网络近似并学习化（amortize）的早期算法：wake 用 q_ϕ 推动生成模型，sleep 用模型样本训练 q_ϕ 逼近后验（以 $(p_\theta\|q_\phi)$ 的方向）。VAE 把同一思想放进统一 ELBO 最大化框架；GAN 展示了另一条“用辅助网络 + 代理目标”实现不可直接优化目标的路线（但与 ELBO 逻辑不同）。**

10.8 第八层：KL 距离的统一

统一恒等式：对数似然 = 变分下界 + KL gap 设联合分布 $p_\theta(x,z)$ 给定，任取任意分布 $q(z)$（只要在 $p_\theta(x,z)>0$ 的地方 q 绝对连续），定义变分下界（ELBO）

$$\mathcal{L}(q,\theta) := \mathbb{E}_q\big[\log p_\theta(x,z)\big] - \mathbb{E}_q\big[\log q(z)\big] = \mathbb{E}_q\big[\log p_\theta(x,z) - \log q(z)\big].$$

则有恒等式（又称 *ELBO decomposition / variational identity*）

$$\log p_\theta(x) = \mathcal{L}(q,\theta) + \big(q(z) \,\|\, p_\theta(z \mid x)\big), \tag{6}$$

其中

$$(q\|p) := \mathbb{E}_q\Big[\log \frac{q(z)}{p(z)}\Big] \geq 0.$$

因此 $\mathcal{L}(q,\theta) \leq \log p_\theta(x)$，并且当且仅当 $q(z) = p_\theta(z \mid x)$（几乎处处）时取等号。

一张“地图”：所有变体都在处理同一个 gap 恒等式 (6) 的作用是把“你到底在做什么”说清楚：

- 你想提升 $\log p_\theta(x)$，但它通常难直接优化；
- 于是你转而提升 $\mathcal{L}(q,\theta)$（一个可操作的下界）；
- 两者的差距就是 $(q\|p_\theta(\cdot \mid x))$，即 “posterior gap”。

不同算法的差异，本质上就是：**你允许 q 有多自由？你能把 KL gap 压到多小？以及你在 θ 上如何更新？**

标准 EM：坐标上升，E-step 把 gap 置零 EM 的 E-step 在当前 $\theta^{(t)}$ 下取

$$q^{(t)}(z) = p_{\theta^{(t)}}(z \mid x),$$

从而

$$\big(q^{(t)}\|p_{\theta^{(t)}}(\cdot \mid x)\big) = 0, \qquad \mathcal{L}\big(q^{(t)},\theta^{(t)}\big) = \log p_{\theta^{(t)}}(x).$$

随后 M-step 提升下界

$$\theta^{(t+1)} \in \arg\max_\theta \ \mathcal{L}\big(q^{(t)},\theta\big),$$

并由 (6) 得到单调性：

$$\log p_{\theta^{(t+1)}}(x) \ \geq \ \mathcal{L}\big(q^{(t)},\theta^{(t+1)}\big) \ \geq \ \mathcal{L}\big(q^{(t)},\theta^{(t)}\big) = \log p_{\theta^{(t)}}(x).$$

所以 EM 是对 (q,θ) 的**坐标上升**（coordinate ascent）算法：E-step 精确对齐后验，M-step 提升下界。

Hard EM / K-Means：把 q 限制为退化分布，牺牲 KL 换可解性 Hard EM 把变分族限制为

$$\mathcal{Q}_{\text{hard}} = \Big\{ q(z) = \delta_{z=\hat{z}} \Big\},$$

也就是“每个样本一个确定标签”的点质量分布。此时 E-step 不再能令 KL 为零（除非后验本来就退化），而是做

$$q^{(t)} \in \arg\max_{q \in \mathcal{Q}_{\text{hard}}} \ \mathcal{L}\big(q, \theta^{(t)}\big),$$

等价于挑选最可能的隐变量配置（MAP 分配）。在同方差球形 GMM 的特殊化下，$\mathcal{L}$（去常数）正好对应 K-Means 的平方误差目标，从而得到“Assign + Update”的交替最小化。

VBEM：把 q 限制为可计算族，目标是最小化 KL gap（但一般不为零） VBEM 选择一个可计算的变分族（例如均值场）

$$\mathcal{Q}_{\text{mf}} = \Big\{ q(z) = \prod_j q_j(z_j) \Big\},$$

并进行两步交替上升：

$$q^{(t)} \in \arg\max_{q \in \mathcal{Q}_{\text{mf}}} \ \mathcal{L}\big(q, \theta^{(t)}\big), \qquad \theta^{(t+1)} \in \arg\max_{\theta} \ \mathcal{L}\big(q^{(t)}, \theta\big).$$

与标准 EM 相比，E-step 只能在受限集合里“尽量贴近后验”，因此一般有

$$\big(q^{(t)} \| p_{\theta^{(t)}}(\cdot \mid x)\big) > 0.$$

但关键是：**VBEM 仍然在最大化同一个下界** $\mathcal{L}$，只是牺牲了“KL=0”的精确性来换取 tractability。

GEM / 近似 EM：不要求精确最优，只要求下界（或 Q）上升 GEM（generalized EM）把“arg max”放松为“让目标变好就行”：

$$\mathcal{L}\big(q^{(t)}, \theta^{(t+1)}\big) \ \geq \ \mathcal{L}\big(q^{(t)}, \theta^{(t)}\big),$$

或等价地只要求 $Q(\theta \mid \theta^{(t)})$ 上升。这样就允许梯度步、牛顿步、近似内点解等工程实现。地图语言是：**你仍在努力压低 gap / 抬高下界，只是每步不做精确坐标最优**。

WS / 工程化变体：把“提升 $\mathcal{L}$”替换成更容易优化的代理（surrogate） 很多工程算法会引入额外变量、对偶形式或更强的松弛，使得每一步优化的是某个 surrogate objective $S_t(\cdot)$：

$$S_t(\theta^{(t+1)}) \ \geq \ S_t(\theta^{(t)}), \qquad \text{且} \quad S_t(\theta) \ \leq \ \log p_\theta(x) \text{ 或与之紧密相关.}$$

从 (6) 的角度，它们的共同点是：**仍在试图让“可优化的量”逼近（或下界住）对数似然，只是采用不同的下界/对偶/松弛，间接控制某种“gap”**。（是否严格对应 KL gap，取决于具体构造；但“引入辅助块 + 交替更新”的精神与 EM 同源。）

Gibbs / MCMC：不显式写 q，而让时间平均充当近似后验 变分方法显式选一个 q 并最小化 $(q\|p)$；MCMC（如 Gibbs）则构造马尔可夫链，其平稳分布为后验 $p_\theta(z \mid x)$：

$$z_j^{(t+1)} \sim p_\theta\big(z_j \mid z_{-j}^{(t)}, x\big).$$

如果链混合良好，则经验分布

$$\hat{q}_T(z) := \frac{1}{T}\sum_{t=1}^{T} \delta_{z=z^{(t)}}$$

在 $T \to \infty$ 时逼近后验。因此它与 (6) 的联系可以这样理解：**你没有显式优化 KL gap，而是用采样把（隐式的）q 逐渐推向后验，让 gap 通过时间平均被“消化”**。

VAE：把 VBEM 的 q 参数化并摊销（amortize），直接最大化 $\mathcal{L}$ VAE 选择 $q_\phi(z \mid x)$ 并最大化

$$\max_{\phi,\theta} \ \mathcal{L}\big(q_\phi(\cdot \mid x), \theta\big),$$

把“每个样本都要做一次 VB E-step”替换为“训练一个推断网络一次性给出近似后验”。在地图上，VAE 是 VBEM 的深度化版本：**仍是下界优化，gap 仍是 KL，只是 q 由神经网络产生**。

一句话总结：你可以用“(i) 变分族大小；(ii) 是否显式优化 q；(iii) θ 更新是否精确”来给所有方法归类 在 (6) 这张地图上：

- EM：$\mathcal{Q}$ 最大，E-step 令 KL$=0$，M-step 精确提升下界；
- Hard EM / K-Means：$\mathcal{Q}$ 极小（退化分布），KL gap 通常不为零但计算简单；
- VBEM / VAE：$\mathcal{Q}$ 受限（可计算/可学习），直接最大化下界；
- GEM / WS / 各类近似：不一定做精确坐标最优，但保持代理目标改进；
- Gibbs：用采样让隐式 q 随时间逼近后验，把“优化 gap”交给混合。

这就是“KL 距离统一”：不同算法看似千姿百态，但都在同一恒等式下选择不同的“q 的可行性”与“gap 的处理方式”。

本节小结：一口气记住 EM 的“内核”

- EM 的内核不是“两个步骤”，而是：**对数似然的下界 + 坐标上升**；
- E-step 是在选 q（让下界更紧 / 让 KL 更小），M-step 是在选 θ（让下界更高）；
- Hard EM、VBEM、GEM、Gibbs、VAE 等都可以理解为：**在“如何处理不可解后验/不可解 M-step”上做不同的近似与松弛**；
- KL/ELBO 公式 (**??**) 是最好的“统一语言”。

11 一致大数定律（Uniform Laws of Large Numbers）

经验过程与 ULLN：记号与关键概念（带英文对照） 令 $Z_1,\dots,Z_n \overset{\text{i.i.d.}}{\sim} P$ 取值于可测空间 $(\mathcal{Z},\mathcal{A})$。对任意可测函数 $f:\mathcal{Z}\to\mathbb{R}$，记

$$Pf := \mathbb{E}[f(Z)] \ （\textbf{population expectation}）, \qquad P_n f := \frac{1}{n}\sum_{i=1}^{n} f(Z_i) \ （\textbf{empirical mean}）.$$

于是 $P_n - P$ 就是经验过程（**empirical process**）作用在函数 f 上的偏差

$$(P_n - P)f = P_n f - Pf.$$

函数类与一致偏差（uniform deviation）. 给定函数类（**function class**）$\mathcal{F}$，定义经验过程在 $\mathcal{F}$ 上的（半）范数

$$\|P_n - P\|_{\mathcal{F}} := \sup_{f\in\mathcal{F}} \big|(P_n - P)f\big| = \sup_{f\in\mathcal{F}} \big|P_n f - Pf\big|.$$

它刻画了：*在整个函数类 $\mathcal{F}$ 上，经验均值对总体均值的最坏偏差*，因此也常被称为 **uniform law of large numbers (ULLN)** 中的核心量。

一致大数定律（uniform law of large numbers, ULLN）. 称 $\mathcal{F}$ 满足（弱）一致大数定律（**(weak) ULLN**），若

$$\|P_n - P\|_{\mathcal{F}} \xrightarrow{P} 0. \tag{ULLN}$$

若进一步有 $\|P_n - P\|_{\mathcal{F}} \to 0$ a.s.，则称满足（强）ULLN（**strong ULLN**）。

为什么要看 "sup"？ 点态 LLN 只告诉我们：对固定的 f，$P_n f \to Pf$。但学习算法输出的 $\hat{f}$ 是*数据依赖*（*data-dependent*）的，因此必须在一个足够大的候选集合上做*一致*（*uniform*）控制，这就是 $\sup_{f\in\mathcal{F}}$ 的来源。

从 ERM 到一致偏差：泛化界的标准链路（generalization via uniform convergence） 设学习问题由损失（**loss function**）$\ell(\theta;Z)$ 描述，参数 $\theta\in\Theta$。为统一记号，把每个参数 θ 对应到一个损失函数

$$f_\theta(\cdot) := \ell(\theta;\cdot), \qquad \mathcal{F} := \{f_\theta : \theta\in\Theta\}.$$

定义

$$R(\theta) := Pf_\theta = \mathbb{E}[\ell(\theta;Z)] \ （\ \textbf{population risk}）, \hat{R}_n(\theta) := P_n f_\theta = \frac{1}{n}\sum_{i=1}^{n}\ell(\theta;Z_i) \ （\textbf{empirical risk}）.$$

经验风险最小化（empirical risk minimization, ERM）. ERM 选择

$$\hat{\theta} \in \arg\min_{\theta\in\Theta} \hat{R}_n(\theta).$$

记总体最优（**population minimizer**）

$$\theta^\star \in \arg\min_{\theta\in\Theta} R(\theta), \qquad R^\star := \inf_{\theta\in\Theta} R(\theta).$$

ERM 的目标是让超额风险（**excess risk**）

$$R(\hat{\theta}) - R^\star$$

尽可能小。

从三项分解到一致偏差：ERM 的关键一步（uniform deviation）. 由三项分解

$$R(\hat{\theta}) - R(\theta^\star) = \big(R(\hat{\theta}) - \hat{R}_n(\hat{\theta})\big) + \underbrace{\big(\hat{R}_n(\hat{\theta}) - \hat{R}_n(\theta^\star)\big)}_{\le 0} + \big(\hat{R}_n(\theta^\star) - R(\theta^\star)\big),$$

第二项因 ERM 最优性而非正。对其余两项分别用“用最坏情形控制特定点”的上界

$$R(\hat{\theta}) - \hat{R}_n(\hat{\theta}) \le \sup_{\theta\in\Theta}\big|R(\theta) - \hat{R}_n(\theta)\big|, \qquad \hat{R}_n(\theta^\star) - R(\theta^\star) \le \sup_{\theta\in\Theta}\big|R(\theta) - \hat{R}_n(\theta)\big|,$$

相加便得

$$R(\hat{\theta}) - R^\star \le 2\sup_{\theta\in\Theta}\big|R(\theta) - \hat{R}_n(\theta)\big| = 2\|P_n - P\|_{\mathcal{F}}, \qquad R^\star := \inf_{\theta\in\Theta} R(\theta). \tag{7}$$

因此，*ERM* 的超额风险（*excess risk*）之所以可控，关键在于整个损失类上的一致偏差（*uniform deviation / uniform convergence*）是否足够小：

$$\|P_n - P\|_{\mathcal{F}} = \sup_{f\in\mathcal{F}} |(P_n - P)f|.$$

为什么下一步会出现覆盖数（covering number）？因为 $\|P_n - P\|_{\mathcal{F}}$ 是对无限函数类取上确界的随机量，它本身不是“可直接计算”的。经验过程理论的典型做法是：

> 先用一个有限的 ϵ-网格逼近 $\mathcal{F}$（离散化），再用并联合界与集中不等式控制有限个点上的偏差，最后把离散化误差吸收回去。

而“需要多少个网格点”正由覆盖数/熵数刻画：

- 覆盖数（**covering number**）与熵数（**entropy**）给出离散化复杂度；
- 集中不等式（**concentration inequalities**）给出有限网格上的尾界；
- 合并两者得到 $\|P_n - P\|_{\mathcal{F}}$ 的 *covering-tail bound*，从而推出 ULLN（uniform law of large numbers）乃至学习率。

本书后续采用的主线就是：

$$\text{一致偏差}\|P_n - P\|_{\mathcal{F}} \implies \text{covering-tail} \implies \text{覆盖数上界（可检验的增长率条件）}.$$

11.1 $L_1(Q)$ 覆盖数、经验覆盖数与一致覆盖数

$L_1(Q)$ 距离与覆盖数（一致定义） 令 $\mathcal{G}$ 为可测函数类 $g : \mathcal{Z} \to \mathbb{R}$。对任意概率测度 Q 定义

$$d_{1,Q}(g,h) := \int |g - h|\, dQ,$$

$$N(\eta, \mathcal{G}, L_1(Q)) := \min\Big\{ m : \exists g_1, \dots, g_m \in \mathcal{G},\ \forall g \in \mathcal{G},\ \min_{j \le m} d_{1,Q}(g, g_j) \le \eta \Big\}.$$

为简记，下文写

$$N_1(\eta, \mathcal{G}, Q) := N(\eta, \mathcal{G}, L_1(Q)).$$

经验覆盖数（empirical covering number） 给定点列 $z_{1:n} = (z_1, \dots, z_n)$，定义经验测度

$$Q_{z_{1:n}} := \frac{1}{n} \sum_{i=1}^{n} \delta_{z_i}.$$

则经验 L_1 距离

$$d_{1,z_{1:n}}(g, h) := \frac{1}{n} \sum_{i=1}^{n} |g(z_i) - h(z_i)|$$

满足

$$d_{1,z_{1:n}}(g, h) = d_{1,Q_{z_{1:n}}}(g, h),$$

从而经验覆盖数可一致写为

$$N_1(\eta, \mathcal{G}, z_{1:n}) := N(\eta, \mathcal{G}, d_{1,z_{1:n}}) = N_1(\eta, \mathcal{G}, Q_{z_{1:n}}).$$

当 $z_{1:n} = Z_{1:n}$ 为随机样本时，$N_1(\eta, \mathcal{G}, Z_{1:n})$ 为随机变量。

一致（分布无关）覆盖数（distribution-free covering number） 定义

$$N_1^{\mathrm{unif}}(\eta, \mathcal{G}) := \sup_{Q} N_1(\eta, \mathcal{G}, Q),$$

其中上确界可限制在有限支持的 Q（因为经验测度本身就是有限支持，且 $\sup_Q$ 的极端情形可由离散测度逼近）。

引理 11.1 (经验覆盖数由一致覆盖数控制)**.** *对任意点列 $z_{1:n}$ 与任意 $\eta > 0$,*

$$N_1(\eta, \mathcal{G}, z_{1:n}) \le N_1^{\mathrm{unif}}(\eta, \mathcal{G}).$$

因此对任意随机样本 $Z_{1:n}$,

$$\mathbb{E}[N_1(\eta, \mathcal{G}, Z_{1:n})] \le N_1^{\mathrm{unif}}(\eta, \mathcal{G}).$$

证明. 由 $N_1(\eta, \mathcal{G}, z_{1:n}) = N_1(\eta, \mathcal{G}, Q_{z_{1:n}})$,且 $N_1^{\mathrm{unif}}(\eta, \mathcal{G}) = \sup_Q N_1(\eta, \mathcal{G}, Q)$,故逐点有 $N_1(\eta, \mathcal{G}, z_{1:n}) \le N_1^{\mathrm{unif}}(\eta, \mathcal{G})$。对 $z_{1:n} = Z_{1:n}$ 取期望即得。 □

11.2 单尺度覆盖数尾界

定理 11.1 (单尺度覆盖数界（对称化 + 覆盖）)**.** *令 $Z_1, \dots, Z_n \overset{\text{i.i.d.}}{\sim} P$，且 $\mathcal{G}$ 为函数类 $g : \mathcal{Z} \to [0, B]$。则对任意 $\epsilon > 0$,*

$$\Pr\left(\sup_{g \in \mathcal{G}} \big| P_n g - P g \big| > \epsilon \right) \ \le\ 8\, \mathbb{E}\Big[N_1\Big(\frac{\epsilon}{8}, \mathcal{G}, Z_{1:n} \Big) \Big] \exp\Big(-\frac{n \epsilon^2}{128 B^2} \Big). \tag{8}$$

证明. 记

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

Step 1 (ghost sample：用独立样本替换期望). 取一组独立“幽灵样本” $Z_1', \dots, Z_n' \overset{\text{i.i.d.}}{\sim} P$，并与 $Z_{1:n}$ 独立，记

$$P_n' g := \frac{1}{n}\sum_{i=1}^{n} g(Z_i').$$

定义事件

$$A := \left\{ \sup_{g\in\mathcal{G}} |P_n g - Pg| > \epsilon \right\}.$$

在 A 上存在（可取为关于 $Z_{1:n}$ 的可测选择）$g^\star \in \mathcal{G}$ 使得 $|P_n g^\star - Pg^\star| > \epsilon$。再定义

$$B := \left\{ \sup_{g\in\mathcal{G}} |P_n g - P_n' g| > \frac{\epsilon}{2} \right\}, \qquad C := \left\{ |P_n' g^\star - Pg^\star| > \frac{\epsilon}{2} \right\}.$$

若 $\omega \in A \cap C^c$，则由三角不等式

$$|P_n g^\star - P_n' g^\star| \geq |P_n g^\star - Pg^\star| - |P_n' g^\star - Pg^\star| > \epsilon - \frac{\epsilon}{2} = \frac{\epsilon}{2},$$

故 $\omega \in B$。因此

$$A \cap C^c \subseteq B \quad \Longrightarrow \quad A \subseteq B \cup (A \cap C).$$

取概率并用并集上界得

$$\Pr(A) \leq \Pr(B) + \Pr(A \cap C). \tag{9}$$

下面控制 $\Pr(A \cap C)$。注意 $g^\star = g^\star(Z_{1:n})$ 仅依赖于原样本，而 $Z_{1:n}'$ 与 $Z_{1:n}$ 独立。对给定 $Z_{1:n}$，$g^\star$ 为确定函数，且 $g^\star(Z_1'), \dots, g^\star(Z_n')$ i.i.d. 取值于 $[0, B]$。于是条件于 $Z_{1:n}$ 有

$$\mathbb{E}[P_n' g^\star \mid Z_{1:n}] = Pg^\star, \qquad \mathrm{Var}(P_n' g^\star \mid Z_{1:n}) = \frac{1}{n}\mathrm{Var}(g^\star(Z) \mid Z_{1:n}) \leq \frac{B^2}{n}.$$

由（条件）Chebyshev 不等式，

$$\Pr(C \mid Z_{1:n}) = \Pr\Big(|P_n' g^\star - Pg^\star| > \frac{\epsilon}{2} \;\Big|\; Z_{1:n} \Big) \leq \frac{\mathrm{Var}(P_n' g^\star \mid Z_{1:n})}{(\epsilon/2)^2} \leq \frac{4B^2}{n\epsilon^2}.$$

因此

$$\Pr(A \cap C) = \mathbb{E}[\mathbf{1}_A \Pr(C \mid Z_{1:n})] \leq \frac{4B^2}{n\epsilon^2} \Pr(A).$$

代回 (9) 得

$$\Pr(A) \leq \Pr(B) + \frac{4B^2}{n\epsilon^2} \Pr(A).$$

若 $n\epsilon^2 \geq 8B^2$，则 $\frac{4B^2}{n\epsilon^2} \leq \frac{1}{2}$，从而

$$\Pr(A) \leq \Pr(B) + \frac{1}{2}\Pr(A) \quad \Longrightarrow \quad \Pr(A) \leq 2\Pr(B). \tag{S1}$$

（当 $n\epsilon^2 < 8B^2$ 时，用平凡界 $\Pr(A) \le 1$，而 (8) 右端是一个正数界，可通过调整常数吸收该情形；下文只写主情形。）

Step 2（对称化：差分均值 $\Rightarrow$ Rademacher 过程）. 由 B 的定义，

$$\Pr(B) = \Pr\left(\sup_{g\in\mathcal{G}} \left| \frac{1}{n}\sum_{i=1}^{n} \big(g(Z_i) - g(Z_i')\big) \right| > \frac{\epsilon}{2} \right).$$

令 $\sigma_1, \dots, \sigma_n$ 为独立 Rademacher 变量（± 1 等概率），并与所有样本独立。条件于 $(Z_{1:n}, Z'_{1:n})$，对每个固定 g，

$$\frac{1}{n}\sum_{i=1}^{n} (g(Z_i) - g(Z_i')) \;\overset{d}{=}\; \frac{1}{n}\sum_{i=1}^{n} \sigma_i \, (g(Z_i) - g(Z_i')),$$

从而对任意 $t > 0$，

$$\Pr\left(\sup_{g} \left| \frac{1}{n}\sum_{i=1}^{n} (g(Z_i) - g(Z_i')) \right| > t \;\middle|\; Z_{1:n}, Z'_{1:n} \right) = \Pr\left(\sup_{g} \left| \frac{1}{n}\sum_{i=1}^{n} \sigma_i (g(Z_i) - g(Z_i')) \right| > t \;\middle|\; Z_{1:n}, Z'_{1:n} \right).$$

对 $(Z_{1:n}, Z'_{1:n})$ 取期望得到无条件等式

$$\Pr\left(\sup_{g} \left| \frac{1}{n}\sum_{i=1}^{n} (g(Z_i) - g(Z_i')) \right| > t \right) = \Pr\left(\sup_{g} \left| \frac{1}{n}\sum_{i=1}^{n} \sigma_i (g(Z_i) - g(Z_i')) \right| > t \right).$$

再用三角不等式拆分：

$$\sup_{g} \left| \frac{1}{n}\sum_{i=1}^{n} \sigma_i (g(Z_i) - g(Z_i')) \right| \le \sup_{g} \left| \frac{1}{n}\sum_{i=1}^{n} \sigma_i g(Z_i) \right| + \sup_{g} \left| \frac{1}{n}\sum_{i=1}^{n} \sigma_i g(Z_i') \right|.$$

因此当左侧 $> t$ 时，至少有一项 $> t/2$，故

$$\Pr\left(\sup_{g} \left| \frac{1}{n}\sum_{i=1}^{n} \sigma_i (g(Z_i) - g(Z_i')) \right| > t \right) \le \Pr\left(\sup_{g} \left| \frac{1}{n}\sum_{i=1}^{n} \sigma_i g(Z_i) \right| > \frac{t}{2} \right) + \Pr\left(\sup_{g} \left| \frac{1}{n}\sum_{i=1}^{n} \sigma_i g(Z_i') \right| > \frac{t}{2} \right).$$

由于 $(Z_{1:n}, \sigma_{1:n})$ 与 $(Z'_{1:n}, \sigma_{1:n})$ 同分布，两项相等，从而

$$\Pr\left(\sup_{g} \left| \frac{1}{n}\sum_{i=1}^{n} (g(Z_i) - g(Z_i')) \right| > t \right) \le 2 \Pr\left(\sup_{g} \left| \frac{1}{n}\sum_{i=1}^{n} \sigma_i g(Z_i) \right| > \frac{t}{2} \right).$$

取 $t = \epsilon/2$ 得

$$\Pr(B) \le 2 \Pr\left(\sup_{g\in\mathcal{G}} \left| \frac{1}{n}\sum_{i=1}^{n} \sigma_i g(Z_i) \right| > \frac{\epsilon}{4} \right). \tag{S2}$$

Step 3（覆盖：把无限 supremum 变成有限并联）. 条件于样本 $Z_{1:n}$。令 $\eta := \epsilon/8$，取 $\{g_1, \dots, g_M\} \subset \mathcal{G}$ 为 $L_1(Q_{Z_{1:n}})$ 意义下的 η-net，其中

$$M = N_1(\eta, \mathcal{G}, Z_{1:n}).$$

即对任意 $g \in \mathcal{G}$ 存在 $j(g)$ 使得

$$\frac{1}{n}\sum_{i=1}^{n} |g(Z_i) - g_{j(g)}(Z_i)| \le \eta.$$

于是对任意 g，

$$\begin{aligned}\left|\frac{1}{n}\sum_{i=1}^{n}\sigma_i g(Z_i)\right| &\le \left|\frac{1}{n}\sum_{i=1}^{n}\sigma_i g_{j(g)}(Z_i)\right| + \left|\frac{1}{n}\sum_{i=1}^{n}\sigma_i (g(Z_i) - g_{j(g)}(Z_i))\right| \\ &\le \left|\frac{1}{n}\sum_{i=1}^{n}\sigma_i g_{j(g)}(Z_i)\right| + \frac{1}{n}\sum_{i=1}^{n}|g(Z_i) - g_{j(g)}(Z_i)| \\ &\le \left|\frac{1}{n}\sum_{i=1}^{n}\sigma_i g_{j(g)}(Z_i)\right| + \eta.\end{aligned}$$

对 g 取上确界得到

$$\sup_{g\in\mathcal{G}}\left|\frac{1}{n}\sum_{i=1}^{n}\sigma_i g(Z_i)\right| \le \max_{1\le j\le M}\left|\frac{1}{n}\sum_{i=1}^{n}\sigma_i g_j(Z_i)\right| + \eta.$$

因此事件包含

$$\left\{\sup_{g\in\mathcal{G}}\left|\frac{1}{n}\sum_{i=1}^{n}\sigma_i g(Z_i)\right| > \frac{\epsilon}{4}\right\} \subseteq \left\{\max_{1\le j\le M}\left|\frac{1}{n}\sum_{i=1}^{n}\sigma_i g_j(Z_i)\right| > \frac{\epsilon}{8}\right\}. \tag{S3}$$

Step 4（Hoeffding + union bound：对有限网并联）. 固定 j 并条件于 $Z_{1:n}$。记

$$S_j := \frac{1}{n}\sum_{i=1}^{n}\sigma_i g_j(Z_i).$$

则 $\mathbb{E}[S_j \mid Z_{1:n}] = 0$，且每项满足

$$\left|\frac{1}{n}\sigma_i g_j(Z_i)\right| \le \frac{B}{n}.$$

由 Hoeffding 不等式（条件于 $Z_{1:n}$）得

$$\Pr\Big(|S_j| > \frac{\epsilon}{8} \;\Big|\; Z_{1:n}\Big) \le 2\exp\left(-\frac{n\epsilon^2}{128B^2}\right).$$

对 $j = 1, \dots, M$ 用 union bound，

$$\Pr\left(\max_{1\le j\le M}|S_j| > \frac{\epsilon}{8} \;\middle|\; Z_{1:n}\right) \le 2M\exp\left(-\frac{n\epsilon^2}{128B^2}\right).$$

结合 (S3) 得到（仍条件于 $Z_{1:n}$）

$$\Pr\left(\sup_{g\in\mathcal{G}}\left|\frac{1}{n}\sum_{i=1}^{n}\sigma_i g(Z_i)\right| > \frac{\epsilon}{4} \;\middle|\; Z_{1:n}\right) \le 2N_1\Big(\frac{\epsilon}{8}, \mathcal{G}, Z_{1:n}\Big)\exp\left(-\frac{n\epsilon^2}{128B^2}\right).$$

对 $Z_{1:n}$ 取期望得

$$\Pr\left(\sup_{g\in\mathcal{G}}\left|\frac{1}{n}\sum_{i=1}^{n}\sigma_i g(Z_i)\right|>\frac{\epsilon}{4}\right)\le 2\,\mathbb{E}\Big[N_1\Big(\frac{\epsilon}{8},\mathcal{G},Z_{1:n}\Big)\Big]\exp\Big(-\frac{n\epsilon^2}{128B^2}\Big). \tag{S4}$$

合并 Step 1–4. 由 (S1)、(S2) 与 (S4)，

$$\Pr(A)\le 2\Pr(B)\le 4\,\Pr\left(\sup_{g\in\mathcal{G}}\left|\frac{1}{n}\sum_{i=1}^{n}\sigma_i g(Z_i)\right|>\frac{\epsilon}{4}\right)\le 8\,\mathbb{E}\Big[N_1\Big(\frac{\epsilon}{8},\mathcal{G},Z_{1:n}\Big)\Big]\exp\Big(-\frac{n\epsilon^2}{128B^2}\Big),$$

即为 (8)。 □

把随机复杂度变成确定量 由 Lemma 11.1，对任意 $\eta>0$ 有 $N_1(\eta,\mathcal{G},Z_{1:n})\le N_1^{\mathrm{unif}}(\eta,\mathcal{G})$，故

$$\mathbb{E}\Big[N_1\Big(\frac{\epsilon}{8},\mathcal{G},Z_{1:n}\Big)\Big]\le N_1^{\mathrm{unif}}\Big(\frac{\epsilon}{8},\mathcal{G}\Big).$$

代回 (8) 得

$$\Pr\left(\sup_{g\in\mathcal{G}}|P_ng-Pg|>\epsilon\right)\ \le\ 8\,N_1^{\mathrm{unif}}\Big(\frac{\epsilon}{8},\mathcal{G}\Big)\,\exp\Big(-\frac{n\epsilon^2}{128B^2}\Big). \tag{9.1'}$$

11.3 VC-subgraph / 伪维数给出可计算覆盖数

本小节的目标是把上一节得到的“分布无关覆盖数” $N_1^{\mathrm{unif}}(\eta,\mathcal{G})$ 进一步变成一个可直接写成 n,η,V 的显式上界。典型做法是用 *VC-subgraph 维数*（对指示类）或 *伪维数*（对实值类）控制覆盖数/打包数，再通过 packing-to-covering 关系得到 $L_1(Q)$ 覆盖数的多项式增长。

定义 11.1 (VC-subgraph 维数)**.** 给定实值函数类 $\mathcal{G}$，定义其 *subgraph* 类

$$\mathsf{SG}(\mathcal{G}):=\big\{(z,t)\in\mathcal{Z}\times\mathbb{R}:\ t<g(z)\text{ for some }g\in\mathcal{G}\big\}.$$

更精确地说，把每个 $g\in\mathcal{G}$ 对应的集合

$$\mathrm{sub}(g):=\{(z,t):t<g(z)\}\subseteq\mathcal{Z}\times\mathbb{R},$$

则 $\mathrm{sub}(\mathcal{G}):=\{\mathrm{sub}(g):g\in\mathcal{G}\}$ 是一族集合类。若 $\mathrm{sub}(\mathcal{G})$ 的 VC 维有限，记为

$$V_{\mathrm{sg}}:=\mathrm{VCdim}\big(\mathrm{sub}(\mathcal{G})\big),$$

则称 $\mathcal{G}$ 为 VC-subgraph 类，维数为 V_{sg}。

定义 11.2 (伪维数（pseudo-dimension）)**.** 实值函数类 $\mathcal{G}$ 的*伪维数* $\mathrm{Pdim}(\mathcal{G})$ 定义为：存在 m 个点 $z_1,\dots,z_m\in\mathcal{Z}$ 与阈值 $t_1,\dots,t_m\in\mathbb{R}$，使得对任意符号向量 $s\in\{0,1\}^m$，都存在 $g\in\mathcal{G}$ 满足

$$\mathbf{1}\{g(z_i)\ge t_i\}=s_i,\qquad i=1,\dots,m.$$

最大可行的 m 记为 $\mathrm{Pdim}(\mathcal{G})$。

注 11.1 (VC-subgraph 与伪维数的关系). 对 $[0,B]$-值函数类，$\mathrm{Pdim}(\mathcal{G})$ 与 $\mathrm{VCdim}(\mathrm{sub}(\mathcal{G}))$ 在常

数意义下等价：通常可把两者任取其一作为“组合复杂度参数”，并在覆盖数界里以 V 表示。因此下文将 $V := \mathrm{Pdim}(\mathcal{G})$（或等价的 V_{sg}）一致记号。

为了把 VC/伪维数转为 $L_1(Q)$ 覆盖数，我们先引入一个“把 L_1 覆盖化为离散逼近”的中间引理。这一步避免了“覆盖数结论看起来像凭空来”的问题：它说明覆盖数界本质上来自 *用有限精度量化（discretization）把实值类变成有限指示类，再用 Sauer–Shelah 控制生长函数*。

引理 11.2 (量化与 VC-subgraph：$L_1(Q)$ 覆盖数的离散化归约)**.** *令 $\mathcal{G} \subseteq [0,B]^{\mathcal{Z}}$，取任意概率测度 Q。对任意 $m \in \mathbb{N}$，令量化步长 $\Delta := B/m$，并定义量化算子*

$$\mathcal{Q}_m(x) := \Delta \cdot \left\lfloor \frac{x}{\Delta} \right\rfloor, \qquad x \in [0,B].$$

定义量化后的函数类 $\mathcal{G}_m := \{\mathcal{Q}_m \circ g : \ g \in \mathcal{G}\}$。则对任意 $g \in \mathcal{G}$ 有

$$\|g - \mathcal{Q}_m \circ g\|_{L_1(Q)} \le \Delta = \frac{B}{m}.$$

从而

$$N(2\Delta, \mathcal{G}, L_1(Q)) \le |\mathcal{G}_m|_Q,$$

其中 $|\mathcal{G}_m|_Q$ 表示 $\mathcal{G}_m$ 在 Q-几乎处处意义下的等价类数量（当 Q 离散时就是在支持点上的不同取值模式数）。

证明. 第一条不等式来自逐点界：对任意 z，$0 \le g(z) - \mathcal{Q}_m(g(z)) < \Delta$，故

$$\|g - \mathcal{Q}_m \circ g\|_{L_1(Q)} = \int (g - \mathcal{Q}_m \circ g)\, dQ \le \Delta.$$

因此任意两个 $g, h \in \mathcal{G}$ 满足

$$\|g-h\|_{L_1(Q)} \le \|g-\mathcal{Q}_m g\|_{L_1(Q)} + \|\mathcal{Q}_m g - \mathcal{Q}_m h\|_{L_1(Q)} + \|\mathcal{Q}_m h - h\|_{L_1(Q)} \le 2\Delta + \|\mathcal{Q}_m g - \mathcal{Q}_m h\|_{L_1(Q)}.$$

换言之，若 $\{\tilde g_1, \dots, \tilde g_M\} \subset \mathcal{G}_m$ 覆盖 $\mathcal{G}_m$ 在 $L_1(Q)$ 距离下精度为 0（即穷举所有不同等价类），则把每个 $\tilde g_j$ 选取其来源 $g_j \in \mathcal{G}$，得到 $\{g_1, \dots, g_M\}$ 是 $\mathcal{G}$ 的 2Δ-cover。故 $N(2\Delta, \mathcal{G}, L_1(Q)) \le |\mathcal{G}_m|_Q$。 □

接下来需要用 VC-subgraph/伪维数去控制 $|\mathcal{G}_m|_Q$。直观上：量化把 $[0,B]$ 切成 m 个水平，函数在每个点的取值只落在有限集合，于是“不同取值模式”的数量由 subgraph 的 VC 维控制。

引理 11.3 (生长函数控制：VC-subgraph $\Rightarrow$ 量化模式数多项式)**.** *设 $\mathcal{G} \subseteq [0,B]^{\mathcal{Z}}$ 为 VC-subgraph 类，令 $V = \mathrm{VCdim}(\mathrm{sub}(\mathcal{G})) < \infty$。则对任意离散测度 $Q = \frac{1}{n}\sum_{i=1}^n \delta_{z_i}$ 与任意 $m \in \mathbb{N}$，量化类 $\mathcal{G}_m$ 在 $\{z_1, \dots, z_n\}$ 上的不同取值模式数满足*

$$|\mathcal{G}_m|_Q \ \le \ (c\,m)^V,$$

其中 $c \ge 1$ 为绝对常数。

注 11.2. 这一结论是标准 VC 理论的“多水平扩展”：当 $m = 2$ 时，$\mathcal{G}_m$ 退化为阈值指示类，其模式数由 Sauer–Shelah 引理给出；一般 m 可通过把多水平量化写成 $m-1$ 个阈值切片的组合来控制，仍得到 $(cm)^V$ 型多项式上界。不同教材对常数 c 的给法略有差异，但幂次 V 是关键。

将上两条引理串起来，得到我们真正需要的“一致 L_1 覆盖数”上界：

引理 11.4（伪维数（或 VC-subgraph）⇒ 一致 L_1 覆盖数多项式上界）. 令 $\mathcal{G} \subseteq [0,B]^{\mathcal{Z}}$ 且 $\mathrm{Pdim}(\mathcal{G}) = V < \infty$（等价地可假设 *VC-subgraph* 维数为 V）。则存在绝对常数 $C \geq 1$，使得对任意 $0 < \eta \leq B$，

$$N_1^{\mathrm{unif}}(\eta, \mathcal{G}) = \sup_Q N(\eta, \mathcal{G}, L_1(Q)) \leq \left(\frac{CB}{\eta}\right)^V.$$

证明. 由定义 $N_1^{\mathrm{unif}}(\eta,\mathcal{G}) = \sup_Q N(\eta,\mathcal{G},L_1(Q))$，又由上一节 remark 可把上确界限制在有限支持的离散测度 Q。固定任意离散 Q，取 $m := \lceil 2B/\eta \rceil$，则 $\Delta = B/m \leq \eta/2$。由引理 11.2，

$$N(\eta, \mathcal{G}, L_1(Q)) \leq N(2\Delta, \mathcal{G}, L_1(Q)) \leq |\mathcal{G}_m|_Q.$$

再由引理 11.3，

$$|\mathcal{G}_m|_Q \leq (cm)^V \leq \left(\frac{CB}{\eta}\right)^V$$

（将 $m \asymp B/\eta$ 吸收到常数 C 中）。对 Q 取上确界即得结论。 □

最终可计算的单尺度尾界 将 Lemma 11.4 代入 (9.1')，得到存在绝对常数 $c_1, c_2 > 0$，对任意 $\epsilon \in (0, B]$，

$$\Pr\left(\sup_{g\in\mathcal{G}} |P_n g - Pg| > \epsilon\right) \leq c_1 \left(\frac{B}{\epsilon}\right)^V \exp\left(-c_2 \frac{n\epsilon^2}{B^2}\right). \tag{VC-tail}$$

（单尺度）典型误差量级：由尾界反解（更精细） 由 (VC-tail)，对任意 $\epsilon \in (0, B]$，

$$\Pr\left(\sup_{g\in\mathcal{G}} |P_n g - Pg| > \epsilon\right) \leq c_1 \exp\left(V \log \frac{B}{\epsilon} - c_2 \frac{n\epsilon^2}{B^2}\right).$$

因此令右端 $\leq \delta$ 等价于

$$c_2 \frac{n\epsilon^2}{B^2} \;\geq\; V \log \frac{B}{\epsilon} + \log \frac{c_1}{\delta}. \tag{Inv}$$

这是一个隐式不等式（ϵ 同时出现在左边的二次项与右边的对数中）。下面给出两种“更精细”的处理方式：一种保持隐式、便于后续局部化；另一种给出显式可用界。

(a) 保留隐式形式（推荐在后续局部化/peeling 中使用）. 直接把 (Inv) 作为“解 ϵ 的判据”即可：任何满足 (Inv) 的 ϵ 都保证 $\Pr(\sup_g |P_n g - Pg| > \epsilon) \leq \delta$。这种写法最干净，因为你不会在这里提前把 $\log(B/\epsilon)$ 粗暴上界为 $\log n$，从而避免引入多余的 $\log n$ 损失。

(b) 显式上界：一个标准的 self-consistent 解法（写给读者看得懂的）.

从隐式条件 (Inv) 出发，记

$$A := V \log(2n) + \log \frac{c_1}{\delta}.$$

我们希望把右端的 $\log(B/\epsilon)$ 也用同一个 A 控制住。为此，只要能保证

$$\log \frac{B}{\epsilon} \;\leq\; \log(2n) \tag{SC}$$

就够了：因为那样

$$V\log\frac{B}{\epsilon}+\log\frac{c_1}{\delta}\ \le\ V\log(2n)+\log\frac{c_1}{\delta}=A.$$

于是 (Inv) 会被一个更强但更容易验证的充分条件替代：

$$c_2\frac{n\epsilon^2}{B^2}\ \ge\ A. \tag{Suff}$$

现在我们只需选 ϵ 同时满足 (SC) 与 (Suff)。

候选解（由 (Suff) 直接“解出来”）. 令

$$\epsilon:=\frac{B}{\sqrt{c_2}}\sqrt{\frac{A}{n}}.$$

则 (Suff) 以等号成立。

验证 *self-consistency*：检查 (SC). 由 $A\ge V\log(2n)\ge\log(2n)$（当 $V\ge1$ 且 $n\ge2$）可得

$$\epsilon=\frac{B}{\sqrt{c_2}}\sqrt{\frac{A}{n}}\ \ge\ \frac{B}{\sqrt{c_2}}\sqrt{\frac{\log(2n)}{n}}\ \ge\ \frac{B}{2n},$$

其中最后一步用到对 $n\ge2$，$\sqrt{\log(2n)/n}\ge1/(2n)$（这是一个可直接检验的数值不等式）。于是

$$\frac{B}{\epsilon}\ \le\ 2n,\quad\text{即}\quad\log\frac{B}{\epsilon}\le\log(2n),$$

从而 (SC) 成立。于是我们确实把原本的 $\log(B/\epsilon)$ 关进了同一个 $\log(2n)$，并闭合了推导链条。

因此存在绝对常数 $C>0$（吸收 $c_2^{-1/2}$ 等常数）使得

$$\epsilon\ \le\ C\,B\sqrt{\frac{V\log(2n)+\log(c_1/\delta)}{n}}$$

满足 (Inv)，从而得到显式高概率界。

推论 11.1 (VC/伪维数类的高概率一致收敛：隐式形式与可检算显式形式)**.** 在 (VC-tail) 的条件下（即存在常数 $c_1,c_2>0$ 使 $\Pr(\sup_{g\in\mathcal{G}}|P_ng-Pg|>\epsilon)\le c_1(B/\epsilon)^V\exp(-c_2n\epsilon^2/B^2)$），则对任意 $\delta\in(0,1)$、任意 $n\ge2$：

1. **（隐式最精细形式）**任取 $\epsilon\in(0,B]$，若

$$c_2\frac{n\epsilon^2}{B^2}\ \ge\ V\log\frac{B}{\epsilon}+\log\frac{c_1}{\delta}, \tag{Inv}$$

则

$$\Pr\left(\sup_{g\in\mathcal{G}}|P_ng-Pg|>\epsilon\right)\le\delta.$$

2. **（显式可检算形式）**存在绝对常数 $C>0$（只依赖于 c_1,c_2），使得

$$\Pr\left(\sup_{g\in\mathcal{G}}|P_ng-Pg|>C\,B\sqrt{\frac{V\log(2n)+\log(c_1/\delta)}{n}}\right)\le\delta. \tag{10}$$

证明. (1) 由 (VC-tail) 令右端 $\le \delta$，取对数并整理即得 (Inv)；因此满足 (Inv) 的任意 ϵ 都保证尾概率 $\le \delta$。

(2) 令

$$A := V\log(2n) + \log\frac{c_1}{\delta}, \qquad \epsilon := \frac{B}{\sqrt{c_2}}\sqrt{\frac{A}{n}}.$$

则 $c_2 n\epsilon^2/B^2 = A$。另一方面，由 $A \ge \log(2n)$（当 $V \ge 1$ 且 $n \ge 2$）可得

$$\epsilon \ge \frac{B}{\sqrt{c_2}}\sqrt{\frac{\log(2n)}{n}} \ge \frac{B}{2n},$$

从而 $\log(B/\epsilon) \le \log(2n)$。于是

$$V\log\frac{B}{\epsilon} + \log\frac{c_1}{\delta} \le V\log(2n) + \log\frac{c_1}{\delta} = A = c_2\frac{n\epsilon^2}{B^2},$$

即 ϵ 满足 (Inv)，再由 (1) 得 $\Pr(\sup_g |P_n g - Pg| > \epsilon) \le \delta$。把 ϵ 写成 (10) 的形式并把 $c_2^{-1/2}$ 吸收到常数 C 中即可。 □

11.4 从单尺度到 chaining：熵积分与 γ_2 的多尺度复杂度

上一节的结论属于**单尺度**控制：选定一个精度 ϵ 的 net 覆盖 $\mathcal{G}$，再对 net 元素做 union bound，因此复杂度只以 $\log N(\epsilon, \mathcal{G}, d)$ 在单一尺度进入尾界。这往往会在反解误差时留下额外的对数项（例如 $\log(B/\epsilon)$ 被粗略吸收为 $\log n$）。

chaining 的核心是把“并联代价”分摊到多尺度上：用一列尺度 $\eta_0 > \eta_1 > \cdots \downarrow 0$ 逐层逼近每个 g，将过程写成“粗到细”的增量和，并逐层控制这些增量；系统性论述见 Ledoux et al.[12]，Talagrand[18]。在**次高斯增量**（相对于某个半度量 d）的情形下，Dudley 的**熵积分**给出经典上界

$$\mathbb{E}\sup_{g\in\mathcal{G}} X_g \lesssim \int_0^{\mathrm{diam}(\mathcal{G},d)} \sqrt{\log N(u, \mathcal{G}, d)}\, du, \tag{11}$$

参见 Dudley[7] 以及经验过程语境下的系统处理 van der Vaart et al.[20]。更进一步，generic chaining 用 $\gamma_2(\mathcal{G}, d)$ 刻画次高斯过程上确界的正确复杂度[18]。

相较单尺度，chaining 的主要优势在于：复杂度由“单点的 $\log N(\epsilon)$”升级为“多尺度累积”（熵积分或 γ_2），从而在许多熵随尺度变化剧烈的函数类上能削弱或避免由粗并联带来的额外对数损失；并且它更自然地与局部化技巧（peeling / local Rademacher complexity）衔接，用局部复杂度替代全局复杂度，得到更锋利（且常更自适应）的界[3,11,20]。

12 神经网络的相合性：让 ULLN 技术链路“落地”

在上一章中，我们已经把

超额风险（excess risk） ⇒ 一致偏差（uniform deviation） ⇒ 覆盖数（covering number）

这条链路写成了可复用模板：

$$\text{ERM} \xRightarrow{(7)} R(\hat{\theta}) - \inf_{\Theta} R(\theta) \;\le\; 2\|P_n - P\|_{\mathcal{F}}, \qquad \|P_n - P\|_{\mathcal{F}} \text{ 由定理 11.1 + 覆盖数控制.}$$

本节将这条模板应用到一族随样本量增长的神经网络函数类上，从而得到回归函数估计的 $L_2(P_X)$ 相合性。整体逻辑只有两块：

(i)（**逼近**）网络类逐渐变“富”，类内最优风险逼近 $f^\star$；

(ii)（**增长率**）网络类的复杂度增长不要太快，使得一致偏差仍能收敛到 0。

相关思路可追溯至 Györfi et al.[10]；本节用本书 §11 的 ULLN 记号将其整理成可复用模板。

12.1 设定、网络函数类与目标

观测 $Z_i = (X_i, Y_i)$，$i = 1, \dots, n$ i.i.d.，其中 $X \in \mathbb{R}^d$，且 $\mathbb{E}[Y^2] < \infty$。记 P_X 为 X 的边缘分布，并定义回归函数

$$f^\star(x) := \mathbb{E}[Y \mid X = x].$$

对任意可测函数 $f : \mathbb{R}^d \to \mathbb{R}$，定义平方风险（总体风险）

$$L(f) := \mathbb{E}\big[(Y - f(X))^2\big], \qquad L^\star := \inf_g L(g).$$

在 $\mathbb{E}[Y^2] < \infty$ 下，$f^\star \in L_2(P_X)$ 且满足（几乎处处意义下的）最小化性质：

$$L(f^\star) = L^\star.$$

更具体地，对任意 $f \in L_2(P_X)$，有正交分解（也可视作条件期望的投影性质）

$$\begin{aligned} L(f) &= \mathbb{E}\big[(Y - f^\star(X) + f^\star(X) - f(X))^2\big] \\ &= \mathbb{E}\big[(Y - f^\star(X))^2\big] + \mathbb{E}\big[(f(X) - f^\star(X))^2\big] + 2\,\mathbb{E}\Big[(Y - f^\star(X))\,(f^\star(X) - f(X))\Big]. \end{aligned}$$

最后一项为 0，因为

$$\mathbb{E}\Big[(Y - f^\star(X))\,(f^\star(X) - f(X))\Big] = \mathbb{E}\Big[\mathbb{E}\big[Y - f^\star(X) \mid X\big]\,(f^\star(X) - f(X))\Big] = 0.$$

因此得到平方损失的恒等式

$$L(f) - L(f^\star) = \mathbb{E}\big[(f(X) - f^\star(X))^2\big] = \|f - f^\star\|^2_{L_2(P_X)}.$$

特别地，对任意序列 (f_n)（使得 $f_n \in L_2(P_X)$），都有

$$L(f_n) \to L(f^\star) \quad \Longleftrightarrow \quad \|f_n - f^\star\|_{L_2(P_X)} \to 0,$$

因为两者差值恰好等于 $\|f_n - f^\star\|^2_{L_2(P_X)}$。

两层网络类（ℓ_1 输出约束） 取一个有界激活函数 $\sigma : \mathbb{R} \to [0,1]$。给定随 n 增长的宽度 k_n 与输出层 ℓ_1 预算 β_n，定义

$$\mathcal{F}_n := \Big\{ x \mapsto c_0 + \sum_{j=1}^{k_n} c_j \, \sigma(a_j^\top x + b_j) : \ a_j \in \mathbb{R}^d, \ b_j \in \mathbb{R}, \ \sum_{j=0}^{k_n} |c_j| \le \beta_n \Big\}. \tag{12}$$

令经验风险最小化器（若最小元不存在则取近似最小元）为

$$\hat{f}_n \in \arg\min_{f \in \mathcal{F}_n} \ \hat{L}_n(f), \qquad \hat{L}_n(f) := \frac{1}{n} \sum_{i=1}^{n} (Y_i - f(X_i))^2.$$

目标： 给出条件使得

$$\|\hat{f}_n - f^\star\|_{L_2(P_X)} \to 0 \quad \text{(in probability 或 a.s.)}.$$

12.2 主定理：逼近 + 增长率 ⇒ ERM 相合性

定理 12.1 (两层网络类的 ERM 相合性（可检验版）)**.** *在上述设定下，令两层网络类 $\mathcal{F}_n$ 如 (12)，并令*

$$\hat{f}_n \in \arg\min_{f \in \mathcal{F}_n} \ \hat{L}_n(f), \qquad \hat{L}_n(f) := \frac{1}{n} \sum_{i=1}^{n} (Y_i - f(X_i))^2.$$

假设：

***(A)*（逼近）** *网络类对 $f^\star$ 渐近可逼近：*

$$\inf_{f \in \mathcal{F}_n} \|f - f^\star\|^2_{L_2(P_X)} \ \to \ 0. \tag{13}$$

***(B)*（增长率：复杂度不超过样本量）** *网络宽度 k_n 与输出层预算 β_n 满足*

$$\frac{k_n \, \beta_n^4 \, \log\big(\beta_n(k_n+1)\big)}{n} \ \to \ 0. \tag{14}$$

则

$$\|\hat{f}_n - f^\star\|^2_{L_2(P_X)} \xrightarrow{P} 0.$$

若进一步加强 (14) 使得后文 (19) 的右端对每个固定 $\epsilon > 0$ 关于 n 可求和，则

$$\|\hat{f}_n - f^\star\|^2_{L_2(P_X)} \to 0 \quad a.s.$$

注 12.1 (定理的两块含义：逼近 vs. 统计复杂度). (A) 是逼近论：网络类是否足够“富”，能逼近 $f^\star$。(B) 是统计复杂度：(14) 保证

$$\|P_n - P\|_{\mathcal{H}_n} \to 0, \qquad \mathcal{H}_n := \{(x,y) \mapsto (y - f(x))^2 : \ f \in \mathcal{F}_n\},$$

从而 ERM 的类内超额风险可由一致偏差控制并收敛到 0。换言之，本节证明的技术核心就是：用覆盖数上界把“增长率条件”转化为一个可检验的 ULLN。

12.3 证明：把 ULLN 模板逐条套到 $\mathcal{F}_n$

我们证明定理 12.1（可检验版）。证明分为两段：先在有界响应情形下闭合

$$\text{ERM} \;\Rightarrow\; \text{一致偏差} \;\Rightarrow\; \text{covering-tail} \;\Rightarrow\; \text{增长率},$$

得到 $\|P_n - P\|_{\mathcal{H}_n} \to 0$ 并推出 $\|\hat{f}_n - f^\star\|_{L_2(P_X)} \to 0$；随后用一个标准截断/规约把一般的 $\mathbb{E}[Y^2] < \infty$ 情形归约到有界响应。

Step 0：先证有界情形（技术规约），再放回一般情形 为便于直接使用定理 11.1 的 $[0, B]$-值函数类 tail bound，我们先在假设

$$|Y| \le M \quad \text{a.s.} \tag{$\star$}$$

下完成证明。最后再说明如何由 $\mathbb{E}[Y^2] < \infty$ 规约到 ($\star$)。

在 ($\star$) 下，取 $\beta_n \ge M$。由 $\sigma \in [0,1]$ 与 $\sum_{j=0}^{k_n} |c_j| \le \beta_n$ 得到

$$\|f\|_\infty \le \beta_n, \qquad \forall f \in \mathcal{F}_n.$$

对任意 $f \in \mathcal{F}_n$ 定义损失函数

$$h_f(x,y) := (y - f(x))^2, \qquad \mathcal{H}_n := \{h_f : \; f \in \mathcal{F}_n\}.$$

则

$$0 \le h_f(x,y) \le (|y| + \|f\|_\infty)^2 \le (M + \beta_n)^2 \le 4\beta_n^2 =: B_n,$$

因此

$$\mathcal{H}_n \subset [0, B_n]^{\mathcal{Z}}.$$

Step 1：ERM ⇒ 类内超额风险由一致偏差控制 令总体/经验平方风险分别为

$$L(f) := Ph_f = \mathbb{E}[(Y - f(X))^2], \qquad \hat{L}_n(f) := P_n h_f = \frac{1}{n}\sum_{i=1}^n (Y_i - f(X_i))^2.$$

由与 (7) 完全同型的分解（把 θ 换为 f，把 $\mathcal{F}$ 换为 $\mathcal{H}_n$），对任意经验最小元 $\hat{f}_n \in \arg\min_{f\in\mathcal{F}_n} \hat{L}_n(f)$ 有

$$L(\hat{f}_n) - \inf_{f\in\mathcal{F}_n} L(f) \le 2\|P_n - P\|_{\mathcal{H}_n}, \qquad \|P_n - P\|_{\mathcal{H}_n} := \sup_{h\in\mathcal{H}_n} |(P_n - P)h|. \tag{15}$$

因此只要能证明 $\|P_n - P\|_{\mathcal{H}_n} \to 0$（概率或几乎处处），就把 $L(\hat{f}_n)$ 推到类内最优风险。

Step 2：covering-tail（套用定理 11.1） 由定理 11.1（把 B 换成 $B_n = 4\beta_n^2$，把 $\mathcal{G}$ 换成 $\mathcal{H}_n$），对任意 $\epsilon > 0$，

$$\Pr(\|P_n - P\|_{\mathcal{H}_n} > \epsilon) \le 8\,\mathbb{E}\Big[N_1\Big(\frac{\epsilon}{8}, \mathcal{H}_n, Z_{1:n}\Big)\Big] \exp\Big(-c\frac{n\epsilon^2}{\beta_n^4}\Big), \tag{16}$$

其中 $c > 0$ 为绝对常数，$N_1(\cdot,\cdot,\cdot)$ 为经验 L_1 覆盖数（见 §11）。

Step 3：损失覆盖数 $\rightsquigarrow$ 函数覆盖数 对任意 $f,g\in\mathcal{F}_n$，在样本点上有

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}|h_f(Z_i)-h_g(Z_i)| &= \frac{1}{n}\sum_{i=1}^{n}\left|(Y_i-f(X_i))^2-(Y_i-g(X_i))^2\right| \\ &= \frac{1}{n}\sum_{i=1}^{n}|f(X_i)-g(X_i)|\cdot|2Y_i-f(X_i)-g(X_i)| \\ &\le 4\beta_n\cdot\frac{1}{n}\sum_{i=1}^{n}|f(X_i)-g(X_i)|,\end{aligned}$$

其中最后一步用到 $|Y_i|\le M\le\beta_n$ 以及 $\|f\|_\infty,\|g\|_\infty\le\beta_n$。因此得到传递不等式

$$N_1\Big(\frac{\epsilon}{8},\mathcal{H}_n,Z_{1:n}\Big)\le N_1\Big(\frac{\epsilon}{32\beta_n},\mathcal{F}_n,X_{1:n}\Big). \tag{17}$$

Step 4：网络类的覆盖数上界（文献给出量级） 由 §11.3 的通用模板（伪维数/VC $\Rightarrow$ 经验 L_1 覆盖数），并结合两层网络类 (12) 的伪维数控制，存在常数 $C_0,C_1>0$（C_1 至多与 d 多项式相关），使得对任意 $\eta\in(0,1)$，

$$N_1(\eta,\mathcal{F}_n,X_{1:n})\ \le\ \Big(\frac{C_0\,\beta_n(k_n+1)}{\eta}\Big)^{C_1k_n}. \tag{18}$$

该量级上界可由 Györfi et al.[10] 的计算推出；我们在此不重复常数推导。

Step 5：增长率 $\Rightarrow$ 一致偏差收敛（指数压过多项式） 将 (18) 与 (17) 代回 (16)，得到存在常数 $C,c>0$，对任意固定 $\epsilon>0$，

$$\Pr(\|P_n-P\|_{\mathcal{H}_n}>\epsilon)\ \le\ C\Big(\frac{\beta_n^2(k_n+1)}{\epsilon}\Big)^{C_1k_n}\exp\Big(-c\frac{n\epsilon^2}{\beta_n^4}\Big). \tag{19}$$

从而只要

$$\frac{k_n\,\beta_n^4\,\log\big(\beta_n(k_n+1)\big)}{n}\to 0, \tag{20}$$

就有（对每个固定 $\epsilon>0$）右端 $\to 0$，即

$$\|P_n-P\|_{\mathcal{H}_n}\xrightarrow{P}0.$$

解释：取对数可见右端的主导项为 $C_1k_n\log(\beta_n^2(k_n+1)/\epsilon)-c\,n\epsilon^2/\beta_n^4$，而 (20) 正是让负的指数项压过前面的多项式项。

若进一步对每个固定 $\epsilon>0$ 有

$$\sum_{n=1}^{\infty}C\Big(\frac{\beta_n^2(k_n+1)}{\epsilon}\Big)^{C_1k_n}\exp\Big(-c\frac{n\epsilon^2}{\beta_n^4}\Big)<\infty,$$

则由 (19) 与 Borel–Cantelli 得 $\|P_n-P\|_{\mathcal{H}_n}\to 0$ a.s.

Step 6：类内最优 + 逼近 $\Rightarrow L_2(P_X)$ 相合（细化推导） 由 Step 1 的 oracle inequality (15) 与 Step 5 得到的 $\|P_n - P\|_{\mathcal{H}_n} = o_P(1)$，

$$L(\hat{f}_n) \le \inf_{f\in\mathcal{F}_n} L(f) + 2\,\|P_n - P\|_{\mathcal{H}_n} = \inf_{f\in\mathcal{F}_n} L(f) + o_P(1). \tag{21}$$

接下来把右端的“类内最优风险”改写成相对 $f^\star$ 的 $L_2(P_X)$ 逼近误差。

(i) 用平方损失恒等式把风险差变成 $L_2(P_X)$ 距离。对任意 f（使得 $f \in L_2(P_X)$），我们已证明

$$L(f) - L(f^\star) = \|f - f^\star\|^2_{L_2(P_X)}. \tag{22}$$

于是对任意 $f \in \mathcal{F}_n$，

$$L(f) = L(f^\star) + \|f - f^\star\|^2_{L_2(P_X)}.$$

对 $\mathcal{F}_n$ 取下确界，得到

$$\inf_{f\in\mathcal{F}_n} L(f) = L(f^\star) + \inf_{f\in\mathcal{F}_n} \|f - f^\star\|^2_{L_2(P_X)}. \tag{23}$$

(ii) 将 (21) 变形为“距离上界”。由 (22)，左端亦可写为

$$L(\hat{f}_n) = L(f^\star) + \|\hat{f}_n - f^\star\|^2_{L_2(P_X)}.$$

将该式与 (23) 代入 (21)，两边同时减去 $L(f^\star)$，得到

$$\|\hat{f}_n - f^\star\|^2_{L_2(P_X)} \le \inf_{f\in\mathcal{F}_n} \|f - f^\star\|^2_{L_2(P_X)} + o_P(1). \tag{24}$$

(iii) 结合逼近假设推出相合。若主定理假设 (A) 成立，即

$$\inf_{f\in\mathcal{F}_n} \|f - f^\star\|^2_{L_2(P_X)} \to 0,$$

则由 (24) 立刻得到

$$\|\hat{f}_n - f^\star\|^2_{L_2(P_X)} \xrightarrow{P} 0.$$

（强一致的推导同理：若 Step 5 给出 $\|P_n - P\|_{\mathcal{H}_n} \to 0$ a.s.，则 (24) 中的 $o_P(1)$ 直接替换为 $o(1)$ a.s.，从而得到 $\|\hat{f}_n - f^\star\|^2_{L_2(P_X)} \to 0$ a.s.）

若逼近条件 (13) 成立，则推出 $\|\hat{f}_n - f^\star\|^2_{L_2(P_X)} \to 0$ （in probability）。强一致情形同理：若 Step 5 给出 a.s. 的 $\|P_n - P\|_{\mathcal{H}_n} \to 0$，则得到 a.s. 的相合性。

Step 7：从有界响应回到 $\mathbb{E}[Y^2] < \infty$（截断规约） 最后说明如何去掉 ($\star$)。令截断响应 $Y^{(M)} := \max\{-M, \min\{Y, M\}\}$，并记对应回归函数 $f^{\star,(M)}(x) := \mathbb{E}[Y^{(M)} \mid X = x]$ 与风险

$$L^{(M)}(f) := \mathbb{E}[(Y^{(M)} - f(X))^2].$$

由于 $|Y^{(M)}| \le M$，对每个固定 M，上述 Step 0–6 可用于 $Y^{(M)}$，从而在增长率条件 (20) 下得到

$$\|\hat{f}_n^{(M)} - f^{\star,(M)}\|_{L_2(P_X)} \to 0,$$

其中 $\hat{f}_n^{(M)}$ 为以 $Y^{(M)}$ 训练得到的 ERM。另一方面，由 $\mathbb{E}[Y^2] < \infty$ 可知 $Y^{(M)} \to Y$ 于 L_2，并且 $\|f^{\star,(M)} - f^\star\|_{L_2(P_X)} \to 0$（条件期望在 L_2 下是收缩映射）。再结合

$$|L(f) - L^{(M)}(f)| \le \mathbb{E}\big[(Y - Y^{(M)})^2\big] + 2\,\|f\|_{L_2(P_X)} \cdot \|Y - Y^{(M)}\|_{L_2},$$

以及 $\|f\|_\infty \le \beta_n$（在本节网络类上有一致控制），可将“截断世界”的结论传回原问题。因此，(20) 与 (13) 蕴含定理 12.1 的结论。

注 12.2 (结构回看：网络只是把模板具体化). Step 1–3 是纯模板：ERM $\Rightarrow$ 一致偏差，且损失类覆盖数可由函数类覆盖数控制；真正的网络特定输入只有 Step 4 的覆盖数上界 (18)，以及逼近性质 (13)。增长率条件 (20) 的含义就是：**复杂度（多项式）被样本量（指数尾）压住。**

文献说明：网络类覆盖数的量级计算与由此导出的增长率条件，可追溯至 Györfi et al.[10]。我们在此采用本书一致的经验过程记号重述，并把证明组织为“covering-tail 控制一致偏差 + 逼近”两块。

13 支持向量机（SVM）：从几何直觉到正则化 ERM

本节做两件事：先把 SVM 作为一个**分类器**完整讲清楚（它到底在解什么优化问题、为什么叫“最大间隔”、软间隔与核从何而来），再把它放回统计学习理论的主线：SVM 本质上是**RKHS 中的正则化经验风险最小化（regularized ERM）**，由此自然引出泛化、复杂度控制与一致性。

13.1 任务设定：二分类与判别函数

考虑二分类问题：样本 $(X, Y) \sim P$，其中 $Y \in \{-1, +1\}$。我们用一个**判别函数** $f : \mathcal{X} \to \mathbb{R}$ 来做预测，并用符号函数给出类别

$$\hat{Y} = \operatorname{sign}(f(X)).$$

因此学习问题的核心不是“拟合概率”而是“找到一个把两类分开的打分函数”。

0-1 风险（我们真正关心的目标） 分类错误率对应的总体风险是

$$R_{01}(f) = \Pr\big(Y \ne \operatorname{sign}(f(X))\big).$$

直接最小化 R_{01} 等价于最小化 0-1 损失，但这是一个**非凸、不可导**的目标，一般难以优化，也不利于推导泛化理论。

SVM 的策略：用一个**凸的替代损失**来近似 0-1 损失，并配合一个明确的复杂度控制（正则化），使得“可优化”和“可泛化”同时成立。

13.2 线性可分：最大间隔的几何直觉

先从最理想的情形出发：两类在某个特征空间中线性可分。设线性打分函数

$$f(x) = w^\top x + b.$$

分类正确要求：对每个样本 (x_i, y_i)，

$$y_i(w^\top x_i + b) > 0.$$

但满足正确分类的 (w, b) 可能有无穷多个。SVM 选择其中“最稳健”的那个：**最大化间隔**。

函数间隔与几何间隔 把约束缩放成

$$y_i(w^\top x_i + b) \geq 1, \quad i = 1, \ldots, n,$$

则两条支持超平面 $w^\top x + b = \pm 1$ 到分离超平面 $w^\top x + b = 0$ 的距离为 $1/\|w\|$，因此总间隔为 $2/\|w\|$。最大化间隔等价于最小化 $\|w\|$。

硬间隔 SVM（线性可分）

$$\min_{w,b} \ \frac{1}{2}\|w\|^2 \quad \text{s.t.} \quad y_i(w^\top x_i + b) \geq 1, \ i = 1, \ldots, n. \tag{HSVM}$$

关键提醒：这里的目标 $\|w\|^2$ 是复杂度控制（让分类边界“平滑/简单”），约束是拟合数据（让样本分类正确）。这已经是“拟合 vs 复杂度”的折中雏形，只是折中通过“硬约束”实现。

13.3 不可分与噪声：软间隔与 hinge 损失

真实数据往往不可分：有噪声、异常点、类间重叠。若仍坚持 (HSVM) 的硬约束，会导致无解或极不稳健。SVM 的修正是允许“少量/程度可控”的违约——引入松弛变量 $\xi_i \geq 0$：

软间隔 SVM（原始形式）

$$\min_{w,b,\xi} \ \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n} \xi_i \quad \text{s.t.} \quad y_i(w^\top x_i + b) \geq 1 - \xi_i, \ \xi_i \geq 0. \tag{SSVM}$$

从约束到损失：hinge 损失的出现 对固定的 (w, b)，最优的 ξ_i 等于

$$\xi_i = \max\{0, \ 1 - y_i(w^\top x_i + b)\}.$$

代回 (SSVM) 得到无约束的经验目标：

$$\min_{w,b} \ \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n} \max\{0, 1 - y_i(w^\top x_i + b)\}. \tag{Hinge-ERM}$$

这就是 **hinge 损失**：

$$L_{\text{hinge}}(y, t) = \max(0, 1 - yt).$$

直观上：如果 $yt \geq 1$（分对且间隔足够大）就不再惩罚；如果 $0 < yt < 1$（分对但太贴边）会被惩罚；若 $yt \leq 0$（分错）惩罚更大。

关键提醒（SVM 的“稳健性”从何而来）： hinge 损失对“已安全分开”的点不给梯度，学习主要由**边界附近**的点决定——这就是“支持向量”的来源。

13.4 对偶问题与支持向量：为什么解是稀疏的

软间隔 SVM 的拉格朗日对偶形式为（线性情形）：

$$\max_{\alpha} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j \, x_i^\top x_j \quad \text{s.t.} \quad 0 \le \alpha_i \le C, \ \sum_{i=1}^{n} \alpha_i y_i = 0. \tag{Dual}$$

最优解满足

$$w = \sum_{i=1}^{n} \alpha_i y_i x_i,$$

从而

$$f(x) = w^\top x + b = \sum_{i=1}^{n} \alpha_i y_i \, x_i^\top x + b.$$

当某个样本的 $\alpha_i = 0$，它对分类面没有贡献；只有 $\alpha_i > 0$ 的点是**支持向量**。这解释了 SVM 计算与解释上的一个重要特征：**解往往由少数关键样本“支撑”**。

13.5 核技巧：把非线性分类写成内积

线性 SVM 的限制在于：只能学到线性边界。解决办法是把输入映射到高维特征空间 $\phi(x)$ 后再做线性分类：

$$f(x) = \langle w, \phi(x) \rangle + b.$$

但显式构造 ϕ 可能代价极高。核技巧的关键是：对偶目标和预测只依赖于内积 $\langle \phi(x_i), \phi(x_j) \rangle$，因此只要定义一个核函数

$$k(x, x') = \langle \phi(x), \phi(x') \rangle,$$

就能在不显式计算 ϕ 的情况下实现非线性分类。

核 SVM 的对偶形式 只需把 (Dual) 中的 $x_i^\top x_j$ 替换为 $k(x_i, x_j)$：

$$\max_{\alpha} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j \, k(x_i, x_j) \quad \text{s.t.} \quad 0 \le \alpha_i \le C, \ \sum_{i=1}^{n} \alpha_i y_i = 0. \tag{K-Dual}$$

预测函数为

$$f(x) = \sum_{i=1}^{n} \alpha_i y_i \, k(x_i, x) + b.$$

一句话总结 **核函数的选择**决定了你允许的非线性形状；**正则强度（C 或 λ）**决定了你对过拟合的容忍度。

13.6 把 SVM 放回 ERM 主线：正则化经验风险最小化

到这里我们已经完整看到了 SVM 的优化问题。现在把它写成统计学习里最常见的形式：**正则化 ERM**。

函数空间表述（RKHS） 令 H 为由核 k 诱导的 RKHS。SVM（hinge）可写为

$$\hat{f}_\lambda \in \arg\min_{f\in H}\left\{\frac{1}{n}\sum_{i=1}^{n}\max\big(0,1-y_i f(x_i)\big)+\lambda\|f\|_H^2\right\}. \tag{SVM-RERM}$$

其中 $\lambda>0$ 与 C 等价地控制正则强度（通常 $C\propto 1/\lambda$，常数因实现细节略有差异）。

关键提醒（这就是 Steinwart 叙事的核心落点）： SVM 不是“某个特殊的几何算法”，而是

$$\text{(凸替代损失)}+\text{(RKHS 范数正则)} \quad\Rightarrow\quad \text{可优化、可控复杂度、可泛化。}$$

13.7 代表定理：无限维优化为何变成有限维

代表定理（Representer Theorem）说明：对形如

$$\min_{f\in H}\Big\{\Phi\big(f(x_1),\dots,f(x_n)\big)+\lambda\|f\|_H^2\Big\}$$

的问题，最优解一定可以写成

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

因此即便 H 是无限维，解也落在样本张成的 n 维子空间里，学习问题化为有限维凸优化。

13.8 为什么正则化 ERM 能学到“真东西”

在 ERM 语言里，学习误差通常拆成两部分：

(1) 估计误差：有限样本导致的偏差 经验目标与总体目标不同，会带来估计误差。它与函数类复杂度有关。在 RKHS 中，$\|f\|_H$ 的约束/惩罚直接控制复杂度：λ 越大，函数越“简单”，估计误差越小。

(2) 逼近误差：函数空间本身的限制 即使有无限数据，我们仍在 H 中找函数；如果真实 Bayes 决策边界不在 H 的可逼近范围内，就会留下逼近误差。核越“丰富”，逼近误差越可能小，但也更容易过拟合，需要更强的正则化。

关键提醒（调参的统计含义）：选核是在决定“你允许什么形状”；选 λ（或 C）是在决定“你对有限样本噪声有多谨慎”。交叉验证本质上是在用数据近似完成这个“逼近—估计”的折中。

13.9 SVM 的一致性：经验过程–比较不等式–校准的统一视角（定理–证明版）

本节把 SVM（更一般：核方法的 Tikhonov 正则化）的一致性证明写成一套与本书 ERM 模板完全兼容的 定理–引理–推论结构。主线是

统一偏差（在有效类上） $\Rightarrow$ 比较不等式（oracle inequality） $\Rightarrow$ 选λ_n 兼顾近似与估计 $\Rightarrow$（分类）校准桥接

覆盖数在这里的角色非常明确：它只服务于第一步，用来把 $\|P_n-P\|_{\mathcal{F}_\lambda}$ 变成可计算的 $B(n,\lambda)$。

13.9.1 设定：正则化 ERM 与有效函数类

观测 $Z=(X,Y)\sim P$，考虑损失族

$$\ell(\theta;Z),\qquad \theta\in\Theta,$$

并写

$$R(\theta):=P\ell(\theta;\cdot),\qquad \hat{R}_n(\theta):=P_n\ell(\theta;\cdot).$$

给定正则项 $J(\theta)\geq 0$ 与 $\lambda>0$，定义**正则化 ERM**

$$\hat{\theta}_\lambda\in\arg\min_{\theta\in\Theta}\Big\{\hat{R}_n(\theta)+\lambda J(\theta)\Big\}. \tag{25}$$

并定义对应的**总体正则化解**

$$\theta_\lambda\in\arg\min_{\theta\in\Theta}\Big\{R(\theta)+\lambda J(\theta)\Big\}.$$

一致性目标是选取 $\lambda=\lambda_n$ 使

$$R(\hat{\theta}_{\lambda_n})\to R^*:=\inf_{\theta\in\Theta}R(\theta),$$

分类时再由 surrogate 风险收敛推出 0-1 风险收敛。

有效类（regularization induces localization）. 由最优化结构，$\hat{\theta}_\lambda$ 必须落在一个由 λ 控制的有效参数集内。在最常见的情形下（例如存在 θ_0 使 $J(\theta_0)=0$ 且 $\ell(\theta_0;z)$ 有界），可以得到

$$J(\hat{\theta}_\lambda)\ \leq\ \frac{\hat{R}_n(\theta_0)-\inf_\theta\hat{R}_n(\theta)}{\lambda}\ \lesssim\ \frac{1}{\lambda},\qquad\Rightarrow\quad \hat{\theta}_\lambda\in\Theta_\lambda:=\{\theta:\ J(\theta)\leq c/\lambda\}. \tag{26}$$

据此定义有效函数类

$$\mathcal{F}_\lambda:=\{\ell(\theta;\cdot):\ \theta\in\Theta_\lambda\}.$$

后续所有经验过程控制都只需在 $\mathcal{F}_\lambda$ 上进行。

13.9.2 比较不等式（oracle inequality）

定理 13.1 (正则化 ERM 的比较不等式)**.** *设损失族 $\mathcal{F}_\lambda$ 可测且（为简洁起见）满足有界性：存在 $B_\lambda<\infty$ 使得对所有 $\theta\in\Theta_\lambda$，$0\leq\ell(\theta;Z)\leq B_\lambda$ a.s. 则对任意 $\lambda>0$，*

$$R(\hat{\theta}_\lambda)-R^*\ \leq\ A(\lambda)\ +\ 2\,\|P_n-P\|_{\mathcal{F}_\lambda}, \tag{27}$$

其中

$$A(\lambda):=\inf_{\theta\in\Theta}\Big\{R(\theta)-R^*+\lambda J(\theta)\Big\}=\big[\inf_\theta\{R(\theta)+\lambda J(\theta)\}\big]-R^*. \tag{28}$$

证明. 由 (25) 的最优性，

$$\hat{R}_n(\hat{\theta}_\lambda)+\lambda J(\hat{\theta}_\lambda)\leq\hat{R}_n(\theta)+\lambda J(\theta),\qquad\forall\theta\in\Theta.$$

两边同时加减总体风险并重排，得

$$\begin{aligned} R(\hat\theta_\lambda) - R^* &= \big(R(\hat\theta_\lambda) - \hat R_n(\hat\theta_\lambda)\big) + \big(\hat R_n(\hat\theta_\lambda) + \lambda J(\hat\theta_\lambda)\big) - \big(R^* + \lambda J(\hat\theta_\lambda)\big) \\ &\le \big(R(\hat\theta_\lambda) - \hat R_n(\hat\theta_\lambda)\big) + \inf_{\theta\in\Theta}\big\{\hat R_n(\theta) + \lambda J(\theta)\big\} - R^*. \end{aligned}$$

对右端的 inf 再加减 $R(\theta)$：

$$\inf_\theta\{\hat R_n(\theta) + \lambda J(\theta)\} \le \inf_\theta\{R(\theta) + \lambda J(\theta)\} + \sup_\theta\big(\hat R_n(\theta) - R(\theta)\big).$$

合并两式，并注意

$$R(\hat\theta_\lambda) - \hat R_n(\hat\theta_\lambda) \le \sup_{\theta\in\Theta_\lambda}\big(R(\theta) - \hat R_n(\theta)\big), \quad \sup_\theta\big(\hat R_n(\theta) - R(\theta)\big) \le \sup_{\theta\in\Theta_\lambda}\big|\hat R_n(\theta) - R(\theta)\big|,$$

得到

$$R(\hat\theta_\lambda) - R^* \le \big[\inf_\theta\{R(\theta) + \lambda J(\theta)\} - R^*\big] + 2\sup_{\theta\in\Theta_\lambda}\big|R(\theta) - \hat R_n(\theta)\big|,$$

即 (27)–(28)。 □

注 13.1 (与 (7) 的关系). 当 $\lambda = 0$ 且 $J \equiv 0$ 时，$A(\lambda) = 0$，(27) 退化为标准 ERM 的 $R(\hat\theta) - R^* \le 2\|P_n - P\|_{\mathcal F}$。正则化的唯一新内容就是：多了一个近似项 $A(\lambda)$，并且经验过程只需在有效类 $\mathcal F_\lambda$ 上控制。

13.9.3 覆盖数控制统一偏差（把 $\|P_n - P\|_{\mathcal F_\lambda}$ 变成可检验条件）

定理 13.2 (统一偏差的 covering-tail 上界（有效类版本）). *在定理 13.1 的有界性条件下，对任意 $\epsilon > 0$，*

$$\Pr\Big(\|P_n - P\|_{\mathcal F_\lambda} > \epsilon\Big) \le 8\,\mathbb E\Big[N_1\Big(\frac{\epsilon}{8}, \mathcal F_\lambda, Z_{1:n}\Big)\Big]\exp\Big(-c\,\frac{n\epsilon^2}{B_\lambda^2}\Big), \tag{29}$$

其中 $c > 0$ 为绝对常数，$N_1(\cdot,\cdot,Z_{1:n})$ 是经验 L_1 覆盖数（与定理 11.1 同记号）。

证明. 直接将本书定理 11.1 应用于函数类 $\mathcal G = \mathcal F_\lambda$ 与上界 $B = B_\lambda$ 即得。 □

注 13.2 (覆盖数在 SVM 一致性中的唯一入口). 定理 13.1 将一致性归约为 $A(\lambda) \to 0$ 与 $\|P_n - P\|_{\mathcal F_\lambda} \to 0$。而定理 13.2 用覆盖数把 $\|P_n - P\|_{\mathcal F_\lambda}$ 变成“多项式复杂度 × 指数尾”的上界，从而导出可检验的增长率条件与学习率。

13.9.4 风险一致性（surrogate 风险）

定理 13.3 (正则化 ERM 的风险一致性（surrogate 风险）). *设对每个 n 选取 $\lambda_n > 0$。若满足*

***(i)*（近似项消失）** $A(\lambda_n) \to 0$;

***(ii)*（估计项消失）** $\|P_n - P\|_{\mathcal F_{\lambda_n}} \xrightarrow{P} 0$,

则

$$R(\hat\theta_{\lambda_n}) - R^* \xrightarrow{P} 0.$$

若进一步对每个固定 $\epsilon > 0$ 有 $\sum_n \Pr(\|P_n - P\|_{\mathcal F_{\lambda_n}} > \epsilon) < \infty$ 且 $A(\lambda_n) \to 0$，则得到 a.s. 收敛：$R(\hat\theta_{\lambda_n}) \to R^*$。

证明. 由定理 13.1，

$$R(\hat{\theta}_{\lambda_n}) - R^* \le A(\lambda_n) + 2\|P_n - P\|_{\mathcal{F}_{\lambda_n}}.$$

右端两项在假设下分别 $\to 0$（概率或几乎处处），结论立得。 □

注 13.3 (如何把 (ii) 变成“增长率条件”). 用定理 13.2，只要能上界 $N_1(\epsilon/8, \mathcal{F}_{\lambda_n}, Z_{1:n})$ 的增长速度（通常通过核的熵数/特征值衰减等），就能保证右端指数项压过覆盖数项，从而推出 (ii)。这一步就是 SVM 学习率理论的来源：λ_n 同时进入 $A(\lambda_n)$ 与覆盖数上界。

13.9.5 分类校准（surrogate ⇒ 0-1）

现在进入二分类。令 $Y \in \{-1, +1\}$，评分函数（margin function）写作 $g_\theta : \boldsymbol{X} \to \mathbb{R}$，对应分类器为 $\hat{y} = sign(g_\theta(x))$。0-1 风险记为

$$R_{01}(g) := \Pr\big(Y \neq sign(g(X))\big), \qquad R_{01}^* := \inf_g R_{01}(g),$$

其中下确界可在所有可测 g 上取，R_{01}^* 即 Bayes 风险。给定 margin-based surrogate $\phi : \mathbb{R} \to [0, \infty)$，其 surrogate 风险为

$$R_\phi(g) := \mathbb{E}\,\phi\big(Y g(X)\big), \qquad R_\phi^* := \inf_g R_\phi(g).$$

(SVM 分类中 ϕ 常取 hinge：$\phi(u) = (1-u)_+$。)

为何需要“校准”？ Step 2–3 只能推出某个 surrogate 风险的收敛 $R_\phi(\hat{g}_n) \to R_\phi^*$。但目标风险是 R_{01}。要把 surrogate 的最优性转成分类最优性，必须使用 分类校准（*classification calibration*）：它刻画“surrogate 的最优解是否在符号上与 Bayes 决策一致”。

定义 13.1 (分类校准（以 margin-based surrogate 为例）). 令 $\eta(x) := \Pr(Y = +1 \mid X = x)$。对任意 $\eta \in [0, 1]$ 与任意标量 $t \in \mathbb{R}$，定义条件 ϕ-风险

$$C_\phi(\eta, t) := \eta\,\phi(t) + (1-\eta)\,\phi(-t), \qquad C_\phi^*(\eta) := \inf_{t \in \mathbb{R}} C_\phi(\eta, t).$$

称 ϕ **分类校准**，若对任意 $\eta \neq \frac{1}{2}$，

$$\inf_{t:\ sign(t) \neq sign(\eta - \frac{1}{2})} C_\phi(\eta, t) \;>\; C_\phi^*(\eta).$$

直观上：当 $\eta > \frac{1}{2}$（应判为 $+1$）时，强行取 $t \le 0$ 会付出严格更大的条件 surrogate 风险；当 $\eta < \frac{1}{2}$（应判为 -1）时，强行取 $t \ge 0$ 亦然。

定理 13.4 (校准/链接不等式（准确表述）). *若 ϕ 是分类校准的 margin-based surrogate，则存在一个单调不减函数 $\psi : [0, \infty) \to [0, \infty)$，满足 $\psi(0) = 0$ 且 $\psi(u) > 0$ 对一切 $u > 0$，使得对任意可测评分函数 g，*

$$\psi\Big(R_{01}(g) - R_{01}^*\Big) \;\le\; R_\phi(g) - R_\phi^*. \tag{30}$$

因此，只要 $R_\phi(g_n) \to R_\phi^*$，必有 $R_{01}(g_n) \to R_{01}^*$。

证明思路（主线可放此，细节可放附录）. 关键是“点态—积分”分解与定义 13.1。记 $\eta(X) = \Pr(Y = 1 \mid X)$，则 0-1 的条件风险为

$$C_{01}(\eta, t) = \eta \cdot \mathbf{1}\{t \le 0\} + (1 - \eta) \cdot \mathbf{1}\{t > 0\}, \qquad C^*_{01}(\eta) = \min\{\eta, 1 - \eta\}.$$

并且有分解

$$R_{01}(g) - R^*_{01} = \mathbb{E}[C_{01}(\eta(X), g(X)) - C^*_{01}(\eta(X))],$$

$$R_\phi(g) - R^*_\phi = \mathbb{E}\big[C_\phi(\eta(X), g(X)) - C^*_\phi(\eta(X))\big].$$

分类校准保证：当 $C_{01}(\eta, t) - C^*_{01}(\eta)$ 为正（即符号犯错或边界处不确定）时，$C_\phi(\eta, t) - C^*_\phi(\eta)$ 也必须付出正的代价，并且这种代价可以通过一个统一的函数 ψ 下界，从而得到 (30)。 □

hinge 的一个“可直接写进正文”的特例（最常用） 对 hinge 损失 $\phi_{\text{hinge}}(u) = (1 - u)_+$，存在一个非常简单、常用且足够强的链接界：

定理 13.5 (hinge ⇒ 0-1：线性链接)**.** 令 $R_{\text{hinge}}(g) := \mathbb{E}(1 - Yg(X))_+$。则对任意 g，

$$R_{01}(g) - R^*_{01} \;\le\; R_{\text{hinge}}(g) - R^*_{\text{hinge}}. \tag{31}$$

证明（可留在正文，足够短）. 首先对任意 g 有逐点上界

$$\mathbf{1}\{Yg(X) \le 0\} \;\le\; (1 - Yg(X))_+,$$

因此 $R_{01}(g) \le R_{\text{hinge}}(g)$。其次，二者的最小值关系为：在所有可测 g 上，$R^*_{\text{hinge}} = 2R^*_{01}$（这是对条件风险逐点最小化可得：$C^*_{\text{hinge}}(\eta) = 2\min\{\eta, 1 - \eta\}$）。于是

$$R_{01}(g) - R^*_{01} \le R_{\text{hinge}}(g) - \tfrac{1}{2}R^*_{\text{hinge}} \le R_{\text{hinge}}(g) - R^*_{\text{hinge}},$$

得到 (31)。 □

注 13.4 (写作建议：主线用 (31) 就够了). 对 SVM 分类而言，最常用的是 hinge，因此正文给出 (31) 这条线性链接可以让读者“一眼看懂” surrogate-to-0-1 的桥接；更一般的 ψ-校准形式（定理 13.4）可以放在注记/附录中作为统一理论背景。

13.9.6 推论：SVM 的分类一致性（最终落点）

推论 13.1 (分类一致性（SVM 的最终落点）)**.** 设使用 *hinge* 损失进行正则化 *ERM*（例如 *SVM*），并选取 λ_n 使得

$$R_{\text{hinge}}(\hat{g}_{\lambda_n}) - R^*_{\text{hinge}} \;\to\; 0 \quad (\textit{in probability 或 a.s.}).$$

则由定理 *13.5*，

$$R_{01}(\hat{g}_{\lambda_n}) - R^*_{01} \;\to\; 0 \quad (\text{同样的收敛方式}).$$

小结（本节与经验过程章节的接口）

- **经验过程/覆盖数只负责估计误差：** 通过控制有效类 $\mathcal{F}_\lambda$ 上的统一偏差（例如用 covering-tail），得到 $B(n, \lambda)$ 并据此选择 λ_n 使 $B(n, \lambda_n) \to 0$；

- **比较不等式把经验过程“嵌入优化”：** oracle inequality 给出 $R_\phi(\hat{g}_\lambda) - R_\phi^* \le A(\lambda) + B(n,\lambda)$，从而 surrogate 风险一致性归约为 $A(\lambda_n) \to 0$ 与 $B(n,\lambda_n) \to 0$；
- **校准只负责对齐风险目标：** 一旦 surrogate 风险收敛，链接不等式（一般形式 (30)，hinge 特例 (31)）立即推出 0-1 风险收敛到 Bayes 风险。

14 Minimax 框架：下界、上界与在密度估计中的运行

本节把“非参数估计能做到多好”组织成可复用的 **minimax** 框架，并以**密度估计**贯穿说明：一旦给定结构类与损失，minimax 风险就被唯一确定。叙事三步：先给出 minimax 的“尺子”，再提示上/下界的两条套路，最后落到密度估计记号。

14.1 从一致收敛到 minimax：两个问题、两把尺子

一致收敛回答什么？(uniform deviation / distribution-free) 学习论常从统一偏差出发：

$$\|P_n - P\|_{\mathcal{F}} = \sup_{f\in\mathcal{F}} \big|(P_n - P)f\big|,$$

从而得到 ERM/正则化 ERM 的（分布无关）超额风险控制 $R(\hat{\theta}) - R^* \lesssim \|P_n - P\|_{\mathcal{F}}$. 这一路线强调对所有分布都成立（distribution-free），但通常不利用“真对象的结构”，因而界可能偏保守。[2,6,16,20]

minimax 回答什么？(worst-case over a structural class) minimax 把问题改写为：若真对象属于一个结构类 $\mathcal{F}$，在最坏情况下能做到的最小平均误差阶是什么？其核心对象是 minimax 风险（minimax risk）

$$R_n(\mathcal{F},\ell) := \inf_{\hat{f}} \sup_{f\in\mathcal{F}} \mathbb{E}_f\, \ell(\hat{f}, f), \tag{32}$$

其中 ℓ 是损失（loss），$\inf_{\hat{f}}$ 对所有可测估计量取下确界。[19]

互补关系（本书主线）

- **一致收敛/复杂度：** 给定算法与函数类，导出非渐近、分布无关上界；
- **minimax：** 给定结构类，刻画信息论最优误差阶（含下界）。

当某个可计算估计量的上界达到 minimax 下界（至多差 $\log n$），就完成“可实现算法 $\approx$ 统计最优”的闭环。[19]

14.2 问题设定：密度估计

观测 $X_1,\dots,X_n \overset{\text{i.i.d.}}{\sim} p$，其中 p 为 $\mathbb{R}^d$ 上密度。目标是构造估计量 $\hat{p} = \hat{p}(X_{1:n})$。

损失（loss） 常用选择包括

$$\ell_2(\hat{p},p) := \|\hat{p} - p\|_2^2 = \int (\hat{p}(x) - p(x))^2\, dx, \qquad kl(\hat{p},p) := KL(p\|\hat{p}) = \int p(x) \log \frac{p(x)}{\hat{p}(x)}\, dx.$$

结构类（model / function class） 用 $\mathcal{P}$ 表示候选密度类（例如 Hölder/Besov 光滑密度）：

$$p \in \mathcal{P}.$$

minimax 风险（density estimation） 对任意损失 ℓ，定义

$$R_n(\mathcal{P}, \ell) := \inf_{\hat{p}} \sup_{p \in \mathcal{P}} \mathbb{E}_p \, \ell(\hat{p}, p). \tag{33}$$

这一定义与 (32) 完全同型：把 f 换成密度 p，把 $\mathcal{F}$ 换成密度类 $\mathcal{P}$ 即可。[19]

14.3 minimax 风险：上下界的两条通用套路

问题设定（密度估计） 观测 $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} p$，其中 p 是定义在 $\mathbb{R}$（或 $[0,1]$）上的概率密度。给定结构类 $\mathcal{P}$（例如 Hölder/Sobolev/Besov 类）与损失 $L(p, \hat{p})$，*minimax* 风险定义为

$$R_n(\mathcal{P}; L) := \inf_{\hat{p}} \sup_{p \in \mathcal{P}} \mathbb{E}_p \big[L(p, \hat{p}) \big],$$

其中 inf 取遍所有基于样本的估计器 $\hat{p} = \hat{p}(X_1^n)$。

上下界要回答的两件事

- **上界（可达性）：** 构造一个显式估计器 $\hat{p}$，证明 $\sup_{p \in \mathcal{P}} \mathbb{E}_p[L(p, \hat{p})] \le C \, \psi_n$；
- **下界（不可超越）：** 证明对任意估计器 $\tilde{p}$ 都有 $\sup_{p \in \mathcal{P}} \mathbb{E}_p[L(p, \tilde{p})] \ge c \, \psi_n$。

两者合并就得到 $R_n(\mathcal{P}; L) \asymp \psi_n$。

14.3.1 下界套路：估计 ⇒ 检验

Step 1：从估计下界到多假设检验 在 $\mathcal{P}$ 中挑选 $M+1$ 个候选密度 $p_0, \ldots, p_M$，使得它们在损失意义下彼此分离，例如存在 $s > 0$ 满足

$$L(p_j, p_k) \ge (2s)^2, \qquad \forall j \ne k.$$

考虑多假设检验 $\psi : \mathcal{X}^n \to \{0, 1, \ldots, M\}$，其最坏误判概率

$$p_{e,M} := \inf_{\psi} \max_{0 \le j \le M} P_{p_j}^{(n)}\big(\psi \ne j\big),$$

其中 $P_{p_j}^{(n)}$ 是样本 X_1^n 的联合分布。

Step 2：检验误差下界推出 minimax 风险下界 若候选族在损失意义下 $2s$-分离，则存在一个常数 $c_0 > 0$（与 n 无关）使得

$$R_n(\mathcal{P}; L) \;\ge\; c_0 \, s^2 \, p_{e,M}.$$

因此：只要能证明“这些候选很难检验”（$p_{e,M}$ 不会太小），就能得到估计的 *minimax* 下界。

常用工具（按使用场景分类）

- **Le Cam 二点法：** 取 $M = 1$，用 TV/Hellinger/KL/χ^2 等距离控制两分布可分辨性；
- **Fano + packing：** 取 $M \to \infty$，用平均 KL 控制多假设检验误差（常搭配 Varshamov–Gilbert）；
- **Assouad 超立方体：** 把候选组织成 $\{0,1\}^m$，将整体难度分解到每一位的两点检验。

14.3.2 上界套路：显式构造 + 偏差–方差分解

统一分解：Bias–Variance 对带平滑参数 λ（带宽/截断层数/分辨率等）的估计器 $\hat{p}_{n,\lambda}$，上界证明通常走：

$$\mathbb{E}_p\|\hat{p}_{n,\lambda} - p\|_2^2 \le \underbrace{\|\mathbb{E}_p\hat{p}_{n,\lambda} - p\|_2^2}_{\mathrm{Bias}^2(\lambda)} + \underbrace{\mathbb{E}_p\|\hat{p}_{n,\lambda} - \mathbb{E}_p\hat{p}_{n,\lambda}\|_2^2}_{\mathrm{Var}(\lambda)}.$$

然后用 λ 平衡两项得到最优速率。

14.4 例子：一维 Hölder 密度的点态 minimax 风险（完整上下界）

结构类 $P(\beta, L)$（本书 Chapter 1 的记号） 令 $\beta > 0$，写 $m := \lfloor\beta\rfloor$。称 $p \in P(\beta, L)$ 若 p 有 m 阶导数且最高阶导数满足 Hölder 条件

$$|p^{(m)}(x) - p^{(m)}(y)| \le L|x - y|^{\beta - m}.$$

（并可附加 $p \ge 0, \int p = 1$ 以及有界性等常规条件。）

14.4.1 上界：核密度估计达到 $n^{-2\beta/(2\beta+1)}$

核密度估计器 取带宽 $h > 0$ 与核函数 K，定义

$$\hat{p}_{n,h}(x) := \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right).$$

定理 14.1 (点态 MSE 上界与最优带宽)**.** *设 $p \in P(\beta, L)$，且 K 是 β 阶核并满足适当可积性条件。则对任意固定点 x_0，存在常数 $C_1, C_2 > 0$ 使得*

$$\sup_{p \in P(\beta,L)} \mathbb{E}_p\big(\hat{p}_{n,h}(x_0) - p(x_0)\big)^2 \;\le\; C_2^2 h^{2\beta} + \frac{C_1}{nh}.$$

因此选择

$$h_n^\star \asymp n^{-1/(2\beta+1)}$$

可得

$$\sup_{p \in P(\beta,L)} \mathbb{E}_p\big(\hat{p}_{n,h_n^\star}(x_0) - p(x_0)\big)^2 \;\lesssim\; n^{-2\beta/(2\beta+1)}.$$

证明. 写偏差 $b(x_0) = \mathbb{E}_p\hat{p}_{n,h}(x_0) - p(x_0)$。由于 K 为 β 阶核，对 $p(x_0 + uh)$ 作 m 阶 Taylor 展开并利用矩消去，得到 $|b(x_0)| \le C_2 h^\beta$，因此 $b(x_0)^2 \le C_2^2 h^{2\beta}$。另一方面，

$$\mathrm{Var}\big(\hat{p}_{n,h}(x_0)\big) = \frac{1}{n}\mathrm{Var}\left(\frac{1}{h}K\left(\frac{X_1 - x_0}{h}\right)\right) \le \frac{C_1}{nh},$$

其中 C_1 由 $\int K^2$ 与 $\|p\|_\infty$ 等控制。合并即得 MSE 上界。最后令 h 平衡 $h^{2\beta}$ 与 $(nh)^{-1}$，得到 $h_n^\star \asymp n^{-1/(2\beta+1)}$ 与相应速率。 □

14.4.2 下界：Le Cam 二点法给出同阶下界

构造两条“很像但在 x_0 不同”的密度 固定 x_0。取一个光滑 “bump” 函数 ψ，满足：(i) 支撑在 $[-1,1]$；(ii) $\int \psi(u)\,du = 0$（保证归一化）；(iii) $\psi(0) = 1$；(iv) ψ 的低阶矩为零到合适阶数（保证 Hölder 正则性与不破坏 $P(\beta, L)$）。令 $h > 0$，定义局部扰动

$$\delta_h(x) := h^\beta \psi\left(\frac{x - x_0}{h}\right).$$

取一个基准密度 $p_0 \in P(\beta, L)$，并令

$$p_1(x) := p_0(x) + c\,\delta_h(x),$$

其中常数 $c > 0$ 足够小以保证 $p_1 \geq 0$ 且 $p_1 \in P(\beta, L)$。则

$$p_1(x_0) - p_0(x_0) = c\,h^\beta.$$

定理 14.2 (点态 minimax 下界)**.** *存在常数 $c_0 > 0$，使得对任意估计器 $T_n(x_0)$，*

$$\sup_{p \in P(\beta,L)} \mathbb{E}_p\big(T_n(x_0) - p(x_0)\big)^2 \;\geq\; c_0\, n^{-2\beta/(2\beta+1)}.$$

证明思路（按 Le Cam 二点法）. 令 $P_0^{(n)}, P_1^{(n)}$ 分别为 p_0, p_1 诱导的样本分布。二点法的关键是把估计误差下界化为两分布不可分辨性：若 $P_0^{(n)}$ 与 $P_1^{(n)}$ 很接近，任何估计器都无法稳定地区分“真的是 p_0 还是 p_1”，从而在 x_0 处至少有 $\asymp (p_1(x_0) - p_0(x_0))^2$ 的误差。

具体地，设 $\Delta := |p_1(x_0) - p_0(x_0)| = ch^\beta$。对任意估计器 $T_n(x_0)$，考虑检验

$$\psi(X_1^n) := \mathbf{1}\{T_n(x_0) \geq (p_0(x_0) + p_1(x_0))/2\},$$

可将“估计误差”与“二点检验误判概率”联系起来，推出

$$\max_{j=0,1} \mathbb{E}_{p_j}\big(T_n(x_0) - p_j(x_0)\big)^2 \;\geq\; \frac{\Delta^2}{4}\, p_{e,1},$$

其中 $p_{e,1}$ 是二点检验的 minimax 误判概率。

接下来控制 $p_{e,1}$：用 KL 或 χ^2（或 Hellinger/TV）上界距离，当两分布足够近时有 $p_{e,1} \geq c_1 > 0$。对 i.i.d. 模型，KL 可写为

$$K(P_0^{(n)}, P_1^{(n)}) = n\,K(P_0, P_1) \approx \frac{n}{2} \int \frac{(p_1 - p_0)^2}{p_0}\,dx.$$

而 $(p_1 - p_0)^2 \asymp h^{2\beta}\psi^2((x - x_0)/h)$，积分给出 $\int (p_1 - p_0)^2 dx \asymp h^{2\beta+1}$，从而

$$K(P_0^{(n)}, P_1^{(n)}) \lesssim nh^{2\beta+1}.$$

选择 $h \asymp n^{-1/(2\beta+1)}$ 使 KL 受常数控制，即可保证 $p_{e,1} \geq c_1$，于是

$$\sup_{p \in P(\beta,L)} \mathbb{E}_p\big(T_n(x_0) - p(x_0)\big)^2 \;\geq\; c_2\,\Delta^2 \asymp h^{2\beta} \asymp n^{-2\beta/(2\beta+1)}.$$

□

14.4.3 结论：点态 minimax 率与核估计的最优性

将定理 14.1（核估计的上界）与定理 14.2（信息论下界）合并，可得

$$c\,n^{-2\beta/(2\beta+1)} \;\leq\; R_n^{\mathrm{pt}}(P(\beta,L);x_0) \;\leq\; C\,n^{-2\beta/(2\beta+1)},$$

其中常数 $c, C > 0$ 只依赖于 β, L（以及核函数的常数），与 n 无关。因此，点态估计的 minimax 收敛阶为

$$R_n^{\mathrm{pt}}(P(\beta,L);x_0) \asymp n^{-2\beta/(2\beta+1)}.$$

更进一步，定理 14.1 给出对核密度估计器 $\hat{p}_{n,h}(x_0)$ 的非渐近风险界

$$\sup_{p \in P(\beta,L)} \mathbb{E}_p\big(\hat{p}_{n,h}(x_0) - p(x_0)\big)^2 \;\leq\; C_2^2 h^{2\beta} + \frac{C_1}{nh},$$

从而当带宽按最优尺度选取

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

两项达到平衡并得到

$$\sup_{p \in P(\beta,L)} \mathbb{E}_p\big(\hat{p}_{n,h}(x_0) - p(x_0)\big)^2 \;\lesssim\; n^{-2\beta/(2\beta+1)}.$$

结合下界（定理 14.2）可知：核密度估计在点态风险意义下达到 *minimax* 速率，因而在误差阶意义下是最优的。

14.5 一维 Hölder 类下 MISE 的 minimax 率：上界与下界闭环

设定与记号 观测 $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} p$，$p$ 为 $[0,1]$ 上密度。记核密度估计器

$$\hat{p}_{n,h}(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right), \qquad h > 0.$$

全局 L_2 风险（MISE）为

$$\mathrm{MISE}(p, \hat{p}_{n,h}) := \mathbb{E}_p \|\hat{p}_{n,h} - p\|_2^2 = \mathbb{E}_p \int_0^1 \big(\hat{p}_{n,h}(x) - p(x)\big)^2\, dx.$$

定义 MISE 的 minimax 风险（全局 L_2）为

$$R_n^{L_2}(\mathcal{P}) := \inf_{\hat{p}_n} \sup_{p \in \mathcal{P}} \mathbb{E}_p \|\hat{p}_n - p\|_2^2.$$

Hölder 类 令 $\beta > 0$，$m = \lfloor \beta \rfloor$。记 $\mathcal{H}^\beta(L)$ 为满足 $p^{(m)}$ 存在且

$$|p^{(m)}(x) - p^{(m)}(y)| \le L|x-y|^{\beta-m}, \qquad \forall x, y \in [0,1]$$

的函数类，并额外假设 $p \ge 0$、$\int_0^1 p = 1$，且 p 有界（例如 $\|p\|_\infty \le B$）。这些边界与有界性条件用于控制核估计的常数项（与速率无关）。

定理 14.3 (一维 Hölder 密度的 MISE minimax 率)**.** 设 $\mathcal{P} = \mathcal{H}^\beta(L)$。若核 K 为 β 阶核并满足 $\int K^2 < \infty$ 等常规条件，则存在常数 $0 < c < C < \infty$ 使得

$$c\, n^{-2\beta/(2\beta+1)} \;\le\; R_n^{L_2}\big(\mathcal{H}^\beta(L)\big) \;\le\; C\, n^{-2\beta/(2\beta+1)}.$$

因此

$$R_n^{L_2}\big(\mathcal{H}^\beta(L)\big) \asymp n^{-2\beta/(2\beta+1)}.$$

证明. 证明分上界与下界两部分。

(A) 上界：第 1 章的 MISE 分解 + 带宽平衡

对任意 p，

$$\mathbb{E}_p\|\hat{p}_{n,h} - p\|_2^2 = \underbrace{\|\mathbb{E}_p\hat{p}_{n,h} - p\|_2^2}_{\mathrm{Bias}^2(h)} + \underbrace{\int_0^1 \mathrm{Var}_p\big(\hat{p}_{n,h}(x)\big)\,dx}_{\mathrm{Var}(h)}.$$

方差项可由直接计算得到（第 1 章的标准结论）

$$\mathrm{Var}(h) \;\le\; \frac{1}{nh}\int K^2.$$

偏差项在 $p \in \mathcal{H}^\beta(L)$ 且 K 为 β 阶核时满足

$$\mathrm{Bias}^2(h) \;\le\; C_\beta\, h^{2\beta}.$$

于是存在常数 $C > 0$ 使得对所有 $p \in \mathcal{H}^\beta(L)$，

$$\mathbb{E}_p\|\hat{p}_{n,h} - p\|_2^2 \;\le\; C\Big(h^{2\beta} + \frac{1}{nh}\Big).$$

令两项平衡，取

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

得到

$$\sup_{p\in\mathcal{H}^\beta(L)} \mathbb{E}_p\|\hat{p}_{n,h} - p\|_2^2 \;\lesssim\; n^{-2\beta/(2\beta+1)}.$$

从而

$$R_n^{L_2}\big(\mathcal{H}^\beta(L)\big) \;\le\; C\, n^{-2\beta/(2\beta+1)}.$$

(B) 下界：第 2 章 testing-based 下界（Assouad：很多个二点检验叠加）

思路：要在全局 L_2 下得到 $n^{-2\beta/(2\beta+1)}$，需要构造一族密度 $\{p_\theta : \theta \in \{0,1\}^m\}$，使得 (i) 相邻（Hamming 距离 1）的两密度难以检验（样本分布很近）；(ii) 但它们在 L_2 意义下的差异可

以在 m 个坐标上累积，最终给出更大的全局下界。这正是第 2 章 Assouad 引理的告诉我们的做法。

构造：取一个光滑 bump ψ，支撑在 $[0,1]$ 内部，满足 $\int \psi = 0$，且 $\psi \in \mathcal{H}^\beta(1)$。令 $h > 0$，把 $[0,1]$ 划分为 $m \asymp 1/h$ 个互不重叠的小区间 $I_1, \dots, I_m$（长度 $\asymp h$），并定义局部 bump

$$\psi_j(x) := h^\beta \, \psi\!\left(\frac{x - a_j}{h}\right), \qquad \operatorname{supp}(\psi_j) \subset I_j,$$

其中 a_j 为第 j 个区间的中心。由缩放可验证 ψ_j 的 $\mathcal{H}^\beta$ 半范数被常数控制，且

$$\|\psi_j\|_2^2 \asymp h^{2\beta+1}.$$

取基准密度 $p_0 \equiv 1$（或一个在内部有余量的光滑密度），并对每个 $\theta = (\theta_1, \dots, \theta_m) \in \{0,1\}^m$ 定义

$$p_\theta(x) := p_0(x) + \varepsilon \sum_{j=1}^{m} (2\theta_j - 1)\, \psi_j(x),$$

其中 $\varepsilon > 0$ 足够小，以保证所有 $p_\theta \geq 0$ 且仍在 $\mathcal{H}^\beta(L)$ 内。

两项关键估计：

(i) L_2 **分离：** 若 θ, θ' 在第 j 位不同（Hamming 距离 1），则

$$\|p_\theta - p_{\theta'}\|_2^2 = 4\varepsilon^2 \|\psi_j\|_2^2 \asymp \varepsilon^2 h^{2\beta+1}.$$

而一般地，若 Hamming 距离为 $d_H(\theta, \theta')$，由于 bump 支撑不交，

$$\|p_\theta - p_{\theta'}\|_2^2 \asymp \varepsilon^2 h^{2\beta+1}\, d_H(\theta, \theta').$$

(ii) **相邻难以检验：** 对相邻的 $\theta, \theta^{(j)}$（只翻转第 j 位），二者的单样本 KL（或 χ^2）距离满足

$$K(P_\theta, P_{\theta^{(j)}}) \;\lesssim\; \varepsilon^2 \|\psi_j\|_2^2 \asymp \varepsilon^2 h^{2\beta+1},$$

于是 n 样本下

$$K(P_\theta^{(n)}, P_{\theta^{(j)}}^{(n)}) = n K(P_\theta, P_{\theta^{(j)}}) \;\lesssim\; n\varepsilon^2 h^{2\beta+1}.$$

选择

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

即可保证相邻假设的 KL 受常数控制，从而对应的二点检验误差被常数下界住（第 2 章二点法/Assouad 框架的标准结论）。

套用 *Assouad*：Assouad 引理把 minimax 风险下界为“每一位二点检验难度 × 每一位造成的损失”之和，从而得到

$$R_n^{L_2}\big(\mathcal{H}^\beta(L)\big) \;\gtrsim\; m \cdot \varepsilon^2 h^{2\beta+1}.$$

代入 $m \asymp 1/h$ 与 $\varepsilon^2 \asymp (n h^{2\beta+1})^{-1}$，得

$$R_n^{L_2}\big(\mathcal{H}^\beta(L)\big) \;\gtrsim\; \frac{1}{h} \cdot \frac{1}{n h^{2\beta+1}} \cdot h^{2\beta+1} = \frac{1}{nh}.$$

最后选取与上界相同的尺度 $h \asymp n^{-1/(2\beta+1)}$，得到

$$R_n^{L_2}\big(\mathcal{H}^\beta(L)\big) \;\gtrsim\; n^{-2\beta/(2\beta+1)}.$$

(C) 合并上下界即可得结论。 □

注释：为什么下界要“很多个二点检验叠加”？ 点态风险的两点构造只需在一个点附近制造差异即可；但全局 L_2 风险会对 $[0,1]$ 上的误差积分，因此最有效的对抗构造是把误差分布在很多互不重叠的小区间上，使 L_2 差异累积。Assouad 正是把这一思想形式化：它把全局估计难度分解为许多坐标上的二点检验难度之和。

15 深度 ReLU 神经网络回归的收敛率：逼近能力、复杂度与（近）minimax 最优

本节用深度 ReLU 网络做一个“minimax 上界套路”的完整闭环例子：在 Hölder 光滑回归模型下，构造一个受控（稀疏、有界、深度/宽度随 n 变）的网络类 $\mathcal{F}_n$，对其做剪裁最小二乘 ERM，并证明其风险满足

$$\sup_{f_0\in\mathcal{H}^\beta([0,1]^d)} \mathbb{E}\|\hat{f}_n - f_0\|_{L_2(P_X)}^2 \;\lesssim\; n^{-\frac{2\beta}{2\beta+d}} \cdot (n),$$

即与 Stone 率同阶（至多差对数因子）。该结论与 Schmidt–Hieber (2020) 的总体结论一致：稀疏连接的深 ReLU 网络通过合适架构选择可达到 minimax 率（至多差 $\log n$ 因子）。[15]

15.1 问题设定与风险

观测 $(X_i, Y_i)_{i=1}^n$ 满足

$$Y_i = f_0(X_i) + \varepsilon_i, \qquad \mathbb{E}[\varepsilon_i \mid X_i] = 0,$$

其中 $X_i \in [0,1]^d$，$f_0 : [0,1]^d \to \mathbb{R}$。记 P_X 为设计分布，考虑平方损失下的预测风险

$$(\hat{f}, f_0) = \|\hat{f} - f_0\|_{L_2(P_X)}^2 = \mathbb{E}\big[(\hat{f}(X) - f_0(X))^2\big].$$

假设（标准回归设定）

(A1) 设计分布 P_X 在 $[0,1]^d$ 上有密度 p_X，且 $0 < p_{\min} \le p_X(x) \le p_{\max} < \infty$；

(A2) 噪声 ε_i 条件于 X_i 次高斯：$\mathbb{E}[\exp(t\varepsilon_i) \mid X_i] \le \exp(\sigma^2 t^2/2)$；

(A3) 回归函数有界：$\|f_0\|_\infty \le F$。

为避免尾部技术，我们对估计器做剪裁（truncation），保证候选函数输出有界，从而便于用覆盖数与经验过程控制统计波动。

15.2 基准：Hölder 光滑回归的 minimax 率（Stone 标尺）

设 $f_0 \in \mathcal{H}^\beta([0,1]^d)$，则在 (A1)–(A2) 等正则条件下，非参数回归的 minimax $L_2(P_X)$ 风险满足 Stone 标尺

$$\inf_{\hat f}\ \sup_{f_0\in\mathcal{H}^\beta} \mathbb{E}\|\hat f - f_0\|_{L_2(P_X)}^2 \ \asymp\ n^{-\frac{2\beta}{2\beta+d}}. \tag{Stone}$$

因此任何“统计最优”的方法都不应在主阶上慢于该速率。[17,19]

15.3 网络类、稀疏约束与剪裁 ERM

ReLU 网络类与参数维度 令 $\sigma(x)=\max\{x,0\}$。给定深度 L、每层宽度向量 $\boldsymbol{p}=(p_0=d,p_1,\dots,p_L,p_{L+1}=1)$，参数集合 $\theta=(W_\ell,b_\ell)_{\ell=0}^{L}$ 定义函数

$$f_\theta(x) = W_L\,\sigma\big(W_{L-1}\sigma(\cdots\sigma(W_0x+b_0)\cdots)+b_{L-1}\big)+b_L.$$

记网络总参数个数

$$D(\boldsymbol{p},L) := \sum_{\ell=0}^{L}(p_{\ell+1}p_\ell + p_{\ell+1}).$$

稀疏与有界权重（Schmidt–Hieber 的统计设定） 为控制复杂度，我们限制

$$\|\theta\|_\infty \le B, \qquad \|\theta\|_0 \le S,$$

其中 $\|\theta\|_0$ 记非零参数个数（sparsity），$\|\theta\|_\infty$ 为所有权重/偏置的绝对值上界。对应函数类记为

$$\mathcal{F}(L,\boldsymbol{p},S,B) := \{f_\theta:\ f_\theta \text{ 为 ReLU 网络，且 } \|\theta\|_0\le S,\ \|\theta\|_\infty\le B\}.$$

剪裁最小二乘 ERM 定义剪裁算子 $_F(u) := \max\{-F,\min\{u,F\}\}$。取 $\mathcal{F}_n := \mathcal{F}(L_n,\boldsymbol{p}_n,S_n,B_n)$，考虑估计器

$$\hat f_n \in \arg\min_{g\in\mathcal{F}_n}\ \frac{1}{n}\sum_{i=1}^{n}\big(Y_i - {}_F\,(g(X_i))\big)^2, \qquad \hat f_n :=_F (\hat f_n).$$

15.4 两块关键输入：逼近引理与复杂度引理

15.4.1 逼近引理：深 ReLU 逼近 Hölder 函数（更精确的参数标度）

下面给一个足够用于统计推导的版本：存在深 ReLU 网络以多项式规模逼近 $\mathcal{H}^\beta$，并且深度只需对数级别增长。该类型结论来自 Yarotsky 关于 Sobolev/Hölder 类的构造性逼近界。[24-25]

引理 15.1 (Hölder 函数的 ReLU 逼近（构造性）). *令 $f_0\in\mathcal{H}^\beta([0,1]^d)$ 且 $\|f_0\|_\infty\le F$。则对任意 $\varepsilon\in(0,1)$，存在常数 C_1,C_2（仅依赖于 β,d 与 Hölder 常数），以及一个 ReLU 网络 f_θ 满足*

$$\|f_\theta - f_0\|_\infty \le \varepsilon,$$

并且该网络可以取

$$L \ \le\ C_1\log(1/\varepsilon), \qquad S\ \le\ C_2\,\varepsilon^{-d/\beta}\log(1/\varepsilon), \qquad \|\theta\|_\infty \le C_2.$$

证明. 证明属于构造性逼近理论：用局部多项式/分片多项式逼近 Hölder 函数，再用 ReLU 网络实现分片与多项式基（必要时实现乘法模块），并通过网络组合实现全局拼接。Yarotsky 给出对 Sobolev/Hölder 类的上界；其中关键标度是：深度只需 $\log(1/\varepsilon)$，有效自由度（可由非零参数数 S 度量）为 $\varepsilon^{-d/\beta}$ 的多项式阶（至多差对数因子）。[24-25] □

15.4.2 复杂度引理：稀疏 ReLU 网络的覆盖数（带上 D, S, B 的显式依赖）

参数到函数的 Lipschitz 控制 为把“参数网格误差”转成“函数输出误差”，我们需要网络对参数的 Lipschitz 性。对有界输入 $x \in [0,1]^d$，若各层权重幅度受控，网络输出对参数变化具有多项式（常写作指数于 L）的放大因子。

引理 15.2 (受控 ReLU 网络的参数 Lipschitz 性). 固定 $(L, \boldsymbol{p})$。存在常数 $C = C(d, \boldsymbol{p})$ 使得：若 $\|\theta\|_\infty \le B$ 且 $\|\theta'\|_\infty \le B$，则对所有 $x \in [0,1]^d$，

$$|f_\theta(x) - f_{\theta'}(x)| \;\le\; C\,(1+B)^L\,\|\theta - \theta'\|_\infty.$$

证明. 逐层递推即可：ReLU 为 1-Lipschitz，线性层在 $\|\cdot\|_\infty$ 下的放大因子由行和范数控制，每层最多带来一个与 $(1+B)$ 同阶的因子；迭代 L 次得出上界。更精确的常数依赖可见网络复杂度文献与 Schmidt–Hieber 的推导。[15] □

引理 15.3 (稀疏 ReLU 网络的覆盖数（显式）). 设 $\mathcal{F} = \mathcal{F}(L, \boldsymbol{p}, S, B)$，并假设 $\sup_{f\in\mathcal{F}} \|f\|_\infty \le F$。则存在常数 $C > 0$ 使得对任意 $\varepsilon \in (0,1)$，

$$\log N\big(\varepsilon, \mathcal{F}, \|\cdot\|_\infty\big) \;\le\; \log \binom{D(\boldsymbol{p}, L)}{S} + S \log\!\Big(\frac{C(1+B)^L}{\varepsilon}\Big) \;\le\; C\,S\,\log\!\Big(\frac{eD(\boldsymbol{p}, L)}{S}\Big) + C\,S\,L\,\log\!\Big(\frac{C(1+B)}{\varepsilon}\Big).$$

证明. 先选非零坐标集合 $I \subset \{1, \dots, D\}$，$|I| = S$，共有 $\binom{D}{S}$ 种可能。对固定的 I，把参数限制在 $[-B, B]^S$ 的 S 维立方体内，用 $\|\cdot\|_\infty$ 网格以步长 $\delta := \varepsilon / \{C(1+B)^L\}$ 覆盖，则网格点数至多 $(2B/\delta + 1)^S$。由引理 15.2，对应函数输出误差不超过 ε。两步相乘并取对数即可；第二个不等式用 $\log \binom{D}{S} \le S \log(eD/S)$。[15] □

15.5 经验过程主模块：剪裁 ERM 的 oracle 不等式（逐步证明）

下面把“平方损失 ERM”化为“经验过程上界”。为使推导自洽，我们把常用步骤拆成三条引理：对称化、收缩、以及 Dudley 熵积分。读者熟悉的话可直接跳到引理 15.7。

引理 15.4 (对称化). 令 $\mathcal{G}$ 为一族可测函数，且 $Z_1, \dots, Z_n$ *i.i.d.* $\sim P$。记 $P_n g = \frac{1}{n}\sum_{i=1}^n g(Z_i)$。则

$$\mathbb{E} \sup_{g\in\mathcal{G}} (P - P_n) g \;\le\; 2\,\mathbb{E} \sup_{g\in\mathcal{G}} \frac{1}{n} \sum_{i=1}^n \xi_i g(Z_i),$$

其中 ξ_i 为独立 *Rademacher* 变量（取值 ± 1 且等概率），并与数据独立。

证明. 标准对称化：引入一份独立副本 $Z_1', \dots, Z_n'$，用 Jensen 与交换期望，将 $P - P_n$ 写成两份经验均值差，再引入 Rademacher 随机符号。详见经验过程教材。[6,20] □

引理 15.5 (平方损失的收缩（剪裁保证有界）)**.** 设 $|Y| \le F + \sigma$ 近似有界（或用次高斯尾界 + 剪裁化归），且 $\mathcal{F}$ 中函数均满足 $\|f\|_\infty \le F$。定义

$$\ell_f(x,y) := (y - f(x))^2 - (y - f_0(x))^2.$$

则存在常数 C_F 使得对任意样本 (X_i, Y_i)，

$$\mathbb{E}_\xi \sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n \xi_i\, \ell_f(X_i,Y_i) \;\le\; C_F\, \mathbb{E}_\xi \sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n \xi_i\,(f(X_i) - f_0(X_i)).$$

证明. 展开 $\ell_f = (f-f_0)^2 - 2(Y-f_0)(f-f_0)$。第一项由 $|f - f_0| \le 2F$ 的有界性与 contraction（或直接用 sup 的平凡界）控制；第二项对 $f - f_0$ 是线性的，系数 $Y - f_0$ 在剪裁/次高斯条件下可由 L_2 或 Orlicz 范数控制，最终得到常数 C_F。严谨版本可在平方损失 ERM 的教材推导中找到。[6,20] □

引理 15.6 (Dudley 熵积分（从覆盖数到 Rademacher 复杂度）)**.** 若 $\mathcal{F}$ 为以 $\|\cdot\|_\infty$ 度量的有界函数类，$\sup_{f\in\mathcal{F}} \|f\|_\infty \le F$，则存在常数 $C > 0$ 使得

$$\mathbb{E} \sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n \xi_i f(X_i) \;\le\; \frac{C}{\sqrt{n}} \int_0^F \sqrt{\log N(\varepsilon, \mathcal{F}, \|\cdot\|_\infty)}\, d\varepsilon.$$

证明. Dudley entropy integral：对 $\mathcal{F}$ 做逐尺度网（chaining），把 Rademacher 和分解为多尺度增量之和，再用每尺度上的覆盖数控制期望上界。[6,20] □

引理 15.7 (剪裁最小二乘的 oracle inequality（严格版）)**.** 在 *(A1)–(A3)* 下，令 $\hat{f}_n$ 为 $\mathcal{F}_n$ 上的剪裁最小二乘 *ERM*。则存在常数 $C > 0$ 使得

$$\mathbb{E}\|\hat{f}_n - f_0\|_{L_2(P_X)}^2 \;\le\; 2 \inf_{g\in\mathcal{F}_n} \|g - f_0\|_{L_2(P_X)}^2 \;+\; C\,\frac{\mathcal{J}(\mathcal{F}_n)^2}{n},$$

其中

$$\mathcal{J}(\mathcal{F}_n) := \int_0^F \sqrt{\log N\big(\varepsilon, \mathcal{F}_n, \|\cdot\|_\infty\big)}\, d\varepsilon.$$

证明. 令 $\ell_f(x,y) = (y - f(x))^2$。由 ERM 基本不等式 $P_n \ell_{\hat{f}_n} \le P_n \ell_g$ $(\forall g \in \mathcal{F}_n)$，得

$$P\ell_{\hat{f}_n} - P\ell_{f_0} \le P\ell_g - P\ell_{f_0} + (P - P_n)(\ell_g - \ell_{f_0}) - (P - P_n)(\ell_{\hat{f}_n} - \ell_{f_0}).$$

取 g 为 $\mathcal{F}_n$ 上的近似最优点并对右侧经验过程项取上确界。对经验过程项用引理 15.4 对称化，再用引理 15.5 把平方损失差收缩到线性类，最后用引理 15.6 由覆盖数给出熵积分上界。把得到的上界代回并整理，即得

$$\mathbb{E}\|\hat{f}_n - f_0\|_{L_2(P_X)}^2 \;\le\; 2 \inf_{g\in\mathcal{F}_n} \|g - f_0\|_{L_2(P_X)}^2 + C\frac{\mathcal{J}(\mathcal{F}_n)^2}{n}.$$

□

15.6 主定理：深 ReLU 回归在 Hölder 类上达（近）minimax 率

定理 15.1 (深 ReLU 回归在 Hölder 类上的（近）Stone 率)**.** *在 (A1)–(A3) 下，设 $f_0 \in \mathcal{H}^\beta([0,1]^d)$ 且 $\|f_0\|_\infty \le F$。令 $\mathcal{F}_n = \mathcal{F}(L_n, \boldsymbol{p}_n, S_n, B_n)$ 为稀疏 ReLU 网络类并取剪裁 ERM $\hat{f}_n$。若*

$$L_n \asymp \log n, \qquad S_n \asymp n^{\frac{d}{2\beta+d}} \log n, \qquad B_n \asymp n^{c_0} \ \ (\text{某个常数} c_0 > 0),$$

且 $D(\boldsymbol{p}_n, L_n)$ 多项式增长（例如 $D(\boldsymbol{p}_n, L_n) \le n^{c_2}$），则存在常数 $C > 0$ 使得

$$\sup_{f_0 \in \mathcal{H}^\beta([0,1]^d)} \mathbb{E}\|\hat{f}_n - f_0\|_{L_2(P_X)}^2 \ \le \ C\, n^{-\frac{2\beta}{2\beta+d}} \cdot (\log n)^{c_1},$$

其中 $c_1 > 0$ 仅依赖于 β, d 与常数项。因此，$\hat{f}_n$ 在主阶上达到 Stone minimax 率（至多差对数因子），与 Schmidt–Hieber (2020) 的结论一致。[15]

证明. 用引理 15.7 把误差拆为“逼近 + 估计”：

$$\mathbb{E}\|\hat{f}_n - f_0\|_{L_2(P_X)}^2 \ \le \ 2 \inf_{g \in \mathcal{F}_n} \|g - f_0\|_{L_2(P_X)}^2 + C\frac{\mathcal{J}(\mathcal{F}_n)^2}{n}.$$

(1) 逼近误差： 令 $\varepsilon_n := n^{-\beta/(2\beta+d)}$（忽略对数项）。由引理 15.1，存在 $g_n \in \mathcal{F}_n$ 使得

$$\|g_n - f_0\|_\infty \le \varepsilon_n \quad \Rightarrow \quad \|g_n - f_0\|_{L_2(P_X)}^2 \le \varepsilon_n^2 = n^{-\frac{2\beta}{2\beta+d}},$$

且该逼近只需

$$S_n \gtrsim \varepsilon_n^{-d/\beta} \log(1/\varepsilon_n) \asymp n^{\frac{d}{2\beta+d}} \log n, \qquad L_n \gtrsim \log(1/\varepsilon_n) \asymp \log n,$$

与定理假设相符。

(2) 估计误差： 由引理 15.3，

$$\log N(\varepsilon, \mathcal{F}_n, \|\cdot\|_\infty) \ \lesssim \ S_n \log\Big(\frac{eD(\boldsymbol{p}_n, L_n)}{S_n}\Big) + S_n L_n \log\Big(\frac{C(1+B_n)}{\varepsilon}\Big).$$

于是熵积分满足（用 $\int_0^F \sqrt{\log(A/\varepsilon)}\, d\varepsilon \lesssim F\sqrt{\log(A/F)}$ 之类粗界）

$$\mathcal{J}(\mathcal{F}_n)^2 \ \lesssim \ F^2\, S_n L_n \log\big(C(1+B_n)D(\boldsymbol{p}_n, L_n)\big).$$

代回引理 15.7 得

$$\mathbb{E}\|\hat{f}_n - f_0\|_{L_2(P_X)}^2 \ \lesssim \ n^{-\frac{2\beta}{2\beta+d}} + \frac{F^2}{n} S_n L_n \log\big(C(1+B_n)D(\boldsymbol{p}_n, L_n)\big).$$

(3) 代入参数增长并整理： 取 $S_n \asymp n^{\frac{d}{2\beta+d}} \log n$、$L_n \asymp \log n$，并用 B_n 与 $D(\boldsymbol{p}_n, L_n)$ 的多项式增长，则

$$\frac{1}{n} S_n L_n \log\big(C(1+B_n)D(\boldsymbol{p}_n, L_n)\big) \ \asymp \ n^{-\frac{2\beta}{2\beta+d}} \cdot (\log n)^{c_1},$$

故

$$\sup_{f_0\in\mathcal{H}^\beta}\mathbb{E}\|\hat{f}_n-f_0\|_{L_2(P_X)}^2 \lesssim n^{-\frac{2\beta}{2\beta+d}}\cdot(\log n)^{c_1}.$$

证毕。 □

本节小结：NN 收敛率在 minimax 叙事中的位置

- Stone 率 $n^{-2\beta/(2\beta+d)}$ 给出 Hölder 回归的 minimax 基准；
- 深 ReLU 的统计上界仍然是“逼近 + 估计”：逼近由构造性逼近理论（引理 15.1）给出，估计误差由覆盖数 + 熵积分 + ERM oracle 不等式（引理 15.3 与 15.7）给出；
- 适当选择深度（$\log n$ 级）与稀疏度（$n^{d/(2\beta+d)}$ 级）即可得到（近）minimax 率，与 Schmidt–Hieber (2020) 的结论一致。[15]

16 过参数化与 double descent：两层神经网络的非渐近理论

本节把 double descent 放回到一个你熟悉的“**结构类 + 正则化 ERM + 非渐近误差界**”框架中。我们以 Wang–Lin（2023）为主线：对两层 ReLU 网络施加 **scaled variation**（可理解为一种“变分范数/路径范数”）正则，可以得到一条同时包含

$$\text{（近似误差）} + \text{（估计误差两段式饱和）}$$

的非渐近上界，从而在同一条不等式里解释“尖峰 + 第二次下降”。[21]

16.1 模型、函数类与 scaled variation 正则

回归模型与风险 观测 $(X_i,Y_i)_{i=1}^n$，满足

$$Y_i=f^*(X_i)+\varepsilon_i,\qquad \mathbb{E}[\varepsilon_i\mid X_i]=0.$$

记总体与经验 L_2 风险

$$\|f-g\|_2^2:=\mathbb{E}\big[(f(X)-g(X))^2\big],\qquad \|f-g\|_n^2:=\frac{1}{n}\sum_{i=1}^n(f(X_i)-g(X_i))^2.$$

本节的“误差”都指 $\|\hat{f}-f^*\|_2^2$。

两层 ReLU 网络与 scaled variation 考虑带常数项的两层 ReLU 网络

$$g(x;\theta)=\sum_{k=1}^m a_k\,\sigma(v_k^\top x+b_k)+c,\qquad \sigma(t)=\max\{t,0\}.$$

为便于把复杂度“抽象成范数”，我们将第一层参数限制在

$$(v_k,b_k)\in\mathbb{S}^{d-1}\times[-1,1],$$

并定义有限宽度的 scaled variation（等价于 Wang–Lin 的设定之一）为

$$\nu(\theta) := \sum_{k=1}^{m} |a_k|.$$

（若你更偏好 $\sum |a_k| \|w_k\|_2$ 的写法，可通过归一化把 $\|w_k\|_2$ 吸收到基函数里；此处不影响“速率/两段式结构”。）

S-norm 与函数空间 $\mathcal{G}$ 将“无限宽网络”写成积分表示：令 μ 为符号测度，

$$f(x) = \int_{\mathbb{S}^{d-1}\times[-1,1]} \sigma(v^\top x + b)\, d\mu(v,b) + c.$$

定义其 S-norm（变分范数）

$$\|f\|_S := \inf\Big\{ \|\mu\|_{\mathrm{TV}} : f(x) = \int \sigma(v^\top x + b)\, d\mu(v,b) + c \Big\},$$

并令

$$\mathcal{G} := \{f : \|f\|_S < \infty\}, \qquad \mathcal{G}(V) := \{f \in \mathcal{G} : \|f\|_S \le V\}.$$

有限宽度的子类记为

$$\mathcal{G}_m(V) := \Big\{ g(\cdot;\theta) : \nu(\theta) \le V,\ \text{宽度} \le m \Big\}.$$

注意：若 $g(\cdot;\theta) \in \mathcal{G}_m(V)$，则对应的离散测度 $\mu = \sum_{k=1}^m a_k \delta_{(v_k,b_k)}$ 满足 $\|\mu\|_{\mathrm{TV}} = \sum |a_k| = \nu(\theta)$，因此 $\|g\|_S \le \nu(\theta)$。

正则化 ERM（scaled variation 正则） 定义估计器

$$\hat{f} \in \arg\min_{g \in \mathcal{G}_m} \Big\{ \frac{1}{2n} \sum_{i=1}^{n} (Y_i - g(X_i))^2 + \lambda\, \nu(g) \Big\},$$

其中 $\nu(g)$ 表示某个表示下的 $\nu(\theta)$，并取其下确界（与上面的 $\|\cdot\|_S$ 同理）。

16.2 主定理：一条非渐近界给出 double descent

下面给出“可直接读出 double descent 两段机制”的定理（对应 Wang–Lin 的主结果形态；我们用本章记号重述）。[21]

定理 16.1 (两层 ReLU + scaled variation 正则的两段式非渐近界)**.** 假设：

1. $X_i \overset{i.i.d.}{\sim} \mathrm{Unif}(B_d)$（单位球上均匀分布）；
2. ε_i 独立且次高斯（例如 $\varepsilon_i \sim N(0,\sigma_\varepsilon^2)$）；
3. $f^* \in \mathcal{G}(V)$，即 $\|f^*\|_S \le V$；
4. λ 取

$$\lambda \asymp \min\Big\{ \sqrt{\frac{d\log(en/d)}{n}},\ \frac{m\, d \log(en/d)}{n} \Big\}.$$

则存在与 d 无关的常数 $C>0$ 使得

$$\mathbb{E}\|\hat{f}-f^*\|_2^2 \;\le\; C\Bigg[\underbrace{V^2\, m^{-(d+3)/d}}_{\text{近似误差：随 } m\uparrow\downarrow} \;+\; \underbrace{(\sigma_\varepsilon^2+V^2)\,\min\Big\{\sqrt{\frac{d\log(en/d)}{n}},\; \frac{m\,d\log(en/d)}{n}\Big\}}_{\text{估计误差：两段式}}\Bigg]. \tag{34}$$

如何从 (34) 读出 double descent？ 固定 n,d,V：

- 当 m 较小，估计误差主导且 $\propto m/n$（更精确为 $md\log n/n$），随 m 增大会变差；
- 当 m 足够大，估计误差饱和在 $\sqrt{(d\log n)/n}$，不再随 m 增大；
- 同时近似误差 $V^2m^{-(d+3)/d}$ 仍继续下降，于是出现“第二次下降”。

16.3 证明：三步闭环（基本不等式 + 近似引理 + 估计引理）

证明（分三步）. 记 P 为总体期望、P_n 为经验平均。令平方损失 $\ell_y(g)=\frac{1}{2}(y-g)^2$，则目标函数为 $P_n\ell_Y(g)+\lambda\nu(g)$。

Step 1：基本不等式（regularized ERM 的标准起点）

取任意对照函数 $g\in\mathcal{G}_m$，由 $\hat{f}$ 的最优性有

$$P_n\ell_Y(\hat{f})+\lambda\nu(\hat{f}) \;\le\; P_n\ell_Y(g)+\lambda\nu(g).$$

展开 $Y=f^*+\varepsilon$ 并整理（把 P_n 的平方展开成“平方差 + 噪声线性项”）可得

$$\frac{1}{2}\|\hat{f}-f^*\|_n^2 \;\le\; \frac{1}{2}\|g-f^*\|_n^2+\underbrace{\frac{1}{n}\sum_{i=1}^n\varepsilon_i\big(\hat{f}(X_i)-g(X_i)\big)}_{(\star)}+\lambda\big(\nu(g)-\nu(\hat{f})\big). \tag{35}$$

因此要控制 $\|\hat{f}-f^*\|_2^2$，只需：(i) 选一个好的 g 使 $\|g-f^*\|$ 小（近似）；(ii) 控制噪声项 $(\star)$（经验过程）；(iii) 选择 λ 使正则项能“吸收” $(\star)$。

Step 2：近似引理（无限宽 $\Rightarrow$ 有限宽，给出 $m^{-(d+3)/d}$）

关键是：$f^*\in\mathcal{G}(V)$ 意味着 f^* 可写成“ReLU ridge function 的变分组合”，于是可用 Maurey-type 抽样/稀疏化把积分表示近似成 m 项求和。

引理 16.1 (scaled variation 类的有限宽近似)**.** *对任意 $f\in\mathcal{G}(V)$，存在 $g_m\in\mathcal{G}_m(CV)$ 使得*

$$\|f-g_m\|_2^2 \;\le\; C\,V^2\,m^{-(d+3)/d}.$$

证明要点：把 f 写成积分表示 $f(x)=\int\phi_w(x)\,d\mu(w)$，其中 $\phi_w(x)=\sigma(v^\top x+b)$。在 $L_2(P_X)$ 这个 Hilbert 空间中，集合 $\{\phi_w\}$ 具有可控的“方向复杂度”（由 d 决定）。对 $\mu/\|\mu\|_{\mathrm{TV}}$ 抽样得到 $w_1,\dots,w_m$，取经验平均并重标度得到

$$g_m(x)=\frac{\|\mu\|_{\mathrm{TV}}}{m}\sum_{k=1}^m s_k\,\phi_{w_k}(x),\quad s_k\in\{\pm1\},$$

则 $\nu(g_m)\le\|\mu\|_{\mathrm{TV}}\le V$，且均方误差由抽样方差 + 几何熵控制，得到 $m^{-(d+3)/d}$ 速率（完整推导见[21] 附录/补充材料）。 □

取 $g = g_m$ 代入 (35)，近似项就给出了定理右侧的第一项。

Step 3：估计引理（噪声项给出“两段式”）

现在控制 (35) 中的噪声线性项 $(\star)$。定义局部类

$$\mathcal{F}(t) := \{h = \hat{f} - g : \ \nu(\hat{f}) \le t, \ \nu(g) \le t\}.$$

用对称化 + Rademacher/高斯复杂度可将

$$\sup_{h \in \mathcal{F}(t)} \Big| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i h(X_i) \Big|$$

上界为“局部复杂度函数”。平方损失的强凸性允许把“函数振幅的复杂度”转成“L_2 误差的快率”，典型形式是

$$\|\hat{f} - f^*\|_2^2 \ \lesssim \ (\text{近似}) \ + \ (\sigma_\varepsilon^2 + V^2) \cdot \Psi(n, d, m),$$

其中 Ψ 由熵积分/局部 Rademacher 给出。

Wang–Lin 的关键观察是：对 $\mathcal{G}_m(V)$ 与 $\mathcal{G}(V)$，覆盖数（metric entropy）在两种尺度下表现不同，从而导出两种复杂度：

1. **有限宽度主导（欠参数化）：** $\mathcal{G}_m(V)$ 的熵近似为 md 量级，配合局部化给出快率 $\Psi \asymp md \log(en/d)/n$；
2. **无限宽度饱和（过参数化）：** 当 m 足够大时，$\mathcal{G}_m(V)$ 已能逼近 $\mathcal{G}(V)$ 的“有效局部复杂度”，而后者的局部复杂度只产生 $\Psi \asymp \sqrt{d \log(en/d)/n}$ 的饱和项（不再随 m 增大）。

将两者取最小，就得到 (34) 中的两段式项。选择

$$\lambda \asymp \min\Big\{ \sqrt{\frac{d \log(en/d)}{n}}, \ \frac{m\, d \log(en/d)}{n} \Big\}$$

可使正则项 $\lambda(\nu(g) - \nu(\hat{f}))$ 吸收噪声项的主导部分，从而闭合不等式。

合并 Step 1–3： 用 Lemma 16.1 控制近似误差，用上面的估计引理控制噪声项并选取 λ，即可得到 (34)。这完成定理证明。 □

16.4 minimax：第二个 valley 的“近最优性”

上面的上界解释了 double descent 的形状；下一步是信息论问题：*在结构类 $\mathcal{G}(V) = \{f : \|f\|_S \le V\}$ 上，任何估计器在最坏情形下至多能做到多好？* Wang–Lin 给出了一个（对数意义下）匹配其过参数化饱和项的 minimax 下界。[21] 为便于与上界对齐，本小节显式写出 V 与噪声方差依赖。

定理 16.2 (在 $\mathcal{G}(V)$ 上的 minimax 下界)**.** *设 $X_i \overset{i.i.d.}{\sim} \mathrm{Unif}(B_d)$ 且 $\varepsilon_i \overset{i.i.d.}{\sim} N(0, \sigma^2)$，令 $Y_i = f(X_i) + \varepsilon_i$。则存在常数 $c > 0$（与 n, V, σ 无关）使得：当 n 足够大时，*

$$\inf_{\tilde{f}} \sup_{f \in \mathcal{G}(V)} \mathbb{E}\|\tilde{f} - f\|_2^2 \ \ge \ c\, \sigma^2\, V^0 \cdot \frac{1}{\sqrt{n \log n}}, \tag{36}$$

其中 $\|\cdot\|_2$ 为 $L_2(P_X)$ 范数（P_X 为 $\mathrm{Unif}(B_d)$）。（注：原文表述与 (36) 等价到同阶；差异体现在常数与对数写法上。）[21]

含义（与第二个 valley 对比） 定理 16.1 的过参数化分支饱和项为

$$(\sigma^2 + V^2)\sqrt{\frac{d\log(en/d)}{n}}.$$

而 (36) 表明：任何方法在 $\mathcal{G}(V)$ 上都不可能把最坏情形风险压到 $o\big((n\log n)^{-1/2}\big)$。因此（忽略维数与对数细节）第二个 valley 达到的量级在该结构类上是“近 minimax 最优”的。

证明（*Fano/packing*：在 $\mathcal{G}(V)$ 内构造大量难区分假设）. 证明采用标准的 Fano 下界套路：构造一族函数 $\{f_\theta : \theta \in \Theta\} \subset \mathcal{G}(V)$，使得 (i) 在 $L_2(P_X)$ 下两两分离；(ii) 诱导的 n 样本分布彼此 KL 距离小。随后由 Fano 引理推出任意估计器的最坏情形风险下界。

Step 1：选择一组“近似正交”的 ReLU 基函数

取 m 个方向 $v_1, \dots, v_m \in \mathbb{S}^{d-1}$，使得两两夹角足够分离（球面 packing）。令

$$\phi_j(x) := \sigma(v_j^\top x), \qquad j = 1, \dots, m.$$

在 $X \sim \mathrm{Unif}(B_d)$ 下，利用旋转对称性与球面 packing 的近正交性，可选取上述 $\{v_j\}$ 使得 Gram 矩阵

$$G_{jk} := \langle \phi_j, \phi_k \rangle_{L_2(P_X)} = \mathbb{E}[\phi_j(X)\phi_k(X)]$$

满足：存在绝对常数 $0 < c_0 < C_0 < \infty$，

$$c_0\, I_m \;\preceq\; G \;\preceq\; C_0\, I_m. \tag{37}$$

这一步的直观含义是：这组 ReLU ridge functions 在 $L_2(P_X)$ 下“有效维数”约为 m。

Step 2：在 $\mathcal{G}(V)$ 内构造一个大的 packing 集

对每个 $\theta \in \{0,1\}^m$，定义

$$f_\theta(x) := \delta \sum_{j=1}^{m} (2\theta_j - 1)\, \phi_j(x),$$

其中 $\delta > 0$ 待定。由于 ϕ_j 都是 ReLU 单元且系数为 $\delta(2\theta_j - 1)$，其 S-norm（变分范数）满足

$$\|f_\theta\|_S \;\le\; \delta \sum_{j=1}^{m} |2\theta_j - 1| \;=\; \delta m.$$

令

$$\delta := \frac{V}{m}, \tag{38}$$

则对所有 θ 都有 $f_\theta \in \mathcal{G}(V)$。

接着计算 L_2 分离。由 (37)，对任意 $\theta \neq \theta'$，

$$\|f_\theta - f_{\theta'}\|_2^2 = \delta^2 \Big\| \sum_{j=1}^{m} 2(\theta_j - \theta'_j)\, \phi_j \Big\|_2^2 \;\ge\; 4c_0\, \delta^2\, d_H(\theta, \theta'),$$

其中 d_H 是 Hamming 距离。由 Varshamov–Gilbert 引理，存在子集 $\Theta \subset \{0,1\}^m$，满足

$$\log|\Theta| \;\geq\; c_1 m, \qquad d_H(\theta,\theta') \;\geq\; \frac{m}{8} \quad (\theta \neq \theta').$$

因此在该子集上两两分离至少为

$$\|f_\theta - f_{\theta'}\|_2^2 \;\geq\; 4c_0\,\delta^2 \cdot \frac{m}{8} \;=\; c_2\,\delta^2 m \;=\; c_2\,\frac{V^2}{m}. \tag{39}$$

Step 3：控制诱导分布的 KL 距离

令 $P_\theta^{(n)}$ 表示在真函数为 f_θ 时 $(X_i,Y_i)_{i=1}^n$ 的联合分布。由于 X_i 的边缘分布与 θ 无关，且在给定 X 下

$$Y_i \mid X_i \sim N(f_\theta(X_i),\sigma^2),$$

两模型的（条件）KL 距离为

$$K(P_\theta^{(n)},P_{\theta'}^{(n)}) = \frac{1}{2\sigma^2}\sum_{i=1}^n \mathbb{E}\big[(f_\theta(X_i)-f_{\theta'}(X_i))^2\big] = \frac{n}{2\sigma^2}\|f_\theta - f_{\theta'}\|_2^2.$$

利用上界方向的 (37)，以及一般 θ,θ' 的 $d_H(\theta,\theta') \leq m$，得到

$$\|f_\theta - f_{\theta'}\|_2^2 \leq 4C_0\,\delta^2\,d_H(\theta,\theta') \;\leq\; 4C_0\,\delta^2 m = 4C_0\frac{V^2}{m}.$$

因此

$$K(P_\theta^{(n)},P_{\theta'}^{(n)}) \;\leq\; \frac{n}{2\sigma^2}\cdot 4C_0\frac{V^2}{m} = C_3\,\frac{nV^2}{\sigma^2 m}. \tag{40}$$

Step 4：选择 m 并应用 Fano 引理

Fano 引理（见你前面 testing 章）告诉我们：若

$$\frac{1}{|\Theta|^2}\sum_{\theta\neq\theta'} K(P_\theta^{(n)},P_{\theta'}^{(n)}) \;\leq\; \alpha\log|\Theta| \quad (\text{某个} 0<\alpha<1),$$

则任意估计器在该集合上的最坏情形 L_2 误差至少为（常数倍）

$$\inf_{\tilde f}\sup_{\theta\in\Theta}\mathbb{E}\|\tilde f - f_\theta\|_2^2 \;\gtrsim\; \text{packing 半径}.$$

由 (40) 与 $\log|\Theta| \geq c_1 m$，充分条件是

$$C_3\,\frac{nV^2}{\sigma^2 m} \;\leq\; \alpha c_1 m, \quad \text{即} \quad m^2 \;\gtrsim\; \frac{nV^2}{\sigma^2}.$$

因此可取

$$m \asymp \sqrt{\frac{nV^2}{\sigma^2\log n}}, \tag{41}$$

其中额外的 $\log n$ 用来吸收 Fano 分母中的常数并保证 $|\Theta|$ 足够大（这正是对数因子出现的来

源）。代入分离半径 (39)：

$$\text{packing 半径} \asymp \frac{V^2}{m} \asymp \frac{V^2}{\sqrt{nV^2/(\sigma^2 \log n)}} = \sigma V \sqrt{\frac{\log n}{n}}.$$

在 Wang–Lin 的 $\mathcal{G}$（未显式固定 V）写法下，可等价重参数化得到

$$\inf_{\tilde{f}} \sup_{f \in \mathcal{G}} \mathbb{E}\|\tilde{f} - f\|_2^2 \gtrsim \frac{1}{\sqrt{n \log n}},$$

即 (36) 的同阶形式（常数/对数写法差异来自 V 的规范化与 m 的取整）。 □

一句话总结 下界证明的本质是：在 $\mathcal{G}(V)$ 内用 m 个“近正交”的 ReLU ridge functions 拼出一个大小为 $\exp(cm)$ 的假设集合；当 m 取到 $\asymp \sqrt{n/\log n}$ 的量级时，这些假设在 L_2 下仍然可分离，但在 n 样本下彼此 KL 足够小、难以可靠区分，从而强迫任何估计器的最坏情形风险至少为 $\asymp (n \log n)^{-1/2}$（到同阶）。

16.5 与偏差–方差的关系：权衡变量被“替换”

传统叙事把复杂度写成“参数维数/自由度”，于是容易得到单峰（U 型）图景。而 (34) 告诉你：在过参数化网络里，

真正控制泛化的是 $\|f\|_S$ 这一类函数复杂度，以及正则/优化路径把解限制在哪个低复杂度区域。

当 m 足够大时，“表达能力”不再稀缺，近似误差持续下降；估计误差是否爆炸，取决于复杂度是否饱和（在这里表现为两段式的 $\min\{\cdot,\cdot\}$），这就是 double descent 的可证明机制。

本节小结

- 两层 ReLU + scaled variation 正则的非渐近界 (34) 在同一条不等式中给出 double descent 的两段机制。[21]
- 过参数化分支的风险饱和在 $\sqrt{(d \log n)/n}$（同阶），并且在 $\mathcal{G}$ 上接近 minimax 下界。[21]
- 这套分析与“经验过程/复杂度控制”完全同构：只是把复杂度从“参数维数”替换为“结构性范数”（S-norm / scaled variation），并用局部化得到两段式饱和。

致谢

（可选）

参考文献

[1] Blau P M, Duncan O D. The American Occupational Structure[M]. New York: Wiley, 1967.

[2] Boucheron S, Lugosi G, Massart P. Concentration Inequalities: A Nonasymptotic Theory of Independence[M]. Oxford: Oxford University Press, 2013.

[3] Boucheron S, Lugosi G, Massart P. Concentration Inequalities: A Nonasymptotic Theory of Independence[M/OL]. Oxford: Oxford University Press, 2013. DOI: 10.1093/acprof:oso/9780199535255.001.0001.

[4] Card D, Krueger A B. Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania[J]. The American Economic Review, 1994, 84(4): 772-793.

[5] Card D, Krueger A B. Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania: Reply[J/OL]. American Economic Review, 2000, 90(5): 1397-1420. DOI: 10.1257/aer.90.5.1397.

[6] Dudley R M. Uniform Central Limit Theorems[M]. 2nd ed. Cambridge: Cambridge University Press, 1999.

[7] Dudley R M. The Sizes of Compact Subsets of Hilbert Space and Continuity of Gaussian Processes[J/OL]. Journal of Functional Analysis, 1967, 1(3): 290-330. DOI: 10.1016/0022-1236(67)90017-1.

[8] Freedman D A. Statistical Models: Theory and Practice[M]. Cambridge: Cambridge University Press, 2009.

[9] Freedman D A, Pisani R, Purves R. Statistics[M]. 4th ed. New York: W. W. Norton & Company, 2007.

[10] Györfi L, Kohler M, Krzyżak A, et al. A Distribution-Free Theory of Nonparametric Regression[M/OL]. New York, NY: Springer, 2002. DOI: 10.1007/0-387-22442-4.

[11] Koltchinskii V. Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems: vol. 2033[M/OL]. Berlin: Springer, 2011. DOI: 10.1007/978-3-642-22147-7.

[12] Ledoux M, Talagrand M. Probability in Banach Spaces: Isoperimetry and Processes[M]. Berlin: Springer, 1991.

[13] Neumark D, Wascher W. Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania: Comment[J/OL]. American Economic Review, 2000, 90(5): 1362-1396. DOI: 10.1257/aer.90.5.1362.

[14] Pearl J. Causality: Models, Reasoning, and Inference[M]. 2nd ed. Cambridge: Cambridge University Press, 2009.

[15] Schmidt-Hieber J. Nonparametric Regression Using Deep Neural Networks with ReLU Activation Function[J]. The Annals of Statistics, 2020, 48(4): 1875-1897.

[16] Shalev-Shwartz S, Ben-David S. Understanding Machine Learning: From Theory to Algorithms[M]. Cambridge: Cambridge University Press, 2014.

[17] Stone C J. Optimal Global Rates of Convergence for Nonparametric Regression[J]. The Annals of Statistics, 1982, 10(4): 1040-1053.

[18] Talagrand M. The Generic Chaining: Upper and Lower Bounds of Stochastic Processes [M]. Berlin: Springer, 2005.

[19] Tsybakov A B. Introduction to Nonparametric Estimation[M]. New York: Springer, 2009.

[20] Van der Vaart A W, Wellner J A. Weak Convergence and Empirical Processes: With Applications to Statistics[M]. New York: Springer, 1996.

[21] Wang H, Lin W. Nonasymptotic Theory for Two-Layer Neural Networks Beyond the Bias–Variance Trade-Off[J/OL]. 2023. https://huiyuan-wang.github.io/files/tradeoff.pdf.

[22] Wright S. Correlation and Causation[J]. Journal of Agricultural Research, 1921, 20: 557-585.

[23] Wright S. The Method of Path Coefficients[J/OL]. The Annals of Mathematical Statistics, 1934, 5(3): 161-215. DOI: 10.1214/aoms/1177732676.

[24] Yarotsky D. Error Bounds for Approximations with Deep ReLU Networks[J]. Neural Networks, 2017, 94: 103-114.

[25] Yarotsky D. Optimal Approximation of Continuous Functions by Very Deep ReLU Networks[J]. Proceedings of Machine Learning Research, 2018, 75: 639-649.

[26] Yule G U. An Investigation into the Causes of Changes in Pauperism in England, Chiefly During the Last Two Intercensal Decades. (Part I.)[J/OL]. Journal of the Royal Statistical Society, 1899, 62(2): 249-286. DOI: 10.1111/j.2397-2335.1899.tb03709.x.

A 附录：常用恒等式与推导细节

（把较长推导放到附录，保持主文流畅。）