2018年春季统计210b（理论统计）-所有讲座

阿迪蒂亚·冈图博伊纳

2018年1月16日

# 目录

1. 第15讲
   1. 经验过程理论的某些方面。. . . . . . . . . . . . . . . . . . . . . . . . . . .5条
      1. 统一大数定律. . . . . . . . . . . . . . . . . . . . . . . . . . . . .5条
2. 第2讲8
   1. 一致中心极限定理. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .8条
   2. 浓度结果. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .11条
3. 第3课13
   1. hoefffding不等式和有界差分集中不等式的证明。. . .13条
      1. 论霍夫丁的不平等。. . . . . . . . . . . . . . . . . . . . . . . . . . …15个
      2. 鞅差分的Hoeffding不等式。. . . . . . . . . . . . . . . . . . …17年
      3. 有界差分集中不等式的证明。. . . . . . . . . . . . …18年
4. 第4课19
   1. 班尼特的不平等. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . …19年
   2. 回到supf的浓度∈F |…|. . . . . . . . . . . . . . . . . . . . . . . . . . . . …21年
5. 第5课23
   1. 期望上确界的界限. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . …24个
   2. Rademacher平均Rn（F（x1，…，xn））的简单界。. . . . . . . . . . . . . . …25个
6. 第6课28
   1. Sauer-Shelah-Vapnik-Chevronenkis引理的证明。. . . . . . . . . . . . . . . . . . …28个
   2. 包装和包装编号。. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . …30个
7. 第7课33
   1. 审查覆盖和包装编号。. . . . . . . . . . . . . . . . . . . . . . . . . . …36个
      1. 欧氏/参数覆盖数. . . . . . . . . . . . . . . . . . . . . . …36个
      2. 非参数函数类。. . . . . . . . . . . . . . . . . . . . . . . . . . . . .37点
      3. 一维光滑类。. . . . . . . . . . . . . . . . . . . . . . . . . . .37点
      4. 一维单调函数。. . . . . . . . . . . . . . . . . . . . . . . . …38个
      5. 多维平滑类。. . . . . . . . . . . . . . . . . . . . . . . . . . .38点
      6. 有界Lipschitz凸函数. . . . . . . . . . . . . . . . . . . . . . . . . . .39点
8. 第8课39
   1. 达德利度量熵界. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . …39个
   2. 达德利的界是无穷T. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .42点
   3. 达德利界在Rademacher平均数中的应用. . . . . . . . . . . . . . . . . . …43年
9. 第9课44
   1. 经验过程的期望上确界。. . . . . . . . . . . . . . . . .45分
   2. 有限维布尔函数类的应用。. . . . . . . . . . . . …47个
10. 第10课50
    1. 回顾上节课的主要经验过程. . . . . . . . . . . . . . . . …50个
    2. M估计问题的应用. . . . . . . . . . . . . . . . . . . . . . . . . . . . .51点
11. 讲座11 56
    1. VC子图维数. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . …56年
    2. 脂肪粉碎的维度. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .58点
12. 讲座12 61
    1. 支架控制。. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .61点
    2. M-估计。. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . …64个
    3. M估计的收敛速度. . . . . . . . . . . . . . . . . . . . . . . . . . . . …65年
13. 讲座13 66
    1. M估计收敛速度的严格推导。. . . . . . . . . . . . . . . . .66点
    2. 有界Lipschitz回归的应用. . . . . . . . . . . . . . . . . . . . . . . . …68年
    3. 回到速率定理. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .70点
14. 讲座14 71
    1. 主要速率定理概述。. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . …71年
    2. 高斯序列模型。. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . …72个
    3. 高斯序列模型中的凸惩罚估计. . . . . . . . . . . . . . …74年
       1. 不等式（118）在稀疏信号估计中的应用。. . . . . . . . . . . . …75个
15. 讲座15 76
    1. 不等式的证明（119）. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . …76年
    2. （119）到f（x）=kxk1的应用. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .77点
    3. 软阈值。. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . …78年
16. 第1681讲
    1. 硬阈值估计器。. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .81点
    2. 线性回归。. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . …83年
    3. θˆλBIC的预测风险. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . …83年
17. 第1786讲
    1. θˆλBIC的预测风险（续）. . . . . . . . . . . . . . . . . . . . . . . . . . . . .86点
    2. θˆλ套索的预测风险. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .88点
       1. 弱稀疏界。. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .89美元
       2. 强稀疏约束. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .90分
18. 第十八讲91
    1. 概述：精确稀疏线性回归. . . . . . . . . . . . . . . . . . . . . . . . . . .91美元
    2. 精确稀疏下套索的预测误差。. . . . . . . . . . . . . . . . . . . . …93年
    3. 检验RE和相容性条件的一个简单充分条件。. . . . . . …95年
    4. 受限等距属性. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . …96年
19. 第十九讲99
    1. 样本中值的极限分布. . . . . . . . . . . . . . . . . . . . . . . . . . . . .99美元
    2. 林德伯格-费勒中心极限定理。. . . . . . . . . . . . . . . . . . . . . . . . . . . . .101号
    3. 回到样本中值的极限分布。. . . . . . . . . . . . . . . . . . . . …101年
    4. 样本模式的极限分布。. . . . . . . . . . . . . . . . . . . . . . . . . . . . …103号
20. 第20课104
    1. 统一的经验过程。. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . …105年
    2. 抽象的结果。. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .106号
    3. 回到统一的经验过程。. . . . . . . . . . . . . . . . . . . . . . . . . . . . .108点
    4. 有可测量性的问题. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . …110个
21. 第21课110
    1. 极大不等式与随机等度连续性. . . . . . . . . . . . . . . . . . . . …112年

## 22讲座22115

22.1一致熵条件下的Donsker定理. . . . . . . . . . . . . . . . . . .117点

22.2 Donsker类的括号条件. . . . . . . . . . . . . . . . . . . . . . . . . . …118页

22.3样本中值收敛速度的应用。. . . . . . . . . . . . . . . . . . . . .119点

22.4 Argmax连续映射定理。. . . . . . . . . . . . . . . . . . . . . . . . . . .121号

## 23讲座23 122

23.1抽象M估计结果。. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .122 23.2适用于MLE。. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 一百二十六

## 24讲座24127

24.1二次均值的可微性。. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . …127年

24.2局部渐近正态性. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . …130个

## 25讲座25 134

25.1决策理论框架。. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . …134页

25.2如何评估决策规则. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .135号

25.3贝耶斯进近。. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .136点

25.4极大极小法。. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .137点

25.5最小最大下界. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . …138个

## 26讲座26 139

26.1稀疏正态均值估计。. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .140英镑

26.2功率约束下的正态平均估计（有限维Pinsker定理）。.143

## 27讲座27 145

27.1多假设检验问题. . . . . . . . . . . . . . . . . . . . . . . . . . . . …145页

27.2相互信息. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . ……148

27.3稀疏正态均值估计的应用。. . . . . . . . . . . . . . . . . . . . . . . . .148

27.4通过数据处理不等式的Fano引理. . . . . . . . . . . . . . . . . . . . . . .150美元

## 28讲座28 151

28.1通过测试得出的最小最大下限。. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . …151年

28.2稀疏正态均值估计。. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .153点

28.3利普希茨回归. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . …153年

28.4吉尔伯特-瓦尔沙莫夫引理. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . …155页

28.5避免F的显式构造的杨巴伦方法。. . . . . . . . . . . . . . . . .156英镑

# 1讲座1

课程的第一个主题是经验过程理论我将在高层次上概述我们计划在课程的经验性过程部分涵盖的内容。

## 1.1经验过程理论的若干方面

经验过程理论通常涉及两个基本问题。

### 1.1.1统一大数定律

第一个问题涉及统一的强大数定律。假设X1，X2，。。。是独立且相同分布的随机对象，取一个集合X中的值。让F表示X上的一类实值函数。可以对随机变量说些什么：

（一）

明确地，

1. 随机变量（5）是否集中在其期望值附近？
2. 就函数F的类和X1，X2，…，的公共分布P而言，我们能提供（5）的有限样本（即，每n个的界）的界吗，。。。？
3. 我们是否可以在F上提供条件，使（5）在概率上或几乎肯定地收敛到零（如果这是真的，我们说一致强数定律成立）？

经验过程理论为这些问题提供了答案为什么这些问题与理论统计有关下面是我们将要详细研究的两个例子。

例1.1（分类）。假设一对具有某种联合分布的随机对象X和Y，其中X取空间中的值X和Y只取两个值：1或+1。分类器是一个函数g:X→{-1，+1}。分类器的误差由L（g）：=P{g（X）6=Y}给出。

分类的目的是基于n i.i.d观测（X1，Y1），…，（X n，Yn）构造一个与（X，Y）具有相同分布的小误差分类器。

对于分类器g，其经验误差（即观察样本上的误差）由下式给出

.

一种自然的分类策略是选择一类分类器C，然后选择对观测样本经验误差最小的分类器C，即：。，

gˆn：=阿格明（g）。

g∈C分类器有多好gˆn，即它的误差有多小：

.

关于L（gˆn），有两个问题是相关的：

1. L（gˆn）是否与infg∈cl（g）相当？也就是说，gˆn的误差是否与C类中可达到的最佳误差相当？
2. L（gˆn）是否与Ln（gˆn）相当？也就是说，gˆn的误差与其“样本内”的经验误差是否具有可比性？

很容易将这两个问题与supg∈C | Ln（g）－L（g）|的大小联系起来。事实上，如果g∗：=argming∈C L（g），那么

L（gˆn）=L（g∗）+L（gˆn）－Ln（gˆn）＋Ln（gˆn）－L（g∗n）

≤L（g）\*+L（gˆn）－Ln（gˆn）＋Ln（g∗）－L（g∗）－L（g∗）－L（g∗）＋2向上| Ln（g）－L（g）|。

g∈C

阿尔索

L（gˆn）≤Ln（gˆn）+L（gˆn）–Ln（gˆn）≤Ln（gˆn）+sup | Ln（g）–L（g）|。

g∈C

因此，回答上述问题的关键是

sup | Ln（g）–L（g）|。

g∈C

现在很容易看出，当F被认为是所有函数I{g（x）6=y}的类时，上面的量是（5）的一个特例，因为g在C上变化，而且（5）中的x is需要被（Xi，Yi）替换。

有时，上面的两个不等式有时可能相当宽松稍后，我们将看到更尖锐的不等式，利用一种称为“局部化”的技术。

例1.2（M-估计量的一致性和收敛速度）。统计学中的许多问题与形式的估计有关

θˆn：=argmax）（2）

θ∈Θ

对于I.Id观测值x1，…，xn取空间x中的值。这里表示参数空间，并且对于每一个th\*，m thy表示X上的实值函数（称为损失或标准函数）。这样的估计称为“T估计”，因为它是通过使目标函数最大化而得到的。M-估计量的最标准例子是：

1. 极大似然估计：对应于X上的一类密度{p ^，th} }的m（x）：= log p（x）。
2. 位置估计器：
   1. 平均值：对应于mθ（x）：=（x-θ）2。
   2. 中值：对应于mθ（x）：=| x−θ|。
   3. 模式：可对应于mθ（x）：=I{| x−θ|≤1}。

估计量θˆn的目标量是

θ0：=argmaxEmθ（X1）。

θ∈Θ

研究M-估计的主要问题是估计θ0的θˆn的精度在渐近框架（n···）中，两个关键问题是：

1. θˆn是否与估计θ0一致，即d（θˆn，θ0）几乎肯定收敛到零还是以概率收敛到n→∞？这里d（·，·）是关于Θ的度量（例如，当Θ是某些k的Rk的子集时，通常的欧几里德度量）。
2. d（θˆn，θ0）的收敛速度是多少例如，它是Op（n-1/2）吗？或Op（n−1/3）？.

要回答这些问题，显然必须研究θ上某种一致意义上的闭性，从而研究{mθ，θ∈Θ}的适当子类F的（5）。

事实上，标准的论据包括第一次写作

（三）

哪里

M（θ）：=Emθ（X1）。

然后一个将（3）的右手边限定为

.

经验过程结果为上述概率提供了界限（需要对M和度量d之间的关系进行一些假设）。

注意，可以通过将左手边替换为

2sup | Mn（θ）–M（θ）|θ∈Θ

但这有时可能过于宽松。

控制策略（5）如下（我们主要关注F是一致有界函数类的情况）：

1. 关键的观察是，随机变量（5）“集中”在其平均值（或期望值）周围。
2. 由于浓度的关系，控制（5）的平均值就足够了平均值将被称为“对称化”的一种称为“rdAMACHER复杂性”的量所限制。
3. RADAMACHER复杂性涉及一个“亚高斯过程”的预期上确界。这是通过一种称为“链接”的技术进一步控制的在这个过程中，我们还将遇到一个称为“Vapnik-Chervonenkis维”的量。

这些主题的最佳参考是Boucheron等人的书[3]。我们应该采取的观点是非共扼性的观点，其中证明了每n的界是成立的。更经典的观点是渐近性的观点，其中声明成立为n→∞。在渐近观点中，假设（5）几乎肯定地收敛为n→∞，则F类是“Glivenko-Cantelli”。利用我们的非共扼边界，可以给F加上适当的条件，在这个条件下F变成

格里文科·坎泰利班。

# 2讲座2

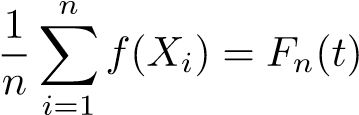
## 2.1一致中心极限定理

现在让我们来描述经验过程理论所解决的第二个基本问题。根据通常的中心极限定理（CLT），我们得到了

!

在平均值为零且方差V ar（f（X1））为n→∞的情况下，在分布上收敛到正态分布这句话对每一个f∈f都是正确的。在一个合理的意义上，这个收敛性是否一致地保持在f类中的f之上为了说明这一点，让我们看看下面的例子。

例2.1假设X1，…，Xn是来自[0,1]上均匀分布的i.i.d观测值同时假设F由所有的指示函数{I（－∞，t）：t∈R}组成在这种情况下，对于f=I–∞，t]，数量



其中Fn是观测值X1，…，Xn的经验分布函数定义

√

Un（t）：=n（Fn（t）－t）对于t∈[0,1]。

当t在[0,1]中变化时，Un（t）表示随机变量的集合随机过程{Un（t），t∈[0,1]}被称为“一致经验过程”很容易看出{Un（t），t∈[0,1]}的每一个实现，被看作是[0,1]上的一个函数，在n个数据点X1，…，Xn处具有跳跃不连续性同样，对于每n，Un（0）=Un（1）=0。

CLT指出，对于每一个t∈[0,1]，实随机变量序列{Un（t）}在分布上收敛到N（0，t−t2）作为N→∞。此外，多元CLT指出，对于每个固定的t1，…，tk，随机向量序列（Un（t1），…，Un（tk））在分布上收敛到多元正态分布，其均值和协方差均为零，由min（ti，tj）－titj给出。

在这一点上，让我们介绍一个名为Brownian Bridge的对象[0,1]上的布朗桥是一个随机过程{U（t），0≤t≤1}，其特征如下：

1. 每个实现都是[0,1]上的一个连续函数，U（0）和U（1）总是固定为等于0。
2. 对于[0,1]中的每个固定t1，…，tk，随机向量（U（t1），…，U（tk））具有多变量正态分布，其均值和协方差均为零，由min（ti，tj）－titj给出。

因此，我们看到过程{Un（t），0≤t≤1}的“有限维分布”收敛到{U（t），0≤t≤1}的“有限维分布”在这里很自然地会问，是否可以在这里声明任何超越有限维收敛的东西。整个过程{Un（t），0≤t≤1}是否收敛到{U（t），0≤t≤1}这是杜布首先提出的猜想，并得到了顿斯克的严格证明。

随机过程序列{Un（t），t∈[0,1]}在分布上收敛到{U（t），t∈[0,1]}的意义是什么？为了理解这一点，让我们首先回顾一下随机向量序列在分布上收敛的一般概念。我们说一个随机向量序列{Zn}取Rk中的值在分布上收敛到Z当且仅当

Eh（Z n）→Eh（Z）为n······

对于每个有界连续实值函数h:Rk→R。

我们可以尝试对此的直接推广，将Un（·）到U（·）的收敛定义为一个随机过程。这些过程的值不是在Rk中，而是在[0,1]上所有有界函数的空间中让我们用`∞（[0,1]）来表示这个空间。这个空间可以用上确界度量来度量：sup0≤t≤1 | g1（t）–g2（t）|。然后我们可以说Gn作为一个随机过程在分布上收敛到U

Eh（U n）→Eh（U）为n····（4）

对于每一个有界的连续实值函数h：`∞[0,1]→R，除了一个测度论问题外，这个定义几乎是有意义的证明了存在随机变量h（UN）不可测的有界连续实值函数h。因此，其中一个用其外部期望E∗h（Un）（稍后正式定义）代替第（4）项中的左侧。

从这个意义上讲，Donsker证明了Un在分布上收敛于Brownian桥。

现在让我们回到一般情况。这里我们考虑随机过程：

! 对于f∈f。

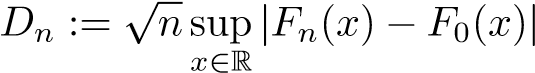
在一个简单的假设下，如supf∈F | F（x）|<∞，对于每个x∈x，函数F 7→Gn（F）属于空间`∞（F）如果随机过程G n（F），F∈F在`∞（F）中收敛到一个过程G（F），F∈F作为n→∞，则一致中心极限定理适用于F极限过程G（f），f∈f将具有这样的性质：对于每一个f1，…，fk∈f，随机向量（G（f1），…，G（fk））将具有与（Gn（f1），…，Gn（fk））相同的协方差的多元正态分布。

我们将刻画分布在`∞（F）中的收敛性，然后在F上得到一些保证一致CLT成立的充分条件。

下面是统一CLTs的一些统计应用。

例2.2（经典动机：拟合优度测试）。假设一个人从分布（cdf）F中观察到i.i.d观测值X1，…，Xn，并想用替代假设H1:F 6=F0来检验无效假设H0:F=F0这里F0是一个固定的分布函数。

Kolmogorov建议通过数量来检验这个假设



其中Fn是数据X1，…，Xn的经验cdf。其思想是在Dn较大时拒绝H0。为了计算该试验的p值，需要确定零分布（即H0下Dn的分布）。Dn的空分布的一个有趣的特性是，只要F0是连续的，空分布就不依赖于F0。（我将把这个事实留作练习；例如，它可以通过分位数变换来证明）。

因此，假设F0是（0,1）上的均匀分布，就可以计算Dn的空分布。在这种情况下，我们可以写

Dn=sup | Un（t）|

0≤t≤1

式中，Un（t）是来自实例180的统一经验过程。

{U n（t），t∈[0,1]}在分布上收敛到布朗桥{U（t），t∈[0,1]}作为n···········

每x>0。

后者的概率可以精确计算（例如，见达德利[5，命题12.3.4]）。因此，一致中心极限定理给出了一种通过Kolmogorov统计量计算拟合优度检验的渐近有效p值的方法。

同样的论点可以用于许多相关的拟合优度统计，例如

1. Cramer Von Mises统计：定义为

Z

Wn:=n（Fn（x）－F0（x））2 dF0（x）。

1. 安德森-达林统计：定义为

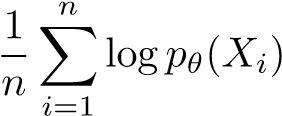
.

1. 斯米尔诺夫统计：定义为

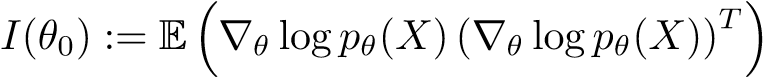
还有。

所有这些统计量的渐近零分布都可以从布朗桥计算出来，这将通过均匀CLT得到验证。

例2.3（MLE的渐近分布）假设X1，…，Xn是一个未知密度pθ0的i.i.d，它属于一个已知的类{pθ：θ∈Θ⊆Rk}。设THSN表示最大值为0的极大似然估计。



超过θ∈Θ。一个经典的结果是，在一些光滑性假设下，在分布上收敛到Nk（0，I-1（θ0）），其中I（θ0）表示k×k Fisher信息矩阵，定义为



式中，梯度∇θ在θ=θ0处计算，期望值取密度pθ0。

为了保持这个结果，需要对pθ，θ∈Θ施加什么样的光滑性假设由于结果涉及包含梯度的信息矩阵I（θ0），所以最小假设似乎是θ7→logpθ（x）需要关于θ可微。此外，由于在I（Th 0）的定义中存在期望值，如果关于关于p p 0的测度零点的导数不存在（关于模型p th\*（x）：＝EXP（ωx（？））），则应该是成立的。

然而，这个结果的经典证明假定这个映射允许两个（有时甚至三个）导数利用一致中心极限定理，我们稍后将提出一个证明，使用最小可微性假设称为二次均值可微性（DQM）。

例2.4（M-估计量的渐近分布结果）。一致中心极限定理也可用于推导其他M估计的极限分布。

例如，假设样本中值定义为：

θˆn：=argmax。

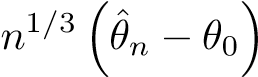
θ∈R

假设观测值的分布函数F在其中值θ0处可微，且具有正导数F（θ0），则可以证明

在分布中收敛到

对于定义为argmax且mθ（x）：=I{x-θ|≤1}且Θ=R的模，渐近

分配要复杂得多结果是



在分布中收敛到

阿格马克斯

h∈R

其中Z是从0开始的标准双边布朗运动，

a2：=p（θ0+1）+p（θ0-1）和。

这里p（·）表示观测值的密度，假设p是单峰对称的，模式θ0，即，对于x<θ0，p0（x）>0，对于x>θ0，p0（x）<0。这一结果是为了说明即使是看起来简单的M-估计的极限分布也是相当复杂的我们稍后将看到如何通过统一的CLTs来证明这些结果。

## 2.2浓缩结果

现在让我们从讨论统一大数定律开始。研究的主要对象是

（五）

其中X1，…，Xn是i.i.d随机对象，取X空间中的值，F是X上实值函数的集合。我们认为（5）集中在它的期望值上。当假定F中的所有函数都有一个正常数B的界时，这是相当容易证明的：

sup | f（x）|≤B，每f∈f（6）

x∈x

在上述假设下，我们将证明（5）的浓度结果这是有界差分集中不等式的结果。

定理2.5（有界差分集中不等式）。假设X1，…，Xn是取X集合中的值的独立随机变量。假设g:X×······································································

SUG（x1，…，xn）-g（x1，…，Xi，1，X0i，Xi+ 1，…，Xn）ωi（7）

x1，…，xn，x0i∈X

对于每i=1，…，n，那么对于每t≥0，我们有

（八）

和

（九）

备注2.1有界差分条件（211）相当于：

|g（x1，…，xn）–g（z1，…，zn）|≤ci

当（x1，…，xn）和（z1，…，zn）在第i个坐标中完全不同时。

它也相当于以下条件：

对于所有x1，…，xn，z1，…，zn∈X。

定理2.5可以看作是对Talagrand的下列定性陈述的量化（见Talagrand[22，第2页]）：依赖于许多自变量（但对其中任何一个变量的影响都不太大）影响的随机变量。数字ci控制第i个变量对函数g的影响。

我们将在下一节课中证明定理2.5。让我们在这里争论一下，它意味着

.

在条件（6）下的确，让

.

我们将在下面证明g满足有界差分假设（211），对于i=1，…，n，ci：=2B/n。要看到这一点，请注意

.

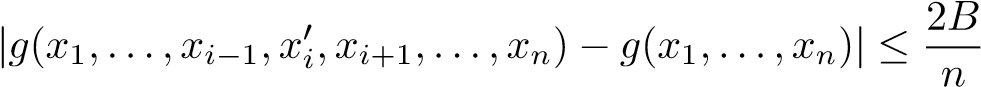
对于每一个f f，通过三角不等式和f（Xi）ωb和

.

在两边取f∈f的上确界，我们得到

.

交换XI和X0i的角色，我们可以推断出



所以（211）在ci=2B/n的情况下成立。定理2.5（特别是不等式（8））给出

每t≥0。

设置

,

我们推导出以下不等式：

（十）

每δ>0，概率至少为1−δ。

这个不等式意味着E（Z）通常是理解Z行为的支配项，这是因为通常E（Z）支配（10）右边的最后一个项。实际上，对于每一个f∈f，

（十一）

因为

,

有理由相信（11）的右手边通常为pvar（f（X1））/n阶。因此（除非var（f（X1））比B2小得多，否则（10）右手边的第一项通常支配第二项，因此为了控制随机变量Z，这就足够关注EZ的期望值了。

# 3讲座3

## 3.1 Hoefffding不等式及有界差分集中不等式的证明

本课程的目的之一是证明有界差分集中不等式。我们将证明另一个称为Hoeffding不等式的标准集中不等式，然后调整Hoeffding不等式的证明，得到有界差集中不等式。

定理3.1（霍夫丁不等式）假设ζ1，…，ζn是独立随机变量假设a1，…，an，b1，…，bn是常数，因此对于每个i=1，…，n，几乎肯定ai≤ζi≤bi。那么对于每个t≥0，我们有

（十二）

和

.

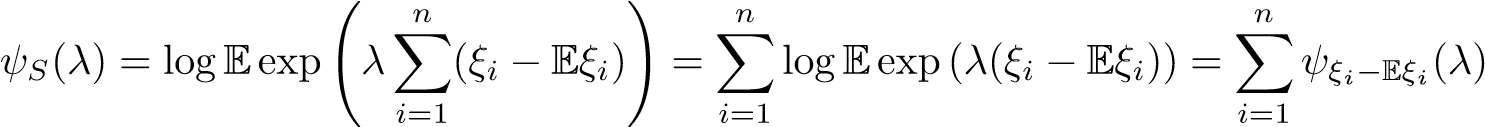
证明设）并写入（对于固定的λ≥0）

P{S≥t}≤P{eλS≥eλt}≤e－∏λtEeλS=exp（－∏λt+ΨS（λ））

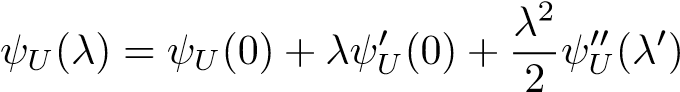
哪里

ΨS（λ）：=被记录人λS

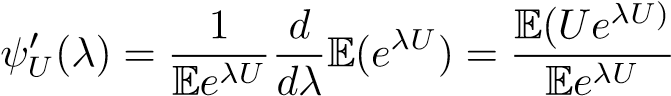
是S的对数矩母函数，现在由ζ1，…，ζn的独立性，



式中，Ψζi-Eζi（·）表示ζi-Eζi的对数矩母函数。固定1≤i≤n，设U：=ζi-Eζi。我们将在下面限制ΨU（λ）我们知道E U=0，而ai-Eζi≤U≤bi-Eζi几乎可以肯定。通过φU（λ）在0附近的二阶泰勒展开，我们可以写出



对于某些0≤λ0≤λ。现在注意，ΨU（0）=logE（1）=0阿尔索



以便

ΨU0（0）=EU=0。

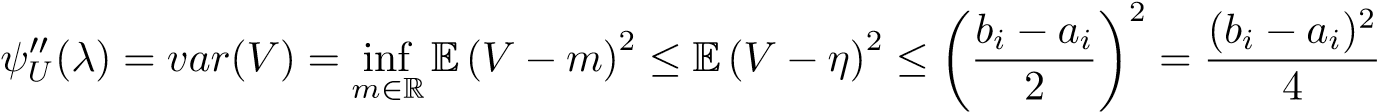
以及

.

现在考虑一个随机变量V，其密度相对于U的分布为eλU/（EeλU），即。，

.

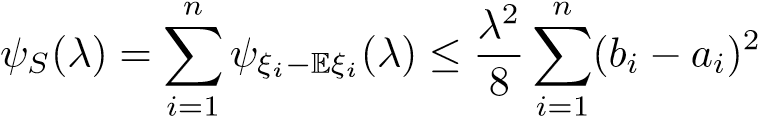
根据上面的计算，显然是0。还要注意，V在区间[ai-Eζi，bi-Eζi]上受支持（因为U在此区间上受支持，PV相对于PU绝对连续）因此，



其中，η是区间的中点[ai-Eζi，bi-Eζi]。因此，我们证明了当λ≥0时，ΨU00（λ）≤（bi-ai）2/4。这与ΨU（0）=0和ΨU0（0）=0一起给出

.

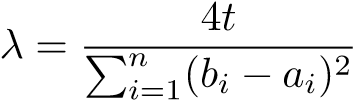
因此



因此

！

每λ≥0我们可以通过设置



证明（12）。要证明下尾不等式，只需将（12）应用到－。

上面给出的证明是以S的矩母函数为界的概率P{S≥t}，这种方法称为Cramer-Chernoff方法。

### 3.1.1关于Hoefffding不等式的注记

考虑Hoeffding不等式的以下特殊情况：假设X1，…，Xn是i.i.d，EXi=μ，var（Xi）=σ2，a≤Xi≤b几乎肯定（a和b是常数）。假设X∏n：=（X1+···+Xn）/n。

霍夫丁不等式给出

所有t≥0时。（13）

这是好的装订吗？这里的“好”是指，上面右边的概率是接近右边的边界，还是边界更宽松。为了回答这个问题，我们当然需要一种近似计算左侧概率的方法。一种自然的方法是调用中心极限定理（假设CLT是有效的）事实上，CLT声明

)作为n→∞

假设Xi的分布（特别是μ，σ2，a和b的数量）不依赖于n（注意Hoeffding不等式不需要这样的假设；特别是，（13）即使μ，σ2，a和b都依赖于n也是有效的）。因此我们可以期待



当n大且CLT成立时什么是P{N（0，σ2）≥t}？我们可以用CramerChernoff方法再次绑定：



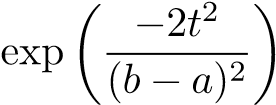
对于每个λ≥0，其中ΨN（0，σ2）是N（0，σ2）的对数矩母函数。通过一个简单的计算，可以看出ΨN（0，σ2）（λ）正好等于λ2σ2/2因此

每t≥0（14）

这个界限准确吗？从下面的不等式可以看出这是很好的（例如，见Feller[7，第7.1节]）：

.

因此exp是控制P{N（0，σ2）≥t}行为的正确指数项。现在让我们把霍夫丁和束缚（14）比较一下。霍夫丁出界了



正态近似

.

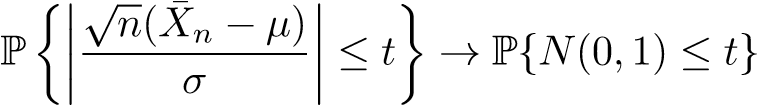
注意，因为a≤X1≤b几乎肯定，

.

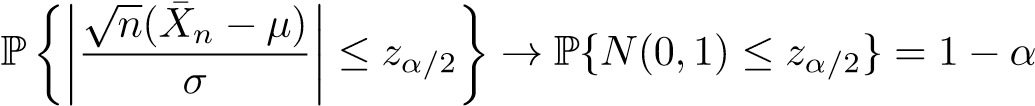
因此，在CLT成立的情况下，Hoeffding是一个更宽松的不等式，其中方差σ2被上界（b-a）2/4所代替。当X1在端点a和b附近放置较少的质量时，这种松散可以非常明显。这里是这种松散的潜在统计含义。

例3.2假设X1，…，Xn是i.i.d，EXi=μ，var（Xi）=σ2，a≤Xi≤b几乎肯定（a和b是常数）假设当μ未知时σ2、a和b已知，并且我们寻求μ的置信区间。解决这个问题有两种方法。

第一种方法使用CLT（正态近似）。事实上，CLT：



作为n··········因此



其中zα/2的定义使得上述最后一个等式成立这将导致以下μ的C.I：

（十五）

注意，这是μ的“渐近有效”100（1-－α）%置信区间另一方面，它的有限样本覆盖率可能不是100（1α）%。

构造μ的置信区间的第二种方法使用hoefffding不等式，该不等式指出

每t≥0。

因此，通过服用，

,

获得以下μ的置信区间：

（十六）

这个不等式保证了有限样本覆盖率100（1α）%。但这个间隔可能比（15）大得多。你喜欢哪两种间隔（15）和（16）？

### 3.1.2鞅差分的Hoeffding不等式

定理3.3（鞅差的Hoeffding不等式）假设F1，…，Fn是增大的σ-场，并且假设ζ1，…，ζn是随机变量，其中ζi是Fi可测的假设

E（ζi-Eζi-Fi-1）=0几乎肯定（17）

对于所有i=1，…，n。还假设，对于每个1≤i≤n，给定Fi-1的ζi的条件分布在一个区间上，该区间的长度从上到下由确定量Ri限定。那么

（十八）

和

每t≥0。

备注3.1。假设（17）意味着（Sj，Fj），j=1，…，n是鞅，其中Eζi）。因此序列{ζi-Eζi，i=1，…，n}是鞅差分序列。

证明出租）如前所述，对于t≥0和λ≥0，

P{S≥t}≤exp（－λt+ΨS（λ））

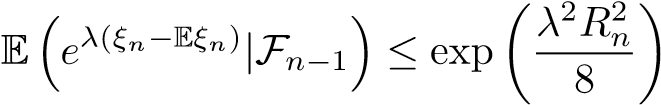
具有

.

现在请注意

.

现在因为Eζn=E（ζn | Fn−1），我们可以使用与证明独立情况下Hoeffding不等式完全相同的参数（通过对数矩生成函数的二阶Taylor展开）来推导



这给了

.

现在重复上面的论证（通过对Fn-2，然后对Fn-3等进行条件化）来推断

.

这给了

.

在λ上优化以推导（18）为了证明下尾不等式，用–ζi代替ζi。

### 3.1.3有界差集中不等式的证明

我们现在将证明有界差分集中不等式是定理3.3的一个简单结果回顾有界差集中不等式的陈述：

定理3.4（有界差分集中不等式）。假设X1，…，Xn是取X集合中的值的独立随机变量。假设g:X×······································································

N

|  |  |
| --- | --- |
| g（x1，…，xn）-g（Z1，…，Zn）} XCII{Xi 6＝Zi}  i=1  对于某些常数c1，…，cn。那么每t≥0，我们就有 | （十九） |

（二十）

和

.

定理3.4的证明。我们将鞅Hoeffding不等式应用于ζi：=E（g（X1，…，Xn）| X1，…，Xi）-E（g（X1，…，Xn）| X1，…，Xi-1）的i=1，…，n

Fi是X1，…，Xi为i=1，…，n生成的sigma场。显然，ζi是Fi可测量的

Eζi=0阿尔索

E（ζi | Fi-1）=E（ζi | X1，…，Xi-1）

=E（例如（X1，…，Xn）| X1，…，Xi）| X1，…，Xi-1）–E[g（X1，…，Xn）| X1，…，Xi-1]=E[g（X1，…，Xn）| X1，…，Xi-1]–E[g（X1，…，Xn）| X1，…，Xi-1]=0。

因此（ζi，Fi）是鞅差分序列。我们现在要讨论的是，给定Fi-1的ζi的条件分布，是在由ci从上面限定的长度区间上支持的。为此，我们需要研究给定X1，…，Xi-1的ζi的条件分布。让我们在x1，…，Xi，1上修正x1，…，Xi—1。那么，ζi是一个单独的Xi函数，我们需要看看当Xi=x变化时，ζi的值的范围。因此，我们需要审视这些价值观：

x 7°e [g（x1，…，Xn）x1＝x1，…，Xi＝1＝Xi＝1，Xi＝x]ω[g（x1，…，Xn）x1= x1，…，Xi＝1＝Xi＝1 ]

当x变化时，X1，…，Xi—1是固定的。现在，独立于X1，…，Xn，上面的右手边等于

Eg（x1，…，Xi，1，x，Xi+ 1，…，Xn）-常数

其中“常数”项仅取决于X1，…，Xi—1。所以我们可以把Ri

Ri:= Sup\* Eg（x1，…，Xi，1，x，Xi+ 1，…，Xn）Eg（x1，…，Xi，1，x0，Xi+1，…，Xn）

x，x0∈x

±g（x1，…，Xi，1，x，Xi+ 1，…，Xn）-G（x1，…，Xi，1，x0，Xi+1，…，Xn）x，x0∈x

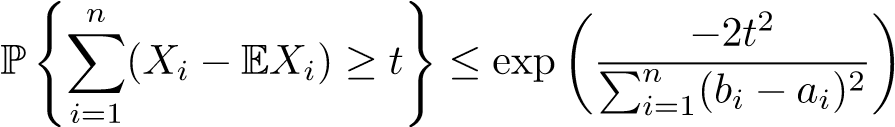
现在很明显，Ri≤ci是由有界差分假设（19）得出的因此我们可以应用定理

3.3用Ri=ci完成定理3.4的证明。

# 第四讲

## 4.1贝内特不等式

让我们回顾一下上节课的霍夫丁不平等。它表明



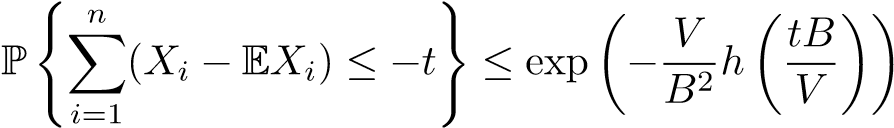
对于每个t≥0，其中X1，…，Xn是独立随机变量，ai≤Xi≤bi几乎肯定。我们注意到当）比4小得多，当CLT保持时，Hoefffding给出的尾部约束可以松脱Bennett不等式试图给出包含方差的尾界。

定理4.1（Bennett不等式）假设X1，…，Xn是具有有限方差的独立随机变量假设对于每个i=1，…，n，Xi≤B几乎肯定（这里B是确定的）。让。那么每t≥0，我们就有

（21个）

|  |  |  |
| --- | --- | --- |
| 哪里 |  |  |
| h（u）：=（1+u）对数（1+u）-u | 当u≥0时。 | （22个） |

备注4.1。如上所述，Bennett不等式只给出了上尾界为了得到下限，需要施加假设Xi<-B。在这种情况下，可以得到



备注4.2对于（22）中定义的函数h，很容易看出h（0）=0，h0（0）=0和h00（0）=1。因此，对于接近零的u，我们有h（u）≈u2/2因此，当tB/V很小时，Bennett不等式给出的界如下：

.

因此，在某些情况下，Bennett不等式给出了类高斯尾。

例如，假设EXi=0，var（Xi）=σ2，Xi≤1然后V=nσ2和Bennett不等式给出

.

当t比√nσ2小时，我们得到高斯型界。

定理4.1的证明在不丧失通用性的情况下，取B=1（使用变量X1/B，…，Xn/B而不是X1，…，Xn）。

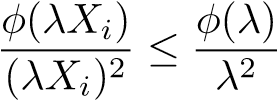
这一证明依赖于以下观察：设φ：R→R表示函数φ（u）：=eu-u-1图u 7→φ（u）/u2随R增大（取φ（0）/02=1/2）。我将把这个事实的核实作为家庭作业留下。

出租）那么对于每个λ≥0

（23个）

这里我们使用了X1，…，Xn的独立性因为Xi≤1，我们得到了λXi≤λ，因此

（利用φ（u）/u2增加的事实），我们推断



这意味着eλXi≤λXi+1+Xi2φ（λ）。

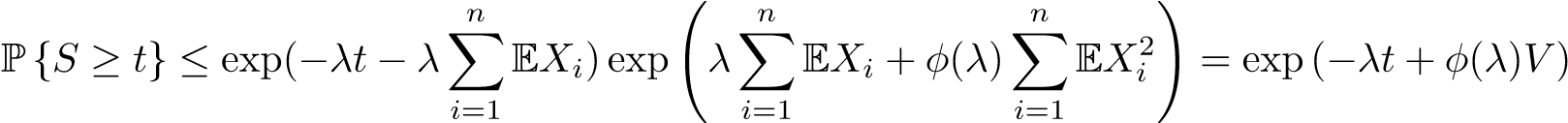
使用（23）右边的这个边界，我们得到

.

我们现在使用平凡不等式（1+x≤ex）



获得



每λ≥0我们现在通过取关于λ的导数并将其设置为零来优化上述界限，以获得：

.

对于这个值λ，可以直接推导（21）。

Bennett不等式中的界的形式可以通过使用以下不等式来简化（其证明留作练习）：

所有u≥0。

这导致了以下结果，即所谓的伯恩斯坦不等式。

定理4.2（伯恩斯坦不等式）。假设X1，…，Xn是具有有限方差的独立随机变量，并且假设对于每个i=1，…，n（B是确定性的），几乎肯定| Xi |≤B设V：=那么每t≥0，我们就有

和

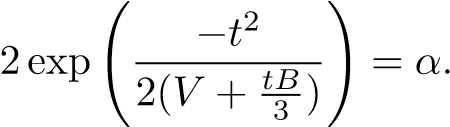
.

备注4.3。有一种伯恩斯坦不等式的版本，它用较弱的矩限制来代替有界性假设。见Boucheron等人。[3，定理2.10]。

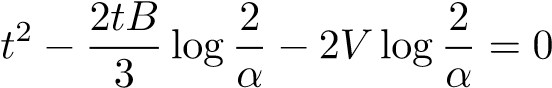
Bernstein不等式中的两个界可以组合成

.

我们现在可以尝试找到t的值，它使得上面右边的界正好等于α，也就是说，我们要解方程



这导致了二次方程

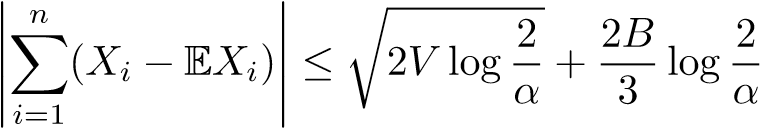


其非负解由



√ √ √

其中，在上一个不等式中，我们使用了a+b≤a+b的事实，因此伯恩斯坦不等式意味着



概率至少为1α如果X1，…，Xn是i.i.d，均值为零，方差为σ2，绝对值为B，则V=nσ2，这就得到了

（24个）

概率至少为1--α。请注意，如果X′n是正态的，那么X′n将被上面右手边的第一个项限定，概率至少为1α。因此，偏差限（24）与正常近似界一致，除了小阶项（IF阶为1／N）；

√

顺序为1/n）。

## 4.2回到supf∈F ||

现在让我们回到

（二十五）

让我们在这里介绍一些符号我们用Pn表示X1，…，Xn的经验测度。i.i.d随机观测的公共分布X1，…，Xn将用P表示

Pf：=Ef（X1）和。

因此，数量可以写成

sup | Pn f–-Pf |或sup |（Pn–-P）f |。

f∈f f∈f

我们通过有界差分不等式证明的集中不等式如下。假设F由B一致限定的函数组成，那么

（26个）

概率为1-α。

我们以前说过，当var（f（X1））比B小时，对于每一个f∈f，这个不等式并不尖锐。在这种情况下，使用Talagrand的集中不平等来证明经验过程的最高点，比（26）强得多，也更深入，更难证明我们将给出这个不等式的陈述，但不给出证明（关于证明，可以参考Boucheron等人[3，第12.4节]）在说明Talagrand不等式之前，让我们来看看一个统计应用程序，在这个应用程序中，有必要处理函数类F，其中与一致界相比，方差很小此应用程序涉及回归问题（它也同样适用于分类问题）。

例4.3（有界回归）。我们有两个随机对象X和Y，分别取X和Y空间中的值。假设Y是实线的有界子区间。问题是在X∈X的基础上预测Y∈Y。预测（或估计）是映射X到R的任意函数g。估计g的（检验）误差由

L（g）：=E（Y−g（X））2。

回归的目的是基于n i.i.d观测值（X1，Y1），…，（X n，Yn）构造一个具有与（X，Y）相同分布的小误差估计量。对于估计量g，其经验误差由

.

一种自然的策略是选择一类预测因子G，然后选择经验误差最小的预测因子G，即：。，

gˆn：=阿格明（g）。

g∈g

现在的关键问题是，就测试误差而言，预测因子gˆn有多好，即其误差有多小：

L（gˆn）：=Eh（Y−gˆn（X））2 | X1，Y1，…，Xn，Yni。

特别是，我们感兴趣的是L（gˆn）与infg∈gl（g）相比有多小。假设这个下确界是在某个g∗∈g.到界L（gˆn）－L（g∗）处得到的，那么很自然地写下：

L（gˆn）－L（g∗）=（Ln（gˆn）－Ln（g∗））+（L（gˆn）－Ln（gˆn））+（Ln（g∗）－L（g∗n）－Ln（gˆn））+（Ln（g∗）－L（g∗）。

我们现在可以使用经验过程符号设P表示（X，Y）的联合分布，Pn表示（X1，Y1），…，（Xn，Yn）的经验分布设F表示所有函数（x，y）7–（y–（g（x））2的类，因为g随g变化。

用这个符号，上面的不等式变成

P（fˆn–-f∗）≤（P–-Pn）（fˆn–-f∗）（27）

式中，fˆn（x，y）：=（y-gˆn（x））2和f∗（x，y）：=（y-g∗（x））2。为了继续前进，我们需要把右手边绑在上面。粗略的界限是

（P-P n）（fˆn-f∗）≤2 sup|Pnf-Pf|。（28）

f∈f

如果我们现在假设一类函数F是由B一致有界的，我们可以使用浓度不等式（26）这将对L（gˆn）－L（g∗n）有一定的约束，前提是可以控制期望（我们稍后将研究如何做到这一点）现在需要注意的是，这种方法永远不会给出更好的界限

L（gˆn）－L（g∗n）大于1/√n。这是因为（26）的右手边已经有一个n-1/2的项。但在回归中，至少对于小类G（如有限维函数类），我们预计测试误差的衰减要比n-1/2（如n-1速率）快得多这种方法不能证明如此快的速度。

为了证明更快的速度，我们需要使用一种叫做“本地化”的技术，而不是粗糙边界（28）。设δˆ表示（27）的左侧，目标是得到δˆ的边界。不平等（27）意味着

δˆ≤sup（PPn）（ff∗）。

f∈f:P（f−f∗）≤δˆ

因此我们真的需要了解如何

sup（PPn）（ff∗）。

f∈f:P（f−f∗）≤δˆ

这有点复杂，因为上确界中的函数类是随机的，依赖于δ。但让我们暂时忽略这一点，专注于获得

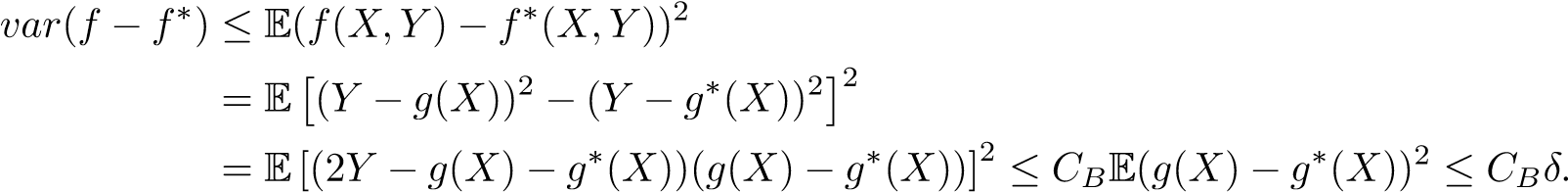
sup（P-Pn）（f-f∗）。（29）

f∈f:P（f−f∗）≤δ

对于一个确定但很小的δ。现在的关键是要认识到这里所涉及的函数有很小的方差（至少在g∗（x）=E（Y | x=x）的明确情况下）事实上，在特定的情况下，我们

.

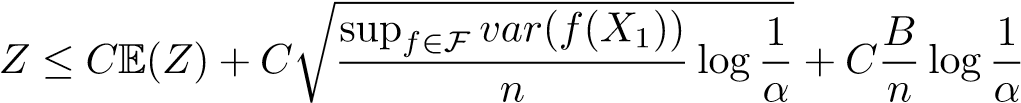
因此，当P（f−f∗）≤δ时，我们有



如果我们用浓度不等式（26）来控制（29），得到的界至少是pB/n，与δ无关这不会导致任何更快的速度。然而，塔拉格兰德的不平等将确保小方差给出一个更好的界在适当的假设下，将期望值的适当界结合起来，可以获得更快的回归率。

除了某些假设外，也可以对分类进行类似的分析。

现在让我们陈述一下经验过程的塔拉格兰德集中不等式。如前所述，假设F由常数B一致地限定。然后，让Z：=supf∈F | Pnf−Pf |，我们得到



概率至少为1α这里C是一个可以显式表示的普适常数。注意，前导项是EZ，第二项只涉及方差最终期限为1/n。

在学习了如何控制EZ之后，我们将回到回归和分类，为不同类别G的测试误差提供明确的误差界。我们将使用Talagrand不等式和局部化。

# 第五讲

本次讲座由迟进主讲。他对笔记做了一些修改（修改后的笔记在文件夹中）。

## 5.1期望上确界

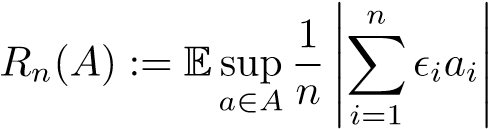
本课程的下一个主要主题涉及限定数量：

E sup | Pnf−Pf |（30）

f∈f

这里的两个主要思想是对称化和链式我们先讨论对称化。

利用F类的RADMACHER复杂性，从上对称化界（30），首先定义RADMACHER复杂性。Rademacher随机变量是一个随机变量，取两个值+1和-1，每个值的概率为1/2对于子集a⊆Rn，其Rademacher平均值定义为



其中，对i.i.d Rademacher随机变量取期望值请注意，首先测量值a1，…，和独立Rademacher噪声之间的“相关性”。这意味着当存在向量（A1，…，A）A适合RADMACHER噪声时，RN（a）是大的。这通常意味着集合A很大在这个意义上，Rn（A）测量集合A的大小。

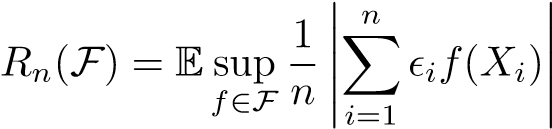
在经验过程设置中，我们有i.i.d随机观测X1，…，Xn取X中的值以及X上的一类实值函数F

F（X1，…，Xn）：={（F（X1），…，F（Xn））：F∈F}。

这是Rn的随机子集，其Rademacher平均数Rn（F（X1，…，Xn））是随机变量。该随机变量相对于x1，…，Xn的分布的期望称为F的RADMACHER复杂性：

Rn（F）：=ERn（F（X1，…，Xn））。

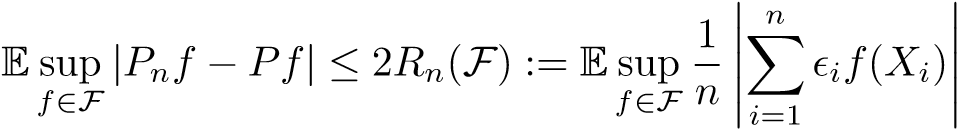
很容易看出



其中期望值是关于和X1，…，Xn，它们都是独立的（i是i.i.d Rademachers，Xi是i.i.d具有分布P）。

下一个结果表明，（30）中的期望是由RADMaCHER复杂性Rn（f）的两倍所限定的。

定理5.1（对称化）我们有



其中左边的期望值是关于X1，…，Xn是i.i.d，分布P，而右边的期望值是关于X和独立rademacher的

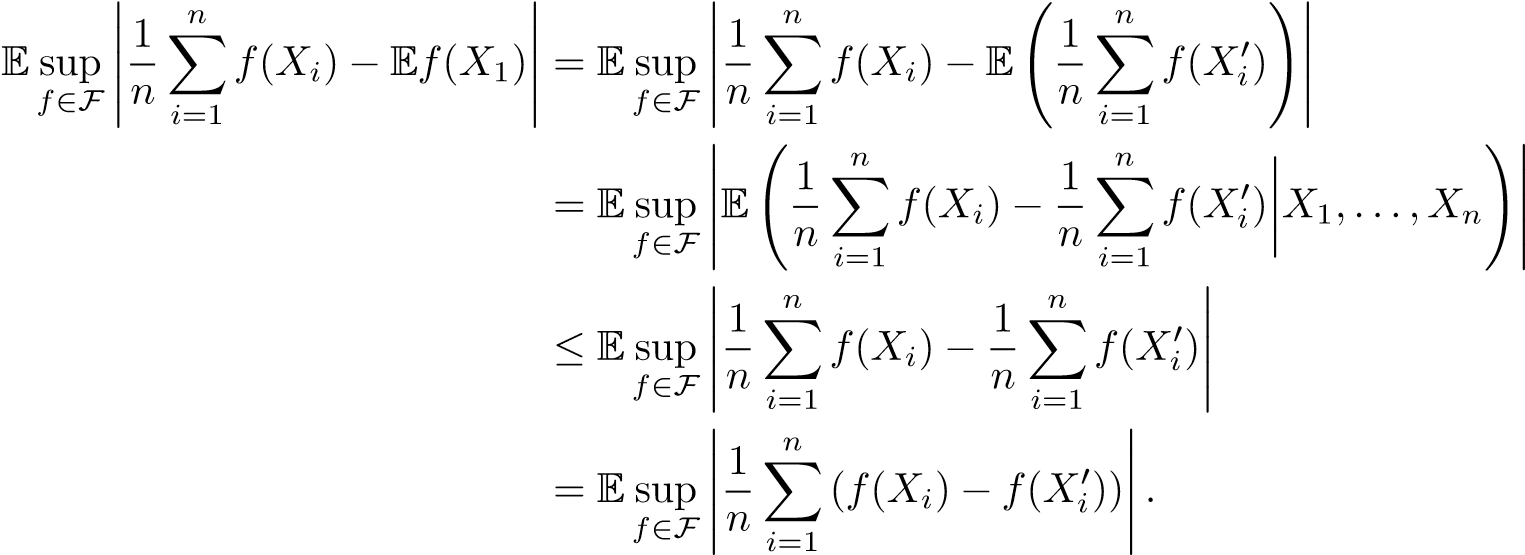
.

证明假设随机变量都是独立的

同样的分布，我们可以写

.

因此，我们



上面使用的方法基本上称为对称化我们现在介绍i.i.d Rademacher变量

. 因为是独立副本，很明显）的分布与）的分布相同因此，我们

.

定理5.1意味着我们可以通过从上面的Rn（F）的边界来控制（30）通常用于限制Rn（F）的策略如下一个首先确定点x1，…，xn∈X并限定集合的Rademacher平均值

F（x1，…，xn）：={（F（x1），…，F（xn））：F∈F}。（31）

如果该Rademacher平均值的上界不依赖于x1，…，xn，则它也自动成为Rn（F）的上界注意，为了使不动点x1，…，xn的Rn（F（x1，…，xn））有界，我们只需要处理使之更易于处理的简单分布。

限制Rn（F（x1，…，xn））的主要技术是链接。然而，在我们开始链接之前，我们将首先研究一个在布尔类F的某些情况下工作良好的更基本的边界。正如我们稍后将看到的，这个边界将不如链接所给出的边界精确。

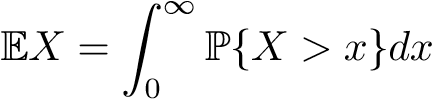
## 5.2 Rademacher平均Rn（F（x1，…，xn））的简单界

这些界限基于以下简单的结果。

提议5.2假设A是具有基数| A |的Rn的有限子集那么

（三十二）

命题5.2的证明。很容易看出，对于每个非负随机变量X

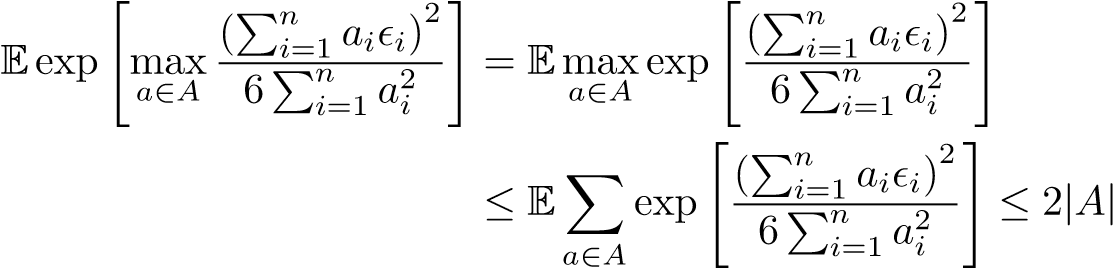


例如，可以通过交换积分和右边的概率来证明我们将在下面使用这个身份。

对于每一个a∈a，我们有

.

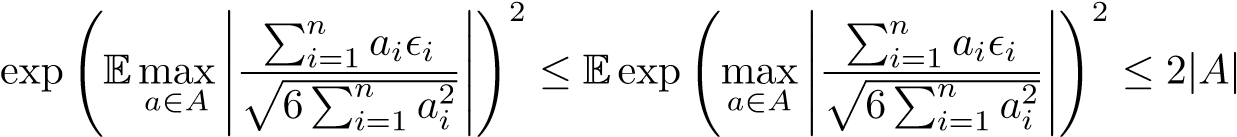
上面的概率界来自霍夫丁不等式。从上面看，我们有



其中A是A的基数，可以重写为

.

现在函数x 7→ex2是凸的（通过计算二阶导数可以很容易地检验），因此Jensen不等式给出了



以便

.

从这里开始，（32）中给出的不等式遵循平凡不等式：

.

现在让我们使用命题5.2来控制布尔函数类的RADAMACHER复杂性。如果F（x）对每个函数F和每个x∈x只取0和1的两个值，则F是一个布尔类。布尔类F是在分类问题中产生的（其中F可以被看作是由I{g（x）6=Y}形式的所有函数F组成）由于历史的原因，它们也很重要：经验过程理论起源于对supt（Fn（t）－F（t））的研究，它对应于取F：={I（－∞，t]：t∈R}。

现在让我们修复一个布尔类F和点x1，…，xn集合F（x1，…，xn）（定义见（31））显然是有限的，我们可以应用命题5.2来控制Rn（F（x1，…，xn））。这给了

.

因为F是布尔型的，所以我们可以把上面的每个F2（XI）绑定在1以上。

（三十三）

现在对于某些类F，对于n个点x1，…，xn∈X的集合，基数F（x1，…，xn）|可以由上面的多项式在n中有界，我们称之为具有多项式判别的类。对于这样的类，我们可以将Rn（F（x1，…，xn））绑定为每个x1，…，xn的p（logn）/n的常数倍数。因为Rn（f）被定义为Rn（f（x1，…，Xn））的期望，我们将得到，对于这样的布尔类，rdAMACHER复杂性由p（log n）/n的常数倍数限定。

定义5.3。布尔函数f的类被称为多项式判别，如果存在多项式r（·），使得对于n个N 1以上的每一组n个点x1，…，xn，x（f，x1，…，xn）的基数是最接近的π（n）。

如何检查给定的布尔类F是否具有多项式判别最流行的方法是通过类的Vapnik Chervonenkis维度（或者简单地说是VC维度）。

定义5.4（VC尺寸）。X上的一类布尔函数F的VC维定义为X满足的一个有限子集{x1，…，xd}的最大整数D。

F（x1，…，xD）={0,1}D。

如果对每个整数D满足上述条件，则将VC维数取为∞。

定义5.5（破碎）。一个X的有限子集{x1，…，xm}被布尔类F所粉碎，如果

F（x1，…，xm）={0,1}m。

根据惯例，我们将粉碎的定义扩展到空子集，并指出空集被每个非空类F粉碎。

从上述两个定义中可以清楚地看出，VC维度的另一个定义是：被F粉碎的有限子集X的最大基数。

VC维数与多项式判别之间的联系是通过以下著名的结果得到的，即Sauer-Shelah引理或VC引理。

引理5.6（Sauer Shelah Vapnik Chevronenkis）假设X上布尔函数类F的VC维数是D，那么对于每n≥1和x1，…，xn∈X，我们有

.

如果n<k，这里取0。如果n≥D，则

.

结合（33）与引理5.6，得到了具有有限VC维数布尔类的RADAMACHER复杂性和期望优越性的下界。

提议5.7。假设F是一个具有VC维D的布尔函数类，那么，对于n≥D，我们有

和

.

这里C是一个普遍的正常数。

备注5.1。结果表明，在上述命题给出的界中不需要对数项我们稍后将看到，通过链式给出的边界没有多余的对数因子。

我们将在下一小节提供引理5.6的证明在此之前，我们给出了两个有限维布尔类的例子。

例5.8设V是X上实函数的D维向量空间，设F：={I（F≥0）：F∈V}。F的VC维数最多为D。

证明对于任何D+1点{x1，…，xD+1}，考虑集合

T={（f（x1），…，f（xD+1）：f∈V}。

由于V是D维向量空间，T是RD＋1的线性子空间，维数D最多。因此，存在Y yRd＋1和y 6＝0，使得y与子空间T正交，即：

)=0表示所有f∈V。（34）

在不丧失一般性的情况下，我们可以假设有一个索引k使得yk>0。现在假设F粉碎{x1，…，xD+1}。那么f∈V满足

F i（Xi）＜0，I i＞Yi＞0；F（Xi）＞0，I i为i＜0；

然后，我们得到了Pi yif（Xi）＝Pi:Yi＜0 yIF（XI）+PI:Yi＞0 yIF（XI）＜0，这与（34）是矛盾的。因此F不能粉碎{x1，…，xD+1}，所以VC维数最多为D。

例5.9让Hk表示Rk中所有封闭半空间的指示符Hk的VC维数正好等于k+1。

证明留作作业题。

# 第六讲

这个讲座是由马克斯·拉比诺维奇主持的。他对笔记做了一些修改（修改后的笔记在文件夹中）。

## 6.1 Sauer-Shelah-Vapnik-Chevronenkis引理的证明

本节包含引理5.6的证明证据使用了一种叫做降档的方法。

用VC维数D修复布尔类F，还修复n≥1和x1，…，xn。为了简化符号，让我们用∏表示集合F（x1，…，xn）首先观察集合∏可以用一个布尔n×n矩阵表示，其中n：=|∏∏实际上，∏的每一个元素都是{0,1}n的元素；所以把这个元素写成矩阵中的一列；把对应于不同元素的所有列附加起来，形成一个n×n矩阵，所有列都是不同的。在引理5.6的证明中，我们将使用∏：作为{0,1}n的子集和布尔n×n矩阵的这两种表示。

对于{x1，…，xn}的一个子集S，让∏S表示通过仅取与S对应的∏的行而形成的∏S∏×N∏子矩阵。例如，如果S：={x1，x5，x8}，则∏S是∏的3×N子矩阵，仅由∏的第一行、第五行和第八行组成。

注意，{x1，…，xn}的一个子集S被F当且仅当{0,1}S |的每一个元素都显示为∏S的列时，才被F粉碎。因为F的VC维数是D，所以{x1，…，xn}的可以被F粉碎的子集的数目显然最多是

.

因此，我们必须证明∏的列数是atmost，{x1，…，xn}的子集数，这些子集被F粉碎。我们可以将其分离为以下结果，该结果仅适用于布尔矩阵。

结果6.1设∏表示具有布尔项的n×n矩阵，其所有列都是不同的假设{1，…，n}的一个子集S被∏粉碎，如果{0，1}S∏的每个元素都显示为∏S的列（空集总是被粉碎）设S（∏）表示{1，…，n}被∏粉碎的子集的数目。那么

N≤S（℃）。

结果证明6.1。这个证据遵循了一个叫做降档的想法选取矩阵的任意一行，比如第一行将该行中的每一个1更改为0，除非更改将创建一个已经存在于∏中的列。这将创建一个新的布尔矩阵，称之为∏0，其所有列都是不同的此操作称为降档。我声称

S（0）≤（35）

这一主张是证据的关键组成部分一旦确定了这一点，剩下的证据就立竿见影了。

为了证明（35），只要证明{1，…，n}的子集S不被∏击碎，它也不被∏0击碎就足够了这意味着，与意味着（35）的∏相比，被∏0粉碎的子集将更少。因此，让我们修正一个不被∏粉碎的子集S是∏行的子集。如果S不包含第一行，那么我们就无需做任何事情，因为除了第一行之外，所有行中的0和0都是相同的所以假设1∈S。实际上，假设，纯粹为了表示简单性，S={1,2,3,4}。

因为S不是被β打碎，所以存在一个Uü{0,1} 4的元素，它不存在于S中。如果U1＝1，那么很明显U也不存在于0S中，因为从0创建了0的降档操作不能创建新的。我们假设u1=0，然后写u=（1，v）u不在∏S中的事实意味着形式（0，v）的元素不以列形式存在于∏S中。这意味着（1，v）不存在于∏0S中。如果不存在，则∏0将包括形式（1，v，x）的列但是∏也必须包含列（1，v，x）。但是，如果∏确实有（1，v，x），那么它将通过降档操作转换为（0，v，x），因为（0，v，x）不总是以列的形式出现在∏中以防止这种移动这就完成了（35）的证明。

现在，考虑0再次，选择任意一行并在0上执行降档重复选择任意行并执行降档的步骤，直到我们得到一个不能通过进一步降档来改变的矩阵。将此矩阵称为∏\*重复应用（35）将意味着S（∏\*）≤S（∏）。现在将通过证明N≤S（&\*）来完成证明。要了解这一点，请考虑∏\*的第一列，让我们作为{1，…，n}中的索引，S\*的第一列中有1我声称S被∏\*粉碎了。要看到这一点，假设S不是空的，因为空的集合总是被粉碎的。为了表示简单，假设S={1,2}。我们需要证明{0,1}2中的所有元素都以列的形式出现在{S中，{0,1}2中只有四个元素：（1,1），（1,0），（0,1）和（0,0）。显然（1,1）出现在∏0的第一列中，也应该出现在某个地方，因为否则，应该可以通过降档来改变∏\*。（0,1）和（0,0）的情况类似。证据是完整的。

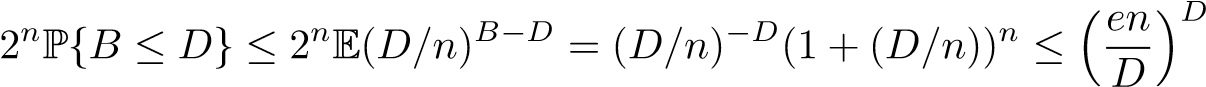
如前所述，结果6.1几乎给出证明引理5.6唯一剩下的就是争论

当n≥D时。

为了证明这一点，让B具有与n次投掷对应的二项分布，成功概率为1/2。那么上面的左手边等于

2nP{B≤D}

函数I{B≤D}从上到下由（D/n）B-D限定，该B-D给出

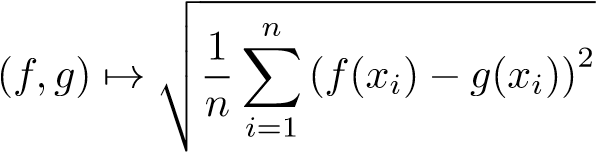


因为1+（D/n）≤eD/n。引理5.6的证明现在已经完成。

## 6.2覆盖和包装编号

如前所述，与命题5.2的简单界相比，链式给出了Rn（F（x1，…，xn））的更好界为了讨论链式，我们需要熟悉覆盖数和包装数的概念。

设T是一个具有伪度量d的集合。伪度量满足（a）d（x，x）=0（对于所有x∈T，（b）d（x，y）=d（y，x），和（c）d（x，z）≤d（x，y）+d（y，z）（对于所有x，y，z）。此外，如果它还满足d（x，y）>0（对于x 6=y），则d（·，·）成为度量。我们需要使用伪度量，因为函数：



通常不是F上的度量（因为它只依赖于F中x1，…，xn处的函数值）。但这是一个有效的伪度量。

定义6.2（包括编号）对于T和δ＞0的一个子集F，δ-覆盖数由NT（δ，f，d）表示，定义为覆盖F所需的闭合δ-球的最小数目。换句话说，NT（δ，f，d）是最小的n，其中每个tf f存在点为t1，…，tn\* t，其中，Mt1，i，n，（t，Ti）ωδ，对于每个t f，中心{Ti}的集合称为δ-网。n（δ，F，d）的对数称为F的δ-度量熵。

备注6.1。注意，中心t1，…，tN不被限制在F中。这与F的覆盖数的定义NT（δ，F，d）中下标T的存在有关。如果我们认为F本身是一个度量空间，而不仅仅是T的子集，那么覆盖数NF（δ，F，d）可能更大，因为中心ti将被强制证明NF（2δ，F，d）≤NT（δ，F，d）和2的附加因子通常没有多大影响。

如果T（δ，T，d）<∞对于每一个δ>0，我们说T是完全有界的。

覆盖数的概念与下面定义的包装数的概念密切相关。

定义6.3。对于δ＞0，f的δ-填充数被定义为最大n，其中每个i＝6 j（点t1，…，tn被称为δ-分离）存在点t1，…，tn f f具有d（Ti，Tj）＞δ。δ-填料数用M（δ，F，d）表示。十

结果表明，覆盖数和包装数是密切相关的。

引理6.4对于每δ>0，我们有

NF（δ，F，d）≤M（δ，F，d）≤NT（δ/2，F，d）≤NF（δ/2，F，d）。（36）

证明。对于（43）中的第一个不等式，让T1，…，Tm·f是m＝m（δ，f，d）f中δ-分离点的极大集。由于极大性，F的每个点都在点T1、…、TM的一个δ内。

这意味着t1，…，tM是F的δ-网，因此NF（δ，F，d）≤M，这证明了

（43）。

对于第二个不等式，再次让T1、…、Tm·f是m＝m（δ，f，d）f中δ-分离点的极大集。现在，如果试图用半径δ/2的闭合球来覆盖F，很明显每个球最多可以包含一个点t1，…，tM。这是因为任意两点ti和tj之间的距离严格大于δ，而δ/2球的直径最多为δ。因此，覆盖F所需的闭合δ/2球数至少为M，这证明了第二个不等式。

第三个不平等是微不足道的。

下面我们看到一些例子，其中覆盖/包装数的显式边界是可能的。

提议6.5。假设k·k表示Rk中的任何范数。例如，它可能是通常的欧几里德范数或n范数。让

BR：={x∈Rn:kxk≤R}。

那么，每一次，我们都

（三十七）

其中d表示与模k·k相对应的度量。

证明。设x1，…，xN表示BR中R分离的任意一组点，即，对于所有i 6=j，则为闭球



对于i=1，…，N是不相交的此外，所有这些球2）都包含在（以原点为中心的半径为2的球）中因此，

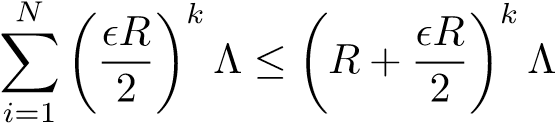
N个

十

音量（Vol(

i=1

其中Vol表示体积（Lebesgue measure）如果让∧表示单位球B1的体积，则上述不等式变成



立即证明（37）。

上面用来证明（37）的参数称为体积参数，因为它基于体积比较。

我们将考虑一些函数类的覆盖/包装数松散地说，函数类可以分为两类：参数类和非参数类参数类的-覆盖数对于某些整数k是有序的，而非参数类的-覆盖数对于某些k是exp（）的形式。这反映了非参数类将比参数类大得多的事实。

下面的命题给出了参数函数类的一个例子。注意，当下面的D和kΓkQ是常数时，由下面的结果给出的覆盖数的形式是。

提议6.6设ΘRk是一个欧氏直径为D的非空有界子集，设F：={Fθ：θ∈Θ}是由Θ索引的X上的一类函数，因此对于某些非负函数Γ：X→R，我们有

|fθ1（x）－fθ2（x）|≤Γ（x）kθ1－θ2k（38）

对于所有x∈x和θ1，θ2∈Θ。这里k·k表示Rk上通常的欧几里德范数。

确定X上的概率测度Q，并让d表示F上的伪度量

s公司

Z d（f，g）：=（f（x）–g（x））2 dQ（x）。

十

那么，每一次，

哪里

证明。条件（46）意味着对于每个θ1，θ2∈Θ，我们有

d（fθ1，fθ2）≤kθ1-θ2kkΓkQ。

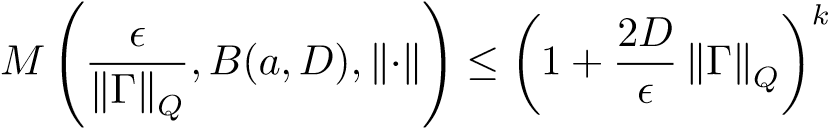
结果，度量d中的每一个分离子集F自动地是Θ的一个分离子集因此

.

为了限制欧几里德包装数，我们将使用这样一个假设：对于每一个a∈Θ，Θ的直径≤D，因此Θ包含在B（a，D）：={x∈Rk:kx−ak≤D}中。因此

.

为了将右手边限制在上面，我们使用命题7.6（注意，我们可以在上面a=0，因为相同半径的球将有相同的包装编号，而不管它们的中心是什么）。这给了



这就完成了7.8命题的证明。

非参数函数类最标准的例子是光滑类。它们的覆盖数是1/的指数我们将首先引入光滑类并在一维中描述它们的覆盖数，然后推广到多维。关于覆盖数结果的证明，见Dudley[6，第8章]。

固定α>0设β表示严格小于α的最大整数。例如，如果α=5，则β=4；如果α=5.2，则β=5。

Sα类定义为由满足以下所有性质的[0,1]上的函数f组成：

1. f在[0,1]上是连续的。
2. f是（0,1）上的可微β倍。
3. |f（k）（x）|≤1，对于所有k=0，…，β和x∈[0,1]，其中f（0）（x）：=f（x）。
4. |f（β）（x）－f（β）（y）|≤| x－y |α－β，所有x，y∈（0,1）。

设ρ表示由定义的Sα上的上确界度量

ρ（f，g）：=sup | f（x）–g（x）|（39）

x∈[0,1]

定理6.7。在α上存在正的常数和C2，所以对于所有的，我们都有。

因此，一维光滑类Sα的度量熵（覆盖数的对数）随时间而增长这里α表示光滑度（α越高，Sα中的函数越光滑）当α=1时，Sα类由[0,1]上的所有有界1-Lipschitz函数组成。

这个结果直接推广到多维。如前所述，α>0和β是严格小于α的最大整数。

对于由非负整数p1，…，pd组成的向量p=（p1，…，pd），设hpi：=p1+···+pd让



Sα，d类定义为由[0,1]d上满足以下条件的所有函数f组成：

1. f在[0,1]d上是连续的。
2. F的所有偏导数DP存在于（0，1）d，对于HPI不β。
3. |d p（x）|≤1，对于hpi≤β，x∈[0,1]d的所有p。
4. |Dpf（x）－Dpf（y）|≤| x－y |α－β，对于hpi＝β和x，y∈（0,1）d的所有p。

再次，我们考虑由ρ（f，g）：=supx∈[0,1]d | f（x）–g（x）|定义的上确界度量。

定理6.8。存在正常数C1和C2仅依赖于α和维数D，因此，对于所有，我们有

.

因此，光滑度α和维数d为的光滑函数类的度量熵为。这随着d的增加而增加，随着α的增加而下降。

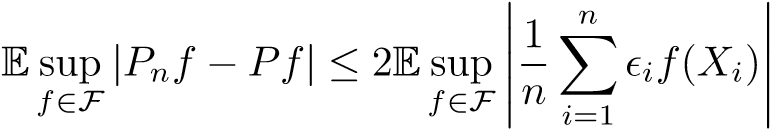
# 第七讲

类中的下一个主要主题是链接在我们开始链接之前，我们将回顾Chi和Max上周讨论的主题。

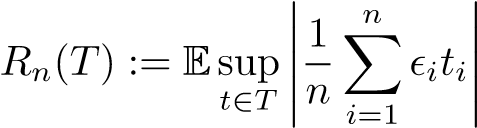
我们正在讨论控制问题：

.

上周引入的对称化技术允许我们将上述内容绑定为：



其中是独立的Rademacher随机变量，也独立于X1，…，Xn。上面右边的期望是关于两者的为了控制上面右边的期望值，通常在X1，…，Xn上有条件地工作。条件期望的形式是



对于Rn的一个子集TRn（T）更容易处理，因为它的期望是关于一个特别简单的分布（i.i.d Rademachers）。

在第五讲中，我们看到了关于Rn（T）的下列基本界。

提案7.1。假设T是具有基数的Rn的有限子集。那么

（四十）

对于一个普遍常数C。

（40）给出的界限有一些缺点。当T是无穷大时，它不会给出任何东西即使T是有限的，在某些特殊情况下它的界也是弱的。例如，当



对于R中的一些不动点x1··········································

.

结果表明，对于这个特定的T，对数项log（n+1）是多余的，Rn（T）的阶数是Cn-1/2。由链式导出的Rn（T）上的界限将为C/n−1/2形式。额外的对数因子是因为（40）的无效性。

尽管存在这些缺点，但界（40）对于推导链式界是重要且关键的在进一步讨论之前，我们将提供提案7.1的证据这一证据将与上周的证明方式略有不同我们将证明（40）的一个更强大的版本。

提案7.2。设T为有限集，设{Xt，T∈T}为随机过程。假设当t∈t且u≥0时，不等式

（41个）

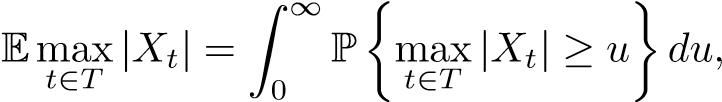
持有。这里∑是一个固定的正实数那么，对于一个普遍的正常数C，我们有

Emax | Xt |≤C∑plog（2 | T |）（42）T∈T

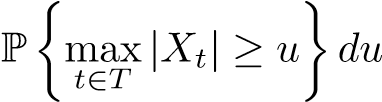
备注7.1请注意，命题7.2实际上是命题7.1的泛化。这是因为对于（t∈t），Hoeffding不等式保证（260）与

.

命题7.2适用于满足（260）的每一组随机变量Xt，此外，它也适用于σ≤∑的Xt∼N（0，σ2）。命题7.2的证明。因为



我们可以通过限制尾概率来控制Emaxt∈T | Xt |



每u≥0。为此，写下

.

这个界限对大u是好的，但对小u不是很好（例如，对u=0是很坏的）。因此，最好只在u≥u0时使用它，以便稍后指定某些u0。这给了

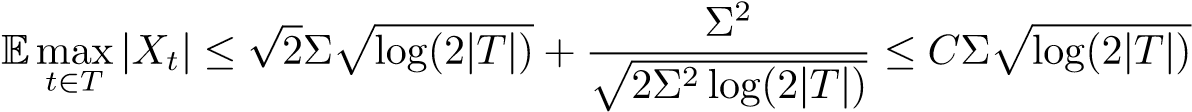
.

我们可以尝试在u0上最小化上述项。一个简单的策略是认识到这里的大项是2 | T |，所以可以通过设置

√

或u0=2∑plog（2 | T |）。

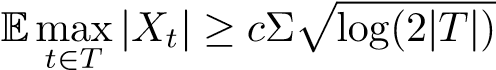
这给了



这证明了结果。

在命题7.2给出的界限是松散的情况下，构造例子并不难。例如，当尾界（260）对于许多t∈t是松散的时，它是松散的（这可能发生在例如对于一些远小于∑2的σ2，当Xt∼N（0，σ2）时）当许多Xt0彼此接近时，它也可能是松散的：例如，在极端情况下，当maxt∈T | Xt |≈Xt0对于单个t0∈T时，在（42）中的束缚是由log | T |因子松散的。

然而，存在（42）中的界紧的例子。最简单的例子如下设Xt，t∈t独立分布为N（0，∑2）。那么可以证明



对于正常数c。因此，在这种情况下，（42）紧到常数因子我将把上述不平等的证明留作家庭作业。这个例子意味着如果没有对过程{Xt，t∈t}的额外假设，命题7.2就无法改进在{Xt，T∈T}不同于（260）的假设下，链式给出了Emaxt∈T | Xt |的改进界假设（260）属于每一个XT的边际分布，但不说任何关于Xt如何接近另一个XS等。相反，对于链接，假设T上存在度量D，这样

.

在这个假设下，链式给出了Emaxt∈T | Xt |上的一个界，它涉及（T，d）的度量性质。在开始链接之前，让我们回顾一下度量空间的覆盖数和包装数的概念。

## 7.1覆层和包装编号的审查

设（T，d）是度量空间或伪度量空间覆盖和包装编号定义如下。

定义7.3（涵盖数字）对于T和δ＞0的一个子集F，δ-覆盖数由NT（δ，f，d）表示，定义为覆盖F所需的闭合δ-球的最小数目。换句话说，NT（δ，f，d）是最小的n，其中每个tf f存在点为t1，…，tn\* t，其中，Mt1，i，n，（t，Ti）ωδ，对于每个t f，中心{Ti}的集合称为δ-网。n（δ，F，d）的对数称为F的δ-度量熵。

备注7.2注意，中心t1，…，tN不被限制在F中。这与F的覆盖数的定义NT（δ，F，d）中下标T的存在有关。如果我们认为F本身是一个度量空间，而不仅仅是T的子集，那么覆盖数NF（δ，F，d）可能更大，因为中心ti将被强制证明NF（2δ，F，d）≤NT（δ，F，d）和2的附加因子通常没有多大影响。

如果T（δ，T，d）<∞对于每一个δ>0，我们说T是完全有界的。

覆盖数的概念与下面定义的包装数的概念密切相关。

定义7.4对于δ＞0，f的δ-填充数被定义为最大n，其中每个i＝6 j（点t1，…，tn被称为δ-分离）存在点t1，…，tn f f具有d（Ti，Tj）＞δ。δ-填料数用M（δ，F，d）表示。

由于下面的结果（在上节课中得到了证明），我们将把覆盖和包装号码视为大致相同。

引理7.5。对于每δ>0，我们有

NF（δ，F，d）≤M（δ，F，d）≤NT（δ/2，F，d）≤NF（δ/2，F，d）。（43）

下面我们看到一些例子，其中覆盖/包装数的显式边界是可能的。了解这些结果是有益的。

7.1.1欧氏/参数覆盖数

### 提案7.6对于R>0，让

.

表示以a.k·k点为中心的半径R的球。这里是常用的欧几里德范数。那么，每一次，我们都

（44个）

和

（45个）

其中d表示常用的欧几里德度量。

这些界限很容易证明（在上一节课中已经证明），而且证明是基于体积比较的（因此，这些界限通常被称为体积界限）。以下是（44）的直接推论

推论7.7假设S⊆Rk包含在某个半径为R的球中

.

d表示通常的欧氏度量。

上述结果表明，Rk中有界集的覆盖数随时间增长等价地，Rk中集合的度量熵增长为）如果k是常数，则度量熵以1/的对数增长对于在Rk中被有界集索引的函数类，只要索引和函数之间的映射是光滑的，通常也可以得出同样的结论。下面的命题提供了一种使这一点精确的方法。

提案7.8。设ΘRk是一个欧氏直径为D的非空有界子集，设F：={Fθ：θ∈Θ}是由Θ索引的X上的一类函数，因此对于某些非负函数Γ：X→R，我们有

|fθ1（x）－fθ2（x）|≤Γ（x）kθ1－θ2k（46）

对于所有x∈x和θ1，θ2∈Θ。这里k·k表示Rk上常用的欧几里德范数。确定X上的概率测度Q，并让d表示F上的伪度量

s公司

Z d（f，g）：=（f（x）–g（x））2 dQ（x）。

十

那么，每一次，

哪里

覆盖数增长为（或度量熵增长为对数（1/））的函数类通常被称为参数、欧氏或有限维。

#### 7.1.2非参数函数类

与参数类相比，非参数函数类要大得多，因为它们的度量熵在（1/）中以多项式形式增长下面提供了一些非参数函数类的标准示例。

#### 7.1.3一维平滑度等级

固定α>0设β表示严格小于α的最大整数。例如，如果α=5，则β=4；如果α=5.2，则β=5。

Sα类定义为由满足以下所有性质的[0,1]上的函数f组成：

1. f在[0,1]上是连续的。
2. f是（0,1）上的可微β倍。
3. |f（k）（x）|≤1，对于所有k=0，…，β和x∈[0,1]，其中f（0）（x）：=f（x）。
4. |f（β）（x）－f（β）（y）|≤| x－y |α－β，所有x，y∈（0,1）。

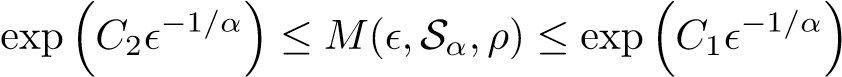
设ρ表示由定义的Sα上的上确界度量

ρ（f，g）：=sup | f（x）–g（x）|（47）

x∈[0,1]

结果表明，Sα的度量熵随时间增长它的证据可以在达德利[6，第8章]中找到。

定理7.9。在α上存在正的常数和C2，所以对于所有的，我们都有。



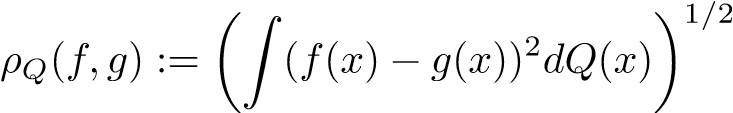
因此，一维光滑类Sα的度量熵（覆盖数的对数）随时间而增长这里α表示光滑度（α越高，Sα中的函数越光滑）当α=1时，Sα类由[0,1]上的所有有界1-Lipschitz函数组成。

#### 7.1.4一维单调函数

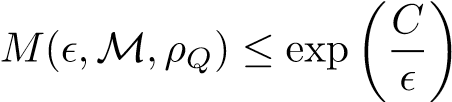
设M表示[0,1]上所有函数f的类

1. f在[0,1]上不递减
2. |f（x）|≤1，对于所有x∈[0,1]。

对于[0,1]上的概率测度Q，让ρQ表示M上的度量，由



那么可以证明



对于[0,1]上的每个概率测度Q存在概率测度Q，下界为

)同时保留包装号。将此结果与定理7.9进行比较，可以清楚地看出M的覆盖数与光滑度等级S1相当，即Sα与α=1因此，有界单调函数具有与有界Lipschitz函数相同的度量熵，即使单调函数不需要连续。

#### 7.1.5多维平滑度等级

与一维情况一样，设α>0，β是严格小于α的最大整数。

对于由非负整数p1，…，pd组成的向量p=（p1，…，pd），设hpi：=p1+···+pd让



Sα，d类定义为由[0,1]d上满足以下条件的所有函数f组成：

1. f在[0,1]d上是连续的。
2. F的所有偏导数DP存在于（0，1）d，对于HPI不β。
3. |d p（x）|≤1，对于hpi≤β，x∈[0,1]d的所有p。
4. |Dpf（x）－Dpf（y）|≤| x－y |α－β，对于hpi＝β和x，y∈（0,1）d的所有p。

再次，我们考虑由ρ（f，g）：=supx∈[0,1]d | f（x）–g（x）|定义的上确界度量。

定理7.10存在正常数C1和C2仅依赖于α和维数D，因此，对于所有，我们有

.

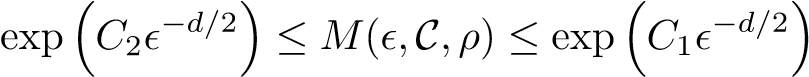
因此，光滑度α和维数d为的光滑函数类的度量熵为。这随着d的增加而增加，随着α的增加而下降。

7.1.6有界Lipschitz凸函数

设C表示0,1]d上所有函数f的类，使得

1. f在[0,1]d上是凸的。
2. |f（x）|≤1，对于所有x∈[0,1]d
3. |f（x）－f（y）|≤kx－yk。

然后可以证明（ρ是[0,1]d上的上确界度量）：



其中C1和C2仅依赖于d。与定理7.10相比，很明显，就度量熵而言，C可与光滑度等级Sd，2相比较这很有趣，因为凸函数在通常意义上不一定是两次可微的然而，它们在度量熵方面具有二阶光滑性的正则性。

# 第八讲

## 8.1达德利度量熵界

今天的主要目标是陈述并证明亚高斯过程上确界的达德利熵界证据包括一个叫做链锁的想法。在开始链接之前，让我们回顾一下上一个类的以下基本结果。

提议8.1。设T为有限集，设{Xt，T∈T}为随机过程。假设当t∈t且u≥0时，不等式

（48个）

持有。这里∑是一个固定的正实数那么，对于一个普遍的正常数C，我们有

Emax | Xt |≤C∑plog（2 | T |）（49）T∈T

正如我们在上节课中所说，在某些情况下，界限（49）可以是紧的（直到乘法常数）例如，这是当Xt，t∈t是i.i.d N（0，∑2）时的情况。因为这个例子，命题8.1如果不在进程{Xt，t∈t}上附加条件，就不能得到改进在（49）很弱的地方构造示例也很容易例如，如果对于某些X0∼N（0，∑2）和Zt，t∈t∼i.i.d N（0,1）和η来说，Xt=X0+ηZt是非常小的，那么很明显，maxt∈t | Xt |≈X0，这样（49）将被对数因子（2 | t |）松脱为了改进（49），我们需要假设XT0之间的距离。达德利的熵界使这一假设变得明确，并为Emaxt∈T | Xt |提供了改进的上界。

我们将首先在指标集T是有限的情况下说明Dudley的界，然后将其改进为T是无限的情况。

定理8.2（有限T的达德利度量熵界）假设（T，d）是一个有限的度量空间，{Xt，T∈T}是一个随机过程，对于每个s，T∈T和u≥0，

（五十）

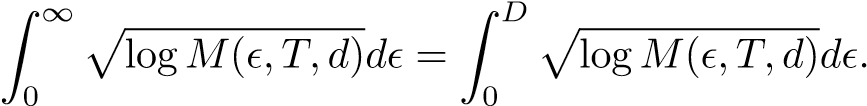
然后，对于一个普遍正常数C，对于每个t0∈T，以下不等式成立：

（51个）

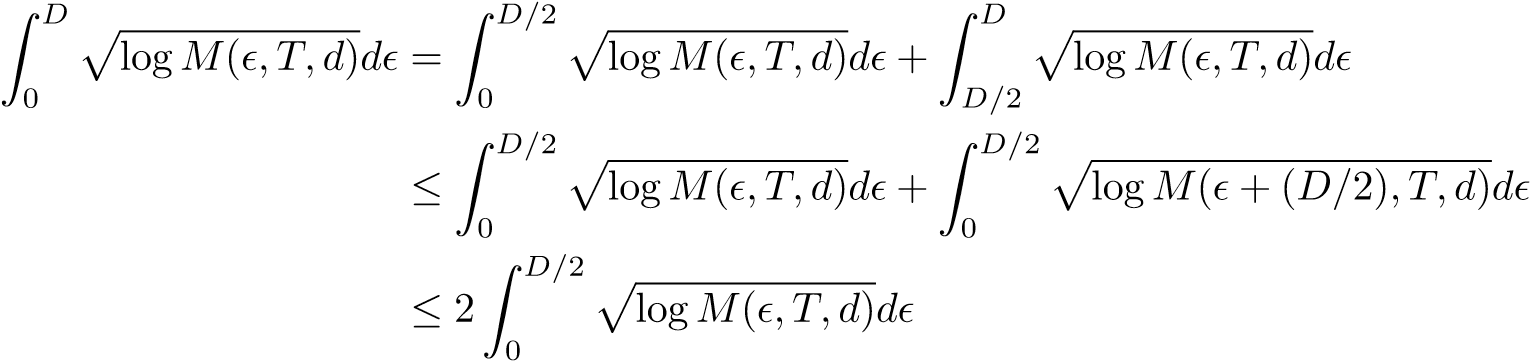
下面的评论提到了书写不平等（51）的一些替代形式，并描述了一些含义。

1. 设D表示度量空间T的直径（即D=maxs，T∈td（s，T））。那么包装号

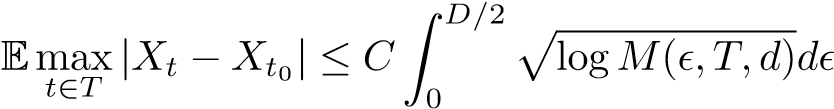
)显然等于1（T中不可能有两点的距离严格大于D）。因此



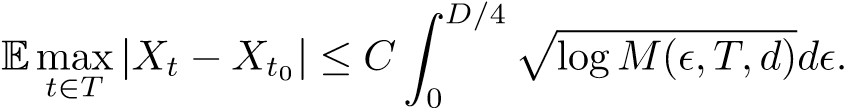
而且



因为）对每一个人因此我们可以说达德利的边界是



式中，上述C等于（51）中常数C的两倍。同样，通过将上述积分分成两部分（大于0到D/4和大于D/4到D/2），我们也可以将达德利的界表示为



上面的常数C是（51）中常数的4倍。

1. （51）中的左手边从下（通过三角形不等式）被Emaxt∈T | Xt |-E | Xt0 |限定因此，（51）意味着

对于每个t0∈T。

1. 如果Xt，t∈t具有平均零且共同为高斯分布，则XtXs是每s，t∈t的平均零正态随机变量，因此（50）与

d（s，t）：=pE（Xs-Xt）2。

1. 从下面的例子可以清楚地看出定理8.2优于命题8.1。假设Xt，t∈t由

Xt=X0+ηZt

对于一些正的但非常小的η和X0∼N（0，∑2和Zt，t∈t∼i.i.d N（0,1）我们之前已经看到，在这个例子中，Emaxt∈T | Xt |的行为应该类似于C∑，但是命题8.1将给出一个额外的对数因子（2 | T |）。另一方面，因为

√

d（s，t）=pE（Xt−Xs）2≤η2，

包装数量）将等于1，除了非常小的值（比如）因此

达德利的界将给出∑+这比命题给出的界要小得多

8.1（因为0很小）。

我们现在给出定理8.2的证明证据将基于一个叫做链式的想法具体地说，我们将把maxt∈T（XtXt0）分裂成链，并在每个链的链中使用命题8.1给出的界。

定理8.2的证明。回想D是T的直径，对于n大于1，使得Tn是T的最大D2～n分集子集，即Min、T\*Tn:S6= T D（S，T）＞D2→N，Tn具有最大的基数受分离约束。T n的基数由包装号M（D2-n，T，d）给出因为最大化，

（五十二）

因为T是有限的，并且对于所有s 6=T，d（s，T）>0，所以当n较大时，集合Tn将等于T。让

N:=min{N≥1:Tn=T}。

对于每个n≥1，让πn:T 7→Tn表示将每个点T∈T映射到Tn中最接近T的点的函数（如果Tn中有多个最接近T的点，则任意选择一个）。换句话说，πn（t）的选择是这样的

.

因此，从（52）开始

|  |  |
| --- | --- |
| d（t，πn（t））对于所有t∈t和n≥1，均≤D2-n。  注意πN（t）=t。最后让T0：={T0}和π0（t）=T0用于所有t∈t。  我们现在注意到 | （53个） |

每t∈t（54）

顺序

t0→π1（t）→π2（t）······································

可以看作是从t0到t的链。这就是给参数命名链的原因到（54），我们得到

北

max | X t−Xt0 |≤max X | Xπn（t）–Xπn−1（t）|≤X max | Xπn（t）–Xπn−1（t）| t yn t yn t yn t n=1 n=1

以便

N个

Emax | X t−Xt0 |≤XEmax | Xπn（t）–Xπn−1（t）|。（55）t∈t∈t

n=1个

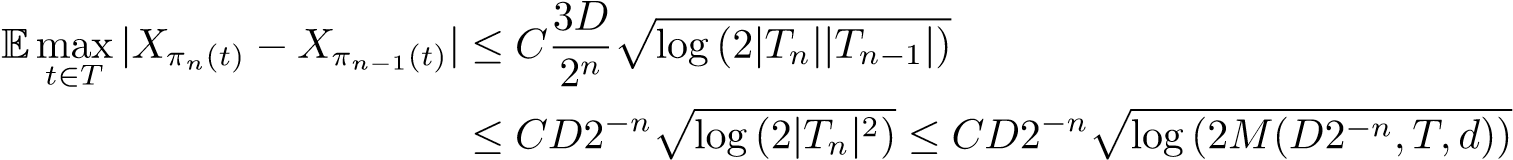
现在，为了使Emaxt∈T | Xπn（T）–Xπn | 1（T）|对于每个1≤n≤n，我们将使用

提议8.1。为此，请首先注意，到（50），我们已经

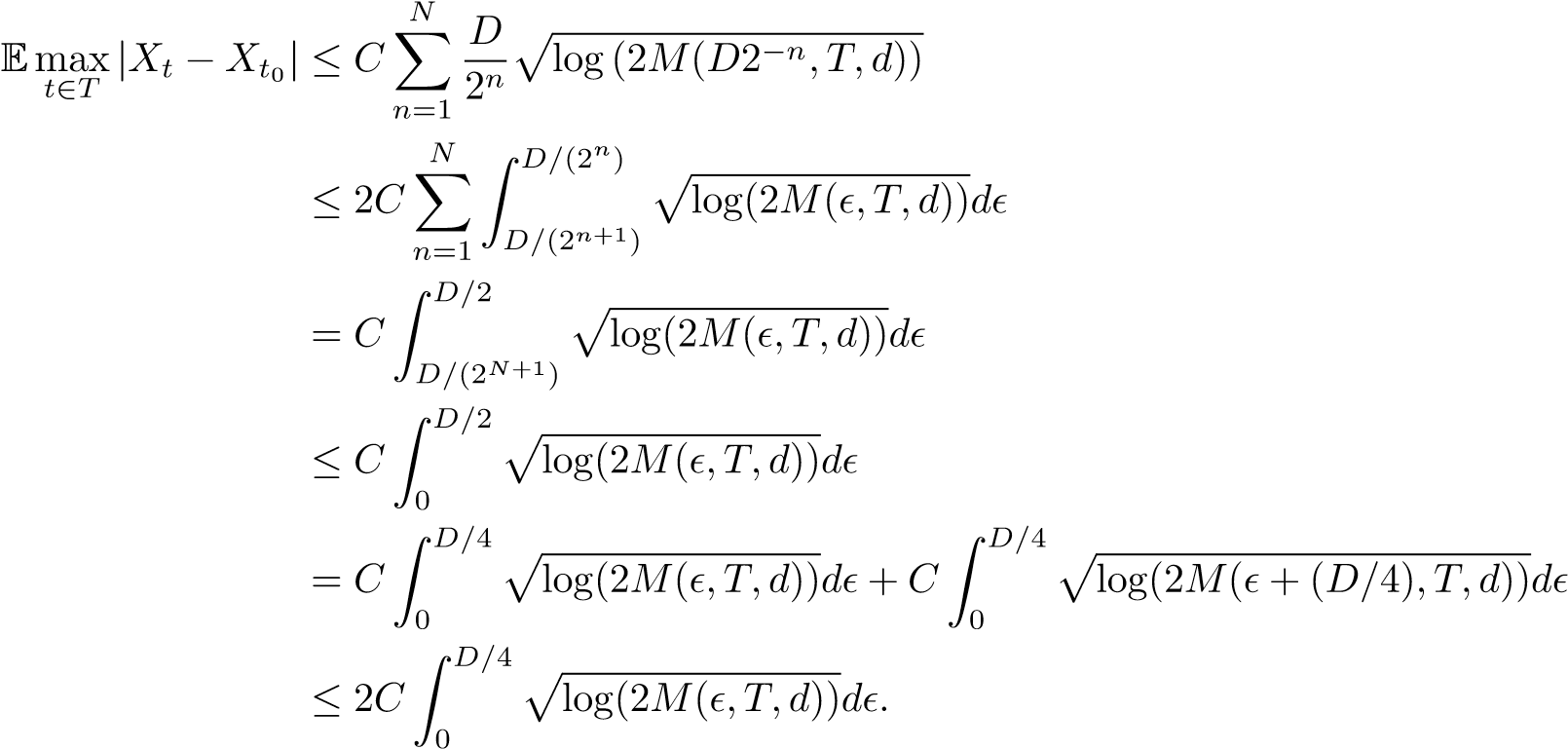
.

现在d（πn（t），πn1（t））≤d（πn（t），t）+d（πn1（t），t）≤D2n+D2（n1）=3D2n。

因此，命题8.1可以用∑：=3D2-n来应用，这样我们就得到了（注意C的值可能会随着事件的发生而变化）



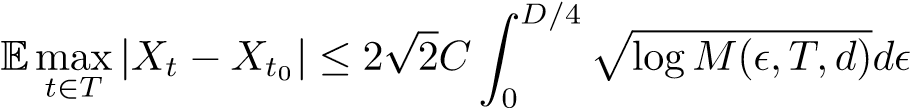
把上面的边界插入（55），我们推断



注意，对于4，包装编号为2

.

我们已经证明了



这证明了（51）。

## 8.2达德利无穷T的界

下一步我们将证明无穷T情形下的达德利界，这需要一个称为可分性的技术假设，这个假设在我们的应用中总是满足的。

定义8.3（可分离随机过程）。设（T，d）为度量空间。如果T存在一个空集n和一个可数子集T，则对于由T索引的随机过程{xt，t}t}是可分的，因此在所有ω/ωn和tt t中，存在一个序列{tn}，在t中具有Limn＝D（tn，t）＝0，并且Limn×ωxtn（ω）＝xt（ω）。

注意，可分性的定义要求T是T的稠密子集，这意味着度量空间（T，d）是可分的（如果度量空间具有可数稠密子集，则称其为可分的）。

如果（T，d）是一个可分度量空间，如果Xt，T∈T有连续的样本路径（几乎肯定），那么Xt，T∈T是可分的Xt、T t具有连续采样路径（几乎肯定）的陈述意味着存在一个零集N，使得对于所有ω/ωn，函数T 7～xt（ω）在t上连续。

如果{Xt，t∈t}是一个可分离的随机过程，那么

几乎可以肯定（56）

对于每个t0∏T，这里T∏是T的一个可数子集，它出现在Xt，T∏T的可分性定义中。

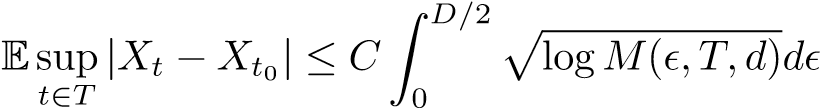
特别地，声明（56）暗示supt∈T | Xt−Xt0 |是可测量的（注意，不可数的上确界一般不保证是可测量的；但这不是可分离过程的问题）。

我们现在来说明可分离过程的达德利定理。这个定理对T没有任何基数限制（它对有限T和无限T都适用）。

定理8.4设（T，d）为可分度量空间，设{Xt，T∈T}为可分随机过程。假设对于每一个s，t∈t和u≥0，我们有

.

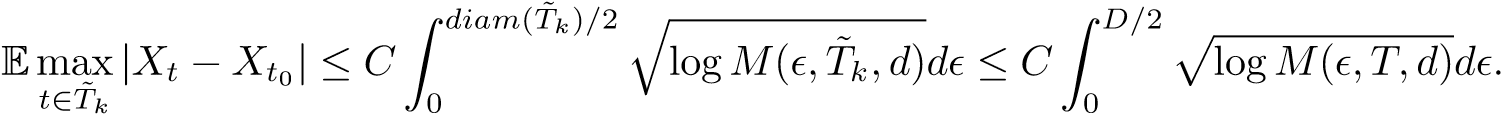
那么对于每个t0∈T，我们有



其中D是度量空间的直径（T，D）。

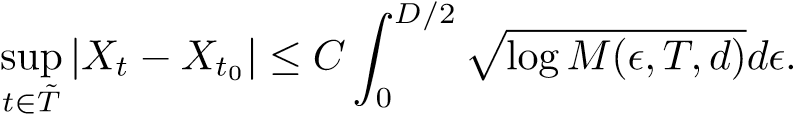
定理8.4的证明设T是T的可数子集，使（56）成立我们可以假设T●包含t0（否则只需将t0添加到T●中）。对于每个k≥1，设T●k为取T●的前k个元素（在T●条目的任意计数中）得到的有限集。我们可以保证每k≥1，T●k含有t0。

将达德利定理（定理8.2）的有限指数集形式应用于{Xt，t∈t∮k}，我们得到



注意，右侧不依赖于k。让k·········，在左侧，我们使用

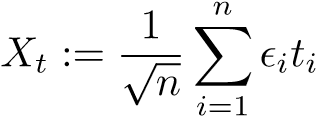
单调收敛定理



证据现在由（56）完成。

## 8.3达德利界在Rademacher平均数中的应用

假设T⊆Rn，并考虑随机过程Xt，T∈T



i.i.d Rademacher随机变量在哪里。

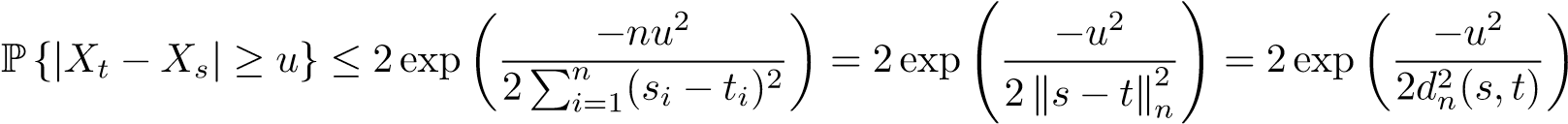
让我们定义关于Rn的以下规范：

.

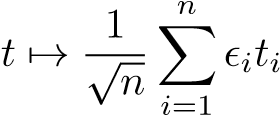
√

换言之，ktkn是t除以n的通常欧氏范数。也让dn（s，t）：=ks−tkn是Rn上的对应度量。

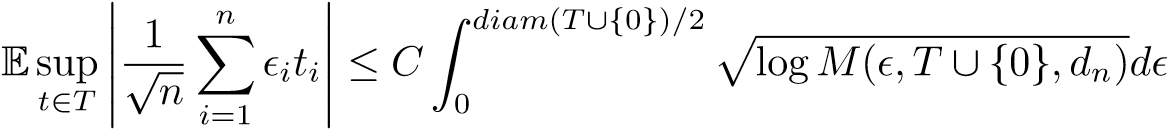
通过Hoefffding不等式，对于每u≥0，



因此，Xt，t∈t满足达德利定理中关于度量dn的假设。还要注意T=Rn是微不足道的可分的，并且映射



是线性的（因此在t中是连续的）这意味着Xt，t∈t是可分的。因此我们可以应用达德利定理我们应用定理8.4，其中t0=（0，…，0）（因为这个向量可能不包含在T中，我们将应用定理8.4到T∪{0}）来获得

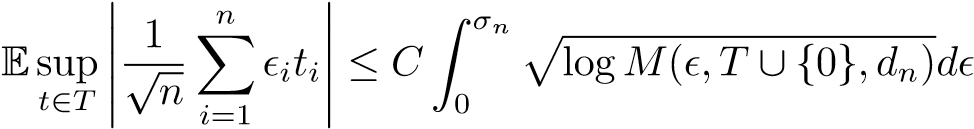


其中，上述直径和填料编号与dn公制有关现在很容易看出

直径（T∪{0}）=sup ks−tkn≤2σn，其中σn：=supktkn。

s，t∈t∪{0}t∈t

因此，我们得到以下上界：



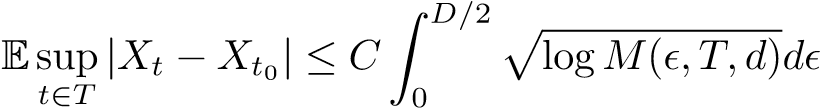
在下一节课中，我们将把上界和对称化技术结合起来，这将给我们一个关于经验过程上确界的重要上界。

# 第九讲

让我们从最后一类的Dudley熵界开始：假设（T，d）是一个度量空间，{Xt，T∈T}是一个满足

对于所有u≥0，s∈T，T∈T。

那么对于每个t0∈T，我们有



其中D表示度量空间的直径（T，D）。

我们应用这个界来控制Rademacher平均值的期望上确界。假设T是Rn的子集。那么

（57个）

哪里

，和dn（s，t）：=ks−tkn。

注意，如果T是有限的，那么



因此边界（57）意味着

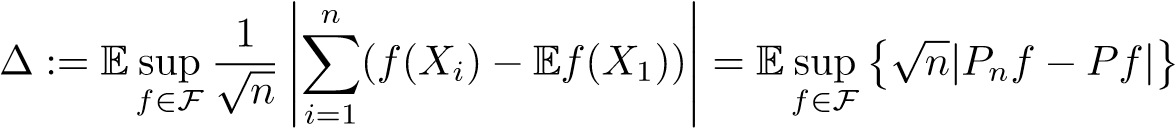
.

注意，我们已经通过对亚高斯随机变量的期望最大值的初等界证明了上述约束。

我们现在将（57）和对称化一起应用，得到经验过程的期望上确界的主界。

## 9.1经验过程的期望上确界

考虑通常的经验过程设置。我们的目标是获得∏的上界，其中



√

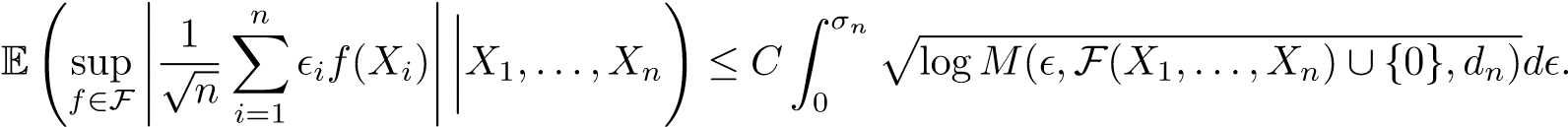
注意在上面的supremums中n的存在我们已经看到对称化给出了

.

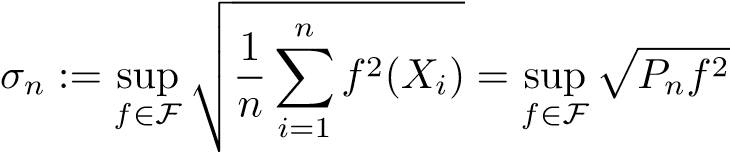
我们把期望写在X1，…，Xn的两部分条件中，得到

.

上面的内在期望可以通过边界（57）来控制。这给了



其中F（X1，…，Xn）：={（F（X1），…，F（Xn））：F∈F}是Rn的子集，



n按√n标度，dn是R上的欧氏度量

我们现在写信



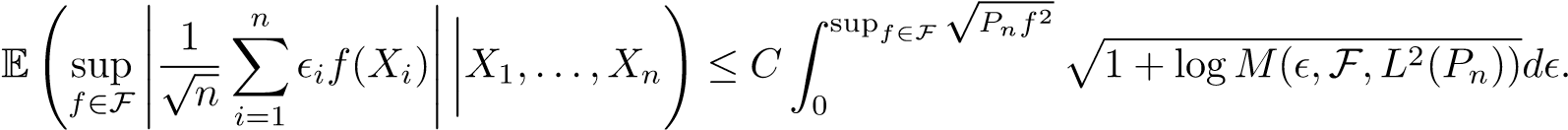
式中，L2（Pn）是指

.

注意平凡不等式



因此我们得到



带着期望，我们得到

.

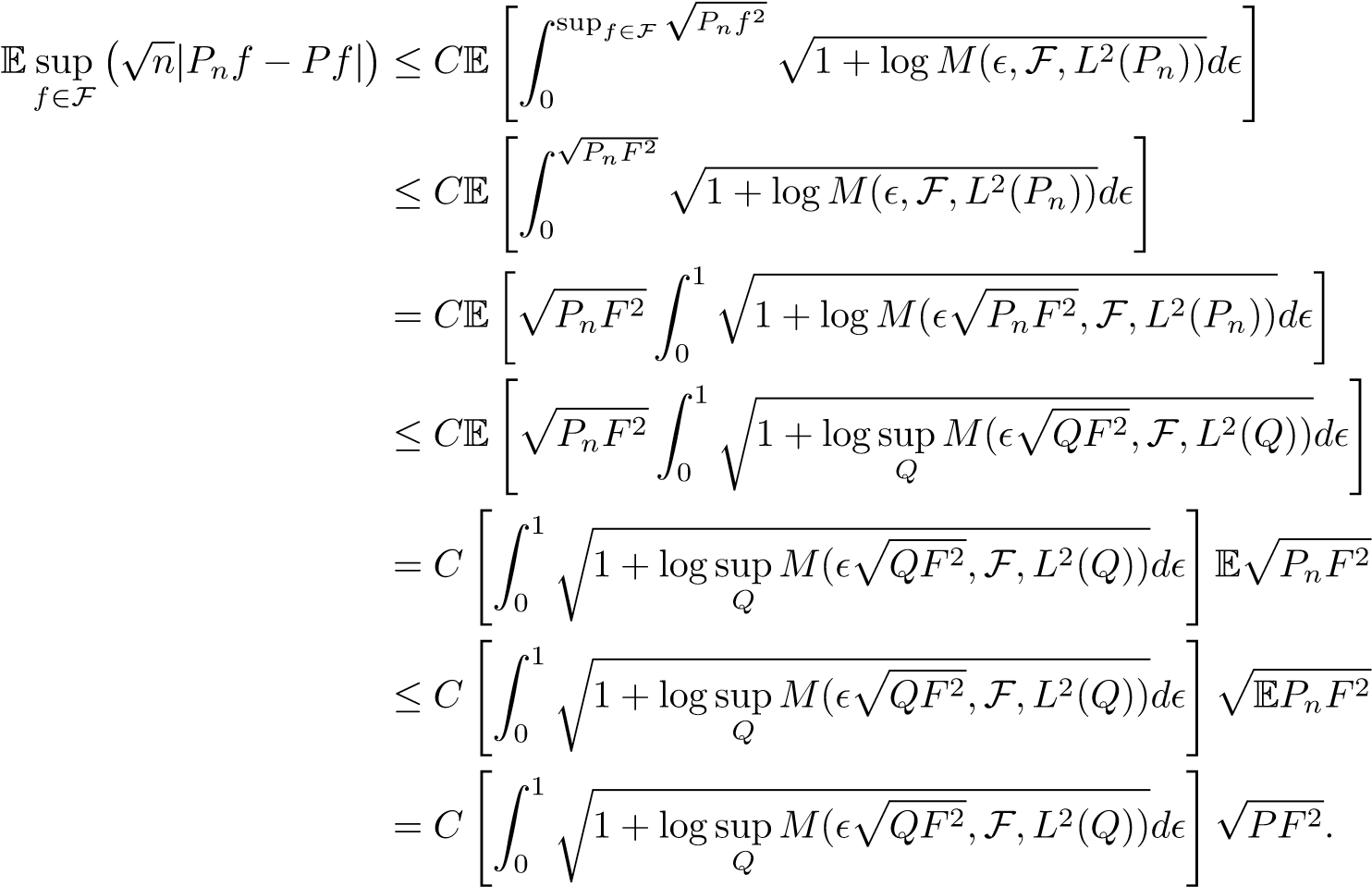
这是我们关于经验过程的期望上确界的第一个界我们可以用信封进一步简化这个界限。我们假设一个非负值函数F:X→[0，∞）是F类if的一个包络

sup | f（x）|≤f（x）对于每个x∈x。

f∈f

√

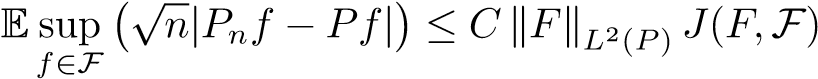
很明显，supf∈F pPnf2≤PnF2



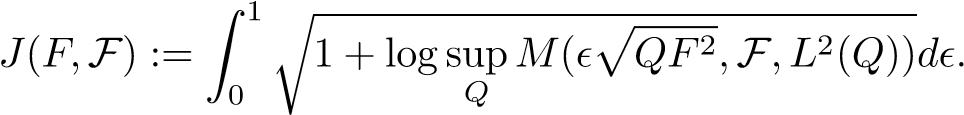
在上述不等式链中，上确界是所有概率测度Q的上确界，Q在X中至多支持一组基数n。PF2也表示EF2（X1）。

因此，我们证明了以下结果。

定理9.1。设F为F类的一个包络，使得PF2<∞。那么



哪里



## 9.2有限维布尔函数类的应用

设F是一个具有有限VC维数的布尔函数类，设D表示其VC维数回想VC维度被定义为X中被F类所粉碎的集合的最大基数，关于VC维度的一个重要事实是Saer-Selay-Vaulnk CelvoNekki引理，它表示

（58）

对于n≥1和x1，…，xn∈X其中

F（x1，…，xn）：={（F（x1），…，F（xn））：F∈F}。

注意在（58）中，如果n<k，则取0。如果n≤D，则（58）的右手等于2D，如果n≥D，则以（en/D）D为界

n≥D（59）

并通过对称化和Rademacher平均上的初等界证明了这个界这个初等界涉及到F（X1，…，Xn）的基数，我们通过（58）来界定它。

然而，事实证明对数因子在（59）中是多余的，并且一个实际上有界

（60个）

这可以作为定理9.1的结果来推导，我们将在本节中演示由于定理9.1给出了包装数的界，因此有必要将F的包装数与其VC维数联系起来。由于达德利的原因，这在下面的重要结果中完成。

定理9.2。假设F是一个具有VC维D的布尔函数类

为所有人。（61）

这里c1和c2是普适正常数，上确界在X上的所有概率测度Q上。

注意，定理9.2给出了-packing数在1时的上界。因为F中的函数只取0和1这两个值，所以很明显）对所有函数都是1

定理9.2的证明。F1，0和1的概率测度Q），并让F1，…，Fn是L2（q）度量中F的一个最大分离子集。这意味着，对于每1≤i 6=j≤N，我们有

.

现在让Z1，Z2是i.i.d.从Q观察到的。根据以上，我们有

.

通过Z1，Z2，…，的独立性，我们推断出对于每一个k≥1，

.

换句话说，这意味着fi和fj在每个Z1，…Zk上达成一致的概率最多在工会的约束下，我们

P{（fi（Z1），…，fi（Zk））=（fj（Z1），…，fj（Zk））约为1。

这马上就给

.

所以如果我们

（62个）

然后

.

因此，对于k的选择（62），存在基数k的子集{Z1，…，ZK}，使得

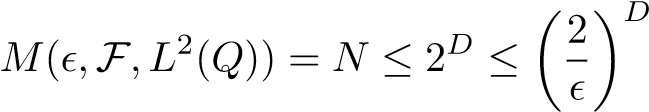
N≤| F（z1，…，zk）|

我们现在应用Sauer-Shelah-VC引理并推断

（63）

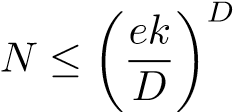
我们现在根据k≤D或k≥D分为两种情况。

情况1:k≤D：这里（63）给出

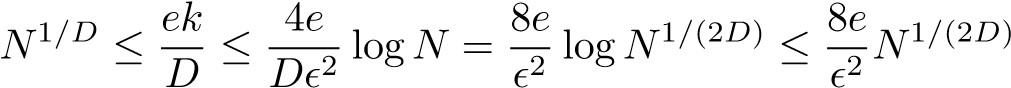


这证明了（61）。

例2:k≥D：这里（63）给出



所以（使用（62）



当我们使用logx≤x时，立即给出

.

定理9.2的证明是完整的。

边界（60）紧跟在定理9.1和定理9.2之后，如下所示。

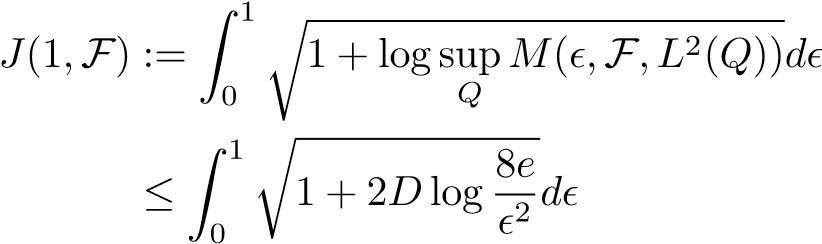
定理9.3。假设F是一个具有VC维D的布尔函数类，那么

（64个）

证明。因为F是一个布尔类，我们可以应用定理9.1，对于所有x，F（x）=1

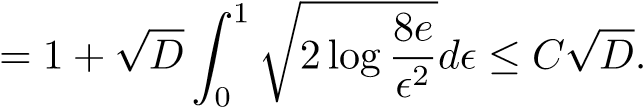
)与

上面的包装数可以由定理9.2限定，定理9.2给出



√ √ √

因为a+b≤a+b



这就完成了定理9.3的证明。

下面是定理9.3的直接应用。

例9.4。假设X1，…，Xn是具有公共cdf的i.i.d实值观测值。让Fn表示数据X1，…，Xn的经验cdf。然后定理9.3立即给出

（六十五）

这是因为布尔类F：={I（－∞，x]：x∈R}具有VC维数1。

我们还可以利用前面讨论的有界差分集中不等式在supx | Fn（x）–F（x）|上获得一个高概率上界这（连同（65））给出

概率≥1α。

例9.5（用VC类分类）。考虑我们观察i.i.d数据（X1，Y1），…，（Xn，Yn）与Xi∈X和Yi∈{0，1}的分类问题。设C是一类从X到{0,1}（这些是分类器）的函数。对于分类器g，我们定义了它的测试错误和训练错误

L（g）=P{g（X1）6=Y1}和

分别是经验风险最小化（ERM）分类器由

gˆn：=阿格明（g）。

g∈C

理解gˆn相对于C类中的最佳测试误差（即L（gˆn）－inf L（g））的测试误差通常是感兴趣的。

g∈C

如果g\*在g∈C上最小化L（g），那么我们可以将上面的差异限制为

L（gˆn）－L（g∗）=Ln（gˆn）－Ln（g∗）+L（gˆn）－Ln（g∗n）＋L（g∗n）－Ln（g∗n）＋L（g∗n）－Ln（g∗n）＋L（g∗n）－Ln（g∗n）

不大于2sup | Ln（g）–L（g）|。

g∈C

上面的最后一个不等式可能相当宽松（我们稍后将研究改进的边界）上述术语可以写成supf∈F | Pnf−Pf |，其中

F：={（x，y）7→I{g（x）6=y}:g∈C}，

P n是（Xi，Yi）的经验分布，i=1，…，n和P是（X1，Y1）的分布。

利用有界差分集中不等式和定理9.3给出的界，我们得到（对于每个α∈（0,1））

（66个）

概率≥1α。

现在可以证明V C（F）≤V C（C）为了证明这一点，有足够的理由认为如果F可以粉碎（x1，y1），…，（xn，yn），那么C可以粉碎x1，…，xn为此，设η1，…，ηn在{0,1}中任意我们需要得到G（Xi）＝εi的函数g\*，I定义δ1，…，δn。

δi=ηiI{yi=0}+（1-ηi）i{yi=1}。

因为F可以粉碎（x1，y1），…，（xn，yn），所以存在f f（Xi，Yi）＝δi的函数f＝1，…，n，如果f（x，y）＝i {g（x）6＝y}，对于某些G c，则现在很容易验证G（Xi）＝εi。这证明C粉碎了X1，…，Xn。V C（F）≤V C（C）的证明是完整的因此，我们从（66）中得到，

prob≥1α。

因此，只要V C（C）=o（n），gˆn相对于C中的最佳测试误差的测试误差收敛到零，即n······。

# 10讲座10

## 10.1总结上节课的主要经验过程

让我们首先回顾一下我们关于经验过程的期望至上的主要界限。如果F是X上一类具有F包络的实值函数，那么（假设PF2<∞），我们有

)（67个）

哪里

这个界的一个重要含义是，当F是一个具有有限VC维数D的布尔函数类时，

（68）

今天，我们将讨论界（67）和（68）在M-估计问题中的应用。我在这里的治疗将是严谨和启发的混合下周我们将回到M-估计，届时所有的启发式论证都将被严格化。

## 10.2 M-估计问题的应用

这可以看作是一个模式估计问题。假设X1，…，Xn是单变量密度p的i.i.d观测值，我们假设p有一个单模θ0，它关于θ0∈R是对称的，另外，我们假设p是光滑有界的，对于x<θ0，p0（x）>0，对于x>θ0，p0（x）<0你可以认为p是平均θ0的法向密度，或者是以θ0为中心的柯西密度。

现在考虑估计θ0的问题。为此，让我们来定义

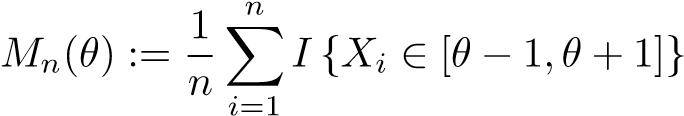
（69）

请注意

M0（θ）=p（θ+1）–p（θ-1）。

由于对p的假设，很明显，M0（θ0）=0，对于θ6=θ0，θ>θ0的M0（θ）<0，对于θ<θ0的M0（θ）>0。这意味着Th＝7 m（Th）在Th 0上具有唯一的最大值。同时M00（θ0）=p0（θ0+1）–p0（θ0–1）<0。

由于Th 0在极大值上最大化M（Th），因此估计Th 0的一个合理的方法是用Tyth-n来估计它，其中Tyth\* n是Mn（Th）上的最大最大值（Th）。



现在自然要问以下问题：

1. θˆn作为θ0的估计量是否一致，即|θˆn-θ0 |在概率上收敛到零。
2. 正如我们将要展示的，在这个例子中有一个是一致的。然后我们可以问：|θˆn-θ0 |到零的收敛速度rn是多少？
3. rn（θˆn-θ0）的极限分布是什么？

下面我们将严格证明θˆn的一致性。我也将提供一个启发性的论据，以找到我们下周将严格制定的利率rn在讨论统一中心极限定理之后，将在几周内讨论极限分布。

分析M-估计量的基本第一步是以下不等式：

0≤M（θ0）－M（θˆn）=Mn（θ0）－Mn（θˆn）+M（θ0）－Mn（θ0）－M（θˆn）+Mn（θˆn）

≤M（θ0）－M n（θ0）－M（θˆn）＋Mn（θˆn）

在这里，我们使用了不等式（Mn（Th 0））Mn（Tyth-n），因为它的最大值是Mn（·）。我们可以用经验过程表示法重写上述不等式。对于θ∈R，我们定义函数mθ：R→R

mθ（x）：=I{θ−1≤x≤θ+1}。

用这个符号，不等式变成

（七十）

这种不平等是如此根本，以至于被称为基本不平等。

为了从（70）导出θˆn的一致性，我们可以粗略地将（70）的右手边定界为

.

现在很容易检查{mθ，θ∈R}是一类具有VC维数2的布尔函数，因此不等式（68）意味着

sup | Pnmθ−Pmθ|→p0。θ∈R

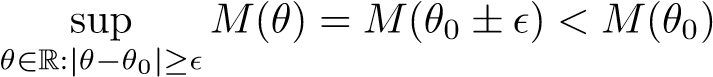
结合（70），我们得到

（71）

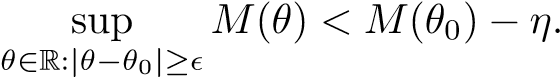
由于我们对p的假设，下面是正确的：

)每一个（72）

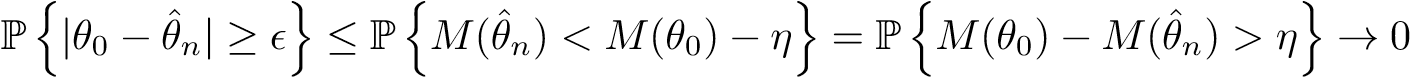
实际上，对于M（θ）如（69），在我们对p的假设下，我们有



因为M（·）在t 0具有唯一的最大值。两个假设（71）和（72）一起暗示|θˆn−θ0 |→p0。要看到这一点，首先固定>0，然后使用（72）获得η>0，这样



接下来就是



其中最后一个（收敛为零）断言来自（71）。这证明了|θˆn-θ0 |→p0。

证明M-估计一致性的这个论点是非常普遍的，并且可以在以下定理中分离出来（例如，在Van der Vaart[24，定理5.7]中可以找到）。

定理10.1（一致性）。设{Mn}为θ∈Θ的随机函数序列，设{M}为θ∈Θ的固定确定性函数。设TH\*N是{Mn（Th），Th }的任意极大值，并设Th 0是{M（Th），T} }的唯一最大化子。假设以下两个条件成立

1. supθΘ| Mn（θ）–M（θ）|→p0。
2. 对每个人来说，不平等。这里的d是以Θ为单位的度量单位。

然后d（θˆn，θ0）→p0作为n····。

假设Suth-TuxFiMn（Th）m（Th）ε- p 0对于一致性来说往往太强（也不总是容易检查），但存在较弱条件的结果。

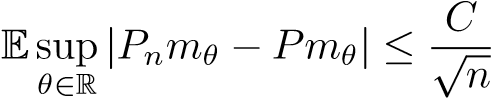
既然θˆn的一致性已经建立，下一个自然问题是关于收敛速度我们可以先重新讨论一致性参数，看看它给出了|θˆn−θ0 |的显式收敛速度我要试探性地争辩。

上面给出的一致性论证是基于不等式：

（73）

为了一致性，我们使用上面的右手边在概率上收敛到零但是不平等

（68）实际上意味着



它给予

.

不等式（73）给出

0≤M（θ0）－M（θˆn）=操作（n－1/2）。（74）

从这里开始，为了得到|θˆn−θ0 |的显式速率，我们需要使用函数M（·）的一些结构。注意M在θ0处的二阶导数等于

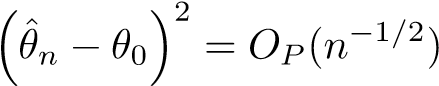
M00（θ0）=p0（θ0+1）–p0（θ0–1）

这是严格负的，因为根据假设，p0（θ0+1）<0，p0（θ0-1）>0其结果是，存在一个常数的C和0的π，因此对于那个邻域中的所有th\*，我们可以写。

|  |  |
| --- | --- |
| M（θ0）－M（θ）≥C（θ－θ0）2。  C值与M00（θ0）有关。用这个，我们可以试探性地写 | （75个） |
| M（θ0）－M（θˆn）≥C（θˆn－θ0）2。 | （76个） |

换句话说，我假设θˆn属于θ0的邻域，其中（75）保持。由于θˆn（我们已经严格证明）的一致性，可以使不等式（76）变得严格。结合

（76）根据（74），我们推断



它给予

.

因此，我们得到了n-1/4作为|θˆn-θ0 |的收敛速度。但事实证明

.

换言之，n-1/4比实际收敛速度慢，反映了我们证明技术中的一些错误宽松的主要根源在于不平等：

)（77个）

原来上面的左手边比右手边小得多。为了启发式地理解上面左手边的大小，让我们首先计算

E |（Pn-P）（mθ-mθ0）|

对于接近θ0的固定θ清楚地

（78个）

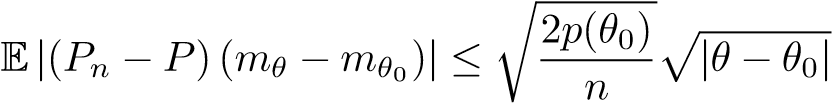
现在对于接近θ0的θ（和θ<θ0），我们有

mθ（x）－mθ0（x）=I{θ－1≤x≤θ0－1}+I{θ+1≤X1≤θ0+1}。

因此

E（mθ（X1）－mθ0（X1））2≤P{θ－1≤X1≤θ0+1}+P{θ+1≤X1≤θ0+1}≤2p（θ0）－θ0 |。

其中，在上一个不等式中，我们使用了这样一个事实，即X1的密度在θ0处有一个模式（这样，在每个其他点的密度都有p（θ0）的限制）。结合这个和（78），我们得到



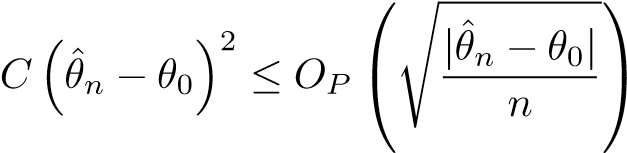
以便

.

对于接近θ0的固定θ，这是正确的试探性地说，这表明

.

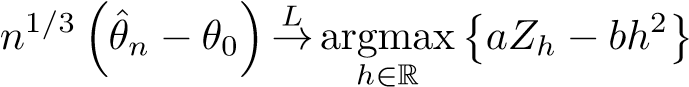
我们稍后将正式证明这一点。请注意，与我们之前的界限（77）相比，这个界限是一个更强的界限将其插入基本不等式（70）的右边，并使用（70）左边的二次界（76），我们推断出



“消去”|θˆn-θ0 | 1/2从两边，我们推断

（七十九）

如前所述，这是θˆn-θ0的正确速率。事实上，事实证明



当n→∞时，其中Zh，h∈R是从0开始的双边布朗运动，a，b是依赖于p和θ0的两个常数我们稍后将证明这个极限结果，但现在它告诉我们n-1/3是正确的收敛速度。

我们将在下周严格计算出价格结果（79）。严格论证中的一个关键因素是建立不平等

（80个）

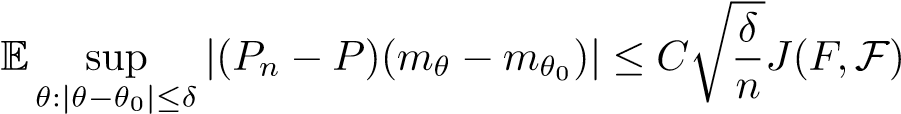
δ足够小。上述不等式可作为（67）的结果导出。实际上，要应用（67），我们首先需要获得类{mθ−mθ0:|θ−θ0 |≤δ}的包络。不难看出

F（x）：=I[θ0～-1～-δ，θ0～-1＋δ]（x）+I[θ0＋1～-δ，θ0＋1＋δ]（x）

是一个包络函数进一步

PF2≤2P{θ0−1-δ≤X1≤θ0−1+δ}+2P{θ0+1-δ≤X1≤θ0+1+δ}≤Cp（θ0）δ≤Cδ。

因此（67）给出



式中F：={mθ–-mθ0：|θ–θ0 |≤δ}。如果J（F，F）<∞，这将证明（80）这将从类{mθ-mθ0}具有有限的VC子图维数（稍后将定义）这一事实出发。注意，这不是一个布尔函数类，所以我们需要VC子图维度而不是VC维度。

现在让我们总结一下关于M-估计率的启发式论证的讨论虽然我们做了一个对应于Mθ（x）：=I{θ-1≤x≤θ+1}的特殊M-估计，但这些思想实际上是相当普遍的。最重要的因素是基本不平等（70）。左手边P（mθ0-mθˆn）从下面由θ7→Pmθ在θ=θ0处的二阶导数的假设限定。右手边可以通过计算来理解

E（mθ（X1）－mθ0（X1））2。

对于mθ（x）=I{θ−1≤x≤θ+1}的具体选择，结果是

E（mθ（X1）－mθ0（X1））2≤C |θ－θ0 |。

对于其他mθ，右侧可能不同（例如，通常在右侧有C |θ−θ0 | 2）。在基本不等式中插入这些界，将得到一个包含|θˆn–θ0 |和n的不等式，该不等式可以被求解以获得该问题中的显式速率（例如n−1/3）。这种试探性的做法将在下周得到证实。在结束本节之前，让我们陈述Van der Vaart和Wellner[25，第294页]关于M-估计收敛速度的一个结果。我们稍后将正式证明这一结果，但基于上述的启发式，其结论应该是相当明显的。

定理10.2（Van der Vaart和Wellner，第294页）设X1，X2假设Θ⊆R是一个开集，设mθ，θ∈是X上的实值函数的集合，这些实值函数被Θ索引假设在x上存在α＞0和函数m，其中PM2 <

|mθ1（x）－mθ2（x）|≤m（x）|θ1－θ2 |α，所有θ1，θ2∈Θ。（81）

设n，n和0，表示在0，0，0，14，0，0，0，0，0，0，0，0，0，0，0，0，0，0，0，0，0，0，0。如果θ→7pmθ在θ0处有两个导数，且二阶导数严格为负，则

|θˆn-θ0 |=OP（n−1/（4−2α））（82）

启发式论证。基于定理中的假设，我们使用上面总结的启发式来证明（82）注意，假设（81）意味着

.

这表明了启发式

.

结合基本不等式和P（mθ0-mθˆn）n上的下界C（θˆn-θ0）2，我们得到了不等式

（θˆn-θ0）2。n−1/2 |θˆn−θ0 |α

它给予

|θˆn-θ0 |n-1/（4-2α）。

注意，当α=1时，定理10.2给出了|θˆn−θ0 |的通常n−1/2速率。

# 11讲座11

对于布尔函数类F，我们已经看到VC维给出了覆盖数的有用上界：

全部0。（83）

对于非布尔函数类会发生什么情况？今天我们将看到一般函数类存在组合维数的两个概念，它们允许通过类似于（83）的边界来控制覆盖数。这是我们今天要讨论的VC子图维数和脂肪破碎维数的概念。

## 11.1 VC子图维数

F的VC子图维数只是取F中函数子图的指标得到的布尔类的VC维数，为了形式化地定义它，我们首先定义函数子图的概念。

定义11.1（子图）对于函数f:X→R，其子图sg（f）是X×R的子集，定义为

sg（f）：={（x，t）∈x×R:t<f（x）}。

换言之，sg（f）由函数f的图形下的所有点组成。

我们现在可以定义F的VC子图维数为：

定义11.2（VC子图维度）。F的VC子图维数定义为布尔类{Isg（F）：F∈F}的VC维数。我们用V C（F）来表示。

VC子图维数与覆盖数的关联方式与（83）相同这在下面的结果中完成。

定理11.3。假设F是一类包络F的函数，设F的VC子图维数等于D，则

全部1（84）

其中c1和c2是通用正常数。

证明。其思想是将F中两个函数之间的L2（Q）范数与子图之间的L2范数联系起来。固定f，g∈f并写入

Z Z轴

|f-g | 2dQ≤2F（x）| f（x）–g（x）| dQ（x）

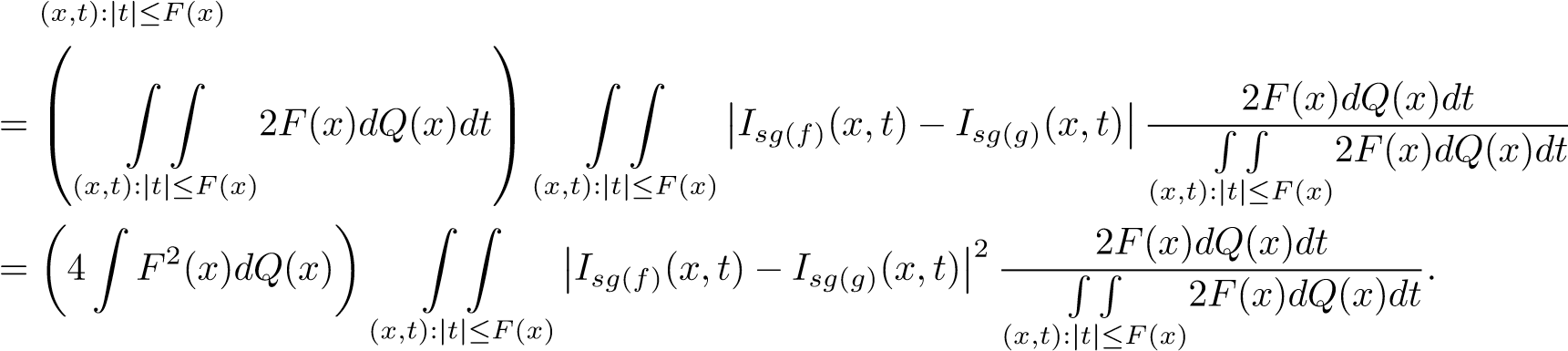
其中我们使用了| f（x）–g（x）|≤2F（x）（这是真的，因为f是f的包络）我们现在使用的事实是，对于每两个实数a和b，我们都有

Z轴

|a-b |=| I{t<a}-I{t<b}dt

这给了

as Isg（f）（x，t）=Isg（g）（x，t）对于| t |>f（x）



我们已经证明了

（85个）

其中Q0是X×R上的概率测度，其密度与Q×Leb成正比

2F（x）I{（x，t）：| t |≤F（x）}。

从（85）推断出



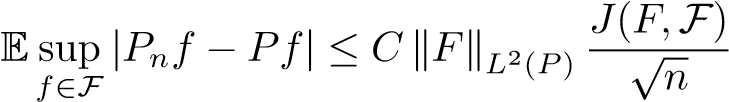
要绑定右侧，我们只需将前面的结果用于布尔类（请参见（83））。这就完成了定理11.3的证明。

下面是定理11.3的直接推论和我们关于经验过程的期望上确界的主要定界。

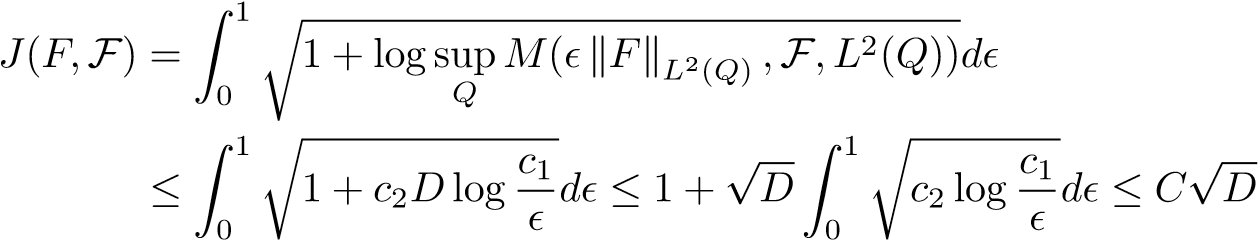
推论11.4如果F有包络F和VC子图的维数D，则

.

证明。我们关于经验过程的期望上确界的主要界给出了

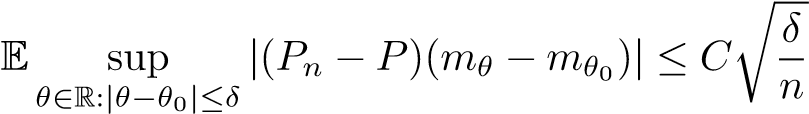


我们把J（F，F）（用定理11.3）定为



这就完成了推论11.4的证明。

例11.5在上节课上，我说



式中mθ（x）：=I{θ-1≤x≤θ+1}在上节课上我对这个事实作了部分证明通过证明函数类



有有限的VC子图维数（≤3？?).

现在让我们看一看VC子图维数的重新定义。这直接根据类F定义维度，而不必转到子图。这一重新制定也将明确与脂肪粉碎维度的联系。

我们假设x的子集{x1，…，xn}是F的子图，如果存在实数T1，…，Tn，使得对于每个子集S {{ 1，…，n}，存在f f f。

|  |  |  |  |
| --- | --- | --- | --- |
| （xs，ts）∈/sg（f）对于s∈s  注意（86）相当于 | 和 | （xs，ts）∈sg（f）表示s/∈s。 | （86个） |
| 对于s∈s，f（xs）≤ts | 和 | f（xs）>ts代表s/∈s。 | （87个） |

在词中，我们说{x1，…，xn}是f的子图，如果存在水平t1，…，tn，使得对于每个子集S {{ 1，…，n}，存在一个函数f，它在每个xs，s s的水平之下，严格地在xs，s/s的水平上。

VC子图维V c（f）被定义为X的有限子集的最大基数，即F所划分的子图。

现在让我们描述一个使用VC子图维数来控制覆盖数的潜在问题设M表示所有非递减函数的类f:R·········································。结果是

为所有人。（88）

也很容易看出M的VC子图维数等于（即，对于每n个1以上，存在一个有限的子集R，即子图被M打碎）。因此，VC子图维数的概念在这里没有用处，定理11.3也没有给出对这类M有意义的任何东西。实际上，可以使用下面讨论的脂肪破碎维数的概念来证明（88）。

## 11.2脂肪粉碎尺寸

脂肪粉碎是一个尺度敏感的概念维度。具体地说，脂肪粉碎维度实际上是（0，~）上的一个函数，即它是为每一个>0定义的。我们将用fat来表示，定义如下。

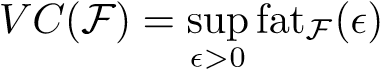
定义11.6（粉碎）。我们说，如果存在实数t1，…，tn，则由F粉碎的子集，使得对于每个S {{ 1，…，n}，存在函数f f，使得

f（xs）≤ts，对于s∈s和（89）

在词中，我们说{x1，…，xn}是f的子图，如果存在水平t1，…，tn，使得对于每个子集S {{ 1，…，n}，存在一个函数f，它在每个xs，s s的水平之下，并且超过xs的水平，注意-shattering概念和子图shattering概念与前一小节的唯一区别是“严格超过”替换为“超过”。

定义11.7（脂肪粉碎尺寸）因为，脂肪粉碎维度FAT被定义为被F粉碎的有限子集X的最大基数。

很明显，脂肪）是一种减少的功能。事实上，可以证明（练习）



其中V C（F）表示VC子图维数这个不等式意味着，特别是，fat）总是由V C（F）从上面限定的。另一个简单的事实是，当F是布尔值时，fat）等于V C（F）每0

现在回想一下最后一个小节中的M类，它包含所有非减缩函数f:R→[-1,1]前面已经提到M的VC子图维数是无穷大的现在我们要证明脂肪是有限的

每个人都很胖。

事实上，上述脂肪破碎的维数界适用于所有函数f:R→R的更大类，其变分为2。下面的结果证明了这一点。回想一下，函数f:R→R的变化是由

n-1个

kfktv=＝Sup xf f（Xi）f（Xi＋1）。

n≥1 x1<x2····<xn i=1

引理11.8固定V>0，让F表示所有函数F:R→R的类，其中kfkTV≤V那么

每个人都很胖。

证明。修复>0。让我们先证明

脂肪（90）

对于这个，-被F粉碎。我们将证明n不能大于（90）的右手边，这将证明（90）。首先指出，因为{x1，…，xn}是破碎的，所以存在满足t-，…，tn的实数t-满足粉碎定义的条件。这意味着，特别是F中存在f1和f2的两个函数，它们满足以下条件：

：对于奇数i

：即使是我

和

：即使是我

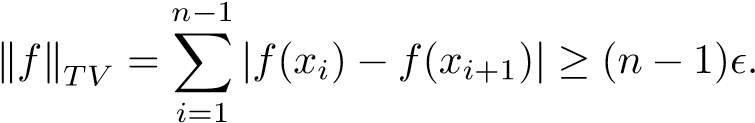
：对于奇数，我现在让f=（f1−f2）/2上述条件共同意味着

2:奇怪的是

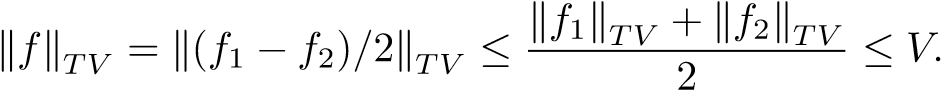
F（XI）是

2:即使是我

这就意味着



另一方面，



结合以上两个不等式，我们得到（这意味着（90）（注意n必须是一个整数，它允许我们把整数部分放在

我们现在证明脂肪）大于或等于（90）的右手边。为了这个，让。考虑任意一组d点y1<···················。设G由2d+1函数组成，这些函数在每个区间I0，…，Id上都是分段常数。设{x1，…，xd+1}是通过从每个d+1区间I0，…，Id中选取一个点而得到的任何有限集。然后很明显，G粉碎了这一点。此外，中每个函数的变化因此F粉碎{x1，…，xd+1}，这意味着

胖。

这就完成了引理11.8的证明。

现在让我们描述一个结果，它将覆盖数绑定到脂肪粉碎维度fat 0这是Mendelson和Vershynin的以下定理[16]。

定理11.9（Mendelson Vershynin）假设F是一类由1一致限定的函数然后存在一个普适的正常数C＝1，使得

为所有人。（91）

让我们看看这个结果对于所有非减缩函数f:R的M类给出了什么样的结果-[-1,1]。这个类中的每个函数最多有2个变量，因此引理11.8意味着

脂肪（92）

这，加上定理11.9，让我们可以推断

对于0。

注意，与指数中的对数（2/）因子（88）相比，该结果较弱我们现在可以问是否有可能通过脂肪粉碎维度导出（88）。如果绑定（91）可以改进为exp（），则这是可能的。请注意，一般情况下不能这样做例如，当F是具有有限VC维数的布尔值时，那么fat）对于每0 1，在这种情况下，显然不能用（91）中的常量替换2/然而，Rudelson和Vershynin[21]已经表明，在关于脂肪的技术正则性假设下（这排除了F是有限维布尔函数的情况），确实有可能改进定理11.9结果如下。

定理11.10（Rudelson Vershynin）。假设F是一类函数。假设a>2和递减函数v:（0，~）—（0，~）是这样的

脂肪和）（93）

为所有人那么

（94个）

因为一切照常都是一个普遍的常数。

注意，正则性条件（93）排除了诸如情况when）为常数的情况。还要注意，F中的函数没有明确的有界性假设；这隐藏在正则性条件中。

现在让我们证明，定理11.10确实暗示了M类非递减函数的结果（88），该类函数被约束为取[-1,1]中的值事实上，对于这门课，我们首先有（92）。另外，由于M中的函数被限制为取–-1,1中的值，因此脂肪破碎维度将大为零事实上，很容易看出3的fat=0。这与（92）一起意味着

大家都胖。

因此，我们可以应用定理11.10。很容易看出条件2满足a＝3。很明显，不平等（88）是（94）的结果。

# 12讲座12

## 12.1支架控制

到目前为止，我们的主要经验过程如下在通常的符号下：

)（95个）

哪里



括号方法提供了Esupf∈F | Pnf | Pf |的另一个上界，我们接下来将描述这个上界这个边界将非常类似于（95），除了sup）将被替换为

-L2（P）中F的括号数。在我们陈述这个结果之前，让我们首先定义括号数字的概念：

1. 给定X上的两个实值函数和u，方括号[`，u]定义为所有函数f:X→R的集合，其中`（X）≤f（X）≤u（X）表示所有X∈X。
2. 给定X上的概率测度P，方括号[`，u]的L2（P）-大小被定义为ku-`kL2（P）。
3. 设F是X上的一类实值函数，当F>0时，方括号数）被定义为每个方括号的最小个数，每个方括号的大小不超过L2（P），使得每个F∈F都属于其中一个方括号。

必须注意的是，括号中的数字大于下面所示的覆盖数字。

### 引理12.1。每一次，

.

这里Fall表示X上所有实值函数的类。

证明。第一个不等式是我们在讨论覆盖数时已经看到的第二个不等式证明如下首先得到括号[其中包括F。

很明显，在L2（P）度量中，中点函数（`i+ui）/2，i=1，…，N构成F的一个/2-网。

接下来，我们提供一个可以显式计算方括号数字的示例。

例12.2。设F：={I（∞，t）：t∈R}设P是R上的一个固定概率测度，然后我们将证明

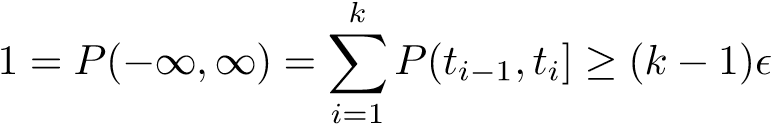
每一个（96）

这是（96）的论点。让t0：=－∞并递归地定义

.

然后，对于足够小的每一个δ>0，很明显（因为否则将是上确界），因此（通过让δ>0），我们推断出。同样，如果ti<~，那么对于每一个δ>0，我们有这样的结果（通过让。

设k≥1为最小整数，其中tk=∞。那么，根据上述，我们得到i=1，…，k-1，所以



这就给了。现在考虑I=1，…，k的括号[I（－∞，ti-1），I（－∞，ti]



我们已经证明了



这一点对所有人来说都是正确的，与（96）相同。

在讨论（95）中包含括号数的类似项之前，让我们先讨论并证明一个简单的经典渐近结果，它表明括号数界可用于控制Esupf∈F | Pnf−Pf |。

提议12.3。假设F是一个函数类，因此对于每个那么

sup | Pnf−Pf |→p0为n→∞。（97）

f∈f

证明。修复>0。设[`i，ui]，i=1，…，N表示覆盖F的L2（P）-大小括号。我们首先要论证

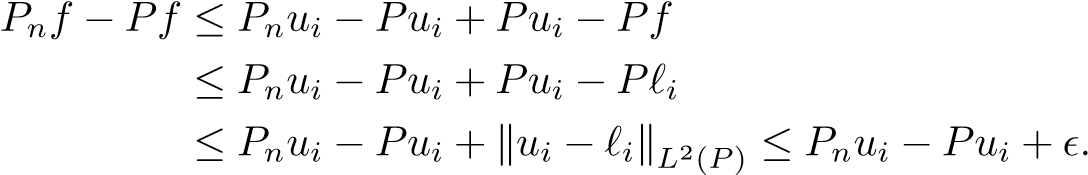
（98个）

让我们先完成（97）的证明，假设（98）是真的要看到这一点，请注意上面的右手边几乎肯定收敛到0，因为n→∞。这是因为，通过强大的大数定律，pNi·PUI和Pn’i’i’i’i几乎收敛到零，对于每个i（n，n，n，n，i，i，i，i，n，n，n，n，i，i，i，n，n，n，n，n，i，i，n，n，n，n，n，n，n，i，i，n，n，n，n，n，n，n，n，n，n，n，i，i，n，n，n，n，n，n，i，n，n，n，n，n，n，n，n，n，n，n，n，n，n，n，n，n，n，n，n，n，n，n，n，n，n，n，n，n，n，n，n，n，n，n，n，n，n，n，n，n，n，n，n，n，n，n，n，n，n，n，n，n，n，n，n，n，n，n，n，n，因此，从（98），我们推断

几乎可以肯定。

将其应用于每一个，并让m·····，就可以推导出（97）。

因此，还有待证明（98）固定f∈f，得到一个包含f的括号[`i，ui]，这意味着对于每个x∈x，i（x）≤f（x）≤ui（x）



同样可以证明这一点。这两个不等式加在一起意味着（98）完成了命题12.3的证明。

我们现在来说明（95）的类似物，包括括号中的数字这将是我们界定经验过程的预期上限的第二个主要结果（第一个主要结果是（95））。

定理12.4。设F为F类的一个包络，使得PF2<∞。那么

)（99个）

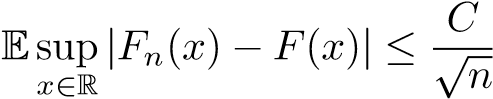
哪里

界（99）与（95）非常相似，唯一的区别是“均匀”覆盖数

))被括号中的数字）替换为-

spect到L2（P）重要的是，注意到在（99）中有Q的上确界，括号中的数字只涉及测量P。相反，如果sup被替换为

例12.5假设X1，…，Xn是具有cdf的i.i.d观测值，设Fn为经验cdf。我们以前见过



每n≥1这是（95）的推论。我们将在这里证明，这也可以通过（99）推导出来。实际上，对于F：={I（－∞，t）：t∈R}，我们已经得到了它的界。我们从中推断出

.

下面给出了一种情况，即包围包围数字比包围均匀覆盖数字更容易处理。

提议12.6。设Θ⊆Rd包含在半径R的球中。设F：={mθ：θ∈Θ}是由Θ索引的函数类假设存在一个具有kMKL2（p）<

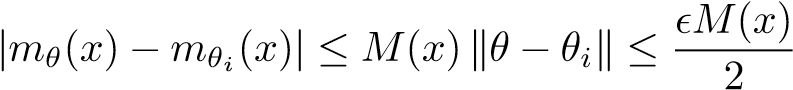
|mθ1（x）－mθ2（x）|≤m（x）kθ1－θ2k（100）

对于所有x∈x和θ1，θ2∈Θ（这里k·k表示通常的欧氏范数）那么每一次，

（101）

证明。设1，…，Th n是欧氏度量中的极大/ 2-填充子集。考虑一下括号请注意

1. 这些括号覆盖F，对于每一个角，都有1个2。然后根据条件（100），



这意味着mθ位于括号中[

1. 这些括号的L2（P）大小最多。这是显而易见的。

由于这两个观察结果，））是从我们先前限定的Θ的欧几里德数/2包数的上方限定的。这就完成了命题12.6的证明。

## 12.2 M-估计

我们现在来讨论本课程的第一个统计学主题：M-估计。基本的抽象设置如下。

设Θ为抽象参数空间。通常，它是参数估计问题Rd的子集，或者是非参数估计问题的函数类我们有两个过程（一个是随机的，一个是确定的）被θ∈Θ索引随机过程通常依赖于“样本”大小n，并用Mn（θ）、θ∈Θ表示确定性过程通常不依赖于n，而仅用M（θ），θ∈Θ表示我们预计n大时Mn接近M。

设Tyth\* n表示Mn（Th）上的最大值（Th），并设Th 0表示M（Th）上的最大值（Th）。M-估计的目的是研究θˆn与θ0的关系。

下面介绍一些具体的M-估计。

1. 经典参数估计：最经典的M估计是极大似然估计（MLE）。这里通常有X中的数据X1，…，Xn，即i.i.d具有分布P。一个也有空间上密度的类{Pθ，θ∈Θ}。MLE最大化Mn（Th）＝PN-Logp Th。这里的过程M（θ）是M（θ）：=P logpθ，然后θ0可以被取为Θ中的参数值，就Kullback-Leibler散度而言，Pθ最接近P。

更一般地，对于其他函数Mθ，可以取Mn（θ）=Pnmθ和M（θ）=Pmθ例如，mθ（x）：=| x－θ|对应于中值估计（这里θˆn是样本中值，θ0是总体中值），mθ（x）：=I{| x－θ|≤1}可被取对应于模式估计。

1. 回归中的最小二乘估计：在回归问题中，一个观察数据（X1，Y1），……，（……，），可以被看作是一个有一些分布的观测。

Mn（θ）：=-Pn（y−θ（x））2

超过θ∈Θ。很自然地把这个πn与0，这是最大化。

M（θ）：=-P（y-θ（x））2。

1. 分类中的经验风险最小化过程：这里观察数据（X1，Y1），…，（Xn，Yn），其中Xi∈X和Yi∈{-1，+1}我们将数据建模为具有分布P的i.i.d.让Θ表示从X到R的一类函数；我们将θ（X）的符号视为分类器的输出考虑一下是很自然的

Mn（θ）：=-PnI{y 6=符号（θ（x））}和M（θ）：=-P{y6=符号（θ（x））}。

在这种情况下，θˆn将是误分类率的经验最小值，θ0将是检验误差的最小值，两者都属于Θ类。因此，比较θˆn和θ0的性能是很自然的。

注意，由于Mn（θ）的最小化是一个组合问题，因此很难计算θˆn。为此，我们还研究了Mn（θ）在分类中的其他选择为了激发这些其他选择，让我们首先将上面的Mn（θ）重写为

Mn（θ）=-Pn I{y 6=符号（θ（x））}=-Pnφ0（-yθ（x）），其中φ0（t）：=I{t≥0}。

为了便于计算，通常用另一个凸的类似于φ0的损失函数φ代替φ0φ的常用选择包括（a）铰链损耗：φ（t）：=（1+t）+，（b）指数损耗：φ（t）：=exp（t），和（c）逻辑损耗：φ（t）：=log（1+et）。注意这三个函数在R上是凸的，它们类似于φ0（注意它们也满足φ（t）≥φ0（t）（对于al t）我们将研究最小化

Mn（θ）：=-Pnφ（-yθ（x））超过θ∈Θ

并将其性能与θ0进行比较。

M-估计理论通常涉及三个问题：（a）一致性，（b）收敛速度，和（c）极限行为。一致性假设θˆn和θ0之间的差异收敛到零，即n→∞收敛速度旨在描述这种收敛的精确速度第三个问题的目标是给出在渐近条件下，在n·································。

一致性通常是成立的，上周我们已经看到了一个关于一致性的定理。我们将主要集中讨论收敛速度问题。在许多情况下，收敛速度的结果自动意味着一致性在其他情况下，我们需要一个初步的一致性结果，这样就可以将注意力集中在θ0的局部邻域上，以确定收敛速度。如果需要初步的一致性，而我们上周的一致性定理还不够充分，我们将提供不同的一致性论证让我们暂时忽略一致性，直接讨论价格为了研究极限行为，我们需要关于统一中心极限定理的理论，我们还没有涉及到这些理论；我们将在几周后回到这些理论上来。

## 12.3 M-估计的收敛速度

在一个抽象的环境中工作是最干净的，其中，Tythn最大化了一个随机过程Mn（Th）超过了T\*，而Th 0最大化了一个确定性过程M（Th）。推导利率的论据是从以下基本不平等开始的：

M（θ0）–M（θˆn）≤hMn（θˆn）–M（θˆn）i–[Mn（θ0）–M（θ0）]

我们已经多次看到这个不等式，它是简单不等式Mn（θˆn）≥

Mn（θ0）。为了方便起见，我们将用（M n-M）（θˆn-θ0）表示上方的右侧，以便

M（θ0）－M（θˆn）≤（Mn－M）（θˆn－θ0）。（102）

我们将用这个不等式来研究θˆn到θ0的收敛速度。我们需要首先修正θˆn和θ0之间的误差设这由d（θˆn，θ0）给出。当Θ是Rd的子集时，将d（·，·）作为通常的欧几里德度量是很自然的一般来说，我们不会要求d（·，·）是一个度量；在这个阶段，我们只要求它是非负的。

注意，差异度量d（·，·）在某种程度上与问题无关，因此，为了理解d（θˆn，θ0）的行为，我们需要将其连接到M（θ）或Mn（θ）。通常的假设是：

M（θ0）–M（θ）和d2（θ，θ0）。（103）

这里，符号a和b表示a≥Cb，表示一个普适常数C（符号a）b的定义类似）。

假设（103）对所有θ∈Θ都是真的。在某些情况下，它只在θ0的邻域中是真的（我们稍后再讨论这个问题）。注意（103）如果我们定义d为

d2（θ，θ0）：=M（θ0）–M（θ）。

这是研究M-估计量的最自然的选择在参数估计问题中，这通常与欧氏度量不对应，因此通常不使用此选项。然而在函数估计问题中，这是一个非常常见的选择。

结合（102）和（103），我们得到

d2（θˆn，θ0）。（M n-M）（θˆn-θ0）。

设δn：=d（θn，θ0）那么上述不平等显然意味着

δˆn2.sup（Mn−M）（θ−θ0）

θ∈Θ：d（θ，θ0）≤δˆn

这表明δn满足以下条件的任何速率δn的δn

（104）

我们将在下一节课中对这种直觉加以严格化。临界不等式（104）给出了在各种问题中确定M-估计收敛速度的一种很好的方法。右边的期望值可以通过我们过去几周研究的经验过程方法来控制。

# 13讲座13

## 13.1 M-估计收敛速度的严格推导

让我们回忆一下设置Θ是一个抽象的参数空间。我们有两个过程（一个是随机的，一个是确定的）被θ∈Θ索引随机过程通常依赖于“样本”大小n，并用Mn（θ）、θ∈Θ表示确定性过程通常不依赖于n，而仅用M（θ），θ∈Θ表示我们预计n大时Mn接近M。

设Tyth\* n表示Mn（Th）上的最大值（Th），并设Th 0表示M（Th）上的最大值（Th）。设d（θˆn，θ0）为测量θˆn和θ0之间间隙的非负差分度量我们假设

M（θ0）–M（θ）和d2（θ，θ0）（105）

对于每个θ∈Θ这里，符号a和b表示a≥Cb，表示一个普遍正常数C（符号a）b的定义类似）根据（105），d的标准选择是

d（θ，θ0）：=pM（θ0）–M（θ）。（106）

当d不是上述的标准选择时，通常发生（105）仅在θ0的邻域中成立我们稍后再谈这个问题。

我们现在严格地求出d（θˆn，θ0）收敛速度的上界形式上，我们认为δn是D（thixn，0）收敛到零的速率，如果对于每个＞0，则存在常数C。

概率（107）

注意，这相当于

Pnd（θˆn，θ0）>2Mδno→0作为M······（108）

应该注意的是（107）和（108）是非同构语句（它们对每个有限n都适用）。特别地，它们意味着渐近速率语句：D（THAN n，Th 0）＝OP（δn），这意味着：对于每一个＞0，存在C和整数N，使得

所有人。（109）

（109）和（107）之间的区别在于（109）对所有n保持不变，而（107）对所有n保持不变。

现在我们来研究概率Pnd（θˆ，θ0）>2Mδno n

对于固定的δn和大的M，我们需要了解当M··········。

写

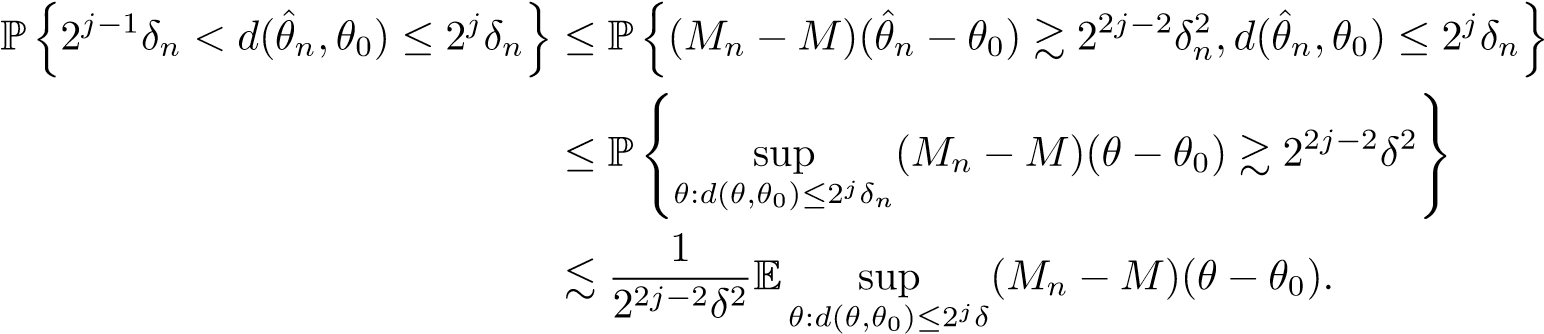
Pnd（θˆn，θ0）>2Mδno=XPn2j−1δn<d（θˆn，θ0）≤2jδno。

焦耳>米

我们现在使用基本不等式（连同条件（105））：

d2（θˆn，θ0）。M（θ0）－M（θˆn）≤（Mn－M）（θˆn－θ0）

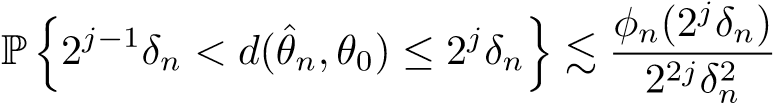
这显然给



假设函数φn（·）是这样的

)每u.（110）

因此我们得到



因此，对每一个j，

.

为了简化上述表达式，通常采用以下假设：存在α＜2；

当c>1和x>0时，φn（cx）≤cαφn（x）（111）

在这个假设下，我们得到

.

当M········因此，如果δn是这样的

,

然后

d（θˆn，θ0）≤2Mδn，概率至少为1-uM

式中，uM→0为M·····。

这给出了以下非共同性收敛速度定理：

定理13.1。假设条件（105）且函数φn（·）满足（110）和（111）然后，对于每M>0，我们得到d（θˆn，θ0）≤2Mδn，概率至少为1-uM，前提是φn（δn）。δn2。这里uM=Pj>M 2j（α-2）→0作为M····。

## 13.2有界Lipschitz回归的应用

假设f0是[0,1]上的未知函数我们观察到根据模型生成的f0上的数据Y1，…，Yn：

对于i=1，…，n

其中是i.i.d N（0,1）随机变量。

假设f0是1-Lipschitz on，在[0,1]上有1的界换言之，我们假设f0∈F，其中F是[0,1]上所有绝对值为1且为1-Lipschitz的函数的集合。在这种假设下，通过

n fˆn=argmin。

i=1

显然fˆn是一个M-估计，我们可以用定理13.1来研究它的一些性质为此，我们首先使用

.

因此

.

为了使用定理13.1，我们可以

.

那是很自然的

.

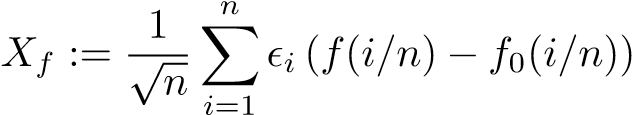
对于差异d，我们可以使用规范选择（106）：

.

为了找到速度，我们需要控制

.

为此，我们可以使用达德利熵界。让



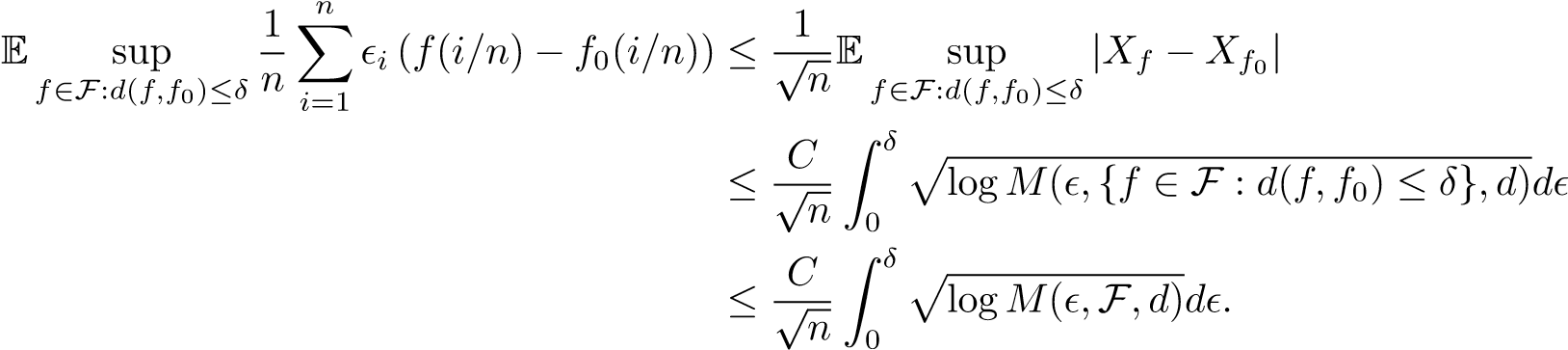
注意到

.

这意味着

所有u≥0。

达德利熵界立即给出



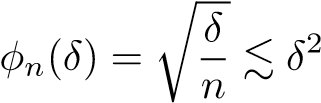
我们之前已经注意到

.

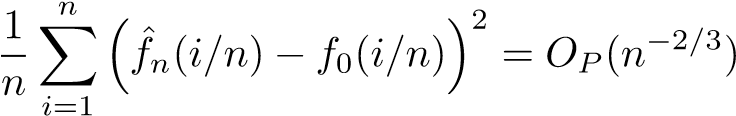
这给了

.

因此，我们可以在定理13.1中取φn（δ）：=pδ/n注意，这显然满足条件φn（c x）≤cαφn（x），α=1/2<2。然后，临界速率确定方程变为：



它给出了δn&n−1/3因此，定理13.1在这里是有效的，其中δn=n−1/3允许我们推导如下：



或者更具体地说，

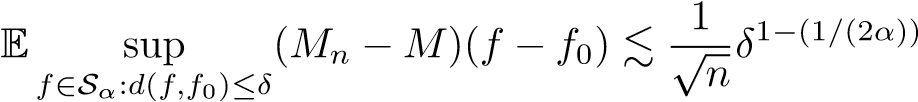
.

因此，我们证明了Lipschitz回归的收敛速度是n 2/3。我们稍后将证明，在极小极大或最坏情况（所有函数f0∈F的最坏情况）意义下，此n-2/3速率不能被任何其他估计量改进。换句话说，n-2/3是[0,1]上Lipschitz函数的极小极大最优收敛速度。

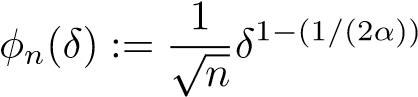
更一般地，假设我们现在假设f0在我们先前定义的平滑类Sα中。在这种情况下，与上述相同的论点导致了不平等：

（112个）

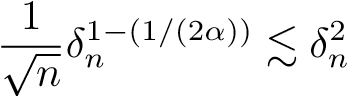
现在假设α>1/2在这种情况下，上面的积分是有限的，我们得到



这样我们就可以



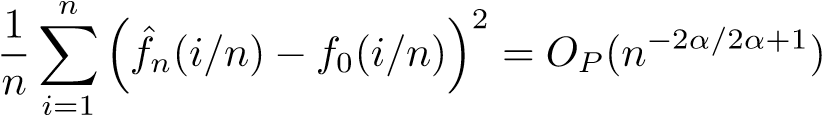
所以决定速率的临界不等式变成



它给予

δn&n-α/（2α+1）

使我们得出结论



结果表明，n-2α/（2α+1）确实是Sα中函数估计的极大极小最优速率。

现在假设α≤1/2。在这种情况下，（112）中的积分是无穷的，所以达德利的界在我们使用的形式中没有给出任何有用的东西。在这种情况下，可以使用达德利界的修正，其中积分的下限不是零，而是严格正的（这是问题6家庭作业3）。然而，d2（fˆn，f0）的结果速率将慢于n-2α/（2α+1）。对于α≤1/2，不知道Sα上的最小二乘估计是否是最小极大最优。

## 13.3回到速率定理

现在让我们回到定理13.1。为了证明这个定理，我们假设

|  |  |
| --- | --- |
| M（θ）–M（θ0）。-d2（θ，θ0）  对于所有θ∈Θ我们还假设 | （113个） |

)（114个）

所有u>0这里φn（u）是（0，~）上的函数，对于某些α<2，它满足φn（c x）≤cαφn（x）。

在这两个假设下，定理13.1断言对于满足φn（δn）的每个δn，d（θˆn，θ0）=OP（δn）。δn2。

在某些情况下，不可能保证（113）对所有θ∈Θ都成立。也不可能确保（114）适用于所有u>0。相反，通常可以保证一个正实数U（不依赖于n）的存在，使得（113）所有d（x，t，0）都服从u u，并且（114）对于所有u u u都成立。在这种情况下，在d（θˆn，θ0）以概率收敛到0的附加假设下，仍然可以断言d（θˆn，θ0）=OP（δn）。这是以下定理的内容（即Van der Vaart和Wellner[25]中的定理3.2.5）。

定理13.2。设u∗为严格正实数（不依赖于n），使得（113）对d（θ，θ∗）≤u∗的所有θ∈Θ都成立设φn为（0，∞）上的函数，满足某些α<2的条件：对于所有c>1和x>0，φn（cx）≤cαφn（x）假设（114）适用于所有0<u≤u∗。假设d（θˆn，θ0）=OP（1）。则对于满足φn（δn）的每个δn，d（θˆn，θ0）=OP（δn）。δn2。

证明。写

P n d（θˆn，θ0）>2Mδn o≤X P n 2j−1δn<d（θˆn，θ0）≤2jδno+P n2d（θˆn，θ0）>u\*o。

j>M:2jδn≤u\*

第一项的有界性与定理13.1的证明完全相同这给了

.

如果选择δn，则上述第一项收敛为零，即M→∞。另一方面，第二项通过假设d（θˆn，θ0）在概率上收敛到0，即n······这就是定理13.2的证明。

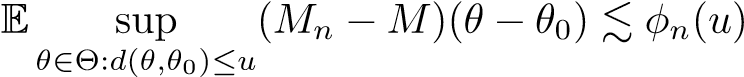
# 14讲座14

## 14.1主要速率定理概述

Tyth-n最大化了随机过程Mn（Th），而Th 0最大化了一个确定性过程M（Th）。假设d（·，·）是这样的

M（θ）–M（θ0）。-d2（θ，θ0）

对于所有θ∈Θ设πn（·）是满足温和条件的函数（存在α＜2），使得ω（Cx）＜c×ωn（x）



所有u>0当满足δn时，随机量d（θˆn，θ0）由δn控制

.

正式地说，我们有

Pnd（θˆn，θ0）>2Mδno.X 2j（α-2）。（115）

焦耳>米

这是这个定理的一些优点和缺点。让我们从优势开始：

1. 它很好地严格了启发式论证事实上，启发性的论点始于基本的不平等：

d2（θˆn，θ0）。M（θ0）－M（θˆn）≤（Mn－M）（θˆn－θ0）。

由此，可以很容易地导出δˆn：=d（θˆn，θ0）满足

.

由此可以合理地推测，d（θˆn，θ0）将由满足

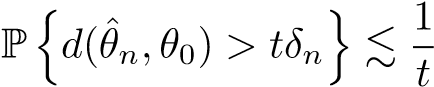
.

这个定理基本上证明了这一点。上面的不等式改为&。这也表明通过求解φn（δn）∼δn2得到的速率应该是d（θn，θ0）的正确速率。

1. 这是非常普遍的这个定理非常普遍，适用于各种问题我们已经看到了这方面的例子，我们将在不久的将来再举一些例子。
2. 很容易证明证明只有几行，不使用任何复杂的机器下面是速率定理的一些重要缺点。
3. 最重要的缺点是，尽管获得的速率通常是正确的，但偏差不相等（115）通常是相当弱的。为此，观察当α=1（例如）时，不等式（115）变为

Pnd（θˆn，θ0）>2Mδ2-M

从中可以看出



对所有人来说，这种不平等是相当弱的，因为它甚至不意味着

.

这种松散性的主要原因是在证明中使用了马尔可夫不等式：

.

这种不平等是相当宽松的。可以使用更复杂的参数（在更特殊的设置下）来代替它，这些参数给出了P{d（θˆn，θ0）>tδn}的改进边界。我们稍后将看到这样的改进结果的一些例子。

1. 通过定理计算速率需要一个

E sup（Mn-M）（θ-θ0）。

θ∈Θ：d（θ，θ0）≤u

虽然存在通用技术，但在特定情况下获得良好的界限仍然是相当困难的。

接下来，我们将应用速率定理来理解高斯序列模型中凸惩罚最小二乘估计量的损失行为在此之前，我们先介绍高斯序列模型。

## 14.2高斯序列模型

有限高斯序列模型是研究许多常用统计方法理论性能的重要模型。关于高斯序列模型下估计理论的综合处理，见Johnstone[11]。

假设我们观察到实值观测值Y1，…，Yn。在高斯序列模型下，我们将观测数据建模为

对于i=1，…，n

是i.i.d N（0,1）。换句话说，）和Y∼N（θ\*，In）。目标是根据损失函数下的数据Y1，…，Yn估算：

.

估计量θˆ的风险定义为

.

通常在未知参数上附加一些结构。以下是此类附加结构的一些标准示例：

1. 无结构：在这种情况下，没有可用的信息直觉上很清楚，我们不能做得比简单的估计量θˆi=Yi，i=1，…，n好得多，这个估计量的风险显然等于1。结果表明，inf sup R（θ，θ●）=1。

θ？θ∈Rn

这意味着简单估计量Y在Rn上是minimax最优的（或者等价地说，每一个其他估计量的θ∈Rn上的最坏情况风险至少为1）。

1. 固定线性子空间：这里假设已知线性子空间S的θ∗∈S。在这种情况下，最自然的估计量是Y到S的投影。这个估计量有风险k/n，其中k是S的维数。S上的极大极小风险等于k/n，即。，

.

1. 平滑度：假设某个函数f∗：[0,1]-[-1,1]的θ∗=（f∗（1/n），…，f∗（1））为1-

利普希茨。在上一类中，我们看到，如果我们考虑估计量θˆ=（fˆ（1/n），…，fˆ（1））其中fˆ是所有1-Lipschitz函数类中以1为界的最小二乘估计量，那么

.

也有可能证明上述界限在预期中也成立，即。，

R（θ，θˆ\*）。n-2/3。

结果表明，n-2/3实际上是所有向量（f（1/n），…，f（1））类上的极小极大风险，其中f:[0,1]-[-1,1]是1-Lipschitz。我们稍后会证明这一点。估计量θˆ实际上是非线性的然而，也可以使用基于Y1，…，Yn的局部平均的线性估计器来实现风险n-2/3。

1. 稀疏性：假设向量θ\*是稀疏的，因为只有少数项是非零的。更精确地假设θ∗k，其中Θk是Rn中所有向量中最多有k个非零项的类。假设k比n小（具体地说，假设k/n→0为n····）在这种情况下，存在满足满足的估计量

（116）

可以证明，在Θk上的极小极大风险也等于上面的右手边（我们稍后将证明这一点）在适当选择调谐参数的情况下，可将获得（116）的估计器θˆ取为套索阈值或软阈值这是一个M-估计，可以通过速率定理来研究。我们稍后再做。

1. 低阶结构：假设θ\*是d×d矩阵a\*的向量化，n=d2。假设A\*的秩至多为r，则基于核范数（奇异值之和）的惩罚估计可被证明为极大极小最优。

## 14.3高斯序列模型中的凸惩罚估计

在具有的高斯序列模型中，一类非常标准的估计器由以下形式的估计器给出：

（117个）

其中f:Rn→R是凸函数，λ>0是适当的调谐参数。当我们认为真实信号是稀疏的时，函数f将是θ的L1范数，在低阶假设下，函数f将是θ对应矩阵的核范数。也使用其他函数f（例如分段常数结构）。

估计量（117）显然是M-估计量，因此我们可以使用我们的一般速率定理来研究：

.

为此，我们首先使用，

.

因此，我们可以写

.

因此，我们可以应用速率定理

)还有。

注意M（Th）在Tax和条件下最大化。

M（θ）–M（θ\*）。-d2（θ，θ\*）

当

d（θ，θ∗）：=kθ−θ∗k。

因此，我们可以应用速率定理，它要求我们限制

.

我们现在用下面的方法来约束这个词因为f是凸的，对于f在θ\*处的每个次梯度s，我们有

f（θ）≥f（θ\*）+hs，θ−θ\*i。因此，

.

在上面的最后一个不等式中，我们使用了柯西-施瓦兹不等式。注意，对于θ\*处f的每个次梯度s，上述情况均成立。f atθ\*的所有次梯度的集合称为f atθ\*的次微分，并用f（θ）表示。因为上面的不等式链对于每个s∈∂f（θ\*）都是真的，所以我们可以取s∈∂f（θ\*）上的一个下确界，它允许我们推导

距离(

因此

.

θ∈Rn:kθ-θ\*k≤u

因此，速率定理意味着由δn控制，δn解



也就是说我们可以

δn：=Edist（。

由速率定理给出的精确不等式对于有界性来说有点弱（如前所述，这在很多情况下都会发生）然而，在这种情况下，可以证明（见Oymak和Hassibi[17]）

（118）

我将在作业中包括一个证明这种不平等的草图不等式（118）意味着如果∂f（θ\*）是Rn中的一个大集合，那么θλ，f的风险将很小注意，不等式（118）适用于每个凸函数f，因此它适用于各种情况。在下一节中，我们将演示它在研究稀疏信号估计中的风险方面的用途。

### 14.3.1不等式（118）在稀疏信号估计中的应用

我们现在将不等式（118）应用于f（θ）=|θ1 |+···+|θn |是θ的L1范数的情况。对于固定的λ，我们需要限制标准高斯向量的平方期望距离）第一步是获得∂f（θ\*）的特征。很容易看出∂f（θ\*）由所有向量（v1，…，vn）组成，使得

|  |  |
| --- | --- |
|   ={1}    vi={-1}  ∈[-1,1] | 如果θi∗>0，如果θi∗<0，如果θi∗=0。 |

因此，λ∂f（θ\*）由所有向量（v1，…，vn）组成，使得

|  |  |
| --- | --- |
|   = {λ}   | 如果θi\*0 |



如果

∈.[-λ，λ]如果。

因此，

距离

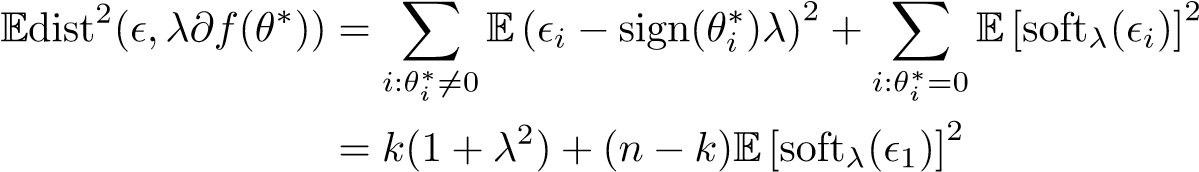
式中）是距离i最近的区间内的点&-，λ，λ。现在很容易看出

0如果

上面的这个函数有一个名字，它被称为i的软阈值，水平为λ。因此我们有

)=柔软。

我们因此获得：



其中k是θ\*中非零项的数目为了更进一步，我们需要计算下一节课我们要做什么。

# 15讲座15

在上节课中，我们研究了下列形式的估计器的性能：

在模型下这里f:Rn→R是凸函数。

对于这个估计，我们上一次注意到下面的不等式是正确的：

（119）

这个不等式是由Oymak和Hassibi[17]引起的；它有一个简单的证明，如下所示。

## 15.1不等式证明（119）

设g（θ）：=ky-θk2/2+λf（θ）表示θ∈Rn。θˆλ，f使g最小的陈述等价于0∈∂g（θˆλ，f）的陈述（这是微不足道的，因为θˆλ，f使g最小等价于g（θ）≥g（θˆλ，f）+D0，θ−θˆλ，fE）。现在很容易检查

∂g（θˆλ，f）=θˆλ，f−Y+λ∂f（θˆλ，f）n

所以我们有

0∈θˆλ，f-Y+λ∂f（θˆλ，f）

或同等

Y-θˆλ，f∈λ∂f（θˆλ，f）。

我们现在使用下面的引理15.1来推导，对于每一个s∈∂f（θ\*），我们都有DY-θˆλs，θˆλ，f-θ\*E≥0。λ，f

写作，我们获得

.

柯西-施瓦兹不等式可用于右上方，其将给出：

.

这对于每个s∈∂f（θ∗）都是正确的，因此我们可以对所有这些s取一个下确界并推导

距离(

这就完成了不等式的证明（119）。

引理15.1。设f:Rn→R是凸函数。那么对于每个θ1，θ2∈Rn和s1∈∂f（θ1），s2∈∂f（θ2），我们有

hθ1-θ2，s1-s2i≥0。（120）

证明。根据亚梯度的定义，我们有

f（θ1）≥f（θ2）+hs2，θ1-θ2i和f（θ2）≥f（θ1）+hs1，θ2-θ1i。

将这两个不等式相加得到（120）。

## 15.2（119）到f（x）=kxk1的应用

在上一节课中，我们把（119）应用到f（x）=kxk1=| x1 |+···+| xn |的情况，并认为

（121）

哪里

|  |  |
| --- | --- |
| y-λ  软的  y+λ | 如果y>λ，如果−λ≤y≤λ，如果y<--λ。 |

我们现在继续通过（低于φ（x）=（2π）-1/2e-x2/2是标准高斯密度）

.

我们现在应用上面第一个积分中的部分积分（u=x和dv=xφ（x）dx），以闭合形式计算第二个积分，并保留第三个积分



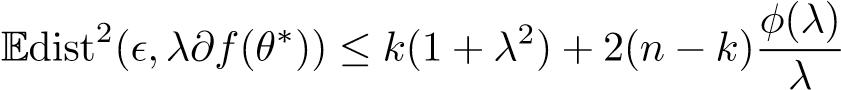
其中Φ（是高斯分布函数我们现在使用标准密尔比高斯界：

对于λ>0

获得

（122）

在（121）中使用这个，我们得到



这意味着，通过不平等（119），

（123）

如果我们现在做出选择

，（124）

我们得到了风险界

（1））（125）

当n·······提供k/n→0时。因此，具有惩罚（124）的套索达到风险2k对数（n/k）。

注意，如果θ\*中非零项的位置是已知的，那么通过Yi估计非零项的朴素估计量和通过0估计零项的实现风险等于k的朴素估计量。相对于（125），这意味着带调谐的套索（124）由于不知道非零位置而付出了2log（n/k）的代价。我们稍后将证明，每一个估计员都必须在极小极大意义上支付这个价格。这是√

凭直觉不难看出。例如，如果k=1且非零信号的幅度在某些c<2时为clogn，则数据中的噪声将淹没该信号，因此每个估计器极有可能丢失该信号并导致clogn损失。我们稍后会把这一点说清楚的。

为了使用调谐参数λ的选择（124），我们需要k的知识

λ=p2logn（126）

它不依赖于k。通过这个选择，界（123）给出

（1））（127）

这只比（125）稍差。如果k是常数阶，那么（125）和（127）之间没有太大的差别，但是对于k=n/（logn），有差别。

套索的界限（123）实际上可以通过更直接的方法得到，而不依赖于不等式（119）。这是因为估计器可以通过软阈值算子以闭合形式写入。这是下一步。

## 15.3软阈值

估计器

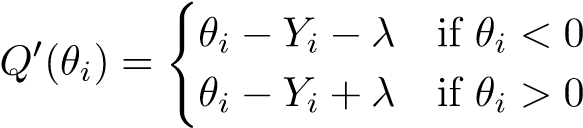
!

可以用封闭的形式书写。确实，首先观察到如果θˆλ=（θˆλ（1），…，θˆλ（n）），那么

θˆλ（i）=argmin。

我∈

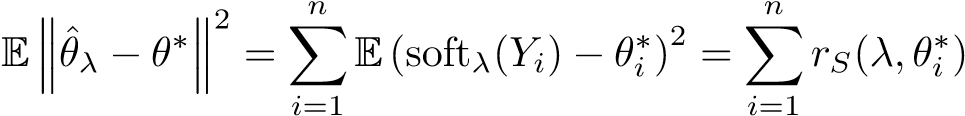
函数Q（θi）：=（Yi-θi）2/2+λ|θi |是凸的



因此，导数Q0（θi）与正斜率呈分段线性关系，除了θi=0处2λ的向上跳跃因此，Q0（θi）在一个点上正好有一个从负到正的符号变化，因此它必须是Q（θi）的最小值根据Yi的值，这个交叉点是正的，0

|  |  |
| --- | --- |
| 或者是否定的，然后我们可以检查 |  |
| i-λ  Yθˆλ（i）=0  Yi+λ  换句话说， | 如果Yi>λ如果−λ≤Yi≤λ如果Yi<-λ |
| θˆλ（i）=软λ（Yi） | 对于i=1，…，n。 |

利用这个，我们可以直接研究θˆλ的风险如下：



哪里

rS（λ，μ）：=E（软的λ（y）－μ）2和y∼N（μ，1）。

这个量rS（λ，μ）是在数据y∼N（μ，1）的单变量问题中软阈值估计器（在阈值λ处）的风险。我们可以显式地将其写成rS（λ，μ）：=E（软的λ（y）－μ）2

ZλZ--λZ∞

=μ2φ（x-μ）dx+（x+λ-μ）2φ（x-μ）dx+（x-λ-μ）2φ（x-μ）dx

−λ −∞ λ

Zλ－－－μZ－－λ－－μZ∞

=μ2φ（z）dz+（z+λ）2φ（z）dz+（z-λ）2φ（z）dz。

−λ−µ −∞ λ−µ

以下是rS（λ，μ）的一些基本性质：

1。函数μ7→rS（λ，μ）在[0，∞上增加通过计算，这是直观且易于检查的：

))（128个）

当μ>0时为正值。2。当μ=0时，我们有

对于所有的λ>0。

我们在（122）中证明了这一点。

三。当μ接近±∞时，风险rS（λ，μ）表现为1+λ2：

极限rS（λ，μ）=1+λ2。

µ→∞

我将把这个作为证明这一点的练习。直观地说，这是显而易见的，因为对于大的μ，最有可能是软的λ（y）=y-λ和E（y-λ-μ）2=1+λ2结合μ→7rs（λ，μ）在[0，∞上增加的事实，我们可以推断

suprS（λ，μ）=1+λ2。（129）

μ∈R

这些事实表明，与朴素估计量y相比，软阈值估计量在μ=0时的风险要小得多，而其最坏情况下的风险要大得多。因此，只有当μ为零或很小时才使用它才有意义。

利用上述观察，我们可以给出套索风险界（123）的另一种证明。的确，我们可以写

.

利用上述第二和第三个事实，我们得到

.

这证明了（123），我们已经在最后一小节中看到，它允许我们分别在调谐参数λ的选择（124）和（126）下推导速率结果（125）和（127）。注意在上面的范围中，我们使用

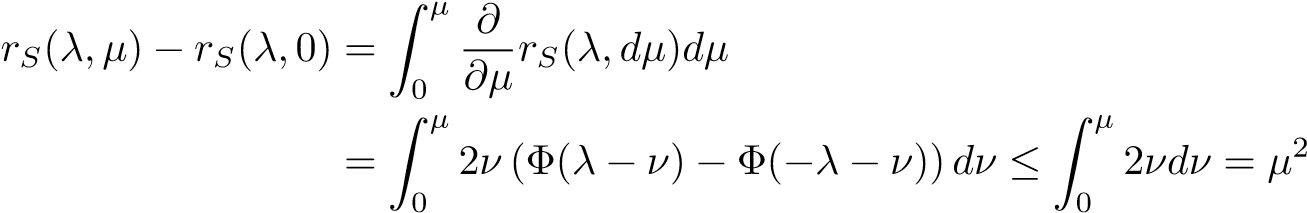
rS（λ，θi∗）≤suprS（λ，μ）（130）μ∈R

当θi\*6=0时如果提供更多关于θ\*的信息，那么这可能不是一个很好的界限。例如，在以下假设下，研究套索的表现也是很常见的：

kθ\*k1≤Cn（131）

对于一些Cn>0在这个假设下，所有的势都可以是非零的，所以使用（130）会给出很差的界限注意，即使在（131）项下，θ\*的所有项都可以是非零的，但它们必须满足绝对值的第j个最大项（用|θ\*表示）应以Cn/j为界的性质。这意味着θ\*的项必须满足某种衰减因此，假设（131）可以被认为是关于θ\*的某种形式的弱稀疏性假设。

现在让我们在假设（131）下研究θˆλ的风险如前所述，对于μ6=0，我们需要一些rS（λ，μ）的界，该界优于1+λ2为此，我们使用不等式（128）来编写



它给予

rS（λ，μ）≤rS（λ，0）+μ2。

结合这个和（129），我们推断

rS（λ，μ）≤rS（λ，0）+min（μ2,1+λ2）。

这意味着套索的风险是由

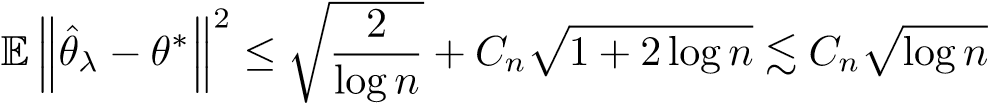
.

我们将在假设（131）下进一步约束这一点因为对于a，b≥0，min（a2，b2）≤ab，我们得到

.

√

选择λ=2logn将导致



这进一步给出

.

. 打√p

例如，如果Cn n，则上述界限变为（logn）/n。我们稍后将看到，在假设（131）下，该速率为极小极大。

# 16讲座16

在上一课中，我们研究了高斯序列模型Y∼Nn（θ\*，In）中的软阈值估计器的行为。我们注意到，这个估计量和LASSO一样，在θ\*具有精确稀疏性和弱稀疏性的情况下，研究了其风险的界我们从硬阈值开始，它有许多类似于软阈值估计器的性质。

## 16.1硬阈值估计器

硬阈值函数定义为：

硬λ（y）：=yI{| y |>λ}或硬λ（y）：=yI{| y |≥λ}

让我们把上面的第一个定义定为具体性。也就是说，当--λ≤y≤λ时，硬λ（y）等于0，否则等于y。当| y |≤λ时，两者均等于0，这与软λ（y）类似。然而，它的不同之处在于它在y中是不连续的，而软的λ（y）是连续的。

高斯序列模型Y∼N（θ∼in）中θ∼的硬阈值估计量θˆλH由下式给出

θˆλH=（硬λ（Y1），…，硬λ（Yn））。

很容易看出θˆλH是优化问题的解：

.

哪里。

注意这种优化与软阈值（或套索）的相似性和相异性，即

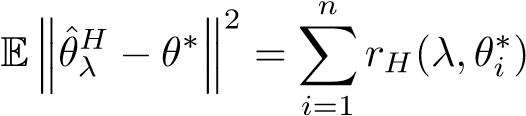
.

注意，θˆλH中的调谐参数是λ2，而θˆλS中的调谐参数是λ2。此外，θˆλH中的平方项之和没有系数1/2。

硬阈值估计量θˆλH在稀疏情况下与软阈值估计量具有非常相似的性质。例如，我们将在下面说明，当θ\*是k稀疏（即kθ\*k0=k）且k/n→0时，

（1））提供λ=p2log（n/k）。（132）

要看到这个，写下



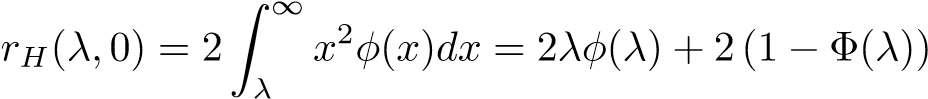
哪里

rH（λ，μ）：=E（硬的λ（y）－μ）2，y∼N（μ，1）。

因此

.

现在很容易看出



上面的积分是通过部分积分计算的标准米尔斯比率现在给出了

)对于所有的λ>0。

也可以显示（家庭作业）

suprH（λ，μ）≤1+λ2，对于所有λ>0μ∈R

与软阈值不同，函数μ7→rH（λ，μ）在μ>0时不是单调增加的。事实上，很容易看出，Lim-Youth-Fuffic Rh（Sm，γ）＝1，但在一定的有限（近α）处达到最大值。由于上述事实，因此

.

选择λ=p2log（n/k）很容易得出（132）也就是说

（1））对于λ=p2logn。

同样，θˆλH的工作原理与假设kθ\*k1≤Cn（这是家庭作业）下的软阈值类似。

## 16.2线性回归

我们现在来研究线性回归模型。在这里，我们再次观察到一个数据向量Y∈Rn，我们将其建模为

具有

对于某些已知的确定性n×p矩阵X，p×1参数θ\*是感兴趣的未知参数。给出θ\*的估计量θˆ，我们将对预测风险感兴趣：

（133）

根据具体情况，研究可能或多或少是自然的，但这通常需要对X有更多的假设，我们将不研究这个损失函数。

硬阈值估计和软阈值估计都可以直接推广到线性回归情形。硬阈值估计器的扩展为：

（134）

p√

注意，当X=I时，我们在具有λ=2logn（n/k）或λ=2logn的稀疏设置下使用该估计量。因此，在（134）中乘以kθk0的项将涉及logn因此，我们将此估计器称为BIC估计器（例如，用θˆλBIC表示）。

软阈值估计器的扩展直接给出了LASSO估计器：

θˆλ套索：=argmin.（135）

我们将从预测风险（133）的角度分析这两种估计。我们将主要关注精确的稀疏性设置，其中k:=kθ\*k0与p和n相比很小。从计算的角度来看，很容易看出（135）可以通过凸优化计算，而（134）根据X很难计算（在最坏的情况下，计算（134）是NP难）。

让我们从分析BIC估计量（134）开始。

## 16.3θˆλBIC的预测风险

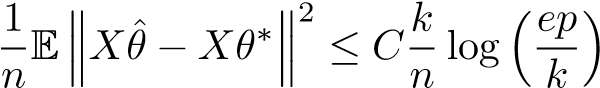
关键问题是：当θ\*为k-稀疏时，适当正则化的BIC估计是否具有以（k/n）（log（ep/k））的常数倍数为界的预测风险？我们将看到，这将是真实的，没有任何假设的设计矩阵X。

在回答这个问题之前，让我们首先分析一个简单的估计器，它由

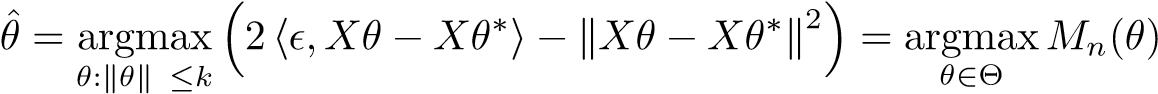
θˆ：=argmin。

θ：kθk0≤k

这个估计器简单地最小化了所有θ上最多有k个项的平方和。记住k是真向量θ\*的L0范数，因此kθ\*k0需要知道是否使用了这个估计量。我们将证明对于这个估计器，



对于一个普适常数C，让我们先用速率定理来解释为什么这是真的。我们可以把θˆ写成



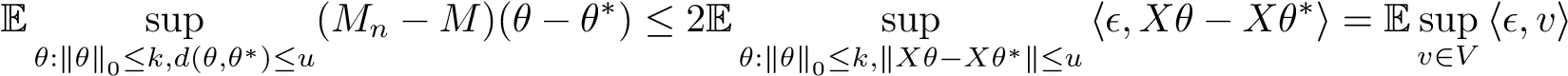
零

式中：={θ∈Rn:kθk0≤k}和

.

我们将对这个Mn和

M（θ）：=-k Xθ~-Xθ\*k2和d（θ，θ\*）：=kXθ~-Xθ\*k。因此，要通过速率定理获得速率，我们需要对



其中V由k V k≤u的所有向量V∈Rn组成，并且对于kθk0≤k的某个θ，V=X（θ−θ\*）满足。我们将使用达德利熵界来控制上面的期望上确界：

直径（V）/2

第页

（136个）

其中）和直径（V）是Rn上常用欧几里德度量中的填充数和直径。

现在注意，对于kθk0≤k的每个θ，我们有kθ−θ\*k0≤2k（因为kθ\*k0≤k）因此，我们可以

V⊆∪{V S:S⊆{1，…，p}，| S |≤2k}

（式中，| S |表示S的基数）式中，VS表示所有向量v∈Rn的集合，其中kvk≤u，v=Xβ表示在S上支撑的某个β（即{i:βi 6=0}⊆）因此，我们推断

.

因为，

VS:={v∈Rn:kvk≤u，v∈C（XS）}

其中X S是仅包含索引属于S的X列而形成的矩阵，C（XS）表示XS的列空间这意味着VS是一个在线性子空间中的球，其维数不超过2k，因此（根据先前关于在线性空间中填充球数的结果）

.

因此，

.

把这个插入（136），我们得到

.

把这个等同于u2，我们会得到

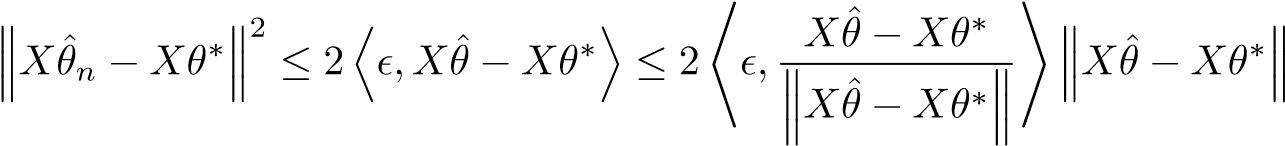
.

因此，这表明

（137）

和往常一样，我们的速率定理不足以给出这种期望控制。为了证明上述的界限，我们可以论证如下与M-估计量θˆ相对应的基本不等式为：M（θ∗）－M（θˆ）≤

（Mn-M）（θˆ-θ\*）与



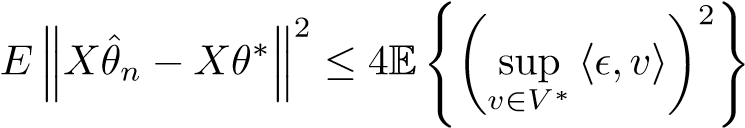
以便

.

哪里

V∗：={V:k V k≤1，V=Xβ，kβk0≤2k}。

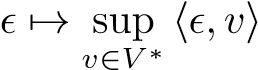
这给了



请注意，我们刚刚严格证明

（138）

从这里开始



是Lipschitz函数（Lipschitz常数为2），可以证明

（139个）

证明了（137）。

了解（138）如何表示（139）的一种方法是通过以下关于高斯随机向量的Lipschitz函数集中的重要结果。

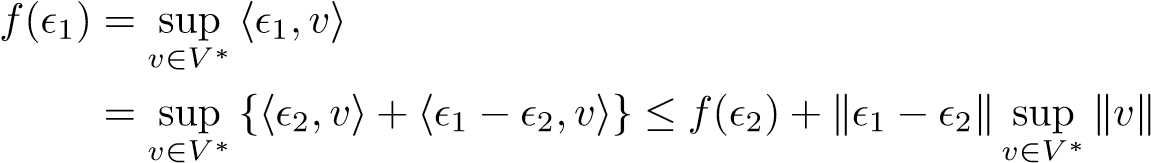
定理16.1。设f:R N→R是L-Lipschitz函数，即| f（x）－f（y）|≤Lkx－yk，设Z∼N（0，In）那么对于所有t≥0，

和（140）

定理16.1可用于从（138）导出（139）。的确，让

.

很容易看出f是1-Lipschitz。的确，



由于V\*中的每个向量最多有1个范数，我们得出f是1-Lipschitz的结论因此，在（140）之前，

.

所以

.

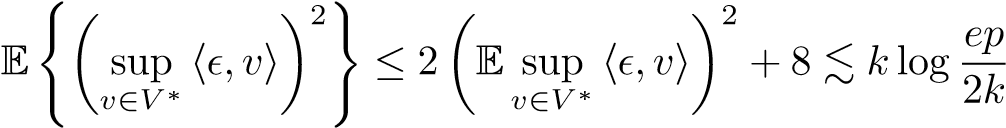
因此

.

这给了

.

结合（138），我们可以推断



证明了（139）。

# 17讲座17

## 17.1θˆλBIC的预测风险（续）

我们将研究估计器的预测风险：

（141）

在线性回归模型中，其中）回想一下，当X=In时，这是硬阈值估计器。θˆλBIC的以下结果为真：如果

λ：=c1plog（ep）

对于足够大的c1，则

)式中k：=kθ\*k0。（142）

我们不证明上面给出的期望界，但我们将使用率定理，它将暗示

损失将以高概率的pk-log（ep）的常数倍数为界。在我们开始讨论这个论点之前，让我们首先回顾一下，在上一节课中，我们证明了

对于θˆ（k）：=argmin kY−Xθk2。（143）

θ：kθk0≤k

注意，该估计量θˆ（k）可被视为θˆλBIC的约束形式在证明（143）的过程中，我们得出

式中k：=kθ\*k0。（144）

我们将在续集中使用这个不等式。现在让我们通过速率来分析

定理。首先注意到，THα- BIC最大化。

.

所以我们将应用速率定理，其中Mn（θ）和M（θ）取为

M（θ）：=——kXθ——Xθ\*k2。

同时d（θ，θ\*）：=k Xθ−Xθ\*k。要应用速率定理，我们需要

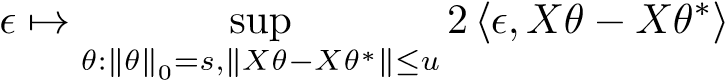
Γ：=E sup（Mn−M）（θ−θ\*）

θ：d（θ，θ\*）≤u

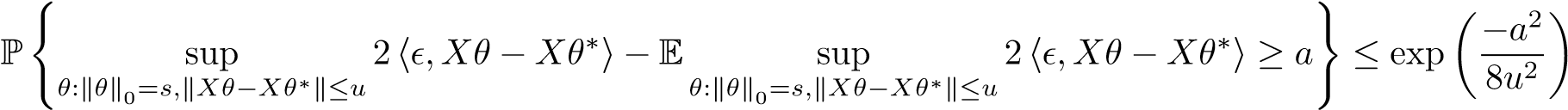
从上面开始，将结果绑定到u2为此，请注意



让我们分别用Γ1和Γ2表示上面的两个项要绑定Γ1，请注意，对于每1≤s≤p，映射



是2u Lipschitz所以根据上次的高斯浓度不等式，我们有



每大于0。因此，我们

Γ1≤铜对数（ep）。

对于边界Γ2，我们简单地使用不等式（144）来获得

.

因此，通过适当地增大常数C，我们得到

氮磷氧化物

Γ=Γ1+Γ2≤最大Cu（s+k）对数（ep）－λs+λk。

1≤s≤p

我们只是利用微积分将S函数最大化。设置成这样

,

我们得到

.

因此，根据速率定理，将由

.

选择λ=c1plog（ep），我们得到

.

如果c1=C2/2，那么我们得到



它给予

u=c1p2k对数（ep）。

因此，这证明了它将以高概率被pk-log（ep）控制此论点不产生可被证明是对上述论点的修改的经验界（142）。

以下是对（142）的两点评论：

1. 它不需要对设计矩阵X进行假设。
2. 如果允许λ依赖于k，则可以导出界（k/n）对数（ep/k）（即（142）中的对数（ep）可以被对数（ep/k）所代替）（见约翰斯通[11，第11章]）。

## 17.2θˆλ套索的预测风险

我们现在研究套索估计器：

θˆλ套索：=argmin.（145）

我们将证明预测误差的两个界

套索

.

n个

第一个界涉及k个k维K1（弱或近似稀疏性），第二个界涉及k个k次K0（强或精确稀疏性）。

### 17.2.1弱稀疏界

这里我们将证明θˆλ套索的预测风险的以下界。设X1，…，X p∈Rn表示n×p设计矩阵X的p列

λ=c1plog（ep）最大kXik（146）

1≤i≤p

对于足够大的c1，则

套索

（147）

对于常数C（c1），仅取决于c1。

请注意，如果取√X=In，则此绑定专门用于我们先前的软阈值绑定。通常，

选择设计矩阵的标度，使max1≤i≤p kXik≤n。在此假设下，调谐参数的选择变为

λ=c1pnlog（ep）

预测风险界（147）变成

套索

.

我们现在证明（147）首先指出THα-拉斯莫最大化了Mn（Th）在TH-RP上的位置。

.

也很容易看出，THEX最大化M（Th），TH\* RP RP。

.

因此，基本的不平等是

M（θ\*）-M（θˆλ套索）≤（Mn-M）（θˆλ套索-θ\*）

变成

套索

.

我们将对上述涉及右手边的条款

套索套索

它给予

套索

（148）

11个

三角形不等式：

拉索拉索

一

11个

现在给

拉索拉索

.

二

因此，如果选择λ，则

，（149）

然后我们会得到套索

.

当如（146）中选择调谐参数λ时，很容易看出（149）具有高概率，然后不等式（147）遵循上述不等式（严格化此参数并完成（147）的证明）。

### 17.2.2强稀疏界

我们现在将尝试用kθ\*k0=k来限制预测误差。在继续之前，让我们介绍以下符号。让我们表示指数i在1，…，p中的集合，其中=0。然后| S |=k。对于向量θ∈Rn，让θS∈Rn表示其第i项等于θi（如果i∈S）且等于0（如果i/∈S）的向量。我们类似地定义θSc。让我们从不等式（148）开始（其中，为了便于记法，我们为θˆλLASSO写θˆ）

.

我们将上面的右手边重写为

.

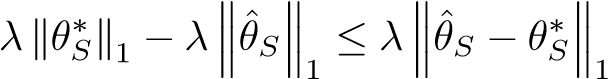
因为θS\*c=0，我们可以简化为

.

重新排列术语，我们推断

.

最后两项可由三角形不等式限定为



所以我们得到

（150）

假设现在我们假设正则化参数λ与之前一样满足（149）。然后我们可以取消这些项，并从上述不等式的两边得到

.

应用右边的Cauchy-Schwarz不等式，我们得到

（151）

进一步

.

这给出（从（151）开始）

.

取消一个，我们得到

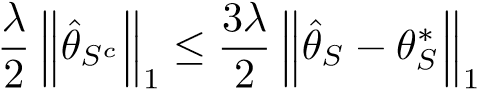
.

我们现在可以将λ取为（146），它确保（149）具有高概率（如果c1足够大），这意味着

套索

具有高概率（也可以导出经验界）。当X=In时，这是我们早期软阈值风险的一种弱形式，其边界是精确稀疏性。然而，这个界的一个主要问题是它依赖于λmin（XTX/n）。当p>n时，我们必然有λmin（XTX/n）=0，因此上面的界限是空泡的。但是，要用较小的量替换λmin（XTX/n），只需对上述参数稍作调整，将λ选择为2。

事实上，如果，那么我们首先可以从（150）中观察到不平等：



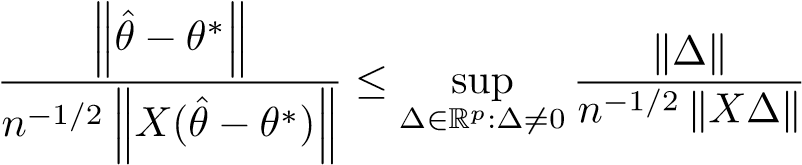
与

.

换言之，向量θˆ-θ\*属于圆锥体：

C：={∏∈Rp:k∏Sck1≤3k∏Sk1}。

通过这个观察，我们可以替换



通过

.

注意，上述两个界限之间的唯一区别是，在后一个界限中，上确界在∏∏C上，而在前一个界限中，上确界在∏∏Rp上。正如我们将在下一节课中看到的，这种改进给出了即使p>n也可能起作用的结果。

# 18讲座18

研究的主要课题是在精确稀疏的情况下，完成对套索在预测风险方面的性能的讨论让我们先回顾一下前几节课的观点，并提出感兴趣的主要问题。

## 18.1概述：精确稀疏线性回归

我们观察到一个数据向量Y∈Rn，我们将其建模为

和。

X是一个确定的n×p矩阵，θ\*是Rp中的一个向量。维数p可以大于样本量n。我们将使用精确稀疏性假设，其中θ\*支持在一个子集S⊂{1，…，p}上，S |=k，k假设小于p和n。估计量θ的预测误差定义为

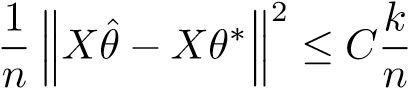
.

我们将研究套索的预测误差定义为

θˆλ套索：=argmin.（152）

在开始之前，让我们首先回顾以下意见：

1. 如果我们知道θ\*的支持度S，那么可以简单地通过X上Y的线性回归来估计θ\*（其中XS是通过删除S中不存在的列从X获得的矩阵）检查该Oracle估计器是否满足预测误差范围是基本的：

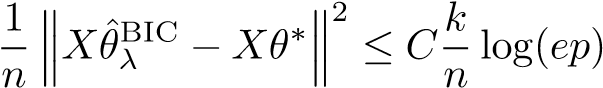


在预期和高概率中。需要注意的是，这里没有关于X矩阵的假设，上面的界限与缩放无关。如果我用一个常数乘以每一列来改变X，那么界限就不会改变。

1. 在上一个类中，我们已经看到由

（153个）

达到预测误差范围



如果选择调谐参数λ作为足够大c1的c1plog（ep），则具有高概率和期望值。因此，与上述Oracle估计器相比，BIC估计器仅支付p中对数的价格。然而，即使对于某些特殊设计矩阵X（例如，当X=i n或当X（i，j）=i{i≥j}对于1≤i，j≤n时）θˆλBIC的有效计算是可能的，在大多数情况下，它是计算困难的。

根据以上两个观察，最有趣的是看看θˆλ套索是否满足

套索

（154）

n个

事实上，与θˆλBIC不同，lasso估计器在计算上是可处理的，并且可以通过对相当大的n和p值进行凸优化而有效地获得。因此，有兴趣看看是否要为这种计算可处理性的预测风险性能（与θˆλBIC相比）付出任何代价。

我们将在下面看到，在X上的一些假设下（154）将是正确的。不幸的是，这些假设是非常限制性的，并且通常在实践中无法检查。在高层次上，这些假设可以理解为X在某种意义上表现为一个身份矩阵。注意，我们已经知道当X是恒等矩阵时（154）是真的（在这种情况下，θˆλLASSO是软阈值估计器）。

## 18.2精确稀疏下套索的预测误差

我们将完成上一节课开始的论点，在X上的一些假设下建立（154）。

作为套索基本不等式的一个简单结果，在上一类中，我们证明了以下不等式（其中，为了简单起见，我们为θˆλ套索编写θˆ）：

（155）

我们现在将调谐参数λ设为

（156）

可以对任何η>0进行以下分析，但通常取η=1。

对λ的选择（156）立即意味着2。在（155）的右边使用这个，我们得到

.

很容易简化为

.

注意，这个不等式同时意味着以下两个不等式：

和（157）

上面（157）中的第一个不等式可以重写为

（158）

因为θS\*c=0（注意S是θ\*的支撑）。这意味着θˆ−θ∗∈CS，其中

CS：={∏∈Rp:k∏Sck1≤3k∏Sk1}。

注意集合Cs是一个凸锥。

（157）中的第二个不等式显然更适用于证明θˆλ套索的预测风险界（154）。事实上，把两边除以n，我们得到

.

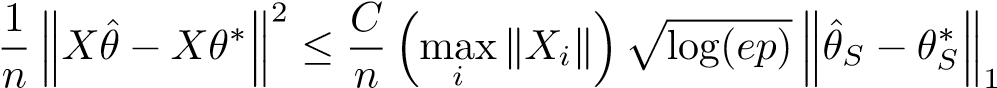
我们现在插入一个λ的显式值事实上，假设

（156）将满足

λ=c1最大值（ep）（159）

我

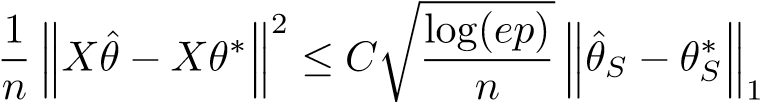
对于足够大的c1值。这里X1，…，Xp表示X的列。将这个值λ插入上面的边界，我们得到



对于依赖于c1的常数C我们现在将对X的列施加一个特定的比例。具体地说，我们假设√

kXik=n，每i=1，…，p.（160）

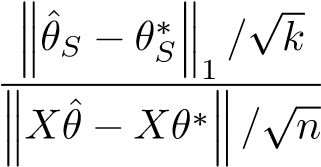
对于这个比例，上面的界限变成



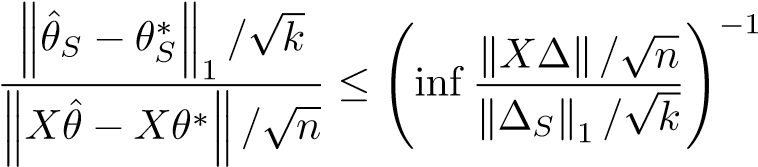
我们可以重写为

.

从这里可以很明显地看出，所需的绑定（154）支持



从上面以常数为界。这正是证明（154）的标准假设。为了使这种假设看起来不那么明显，通常将上述数量限定为：



其中，下确界可以接管Rp中任何包含θˆ-θ\*的集合。因为我们知道当λ满足（156）（见（158））时θˆ-θ∗∈CS，我们可以取∏∈CS上的下确界，它给出

.

数量

称为兼容性因子我们已经证明了

（161）

当取（159）中的λ时，具有高概率和高期望。如果我们现在假设

φ（S）≥φ0>0（162）

对于常数φ0，我们证明了（154）成立上述假设（162）被称为兼容性条件。这是设计矩阵X和集合S（θ\*的支持）的一个性质因为它依赖于未知的θ\*，所以在实践中无法验证。因此，假设infφ（S）≥φ0>0（163），取代它

S⊆{1，…，p}：| S |≤k

这在原则上是可验证的，因为它只依赖于设计矩阵X和未知θ\*的稀疏度k但是，不幸的是，即使k是已知的，验证（163）也需要遍历大小小于k的{1，…，p}的所有子集，这在计算上是困难的。

现在让我们快速讨论一下与相容性因子密切相关的受限特征值根据柯西-施瓦兹不等式，我们有

对于每一个△Rp。

因此

.

这个量γ（S）被称为X的限制特征值。很明显，因为γ（S）小于φ（S），如果用γ（S）代替φ（S），不等式（161）也成立，这意味着（154）在假设

0（164）

对于常数γ0。这称为限制特征值（RE）条件，它意味着（因此比）相容条件弱。不幸的是，检查（164）在计算上也是困难的，在实践中是没有希望的。

## 18.3检验RE和相容性条件的一个简单充分条件

以下引理给出了检验假设（163）和（164）的一个简单充分条件在这个引理中出现的X上的条件（165）有时被称为：“X是ρ-非相干的”引理18.1。假设X是一个n×p矩阵

对于所有1≤i≤n和（165）

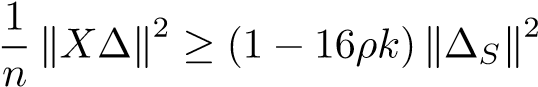
|  |  |
| --- | --- |
| 那么对于每一个具有基数k的S{1，…，p}，我们有 |  |
| 问  γ（S）≥（1−16ρk）+ | （166个） |

注意（165）中的第一个假设仅仅意味着X的每一列都被规范化为有范数

√

也只有当√ρ<1/（16k）时，结论（166）才是非平凡的例如，当ρ≤1/（32k）时，

然后（166）表示γ（S）≥1/2。引理的证明18.1。我们需要证明

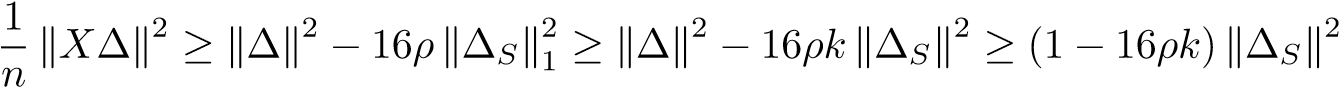


对于满足k∏Sck1≤3k∏Sk1的每∏∏Rp。固定这样的∏并写入

.

现在请注意，在假设k∏Sck1≤3k∏Sk1的情况下，k∏k1=k∏Sck1+k∏Sk1≤4k∏Sk1我们

因此获得



这就完成了证明。

## 18.4受限等距特性

受限等轴测属性（RIP）与RE和兼容性条件有关。定义如下。

定义18.2（RIP）设X是n×p矩阵。对于δ∈（0,1）和k≤p，我们认为X具有RIP（δ，k）性质

（167）

对于所有∏∈Rp

对于S⊆{1，…，p}，让XS表示X的n×| S |子矩阵，X的所有列X i都是通过删除i/∈S而形成的。用这个符号，很容易看出上述RIP的定义等同于下面的定义。

定义18.3（RIP的替代定义）设X是n×p矩阵。对于δ∈（0,1）和k≤p，我们认为X具有RIP（δ，k）性质

（1-δ）2≤λmin（XSTXS/n）≤λmax（XSTXS/n）≤（1+δ）2（168）

对于每一个子集S⊆{1，…，p}，| S |≤k，这里的λmin和λmax分别指最小和最大特征值。

由（168）可知，要使Xn×p满足RIP（δ，k）性质，n≥k是必要的。下面的结果表明，RIP性质意味着RE条件。

引理18.4假设X满足RIP（δ，k+m）。那么对于每一个S⊆{1，…，p}和| S |≤k，我们就有

（169）

从（169），可以很容易地推断出以下内容：

提供。

这意味着

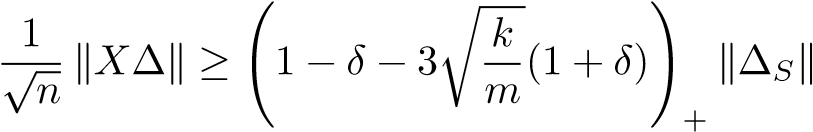
.

例如，通过取δ=1/4和u=1/5，我们得到

.

因此，利用引理18.4，我们可以推导出infS⊆{1，…，p}：| S |≤kγ（S）上的正下界，只要RIP（δ，m）保持一个小常数δ，m等于k的常数（取决于δ）倍数。

引理的证明18.4。我们需要证明



对于满足k∏Sck1≤3k∏Sk1的每∏∏Rp。固定这样一个向量设I1由Sc中与m个最大（绝对值）项对应的指数组成同样，设I2由（S∪I1）c中的指数组成，这些指数对应于∏的m个最大（绝对值）项继续这样定义分区

I1，…，Sc的Il

|I1 |=···＝| Il−1 |=m且| Il |≤m。

我们现在写信

一

沲∏∏S+沲I1+··∏∏Il=沲S∪I1+X沲Ii。

i=2

以便

.

我们现在使用X满足RIP（δ，k+m）的事实，它给出（注意| S∪I1 |≤k+m和| Ii |≤m

所有我）

.

现在，通过构造，在∏Ii中的每个条目的绝对值小于在∏Ii-1中的每个条目的绝对值这尤其意味着∏Ii中每个条目的绝对值小于

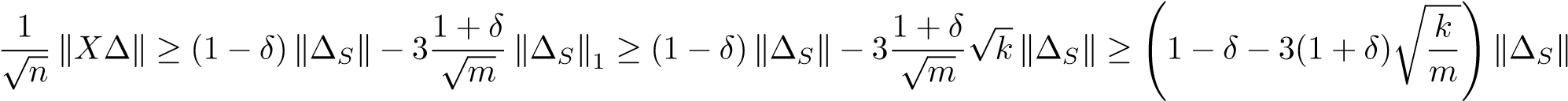
ΔIi-1项绝对值的平均值这给了

每2≤i≤l-1。

因此，我们得到

.

利用锥条件k∏Sck1≤3k∏Sk1，我们进一步推导出



这就完成了证明。

一个重要的事实是，RIP将被某些具有高概率的随机矩阵X所满足。这意味着（通过上面的引理）对于这样的矩阵X，RE条件将以高概率保持。最简单的例子是当X的项是i.i.d标准高斯。下面证明了这一点（证据来自Baraniuk等人。[1]；有关其他随机合集的扩展，请参见论文）。

定理18.5假设n×p矩阵X的项独立于n（0,1）且同分布。然后X满足RIP（δ，k），且概率至少为1-exp（-nδ2/64）

.

证明。很容易看出，因为X的项是i.i.d N（0,1），我们有

.

现在，随机变量满足标准浓度不等式（其证明留作练习）：

对于所有0≤t≤1。

这就立即得出，对于每一个∏∏Rp和δ∏（0,1）

.

因为1+δ/2≤（1+δ/2）2和1-δ/2≤（1-δ/2）2，我们也有

（170）

现在让集合1，…，m是集合的最大δ/4子集（通常欧几里得度量）

{∏：k∏k=1，k∏k0≤k}。

通过一个标准的体积参数，可以证明

.

根据并界，从（170）可以看出

)

对于所有1≤j≤M

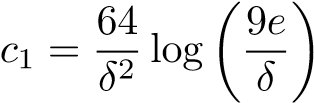
满足界

.

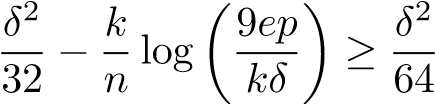
现在假设对于某个常数c1，n≥c1 k log（ep/k）那么

.

因此当



我们有



以便

.

为了完成证明，我们只需要证明

对于所有j=1，…，M（171）

意味着

对于k∏k=1且k∏k0≤k的所有∏（172）

证明这一含义的论点如下。设A为其中的最小数

对于k∏k=1且k∏k0≤k的所有∏（173）

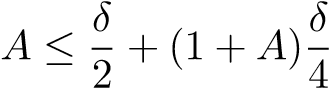
我们将证明A≤δ首先要注意A<∞是因为λmax（XTX/n）<∞。现在用k（1，4，m，m）的填充性质和构造，将k，k，k＝1和k，k，k，k，k，k，k，k，k，k，1，k，k，k，a，k，k，k，k，k，k，k，k，k，k，k，k，k，k，k，k，k，k，k，k，k，k，k，k，k，k，k，k，k，k，k，k，k，k，k，k，k，k，k，k，k，k，k，k，k，k，k，k，k，k，k，k，k，k，k，k，k，k，k，k，k，k，k，



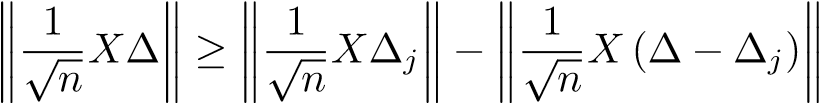
我们用（171）和（173）来得到最终的不等式，其中∏替换为∏-∏j。因为k∏jk=1，k∏-∏jk≤δ/4

.

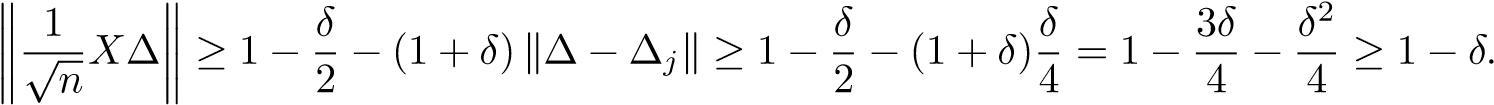
通过与（173）的比较，我们推断出（根据A的定义）



它给出了A≤δ。这证明了（172）中的上不等式为了证明更低的不平等，写下



用（171）和（173）加上A=δ（注意，我们可以选择∏j，这样k∏-∏jk0≤k），我们得到



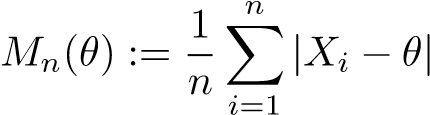
这证明了（172）中的下界，完成了定理的证明。

# 19讲座19

这门课的下一个主题是随机过程的收敛性我们研究这个问题的主要动机是证明M估计的极限分布结果。我们将从两个经典例子开始。

## 19.1样本中值的极限分布

假设X1，…，Xn是正态密度f的i.i.d观测值，平均θ0，方差1。实际上，很明显，下面的结果几乎不需要正规性，更一般地保持不变，但为了简单起见，让我们假设f是N（θ0,1）。设θˆn表示基于X1，…，Xn的样本中值，Xn定义为



在θ∈R上。也设M（θ）：=E | X1−θ|，并注意θ0唯一地最小化了M（θ）在θ∈R上。

我们在其中一个作业中看到θˆn在概率上收敛到θ0，即θˆn是θ0的一致估计量。我们的一般速率定理也可以直接应用于推导θˆn-θ0=OP（n-1/2），即θˆn到θ0的收敛速率为n-1/2。我们现在要讨论的问题是，找到的极限或渐近分布。找到这个极限分布有很多方法，但是我们应该遵循标准的经验过程方法，这种方法很容易推广到其他M-估计。这种方法也强调了研究随机过程收敛性的必要性。

我们求极限分布的方法是基于以下局部化、中心化和重标度的随机过程：

对于h∈R。

√这是一个随机过程，用h∈R表示。它的重要性质（容易看到）是hˆn：=

n（θˆn−θ0）最小化M∮n（h），h∈R i.e。，

.

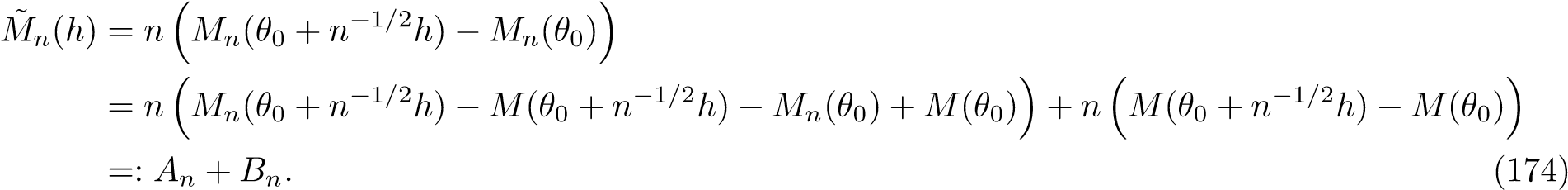
√

这表明以下方法可以找到n（θˆn-θ0）的极限分布。我们研究了过程M∮n（h），h∈R，并证明它在适当意义上收敛于某个极限过程M∮（h），h∈R如果这一过程的收敛性足够强，那么我们有希望认为

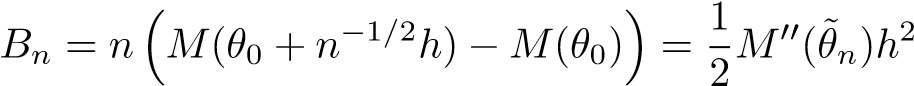
.

∈ ∈

实际上，对于每个固定的h∈R，理解M∮n（h）作为n→∞的行为并不太困难



现在让我们分别分析An和Bn。显然，Bn是一个确定性序列。为了理解这一点，我们将对θ0附近的M（θ0+n−1/2h）使用二阶泰勒解释注意M（θ）：=E | X1−θ|是一个光滑函数。还注意到，M0（Th 0）＝0，因为Th 0最大化m（Th），Th·r。



式中θ？n是θ0和θ0+n−1/2h之间的某个数。显然θ？n→θ0作为n················

如n·····。

现在我们来讨论平均零随机变量为了理解它，让我们首先计算它的方差：

!

其中，我忽略了X1在θ0和θ0+n-1/2h之间的贡献（对于大n不重要；请验证这一点）。这给了

瓦兰≈h2var（I{X1<θ0}-I{X1>θ0}）。

现在因为P{X1<θ0}=P{X1>θ0}（θ0是总体中值），很容易检查上面出现的I{X1<θ0}-I{X1>θ0}的方差是否等于1因此，我们获得了

var（An）→h2为n·····。

事实上可以证明

1）如n→∞。

为此，我们可以使用Lindeberg-Feller中心极限定理（下表）。

## 19.2林德伯格-费勒中心极限定理

定理19.1对于每个n，设Yn1，…，Ynkn为独立于kn的随机向量，对于每个i=1，…，kn，EkYnik2<∞。假设以下两个条件成立：

千牛

|  |  |
| --- | --- |
| 十  Cov（Yni）→∑as n→∞  i=1  其中Cov（Yni）表示随机向量Yni和 | （175个） |

作为n··········

|  |  |  |
| --- | --- | --- |
| 那么  千牛 |  |  |
| X L（Yni-EYni）→N（0，∑） | 如n·····。 | （177） |

i=1

有关此结果的证明，请参见Pollard[20，第181页]很容易看出，这个结果推广了通常的CLT实际上，通常的CLT表示对于i.i.d随机变量X1，X2具有

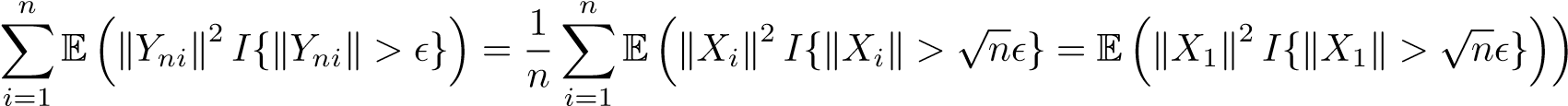
EXi=μ，EkXik2<∞和Cov（Xi）=∑，我们有

（？）作为n·····。

事实上，这可以通过应用定理19.1来证明

.

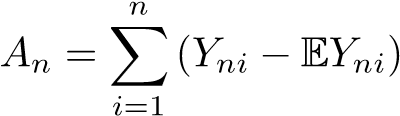
条件（175）是明显的，而对于（176）注意



它明显地通过支配收敛定理收敛到零（在假设EkX1k2<∞下）。

## 19.3回到样本中值的极限分布

从（174）中调用随机变量AnLindeberg-Feller CLT可用来证明An→L N（0，h2）。首先要注意



哪里

.

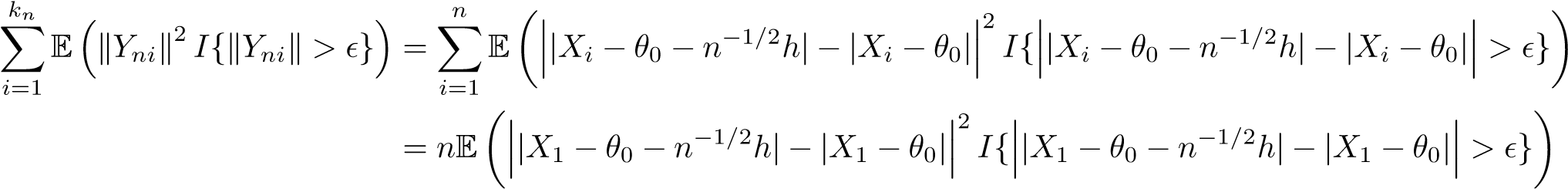
我们已经检查过了

n个

Xvar（Yni）=var（An）→h2为n····。

i=1

要检查（176），请注意



利用平凡不等式

,

我们得到

0为n→∞。

因此，定理19.1的条件成立，我们得到

An→L N（0，h2）为N→∞。

因此，如果我们定义

)对于h∈R

其中Z∼N（0,1），那么我们已经证明了

M∮n（h）→L M∮（h）作为n→∞对于每个h∈R。

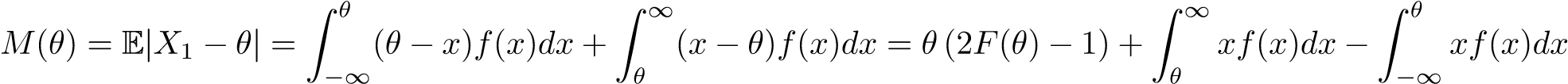
结果表明，对于每一个固定的h∈R，M~n收敛到M~n的过程比在分布上收敛到M~z的过程强，我们稍后将看到这一点这种较强的收敛性使我们可以推断

.

上面的argmax可以用封闭的形式写（注意，我们在h中有一个二次函数），这样我们就可以得到

.

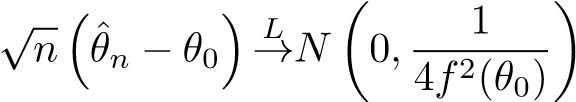
我们可以通过用f（θ0）来写M00（θ0）来稍微简化这一点。确实，先写



其中F是对应于F的cdf

M0（θ）=2θf（θ）+2（f（θ）–1）–2θf（θ）=2（f（θ）–1）

M00（θ）=2f（θ）因此我们有

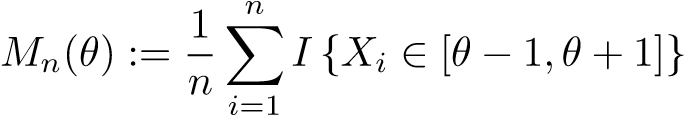


为了使这一论点更为严格，我们必须证明随机过程M~n在足够强的意义上收敛到M~n，这样它们的参数也收敛。

## 19.4采样模式的极限分布

前一节给出的求样本中值极限分布的一般方法非常广泛，也可用于其他M估计为了说明这一点，让我们应用它来确定样本模型的极限分布。设X1，…，Xn de i.i.d从正态密度f观测，平均θ0，方差1同样，结果不需要正态性（如果f是以θ0为中心的Cauchy密度，则也成立），但为了简单起见，我们假设f是N（θ0,1）。

设Stang-n表示任何样本模式，它被定义为任何最大化器。



在θ∈R上。也让

（178）

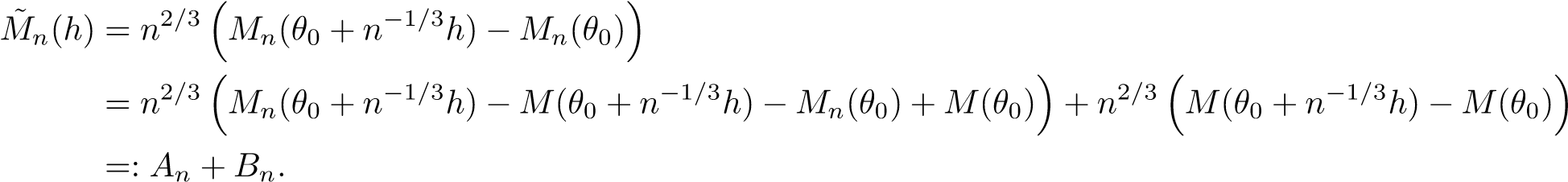
我们以前已经看到θˆn是θ0的一致估计，并且

θˆn-θ0=OP（n-1/3）。

我们现在将试探性地确定n1/3的极限分布（θˆn-θ0）。后面将给出严格化参数所需的必要过程收敛结果。为了研究hˆn：=n1/3（θˆn−θ0），自然要定义过程

对于h∈R

并注意到H\* n最大化了m H n（H），让我们试着了解每一个固定H r m m n（h）为n～ω的行为。



期望项Bn的处理与中值情形完全相同，在θ0处通过光滑函数M（θ0+n−1/3h）的二阶泰勒展开（注意，M0（θ0）=0）来获得

如n·····（179）

对于随机项An，让我们像以前一样，从计算它的方差开始：

.

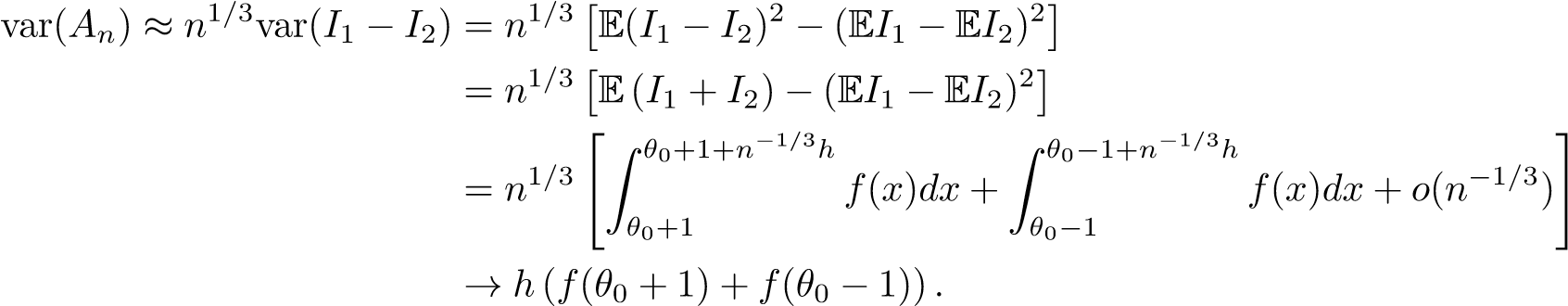
如果h>0，n很大，很容易看出

var（An）≈n1/3var（I1−I2）

哪里

I1：=I{θ0+1≤X1≤θ0+1+n−1/3h}和I2：=I{θ0−1≤X1<θ0+n−1/3h−1}。

因此



如果h<0，则必须将h替换为上面的-h因此，对于每一个h∈R，我们有

var（An）→| h |（f（θ0+1）+f（θ0-1））为n→∞。

事实上，通过Lindeberg Feller CLT（在中值的情况下），可以证明（这是作为家庭作业留下的）

An→L N（0，| h |{f（θ0+1）+f（θ0-1）}）为N→∞。

结合这个和（179），我们得到

.

假设{B h，h∈R}是一个从0开始的双边布朗运动，即B0=0，{Bh，h≥0}是一个标准布朗运动，{B-h，h≥0{是另一个独立于{Bh，h≥0}的标准布朗运动。注意Bh∼N（0，| h |）对于每个h∈R

,

我们有

M∮n（h）→L M∮（h）作为n→∞。

我们的上述论点可以得到加强，以证明（M∏n（h1），…，M∏n（hk））对于每个不动点h1，…，hk∈R在分布上收敛到（M∏（h1），…，M∏（hk））。这是Lindeberg-Feller CLT的结果，作为练习而离开。我们稍后将看到，M∮n在一个比仅对每个固定k≥1和点h1，…，hk∈R强得多的意义上收敛到M∮，这种强收敛使我们可以得出结论

.

∈ ∈

由于（178），很容易看出M00（θ0）=f0（θ0+1）-f0（θ0-1）

（非严格地）的极限分布由

阿格马克斯。h∈R

注意f0（θ0−1）–f0（θ0+1）>0。上述随机变量的分布与切尔诺夫分布有关（见。

在下一节课中，我们将看到更多的过程收敛的例子，并更严格地理解和形式化这个概念。

# 20讲座20

在这堂课中，我们将开始正式研究随机过程的收敛理论为了理解一般的观点，看看统一经验过程的特例是有帮助的。

## 20.1统一经验过程

假设X1，…，Xn是均匀分布在[0,1]上的i.i.d随机变量。对于每个t∈[0,1]，让

（180）

随机变量集合{Un（t）：0≤t≤1}表示由[0,1]索引的随机过程。这个过程的每一个实现（对应于X1，…，Xn的每一个实现）都是[0,1]上的一个有界函数（并且在（0,1）中的每一点都具有左极限的右连续函数）。注意{Un（t）：t∈[0,1]}的实现不是连续函数。

根据通常的多元中心极限定理，对于每k≥1和t1，…，tk∈[0,1]，

（Un（t1），…，Un（tk））→Lnk（0，∑）为n·······

式中，∑由∑（i，j）：=min（ti，tj）－titj给出。

布朗桥是一个随机过程{U（t）：0≤t≤1}是一个由[0,1]索引的随机过程，由以下两个性质定义：

1. U（t），t∈[0,1]的每一个实现在U（0）=U（1）=0的[0,1]上都是连续函数。
2. 对于每个固定的t1，…，tk∈[0,1]，随机向量（U（t1），…，U（tk））具有多变量正态分布，平均向量为0，协方差矩阵为∑（i，j）：=min（ti，tj）－titj。

基于以上，很明显，对于k≥1和t1，…，tk∈[0,1]，我们有

（Un（t1），…，Un（tk））→L（U（t1），…，U（tk））为n→∞（181）

这是通常CLT的结果。根据分布收敛的定义，语句（181）意味着

Eg（U n（t1），…，Un（tk））→Eg（U（t1），…，U（tk））为n→∞（182）

对于每个有界连续函数g:Rk→R，（182）对于所有有界连续函数g:Rk→R，当且仅当（182）对于所有有界Lipschitz函数g:Rk→R，也是正确的。有关此等价性的证明，请参见，例如Pollard[20]。

因此，结果（181）可以用以下方式重新表述：随机过程的任何有界连续函数的期望（仅通过其在[0,1]中的有限点集上的值依赖于Un）收敛到布朗桥U的相应期望。例如，这意味着

Eh（最大值| U n（ti）|）→Eh（最大值| U（ti）|）为n→∞。（183）

1≤i≤k 1≤i≤k

对于每个有界连续函数h:R→R，它等价于

最大值| Un（ti）|→L最大值| U（ti）|为n→∞。

1≤i≤k 1≤i≤k

虽然这是有用的，但人们通常需要处理依赖于整个过程的Un函数，而不仅仅是它在有限点集上的值。例如，有兴趣（例如，通过Kolmogorov-Smirnov检验检验拟合优度的统计应用）询问

sup | U n（t）|→L sup | U（t）| as n→∞

0≤t≤1 0≤t≤1

相当于

Eh（sup | Un（t）|）→Eh（sup | U（t）|）（184）

0≤t≤1 0≤t≤1

对于所有有界连续函数h:R→R。显然，这些函数依赖于Un through在[0,1]上的所有值，而不仅仅是在有限多个值上这里有一个合理的策略来证明（184）。取点0=t0<t1<·····································通过Un（t）的正确连续性，似乎可以选择一个足够大的网格，以便

其中F：={t0，t1，…，tk}。（185）

另外，由于布朗桥{U（t），0≤t≤1}具有连续的样本路径，这似乎是合理的

.

到183年，

如n·····。

把上述三个方程放在一起，似乎可以推断出（184）。对于这种策略，重要的是，近似（185）对于所有大N都在n中保持“一致”，事实上，如果网格F必须随着N的变化而显著改变以保持近似，那么该策略不能工作。从这个讨论看来，从一个随机过程的有限维收敛到无限维收敛的移动应该是可能的，在一个假设中保证在所有大的n值上均匀地逼近该过程。这就是所谓的渐近等度连续性的假设（也称为随机等连续性）在下面所述的抽象结果中表示。

## 20.2抽象结果

对于下一个结果，我们使用以下符号。`∞[0,1]表示[0,1]上所有有界函数的类（即sup0≤t≤1 | f（t）|<∞的所有函数f）我们将`∞[0,1]视为以下度量下的度量空间：

（186）

当我们提到一个连续函数h：`∞[0,1]→R时，我们的意思是h在上面定义的度量中是连续的。

此外，C[0,1]表示[0,1]上所有连续函数的类。

定理20.1。假设对于每个n≥1，{Xn（t），t∈[0,1]}是一个随机过程，其实现是`∞[0,1]中的函数假设{Xt，t∈[0,1]}是另一个随机过程，其实现是C[0,1]中的函数假设以下两个条件成立：

1. 对于k≥1和t1，…，tk∈[0,1]，

（X n（t1），…，Xn（tk））→L（X（t1），…，X（tk））作为n·····（187）

这种假设称为有限维收敛。

1. 对于每一个δ＞0，都存在一个整数和一个有限网格0＝t0<t1<…·tk＝1＝tk＝1这样

为所有人。（188）

这种假设称为随机等连续性或渐近等连续性。

那么对于每个有界连续函数h：`∞[0,1]→R，我们有

Eh（X n）→Eh（X）作为n······（189）

备注20.1我们可以简化假设（188），通过将其转换为：对于每一个图0，存在一个整数n和一个有限的网格0＝1，1。

所有n≥nη（190）

很容易看出（188）和（190）是等价的。的确，（188）显然意味着（190）。同时，（190）表示暗示（188）。

随机等度连续性假设（188）基本上说，Xn（t），0±t，1可以由Xn（t）、t{{t0，t1，…，tk}与所有大N邻接，即只要n大，近似在n中保持一致。

定理20.1的证明。我们将证明（189）对于所有有界且Lipschitz的函数h：`∞[0,1]→R。结果表明，如果（189）适用于所有有界Lipschitz h，那么它也适用于所有有界连续h，但我们将跳过这个证明。

因此，假设h在绝对值上有B的界，并且是L-Lipschitz，即。，

sup | h（u）|≤B和| h（u）－h（v）|≤Lku－vk∞=L sup | u（t）－v（t）|（191）

u∈`∞[0,1]0≤t≤1

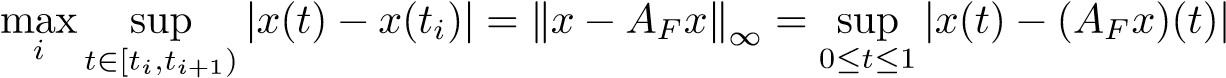
Fix>0，调用随机等连续性假设，取一个整数和一个网格0=t0<t1<····<tk-1<tk=1，这样（188）就成立了设F：={t0，t1，…，tk}并设AF：`∞[0,1]？`∞[0,1]定义为

k-1型

（AFx）（t）：=x x（t I）I{ti≤t<ti+1}，x∈`∞[0,1]和t∈[0,1]

i=0

和（AFx）（1）=x（1）。很容易检查每个x∈`∞[0,1]，



因此

为所有人。

我们现在改变网格F，使得上述不等式也适用于过程X（t），0≤t≤1（注意X有连续的采样路径）为此，让S={s0，s1，s2，…}成为

[0,1]其中s0=0，s1=1。那么，对于每个x∈C[0,1]，

.

因为X有连续的样本路径，所以我们有

0几乎可以确定为m····。

因此对于所有的大m，我们有



取这么大的m“合并”两个网格{s0，s1，…，sm}和{t0，t1，…，tk}。对于生成的合并网格，比如说T，我们有

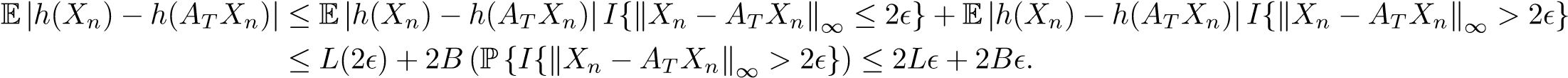
和

注意，在概率内变为2（这是因为，当我们使用| Xn（t）–Xn（sj）|≤| Xn（t）–Xn（ti）|+| Xn（ti）–Xn（sj）|时，对X也同样如此）。

现在对于满足函数h（191），我们可以编写

|E h（Xn）－Eh（X）|≤E | h（Xn）－h（ATXn）|+E | h（X）－h（ATX）|+| Eh（ATXn）－Eh（ATX）|（192）

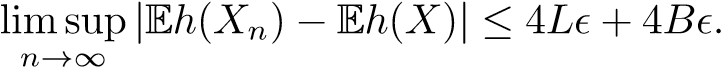
对于上面右边的第一个学期，我们认为



同样的上界也适用于（192）中的第二项对于（192）中的第三项，使用有限维收敛假设（注意网格T不依赖于n）来声明

|Eh（ATX n）－Eh（ATX）|→0作为n→∞。

我们已经证明了



因为>0是任意的，我们已经证明了（189）。

## 20.3回到统一的经验过程

回顾统一经验过程Un in（180）和Brownian桥U。然后，正如我们所看到的，多元CLT意味着有限维收敛我们在这里也要证明，联合国也满足随机等连续性为此，让我们首先注意到随机等连续性来自

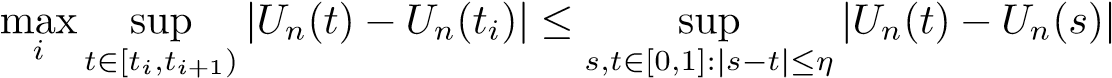
E sup | Un（s）–Un（t）|→0作为n→∞和η→0。（193）

s，t∈[0,1]：| s−t |≤

事实上，如果（193）成立，则给出＞0和δ＞0，则存在着ε＞0和一个整数，使得我们每个人都有。

为所有人。

设0=t0<t1<·····<tk=1为[0,1]中间距为η的均匀网格。那么很明显

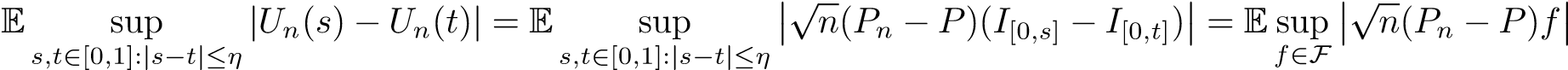


因此，通过马尔可夫不等式，我们得到

为所有人

它给出了随机的等连续性。

我们现在将核实（193）这将通过我们在第9课中所研究的经验过程的期望上确界来完成设Pn表示X1，…，Xn的经验测度，设P表示[0,1]上的一致测度。那么



哪里

.

然后我们使用下面的不等式（在第9课中证明）

.

类F有一个平凡的包络，所以我们得到（这里有一个错误）。我们无法推断

从supf∈F pPnf2≤1那；不等式实际上会走另一条路，因为-包装数随着减少而增加；正确的论点见下节课我把这个错误的论点留在这里，所以我们知道它不起作用。）

.

将F的VC子图维数绑定，并用包装数与VC子图维数的关系表示



它给予

!

√ √ √

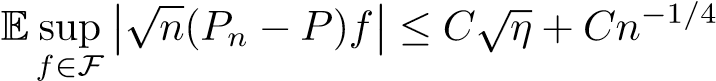
其中我们使用了平凡不等式a+b≤a+b。对于每一个f=I[0，s]-I[0，t]∈f，我们有



所以我们得到

.

现在证明{f2:f∈f}是一个VC维数最多为2的布尔类，这样



作为η→0和n··············。

Un到U的有限维收敛性以及随机等连续性意味着（通过定理20.1）

Eh（U n）→Eh（U）为n····（194）

对于每个有界连续函数h：`∞[0,1]→R。

## 20.4可测量性问题

断言（194）存在一些可测量性问题。事实证明，对于每个有界连续函数h：`∞[0,1]→R，h（Un）是不可测的。下面让我们举例说明当n=1时的情况（参数也可以扩展为n的更高值；参见Pollard[18，问题1，第86页]）。

首先注意，随机过程U1仅依赖于X1。实际上，`∞[0,1]中的函数U1正好等于

U1=I[X1,1]——内径

其中Id是函数Id（t）=t。如果可能的话，我们假设h（U1）=h（I[X1,1]-Id）对于每一个有界且连续的h：`∞[0,1]→R是可测的从而产生矛盾很容易看出，上述断言相当于

h（I[X1,1]）对于每一个有界且连续的h：`∞[0,1]→R是可测的。（195）

通过连续函数和闭集之间的标准连接（例如，见Billingsley[2，第1章]），可以证明（195）意味着

对每个开集O⊆`∞[0,1]都是可测的。

这里，`∞[0,1]中的开集是相对于度量（186）定义的。现在可以验证，对于每个子集A⊆[0,1]，以下是正确的：

式中O：=∪s∈AB（I[s，1]，1/2）

其中B（I[s，1]，1/2）是指以I[s，1]为中心、半径为1/2的`∞[0,1]中的开球（相对于度量（186））因为开放集的任意并集是开放的，所以上面定义的集合O是开放的因此，我们得到了，作为（195）的结果，I{X1∈a}对于[0,1]的每一个子集a都是可测的因为X1是根据[0,1]上的均匀分布分布分布的，这意味着可以在[0,1]的所有子集的集合上定义概率测度，使得每个区间的概率等于区间的长度在选择的公理下，这是不可能发生的。

由于上述矛盾，h（U1）不可能对所有有界连续函数h：`∞[0,1]→R都是可测的，也可以证明h（Un）对每n≥1都是相同的。因此，这意味着我们不能真正谈论Eh（Un）。作为一个固定办法，我们在这里考虑外部期望（用E\*h（Un）表示）。幸运的是，这个理论通过了这个修正更多细节将在下节课中提供。

# 21讲座21

在最后一类中，我们证明了下列过程收敛定理。

定理21.1。假设对于每个n≥1，{Xn（t），t∈[0,1]}是一个随机过程，其实现是`∞[0,1]中的函数假设{Xt，t∈[0,1]}是另一个随机过程，其实现是C[0,1]中的函数假设以下两个条件成立：

1. 对于k≥1和t1，…，tk∈[0,1]，

（X n（t1），…，Xn（tk））→L（X（t1），…，X（tk））作为n·····（196）

这种假设称为有限维收敛。

1. 对于每一个δ＞0，都存在一个整数和一个有限网格0＝t0<t1<…·tk＝1＝tk＝1这样

为所有人。（197）

这种假设称为随机等连续性或渐近等连续性。

然后对于每一个有界连续函数h：`∞[0,1]→R（这里我们把`∞[0,1]看作是一致度量下的度量空间），我们得到

Eh（X n）→Eh（X）作为n·····（198）

这里有一些关于这个定理的评论。

√

1. 如果Xn的采样路径有跳跃（例如Xn（t）=n（Fn（t）－t）），则（如前一类中所述）h（Xn）不必对每个有界连续函数h：`∞[0,1]→R都是可测的。在这种情况下，Eh（Xn）可能没有正确的定义。可以通过将E h（Xn）替换为其外部期望E\*h（Xn）来解决这个问题

E\*H（Xn）：＝INF{eB:B是可测量的，B超过H（Xn），EB存在}。

如果用E\*h（Xn）代替Eh（Xn），则定理的结果是真的在我们的治疗中，我们将忽略这些可测量的问题。有关详细分析，请参见加藤[12]。

1. 我们取（198）作为`∞[0,1]中随机过程序列{Xn}到X收敛性的定义。我们把它写成X n→L X作为n·····。与在欧氏空间上分布收敛的情况一样，以下是Xn→L X的等价定义：
   1. 每个有界Lipschitz函数h的E∗h（Xn）→Eh（X）：`∞[0,1]→R。
   2. 对于`∞[0,1]中的每个开集G，我们都有liminfn→∞P∗{Xn∈G}≥P{X∈G}。
   3. 对于`∞[0,1]中的每一个闭集F，我们有limsupn→∞P∗{Xn∈F}≤P{X∈F}。

过程收敛的一个重要结果是连续映射定理：假设Xn→L X和g：`∞[0,1]→Rk是连续的，则g（Xn）→lg（X）这是过程收敛定义的一个微不足道的结果。

1. 有限维收敛通常是Lindeberg-Feller中心极限定理的结果四。随机等连续性隐含于

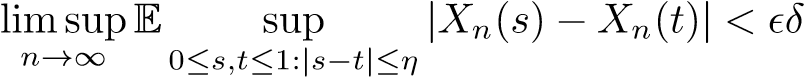
=0，每δ>0

进一步暗示

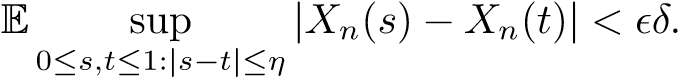
四肢支撑| Xn（s）–Xn（t）|=0.（199）

η≤0n→∞0≤s，t≤1：| s−t |≤

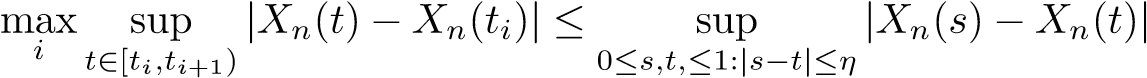
例如，看到（199）意味着（197），注意每一个> 0和三角形> 0，有0个这样的



这进一步意味着一个整数的存在，对于所有人来说，



设0=t0<t1<·····<tk=1为[0,1]中的均匀网格，间距η为



所以通过马尔可夫不等式，我们得到

为所有人

这证明了（197）。

在定理21.1中，区间[0,1]可以被R的任何其他紧子区间[a，b]所代替。事实上，它可以被任何抽象集T所代替。在这种情况下，我们将跳过下列定理的证明（证明可以在Kato[12，定理11]中找到）。

设`∞（T）表示T上所有有界函数的空间，将其视为具有度量（f1，f2）7→supt∈T | f1（T）－f2（T）|的度量空间。

定理21.2。对于每个n≥1，设Xn（t），t∈t是一个在`∞（t）中实现的随机过程。假设

1. 对于k≥1和t1，…，tk∈T，随机向量序列（Xn（t1），…，Xn（tk））在分布上收敛到一定的极限。
2. t上存在一个半度量D，它（t，d）是完全有界的，并且

四肢支撑| Xn（s）–Xn（t）|=0。

η≤0n→∞0≤s，t≤1:d（s，t）≤

然后，存在一个随机过程x（t），t't，它的实现是T（关于度量D）上的连续函数，使得在每个有界连续函数H：‘0’（t）~r中，“n”（t）中的xn×l x，或等价地，E＊H（xn）-ε（x），作为n＝~（0）。

注意定理21.2不是以极限对象X开始的，而是用连续样本路径（相对于度量n满足Xn满足随机等度连续性）来判定过程x（t）、t（t）的存在性。注意，极限对象必然满足（Xn（t1），…，Xn（T k））在分布上收敛到（X（t1），…，X（tk））的性质，对于每k≥1和t1，…，tk∈T。

## 21.1个极大不等式与随机等度连续性

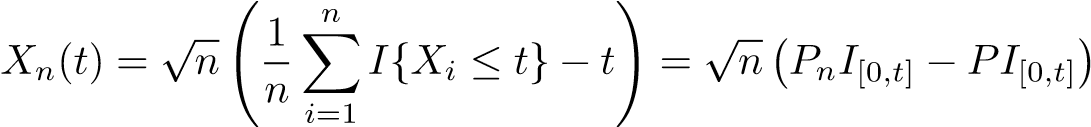
如我们所见，过程收敛的关键条件是随机等连续性。为了证明这一点，我们显然需要

E sup | Xn（s）–Xn（t）|。

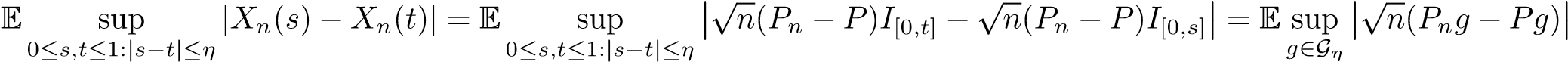
0≤s，t≤1:d（s，t）≤

在大多数应用中，Xn将与经验过程相关，因此我们得到了经验过程的期望上确界。

让我们先用一个例子来说明总体思路。假设指数集T=[0,1]和Xn是统一的经验过程，即。，



式中，Pn是与观测值X1，…，Xn相对应的经验测量值，其i.i.d在[0,1]上是一致的（P也是[0,1]的分布）在这里我们将试图证明Xn满足随机等连续性为此，首先要注意



我们使用以下符号：

对于η∈[0,1]。

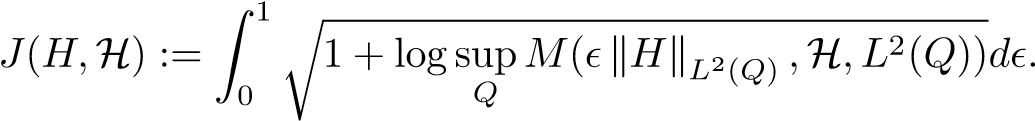
我们可以利用经验过程的期望上确界来控制

.

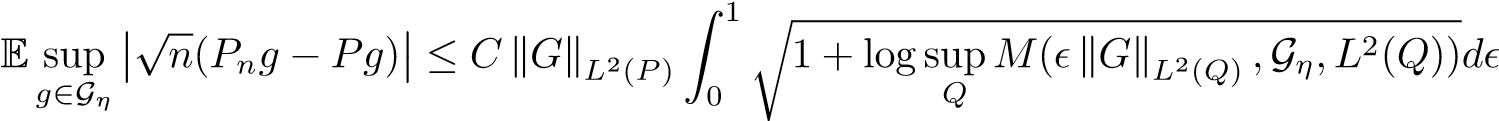
经验过程的期望上确界的一个主要界限（从第9课开始）是

)（200个）

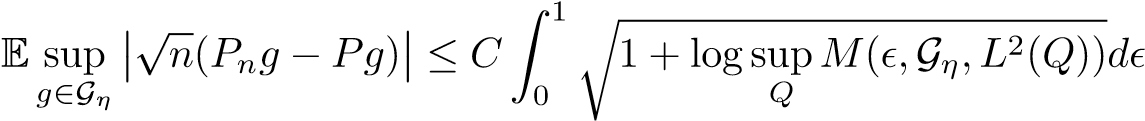
哪里



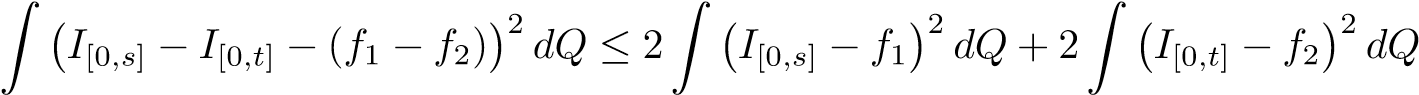
其中H是H的包络，定义为H（x）：=suph∈H | H（x）|将其应用于H=Gη，我们得到



其中G是Gη的包络现在很容易看出G是等于1的常数函数。注意，PG2=1，而PG2≤η，对于每个g∈gη（换句话说，PG2远大于supg∈gηPG2）因为G∏1，上面的界限变成



我们现在证明上面的积分是由一个常数从上面限制的。至少有两种方式可以展示这一点对于第一种方法，证明类{I[0，s]-I[0，t]：0≤s，t≤1}具有有限的VC子图维数（最多3？？）并使用我们先前的包装数和VC子图维数之间的关系第二种方式，平凡的不平等



它适用于每一个s，t和每一对函数f1和f2，这意味着

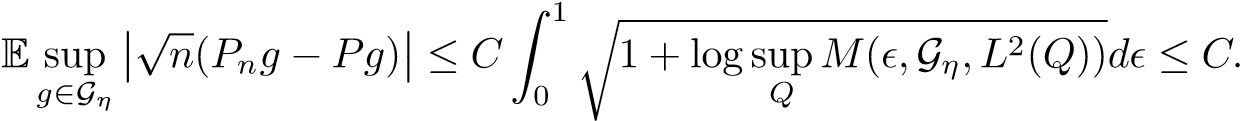
.

这是因为我们可以在L2（Q）距离中覆盖函数I[0，s]-I[0，t]到2δ内，方法是覆盖单个函数I[0，t]到δ内，然后在覆盖中获取所有函数对。这给了

（201年）

其中c和c是正常数（后一个不等式是根据F是一个具有VC维数1的布尔函数类这一事实得出的）。

因此，我们已经证明

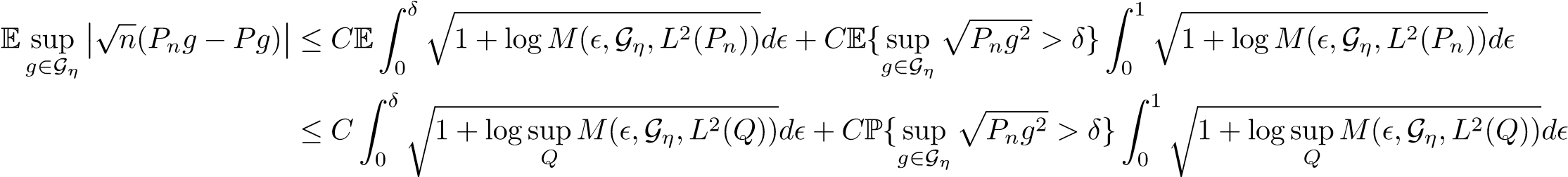


注意，上面的第二个不等式不能得到显著的改进，因为积分至少是1不幸的是，上面的约束不够强，无法屈服

（202）

为了改进我们的边界以推导出上述结果，我们需要使用优于（200）的边界。回想一下（从第9课开始），虽然（200）被称为一个主界，但它实际上遵循以下不等式：

（203）从这个界限，我们可以论证如下对于每个δ∈[0,1]，我们可以



通过（201），我们可以用常数代替第二个积分，用覆盖数F的积分代替第一个积分，得到

（204）最后一个期望上确界可以通过

!

其中我们使用了平凡不等式√a+b≤√a+√b。如前所述，supg∈GηPg2≤η

.

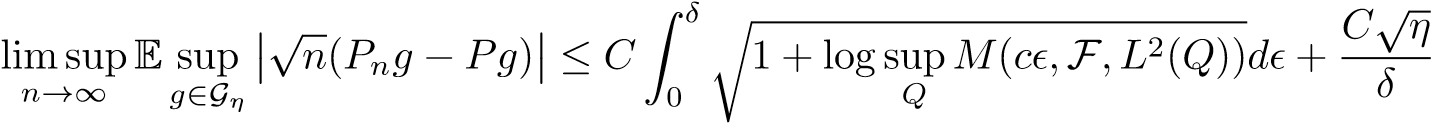
现在注意{g2:g∈gη}是一个VC维数最多为2的布尔类，因此

.

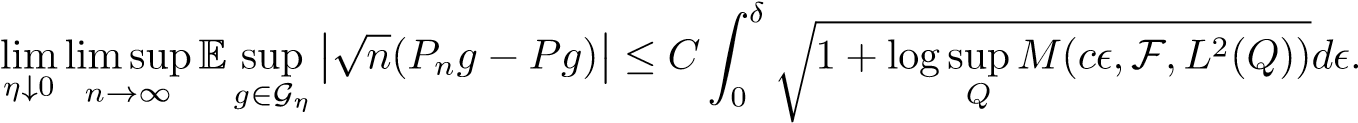
结合（204），我们得到

.

因此，



进一步



因为对于每一个δ>0都是这样，所以如果我们把右手边的极限取为δ→0，不等式也成立。现在很容易证明这个极限是零（这是因为从0到1的积分是有限的，所以从0到δ的积分应该作为δ→0由主导收敛定理变为零）从而证明了（202）与Xn（t），t∈[0,1]的随机等连续性相同。结合有限维收敛，证明了均匀经验过程在分布上收敛于布朗桥这是一致经验过程的Donsker定理。

我们必须做一点以上的工作，从（203）到（202）。存在其它极大不等式，使得人们更容易推断出（202）。下面的定理（摘自Kato[12，定理8]）就是这样一个结果。

定理21.3设H是H类的一个包络，其PH2<∞那么

（205）

每δ满足

.

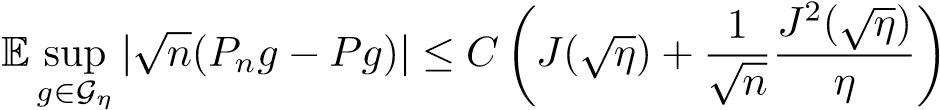
在这里

r和Γ：=E最大H2（Xi）。

1≤i≤n

现在让我们证明定理21.3很容易得到（202）实际上，应用不平等（205）

√2≤η，PG2=1），得到H=Gη（包络H∏1）和δ=η（注意supg∈GηPg



这意味着



如前所述，变为0，即η≤0这给出了（202）的较短证明（尽管依赖于定理21.3的非平凡结果）。

# 22讲座22

在上一课中，我们证明了一致经验过程在`∞[0,1]中收敛于布朗桥的分布。其主要内容是证明均匀经验过程的随机等连续性条件。这个条件说明

四肢支撑| Xn（t）–Xn（s）|=0

η￥0 n→∞s，t∈[0,1]：| s−t |≤

其中Xn（t）是统一的经验过程我们上一次观察到这句话相当于

=0（206）

哪里

.

√

在这堂课中，我们将首先把这个论点推广到更一般的过程{n（Pnf-Pf）：f∈f}。具体地说，让F是一类函数

还有。

我们将在F上提供一个（206）成立的充分条件。为了控制（206）中的期望值，我们将使用下面的最大不等式（在前面的课中也有这样的表述）：

定理22.1。设H是H类的一个包络，其PH2<∞那么

（207）

每δ满足

.

在这里

r和Γ：=E最大H2（Xi）。

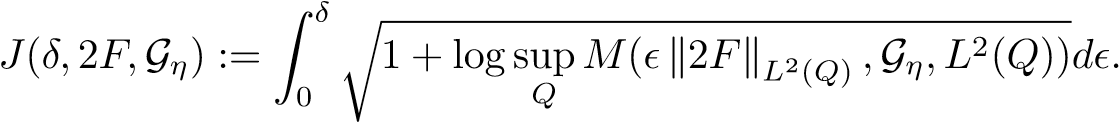
1≤i≤n

我们将定理22.1应用于H=Gη假设F是F的包络，那么很明显，2F是Gη的包络因此，我们在应用定理22.1时取H=2F我们得到

!

（208个）

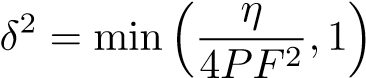
哪里



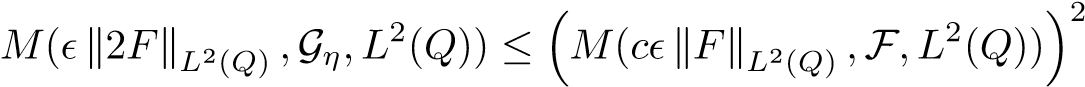
这里，如定理22.1所示，δ是满足

.

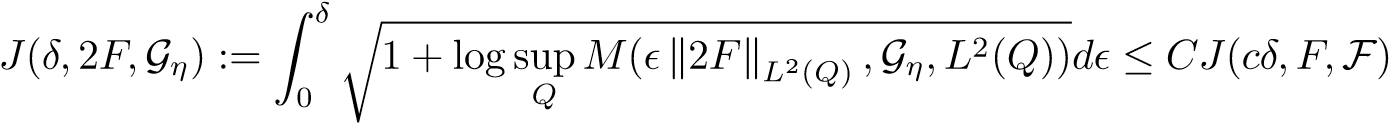
因为supg∈GηPg2≤η，我们可以



所以δ/-0等于η/-0。因为Gη是FF的子集（FF是所有函数的类{f1f2:f1，f2∈F}），我们可以用F的填充数的平方来平凡地限定Gη的填充数，



对于正常数c，这给出



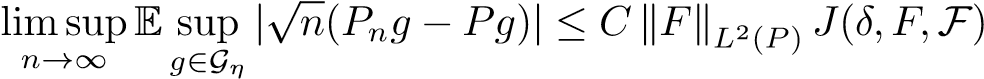
对于两个正常数c和c，将其插入（208），我们得到

.

因此，我们得到

.

下面的引理22.2则意味着上面右手边的limsup项等于零（只要J（cδ，F，F）<∞）



如果我们现在假设limδ＊0j（δ，F，F）=0，那么我们建立（206）。注意limδ≤0 J（δ，F，F）=0是

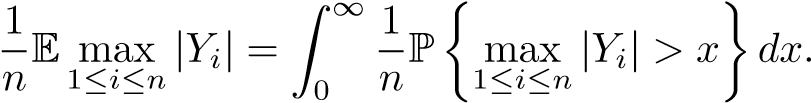
（209）

它有待于陈述和证明引理22.2。

引理22.2。假设Y1，…，Yn是E | Y1 |<∞的同分布随机变量（这里没有独立性假设）。那么

（210）

引理的证明22.2这是支配收敛定理的结果。我们写作



对于每个固定的x，很明显，上面的被积函数收敛为零，即n→∞此外，被积函数的界是（由并界和相同的分布假设）P{Y1 |>x}，P{Y1 |<∞积分到E | Y1 |<∞。因此，语句（210）遵循支配收敛定理。

## 22.1一致熵条件下的Donsker定理

因此，我们证明在条件（209）（称为均匀熵条件）下

√

随机过程Xn（f）：=n（Pnf-Pf）满足关于度量（f，g）→k7 f-gkL2（P）的随机等连续性条件。另一方面，我们也通过通常的多元中心极限定理在这里有有限维收敛，即。，

一

（Xn（f1），…，Xn（fk））→N（0，∑）

式中∑（i，j）=Cov（fi（X1），fj（X1））对于每k≥1和f1，…，fk∈F，我们可以应用上一类的过程收敛结果，得到Xn（F），F∈F在`∞（F）上的分布收敛该定理还保证了极限过程X（f），f∈f是一个高斯过程（即（X（f1），…，X（f k））对于每k≥1和f1，…，fk∈f）具有相对于度量（f，g）7→kfgkL2（P）的连续采样路径。这些结论在下面的定理中得到了重申。

定理22.3。假设均匀熵条件（209）。然后，存在一个具有连续样本路径的高斯过程X（f），F\*过程F（关于由Xn（f）定义的度量{xn}：=πn（pN-kf＝gkL2（p）），使得随机ρ（f）的序列。

f-Pf），f∈f在分布上收敛到X-in`

定义22.4（Donsker函数类）。假设一类函数F相对于

√如果随机过程X n（f）：=n（Pnf−Pf），则概率测度P（也称为P-Donsker），f∈f在分布上收敛于高斯过程X（f），f∈f相对于度量（f，g）7→kf−gkL2（P）具有连续样本路径。

因此，定理（22.3）指出，如果F满足一致熵条件（209），那么对于每个概率测度P，F是PDonsker。

## 22.2 Donsker类的括号条件

用包围熵数代替（209）中的均匀熵，得到了P-Donsker的另一个充分条件。具体来说，假设

（211）

注意，这个条件取决于概率度量P（与（209）不同）。下面的定理证明了在上述条件下，F是P-Donsker。

定理22.5如果F满足概率测度P的括号条件（211），则F是PDonsker。

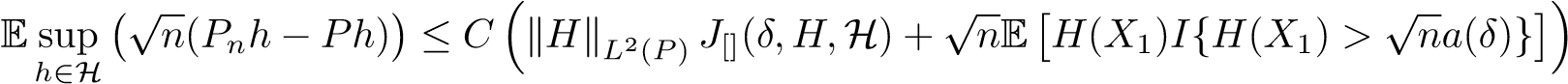
√

为了证明这一定理，足以证明过程Xn（f）=n（Pnf-Pf）在（211）下满足随机等连续性条件。为此，我们将从范德法特〔24，引理19.34〕中使用下列极大不等式。

定理22.6假设H是一类函数H的包络，并假设kHkL2（P）<∞。那么对于满足δ>0的每一个

,

以下不等式成立：



哪里

和

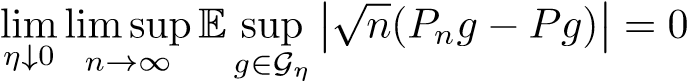
定理22.6与定理21.3中关于带括号的熵的类似。此外，定理22.6可以看出我们以前的基于包围的最大不等式的改进：

.

当δ很小时，定理22.6给出的界比上述界好得多。

我们现在用定理22.6证明定理22.5（关于定理22.6的证明，请参考Van der Vaart[24，引理19.34]）。

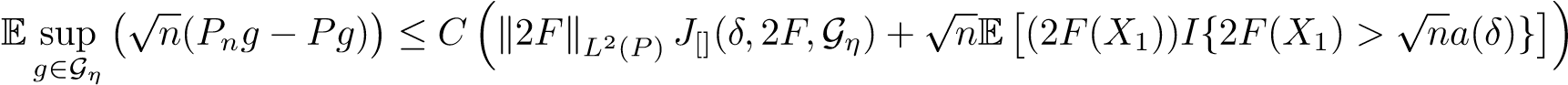
定理22.5的证明证明F是P-Donsker的关键是证明随机等连续性（有限维收敛遵循通常的中心极限定理）。对于随机等连续性，我们需要证明



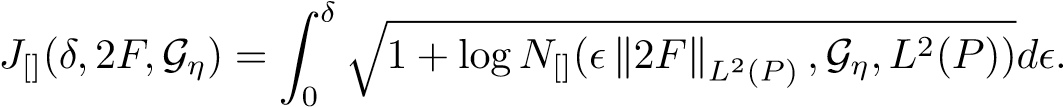
哪里

.

为此，我们使用定理22.6，其中H：=Gη和H=2F（其中F是F的包络）来获得



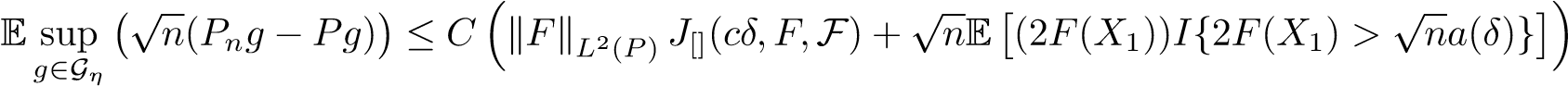
具有



因为{f-g:f，g∈f}的括号数可以被f的括号数的平方所限制，我们得到



对于两个正常数c和c，我们得到



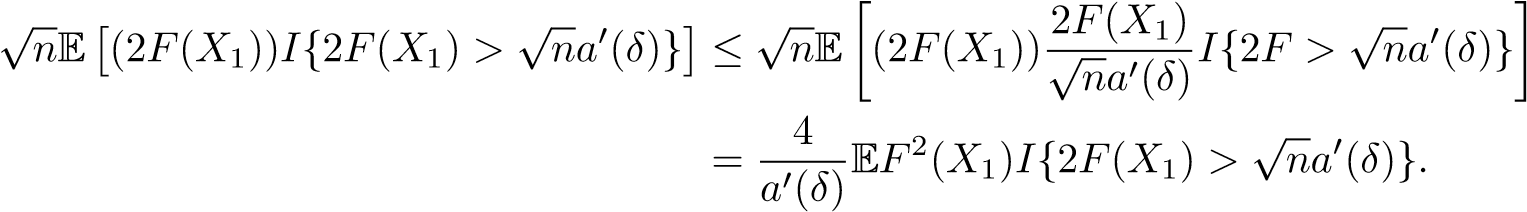
阿尔索

.

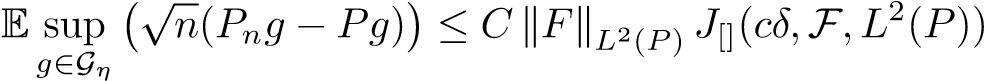
因此我们有

.

注意a0（δ）不依赖于n。上面的第二个项可以有界于



利用支配收敛定理，上述期望值收敛到零，即n········（注意，我们假设EF2（X1）·····）我们已经证明了



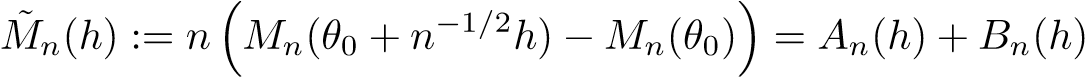
每δ>0在假设（211）下，上述右侧收敛为0，即δ→0。这证明了随机等连续性，从而证明了F是P-Donsker。

## 22.3样本中值收敛速度的应用

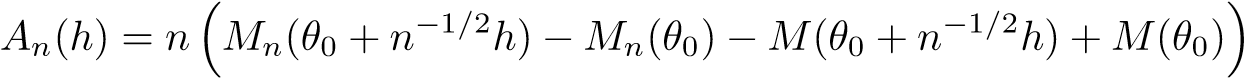
我们给出了一个例子来证明样本中值的收敛速度是研究过程收敛性的动机之一既然我们已经了解了什么是过程收敛，那么让我们回顾一下这个例子，并严格说明这个论点。设置如下我们有i.i.d数据X1，…，Xn是由N（θ0，1）分布产生的让

M（θ）：=E | X1−θ|

对于Tyr\* R，估计Tyth-n定义为Mn（Th）上的最小极小值，同时也证明了Th 0唯一地最大化M（Th），Th·r。为了得到θˆn的极限分布，我们考虑了以下过程：



哪里



和

.

我们早先已经看到，A N（h）→la（h）：=h Z，其中Z∼N（0,1）和Bn（h）→B（h）：=M00（θ0）h2/2，对于每个固定的h∈R。第一个收敛是Lindeberg-Feller CLT的应用，第二个收敛是从Taylor展开到二阶的。这些收敛性声明实际上可以得到很大的加强。事实上，它认为

An→L A in`∞–[－Γ，Γ]对于每个固定的Γ>0（212）

和

Bn→B均匀地覆盖在每∏>0（213）的∏Γ，Γ]上

上述一致收敛意味着sup | h |≤Γ| B n（h）-B（h）|→0作为n→∞语句（213）通过通常的泰勒展开论证（左为练习）很容易证明。我们将在下面简述（212）的论点过程收敛性声明（212）需要两个要素：有限维收敛性和随机等连续性对于有限维收敛，我们需要证明

（An（h1），…，An（hk））→L（A（h1），…，A（hk））

对于每k≥1和h1，…，hk∈[-Γ，Γ]。这可以通过多元Lindeberg-Feller中心极限理论（左为练习）来证明。对于随机等连续性，我们需要证明

（214）

为此，请注意

E sup | An（h1）－An（h2）|＝E sup（n | Png－Pg |）h1，h2∈[Γ，Γ]：| h1－h2 |≤ηg∈gη

哪里

Gη：=nx→| 7X−θ0−n−1/2h1 |−x−θ0−n−1/2h2 |：h1，h2∈[——Γ，Γ]，| h1−h2 |≤ηo。

然后，可以使用经验过程的期望上确界（例如基于括号数字的界）中的一个界来证明语句（214）。这又是作为练习留下的。

可以将两个语句（212）和（213）添加到yield（验证这一点）：

对于每一个固定的Γ>0，在`∞[－Γ，Γ]中。

√因此，M∮n的极限过程为M∮（h）：=hZ+M00（θ0）h2/2，对于h∈R

如果我们能证明argminh∈R M∮n（h）在分布上收敛于argminh∈rm∮（h），则n（θˆn∏θ0）现在紧随其后这可以从一个一般的argmax连续映射定理中推导出来。

## 22.4 Argmax连续映射定理

定理22.7。设H为度量空间。设{Mn（h），h∈h}和{M（h），h∈h}为h指标的随机过程，假设下列条件成立：

1. H的每个紧子集K的M→lm在`∞（K）中。
2. M的每个实现在H上是连续的。
3. H H使H（H）上的Mn（H）最大化。
4. 设H是H（H）上M（H）的唯一极大化子。
5. 紧密性：对于每一个，存在一个紧凑子集，使得

和

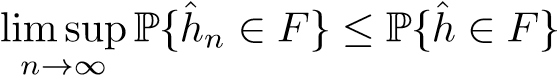
也就是说，对于每个有界连续函数f:H→R，我们有Ef（Hˆn）→Ef（Hˆ）作为n→∞。

备注22.1。通常情况下，定理22.7适用于该过程

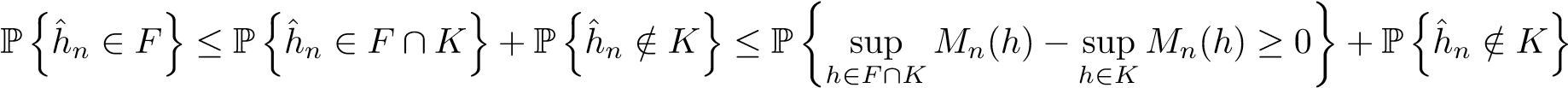
和M？作为M？n的极限过程。在这种情况下，请注意，以及由此产生的紧密性条件

相当于θˆn-θ0=OP（rn-1）。因此，在应用定理22.7获得rn（θˆn-θ0）的渐近分布之前，需要证明一个初步的速率结果。

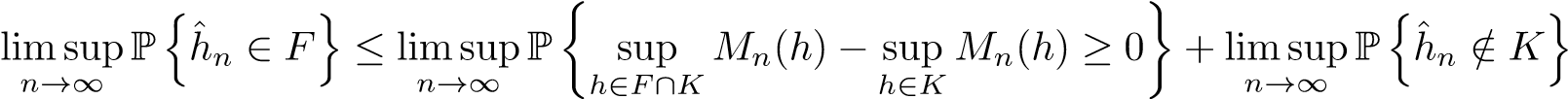
备注22.2我们可以将定理22.7应用于M n（θ）、θ∈Θ和M（θ）、θ∈Θ（而不是M∮n和M∮）这通常会导致θˆn的一致性结果。定理22.7的证明。足以证明



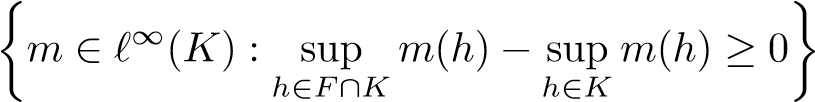
对于H的每个闭子集F，固定闭子集F⊆H，并在H中固定任意紧集K



它给予



现在请注意



是`∞（K）的闭子集。这是因为如果suph∈F∩K mk（h）－suph∈K mk（h）＞0，且mk→m在K中一致，则suph∈F∩K m（h）－suph∈K m（h）＞0。因此，从Mn在`∞（K）中收敛到M，我们得到

.

因此我们得到

.

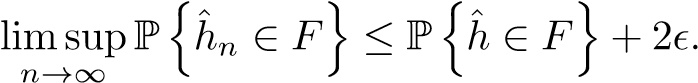
我们现在宣称

.

原因是当右手握着时，我们有suph∈F∩K M（h）≥suph∈km（h）≥suph∈hm（h）M的样本路径的连续性和F的封闭性（这意味着F k是紧的）意味着SF H f k k（H）在F k的某个点上实现。M上的唯一最大假设将意味着f k k中实现m的最大值的点必须等于H，这意味着H.f f。

.

注意，这对于H的每个闭子集F和H的每个紧子集K都是正确的。现在固定>0并选择K作为紧性条件。这会给



让趋向于零来完成证明。

将此结果与上一节的过程收敛结果一起使用，以完成样本中值极限分布结果的证明。

# 23讲座23

本课程的主要目的是证明下列定理（范德法特[24，定理5.23]），这些定理证明了M-估计在某些一般条件下的渐近正态性为了稍微简化证明，我对定理做了一些简化（例如假设标准函数用R索引；Van der Vaart[24，定理5.23]中的完整定理适用于标准函数用Rk中的一个开集为固定k索引的情况）。

## 23.1抽象M估计结果

定理23.1假设{mθ，θ∈R}是一类由R.给定i.i.d观测值索引的函数

x1，…，具有分布p的Xn，我们将估计的πn定义为pNM上的任何极大值，超过π^ r。设TH 0是定义为Ph th\*上的任何最大化的π-π的人口类似物。

1. 假设θ7→mθ（x）在θ0处可微，导数m˙θ0（x）几乎是确定的x（w.r.t P）。
2. 假设存在一个函数p（x），p p＜2＜ω（即λL2（p））

|mθ1（x）－mθ2（x）|≤Γ（x）|θ1－θ2 |（215）

对于所有θ1、θ2和x。

1. 假设θ7→M（θ）：=Pmθ在θ0处是两个连续可微的，M00（θ0）<0。
2. θˆn对于θ0是一致的，即θˆn→Pθ0作为n→∞。

然后得出以下两个结论：

1. θˆn到θ0的收敛速度为n−1/2，即|θˆn−θ0 |=OP（n−1/2）。
2. 以下情况成立：

如n·····（216）

在开始证明这个定理之前，让我们先看一下下面的注释。

1. 该定理的条件在mθ（x）=––| x–θ|时成立，因此该定理可以看作是我们对上一类样本中值的极限分布结果的推广。
2. 注意，假设标准函数θ7→mθ（x）仅在θ0处（几乎肯定是关于x）对θ可微一次但极限函数M（θ）=Pmθ假定为两次可微。如果我们坚持准则函数是二次可微的，那么这个定理将不再适用于mθ（x）=––| x–|θ|等函数。然而，M估计的渐近正态性的经典证明将泰勒展开到二阶，并且这些论据需要存在二阶导数（以及一些附加正则性）。
3. Th 0不被认为是M（Th），Th，r的唯一最大值，而不是这样，则假定Tyth-n对于Th 0是一致的，即它在概率上收敛到Th 0。这意味着θˆn将接近θ0，为了得到θˆn的渐近图，我们可以关注θ0的局部区域因此，这个定理是一个局部结果，所有的注意力都集中在θ0的局部区域上。

我们现在证明定理23.1。它将使用到目前为止我们在本课程中看到的一些想法和结果。

定理23.1的证明。第一个任务是证明收敛速度是n-1/2。为此，我们可以直接使用速率定理。设Mn（θ）：=Pnmθ和d（θ，θ0）=。为了确定利率，我们必须

E sup（Mn-M）（θ-θ0）

θ：|θ-θ0 |≤δ

然后把束缚等于δ2以上数量等于

E sup（Pn-P）（mθ-mθ0）≤E sup（Pn-P）（mθ-mθ0）|。

θ：|θ-θ0 |≤δθ：|θ-θ0 |≤δ

为了控制上述期望的上确界，我们使用括号边界（从第12课开始）：

（217）

这里的相关类H是{mθ–-mθ0:|θ–-θ0 |≤δ}，其包络（根据Lipschitz条件（215））可以被认为是H（x）：=Γ（x）δ。因此我们有

.

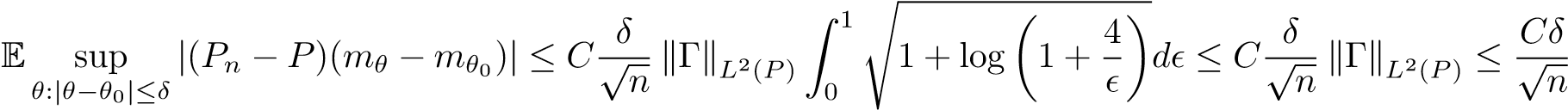
为了控制上面的括号数，我们使用第12课的这个结果：如果Θ⊆Rd包含在半径R的球中，并且如果{gθ：θ∈Θ}是满足| gθ1（x）–gθ2（x）|≤Υ（x）kθ1−θ2k的函数类，对于所有x和θ1，θ2∈Θ。如果Υ∈L2（P），则

每（218）

将这个结果与gθ：=mθ-mθ0和Υ=Γ的{gθ：|θ−θ0 |≤δ}类结合，我们得到了

.

因此我们得到

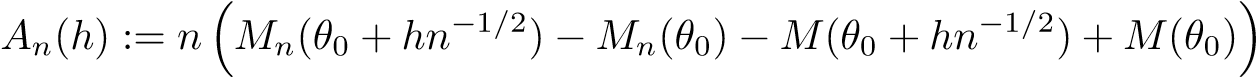


因为kΓkL2（P）是有限的因此，为了得到|θˆn－θ0 |的速率上限，我们可以解出δn−1/2=δ2，得到δ=n−1/2因此，我们证明了|θˆn-θ0 |=OP（n-1/2）。

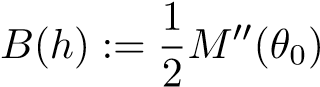
现在我们将试图证明（216）为此，我们考虑



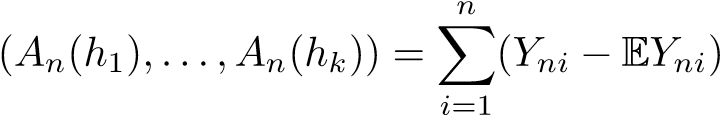
以h∈R为指标，可分解为M∮n（h）=An（h）+Bn（h），其中



还有通过对m 0（Th）的二阶泰勒展开，证明了m（Th）在其最大值为0的点（m＝0（t＝0）＝0）的情况下，它的最大值是两次连续可微的，这也意味着BN（H）收敛于



对于每个固定的h∈R，我们现在将证明，对于每个固定的K，a（h）在`∞[-K，K]中收敛于随机过程a（h）。为了证明这一点，第一步是建立有限维收敛，即（An（h1），…，An（hk））在分布上收敛于固定的K和h1，…，hk。为此，我们将使用林德伯格费勒CLT请注意



哪里

.

因为X1，…，Xn是i.i.d，我们有Cov（Cov（Yn1）。现在对于每个固定的h∈R，



现在通过对θ0处的准则函数mθ（x）的几乎确定的一阶导数假设，我们得到

)几乎可以肯定。

同样根据Lipschitz假设（215），我们有

.

因此通过控制收敛定理，我们得到

))如n·····。

同样，对于每个固定的h1和h2，我们都有nCov

因此

n个

十

Cov（Yni）→Cov（A（h1），…，A（hk））为n······

i=1

哪里

A（h）：=Zhpvar（˙mθ0（X1）），其中Z∼N（0,1）。

此外，请注意

.

因此



通过支配收敛定理（注意，我们假设

EΓ2（X1）<∞）。Lindeberg Feller CLT的假设都满足，因此我们可以得出结论

（A n（h1），…，An（hk））→L（A（h1），…，A（hk））作为n→∞

对于每个固定k≥1和h1，…，hk∈R。

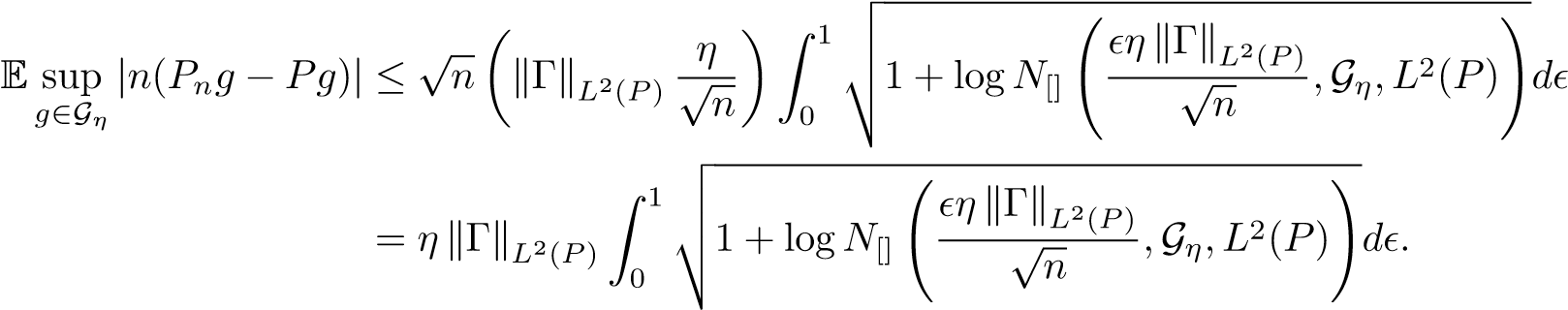
为了将每个固定K的有限维收敛转化为`∞[-K，K]的过程级收敛，我们需要证明随机等连续性

E sup | An（h1）－An（h2）|=E sup | n（Png－Pg）| h1，h2∈[－K，K]：| h1－h2 |≤ηg∈gη

哪里

.

根据Lipschitz假设（215），很明显，函数x 7→ηn-1/2Γ（x）是Gη的包络。因此，边界（217）给出



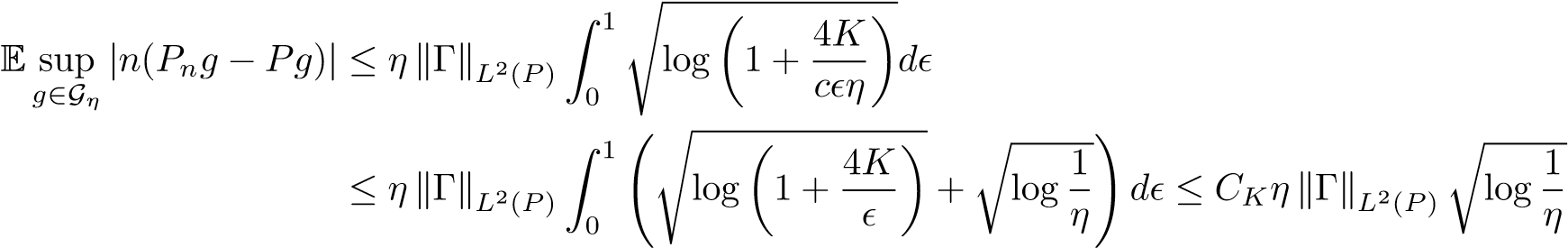
很容易看出，对于一个足够小的正常数c，

.

因此，通过使用（218）来控制上面右边的括号数字，我们得到

.

因此我们得到



式中，CK是一个仅依赖于K的常数。上面的右手边显然是以η→0的形式归零这证明了{An（h），-K≤h≤K}的随机等连续性。结合前面建立的有限维收敛结果，我们可以推导出，对于每一个K≥0，在`∞[-K，K]中的An→L A。

还证明了Bn（h）到B（h）在每个固定h∈R上的早期收敛可以改进为在[-K，K]上的一致收敛。这是M在θ0处两次连续可微性的结果。

利用`∞.[-K，K]中的A n→L A和`.[-K，K]上的Bn→B一致，我们可以推导出M∮n=An+Bn→L M∮：=A+B in `∞.[-K，K]。因此，我们可以使用argmax连续映射定理（所有条件都满足）得出结论

.

这就完成了定理23.1的证明。

## 23.2 MLE申请

定理23.1适用于极大似然估计。设P={Pθ，θ∈Θ}表示一类概率测度，其中Θ是R的开子集，并假定X1，…，Xn是来自Pθ0的i.i.d观测。假设每个Pθ具有关于公共支配测度μ的密度Pθ。在这种情况下，

定理23.1适用于mθ（x）=logpθ（x）和P=Pθ0。如果定理23.1的假设成立，那么

√

因此，每个MLEθˆn都有n个收敛速度和

.

这个结果的优点是它只需要logpθ在θ0处可微一次就可以几乎确定x（函数˙mθ0（x）称为得分函数）。相比之下，MLE的渐近正态性的传统结果需要在Th 0上存在至少两个Logp Th的导数。渐近方差由

.

这里的分子是Fisher信息I（θ0）。在˙mθ0（x）的附加光滑性假设下，可以证明M00（θ0）=-I（θ0），因此渐近方差是熟悉的1/I（θ0）。涉及到额外光滑性的最清楚的假设是Le-Cam在二次均值下的可微性。

定义23.2（二次均值可微性（DQM））。我们认为{p ^，th} }在二次均值为0°的情况下是可微的，如果存在一个函数α^ 0×L2（p ^ 0），则

作为θ→θ0。

在DQM假设下，函数`˙θ0扮演得分函数的角色，Fisher信息将由I（θ0）=varPθ0（`˙θ0（X1））定义（更多细节将在下节课中给出）下一个结果断言了DQM下MLE的N（0,1/I（θ0））渐近分布和logpθ上的一个附加Lipschitz假设这是范德法特[24，定理5.39]。

定理23.3设Θ是θ0∈Θ的开集假设{Pθ，θ∈Θ}在θ0处满足DQM同时假设

|对数pθ1（x）－对数pθ2（x）|≤Γ（x）|θ1－θ2 |（219）

对于所有x和θ1，θ2在θ0的邻域中，Pθ0Γ2<∞如果I（θ0）>0且θˆn与θ0一致，则

.

证明定理23.3的一个关键因素是，DQM性质意味着另一个称为局部渐近正态性（LAN）的性质我们说{Pθ，θ∈Θ}在θ0满足LAN

（1）（220）

其中Sn在分布上收敛到Pθ0下的N（0，I（θ0））。下节课将展示，θ0处的DQM表示I=I（θ0）的LAN应该清楚的是（220）加上额外的Lipschitz假设（219）以及√θˆn的一致性意味着（23.3）首先要注意的是

（220）为M●n（h）如果我们定义了M（h）=h Z I-h2I/2，其中Z∼N（0,1）和I=I（θ0），则（220）意味着M（N）的有限维分布收敛于M（N）的有限维分布。在Lipschitz假设（219）下，这种有限维收敛可以用过程收敛来补充，从而在`∞.[-K，K]中对每一个固定的K≥0产生收敛然后可以使用argmax连续映射定理来产生（23.3）这将完成定理23.3的证明。因此，在DQM假设下建立（220）是证明定理23.3的关键。我们将在下节课中证明这一重要事实（DQM意味着LAN）。

# 24讲座24

本课程将讨论二次平均数（DQM）和局部渐近正态性（LAN）的可微性。我正在关注Pollard[19]中非常干净的处理方法，我建议您阅读这篇漂亮的文章。

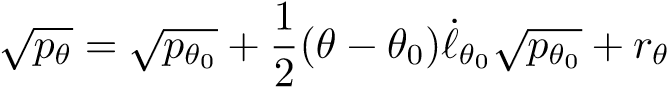
## 24.1二次均值的可微性

基本设置如下我们有一类概率测度P：={Pθ，θ∈Θ}在一些空间上，这些空间被R的子集Θ索引（对于固定k≥1，扩展到Θ⊆Rk的情形是可能的，但是为了简单起见，我们将限制到k=1）。假设存在一个单西格玛有限测度μ，每个Pθ具有将由Pθ表示的密度以下是DQM的定义。

定义24.1（二次均值可微性（DQM））我们认为P在二次均值为0°时是可微的，如果存在一个函数α^ 0×L2（p ^ 0）

作为θ→θ0。

换句话说，如果P在θ0处满足DQM，那么我们得到了展开式：



其中rθ满足

.

LE-CAM表明，在DQM下，可以证明统计中的经典渐近结果（如极大似然估计的渐近正态性），而不要求密度Th 7×p（x）在0°上是两次或三次可微的。

以下引理表明，如果P在θ0处满足DQM，并且如果θ7→Pθ（x）在θ0处在通常意义上可微，则DQM给出的函数`˙θ0与logpθ（x）的通常导数一致。

引理24.2。设P在θ0∈Θ满足DQM，且函数`˙θ0。还假设θ7→pθ（x）在θ0处可微，导数p˙θ0（x）几乎确定x相对于测度μ那么

`˙θ0（x）pθ0（x）=a.s x（w.r.tμ）的p˙θ0（x）。

证明。假设{θn}是收敛到θ0的序列。DQM假设允许我们编写

)（221个）

对于所有x和n

（222）

如果有必要的话，我们假设

（223）

例如，这可以通过将{θn}替换为选择{nk}的子序列{θnk}来实现，以便

.

我们现在使用这个事实，它意味着几乎肯定（通过单调收敛），这进一步意味着fi→0几乎肯定为i····因此，假设（223）

0 a.s（w.r.tμ）作为n→∞。

因此，我们可以将（221）重写为

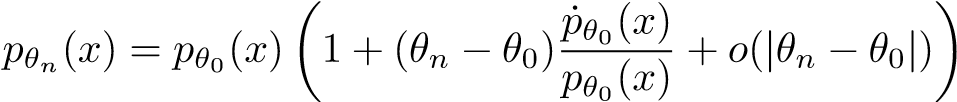
)a.s（w.r.tμ）为n→∞。（224）

现在让我们利用θ7→pθ（x）在θ0处可微，导数˙pθ0（x）几乎肯定是关于μ的，并写出

pθn（x）=pθ0（x）+（θn−θ0）p˙θ0（x）+o（|θn−θ0 |）a.s（w.r.tμ）as n→∞。（225）

观察（224）和（225）都保持几乎确定的x（关于μ）。

我们现在将处理两个不同的案件第一种情况是pθ0（x）>0在这种情况下，我们可以将（225）重写为（注意pθ0（x）不依赖于n，因此o（|θ−θ0 |/pθ0（x））=o（|θ−θ0 |）：



取两边的平方根，我们得到

.

√

通过x=0时x 7→x的泰勒展开到一阶，我们从上面推导出

.

通过与（224）的比较，我们推断

.

现在让我们考虑pθ0（x）=0的情况。在这种情况下，（224）和（225）分别变成

pθ（x）=o(

p n |θn−θ0 |）=•？pθn（x）=o（|θn−θ0 | 2）

pθn（x）=（θn-θ0）pθ0（x）+o（|θn-θ0）。

将上述两个方程等价，我们得到

o（|θn-θ0 | 2）=（θn-θ0）p˙θ0（x）+o（|θn-θ0 |）。

除以|θn－θ0 |，让n→∞，我们得到˙pθ0（x）=0，即在这种情况下，方程˙θ0（x）pθ0（x）=p˙θ0（x）也满足这就完成了证明上面的引理暗示

当pθ0（x）>0时。

上面的右边是经典的分数函数。因此，当DQM成立时，我们将函数`˙θ0称为得分函数。

经典分数函数的一个标准事实是，它对概率测度Pθ0的期望等于零。这方面的经典证明涉及交换微分阶w.r.tθ和积分：

.

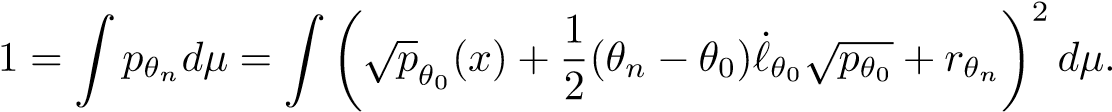
以下引理表明，DQM假设直接暗示了这一事实。

引理24.3。假设P在θ0处满足DQM，且得分函数为`˙θ0那么

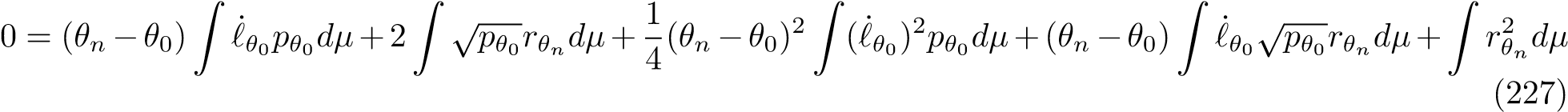
Z轴

`˙θ0（x）pθ0（x）dμ（x）=0。（226）

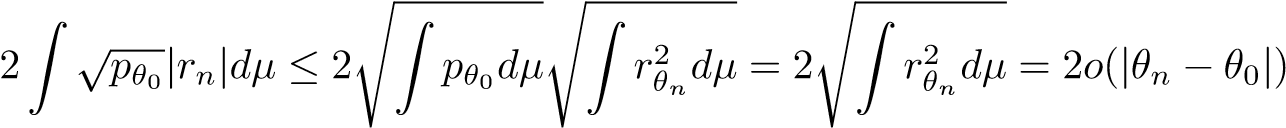
证明。设θn是收敛到θ0的序列通过DQM表示，我们可以写出（221），余项rθn满足（222）请注意



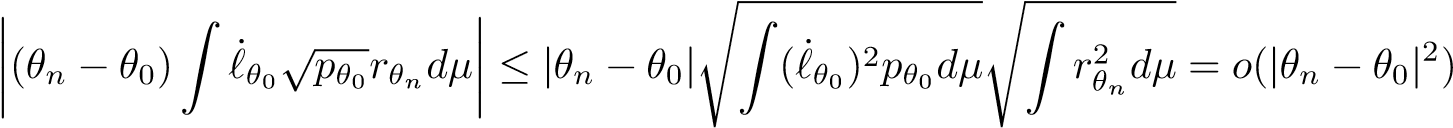
我们现在展开右边的正方形，上面将引出六个条件其中一项等于R pθ0=1，与左手边相消因此我们得到



上面右边的第一项在绝对值中显然是O（|θn−θ0 |）。第三项是O（（θn--θ0）2）最后一项（by（222））等于o（（θn-）θ0）2）。剩下的两项（第二项和第四项）可以通过Cauchy-Schwarz不等式控制为



到（222）和



再次通过（222）因此，很明显（227）中右边的前导项是第一项通过将方程（227）除以|θn−θ0 |，并让n→∞，我们推导出（226）。

我们现在将定义Fisher信息。假设P满足DQM且得分函数为˙θ0然后θ0处的Fisher信息由

.

引理24.3的证明中使用的论点引出了一个有趣而重要的事实，涉及`˙θ0和Fisher信息。因为R`˙θ0pθ0dμ=0，我们可以将其插入（227）中以获得（同样使用（227）中的最后两个项是o（|θn−θ0 | 2））：

.

在R（`˙θ0）2pθ0dμ=I（θ0）的情况下，我们得到

（228）

这一事实对于确定DQM意味着LAN是至关重要的关于（228）的有趣方面如下。语句（222）意味着krθnkL2（μ）=o（|θn--θ0 |）因此，如果我们使用上面左边的Cauchy-Schwarz不等式，我们得到左边是o（|θn−θ0 |）但是上面的等式意味着右手边是O（（θn-，θ0）2），这是一个更有力的结论，可以从

与Rθn.Pollard-Cauchy-Schwarz的L2（μ）范数相比，Rθn√pθ0dμ要小得多。因此

[19]将这种现象归因于函数√pθn2（μ）都有范数1（这在L中是清楚的

从以上（228）的证明中，并认为这是DQM魔力背后的主要原因。

## 24.2局部渐近正态性

DQM是关于密度pθ在θ0处的第一可微性的一个陈述。当然在DQM中没有提到二阶可微性。然而，值得注意的是，DQM假设意味着对数似然函数在θ0附近有一个二阶泰勒展开式，其尺度为n-1/2。这种局部泰勒展开称为局部渐近正态性（LAN），并在下面的结果中得到了证明。

定理24.4假设P在θ0处满足DQM，且得分函数为˙θ0，Fisher信息为I（θ0）。那么对于每个固定的h∈R，我们有

如n·····。

同样地，上述定理的结论可以写成

（1）作为n→∞。（229）

我们说P在θ0处满足LAN性质，如果对每个h∈R都满足，为什么这称为局部渐近正态性？要看到这个，首先要注意，在CLT，我们有

.

因此，作为（229）的结果，我们得到每h∈R，

)在X1，…，Xn∼i.i.d Pθ0下。

现在考虑一个第二估计问题，其中有一个观测Y，其密度属于{Qh，h∈R}如果Qh具有密度qsh，该密度qsh是具有平均h和方差1/I（θ0）的正态分布的密度很容易看出

)在Y∼N（0,1/I（θ0））下。

因此（229）有效地说{Pθ，θ∈Θ}（只要P满足DQM，它可以是任意的）的似然比类似于正常实验{Qh，h∈R}的似然比，其中Qh=N（h，1）。因此，在n-1/2尺度上，在θ0附近渐近地，原始统计问题P变成了一个正态平均估计问题。这就是为什么（229）被称为局部渐近正态性。

我们现在证明定理24.4。

定理24.4的证明。这个证明中的所有期望和概率都与概率测度Pθ0有关。写

哪里

我们会利用这个事实

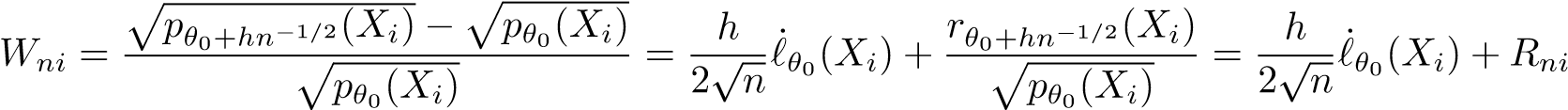
)林在哪里

y→0

或等价地，β（y）=o（y2）作为y→0这给了

.

使用DQM表示，我们可以编写



哪里

.

因此我们得到

.

现在通过DQM观察，我们知道关于随机变量Rni的以下内容：

.

这给出了L1中的（1）和0（Pθ0），这进一步意味着

0个。另外，根据柯西-施瓦兹不等式，我们有

.

因此我们有

.

我们稍后会证明

（1）（230）

所以我们有

.

上面右手边的第三项在概率上明显收敛到−h2I（θ0）/4（根据强大数定律），因此为了完成定理24.4的证明，我们只需要证明

（231）

为此，写下

不，不

2XRni=2XEθ0Rni+2X（Rni-ERni）。

i=1 i=1 i=1

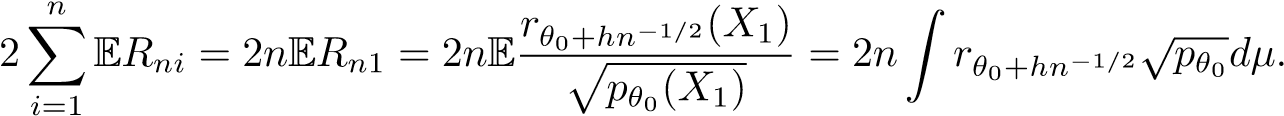
因为

,

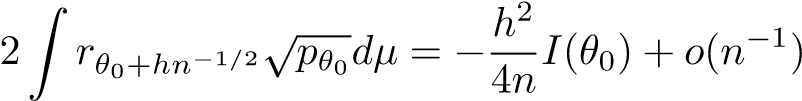
我们得到

（1）如n·····（232）

请注意



我们现在用事实（228）来说明



以便

.

结合（232），我们得到（231）。为了完成定理24.4的证明，我们只需要验证（230）。这主要是β（y）=o（y2）作为y→0的结果事实证明，为了证明（230），这足以证明

（1）和max | Rni |=oPθ0（1）。（233）

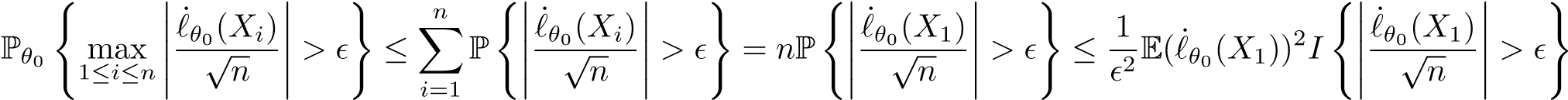
1≤i≤n

实际上，如果这些语句成立，那么（作为β（y）=o（y2）），我们可以写（严格化这个）：

.

我们现在将通过证明（233）中的断言来完成证明对于（233）中的第一个断言，请编写

（对于大于0的固定值），



利用控制收敛定理，以n·····························。

对于（233）中的第二个断言，请编写

.

这就完成了定理24.4的证明。

# 25讲座25

这个类中的下一个（也是最后一个）主题是关于极大极小下界在这堂课中，我们将主要推动极小极大下界的研究。我们从研究这些问题的基本决策理论出发。

## 25.1决策理论框架

在一般抽象决策理论框架下，可以研究极小极大。这里描述了这个框架。这本书的经典参考文献是《弗格森》[8，第1章和第2章]。

我们有一个未知的参数θ。θ可以是实数、向量、函数或矩阵等，我们假设θ取已知集合Θ中的值，我们称之为参数空间。

数据一般用X表示。X也可以是实数、向量、函数或矩阵等。我们假设X取集合X中的值，称为样本空间。

X与θ之间的联系是X的分布依赖于通过已知概率测度Pθ得到的θ。所有概率测度的类{Pθ，θ∈Θ}将用P来表示。我们假设每个Pθ对于一个单西格玛有限测度μ都有一个密度Pθ。

接下来，我们有一个行动空间A，它对应于统计学家在问题中需要采取的行动（例如，在估计问题中，A将等于或大于Θ，在用空假设和替代假设检验问题时，A将对应于两个假设等。具体例子如下）。

损失函数L是定义在Θ×a上的非负函数，即对于每个参数θ∈Θ和作用a∈a，存在一个非负损失L（θ，a）。

非随机决策规则d是从X到A的函数，换句话说，d将一个动作关联到每个X∈X，决策规则d在特定参数值θ处的风险由

R（θ，d）：=EθL（θ，d（X））

上面的期望值是关于X∼Pθ的。

统计学家在决策问题中的目标是选择风险R（θ，d）较小的决策规则d然而，由于风险R（θ，d）依赖于未知的θ，所以“小风险”的说法需要进一步限定。换句话说，决策规则d的风险取决于未知参数值（或自然状态）是什么。所以当我们说小风险的时候，我们需要指定是指θ以上的一致小风险，还是指小平均风险，还是指小最坏情况风险。我们将在看到一些决策理论问题的例子后不久再讨论这个问题。

例25.1考虑在平方误差损失下由观测Y∼Nn（θ，In）估计向量θ∈Rn的问题假设已知θ是k-稀疏的，即θ中的非零项的数目最多为k。这可以放在上面概述的决策理论框架中，将Θ取为Rn中所有k-稀疏向量的集合，X=Rn，A=Θ或A=Rn（取决于我们是否希望估计量是k-稀疏的）和L（θ，A）=kθ-ak2。Pθ也是Nn（θ，In）分布决策规则是θ的简单估计，估计量θˆ的风险由

.

例25.2。考虑与上一个问题相同的设置，但现在假设我们要估计θ的L1范数（而不是整个向量θ）：kθk1：=|θ1 |+··+|θn |。然后，Θ、X和Pθ与上例中相同，但A=R，损失函数为L（θ，A）=（kθk1−A）2。

例25.3。考虑由独立观测值Y1，…，Yn和Yi∼N（f（i/N），1）估计Lipschitz函数f:[0,1]→R的问题，对于i=1，…，N，在这个问题中，我们可以把Θ取为从[0,1]到R的所有Lipschitz函数的类，对于f∈Θ，概率测度Pf是均值为（f（1/N），…，的多元正态分布，。。。，f（n/n））和协方差矩阵作用空间可以取为[0,1]上所有实值函数的空间，损失函数可以是：

或者。

例25.4。考虑从n个独立观测值X1，…，Xn（θ，1）中对H0:θ=0和H1:θ=1进行检验的问题因为我们正在测试θ=0和θ=1，所以我们认为只有这两个值是可能的，所以我们取Θ={0,1}。然后操作空间也是A={0,1}。自然损耗函数是L（θ，A）=I{θ6=A}给定一个决策规则d（test），其风险由

R（θ，d）=Pθ{θ6=d（X）}。

注意，如果θ=0，那么风险由R（0，d）=P0{d（X）=1}给出，这是通常的I型错误当θ=1时，风险由R（1，d）=P1{d（X）=0}给出，这是第二类错误。

例25.5。考虑基于n（θ，1）分布的n i.d观测X1，…，Xn检验H0:θ∈Θ0对H1:θ∈Θ1的假设的问题。在这种情况下，A={0,1}和

L（θ，a）=I{θ∈0，a=1}+I{θ∈1，a=0}。

决策规则（检验）d的风险由R（θ，d）=Pθ{d（X）=1}（如果θ∈0）和R（θ，d）=Pθ{d（X）=0}（如果θ∈1）给出。这些错误可分别被视为I类和II类错误注意，这取决于θ（即，对于每个θ∈Θ0有一个i型误差族，对于每个θ∈Θ1有一个II型误差族）。

## 25.2如何评价决策规则

如前所述，决策规则d的风险R（θ，d）取决于θ事实证明，通常不可能找到一个单一的决定规则d∗，以便

R（θ，d∗）≤R（θ，d）对于每个θ∈Θ和决策规则d。（234）

例如，在具有Θ=A⊆Rk和L（θ，A）=kθ−ak2的估计问题中。考虑固定θ0∈Θ的估计量d0（X）=θ0该估计量在θ=θ0（即R（θ0，d0）=0）时明显具有等于0的风险因此，如果存在满足（234）的判定规则D，则R（Th 0，D））r（Th 0，D0）＝0。因为θ0∈Θ在这里是任意的，这意味着

R（θ，d∗）=Eθkθ∗d∗（X）k2=0

对于每个θ∈Θ这意味着对于每一个θ∈Θ，在Pθ下d∗（X）=θ几乎是确定的对于一般类{Pθ，θ∈θ}，这显然是不可能的。

因此，我们不能指望有一个强有力的最佳决策规则d\*（234）。获得最优性的放松概念有三种常见方法：

1. 第一种方法是限制所有决策规则的一个子类例如，在参数估计问题中，通常将注意力限制在无偏或等变估计上在参数检验问题中，很自然地限制了对水平α检验或无偏水平α检验的关注。这种方法是在STAT 210A中采用的，我们不会在这里继续。
2. 贝叶斯方法
3. 极大极小法

我们将在这里详细研究Bayes和Minimax方法。

## 25.3贝叶斯方法

在这里，我们在Θ上确定了一个概率测度w，并通过它们的平均风险来评估决策规则（其中平均值是相对于w进行的）。换句话说，我们根据决策规则的平均风险来评估决策规则d：

Z轴

R（θ，d）dw（θ）（235）

关于概率测度w，概率测度w也被称为适当先验或简单先验。最小可实现的平均风险称为相对于w的Bayes风险，并用Z表示

RBayes（w）：=inf R（θ，d）w（dθ）。

dΘ

最小化（235）的估计量d被称为关于w的Bayes估计量。

这种评估决策规则的方法的明显问题是它依赖于先验w，在许多情况下，不清楚先验的合理选择是什么。例如，在例25.3的Lipschitz回归问题中，需要在[0,1]上的所有Lipschitz函数的类上选择一个先验，但不清楚如何做到这一点。

尽管存在上述问题，但Bayes方法的重要优点在于，找到Bayes规则（最小化（235）的规则）在原则上是可处理的。实际上，我们可以把（235）写成

Z Z Z Z Z Z Z Z

R（θ，d）dw（θ）=EθL（θ，d（X））dw（θ）=L（θ，d（X））d pθ（X）dw（θ）=L（θ，d（X））pθ（X）dμ（X）dw（θ）。

ΘXΘX

我们现在交换上面的积分顺序（这是允许的，因为损失函数是非负的）来得到

.

从简单的不等式

Z Z轴

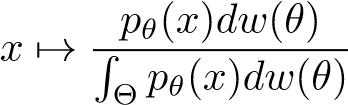
L（θ，d（x））pθ（x）dw（θ）≥inf L（θ，a）pθ（x）dw（θ）

Θa∈aΘ

对于每个x∈x，应该清楚的是，最小化（235）的规则是由

.

密度



简单地说，在X |θ∼pθ和θ∼w模型中，给定X=X的θ的后验密度。因此，我们得到了众所周知的事实，即Bayes规则最小化了损失函数的后验期望。例如，在平方误差损失L（θ，a）=kθ-ak2的情况下，贝叶斯规则只是后验分布的期望。

上述计算也给出了关于w的Bayes风险的精确表达式：

亚拜（236）

## 25.4极大极小法

在minimax方法中，我们根据决策规则在θ∈Θ上的最坏情况（上确界）风险来评估决策规则。换言之，我们的目标是选择supθ∈R（θ，d）较小的决策规则d。这种方法的优点是不需要选择特定的先验分布。缺点是它只关注最坏的情况，可能被认为过于悲观。然而，这是目前应用最广泛的最优性准则。

最小最大风险定义为

RMinimax:=inf supEθL（θ，d（X））

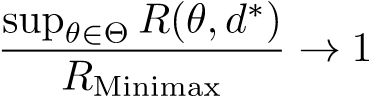
dθ∈Θ

其中，下确界接管了所有决策规则d。如果

.

发现极大极小估计在许多问题中是很困难的，所以通常是近似极小的。有两种常用的近似极小值的概念。它们是根据“样本大小”或“维度”参数n定义的，该参数存在于大多数决策问题中。具体地说，我们假设决策问题的所有成分（即Θ，A，L（θ，A）和Pθ）可能都依赖于样本大小或维度参数n，并且我们只对n的大值感兴趣。在这种情况下，我们有以下两个定义：

1. Sharp Asymptotically Minimaxity：我们认为决策规则D是尖锐渐近极大极小。

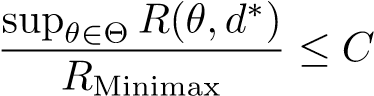


如n·····这相当于

supR（θ，d∗）=RMinimax（1+o（1））作为n→∞。

θ∈Θ

1. 速率最小化：我们认为决策规则D是速率极大极小



对于不依赖于n的常数C，这相当于

supR（θ，d∗）=O（RMinimax）作为n→∞θ∈Θ

现在考虑以下情况。假设我们构造了一个决策规则d\*（比如用M-估计方法），并且我们很好地理解了它的性能，因为我们在它的最高风险上有一个上界un，也就是说，我们知道

supR（θ，d∗）≤un。

θ∈Θ

那么我们如何证明d\*是minimax（在上述意义中的一个：minimax，sharp不对称minimax或rate minimax）显然，为了做到这一点，我们需要从下面绑定RMinimax实际上，如果我们证明了极小极大下界：

RMinimax≥n（237）

然后我们可以断言

1. 最小n，如果n＝unn。
2. 如果UN/N n＝1为n～ω，则为渐近渐近极大极小。
3. 如果un/`n=O（1）为n～ω，则速率极小。

当然，做到这一点的关键是能够证明极小极大下界（即形式（237）的界），我们现在将研究这个下界。

## 25.5极大极小下界

基本上只有一种方法可以证明极大极小风险的下界。这涉及到用Bayes风险从下面限制minimax风险。的确

Θ上每个概率测度w的RMinimax≥RBayes（w）。（238）

如前所述，Bayes风险RBayes（w）是一个更容易处理的对象（与RMinimax相比），它有精确的表达式（236）。

不等式（238）可以重写为

贝耶斯

其中对所有概率测度w取上确界。另一方面，很容易看出极小极大风险满足：

.

因此我们有

贝耶斯。

现在假设

,

然后我们将得到RMinimax=supw RBayes（w），这意味着从下面限制最小最大风险的唯一方法是通过Bayes风险来确定适当的先验w。显然，内界和上界不能总是互换的；例如，

RMinimax=inf supEθL（θ，d（X））6=supinf EθL（θ，d（X））=0。

dθ∈Θθ∈Θd

但是，在某些情况下，它们可以互换。保证交换的定理称为极大极小定理。文献中存在这样的一个极小极大定理，其中一个例子是如下（已知的某个时间点）：

定理25.6设K是向量空间X的凸子集，L是Hausdorff拓扑向量空间Y的紧凸子集。设f:K×L→R是这样一个函数

1. x 7→f（x，y）对于每个固定的y∈L是凸的。
2. y 7→f（x，y）是凹的，对于每个固定的x∈K是连续的。

那么

inf sup f（x，y）=sup inf（x，y）。

x∈K y∈ly∈lx∈K

这个定理可以用来验证（25.5）。为此，我们可以取K为所有决策规则d的类，L为所有概率测度w的集合我们需要

Z轴

（d，w）7→EθL（θ，d（X））dw（θ）

对于每个固定的W，D在D中是凸的，在W中凹的W是正确的，但是为了保证D中的凸性，我们需要切换到随机决策规则，并将风险的概念推广到随机化的决策规则。为了验证紧性假设，需要在所有概率测度的空间上放置一个拓扑这些可以在相当普遍的情况下完成，但细节是相当复杂的您可以查看Le Cam and Yang[15]或Le Cam[14]了解详细信息。

总结这一节，不等式（238）总是正确的通常，RMinimax=supw RBayes（w）so（238）是获得极大极小下界的唯一方法因为Bayes风险的确切表达式（236），我们有

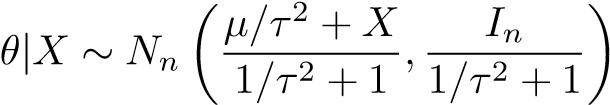
RMinimax≥RBayes（239）

我们将在续集中看到许多（239）的例子。一个简单的例子是下面我们可以使用（239）来证明精确的最小值。

例25.7（多元正态模型）考虑损失下由X∼Nn（θ，In）估计θ∈Rn的问题

.

在这种情况下，参数空间是Θ=Rn。估计量d（X）=X的风险等于1。结果表明，这是最小最大风险为此，设w为Rn上的正态分布，平均向量μ，协方差矩阵τ2In然后可以显式地计算Bayes风险RBayes（w）要了解这一点，请注意后验分布由



以便

亚拜。

这给了

每τ>0。

设τ····，得到RMinimax≥1。因为X在Rn上的最高风险不超过1，这证明了X是极大极小的。

# 26讲座26

在上节课中，我们看到了重要的极大极小下界：

RMinimax≥RBayes（w），对于Θ上的每个概率测度w。

在本课中，我们将讨论两个非平凡的例子，其中的上界可以用来建立自然估计的尖锐渐近最大值。第一个例子涉及稀疏正态均值估计第二个例子涉及在幂（L2范数）约束下的正态平均值的估计（这个结果被称为有限维Pinsker定理）。

## 26.1稀疏正态均值估计

考虑在平方误差损失下由观测Y∼Nn（θ，In）估计k-稀疏向量θ∈Rn的问题。该估计问题可以放在决策理论框架中，其中Θ是Rn中所有k-稀疏向量的集合，A=Rn和L（θ，A）=kθ-ak2Pθ也是概率测度Nn（θ，In）在本节中，我们假设稀疏水平k满足k=o（n）（即k/n→0作为n·····）。

从我们先前的结果，我们已经看到套索（这是相同的软阈值）估计器

具有调谐参数λ=p2log（n/k）满足

supEθL（θ，θˆλ）≤（2k对数（n/k））（1+o（1））为n→∞。

θ∈Θ

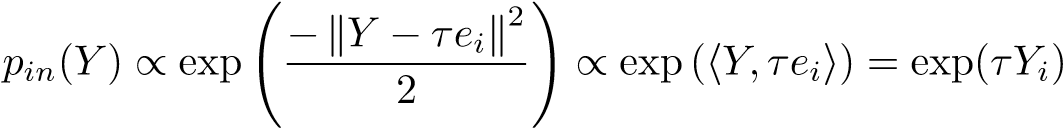
我们现在证明这个估计量是尖锐的渐近极大极小

RMinimax≥（2k对数（n/k））（1+o（1））为n→∞。（240）

以下论点摘自约翰斯顿[11，第8.6节]。

我们首先研究k=1（然后讨论一般的k=o（n））。对于k=1，参数Θ特别简单，由Rn中的1-稀疏向量组成这里自然先验是有限参数集{τe1，…，τen}上的一致先验，其中ei是通常的标准单位向量，τ≥0将被适当地选择（取决于n）我们先用w表示这个。

设后验分布用（p1n（Y），p2n（Y），…，pnn（Y））表示（即pin（Y）是与τe i相关的后验概率）。很容易看出



因此

对于i=1，…，n。

因此，后验均值（即Bayes估计量）由（τp1n（Y），…，τpnn（Y））给出。因此，在此之前的Bayes风险

亚拜。

通过用i=i对应的项替换上面的内和，我们得到了下界

亚拜

通过对称，上面的每一项都取相同的值，因此我们得到

亚拜。

为了计算上述期望值（注意，τ不是常数，但它随n变化），我们通过取Y1=τ+z1和Yi=zi作为i≥2，切换到标准高斯随机变量z1，…，zn然后我们需要计算：

.

我们将证明这一点

,

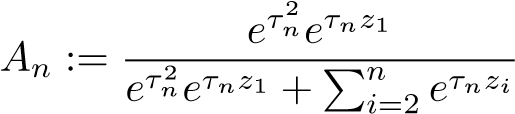
数量S收敛到1这就意味着

RMinimax≥RBayes

这将证明k=1所需的下限（240）。

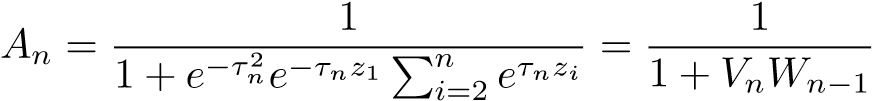
√

为了便于记法，让我们表示λn=2logn，这样exp（。为了证明τn：=λn−logλn的S=1+o（1），我们只需要证明随机变量序列



概率收敛到0注意，a正是τe1的后验概率（我们在真值为τe1的情况下工作）因此，概率收敛到零意味着τe1的后验概率（当真值为τe1时）变为零，这直观地意味着将错过尖峰。

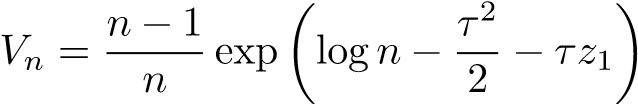
为了证明An→P 0，重写



哪里

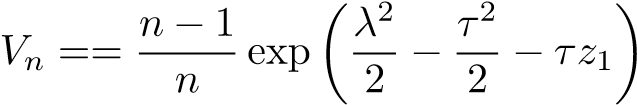
Vn=（n−1）e负极τ2/2e负极τz1和。

很容易看出，Vn在概率上收敛到＋∞要看到这个，写下



√

因为λ=2对数，



现在



因为λn-τn················（几乎肯定）。

因此，为了证明An在概率上变为0，我们只需要证明Wn-1在概率上变为1为简单起见，重新编制索引并调用n-1作为n。

.

这只是i.i.d平均值，是一个随机变量但是τn依赖于n，所以我们不能应用通常的弱大数定律但是我们可以使用下面的弱定律。

定理26.1。对于每个n，设Xnk，1≤k≤n为独立随机变量设bn>0，bn····。假设n→∞

1. ，和
2. .

设Sn：=Xn1+···+Xnn并置。那么

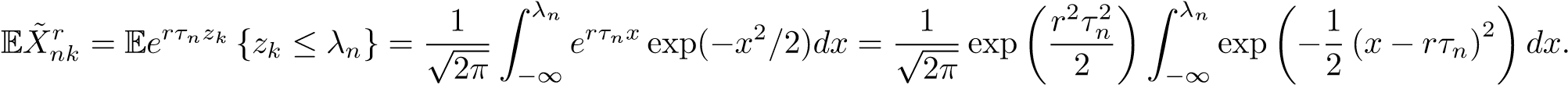
概率为n·····。

√

我们应用上述定理，Xnk：=eτnzk，bn=eτnλn（回忆一下λn：=2logn）。定理26.1中的第一个条件可以检查如下。注意{Xnk |≤bn}={zk≤λn}，这样

.

为了验证定理26.1中的第二个条件，我们需要计算r=2：



因此

EX∮nkr=er2τn2/2Φ（λn-rτn）。（241）

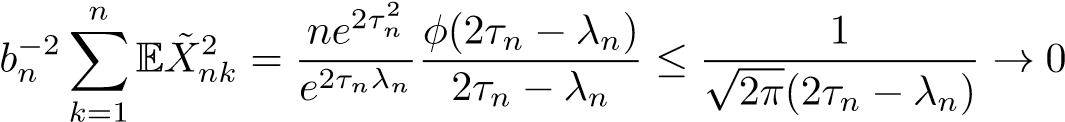
因此

.

观察到λn最终将小于2τn，从而使λn-2τn为负因此

.

因此

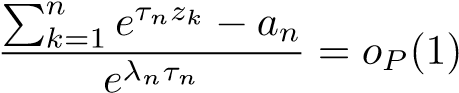


如n·····。

定理26.1的两个条件已经被证明，我们现在可以应用它了我们需要计算一个，我们可以简单地使用（241），r=1这会给

)（242个）

因此定理26.1给出



这和

.

因此

.

从（242）开始，我们有

.

因为λn-τn········和λn+τn·········

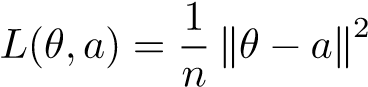
Wn=1+oP（1）

这就是我们想要证明的。这证明了k=1时的（240）。

为了证明（240）对于一般k=o（n），其思想是使用独立块先验将索引{1，…，n}分成大小为m=mn=bn/knc的kn块在每个块上，在k=1的情况下，使用单个峰值总的来说，先验知识将使这些kn块独立。由于独立性，Bayes风险加起来，得到了2k对数（n/k）（1+o（1））的整体下界。我们需要假设k/n→0，因为在每个块中，我们需要观测数mn去无穷大。这就完成了（240）的证明。

## 26.2功率约束下的正态平均估计（有限维Pinsker定理）

考虑由损失函数中的Y∼Nn（θ，In）估计向量θ∈Rn的问题



在以下约束下：

对于固定的c>0。（243）

设Θ表示满足上述约束（在信号处理和信息论中，该约束常被称为功率约束）的所有θ∈Rn的类设A=Rn，通常Pθ是Nn（θ，In）分布。估计量θˆ的风险由下式给出

.

幂约束（243）下θ的最佳候选估计是什么？最自然的估计量是Y在Θ上的投影。这是一个M-估计，可以用我们以前的技术来分析。实际上，在本例中，投影具有显式形式

如果

如果是

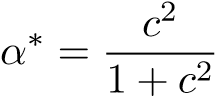
注意，这是一个非线性估计量结果表明，在这个问题中，对于适当的α>0，αY形式的简单线性估计有很好的性能这将在下面演示。首先，注意αY的风险由

.

在权力约束下（243），我们有

supR（θ，αY）=（1-α）2c2+α2。

θ∈Θ使右手边最小的α的值是



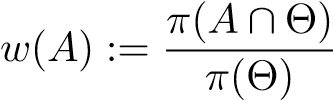
其风险由

.

结果表明，在这个问题中，线性估计量α\*Y是尖锐的渐近极大极小为了证明这一点，我们将在下面展示

（244）

证明（244）的第一步是在Θ上选择适当的先验w。对w来说最自然的选择可能是在Θ之前的制服但是，使用下面的previor稍微简单一些。设π表示Rn上的Nn（0，δ2c2In）分布，其中δ∈（0，1）。我们取w为π条件下的条件概率测度，即。，



对于Rn的Borel子集A。设θˆB（w）表示相对于前w的Bayes估计量。注意，由于w在凸集Θ上受支持，估计量θˆB（w）将属于具有概率1的Θ然后我们有了RBayes

贝耶斯。

由于π是Nn（0，δ2c2In）先验，其Bayes风险很容易用封闭形式计算为

亚拜

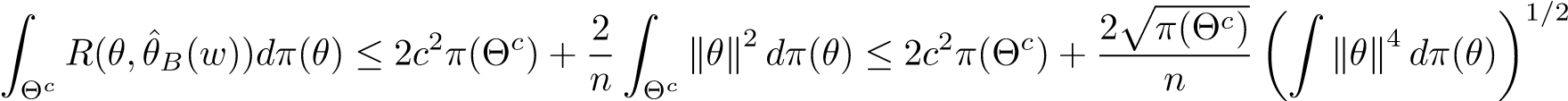
所以我们得到

亚拜。

现在注意，通过初等不等式ka−bk2≤2kak2+2kbk2，我们得到

.

这给了



通过柯西-施瓦兹不等式现在在π下，

以便

.

把东西放在一起，我们得到

!

亚拜

因为π（Θ）≤1我们完成了这个论证，我们只需要π（Θc）的下界为此，请注意

.

利用卡方浓度不等式

对于0≤t≤1，

我们得到

! 什么时候。

因此，对于0.5≤δ2≤1，我们得到

亚拜。

因此，我们

liminf RBayes n→∞

这意味着

.

由于δ2∈[0.5,1]是任意的，我们可以让δ2→1得到（244）。这就证明了线性估计量γy的尖锐渐近极大性。

从以上两个尖锐渐近最小轴的例子中，应该清楚的是，这些参数的关键是选择适当的先验W。而且一旦选择了W，参数通常是错综复杂的，因为我们甚至不想在n中丢失常数因子。

接下来我们将研究速率极小值的参数，在失去最小因子的同时，从下界中得到极大极小风险。这些参数简单得多，并且使用离散的prior（通常是参数空间的有限子集上的一致prior）下周我们将研究这些（与多假设检验问题的界有关的）。

# 27讲座27

我们将在今天的讲座中讨论多假设检验问题中的一致Bayes风险下限事实证明，在下一堂课中，一般的决策理论问题中的极小极大风险总是可以由下面的测试风险来限定的，这对于建立利率模型是非常有用的。

## 27.1多假设检验问题

假设我们观察到数据X取空间X中的值（通常，X可以是向量、矩阵、函数等）。对于X的分布，我们有以下N个假设：

H1:X∼P1，H2:X∼P2，…HN:X∼PN。

这里P1，…，PN是X上的概率测度，我们需要根据观测值X选择其中一个假设，检验T是X到{1，…，N}之间的任意函数。给定一个测试T，其类型i误差由i=1，…，n的Pi{T 6=i}定义。我们将根据i=1，…，n的测试类型i误差的平均值来评估测试。

具体来说，让

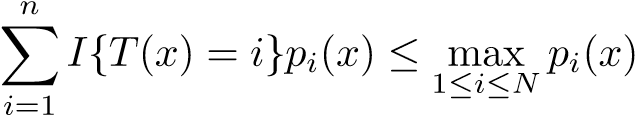
.

我们可以在一般的决策理论框架下，通过取Θ=A={1，…，N}和L（θ，A）=I{θ6=A}来处理这个问题。在这种情况下，R（T）将简单地是关于Θ之前离散一致性的试验T平均值的平均风险。

很容易看到（如下所示）最大似然检验给出了最小化r（t）的测试t。要了解这一点，让pi表示pi相对于公共支配度量μ的密度我们可以写

.

很容易看出，对于每个测试T和x∈x，我们有



用t（x）定义的最大似然检验实现相等：＝ARGMAX1±iπ（x）。这证明了T\*最小化R（T），并且

（245）

通常很难精确计算B（P1，…，PN）。我们将重点讨论B（P1，…，PN）的下界如后文所述，这些下界将产生一般决策理论问题中极大极小风险的下界。

为了激发B:=B（P1，…，PN）的下界，让我们首先为B提供一个直观的含义。因为B是测试问题中可能的最小平均误差（Bayes风险），所以应该清楚，在某种意义上，它度量概率度量P1，…，PN之间的分离程度。实际上，如果P1，…，PN彼此相距很远，那么测试问题应该更容易，而B将更小另一方面，如果P1，…，PN彼此接近，则测试问题将更难，B将更大还要注意，我们总是有Z

1≤maxpi（x）dμ（x）≤N

我

以便

.

而且，当P1= Fo.= PN时，很容易看出B（P1，…，PN）取最大可能的值1（1／N）。这是有意义的，因为当P1= Fo..Pn时，基于Xπ的I是不可能的，因此测试Bayes风险B（P1，…，PN）取其最大可能的值。

另一方面，当P1，…，Pn是相互奇异的时，B（P1，…，PN）取其最小值0（使得MaXi-Pi= P1+…PN几乎肯定W.R.T）。在这种情况下，可以根据X∼Pi完美地识别i，从而使测试Bayes风险处于其可能的最低值。

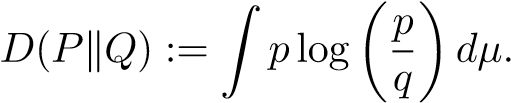
B（P1，…，PN）测量P1，…，PN之间分离度的直觉表明，我们可以通过其他自然量来限制它，以测量P1，…，PN的分离度或扩散度。对于实数a1…..aN，最自然的测量它们的分布的方法是它们的方差：

.

我们可以通过定义

)有。（246）

这里的D是指概率度量之间的差异/发散的概念（类似于实数之间的平方欧几里德距离）对D的各种选择是可能的，但最常见的是Kullback-Leibler散度给定两个概率测度P和Q，其密度分别为P和Q，相对于公共支配测度μ，它们之间的Kullback-Leibler散度定义为



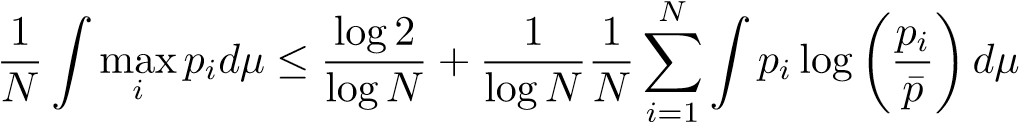
基于上述讨论，应该清楚的是，B（P1，…，PN）和I（P1，…，PN）之间应该存在某种联系（因为两者都在测量P1，…，PN之间的扩散或分离程度）以下引理描述了两者之间的简单关系。这在统计学文献中经常被用来证明极小极大下界，这里它被称为Fano不等式或Fano引理。事实上，这是Fano不等式的一种弱形式；今天晚些时候我们将描述一个强形式。

引理27.1。以下不等式适用于每N≥1和概率测度P1，…，PN：

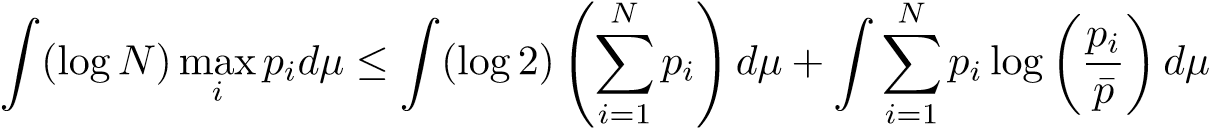
（247）

引理的证明27.1。这个优雅的证据是由凯普曼[13，135页]提供的。

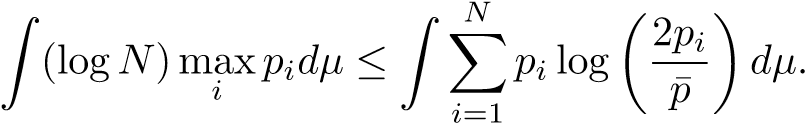
使用B（P1，…，PN）的公式（245）和I（P1，…，PN）的定义，很明显（247）等于



很容易看出，这进一步等同于（将上述两边乘以N logN，并使用R（Pi Pi）dμ=N），



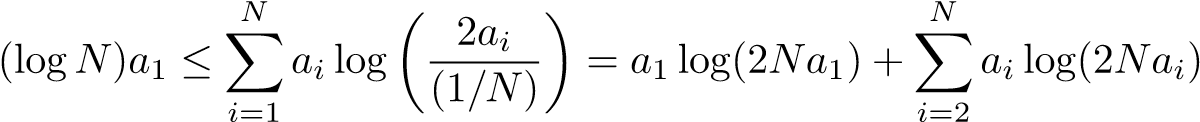
与



为了证明这一点，显然足以证明以下涉及非负实数的事实对于每一组非负实数a1，…，aN，以下不等式成立：

（248）

为了证明这种不等式，首先我们可以假设，在不损失一般性的情况下，＝1（因此A1，…，AN变为概率向量），A1＝MaXi-AI。不等式（248）等价于



重新排列为

N个

a1对数（2a1）+Xai对数（2Nai）≥0。

i=2

上述不等式是真的，因为（注意2N≥2（N−1））

.

这就完成了给出（247）的（248）的证明。

## 27.2相互信息

数量I（P1，…，PN）（在（246）中定义）被称为互信息。互信息是信息论中的一个术语。通常定义为一对随机变量Y和Z。形式上，Y和Z之间的互信息I（Y，Z）被定义为Y和Z的联合分布与Y和Z的边际分布乘积之间的Kullback-Leibler散度，

I（Y，Z）：=D（P（Y，Z）kPY×PZ）。

现在考虑两个随机变量Θ和X，使得Θ在{1，…，N}上均匀分布，并且给定的X的条件分布是π。很容易看出

（249）

因此，在（246）中定义的I（P1，…，PN）只是Θ和X之间的互信息。因此，I（P1，…，PN）被称为互信息项因此，Fano不等式给出了相互信息条件下Bayes风险的一个界。

以下关于I（P1，…，PN）的事实将在续集中有用。

引理27.2对于每N≥1和概率测度P1，…，PN，我们有

)（250个）

其中，对所有概率测度取下下确界证明很简单，基于以下身份：

其结果是：

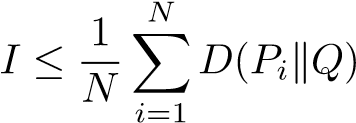
这里∏p=（p1+···+p N）/N是p′相对于μ的密度，q是q相对于μ的密度。

## 27.3稀疏正态均值估计的应用

设n≥1，设Pi为i=1，…，n的Nn（τei，In）分布。这里τ是一个依赖于n的正实数，ei是第i个位置为1，其他位置为0的向量。我们对B：=B（P1，…，Pn）以及它如何依赖于n和τ感兴趣Fano不等式（特别是不等式（247））给出

其中I：=I（P1，…，Pn）。

为了从这里得到一个显式的界，我们需要I的上界。这里引理27.2非常有用，因为它指出



对于每个概率测度Q，Q的一个自然选择是Q=Nn（0，In）。然后，我们得到（利用D（Nn（μ1，∑）kNn（μ2，∑））=（μ1～（2））T～-1（μ1～（2））/2）

.

这让我们可以推断

（251）

这给出了有趣的推论，例如：

对于τ=plogn。

√

然而，当τ接近λn:=2logn时，不等式（251）不足以产生任何非平凡的结果。例如，当τn：=λn−log（λn）时，则（251）不会给出任何有用的结果。但是，通过直接计算（如下所示），可以表明

=1，当τn=λn-对数（λn）时。（252）

这表明了利用Fano不等式获得B.to证明（252）的下界的一个重要弱点，首先要注意

哪里。

从中我们可以得到

哪里。

因此，如果z1，…，zn是独立的标准正态随机变量，那么

.

为了进一步将这个从下界绑定，我们需要从上面所做的绑定EMAXI E TAZI ZI。

方法（回忆λn=2logn）：

.

注意，上面的最后一个不等式（Mill的比率界）要求τ<λ因此我们得到

.

对于τ=λ-对数（λ），我们将λ-τ··········。

## 27.4通过数据处理不等式的Fano引理

数据处理不等式是关于Kullback-Leibler散度的一个标准事实它声明如下。假设P和Q是空间X上的两个概率测度，设Γ：X→Y为任意函数。

设PΓ-1表示在映射Γ下的概率测度P的图像，即。，

PΓ-1（A）：=P{j∈A}。

类似地定义QΓ-1数据处理的不平等表明

.

对于每一对概率测度P和Q以及每一个函数Γ，这都是正确的。我们在这里不会给出这个事实的证明（这是标准的，在很多地方都可以找到）我们将在引理27.1中概述Fano不等式的一个简单证明（实际上我们将通过数据处理不等式导出Fano不等式的一个比（247）更强的版本）。

考虑Fano不等式的设置，其中我们有N个概率测度P1，…，PN，在分别具有密度P1，…，PN的空间X上。考虑两个随机变量Θ和X，使得Θ在{1，…，N}上均匀分布，并且给定的X的条件分布为π。设P为Θ和X的联合分布，也设Q为Θ和X的边际分布的乘积的联合分布，我们在（249）中看到

I（P1，…，PN）=D（PkQ）。

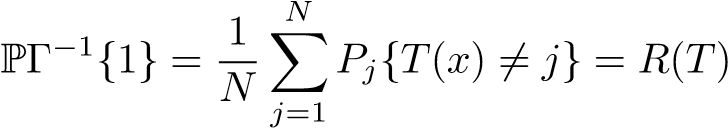
现在修复一个测试T，即T是X到{1，…，N}之间的函数然后我们将数据处理不等式应用于由

Γ（j，x）：=I{T（x）6=j}对于j∈{1，…，N}和x∈x。

数据处理不等式将给出

.

现在很容易看出



和

.

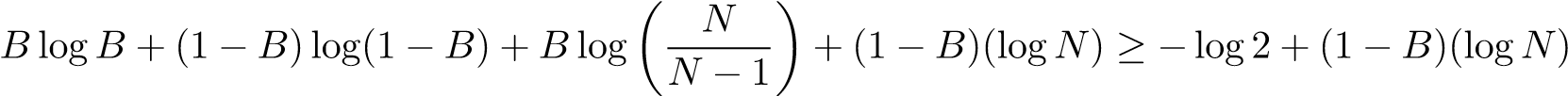
因此，我们已经证明，对于每一个测试T，

.

因为这对于每个测试T都是成立的，所以我们可以取t＝t（最大似然检验，它使r（t）在所有t上最小），使得r（t）＝b＝b（p1，…，pn）。这样就可以

（253）

这种不平等可以看作是法诺不平等的一个更强有力的版本。很容易证明（253）意味着（247）要看到这个，只要注意（253）的右边等于：



因为infx∈（0,1）（x log x+（1x）log（1x））=-log2和log（N/（N1））≥0。

这种证明Fano不等式（通过数据处理不等式）的一个重要优点是它推广到f-发散f-发散是概率测度之间的一类一般发散，包括作为特例的Kullback-Leibler发散。它们的定义如下。设f：（0，∞）→R是f（1）=0的凸函数然后很容易表明，存在以下限制（即使它们可以是+的）。假设P和Q是空间X上的两个概率测度，其密度P和Q相对于公共支配测度μ。P和Q之间的f-散度用Df（PkQ）表示，定义如下：

.

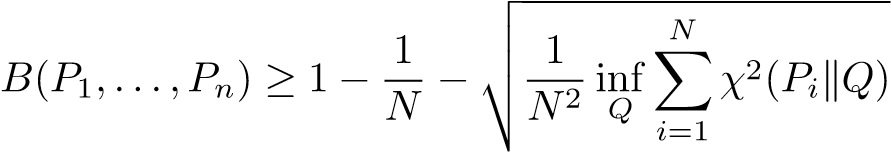
不同的f选择导致不同的具体分歧例如，KL散度对应于f（x）=xlogx，总变化距离对应于f（x）=| x−1 |/2，平方海林格距离对应于

对于f（x）=1-√x或f（x）=（√x−1）2/2，卡方散度对应于f（x）=x2−1，依此类推。

结果表明，每一个f-散度都满足数据处理不等式。利用这一点，可以证明Fano不等式对每一个f-散度的以下推广：

.

对于f的特定选择，这将导致B的更明确的下界。例如，对于f（x）=x2-1，可以得到



式中，对于f（x）=x2-1，x2（PkQ）=Df（PkQ）。见Gushchin[10]或Guntuboyina[9]和Chen等人[4]了解更多详细信息。

# 28讲座28

我们将研究通过FANE不等式证明速率极小结果的方法。第一步是通过测试问题中的Bayes风险，从一般决策理论问题中的极大极小风险的下界。

## 28.1经检验的最小最大下界

考虑具有参数空间Θ、作用空间a和非负损失函数L（θ，a）的一般决策理论设置。我们观察到分布属于{Pθ，θ∈Θ}族的数据X。

设F是Θ的有限子集。我们说F是由一个正实数分隔的

inf（L（θ1，a）+L（θ2，a））对于每个θ1，θ2∈F，θ16=θ2≥η。

a∈a

设w表示F上的一致先验，下面的引理表明，当F是η-分离时，Bayes风险RBayes（w）从下有界于（η/2）倍于与概率测度Pθ，θ∈F相对应的测试问题中的Bayes风险。

引理28.1。假设F是分开的。那么

亚拜（254）

请注意

RBayes）和

其中，下确界在RBayes（w）中的所有决策规则d上，在B（{Pθ，θ∈F}）中的所有测试T（即X到F的函数）上。

引理的证明28.1利用L（θ，a）≥（η/2）I{L（θ，a）≥η/2}，我们得到

亚拜。

对于每个决策规则d，我们现在按以下方式关联测试T。将T（x）定义为等于Th，如果存在一个Th f，使得L（Th，d（x））＜2（注意到，因为f是ε-分离的，最多存在一个Th f f，使得L（Th，D（x））＜ε/ 2）。如果没有这样的θ∈F，那么我们把T（X）取为F中的任意点

对于每个θ∈F。

从这里开始，不平等（254）紧接着出现。

在上一类中，我们证明了以下不等式（称为Fano不等式）

.

结合引理28.1和每个先验w的RMinimax≥RBayes（w）的事实，我们得到了以下极大极小下界：

对于每个有限F⊆（255）

我们将在下面看到这一界限的两个例子：稀疏正态均值估计和Lipschitz回归。使用（255）的主要挑战是选择适当的F。

在许多应用中，对于A上的一些伪度量d，A和L（θ，A）=d2（θ，A）

最小d（θ1，θ2）≥τ=？F是（τ2/2）-分离的。（256）

θ1，θ2∈F:θ16=θ2

这是因为对于每一个θ1，θ2∈F和θ16=θ2和a∈a，我们有

.

## 28.2稀疏正态均值估计

考虑由Y∼Nn（θ，in）估计1-稀疏向量θ∈Rn的平方欧氏损失问题这里Θ是Rn中所有1-稀疏向量的类，A=Rn，L（θ，A）是θ和A之间的平方欧氏距离，Pθ是Nn（θ，in）分布。

这里很自然地用√F={τe1，…，τen}来表示一些τ>0（稍后选择）

kτei-τejk=τ2，对于每个i 6=j，得出F是与η=τ2分开的η（使用（256））。不等式（255）给出

其中I=I（Pθ，θ∈F）上节课我们看到了

这给了

.

取τ2=logn时，RMinimax≥clogn，表示正常数√c（对于n大）这个结果很好

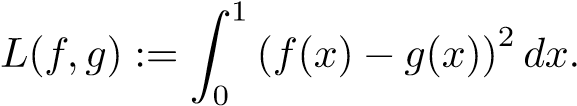
用La= 2Logn得到软阈值的速率极小。然而，它不够强，不能产生尖锐的渐近极大极小。

## 28.3利普希茨回归

设F表示在绝对值上被1和1-Lipschitz限制的所有函数F:[0，.1]→R的类。考虑从i.i.d观测值（X1，Y1），…，（Xn，Yn）估计f∈f的问题，其中

Xi∼Unif[0,1]和Yi | Xi∼N（f（Xi），1）。

我们取Θ=F，A为[0，1]上所有实值函数的类，并使用损失函数



我们用Pf表示（X1，Y1），…，（Xn，Yn）的联合分布注意，Pf在[0,1]n×Rn上具有以下密度：

.

我们对最小最大风险感兴趣：

.

可以证明，对于一个普适正常数C，RMinimax≤Cn-2/3。这可以通过研究F上的最小sqaures估计来实现（使用我们之前在类中看到的方法）我们还可以考虑更简单的核回归估计（例如，参见Tsybakov[23，第1章]）。在这里，我们将证明（使用（255））对于正常数c，RMinimax≥cn-2/3。这将特别证明，对于F中的函数估计，最小二乘估计是最小最大速率最优的。

主要的挑战是构造一个合适的F的有限子集F固定一个小的δ>0。对于长度δ[0,1]的闭子区间i，让Ti：i〔0，δ〕表示分段线性帐篷函数，它等于区间I的中点处的最大值δ（特别是i在i的中点处从i的0端到δ线性增加，然后在i的右端点线性减小到0）。现在考虑m间隔：

Ij：=[（j−1）δ，jδ]对于j=1，…，m：=b1/δc&1/δ。

我们现在在F中构造2m个函数，这些函数用τ∈{0,1}m索引，用{Fτ，τ∈{0,1}m}表示具体地说，对于每个τ∈{0,1}m，如果τj=1，我们定义fτ等于区间Ij上的帐篷函数TIj，如果τj=0，则定义fτ等于Ij上的零。此外，每个fτ在∪jIj外等于零。

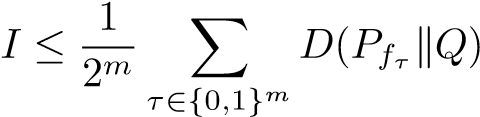
我们将对这个集合{Pfτ：τ∈{0,1}m}应用（255）。第一步是找到一个合适的值η，F：={Fτ，τ∈{0,1}m}分开。为此，用τ=6τ0固定τ，τ0∈{0,1}m然后τj 6=τj0，对于某些j∈{1，…，m}很容易看出

.

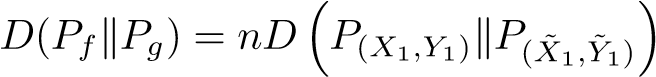
由此（从（256）可知，F与η和δ3分开不平等（255）那么说

其中I：=I（Pfτ，τ∈{0,1}m）。

我们下一步需要把我从上面绑起来。为此，我们将使用



Q=P0时（即f∏0对应的Pf）。注意，对于两个函数f，g，

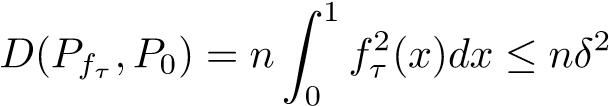


哪里

X1∼统一[0,1]；Y1 | X1∼N（f（X1），1）和X●1∼统一[0,1]；Y●1 | X●1∼N（g（X●1），1）我们也可以计算出

.

因此我们有



因此I≤nδ2。因此我们有

.

因为m≥c1/δ对于正常数c1，我们有

.

取δ3=（c1 log2）/（2n），我们得到了所有大n的RMinimax&n−1。然而，请注意，我们开始证明RMinimax&n−2/3，因此上述参数得出了RMinimax的次优下界。这个论点变弱的原因是在分离计算中事实上，我们利用了

对于每个τ，τ0∈{0,1}m，其中τ6=τ0。

当τ和τ0仅在一个坐标上不同时（即H（τ，τ0）：=Pj i{τj 6=τj0}正好等于1），也很容易看出上述界限紧到常数乘因子然而，一般来说，fτ和fτ0之间的L2距离取决于H（τ，τ0）很明显

)对于所有τ，τ0∈{0,1}m.（257）

Gilbert Varshamov Lemma（下一步证明）证明了{0,1} m的一个子集W的存在性，其基数为W＞EXP（m／8），并且对于每一个{（4，0）＞m／0，具有0＝T＝0的TAW 0 W。其思想是将（255）应用于F={Fτ：τ∈W}，而不是F={Fτ：τ∈{0,1}m}。上面的不等式（257）以及τ的H（τ，τ0）&m&（1/δ），τ，τ0∈W，τ6=τ0意味着F={Fτ：τ∈W}与η&δ2分开相互之间的信息边界保持不变。我们会得到

.

选择δ3=（c1 log2）/（2n）将给出所有大n的RMinimax&n−2/3。

## 28.4 Gilbert-Varshamov引理

引理28.2对于每m m 1，存在{0,1} m的一个子集W，其基数W＞Exp（m／8），使得每一个顶点的H（TAU，TAL 0）＞M／4，TAW 0＝TH0＝0。

证明。这里将使用以下基本概率界：

（258）

要证明（258），请注意（第一个等式遵循对称性）

.

取λ=log3，我们得到

.

现在设W是{0,1} m的最大子集，对于每一个顶点，H（TAU，TAU 0）＞M／4，TAW 0＝W，其中TAU 6＝TAI 0。这里的最大值意味着如果{0,1} M的任何其他元素被加到W，则分离条件将被违反。这意味着[BH（TAG，M/4）＝{0,1} M，其中B（TAm，M/4）：{{ω{0,1}M:H（TAG，ω））小于m/4 }h。

τ∈W

以便

2m=X | BH（τ，m/4）|≤| W | max | BH（τ，m/4）|（259）

τ∈W

τ∈W现在对于每个A⊆{0,1}m，我们有

|A|=2mP{（T1，…，Tm）∈A}，其中T1，…，Tm是i.i.d Ber（1/2）。

因此

2-m | BH（τ，m/4）|=P{（T1，…，Tm）∈BH（τ，m/4）}

（米）

=P XI{Ti 6=τi}≤m/4=P{Bin（m，1/2）≤m/4}≤exp（m/8）。

i=1

不等式（259）立即给出| W |≥exp（m/8），从而完成引理28.2的证明。

## 28.5避免F的显式构造的Yang-Barron方法

如前所述，应用（255）的主要困难是构造Θ的有限子集F。Yang和Barron[26]有一个避免F的显式构造的好主意，前提是可以得到关于包装和覆盖数Θ的结果下面是这个想法背后的细节。

对于ε＞0，假设n（ε，ε）是任意正实数，因此存在一个具有基数f（n，ε）的ε-分离有限子集F。将不等式（255）应用于这样一个F，我们得到

（260）

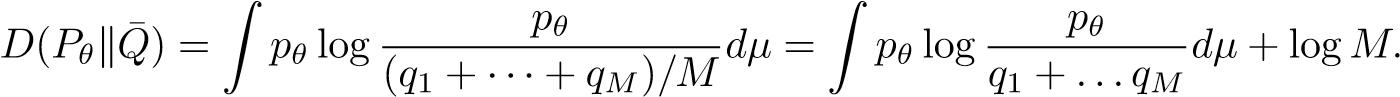
我们现在用下面的方法从上面对I=I（{Pθ，θ∈F}）进行定界我们知道

)（261）

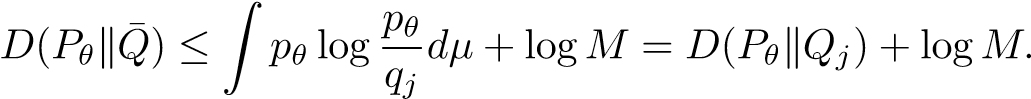
对于X上的每个概率测度Q，假设Q1，…，Q M是X上的任意概率测度，并应用（261），Q=Q’=（Q1+···+QM）/M，这给出了

.

现在对于每个θ∈F，如果q1，…，qM分别表示q1，…，qM w.r.tμ的密度（而pθ表示pθw.r.tμ的密度），那么



现在每1≤j≤M，我们就有q1+···+qM≥qj

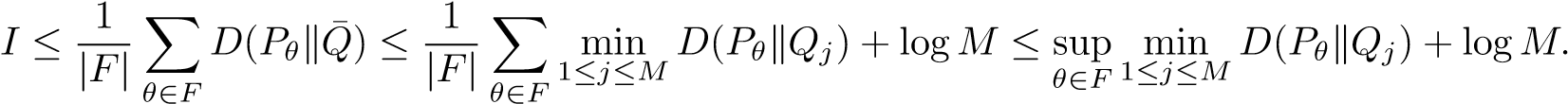


因为这对于每1≤j≤M是真的，我们推断

D（PθkQ’）≤min D（PθkQj）+logM。

1≤j≤M

因为这对于每个θ∈F都是正确的，所以我们得到



0和Θ的子集S，let）表示X上概率测度Q1，…，QM的最小数M，使得

.

上面的论证给出了

)为每一个人。

利用（260）中的这个界，我们得到，对于每个η>0和>0，

.

不等式Θ（这仅在以下情况下有用）给出

（262）

这个界的优点是它只依赖于Θ的性质。例如，在Lipschitz回归例子中，已知（第6课）在L2度量下F的包装数（这里F是[0,1]上所有实值函数的类，它们是1-Lipschitz且有1的边界）满足：

.

使用这个，很容易证明我们可以

和日志

在（262）。这给了

.

从这里取和η∼n−2/3（并适当调整基础常数），我们可以立即得到RMinimax&n−2/3。注意，此参数中没有使用有限子集F的显式构造（这项工作隐式地在包装数界限的证明中完成）。

更一般地，回想一下第6课中的平滑度等级Sdα。考虑从n i.i.d观测值（X1，Y1），…，（Xn，Yn）对函数f∈Sd，α的估计

Xi∼Unif[0,1]和Yi | Xi∼N（f（Xi），1）。

考虑[0,1]d上的积分L2损失函数：

Z轴

L（f，g）=（f（x）–g（x））2倍。

[0,1]天

在这种情况下，使用Sd，α的包装数满足的事实（在第6课中陈述）：

,

我们可以（这里的常数都取决于d）

和日志

在（262）这给了

.

取η=n−2α/（2α+d），我们得到

RMinimax≥n−2α/（2α+d）。

# 工具书类

1. Baraniuk，R.，M.Davenport，R.DeVore和M.Wakin（2008年）随机矩阵限制等距性质的一个简单证明。施工约28（3），253–263。
2. 比林斯利，P.（1968）概率测度的收敛性。纽约：威利。
3. Boucheron，S.，G.Lugosi和P.Massart（2013年）。集中不平等：一个非共同性的独立理论。牛津大学出版社。
4. Chen，X.，A.Guntuboyina和Y.Zhang（2016年）。关于Bayes风险下限。机器学习研究杂志17（219），1-58。
5. 达德利，R.M.（1989）真实分析和概率加州贝尔蒙特：沃兹沃斯。
6. 达德利，R.M.（1999）。一致中心极限定理剑桥大学出版社。
7. Feller，W.（1968年）概率论及其应用导论（第三版），第一卷。纽约：威利。
8. Ferguson，T.S.（1967年）数理统计：一种决策理论方法。波士顿：学术

按。

1. Guntuboyina，A.（2011年）。利用f-发散的极大极小风险下界及其应用关于信息理论的IEEE交易572386-2399。
2. Gushchin，A.A.（2003年）。关于Fano引理和极大极小风险的类似不等式。西奥。概率和数学。统计学家。67、29-41岁。
3. 约翰斯顿，I.M.（2017）高斯估计：序列和小波模型。手稿，2017年8月。可在
4. 加藤，K.（2017）经验过程理论课堂讲稿可在
5. Kemperman，J.H.B.（1969年）论信息的最佳传输速率。在概率论和信息论中斯普林格·维拉格。数学课堂讲稿，89，第126-169页。
6. Le Cam，L.（1986年）。统计决策理论中的渐近方法。纽约：斯普林格·维拉格。
7. Le Cam，L.和G.L.Yang（2000年）。统计学中的渐近线：一些基本概念（第2版）。斯普林渥弗拉格。
8. Mendelson，S.和R.Vershynin（2003年）。熵与组合维数。发明数学152，37-55。
9. Oymak，S.和B.Hassibi（2013年）。用于近去噪的尖锐MSE边界。计算数学基础，1-65。
10. 波拉德，D.（1984）。随机过程的收敛性。纽约：斯普林格。
11. Pollard，D.（1997年）。再看二次均值的可微性在D.Pollard，E.Torgersen和G.L.Yang（编辑），Lucien Le Cam的Festschrift，第305-314页。纽约：斯普林格·维拉格。
12. Pollard，D.（2001年）。测量理论概率的用户指南剑桥大学出版社。
13. Rudelson，M.和R.Vershynin（2006年）。随机过程和凸体截面的组合数学。数学年鉴，603-648。
14. Talagrand，M.（1996年）。重新审视独立。概率年鉴24，1-34。
15. Tsybakov，A.（2009年）。非参数估计导论斯普林格·维拉格。
16. Van der Vaart，A.（1998年）。渐近统计。剑桥大学出版社。
17. Van der Vaart，A.和J.A.Wellner（1996年）弱收敛与经验过程：在统计学中的应用。斯普林格·维拉格。
18. Yang，Y.和A.Barron（1999年）。极小极大收敛速度的信息论确定。统计年鉴271564-1599。