

翻译并解释以下内容，翻译不能有遗漏，解释请基础且详尽，因为我的数学基础不太好。

## 6 Frequentist statistics

### 6.1 Introduction

The approach to statistical inference that we described in Chapter 5 is known as Bayesian statistics. Perhaps surprisingly, this is considered controversial by some people, whereas the application of Bayes rule to non-statistical problems — such as medical diagnosis (Section 2.2.3.1), spam filtering (Section 3.4.4.1), or airplane tracking (Section 18.2.1) — is not controversial. The reason for the objection has to do with a misguided distinction between parameters of a statistical model and other kinds of unknown quantities.

Attempts have been made to devise approaches to statistical inference that avoid treating parameters like random variables, and which thus avoid the use of priors and Bayes rule. Such approaches are known as frequentist statistics, classical statistics or orthodox statistics. Instead of being based on the posterior distribution, they are based on the concept of a sampling distribution. This is the distribution that an estimator has when applied to multiple data sets sampled from the true but unknown distribution; see Section 6.2 for details. It is this notion of variation across repeated trials that forms the basis for modeling uncertainty used by the frequentist approach. By contrast, in the Bayesian approach, we only ever condition on the actually observed data; there is no notion of repeated trials. This allows the Bayesian to compute the probability of one-off events, as we discussed in Section 2.1. Perhaps more importantly, the Bayesian approach avoids certain paradoxes that plague the frequentist approach (see Section 6.6). Nevertheless, it is important to be familiar with frequentist statistics (especially Section 6.5), since it is widely used in machine learning.

好的，我会先逐句翻译，然后再用通俗、详细的方式解释。

## 翻译

### 6 频率学派统计学

#### 6.1 引言

我们在第 5 章中介绍的统计推断方法被称为贝叶斯统计。也许令人惊讶的是，这种方法被一些人认为是有争议的，而将贝叶斯法则应用在非统计问题上——例如医学诊断（第 2.2.3.1 节）、垃圾邮件过滤（第 3.4.4.1 节）、或飞机跟踪（第 18.2.1 节）——却并不具争议性。

这种反对的原因，和一个错误的区分有关：即把统计模型的参数和其他种类的未知量区分开来。

人们尝试设计一些统计推断的方法，来避免把参数当作随机变量，从而避免使用先验分布和贝叶斯法则。这样的方式被称为频率学派统计（frequentist statistics）、经典统计

（classical statistics）或正统统计（orthodox statistics）。这种方法不是基于后验分布，而是基于“抽样分布”的概念。所谓抽样分布，就是当我们从真实但未知的分布里，抽取多个数据集并分别计算估计量时，这个估计量本身会形成的分布；详细解释见第 6.2 节。正是这种“在重复实验中产生的变动”的思想，构成了频率学派处理不确定性的基础。

相比之下，在贝叶斯方法中，我们只考虑实际观察到的数据；不存在“重复实验”的概念。这让贝叶斯学派能够计算某些一次性事件的概率，正如我们在第 2.1 节讨论过的。或许更重要的是，贝叶斯方法能避免一些困扰频率学派的悖论（见第 6.6 节）。

尽管如此，熟悉频率学派统计学仍然很重要（尤其是第 6.5 节），因为它在机器学习中被广泛使用。

---

## 解释（通俗版，基础且详尽）

### 1. 两种统计学派

- **贝叶斯统计（Bayesian statistics）**

在贝叶斯方法中，我们把未知的参数（比如均值、方差）当成“随机的”，它们有概率分布。我们会先用“先验分布”表达对参数的信念，然后结合观测到的数据，通过贝叶斯公式算出“后验分布”，来更新我们对参数的认识。

👉 比喻：好像你先有一个模糊的猜测（先验），看到数据后进行修正，得到新的看法（后验）。

- **频率学派统计（Frequentist statistics）**

在频率学派方法里，参数是固定的，但未知；它不是“随机”的。随机性来自于数据本身（因为每次采样结果不同）。所以，频率学派关心的是“如果我在同样条件下反复做实验、重复采样，那么某个统计量（比如样本均值）的表现会怎样”。这种“在无限次重复实验下，统计量会形成一个分布”的东西，就叫“抽样分布”。

👉 比喻：好像你在扔骰子，骰子本身（参数）是固定的，但因为你多次掷骰子，每次结果不同，所以形成了概率分布。

---

## 2. 为什么有人觉得贝叶斯有争议

- 在非统计问题（比如垃圾邮件识别、医学诊断），大家用贝叶斯定理都很自然，不会觉得奇怪。
  - 但在统计建模里，有些人认为“参数不是随机的”，不能给它们设概率分布，所以他们不喜欢贝叶斯方法。
- 

## 3. 关键区别

- **频率学派**：不考虑“先验”，不把参数当随机变量，而是通过假设“无限次重复实验”来研究估计值的波动性，用“抽样分布”表达不确定性。
- **贝叶斯学派**：只关心你“实际观察到的数据”，把参数也看作随机变量，可以直接算“一次性事件的概率”。

例如：

- 频率学派的问题是：如果我反复做 100 次抛硬币实验，每次算“正面出现的比例”，这个比例的分布是什么？
  - 贝叶斯学派的问题是：在我手头这一次实验中，看到 60% 的正面，硬币偏向正面的概率是多少？
- 

## 4. 为什么要学频率学派

即使贝叶斯方法在逻辑上更自然、更能避免悖论，但频率学派方法在机器学习、科学研  
究、工业实践里依然很常见。比如大家熟悉的 **p 值**、**置信区间**、**假设检验** 都是频率学派的  
产物。

所以即使偏好贝叶斯，也要理解频率学派的概念。

---

要不要我再帮你用一个“**扔硬币的例子**”，来分别演示频率学派和贝叶斯学派在同一个问题  
上的不同思路？这样你可能会更直观。



已上传的图片

翻译并解释以下内容，翻译不能有遗漏，解释请基础且详尽，因为我的数学基础不太好。

Figure 6.1 A bootstrap approximation to the sampling distribution of  $\hat{\theta}$  for a Bernoulli distribution. We use  $B = 10,000$  bootstrap samples. The  $N$  data cases were generated from  $\text{Ber}(\theta = 0.7)$ . (a) MLE with  $N = 10$ . (b) MLE with  $N = 100$ . Figure generated by `bootstrapDemoBer`.

## 6.2 Sampling distribution of an estimator

In frequentist statistics, a parameter estimate  $\hat{\theta}$  is computed by applying an estimator  $\delta$  to some data  $D$ , so  $\hat{\theta} = \delta(D)$ . The parameter is viewed as fixed and the data as random, which is the exact opposite of the Bayesian approach. The uncertainty in the parameter estimate can be measured by computing the sampling distribution of the estimator. To understand this concept, imagine sampling many different data sets  $D(s)$  from some true model,  $p(\cdot|\theta^*)$ , i.e., let  $D(s) = \{x(s)_i\}_{i=1}^N$ , where  $x(s)_i \sim p(\cdot|\theta^*)$ , and  $\theta^*$  is the true parameter. Here  $s = 1: S$  indexes the sampled data set, and  $N$  is the size of each such dataset. Now apply the estimator  $\hat{\theta}(\cdot)$  to each  $D(s)$  to get a set of estimates,  $\{\hat{\theta}(D(s))\}$ . As we let  $S \rightarrow \infty$ , the distribution induced on  $\hat{\theta}(\cdot)$  is the sampling distribution of the estimator. We will discuss various ways to use the sampling distribution in later sections. But first we sketch two approaches for computing the sampling distribution itself.

### 6.2.1 Bootstrap

The bootstrap is a simple Monte Carlo technique to approximate the sampling distribution. This is particularly useful in cases where the estimator is a complex function of the true parameters. The idea is simple. If we knew the true parameters  $\theta^*$ , we could generate many (say  $S$ ) fake

datasets, each of size  $N$ , from the true distribution,  $xs[i] \sim p(\cdot|\theta^*)$ , for  $s = 1: S$ ,  $i = 1: N$ . We could then compute our estimator from each sample,  $\hat{\theta}[s] = f(xs[1:N])$  and use the empirical distribution of the resulting samples as our estimate of the sampling distribution. Since  $\theta$  is unknown, the idea of the parametric bootstrap is to generate the samples using  $\hat{\theta}(D)$  instead. An alternative, called the non-parametric bootstrap, is to sample the  $xs[i]$  (with replacement) from the original data  $D$ , and then compute the induced distribution as before. Some methods for speeding up the bootstrap when applied to massive data sets are discussed in (Kleiner et al. 2011). Figure 6.1 shows an example where we compute the sampling distribution of the MLE for a Bernoulli using the parametric bootstrap. (Results using the non-parametric bootstrap are essentially the same.) We see that the sampling distribution is asymmetric, and therefore quite far from Gaussian, when  $N = 10$ ; when  $N = 100$ , the distribution looks more Gaussian, as theory suggests (see below). A natural question is: what is the connection between the parameter estimates  $\hat{\theta}[s] = \hat{\theta}(xs[1:N])$  computed by the bootstrap and parameter values sampled from the posterior,  $\theta[s] \sim p(\cdot|D)$ ? Conceptually they are quite different. But in the common case that the prior is not very strong, they can be quite similar. For example, Figure 6.1(c-d) shows an example where we compute the posterior using a uniform Beta(1,1) prior, and then sample from it. We see that the posterior and the sampling distribution are quite similar. So one can think of the bootstrap distribution as a “poor man’s” posterior; see (Hastie et al. 2001, p235) for details. However, perhaps surprisingly, bootstrap can be slower than posterior sampling. The reason is that the bootstrap has to fit the model  $S$  times, whereas in posterior sampling, we usually only fit the model once (to find a local mode), and then perform local exploration around the mode. Such local exploration is usually much faster than fitting a model from scratch.

### 6.2.2 Large sample theory for the MLE

In some cases, the sampling distribution for some estimators can be computed analytically. In particular, it can be shown that, under certain conditions, as the sample size tends to infinity, the sampling distribution of

the MLE becomes Gaussian. Informally, the requirement for this result to hold is that each parameter in the model gets to “see” an infinite amount of data, and that the model be identifiable. Unfortunately this excludes many of the models of interest to machine learning. Nevertheless, let us assume we are in a simple setting where the theorem holds. The center of the Gaussian will be the MLE  $\hat{\theta}$ . But what about the variance of this Gaussian? Intuitively the variance of the estimator will be (inversely) related to the amount of curvature of the likelihood surface at its peak. If the curvature is large, the peak will be “sharp”, and the variance low; in this case, the estimate is “well determined”. By contrast, if the curvature is small, the peak will be nearly “flat”, so the variance is high. Let us now formalize this intuition. Define the score function as the gradient of the log likelihood evaluated at some point  $\hat{\theta}$ :

$$s(\hat{\theta}) = \nabla \log p(D|\theta)|_{\hat{\theta}} \quad (6.1)$$

Define the observed information matrix as the gradient of the negative score function, or equivalently, the Hessian of the NLL:

$$J(\hat{\theta}(D)) = -\nabla s(\hat{\theta}) = -\nabla^2 \theta \log p(D|\theta)|_{\hat{\theta}} \quad (6.2)$$

In 1D, this becomes

$$J(\hat{\theta}(D)) = -d \frac{d}{d\theta} \log p(D|\theta)|_{\hat{\theta}} \quad (6.3)$$

This is just a measure of curvature of the log-likelihood function at  $\hat{\theta}$ .

Since we are studying the sampling distribution,  $D = (x_1, \dots, x_N)$  is a set of random variables. The Fisher information matrix is defined to be the expected value of the observed information matrix:

$$IN(\hat{\theta}|\theta^*) = E_{\theta^*} [J(\hat{\theta}|D)] \quad (6.4)$$

where  $E_{\theta^*} [f(D)] = \frac{1}{N} \sum_{i=1}^N f(x_i)$  is the expected value of the function  $f$  when applied to data sampled from  $\theta^*$ . Often  $\theta^*$ , representing the “true parameter” that generated the data, is assumed known, so we just write  $IN(\hat{\theta}) = IN(\hat{\theta}|\theta^*)$  for short. Furthermore, it is easy to see that  $IN(\hat{\theta}) = NI_1(\hat{\theta})$ , because the log-likelihood for a sample of size  $N$  is just  $N$  times “steeper” than the log-likelihood for a sample of size 1. So we can drop the 1 subscript and just write  $I(\hat{\theta}) = I_1(\hat{\theta})$ . This is the notation that is usually used. Now let  $\hat{\theta} = \hat{\theta}_{MLE}(D)$  be the MLE, where  $D \sim \theta^*$ . It can be shown that

$$\hat{\theta} \sim N(\theta^*, IN(\theta^*)^{-1}) \quad (6.5)$$

as  $N \rightarrow \infty$  (see e.g., (Rice 1995, p265) for a proof). We say that the sampling distribution of the MLE is asymptotically normal. What about the variance of the MLE, which can be used as some measure of confidence in the MLE?

Unfortunately,  $\theta^*$  is unknown, so we can't evaluate the variance of the sampling distribution. However, we can approximate the sampling distribution by replacing  $\theta^*$  with  $\hat{\theta}$ . Consequently, the approximate standard errors of  $\hat{\theta}_k$  are given by

$$se_k = \sqrt{N(\hat{\theta}) - 1} / \sqrt{N}$$
 (6.6)

For example, from Equation 5.60 we know that the Fisher information for a binomial sampling model is

$$I(\theta) = N\theta(1 - \theta)$$
 (6.7)

So the approximate standard error of the MLE is

$$se = \sqrt{N(\hat{\theta})} = \sqrt{N} / \sqrt{N\theta(1 - \theta)} = \sqrt{N} / \sqrt{N\theta(1 - \theta)}$$
 (6.8)

where  $\hat{\theta} = \bar{x}$ . Compare this to Equation 3.27, which is the posterior standard deviation under a uniform prior.

已思考 33s

下面先逐字逐句翻译，随后给出通俗而详细的解释（含小例子）。我会保留原式子/符号；个别地方原文有轻微排版瑕疵，我会在解释里说明其含义。

## 翻译

**图 6.1** 对伯努利分布中  $\hat{\theta}$  的抽样分布的自助法 (bootstrap) 近似。我们使用  $B = 10,000$  个自助样本。 $N$  个数据样本来自  $Ber(\theta = 0.7)$ 。  
 (a)  $N = 10$  时的极大似然估计 (MLE)。  
 (b)  $N = 100$  时的极大似然估计。图由 `bootstrapDemoBer` 生成。

## 6.2 估计量的抽样分布

在频率学派统计中，参数估计  $\hat{\theta}$  是通过把一个估计器  $\delta$  作用到某些数据  $D$  上计算得到的，即  $\hat{\theta} = \delta(D)$ 。参数被视为固定的，而数据被视为随机的，这与贝叶斯方法恰好相反。参数估计的不确定性可以通过计算估计器的抽样分布来度量。要理解这一概念，设想从某个真实但未知的模型  $p(\cdot | \theta^*)$  中抽取许多不同的数据集  $D^{(s)}$ ，即令  $D^{(s)} = \{x_i^{(s)}\}_{i=1}^N$ ，其中  $x_i^{(s)} \sim p(\cdot | \theta^*)$ ，并且  $\theta^*$  是真实参数。这里  $s = 1 : S$  给出了被抽取的数据集的索引，而  $N$  是每个数据集的大小。现在把估计器  $\hat{\theta}(\cdot)$  作用到每个  $D^{(s)}$  上以得到一组估计  $\{\hat{\theta}(D^{(s)})\}$ 。当我们让  $S \rightarrow \infty$  时， $\hat{\theta}(\cdot)$  上所诱导的分布就是该估计器的抽样分布。我们将在后续小节讨论使用抽样分布的各种方法。但首先我们概述两种计算抽样分布本身的方法。

### 6.2.1 自助法 (Bootstrap)

自助法是一种用来近似抽样分布的简单蒙特卡洛技术。当估计器是**真实参数**的复杂函数时，这尤其有用。思想很简单：如果我们**知道**真实参数  $\theta^*$ ，就可以从真实分布中生成许多（比如  $S$  个）“假”数据集，每个大小为  $N$ ：

$$x_i^{(s)} \sim p(\cdot | \theta^*), s = 1 : S, i = 1 : N.$$

然后我们可以对每个样本计算估计量  $\hat{\theta}_s = f(x_{1:N}^{(s)})$ ，并把得到的样本的经验分布作为抽样分布的估计。

由于  $\theta$  是未知的，**参数化自助法 (parametric bootstrap)** 的想法是改为使用  $\hat{\theta}(D)$  来生成样本。另一种方法称为**非参数自助法 (non-parametric bootstrap)**，即从原始数据  $D$  中**有放回地**抽取  $x_i^{(s)}$ ，然后像之前一样计算所诱导的分布。关于在处理**海量数据集**时加速自助法的一些方法可见 (Kleiner 等, 2011)。

图 6.1 展示了一个例子：我们用参数化自助法来计算伯努利分布的 MLE 的抽样分布。（使用非参数自助法得到的结果基本相同。）我们看到，当  $N = 10$  时，抽样分布是不对称的，因此离高斯分布相当远；当  $N = 100$  时，分布看起来更像高斯分布，正如理论所暗示的那样（见下面）。一个自然而然的问题是：用自助法计算得到的参数估计  $\hat{\theta}_s = \hat{\theta}(x_{1:N}^{(s)})$  与从后验中采样得到的参数值  $\theta^{(s)} \sim p(\theta | D)$  有什么联系？从概念上说，它们是相当不同的。但在**先验并不强**的常见情形下，它们可能相当相似。例如，图 6.1(c-d) 展示了一个例子：我们使用**均匀的** Beta(1,1) 先验来计算后验，然后从中采样。可以看到，后验与抽样分布相当相似。因此，可以把自助法分布看作是“穷人的后验”；细节见 (Hastie 等, 2001, 第 235 页)。然而，也许有些出人意料的是，自助法可能比后验采样**更慢**。原因在于：自助法必须拟合模型  $S$  次，而在后验采样中，我们通常只拟合**一次**（找到一个局部极大），然后在该极大附近做**局部探索**。这种局部探索通常比从头开始拟合模型快得多。

### 6.2.2 MLE 的大样本理论

在某些情形下，某些估计器的抽样分布可以**解析**计算。具体而言，可以证明在一定条件下，当样本量趋于无穷大时，MLE 的抽样分布变为高斯分布。非正式地说，使该结果成立的要求是：模型中的每个参数都能“看到”无限量的数据，且模型是**可识别的**。不幸的是，这排除了许多机器学习中感兴趣的模型。尽管如此，我们先假设处于定理成立的一个简单场景中。此时高斯分布的中心是 MLE  $\hat{\theta}$ 。那么这个高斯的**方差**是什么呢？直观上，估计量的方差将（反比地）与似然函数在其峰值处的**曲率**有关。如果曲率很大，峰值会很“尖锐”，方差就低；在这种情况下，估计是“确定得较好”的。相反，如果曲率很小，峰值几乎“平坦”，方差就高。现在把这种直觉形式化。

把\*\*得分函数 (score) \*\*定义为在某一点  $\hat{\theta}$  处对对数似然的梯度：

$$s(\hat{\theta}) = \nabla_{\theta} \log p(D | \theta) \Big|_{\hat{\theta}} \quad (6.1)$$

把**观测信息矩阵 (observed information)** 定义为**负得分函数的梯度**，或等价地，\*\*负对数似然 (NLL)\*\* 的 Hessian：

$$J(\hat{\theta}(D)) = -\nabla_S(\hat{\theta}) = -\nabla_{\theta}^2 \log p(D|\theta) \Big|_{\hat{\theta}} \quad (6.2)$$

在一维情形，这变为

$$J(\hat{\theta}(D)) = -\frac{d^2}{d\theta^2} \log p(D|\theta) \Big|_{\hat{\theta}} \quad (6.3)$$

这就是在  $\hat{\theta}$  处对数似然函数曲率的度量。

因为我们研究的是抽样分布， $D = (x_1, \dots, x_N)$  是一组随机变量。**费舍尔信息矩阵** (Fisher information) 被定义为观测信息矩阵在数据分布下的期望：

$$I_N(\hat{\theta} | \theta^{*}) = E_{\theta^{*}}[J(\hat{\theta} | D)] \quad (6.4)$$

其中  $E_{\theta^{*}}[f(D)] = \frac{1}{N} \sum_{i=1}^N f(x_i) p(x_i | \theta^{*})$  表示当数据从  $\theta^{*}$  生成时函数  $f$  的期望。（常把  $\theta^{*}$  ——生成数据的“真实参数”——视为已知，因此我们简写为  $I_N(\hat{\theta}) = I_N(\hat{\theta} | \theta^{*})$ 。）此外，很容易看出  $I_N(\hat{\theta}) = N I_1(\hat{\theta})$ ，因为样本量为  $N$  的对数似然比样本量为 1 的对数似然“陡峭  $N$  倍”。因此我们可以去掉下标 1，直接写  $I(\hat{\theta}) = I_1(\hat{\theta})$ 。这也是通常使用的记号。

现在令  $\hat{\theta} = \hat{\theta}_{\text{mle}}(D)$  是 MLE，且  $D \sim \theta^{*}$ 。可以证明

$$\hat{\theta} \xrightarrow{D} N(\theta^{*}, I_N(\theta^{*})^{-1}) \quad (6.5)$$

当  $N \rightarrow \infty$ （证明见如 Rice 1995, 第 265 页）。我们称 MLE 的抽样分布渐近正态。那么 MLE 的方差如何呢？它可以用作对 MLE 置信度的一种度量。不幸的是  $\theta^{*}$  未知，因此我们无法直接计算抽样分布的方差。不过，我们可以用  $\hat{\theta}$  替换  $\theta^{*}$  来近似抽样分布。因此， $\hat{\theta}_k$  的近似标准误 (standard error) 为

$$\text{se}_{\hat{\theta}} = (I_N(\hat{\theta})^{-1})_{kk}^{1/2} \quad (6.6)$$

例如，由式 (5.60) 我们知道二项采样模型的费舍尔信息为

$$I(\theta) = \frac{1}{\theta(1-\theta)} \quad (6.7)$$

于是 MLE 的近似标准误是

$$\text{se}_{\hat{\theta}} = \frac{1}{\sqrt{I_N(\hat{\theta})}} = \frac{1}{\sqrt{NI(\hat{\theta})}} = \sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{N}} \quad (6.8)$$

其中  $\hat{\theta} = \frac{1}{N} \sum_i X_i$ 。将其与式 (3.27) 比较，后者是在均匀先验下的后验标准差。

# 解释（基础且详尽）

## 1) 图 6.1 在讲什么？

- 场景：每个样本都是**伯努利** (0/1) 数据，真实成功率  $\theta = 0.7$ 。
- 我们想知道：用**极大似然估计 (MLE)** 得到的  $\hat{\theta}$  (对伯努利/二项，MLE 就是样本成功比例) 在重复抽样下会怎样波动——这就叫**抽样分布**。
- 直接反复做真实实验很贵，所以用 **bootstrap** 来模拟：
  1. 先用手里这份数据算出  $\hat{\theta}$ ；
  2. 再用  $\hat{\theta}$  当作“近似真实参数”，反复生成很多“假数据”（本图  $B=10,000$  次）；
  3. 每次都重新算一个  $\hat{\theta}$ ，把这些  $\hat{\theta}$  画直方图，就得到了抽样分布的近似。
- 左图 (a)  $N=10$ ：分布**不对称**、离正态挺远；右图 (b)  $N=100$ ：分布更接近**正态**（高斯）。这符合“大样本正态性”——样本越大， $\hat{\theta}$  的分布越像正态，越“集中”。

小提示：图标标题里出现的 “mle=...， se=...” 是那次仿真的输出示例，具体数值依赖于那一批数据；理解这张图的重点是**形状**：小样本偏、扁、歪；大样本更像钟形。

## 2) 什么是“抽样分布”与“频率学派”的不确定性？

- 在频率学派里，**参数是固定的**（只是我们不知道），**数据是随机的**（因为每次抽样不同）。
- 不确定性来自“如果我**重复**做同样的实验，我的估计会怎么变？”——把这种在重复实验下的估计结果收集起来形成的分布，就叫**抽样分布**。
- 这与贝叶斯不同：贝叶斯把参数也当随机的，用先验+数据得到后验；频率学派不这么做，它用“重复实验的波动”来衡量不确定性。

## 3) 两种 Bootstrap：参数化 vs 非参数化

- **参数化自助法**：先用当前数据算  $\hat{\theta}$ ，再按某个模型（这里是伯努利( $\hat{\theta}$ )）去模拟数据。
- **非参数自助法**：直接从原数据里有放回抽样，省去“先假定一个模型”。
- 在本例里，两者效果“本质相同”（所以图里只展示了参数化版本）。

## 4) Bootstrap 与贝叶斯后验的关系

- 从概念上：自助法的  $\{\hat{\theta}^{(s)}\}$  来自**重复拟合**；贝叶斯的  $\{\theta^{(s)}\}$  来自**后验抽样**。
- 当先验很弱（比如 Beta(1,1)），这两种分布经常**很相似**。因此有人把 bootstrap 的分布戏称为“**穷人的后验**”。
- 但速度上可能相反：bootstrap 需要**拟合 S 次**；而贝叶斯往往只需**一次拟合找到高峰**，再在附近**局部探索**，通常更快。

## 5) 为什么“大样本下 MLE 近似正态”？

- 直觉：在正确模型、足够数据、可识别等条件下， $\hat{\theta}$  会围绕真实  $\theta^*$  摆摆，且按高斯方式集中；这叫渐近正态性：

$$\hat{\theta} \approx N(\theta^*, I_N(\theta^*)^{-1})$$

- 这里  $I_N$  是费舍尔信息：可以把它理解为“数据里关于参数的信息量”；信息越多，不确定性越小。

## 6) 曲率、得分、信息：怎么量化“不确定性”？

- 得分  $s(\hat{\theta})$ ：对数似然在  $\hat{\theta}$  的梯度。
- 观测信息  $J(\hat{\theta})$ ：对数似然在  $\hat{\theta}$  的负二阶导（负 Hessian）。它本质就是曲率——峰越尖，曲率越大。
- 费舍尔信息  $I_N$ ：把  $J(\hat{\theta})$  对“在真实参数  $\theta^*$  下可能出现的数据”取期望。结果与样本量  $N$  成正比 ( $I_N = NI_1$ )。
- 渐近方差是  $I_N(\theta^*)^{-1}$ ，所以曲率越大（信息越多） $\rightarrow$  方差越小，估计更稳定。

## 7) 把式子落到伯努利/二项上（最常见的例子）

- 二项/伯努利的费舍尔信息（每个样本）：

$$I(\theta) = \frac{1}{\theta(1-\theta)}$$

- 总信息（N 个样本）： $I_N(\theta) = N I(\theta) = \frac{N}{\theta(1-\theta)}$
- 用  $\hat{\theta}$  代替  $\theta^*$  得到\*\*标准误（SE）\*\*近似：

$$se_{\hat{\theta}} \approx \sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{N}}$$

这也是大家熟悉的“样本比例的标准误”。

### 数值直观

- 若  $N = 10$  且某次样本给出  $\hat{\theta} = 0.9$ ，  
 $se \approx \sqrt{0.9 \times 0.1/10} \approx 0.095$ （分布很宽）。
  - 若  $N = 100$  且  $\hat{\theta} \approx 0.7$ ，  
 $se \approx \sqrt{0.7 \times 0.3/100} \approx 0.046$ （分布更窄）。
- 这与图 6.1 的直观完全一致：样本多，波动小，柱子更集中、更像钟形曲线。

## 8) 与贝叶斯结果怎么对比？

- 在均匀先验  $Beta(1,1)$  下， $\theta$  的后验是  $Beta(1 + \text{成功数}, 1 + \text{失败数})$ 。它的标准差与上面的频率学派 SE 形式非常相近。

- 这解释了为什么在先验很弱时，图中“后验分布”会与“bootstrap 近似的抽样分布”很像。

## 小结

- **抽样分布**: 频率学派用“重复试验的波动”来刻画估计的不确定性。
- **Bootstrap**: 用仿真来近似抽样分布；小样本时分布常常偏斜，样本大了更接近正态。
- **大样本理论**: 在合适条件下，MLE 渐近正态，方差约为费舍尔信息的逆。
- **伯努利/二项**:  $se \approx \sqrt{\hat{\theta}(1 - \hat{\theta})/N}$ 。
- **与贝叶斯**: 先验弱时，bootstrap 分布与贝叶斯后验常常类似；但在计算上，bootstrap 可能更慢。

如果你愿意，我可以用几行“掷硬币/点击率”的模拟代码，把图 6.1 的效果重现给你看，并把 SE 的计算对上数值。

翻译并解释以下内容，翻译不能有遗漏，解释请基础且详尽，因为我的数学基础不太好。

### 6.3 Frequentist decision theory

In frequentist or classical decision theory, there is a loss function and a likelihood, but there is no prior and hence no posterior or posterior expected loss. Thus there is no automatic way of deriving an optimal estimator, unlike the Bayesian case. Instead, in the frequentist approach, we are free to choose any estimator or decision procedure  $\delta : X \rightarrow A$  we want.

Having chosen an estimator, we define its expected loss or risk as follows:

$$R(\theta^*, \delta) = E_p(\tilde{D} | \theta^*) [L(\theta^*, \delta(\tilde{D}))] = L(\theta^*, \delta(\tilde{D})) p(\tilde{D} | \theta^*) d\tilde{D} \quad (6.9)$$

where  $\tilde{D}$  is data sampled from “nature’s distribution”, which is represented by parameter  $\theta^*$ . In other words, the expectation is wrt the sampling distribution of the estimator. Compare this to the Bayesian posterior expected loss:

$$\rho(a | D, \pi) = E_p(\theta | D, \pi) [L(\theta, a)] = \int \theta L(\theta, a) p(\theta | D, \pi) d\theta \quad (6.10)$$

We see that the Bayesian approach averages over  $\theta$

(which is unknown) and conditions on D (which is known), whereas the frequentist approach averages over  $D^*$  (thus ignoring the observed data), and conditions on  $\theta^*$  (which is unknown). Not only is the frequentist definition unnatural, it cannot even be computed, because  $\theta^*$  is unknown. Consequently, we cannot compare different estimators in terms of their frequentist risk. We discuss various solutions to this below.

6.3.1 Bayes risk How do we choose amongst estimators? We need some way to convert  $R(\theta^*, \delta)$  into a single measure of quality,  $R(\delta)$ , which does not depend on knowing  $\theta^*$ . One approach is to put a prior on  $\theta^*$ , and then to define Bayes risk or integrated risk of an estimator as follows:

$$RB(\delta) = E_p(\theta^*) [R(\theta^*, \delta)] = R(\theta^*, \delta)p(\theta^*)d\theta^* \quad (6.11)$$

A Bayes estimator or Bayes decision rule is one which minimizes the expected risk:

$$\delta_B = \operatorname{argmin}_{\delta} RB(\delta) \quad (6.12)$$

Note that the integrated risk is also called the preposterior risk, since it is before we have seen the data. Minimizing this can be useful for experiment design. We will now prove a very important theorem, that connects the Bayesian and frequentist approaches to decision theory.

Theorem 6.3.1. A Bayes estimator can be obtained by minimizing the posterior expected loss for each  $x$ .

Theorem 6.3.2 (Wald, 1950). Every admissible decision rule is a Bayes decision rule with respect to some, possibly improper, prior distribution.

This theorem shows that the best way to minimize frequentist risk is to be Bayesian! See (Bernardo and Smith 1994, p448) for further discussion of this point.

已思考 14s

下面先逐句翻译，再给出基础且详尽的解释（含直观示例）。我会把公式用更清晰的数学写法重排，保证不遗漏任何内容。

## 翻译

### 6.3 频率学派决策理论

在频率学派（或经典）决策理论中，存在**损失函数**和**似然**，但没有先验，因此也就没有后验或“后验期望损失”。因此，与贝叶斯情形不同，这里没有一种**自动**导出最优估计器的方法。相反，在频率学派的方法里，我们可以自由选择任意我们想要的估计器或决策过程  $\delta : X \rightarrow A$ 。

一旦选择了一个估计器，我们如下定义它的**期望损失**或**风险**（risk）：

$$R(\theta^*, \delta) = E_{p(\tilde{D}|\theta^*)}[L(\theta^*, \delta(\tilde{D}))] = \int L(\theta^*, \delta(\tilde{D})) p(\tilde{D} | \theta^*) d\tilde{D} \quad (6.9)$$

其中  $\tilde{D}$  是从“**自然的分布**”中抽到的数据，该分布由参数  $\theta^*$  表示。换句话说，**期望**是相对于估计器的**抽样分布**而言的。把这个与贝叶斯的**后验期望损失**对比：

$$\rho(a | D, \pi) = E_{p(\theta|D, \pi)}[L(\theta, a)] = \int_{\Theta} L(\theta, a) p(\theta | D, \pi) d\theta \quad (6.10)$$

我们看到：贝叶斯方法是对**未知的**  $\theta$  求平均，并且**以已知的数据**  $D$  为条件；而频率学派的方法是对  $\tilde{D}$  求平均（因此**忽略了已观察到的数据**），并且**以未知的**  $\theta^*$  为条件。频率学派的这个定义不仅**不自然**，而且甚至**无法计算**，因为  $\theta^*$  是未知的。因此，我们不能用它来比较不同估计器的频率学派风险。下面我们讨论若干解决方式。

### 6.3.1 贝叶斯风险 (Bayes risk)

我们该如何在估计器之间做选择？需要把  $R(\theta^*, \delta)$  转换成一个不依赖于知道  $\theta^*$  的单一质量度量  $R(\delta)$ 。一种方法是给  $\theta^*$  放一个**先验**，然后把估计器的**贝叶斯风险**（或**积分风险**）定义为：

$$R_B(\delta) = E_{p(\theta^*)}[R(\theta^*, \delta)] = \int R(\theta^*, \delta) p(\theta^*) d\theta^* \quad (6.11)$$

**贝叶斯估计器**或**贝叶斯决策规则**，是指使期望风险最小的那个：

$$\delta_B = \arg \min_{\delta} R_B(\delta) \quad (6.12)$$

注意，积分风险也称为**前后验风险**（preposterior risk），因为它发生在我们看到数据之前。最小化它对**实验设计**是有用的。接下来我们将证明一个把贝叶斯与频率学派的决策理论联系起来的非常重要的定理。

**定理 6.3.1.** 通过对每个  $x$  **最小化后验期望损失**，就能得到一个贝叶斯估计器。

**定理 6.3.2 (Wald, 1950).** 每一个**可容许**（admissible）的决策规则，都是相对于某个（可能是不真（improper）的）先验分布的**贝叶斯决策规则**。

该定理表明：**最小化频率学派风险的最佳方式，就是做贝叶斯！** 关于这一点的进一步讨论参见（Bernardo and Smith 1994, 第 448 页）。

# 解释 (基础且详尽)

## 1) 基本元素：动作、损失、估计器

- 观测空间  $X$ ：数据  $D$  活在这里。
- 动作空间  $A$ ：你要做的决策或给出的估计（例如“报一个参数值”）。
- 损失  $L(\theta, a)$ ：如果真实参数是  $\theta$ ，而你采取了动作  $a$ ，要付出的代价（如平方误差  $(\theta - a)^2$ 、0-1 损失等）。
- 决策规则/估计器  $\delta : X \rightarrow A$ ：看见数据  $D$  后输出动作/估计  $\delta(D)$ 。

## 2) 频率学派的“风险”：为什么尴尬？

频率学派认为 **参数  $\theta^*$  固定但未知**，而**数据是随机的**。于是它把“看到不同样本时可能犯的损失”取期望：

$$R(\theta^*, \delta) = E_{\tilde{D} \sim p(\cdot | \theta^*)}[L(\theta^*, \delta(\tilde{D}))].$$

问题在于：

- 它对未观察到的样本  $\tilde{D}$  求平均，忽略了手上真实的  $D$ ；
  - 它以未知的  $\theta^*$  为条件，所以**没法直接算**（我们不知道  $\theta^*$ ）。
- 因此，单看  $R(\theta^*, \delta)$ ，不同  $\delta$  的“好坏”无法在现实中直接比较。

## 3) 贝叶斯的“后验期望损失”：可计算、贴数据

贝叶斯把  $\theta$  当随机变量，有先验  $\pi(\theta)$ ，看到数据后有后验  $p(\theta | D, \pi)$ 。于是在已观测的数据  $D$  上，对  $\theta$  取后验平均：

$$\rho(a | D, \pi) = E_{\theta | D, \pi}[L(\theta, a)].$$

这可以**立刻计算**，并且与**我们手头的  $D$** 直接相关。

## 4) “贝叶斯风险”把频率学派问题补齐

为了把频率学派的  $R(\theta^*, \delta)$  变成一个**不依赖  $\theta^*$** 的整体指标，引入先验  $p(\theta^*)$  并做一次“再平均”：

$$R_B(\delta) = E_{\theta^*}[R(\theta^*, \delta)].$$

这等价于在实验之前、对“自然可能给的参数值”做加权平均，所以也叫**前后验风险，在实验设计**（比如该收多少样本、设计怎样的采样方案）里尤其有用。

直观比喻：频率学派的风险是“如果真相是  $\theta^*$ ，长期平均的损失是多少？”

贝叶斯风险是“真相可能是很多个  $\theta$ （按先验加权），综合考虑这些可能性后，长期平均

的损失是多少？”

## 5) 定理 6.3.1 的要点 (为什么“逐点最小化”就行)

把贝叶斯风险展开并用**全期望定理** (以及贝叶斯定理) 交换积分顺序, 可得

$$R_B(\delta) = E_D \left[ \underbrace{E_{\theta|D}[L(\theta, \delta(D))]}_{\text{后验期望损失}} \right].$$

因此, 要使整体期望最小, 只需对**每个可能出现的数据  $x$** , 让  $\delta(x)$  选择能**最小化后验期望损失**的动作即可。

- 若损失是**平方误差**, 最优动作是**后验均值**  $E[\theta | D]$ 。
  - 若损失是**绝对误差**, 最优是**后验中位数**。
  - 若损失是0-1 (分类), 最优是**后验众数/最大后验 (MAP)**。
- 这就是“贝叶斯估计器可通过逐点最小化后验风险得到”的含义。

## 6) 定理 6.3.2 (Wald) : 可容许 = 某个先验下的贝叶斯

- **可容许 (admissible)** : 不存在另一个规则在所有  $\theta$  上**风险都不更差**, 并且在至少一个  $\theta$  上**更好**。若有这样的对手规则, 你的规则就是**可劣 (inadmissible)**。
- Wald 的定理说: **每一个可容许的规则, 都是某个先验 (哪怕是不真先验) 下的贝叶斯规则。**

含义: 想要得到不被“统统吊打”的好规则, **等价于**在某个先验下做贝叶斯最优化。因此, 从“如何让频率学派风险好”的角度看, **做贝叶斯就是正道**。

## 7) 一个小例子 (估计伯努利成功率)

- 损失: 平方误差  $L(\theta, a) = (\theta - a)^2$ 。
- 频率学派: 你可以选任何估计器 (如样本均值、加权均值等), 然后关心  $R(\theta^*, \delta)$ , 但它依赖未知  $\theta^*$ , 难比较。
- 引入先验 (如  $Beta(\alpha, \beta)$ ) 后:
  - 后验是  $Beta(\alpha + S, \beta + F)$ ;
  - 逐点最小化后验期望损失得到 **后验均值**  $\frac{\alpha+S}{\alpha+\beta+N}$ ;
  - 这就是**贝叶斯估计器**, 并且按 Wald 的定理, 它是某种意义上**可容许的好选择**。

## 关键要点回顾

- 频率学派有损失与似然, 但**无先验/后验**, 因此“最优估计器”不再自动给出; 要比较规则需要风险  $R(\theta^*, \delta)$ , 可它**依赖未知真值**。

- **贝叶斯风险**通过给  $\theta^*$  放先验并积分，得到**不依赖真值的总体度量**，可用于设计与选择规则。
- **定理 6.3.1：**贝叶斯估计器 = 对每个数据点**最小化后验期望损失**的选择。
- **Wald 定理：**每个可容许的规则都是某个先验下的**贝叶斯规则**  $\Rightarrow$  想把频率学派风险做到“不可被统治”，**当贝叶斯就对了**。

如果你想，我可以用一个具体数据（比如 10 次投币、观测到 7 次正面）演示：计算后验、给出对应的贝叶斯估计，以及和频率学派风险的关系。



翻译并解释以下内容，翻译不能有遗漏，解释请基础且详尽，因为我的数学基础不太好。

Figure 6.2 Risk functions for two decision procedures,  $\delta_1$  and  $\delta_2$ . Since  $\delta_1$  has lower worst case risk, it is the minimax estimator, even though  $\delta_2$  has lower risk for most values of  $\theta$ . Thus minimax estimators are overly conservative.

6.3.2 Minimax risk Obviously some frequentists dislike using Bayes risk since it requires the choice of a prior (although this is only in the evaluation of the estimator, not necessarily as part of its construction). An alternative approach is as follows. Define the maximum risk of an estimator as

$$R_{\max}(\delta) = \max_{\theta^*} R(\theta^*, \delta) \quad (6.18)$$

A minimax rule is one which minimizes the maximum risk:  
 $\delta_{MM} = \operatorname{argmin}_{\delta} R_{\max}(\delta) \quad (6.19)$

For example, in Figure 6.2, we see that  $\delta_1$  has lower worst-case risk than  $\delta_2$ , ranging over all possible values of  $\theta^*$ , so it is the minimax estimator (see Section 6.3.3.1 for

an explanation of how to compute a risk function for an actual model). Minimax estimators have a certain appeal. However, computing them can be hard. And furthermore, they are very pessimistic. In fact, one can show that all minimax estimators are equivalent to Bayes estimators under a least favorable prior. In most statistical situations (excluding game theoretic ones), assuming nature is an adversary is not a reasonable assumption.

已思考 13s

下面先逐字翻译，再给出基础且详尽的解释（配合图意）。

---

## 翻译

**图 6.2** 两个决策过程  $\delta_1$  与  $\delta_2$  的风险函数。由于  $\delta_1$  的**最坏情形**风险更低，所以它是极小极大 (minimax) 估计器，尽管对大多数  $\theta$  的取值， $\delta_2$  的风险更低。因此，极小极大估计器往往**过于保守**。

### 6.3.2 极小极大风险 (Minimax risk)

显然，一些频率学派研究者不喜欢使用**贝叶斯风险**，因为它需要选择一个先验（尽管这只是在**评价**估计器时需要，并不一定是其**构造**的一部分）。一种替代方法如下。把某个估计器的**最大风险**定义为

$$R_{\max}(\delta) = \max_{\theta^*} R(\theta^*, \delta) \quad (6.18)$$

**极小极大规则**就是使最大风险最小的规则：

$$\delta_{\text{MM}} = \arg \min_{\delta} R_{\max}(\delta) \quad (6.19)$$

例如，在图 6.2 中，我们看到  $\delta_1$  在所有可能的  $\theta^*$  取值范围上，其**最坏情形**风险都比  $\delta_2$  更低，因此它是极小极大估计器（关于怎样为一个实际模型计算风险函数，见第 6.3.3.1 节）。极小极大估计器有一定吸引力。然而，计算它们可能很困难。而且，它们也非常悲观。事实上，可以证明：**所有**极小极大估计器都等价于在某个\*\*最不利先验 (least favorable prior) \*\*下的贝叶斯估计器。在大多数统计情境（博弈论情形除外）中，把自然假设为对手并不合理。

---

## 解释（基础且详尽）

## 1) 先回顾：风险函数 $R(\theta^*, \delta)$

- 它表示：当真实参数是  $\theta^*$  时，按照规则  $\delta$  去估计或决策，在重复抽样意义下的平均损失（期望损失）。
- 图里纵轴是  $R$ ，横轴是  $\theta$ 。两条曲线分别是两条规则的风险随  $\theta$  变化的样子。

## 2) 最大风险与极小极大思想

- **最大风险  $R_{\max}(\delta)$** ：看这条曲线的最高点——也就是“最坏情形”下你会有多糟。
- **极小极大规则**：在所有候选  $\delta$  里，挑那条曲线的最高点尽可能低者。  
直观：把自然当作“可能出最坏牌的对手”，你挑一个在最坏情况下也损失最小的策略。

## 3) 读图 6.2 的结论

- 蓝色虚线 ( $\delta_2$ ) 在大多数  $\theta$  区域都更低（平均表现更好），但在某些边缘区间会飙高；
- 红色实线 ( $\delta_1$ ) 整体较平、最高点更低。
- 于是按定义， $\delta_1$  的**最大风险更小** → 它是**极小极大估计器**。
- 但这也说明：**极小极大很保守**——为了压低最坏点，你牺牲了大多数情况下的表现。

## 4) 为何有人不爱用“贝叶斯风险”，转而用极小极大？

- 贝叶斯风险  $R_B(\delta) = E_{\theta^*}[R(\theta^*, \delta)]$  需要指定**先验  $p(\theta^*)$** 。
- 有些频率学派不想“主观”选先验，于是选择完全不依赖先验的**极小极大准则**。
- 好处：给出**分布无关**的保证（不管真相是哪一个  $\theta^*$ ，最坏也不至于太差）。
- 坏处：常常过于悲观，现实中真相未必总在“最坏处”；而且实际求解困难（要在规则和参数之间做  $\min_{\delta} \max_{\theta^*}$  的博弈）。

## 5) 与贝叶斯的深层联系——“最不利先验”

- 结论：**任何极小极大估计器，都等价于某个先验下的贝叶斯估计器**；这个先验叫**最不利先验**。
- “最不利”的意思：在这个先验下，最优贝叶斯规则的贝叶斯风险被**推到最高**，并与极小极大问题的“鞍点”对上。
- 含义：即便你宣称“我不想用先验”，极小极大其实隐含了一个**对抗性的先验视角**。
- 因此，在非博弈的普通统计任务里，把“自然”当成敌人往往不现实；此时极小极大就显得**太保守**。

## 6) 一个通俗小例子（只看概念）

假设我们估计一枚硬币正面概率  $\theta$ ，损失用平方误差。

- 规则  $\delta_2$ : 用样本比例  $\hat{\theta}$  (无偏, 平均表现好), 但当样本极小且  $\theta$  很接近 0 或 1 时, 风险会冲高。
  - 规则  $\delta_1$ : 做一点“收缩”(例如把比例往 0.5 拉一点)。这会让多数  $\theta$  处略差, 但能显著降低端点附近的最坏风险。
  - 因此  $\delta_1$  可能成为**极小极大**; 而  $\delta_2$  虽整体更好, 却输在**最坏点**。
- 

## 小结

- **极小极大**: 选择使**最坏情形风险**最低的规则; 优点是稳妥、无先验; 缺点是**悲观、难算**、常牺牲平均表现。
- 从理论上, 极小极大规则 = 某个**最不利先验下的贝叶斯规则**; 因此想靠极小极大来“避免贝叶斯”, 并不真正做到了“去先验”。
- 在大多数统计应用中, 更常见的做法是: 用合适的先验/模型假设进行贝叶斯或频率学派的权衡, 而不是把“自然”当对手。



翻译并解释以下内容, 翻译不能有遗漏, 解释请基础且详尽, 因为我的数学基础不太好。

Figure 6.3 Risk functions for estimating the mean of a Gaussian using data sampled  $N(\theta^*, \sigma^2 = 1)$ . The solid dark blue horizontal line is the MLE, the solid light blue curved line is the posterior mean when  $\kappa = 5$ . Left:  $N = 5$  samples. Right:  $N = 20$  samples. Based on Figure B.1 of (Bernardo and Smith 1994). Figure generated by riskFnGauss.

### 6.3.3 Admissible estimators

The basic problem with frequentist decision theory is that it relies on knowing the true distribution  $p(\cdot | \theta^*)$  in order to evaluate the risk. However, It might be the case that some estimators are worse than others regardless of the

value of  $\theta^*$ . In particular, if  $R(\theta, \delta_1) \leq R(\theta, \delta_2)$  for all  $\theta \in \Theta$ , then we say that  $\delta_1$  dominates  $\delta_2$ . The domination is said to be strict if the inequality is strict for some  $\theta$ . An estimator is said to be admissible if it is not strictly dominated by any other estimator.

### 6.3.3.1 Example

Let us give an example, based on (Bernardo and Smith 1994). Consider the problem of estimating the mean of a Gaussian. We assume the data is sampled from  $x_i \sim N(\theta^*, \sigma^2 = 1)$  and use quadratic loss,  $L(\theta, \hat{\theta}) = (\theta - \hat{\theta})^2$ . The corresponding risk function is the MSE. Some possible decision rules or estimators  $\hat{\theta}(x) = \delta(x)$  are as follows:

$\delta_1(x) = \bar{x}$ , the sample mean

$\delta_2(x) = \tilde{x}$ , the sample median

$\delta_3(x) = \theta_0$ , a fixed value

$\delta_K(x)$ , the posterior mean under a  $N(\theta|0, \sigma^2/\kappa)$  prior:

$$\delta_K(x) = N N + \kappa x + \kappa N + \kappa \theta_0 = w\bar{x} + (1-w)\theta_0 \quad (6.20)$$

For  $\delta_K$ , we consider a weak prior,  $\kappa = 1$ , and a stronger prior,  $\kappa = 5$ . The prior mean is  $\theta_0$ , some fixed value. We assume  $\sigma^2$  is known. (Thus  $\delta_3(x)$  is the same as  $\delta_K(x)$  with an infinitely strong prior,  $\kappa = \infty$ .) Let us now derive the risk functions analytically. (We can do this since in this toy example, we know the true parameter  $\theta^*$ .) In Section 6.4.4, we show that the MSE can be decomposed into squared bias plus variance:

$$MSE(\hat{\theta}(\cdot)|\theta^*) = \text{var } \hat{\theta} + \text{bias}^2(\hat{\theta}) \quad (6.21)$$

The sample mean is unbiased, so its risk is

$$MSE(\delta_1|\theta^*) = \text{var } [\bar{x}] = \sigma^2 N \quad (6.22)$$

The sample median is also unbiased. One can show that the variance is approximately  $\pi/(2N)$ , so

$$MSE(\delta_2|\theta^*) = \pi/2N \quad (6.23)$$

For  $\delta_3(x) = \theta_0$ , the variance is zero, so

$$MSE(\delta_3|\theta^*) = (\theta^* - \theta_0)^2 \quad (6.24)$$

Finally, for the posterior mean, we have

$$MSE(\delta_K|\theta^*) = E[(w\bar{x} + (1-w)\theta_0 - \theta^*)^2] \quad (6.25)$$

$$= E[(w(x - \theta^*) + (1-w)(\theta_0 - \theta^*))^2] \quad (6.26)$$

$$= w^2 \sigma^2 N + (1-w)^2 (\theta_0 - \theta^*)^2 \quad (6.27)$$

$$= 1(N + \kappa)^2 N \sigma^2 + \kappa^2 (\theta_0 - \theta^*)^2 \quad (6.28)$$

These functions are plotted in Figure 6.3 for  $N \in \{5, 20\}$ .

We see that in general, the best estimator depends on the value of  $\theta^*$ , which is unknown. If  $\theta^*$  is very close to  $\theta_0$ , then  $\delta_3$  (which just predicts  $\theta_0$ ) is best. If  $\theta^*$  is within some reasonable range around  $\theta_0$ , then the posterior

mean, which combines the prior guess of  $\theta_0$  with the actual data, is best. If  $\theta^*$  is far from  $\theta_0$ , the MLE is best. None of this should be surprising: a small amount of shrinkage (using the posterior mean with a weak prior) is usually desirable, assuming our prior mean is sensible. What is more surprising is that the risk of decision rule  $\delta_2$  (sample median) is always higher than that of  $\delta_1$  (sample mean) for every value of  $\theta^*$ . Consequently the sample median is an  $R(\theta^*, \delta)$  inadmissible estimator for this particular problem (where the data is assumed to come from a Gaussian). In practice, the sample median is often better than the sample mean, because it is more robust to outliers. One can show (Minka 2000d) that the median is the Bayes estimator (under squared loss) if we assume the data comes from a Laplace distribution, which has heavier tails than a Gaussian. More generally, we can construct robust estimators by using flexible models of our data, such as mixture models or non-parametric density estimators (Section 14.7.2), and then computing the posterior mean or median.

已思考 18s

下面先完整翻译，再做基础且详尽的解释（含直观读图与推导思路）。我会把原式子整理成清晰的数学写法，保证不遗漏任何内容。

## 翻译

图 6.3 用来自  $N(\theta^*, \sigma^2 = 1)$  的数据估计高斯分布均值时的风险函数。深蓝色的水平实线是 MLE (极大似然估计)，浅蓝色的弯曲实线是在  $\kappa = 5$  时的后验均值。左图： $N = 5$  个样本。右图： $N = 20$  个样本。基于 (Bernardo and Smith 1994) 图 B.1。图由 `riskFnGauss` 生成。

### 6.3.3 可容许 (admissible) 估计器

频率学派决策理论的基本问题在于：它在评估风险时依赖知道真实分布  $p(\cdot | \theta^*)$ 。然而，可能存在这样的情形：不论  $\theta^*$  取何值，某些估计器都劣于另一些。具体地，若对所有  $\theta \in \Theta$  都有  $R(\theta, \delta_1) \leq R(\theta, \delta_2)$ ，则称  $\delta_1$  支配 (dominate)  $\delta_2$ 。如果对某些  $\theta$  严格不等（小于），则称为严格支配。若一个估计器没有被任何其他估计器严格支配，就称它是可容许的 (admissible)。

### 6.3.3.1 例子

给出一个基于 (Bernardo and Smith 1994) 的例子。考虑估计高斯分布的均值。假设数据来自  $x_i \sim N(\theta^*, \sigma^2 = 1)$ , 并采用**二次损失**  $L(\theta, \hat{\theta}) = (\theta - \hat{\theta})^2$ 。相应的风险函数就是**均方误差 (MSE)**。一些可能的决策规则或估计器  $\hat{\theta}(x) = \delta(x)$  如下:

- $\delta_1(x) = \bar{x}$ , 样本均值
- $\delta_2(x) = \tilde{x}$ , 样本中位数
- $\delta_3(x) = \theta_0$ , 某个**固定常数**
- $\delta_\kappa(x)$ , 在先验  $N(\theta | \theta_0, \sigma^2/\kappa)$  下的**后验均值**:

$$\delta_\kappa(x) = \frac{N}{N + \kappa} \bar{x} + \frac{\kappa}{N + \kappa} \theta_0 = w\bar{x} + (1 - w)\theta_0, \quad w = \frac{N}{N + \kappa} \quad (6.20)$$

对  $\delta_\kappa$ , 我们考虑一个**弱先验**  $\kappa = 1$  和一个**强先验**  $\kappa = 5$ 。先验均值是某个固定值  $\theta_0$ 。我们假设  $\sigma^2$  已知。(因此  $\delta_3(x)$  与  $\kappa = \infty$  的  $\delta_\kappa(x)$  是一样的——先验无限强, 只报  $\theta_0$ 。) 现在解析地推导风险函数。(在这个玩具例子里, 我们**知道**真实参数  $\theta^*$ 。)

在第 6.4.4 节里我们会展示: MSE 可以分解为“偏差平方 + 方差”:

$$\text{MSE}(\hat{\theta}(\cdot) | \theta^*) = \text{var}(\hat{\theta}) + \text{bias}^2(\hat{\theta}) \quad (6.21)$$

- 样本均值**无偏**, 所以它的风险是

$$\text{MSE}(\delta_1 | \theta^*) = \text{var}(\bar{x}) = \frac{\sigma^2}{N} \quad (6.22)$$

- 样本中位数也**无偏**。可以证明其方差近似为  $\pi/(2N)$ , 因此

$$\text{MSE}(\delta_2 | \theta^*) = \frac{\pi}{2N} \quad (6.23)$$

- 对  $\delta_3(x) = \theta_0$ , 方差为 0, 因此

$$\text{MSE}(\delta_3 | \theta^*) = (\theta^* - \theta_0)^2 \quad (6.24)$$

- 最后, 对后验均值,

```
\begin{aligned}
\text{MSE}(\delta_\kappa | \theta^*) &= \\
&\mathbb{E}[(\bar{x} + (1-w)(\theta_0 - \theta^*))^2] \tag{6.25} \\
&\mathbb{E}[(\bar{x} - \theta^*)^2 + (1-w)(\theta_0 - \theta^*)^2] \tag{6.26} \\
&= w^2 \frac{\sigma^2}{N} + (1-w)^2 (\theta_0 - \theta^*)^2 \tag{6.27} \\
&= \frac{1}{(N+\kappa)^2} (N\sigma^2 + \kappa(\theta_0 - \theta^*)^2) \tag{6.28}
\end{aligned}
```

这些函数在图 6.3 中对  $N \in \{5, 20\}$  作了绘制。我们看到, 一般而言, **最佳估计器取决于  $\theta^*$  的取值**——而  $\theta^*$  是未知的:

- 如果  $\theta^*$  非常接近  $\theta_0$ , 那么只报  $\theta_0$  的  $\delta_3$  最好;

- 如果  $\theta^*$  位于  $\theta_0$  附近的某个**合理范围内**，那么结合了“先验猜测  $\theta_0$ ”与“实际数据”的**后验均值最好**；
- 如果  $\theta^*$  离  $\theta_0$ **很远**，则 MLE 最好。

这并不令人意外：只要先验均值  $\theta_0$  合理，用**弱先验做一点点收缩**（后验均值）通常是可取的。更令人意外的是：在这个问题中（假设数据来自高斯），决策规则  $\delta_2$ （样本中位数）的风险在每一个  $\theta^*$  上都高于  $\delta_1$ （样本均值）。因此，样本中位数是这个特定问题的一个**不可容许 (inadmissible) 估计器**。

**在实践中，中位数往往比均值更好，因为它对离群点更鲁棒。**可以证明 (Minka 2000d)，如果我们假设数据来自**拉普拉斯分布**（比高斯有更重的尾部），那么在平方损失下，中位数就是**贝叶斯估计器**。更一般地，我们可以通过使用对数据更灵活的模型（例如**混合模型或非参数密度估计**，见第 14.7.2 节），然后计算后验均值或后验中位数，从而构造鲁棒估计器。

## 解释 (通俗、分步)

### 1) 图该怎么读？

- 横轴是“真实均值”  $\theta^*$ ，纵轴是风险  $R(\theta^*, \delta) = \text{MSE}$ 。
- **深蓝水平线 (MLE:  $\bar{x}$ )** ——其 MSE 是  $\sigma^2/N$ ，与  $\theta^*$ **无关**，所以画成水平线；样本越多，线越低（右图比左图低）。
- **黑色虚线 (固定报  $\theta_0$ )** ——在  $\theta^* = \theta_0$  处  $\text{MSE}=0$ ，离  $\theta_0$  越远越差，呈“U 形抛物线”。
- **浅蓝/绿色曲线 (后验均值)** ——在  $\theta_0$  附近低、远离时升高； $\kappa$  越大（先验越强），曲线越像黑虚线； $\kappa$  小（先验弱）时曲线更接近水平的 MLE。
- **红色虚线 (中位数)** ——近似水平，但位置始终**高于深蓝线 (均值)**，说明在正态模型下，中位数 MSE 始终更大。

左图  $N = 5$ ：各曲线差异大，先验的影响明显；右图  $N = 20$ ：总体风险下降，后验均值曲线更接近 MLE，说明**数据量变大后，先验权重变小**（因为  $w = N/(N + \kappa)$  增大）。

### 2) 式 (6.20) 的含义 (“收缩/折中”)

$$\delta_\kappa = w\bar{x} + (1-w)\theta_0, \quad w = \frac{N}{N + \kappa}.$$

- $\kappa$  是“先验强度”。 $\kappa$  越大，越相信先验  $\theta_0$ ； $\kappa \rightarrow \infty$  就退化成“固定报  $\theta_0$ ”；
- $N$  越大， $w$  越接近 1，**更多相信数据**， $\delta_\kappa$  趋近于 MLE。

### 3) 为什么这些 MSE 公式成立？

- **均值**:  $\text{var}(\bar{x}) = \sigma^2/N$ , 无偏  $\Rightarrow \text{MSE} = \text{方差}$ 。
- **中位数 (正态)** : 渐近方差  $\approx \pi\sigma^2/(2N) \Rightarrow$  这里  $\sigma^2 = 1$  时就是  $\pi/(2N)$ 。
- **固定常数  $\theta_0$** : 没有随机性  $\Rightarrow$  方差 0, 只有偏差  $(\theta^{*} - \theta_0)^2$ 。
- **后验均值**: 把  $\delta_{\kappa} - \theta^{*}$  写成“随机项  $w(\bar{x} - \theta^{*})$  + 常数项  $(1 - w)(\theta_0 - \theta^{*})$ ”；前者均值 0、方差  $(w^2\sigma^2/N)$ , 后者是偏差, 平方后得到式 (6.27)-(6.28)。

#### 4) 可容许性与“中位数为何不可容许 (在高斯假设下) ? ”

- **支配**: 若某规则在所有  $\theta^{*}$  上 MSE 都不更大、且至少在某些  $\theta^{*}$  更小, 就支配另一个。
- 在这里,  $\text{MSE}(\text{中位数}) = \pi/(2N)$  始终大于  $\text{MSE}(\text{均值}) = \sigma^2/N$  (当  $\sigma^2 = 1$ ,  $\pi/2 \approx 1.57 > 1$ ) , 所以**均值支配中位数**  $\Rightarrow$  **中位数不可容许**。
- 但这结论仅在“数据真是高斯”时成立。若数据尾部更重 (如拉普拉斯), 中位数反而在平方损失下是贝叶斯最优。这说明: **可容许/优劣与模型假设密切相关**。

#### 5) 实用建议 (直观)

- **样本均值**: 在近似正态且离群点不多时, MSE 最小;
- **中位数**: 对离群点鲁棒, 但若数据真近似正态, 效率 (MSE) 不如均值;
- **后验均值 (收缩)** : 当你对  $\theta$  的合理范围有先验判断 ( $\theta_0$ ), 尤其在样本少时, 常能**整体降低风险**; 样本大时自动更靠近均值;
- **永远报  $\theta_0$** : 只有在你非常确信  $\theta^{*} \approx \theta_0$  时才好, 否则风险会快速变大。

### 小结

- 图 6.3 直观展示了不同估计器的**风险随  $\theta^{*}$  的变化**: 水平 (MLE) 、U 形 (固定值) 、介于两者之间的收缩 (后验均值) 、以及始终高一些的中位数 (在正态假设下) 。
- **可容许性**: 若一个规则在所有参数值上都不更差且在某些点更好, 就**支配**另一个; 被**支配者不可容许**。在正态模型下, **均值支配中位数**。
- **模型很重要**: 一旦分布尾部变重 (如拉普拉斯), 中位数就可能成为**贝叶斯最优**。选哪种估计器, 取决于你对数据分布与先验信息的判断。

θ和估计器是什么关系

已思考 23s

简短说:

- **$\theta$  (theta) 是模型里的真实参数**, 比如硬币正面概率、正态分布的真实均值等。
- **估计器 (estimator, 记作  $\delta$ ) 是一个规则/函数**, 把数据  $D$  映射成一个数 (或向量) 当作对  $\theta$  的猜测,  $\hat{\theta} = \delta(D)$ 。
- **估计值 (estimate) 是估计器在你这份具体数据上算出来的具体数值**, 也写成  $\hat{\theta}$ 。

下面把两者的关系讲清楚 (分频率学派与贝叶斯两种观点) :

---

## 1) 频率学派 (Frequentist)

- 认为  $\theta$  是固定但未知; 随机性来自数据。
- 估计器  $\delta(D)$  因为依赖随机数据而是**随机变量**, 在“重复抽样”下有自己的**抽样分布**。
- 估计器与  $\theta$  的关系用若干性质刻画:
  - **偏差**:  $\text{bias}(\delta) = E[\delta(D)] - \theta$ 。无偏就等于 0。
  - **方差/标准误**:  $\text{var}(\delta)$ 、 $\text{se} = \sqrt{\text{var}}$ , 描述估计值围绕  $\theta$  摆摆的大小。
  - **均方误差 (MSE)** :  $\text{MSE} = \text{var} + \text{bias}^2$ , 综合“准不准”和“稳不稳”。
  - **一致性**: 样本量  $N \rightarrow \infty$  时,  $\delta(D) \rightarrow \theta$  (概率意义)。
  - **效率**: 在无偏估计器中, 方差最小者最好 (受 Cramér–Rao 下界约束)。
- 例子 (伯努利) : 观测  $X_i \sim \text{Ber}(\theta)$ , 取

$$\hat{\theta} = \delta(D) = \frac{1}{N} \sum_i X_i$$

则  $E[\hat{\theta}] = \theta$  (无偏),  $\text{var}(\hat{\theta}) = \theta(1 - \theta)/N$ ,  $N$  越大,  $\hat{\theta}$  越靠近  $\theta$ 。

---

## 2) 贝叶斯 (Bayesian)

- 把  $\theta$  当作**随机变量** (有先验  $\pi(\theta)$ ), 看到数据后得到**后验**  $p(\theta | D)$ 。
- “估计器”是一个**决策规则**, 通常通过**最小化后验期望损失**得到:

$$\delta_B(D) = \arg \min_a E_{\theta|D}[L(\theta, a)].$$

- 若损失是平方误差, 最优是**后验均值**  $E[\theta | D]$ 。
  - 若是绝对误差, 最优是**后验中位数**; 若是 0–1 损失, 最优是 MAP。
  - 这里估计器与  $\theta$  的关系是:  $\delta_B(D)$  用后验把  $\theta$  的不确定性“平均”起来, 给出对  $\theta$  的最佳动作/数值。
-

### 3) 一句话对比

- $\theta$ : 问题的“真相”。
  - **估计器  $\delta$** : 用数据去“读取真相”的**工具/算法**; 输出  $\hat{\theta}$  作为对  $\theta$  的近似。
  - 频率学派以“反复抽样”评估  $\delta$  围绕  $\theta$  的表现; 贝叶斯用“后验”直接对  $\theta$  的不确定性做决策。
- 

### 4) 再给一个直观类比

- $\theta$  像“室温的真实值”。
- 估计器像“温度计+读数规则”。温度计会有随机误差（方差）、可能系统性偏高/偏低（偏差）。我们希望温度计既不偏、抖动又小，或者在贝叶斯视角下，结合先验知识给出最佳读数。

如果你告诉我一个具体场景（比如估计硬币正面率、或正态均值、或回归系数），我可以把对应的  $\delta$ 、偏差、方差、MSE 公式写出来对比给你看。

翻译并解释以下内容，翻译不能有遗漏，解释请基础且详尽，因为我的数学基础不太好。

#### 6.3.3.2 Stein's paradox

Suppose we have  $N$  iid random variables  $X_i \sim N(\theta_i, 1)$ , and we want to estimate the  $\theta_i$ . The obvious estimator is the MLE, which in this case sets  $\hat{\theta}_i = x_i$ . It turns out that this is an inadmissible estimator under quadratic loss, when  $N \geq 4$ . To show this, it suffices to construct an estimator that is better. The James-Stein estimator is one such estimator, and is defined as follows:

$$\hat{\theta}_i = \hat{B}x + (1 - \hat{B})x_i = x + (1 - \hat{B})(x_i - x) \quad (6.29)$$

where  $x = \frac{1}{N} \sum_{i=1}^N x_i$  and  $0 < B < 1$  is some tuning constant. This estimate “shrinks” the  $\theta_i$  towards the overall mean. (We derive this estimator using an empirical Bayes approach in Section 5.6.2.) It can be shown that this shrinkage estimator has lower frequentist risk (MSE) than the MLE (sample mean) for  $N \geq 4$ . This is known as Stein's paradox. The reason it is called a paradox is illustrated by the following example. Suppose  $\theta_i$  is the “true” IQ of student  $i$  and  $X_i$  is his test score. Why should my estimate

of  $\theta_i$  depend on the global mean  $x$ , and hence on some other student's scores? One can create even more paradoxical examples by making the different dimensions be qualitatively different, e.g.,  $\theta_1$  is my IQ,  $\theta_2$  is the average rainfall in Vancouver, etc. The solution to the paradox is the following. If your goal is to estimate just  $\theta_i$ , you cannot do better than using  $x_i$ , but if the goal is to estimate the whole vector  $\theta$ , and you use squared error as your loss function, then shrinkage helps. To see this, suppose we want to estimate  $\|\theta\|^2$  from a single sample  $x \sim N(\theta, I)$ . A simple estimate is  $\|x\|^2$ , but this will overestimate the result, since

$$E \|x\|^2 = E \sum_{i=1}^N x_i^2 = N + \|\theta\|^2 \quad (6.30)$$

Consequently we can reduce our risk by pooling information, even from unrelated sources, and shrinking towards the overall mean. In Section 5.6.2, we give a Bayesian explanation for this. See also (Efron and Morris 1975).

已思考 37s

下面先完整翻译，然后做通俗、详细的解释（含直观推理与小算式）。我会保留原式子并把符号写清楚。

## 翻译

### 6.3.3.2 斯坦 (Stein) 的悖论

设我们有  $N$  个相互独立同分布的随机变量  $X_i \sim N(\theta_i, 1)$ ，目标是估计各个  $\theta_i$ 。显而易见的估计器是 MLE，在本情形它令  $\hat{\theta}_i = x_i$ 。事实证明，在**二次损失**（平方误差）下，当  $N \geq 4$  时，这个估计器是**不可容许的**。要证明这一点，只需构造一个更好的估计器即可。

**James–Stein 估计器**就是这样一个估计器，定义如下：

$$\hat{\theta}_i = \hat{B}\bar{x} + (1 - \hat{B})x_i = \bar{x} + (1 - \hat{B})(x_i - \bar{x}) \quad (6.29)$$

其中  $\bar{x} = \frac{1}{N} \sum_{i=1}^N x_i$ ，并且  $0 < B < 1$  是某个调节常数。这个估计把  $\theta_i$  “**向总体均值收缩**”。（我们在 5.6.2 节用**经验贝叶斯**的方法推出这个估计器。）可以证明，对  $N \geq 4$ ，这种收缩估计器的**频率学派风险 (MSE)** 比 MLE (样本均值) 更低。这就叫**Stein 的悖论**。之所以称为“悖论”，可用下面的例子说明：假设  $\theta_i$  是第  $i$  个学生的“真实 IQ”，而  $X_i$  是他的测试分数。为什么我对  $\theta_i$  的估计要依赖于全体均值  $\bar{x}$ ，从而依赖于**其他学生的分数**呢？把不同维度设成性质完全不同的量会显得更悖论，比如  $\theta_1$  是我的 IQ， $\theta_2$  是温哥华的平均降雨量，等等。

解决这个悖论的关键是：如果你的目标只是估计某一个  $\theta_i$ ，那你做不到比使用  $x_i$  更好；但如果目标是估计整个向量  $\theta$ ，并且你的损失函数是平方误差，那么收缩就会有帮助。为说明这一点，设我们想用一次观测  $x \sim N(\theta, I)$  来估计  $\|\theta\|_2^2$ 。一个简单估计是  $\|x\|_2^2$ ，但它会高估结果，因为

$$E[\|x\|_2^2] = E\left[\sum_i x_i^2\right] = \sum_i (1 + \theta_i^2) = N + \|\theta\|_2^2 \quad (6.30)$$

因此，即使是来自不相关来源的信息，通过汇总并向总体均值收缩也能降低我们的风险。在 5.6.2 节，我们给出这一点的贝叶斯解释。亦可参见 (Efron and Morris, 1975)。

---

## 解释（基础且详尽）

### 1) 问题设定与“悖论”是什么

- 有  $N$  个量要同时估计： $\theta_1, \dots, \theta_N$ 。每个量我们只看到一次观测： $X_i \sim N(\theta_i, 1)$ 。
- **朴素做法 (MLE)**：就用各自的观测当估计： $\hat{\theta}_i = x_i$ 。这对单个参数来说确实合理：无偏、方差 1。
- **惊人结论**：当一起评价“整组估计”的好坏（把每个坐标的平方误差加总）时，存在一种“把各坐标向总体均值  $\bar{x}$  拉一拉（收缩）”的办法，使得总的平均平方误差更小——统一好于逐个用  $x_i$ 。这就是 Stein 的悖论。

直观：我们允许引入一点点“系统性偏差”（把估计往  $\bar{x}$  拉），换回每个坐标方差的明显下降；在高维时，总体方差的下降超过了偏差平方的增加，因而总风险下降。

### 2) James–Stein 估计器在做什么

$$\hat{\theta}_i = \hat{B} \bar{x} + (1 - \hat{B}) x_i = \bar{x} + (1 - \hat{B})(x_i - \bar{x}), \quad 0 < \hat{B} < 1.$$

- 当  $\hat{B} = 0$ ：退化为 MLE（完全相信各自的观测）。
- 当  $\hat{B} \rightarrow 1$ ：所有估计都变成同一个数  $\bar{x}$ （强烈收缩）。
- 实际上  $\hat{B}$  由数据自适应确定（在经验贝叶斯里从数据估出来），通常位于中间，使估计适度地向整体均值靠拢。

这种“借力”“借数”也称 borrowing strength：每个坐标的估计不仅用自己的信息，也利用“总体水平”的信息。

### 3) 为何“整体更好”而“单个不会更好”

- 若只看某个固定的  $i$ ， $\hat{\theta}_i = x_i$  在正态+平方损失下已经是最佳无偏选择；任何强行往  $\bar{x}$  拉的动作，都会对这个坐标引入偏差，不一定更好。

- 但把所有坐标的平方误差求和作为损失时（这是章节里默认的目标），收缩会整体降低方差之和，且这一下降在  $N$  较大时更明显，于是总和的 MSE 下降  $\Rightarrow$  MLE 被严格支配  $\Rightarrow$  不可容许。

#### 4) 式 (6.30) 直观演示“为什么要收缩”

我们想估计  $\|\theta\|_2^2 = \sum \theta_i^2$ 。最直接的估计是  $\sum x_i^2$ 。但

$$E[x_i^2] = \text{var}(x_i) + (Ex_i)^2 = 1 + \theta_i^2,$$

所以

$$E[\sum x_i^2] = N + \sum \theta_i^2 = N + \|\theta\|_2^2.$$

这说明朴素估计偏大，多出来的正是噪声方差在每个坐标上各加了 1。

收缩的作用：把各  $x_i$  往整体均值拉一点，等价于从每个坐标里“减掉一些噪声”，从而把系统的“+1”偏差抵消一部分，降低了总风险。

#### 5) 为什么被称为“悖论”

- 从“单个学生 IQ”角度看，凭什么要参考其他学生的分数？
- 但从“同时估计全班所有人的 IQ，并且用平方误差求和来打分”的角度，适度“向班级平均分收缩”能让总误差更小。
- 甚至当不同维度完全不相关（如 IQ vs. 温哥华降雨量），只要数学上的损失函数是“把所有维度的平方误差相加”，收缩照样在数学上减少总风险。这就是“悖论感”的来源。

#### 6) 备注与边界

- 结论依赖：正态分布、已知同方差、平方损失、同时评估整个向量。条件变了（如不同方差、非正态、不同损失），形式会变化。
- 实务里常用改进版本（如“正部分” James–Stein），以避免出现负的收缩系数。
- 贝叶斯视角（第 5.6.2 节）：把  $\theta_i$  看作从某个总体（例如以  $\theta_0$  为均值、方差  $\sigma^2/\kappa$ ）中抽来的；则 James–Stein 可看作在该层级模型下的后验均值/经验贝叶斯估计，自然就是一种“向总体 ( $\theta_0$  或  $\bar{x}$ ) 收缩”。

#### 一句话总结

- \*\*单点最优：\*\*估计某个  $\theta_i$  本身， $\hat{\theta}_i = x_i$  很难被全面改进。
- 整体最优：若以总平方误差为目标去估计整个  $\theta$  向量，James–Stein 的收缩能在  $N$  足够大时（文中  $N \geq 4$ ）获得更低风险  $\Rightarrow$  这使得逐点的 MLE 不可容许。
- 直觉：用一点偏差换取更大的方差下降，在高维时合算。

翻译并解释以下内容，翻译不能有遗漏，解释请基础且详尽，因为我的数学基础不太好。

### 6.3.3.3 Admissibility is not enough

It seems clear that we can restrict our search for good estimators to the class of admissible estimators. But in fact it is easy to construct admissible estimators, as we show in the following example.

Theorem 6.3.3. Let  $X \sim N(\theta, 1)$ , and consider estimating  $\theta$  under squared loss. Let  $\delta_1(x) = \theta_0$ , a constant independent of the data. This is an admissible estimator.

已思考 43s

下面先完整翻译，再做基础且详尽的解释（含直观证明）。

## 翻译

### 6.3.3.3 仅有“可容许”还不够

看起来我们可以把寻找好估计器的范围限制在\*\*可容许 (admissible) \*\*估计器这一类上。但事实上，构造可容许估计器非常容易，下面的例子就说明了这一点。

**定理 6.3.3.** 设  $X \sim N(\theta, 1)$ , 并在平方损失下估计  $\theta$ 。令

$\delta_1(x) = \theta_0$  (一个与数据无关的常数)。这是一个可容许的估计器。

## 解释（基础且详尽）

### 1) 先回顾：什么叫“可容许”

- 风险函数  $R(\theta, \delta) = E_\theta[(\delta(X) - \theta)^2]$  (这里就是均方误差)。
- 若对所有  $\theta$  都有  $R(\theta, \delta_1) \leq R(\theta, \delta_2)$ , 且对某个  $\theta$  严格小于, 则称  $\delta_1$  支配  $\delta_2$ 。
- **可容许：**一个估计器没有被任何别的估计器严格支配。

### 2) 为什么“常数估计器” $\delta_1(x) = \theta_0$ 是可容许的?

它的风险是

$$R(\theta, \delta_1) = E_\theta[(\theta_0 - \theta)^2] = (\theta - \theta_0)^2.$$

特别地，在  $\theta = \theta_0$  时，

$$R(\theta_0, \delta_1) = 0.$$

现在假设存在另一个估计器  $\tilde{\delta}$  支配  $\delta_1$ 。那么在  $\theta = \theta_0$  处必须有

$$R(\theta_0, \tilde{\delta}) \leq R(\theta_0, \delta_1) = 0.$$

风险不可能为负，因此  $R(\theta_0, \tilde{\delta}) = 0$ 。在平方损失下，风险为 0 意味着

$$\tilde{\delta}(X) = \theta_0 \quad \text{以概率 1 (在 } \theta = \theta_0 \text{ 时)}.$$

而正态分布在所有实数上都有正密度，所以这强制  $\tilde{\delta}(x)$  对所有  $x$  都等于  $\theta_0$ 。也就是说  $\tilde{\delta}$  与  $\delta_1$  完全相同，因而在任何  $\theta$  下风险都相同，不可能“某处严格更好”。这与“严格支配”的假设矛盾。

⇒ 没有谁能严格支配  $\delta_1$ ，所以  $\delta_1$  是可容许的。

另一个角度： $\delta_1$  是把先验放在  $\theta = \theta_0$  的“点质量”上得到的贝叶斯估计器；而“相对于适当先验的贝叶斯估计器都是可容许的”。点质量先验当然是适当的，所以它可容许。

### 3) 但它“好”吗？——这正是标题想说的

- 尽管  $\delta_1$  可容许，它根本不看数据。除非真实值恰好非常接近  $\theta_0$ ，否则风险  $(\theta - \theta_0)^2$  会很大。
- 可容许性只保证“没有一个估计器在所有  $\theta$  上都比它好”，并不保证它在大多数  $\theta$  上表现好、或整体风险小。
- 因而，“仅凭可容许来选估计器”太弱；很多很差但不可被完全压制的估计器（比如这种常数估计器）都会通过这道筛子。

### 4) 一个小数值直观

设  $\theta_0 = 0$ ，样本均值（MLE）在单样本时的风险是 1 ( $\text{var}(X) = 1$ )。

- 常数估计器的风险是  $\theta^2$ 。它只有在  $|\theta| \leq 1$  时才不劣于 MLE；当  $|\theta| > 1$  时明显更差。
- 但仍不可被严格支配，因为在  $\theta = 0$  它的风险是 0，任何非平凡估计器都做不到比 0 更小。

### 5) 启示：需要更强的准则

在实践/理论上我们通常会再加入其他标准来挑选估计器，例如

- 贝叶斯风险**（选择合理先验，最小化总体期望风险）；
- 极小极大风险**（控制最坏情形）；
- 一致性、无偏/小偏、方差/效率、稳健性等性质**；

- 等变性/对称性限制，排除“怪异”的规则。

这些标准能把像  $\delta_1$  这样的“虽然可容许但没什么用”的估计器排除掉。

### 一句话总结：

“可容许”只是不被任何规则在所有参数值上全面碾压；它并不等于“好”。常数估计器  $\delta_1(x) = \theta_0$  就是可容许但通常很差的典型例子，因此仅靠可容许性远远不够。

翻译并解释以下内容，翻译不能有遗漏，解释请基础且详尽，因为我的数学基础不太好。

#### 6.4 Desirable properties of estimators

Since frequentist decision theory does not provide an automatic way to choose the best estimator, we need to come up with other heuristics for choosing amongst them. In this section, we discuss some properties we would like estimators to have. Unfortunately, we will see that we cannot achieve all of these properties at the same time.

##### 6.4.1 Consistent estimators

An estimator is said to be consistent if it eventually recovers the true parameters that generated the data as the sample size goes to infinity, i.e.,  $\hat{\theta}(D) \rightarrow \theta^*$  as  $|D| \rightarrow \infty$  (where the arrow denotes convergence in probability). Of course, this concept only makes sense if the data actually comes from the specified model with parameters  $\theta^*$ , which is not usually the case with real data.

Nevertheless, it can be a useful theoretical property. It can be shown that the MLE is a consistent estimator. The intuitive reason is that maximizing likelihood is equivalent to minimizing  $KL + p(\cdot|\theta^*)||p(\cdot|\hat{\theta})$ , where  $p(\cdot|\theta^*)$  is the true distribution and  $p(\cdot|\hat{\theta})$  is our estimate. We can achieve 0 KL divergence iff  $\hat{\theta} = \theta^*$ .

已思考 7s

下面先完整翻译，再给出基础且详尽的解释（我会把排版小错误更正为标准写法）。

# 翻译

## 6.4 估计器的期望性质

由于频率学派决策理论并不能自动给出“最佳估计器”，我们需要提出其他的启发式方法来在不同估计器之间做选择。本节讨论我们希望估计器具备的一些性质。不幸的是，我们会看到这些性质不可能同时全部满足。

### 6.4.1 一致估计 (Consistent estimators)

如果当样本量趋于无穷大时，估计器最终能够恢复生成数据的真实参数，则称该估计器是一致的；也就是说，当  $|D| \rightarrow \infty$  时，

$$\hat{\theta}(D) \xrightarrow{P} \theta^*$$

(箭头表示依概率收敛)。当然，只有当数据确实来自“用参数  $\theta^*$  指定的那个模型”时，这个概念才有意义；而在真实数据里，这往往并不成立。尽管如此，这仍是一个有用的理论性质。

可以证明，极大似然估计 (MLE) 是一个一致估计。直观原因是：最大化似然等价于最小化下面这个 Kullback–Leibler 散度

$$\text{KL}(p(\cdot | \theta^*) \| p(\cdot | \hat{\theta})),$$

其中  $p(\cdot | \theta^*)$  是真实分布， $p(\cdot | \hat{\theta})$  是我们的估计。只有当  $\hat{\theta} = \theta^*$  时，KL 散度才能为 0。

## 解释 (基础且详尽)

### 1) “一致”是什么意思？

- 口语版：数据越来越多时，你的估计会越来越接近真值。
- 数学版：对任意很小的  $\varepsilon > 0$ ，当样本量  $n$  足够大时，

$$\Pr(|\hat{\theta}_n - \theta^*| > \varepsilon) \rightarrow 0.$$

这叫“依概率收敛”(符号  $\xrightarrow{P}$ )。

小提醒：一致性是大样本性质；它不保证小样本表现好，也不要求“无偏”。一个估计器可以有小偏差，但只要样本增大时能靠近真值，就是一致的。

### 2) 为什么说“只有模型真正确”时才有意义？

- 定义里用了“真实分布就是  $p(\cdot | \theta^*)$ ”的假设。

- 现实中模型常常只是近似；这时我们讨论的是**拟合到的最优近似**，而不是“真实  $\theta^{*}$ ”。  
(扩展名词叫“伪真值”。)

### 3) 为什么 MLE 一致？（直觉）

- 对数似然的**平均值**随样本增大会靠近它的**期望**（大数定律）。
- 这个期望等于：常数  $E_{\theta^*}[\log p(X | \theta^*)]$  **减去**  $KL(p(\cdot | \theta^*) \| p(\cdot | \theta))$ 。
- 因此，最大化期望对数似然  $\Leftrightarrow$  **最小化 KL**。KL 的最小值是 0，且**只在**两分布相同 ( $\theta = \theta^*$ ) 时达到。
- 所以，当样本足够大时，MLE 会选到让 KL 最小的参数，因而**逼近真值**。

这需要一些常见的技术条件：模型**可识别**、真值在参数空间内部、对数似然够“好”（可微/上确界可达）等。

### 4) 一个直观小例子

- 正态均值未知、方差已知： $X_i \sim N(\theta^*, \sigma^2)$ 。
- MLE 是样本均值  $\bar{X}$ 。
- $E[\bar{X}] = \theta^*$ ,  $\text{Var}(\bar{X}) = \sigma^2/n$ 。方差随  $n$  以  $1/n$  缩小，因此  $\bar{X} \xrightarrow{P} \theta^*$ 。
- 这正是“一致”的典型示例：样本越多，估计越稳、越准。

### 5) 一致性 vs. 无偏性 vs. 方差

- 无偏性**:  $E[\hat{\theta}] = \theta^*$  (一锤子买卖的平均不偏)。
- 一致性**: 样本增大时概率上靠近真值（长期表现）。
- 标准误/方差**: 波动大小，通常随  $n$  以  $1/\sqrt{n}$  缩小。
- 一个估计器可以**有偏但一致**；也可能**无偏但不一致**（极端构造）。选择时要看整体权衡（如 MSE）。

## 小结

- 因为频率学派没有“自动最优”，我们看重一些**期望性质**；“一致性”是核心之一：数据越多，估计越接近真值。
- 在模型正确且常见条件下，**MLE 是一致的**：它本质在于“让估计分布接近真实分布”，即最小化 KL。
- 但一致性是**渐近概念**；实践中还要关注小样本下的偏差、方差、稳健性等。

翻译并解释以下内容，翻译不能有遗漏，解释请基础且详尽，因为我的数学基础不太好。

6.4.2 Unbiased estimators The bias of an estimator is defined as

$$\text{bias}(\hat{\theta}(\cdot)) = E_p(D|\theta^*) \hat{\theta}(D) - \theta^* \quad (6.32)$$

where  $\theta^*$  is the true parameter value. If the bias is zero, the estimator is called unbiased. This means the sampling distribution is centered on the true parameter. For example, the MLE for a Gaussian mean is unbiased:

$$\text{bias}(\hat{\mu}) = E[x] - \mu = E \frac{1}{N} \sum_{i=1}^N x_i - \mu = N\mu/N - \mu = 0 \quad (6.33)$$

However, the MLE for a Gaussian variance,  $\hat{\sigma}^2$ , is not an unbiased estimator of  $\sigma^2$ . In fact, one can show (Exercise 6.3) that

$$E \hat{\sigma}^2 = N - 1 N \sigma^2 \quad (6.34)$$

However, the following estimator

$$\hat{\sigma}^2 = \frac{1}{N-1} \sum_{i=1}^N (x_i - \bar{x})^2 \quad (6.35)$$

is an unbiased estimator, which we can easily prove as follows:

$$E \hat{\sigma}^2 = E \frac{1}{N-1} \sum_{i=1}^N (x_i - \bar{x})^2 = \frac{1}{N-1} \sum_{i=1}^N \sigma^2 = \sigma^2 \quad (6.36)$$

In Matlab, `var(X)` returns  $\hat{\sigma}^2$ , whereas `var(X,1)` returns  $\hat{\sigma}^2$  (the MLE). For large enough  $N$ , the difference will be negligible. Although the MLE may sometimes be a biased estimator, one can show that asymptotically, it is always unbiased. (This is necessary for the MLE to be a consistent estimator.) Although being unbiased sounds like a desirable property, this is not always true. See Section 6.4.4 and (Lindley 1972) for discussion of this point.

已思考 27s

下面先完整翻译，再给出基础且详尽的解释（含直观例子）。我会把原有的公式排版为标准写法，且不遗漏任何内容。

## 翻译

### 6.4.2 无偏估计 (Unbiased estimators)

估计器的偏差 (bias) 定义为

$$\text{bias}(\hat{\theta}(\cdot)) = \mathbb{E}_{p(D|\theta^*)} [\hat{\theta}(D) - \theta^{*}] \quad (6.32)$$

其中  $\theta^*$  是真实参数值。若偏差为 0，则称该估计器是无偏的。这意味着其**抽样分布**以真实参数为中心。

例如，高斯分布均值的 MLE 是无偏的：

$$\text{bias}(\bar{x}) = \mathbb{E}[\bar{x}] - \mu = \mathbb{E}\left[\frac{1}{N} \sum_{i=1}^N x_i\right] - \mu = \frac{N\mu}{N} - \mu = 0 \quad (6.33)$$

然而，高斯方差的 MLE（记作  $\hat{\sigma}^2$ ）并不是  $\sigma^2$  的无偏估计。事实上，可以证明（习题 6.3）

$$\mathbb{E}[\hat{\sigma}^2] = \frac{N-1}{N} \sigma^2 \quad (6.34)$$

不过，下面这个估计量

$$\hat{\sigma}_{N-1}^2 = \frac{N}{N-1} \hat{\sigma}^2 = \frac{1}{N-1} \sum_{i=1}^N (x_i - \bar{x})^2 \quad (6.35)$$

是一个无偏估计。我们可以很容易地证明：

$$\mathbb{E}[\hat{\sigma}_{N-1}^2] = \mathbb{E}\left[\frac{N}{N-1} \hat{\sigma}^2\right] = \frac{N}{N-1} \cdot \frac{N-1}{N} \sigma^2 = \sigma^2 \quad (6.36)$$

在 Matlab 中，`var(x)` 返回  $\hat{\sigma}_{N-1}^2$ ，而 `var(x,1)` 返回  $\hat{\sigma}^2$ （即 MLE）。当  $N$  足够大时，两者差别可以忽略。尽管 MLE 有时是有偏估计，但可以证明渐近地它总是无偏的。（这对于 MLE 成为一致估计是必要的。）尽管“无偏”听起来是个理想性质，但并非总是如此；见第 6.4.4 节和 (Lindley 1972) 的讨论。

## 解释（基础且详尽）

### 1) 什么是“偏差 (bias) ”？

- 定义：**偏差 = “平均估计值 – 真实值”。这里的“平均”是在**重复抽样下的期望**。
- 直观：**如果我们一遍又一遍地做相同实验、每次都用同样方法估计，所有估计值的平均恰好落在真值上  $\Rightarrow$  无偏；否则就是**有偏**（系统性偏高或偏低）。

### 2) 均值的 MLE 为什么无偏？（式 6.33）

- 对正态  $X_i \sim N(\mu, \sigma^2)$ ，样本均值  $\bar{x} = \frac{1}{N} \sum x_i$ 。
- 期望是线性的： $\mathbb{E}[\bar{x}] = \frac{1}{N} \sum \mathbb{E}[x_i] = \frac{1}{N} \cdot N\mu = \mu$ 。

- 所以  $\text{bias}(\hat{\mu}) = 0$ ; 抽样分布的“中心”就在  $\mu$ 。

### 3) 方差的 MLE 为什么有偏? (式 6.34)

- 常见的 (MLE) 样本方差形式是  $\hat{\sigma}^2 = \frac{1}{N} \sum (x_i - \bar{x})^2$ 。
- 由于<sup>\*\*</sup> $\bar{x}$  也是从数据估出来的<sup>\*\*</sup>, 它“吃掉”了 1 个自由度, 导致  $\hat{\sigma}^2$  的期望略小于  $\sigma^2$ :  $E[\hat{\sigma}^2] = \frac{N-1}{N} \sigma^2$ 。
- 这就是著名的 **Bessel 校正**: 把分母换成  $N - 1$  来补偿这个偏差。

### 4) 无偏的样本方差 (式 6.35–6.36)

- 定义  $\hat{\sigma}_{N-1}^2 = \frac{1}{N-1} \sum (x_i - \bar{x})^2$ 。
- 用上面结论一乘一除即可得  $E[\hat{\sigma}_{N-1}^2] = \sigma^2$ 。
- 软件差异**:
  - Matlab: `var(x)` 用  $N - 1$  (无偏), `var(x,1)` 用  $N$  (MLE)。
  - (额外常识) NumPy: `np.var(x)` 默认分母  $N$  (MLE), `np.var(x, ddof=1)` 用  $N - 1$ 。

**大小比较直观**: 若  $N = 5$ ,  $(N - 1)/N = 4/5 = 0.8$ 。真方差  $\sigma^2 = 10$  时, MLE 的期望只有 8, 偏小; 用  $N - 1$  修正后, 期望回到 10。

### 5) “渐近无偏”与“一致性”的关系

- 渐近无偏**: 随着样本量  $\rightarrow \infty$ , 偏差  $\rightarrow 0$ 。
- 一致性**: 估计值以概率收敛到真值。
- 一般来说, **一致性要求偏差消失** (否则估计不会收敛到真值)。在常见的正则条件下, MLE 渐近无偏且一致。

### 6) 为什么“无偏”不一定最好?

- 评价好坏常看 MSE (均方误差) = 方差 + 偏差<sup>2</sup>。
- 一个估计器即使有一点偏差, 只要能显著降低方差, MSE 可能更小 (整体更准)。
  - 典型例子: **收缩估计** (如 James-Stein) 会引入偏差, 但能降低方差, 总体 MSE 反而更优。
- 所以“无偏”只是众多理想性质之一; 实践中要权衡**偏差-方差折中** (见 6.4.4)。

## 小结

- 偏差度量**“平均意义上”估计是否系统性偏离真值; 无偏表示抽样分布以真值为中心。
- 正态均值的 MLE 无偏; 正态方差的 MLE 有负偏 (期望为  $(N - 1)/N \sigma^2$ ) , 用  $N - 1$  校正可得无偏估计。

- MLE 虽可能小样本有偏，但在一般条件下渐近无偏且一致。
- “无偏”并非总是更好；实际更看重 MSE 与稳健性/效率等综合表现。

翻译并解释以下内容，翻译不能有遗漏，解释请基础且详尽，因为我的数学基础不太好。

6.4.3 Minimum variance estimators It seems intuitively reasonable that we want our estimator to be unbiased (although we shall give some arguments against this claim below). However, being unbiased is not enough. For example, suppose we want to estimate the mean of a Gaussian from  $D = \{x_1, \dots, x_N\}$ . The estimator that just looks at the first data point,  $\hat{\theta}(D) = x_1$ , is an unbiased estimator, but will generally be further from  $\theta^*$  than the empirical mean  $\bar{x}$  (which is also unbiased). So the variance of an estimator is also important. A natural question is: how long can the variance go? A famous result, called the CramerRao lower bound, provides a lower bound on the variance of any unbiased estimator. More precisely, Theorem 6.4.1 (Cramer-Rao inequality). Let  $X_1, \dots, X_n \sim p(X|\theta_0)$  and  $\hat{\theta} = \hat{\theta}(x_1, \dots, x_n)$  be an unbiased estimator of  $\theta_0$ . Then, under various smoothness assumptions on  $p(X|\theta_0)$ , we have

$$\text{var } \hat{\theta}^* \geq \frac{1}{n} I(\theta_0) \quad (6.37)$$

where  $I(\theta_0)$  is the Fisher information matrix (see Section 6.2.2). A proof can be found e.g., in (Rice 1995, p275). It can be shown that the MLE achieves the Cramer Rao lower bound, and hence has the smallest asymptotic variance of any unbiased estimator. Thus MLE is said to be asymptotically optimal.

已思考 12s

下面先完整翻译，再给出基础且详尽的解释（含直观例子与要点提示）。原文里“how long can the variance go?” 应为“how low can the variance go?”（应是排版/措辞小瑕疵），我在解释部分会按“方差能最小到什么程度”来理解。

## 翻译

### 6.4.3 最小方差估计器 (Minimum variance estimators)

直觉上我们希望估计器是无偏的（尽管我们稍后会给出反对这一主张的一些论据）。然而，仅仅无偏还不够。举例来说，假设我们要用  $D = \{x_1, \dots, x_N\}$  来估计高斯分布的均值。那个只看第一个数据点的估计器， $\hat{\theta}(D) = x_1$ ，是一个无偏估计器，但通常会比经验均值  $\bar{x}$ （它也是无偏的）离  $\theta^*$  更远。所以估计器的**方差**也很重要。一个自然的问题是：**方差能多小？**一个著名结果，称为 Cramér–Rao 下界，给出了任何无偏估计器的方差的下界。更精确地说，

**定理 6.4.1 (Cramér–Rao 不等式)**。设  $X_1, \dots, X_n \sim p(X | \theta_0)$ ，并且  $\hat{\theta} = \hat{\theta}(x_1, \dots, x_n)$  是  $\theta_0$  的无偏估计器。则在关于  $p(X | \theta_0)$  的若干光滑性假设下，有

$$\text{var}(\hat{\theta}) \geq \frac{1}{n I(\theta_0)} \quad (6.37)$$

其中  $I(\theta_0)$  是费舍尔信息矩阵（见第 6.2.2 节）。一个证明可见例如 (Rice 1995, 第 275 页)。可以证明，MLE 达到了 Cramér–Rao 下界，因此在任意无偏估计器中具有**最小的渐近方差**。于是说 MLE 是**渐近最优的**。

---

## 解释 (基础且详尽)

### 1) 为什么“无偏还不够”？

- $\hat{\theta}_1(D) = x_1$  与  $\hat{\theta}_2(D) = \bar{x}$ （样本均值）都是**无偏**： $E[\hat{\theta}] = \theta^*$ 。
- 但  $\text{var}(x_1) = \sigma^2$ ，而  $\text{var}(\bar{x}) = \sigma^2/N$ 。通常  $\bar{x}$  离真值更近（波动更小）。
- 结论：除了“平均不偏”，我们还要控制**方差**——这关乎估计的“稳不稳”。

### 2) Cramér–Rao 下界 (CRLB) 说了什么？

- 在满足常见**正则条件**（模型可微、积分交换合法、支持集与  $\theta$  无关等）时，任何**无偏**估计器的方差都有一个**不可逾越的下界**：

$$\text{var}(\hat{\theta}) \geq \frac{1}{n I(\theta_0)}.$$

- 这里  $I(\theta_0)$  是**每个样本**的费舍尔信息；总信息量是  $nI(\theta_0)$ （信息随样本数线性增加）。
- **含义**：信息越多 ( $I$  越大)，下界越低；样本越多 ( $n$  越大)，方差下界按  $1/n$  降低。
- **多维参数**时，形式是协方差矩阵不等式： $\text{Cov}(\hat{\theta}) \geq (n I(\theta_0))^{-1}$ （半正定意义）。

直觉图景：费舍尔信息  $\approx$  似然在真值附近的**曲率**；曲率越大，“峰”越尖，参数越容易被

锁定，方差就越小。

### 3) MLE 为何“渐近最优”？

- 在适当条件下，MLE **渐近正态**:  $\hat{\theta}_{\text{MLE}} \xrightarrow{d} N(\theta_0, (nI(\theta_0))^{-1})$ 。
- 这等于说：**大样本时**，MLE 的方差正好**达到 CR 下界**；因此在**所有无偏估计器**中，它的**渐近方差最小（渐近有效/efficient）**。

### 4) 一个具体例子：正态均值（方差已知）

- 设  $X_i \sim N(\mu, \sigma^2)$ ,  $\mu$  未知,  $\sigma^2$  已知。
- 每个样本的费舍尔信息  $I(\mu) = 1/\sigma^2$ 。因此下界是

$$\text{var}(\hat{\mu}) \geq \frac{1}{n(1/\sigma^2)} = \frac{\sigma^2}{n}.$$

- 样本均值  $\bar{X}$  的方差正是  $\sigma^2/n$ , 恰好**达到下界** → 最优（在无偏类里）。

### 5) CRLB 的边界与误用提示

- 只对无偏估计器**给出下界。若允许**有偏**，可以出现方差更小但有偏差的估计器；它们的MSE 可能更优（典型：收缩估计、James–Stein）。
- 正则条件**不满足时（如支持集依参数改变、不可微等），CRLB 可能不适用或要换用推广形式。
- 小样本下**，即便 MLE 无偏或近似无偏，它也不一定“恰好等于下界”；“达到下界”是**渐近结论**。

### 6) 把几件事串起来

- 我们希望：**尽量无偏、方差尽量小**。
- CRLB 告诉我们：**在无偏**这条赛道上，“方差小到什么程度”的**终极极限**；MLE 在大样本会“贴着极限跑”。
- 但实践中常更看重**MSE = 方差 + 偏差<sup>2</sup>**的整体表现；适度引入偏差换取大幅降方差，常能得到更小的 MSE（这就是**偏差–方差权衡**）。

## 小结

- 仅无偏不够**；我们还关心方差。
- Cramér–Rao 下界**：任何无偏估计器的方差  $\geq 1/(nI(\theta_0))$ 。
- MLE 渐近有效**：在一般条件下，它的渐近方差等于下界。
- 现实选择估计器，要结合 **MSE、稳健性、先验信息** 等综合考量，而不仅仅“要无偏”。

翻译并解释以下内容，翻译不能有遗漏，解释请基础且详尽，因为我的数学基础不太好。

6.4.4 The bias-variance tradeoff Although using an unbiased estimator seems like a good idea, this is not always the case. To see why, suppose we use quadratic loss. As we showed above, the corresponding risk is the MSE. We now derive a very useful decomposition of the MSE. (All expectations and variances are wrt the true distribution  $p(D|\theta^*)$ , but we drop the explicit conditioning for notational brevity.) Let  $\hat{\theta} = \hat{\theta}(D)$  denote the estimate, and  $\bar{\theta} = E[\hat{\theta}]$  denote the expected value of the estimate (as we vary D). Then we have

$$E[(\hat{\theta} - \theta^*)^2] = \text{var}[\hat{\theta}] + \text{bias}^2(\hat{\theta})$$

In words,

$$\text{MSE} = \text{variance} + \text{bias}^2 \quad (6.42)$$

This is called the bias-variance tradeoff (see e.g., (Geman et al. 1992)). What it means is that it might be wise to use a biased estimator, so long as it reduces our variance, assuming our goal is to minimize squared error.

已思考 8s

下面先完整翻译，再给出基础且详尽的解释（含一步步推导与直观例子）。我会保留所有公式与含义，不做遗漏。

## 翻译

### 6.4.4 偏差-方差折中 (The bias-variance tradeoff)

虽然使用无偏估计器看起来是个好主意，但这并不总是正确的。为说明原因，假设我们采用二次损失。如上所示，相应的风险就是 MSE (均方误差)。下面我们推导一个对 MSE 的非常有用的分解。（所有的期望与方差都是相对于真实分布  $p(D | \theta^*)$  而言的，但为简洁起见，我们在符号上省略这种条件。）令  $\hat{\theta} = \hat{\theta}(D)$  表示估计值，令  $\bar{\theta} = E[\hat{\theta}]$  表示当数据  $D$  变化时估计值的期望。于是有

$$E[(\hat{\theta} - \theta^*)^2] = \text{var}[\hat{\theta}] + \text{bias}^2(\hat{\theta})$$

用文字表述，即

$$\text{MSE} = \text{variance} + \text{bias}^2 \quad (6.42)$$

这被称为**偏差-方差折中**（参见例如 Geman 等，1992）。其含义是：只要我们的目标是最小化平方误差，那么使用一个有偏的估计器也可能是明智的，前提是它能降低方差。

---

## 解释（基础且详尽）

### 1) 背景与符号

- 我们在**二次损失**  $L(\theta, \hat{\theta}) = (\hat{\theta} - \theta)^2$  下讨论频率学派的**风险**：  
 $\text{Risk} = E[(\hat{\theta} - \theta^*)^2] = \text{MSE}$ 。
- 这里所有的  $E[\cdot]$ 、 $\text{var}(\cdot)$  都是对“若干次重复抽样得到的不同数据集  $D$ ”取的期望/方差（条件在真实参数  $\theta^*$  上）。

**记号：**

$\hat{\theta} = \hat{\theta}(D)$  (估计器在数据上的输出)，  $\bar{\theta} = E[\hat{\theta}]$  (抽样平均意义下的估计期望)。

**偏差**:  $\text{bias}(\hat{\theta}) = E[\hat{\theta}] - \theta^* = \bar{\theta} - \theta^*$ 。

**方差**:  $\text{var}(\hat{\theta}) = E[(\hat{\theta} - \bar{\theta})^2]$ 。

### 2) 一步步推导 MSE 分解

把  $\hat{\theta} - \theta^*$  拆成“波动部分 + 系统性差距”：

$$\hat{\theta} - \theta^* = (\hat{\theta} - \bar{\theta}) + (\bar{\theta} - \theta^*).$$

两边平方并取期望：

$$\begin{aligned} E[(\hat{\theta} - \theta^*)^2] &= E[(\hat{\theta} - \bar{\theta})^2] + 2 E[(\hat{\theta} - \bar{\theta})(\bar{\theta} - \theta^*)] + (\bar{\theta} - \theta^*)^2 \\ &= \underbrace{\text{var}(\hat{\theta})}_{\text{波动}} + \underbrace{0}_{\text{交叉项为 } 0, \text{ 因为 } E[\hat{\theta} - \bar{\theta}] = 0} + \underbrace{(\bar{\theta} - \theta^*)^2}_{\text{偏差}^2}. \end{aligned}$$

于是得到

$$\text{MSE} = \text{var}(\hat{\theta}) + \text{bias}^2(\hat{\theta}) \quad (\text{式 6.42}).$$

### 3) 这句话到底在说什么？

- 方差大**: 估计值在不同样本之间**摇摆很厉害**（不稳定、易过拟合）。
- 偏差大**: 估计器系统性地**偏离真值**（模型过于简单、欠拟合）。
- MSE**把两者加总：
  - 若能显著降低方差，即使引入**小偏差**，整体 MSE 也可能**更小**；

- 反之，过度追求无偏（偏差 $\approx 0$ ）可能导致方差很大、MSE 反而变差。  
这就是“折中”：通常不存在同时让“偏差 $\approx 0$  且方差极小”的魔法方案。

## 4) 三个直观小例子

### 例子 A：只看第一个样本 vs. 用样本均值

估计正态均值  $\mu$ , 方差已知为  $\sigma^2$ 。

- 估计器1:  $\hat{\mu}_1 = x_1$ 。无偏，但方差  $\sigma^2$  (大)。
- 估计器2:  $\hat{\mu}_2 = \bar{x}$ 。无偏，方差  $\sigma^2/N$  (小很多)。

**启示：**都无偏，但**方差不同**  $\rightarrow \bar{x}$  的 MSE 更小。

### 例子 B：收缩估计 (带来偏差，换取更小方差)

$$\hat{\mu}_\alpha = \alpha \bar{x} + (1 - \alpha) \mu_0 \quad (0 \leq \alpha \leq 1).$$

- 偏差:  $\text{bias} = E[\hat{\mu}_\alpha] - \mu = \alpha\mu + (1 - \alpha)\mu_0 - \mu = (1 - \alpha)(\mu_0 - \mu)$ 。
- 方差:  $\text{var} = \alpha^2 \sigma^2/N$ 。
- MSE:  $\alpha^2 \sigma^2/N + (1 - \alpha)^2(\mu_0 - \mu)^2$ 。

若  $\mu_0$  选得较接近真值、样本又很小，此时选  $\alpha < 1$  (略收缩) 可显著降方差，即使引入小偏差，总体 MSE 更低。

(这正是岭回归/贝叶斯后验均值/James–Stein 的精神。)

### 例子 C：固定常数

$\hat{\theta} \equiv \theta_0$ 。

- 方差 = 0 (完全稳定)，但偏差 =  $\theta_0 - \theta^*$ 。
- 若  $\theta_0$  与真值相去甚远，则 MSE 很大。

**启示：**把方差压成 0 的代价可能是巨大的偏差。

## 5) 在机器学习里的常见影子

- **高方差**  $\leftrightarrow$  **过拟合**：模型太复杂、对训练数据的噪声也“记住”了。
- **高偏差**  $\leftrightarrow$  **欠拟合**：模型太简单，学不出数据里的规律。
- 常见控方差/引入偏差的手段：**正则化** (L2/L1)、**早停**、**特征降维**、**bagging/随机森林**、**收缩 (shrinkage)** 等。

## 6) 何时追求无偏？何时允许有偏？

- 若存在**无偏且方差已接近 CRLB 的估计器** (如正态均值的样本均值)，那当然理想。
- 但在**小样本、模型复杂或噪声大时**，适度引入偏差以**大幅降方差**，整体 MSE 往往更优  
(这在实践中非常常见)。

## 小结

- 在二次损失下，风险=MSE，并且

$$\text{MSE} = \text{方差} + \text{偏差}^2.$$

- 这揭示了“偏差-方差折中”：为了降低总体误差，有时宁可用一点偏差换取明显更小的方差。
- 选择估计器时，不要只盯着“无偏”，而要看**总体 MSE**与实际稳定性、鲁棒性。



已上传的图片

翻译并解释以下内容，翻译不能有遗漏，解释请基础且详尽，因为我的数学基础不太好。

Figure 6.4 Left: Sampling distribution of the MAP estimate with different prior strengths  $\kappa_0$ . (The MLE corresponds to  $\kappa_0 = 0$ .) Right: MSE relative to that of the MLE versus sample size. Based on Figure 5.6 of (Hoff 2009). Figure generated by samplingDistGaussShrinkage.

6.4.4.1 Example: estimating a Gaussian mean Let us give an example, based on (Hoff 2009, p79). Suppose we want to estimate the mean of a Gaussian from  $x = (x_1, \dots, x_N)$ . We assume the data is sampled from  $x_i \sim N(\theta^* = 1, \sigma^2)$ . An obvious estimate is the MLE. This has a bias of 0 and a variance of

$$\text{var}[x|\theta^*] = \sigma^2 N \quad (6.43)$$

But we could also use a MAP estimate. In Section 4.6.1, we show that the MAP estimate under a Gaussian prior of the form  $N(\theta_0, \sigma^2/\kappa_0)$  is given by

$$\hat{x} = N N + \kappa_0 x + \kappa_0 N + \kappa_0 \theta_0 = w x + (1 - w) \theta_0 \quad (6.44)$$

where  $0 \leq w \leq 1$  controls how much we trust the MLE compared to our prior. (This is also the posterior mean, since the mean and mode of a Gaussian are the same.) The bias and variance are given by

$$E[\tilde{x}] - \theta^* = w\theta_0 + (1-w)\theta_0 - \theta^* = (1-w)(\theta_0 - \theta^*) \quad (6.45)$$

$$\text{var}[\tilde{x}] = w^2 \sigma^2 N \quad (6.46)$$

So although the MAP estimate is biased (assuming  $w < 1$ ), it has lower variance. Let us assume that our prior is slightly misspecified, so we use  $\theta_0 = 0$ , whereas the truth is  $\theta^* = 1$ . In Figure 6.4(a), we see that the sampling distribution of the MAP estimate for  $\kappa_0 > 0$  is biased away from the truth, but has lower variance (is narrower) than that of the MLE. In Figure 6.4(b), we plot  $\text{mse}(\tilde{x})/\text{mse}(x)$  vs  $N$ . We see that the MAP estimate has lower MSE than the MLE, especially for small sample size, for  $\kappa_0 \in \{1, 2\}$ . The case  $\kappa_0 = 0$  corresponds to the MLE, and the case  $\kappa_0 = 3$  corresponds to a strong prior, which hurts performance because the prior mean is wrong. It is clearly important to “tune” the strength of the prior, a topic we discuss later.

已思考 13s

下面先完整翻译，再给出基础且详尽的解释（包含小心纠正一个排版小瑕疵，并给直观结论）。

---

## 翻译

**图 6.4 左：**不同先验强度  $\kappa_0$  下，MAP 估计的抽样分布。（当  $\kappa_0 = 0$  时对应 MLE。）  
**右：**相对于 MLE 的 MSE（均方误差）随样本量变化的曲线。图基于 (Hoff 2009) 的图 5.6。图由 `samplingDistGaussShrinkage` 生成。

### 6.4.4.1 例子：估计高斯均值

给出一个基于 (Hoff 2009, p79) 的例子。假设我们要用  $x = (x_1, \dots, x_N)$  来估计高斯分布的均值。我们假设数据来自  $x_i \sim N(\theta^{*} = 1, \sigma^2)$ 。一个显然的估计是 MLE。它的偏差为 0，方差为

$$\text{var}[\bar{x} \mid \theta^{*}] = \frac{\sigma^2}{N} \quad (6.43)$$

但我们也使用 MAP 估计。在 4.6.1 节我们展示：在高斯先验  $N(\theta_0, \sigma^2/\kappa_0)$  下，MAP 估计为

$$\tilde{x} = \frac{N}{N + \kappa_0} \bar{x} + \frac{\kappa_0}{N + \kappa_0} \theta_0 = w\bar{x} + (1 - w)\theta_0 \quad (6.44)$$

其中  $0 \leq w \leq 1$  控制我们相对“信任 MLE (样本均值) 还是先验”的程度。 (这同时也是**后验均值**, 因为高斯的均值与众数相同。 ) 其偏差与方差为

$$E[\tilde{x}] - \theta^* = w\theta_0 + (1-w)\bar{x} - \theta^* = (1-w)(\theta_0 - \theta^*) \quad (6.45)$$

$$\text{var}[\tilde{x}] = w^2 \frac{\sigma^2}{N} \quad (6.46)$$

因此, 虽然 MAP 估计在  $w < 1$  时是有偏的, 但它具有**更小的方差**。现在假设我们的先验略有错配, 因此取  $\theta_0 = 0$ , 而真实值是  $\theta^* = 1$ 。在图 6.4(a) 中可以看到, 对  $\kappa_0 > 0$  的情形, MAP 估计的抽样分布相对真实值会产生偏移, 但其方差比 MLE 更小 (分布更窄)。在图 6.4(b) 中, 我们绘制了  $\text{mse}(\tilde{x})/\text{mse}(\bar{x})$  随  $N$  的变化。可以看到, 对  $\kappa_0 \in \{1, 2\}$  而言, MAP 的 MSE 小于 MLE, 尤其在**小样本**时更明显。 $\kappa_0 = 0$  对应 MLE,  $\kappa_0 = 3$  对应一个较强的先验, 由于先验均值是错的, 会损害性能。显然, 应当对先验强度进行“调谐” (这一点我们将在后面讨论)。

注: 式 (6.45) 的第一步有排版小瑕疵; 正确推导是  $E[\tilde{x}] = w E[\bar{x}] + (1-w)\theta_0 = w\theta^* + (1-w)\theta_0$ , 因此偏差  $E[\tilde{x}] - \theta^* = (1-w)(\theta_0 - \theta^*)$  (结论与原式一致)。

## 解释 (基础且详尽)

### 1) 设定与关键公式

- 数据:  $x_i \sim N(\theta^*, \sigma^2)$ , 真实均值  $\theta^* = 1$ 。
- MLE (样本均值  $\bar{x}$ ) : 无偏, 方差  $\sigma^2/N$ 。
- MAP/后验均值 (收缩估计) :

$$\tilde{x} = w\bar{x} + (1-w)\theta_0, \quad w = \frac{N}{N + \kappa_0}.$$

- 偏差:**  $\text{bias} = (1-w)(\theta_0 - \theta^*)$ 。
- 方差:**  $\text{var} = w^2 \sigma^2/N$  (比 MLE 小一个  $w^2$  倍)。

这就是**偏差-方差折中**中的标准例子: 引入一点偏差, 换取明显降低的方差。

### 2) 读图 6.4(a): 抽样分布形状

- 先验均值设为  $\theta_0 = 0$ , 真值  $\theta^* = 1$ , 样本数  $n = 5$ 。
- $\kappa_0 > 0$  时,  $\tilde{x}$  的分布向  $\theta_0$  偏, 但更窄 (方差更小) ——这就是**收缩**。
- $\kappa_0 = 0$  就是 MLE (蓝圈), 中心在 1, 分布较宽。

### 3) 读图 6.4(b): 相对 MSE 随样本量

- 纵轴:  $\text{MSE}(\tilde{x})/\text{MSE}(\bar{x})$ ; 小于 1 代表 MAP 优于 MLE。
- 结论:
  - 小样本:**  $\kappa_0 = 1, 2$  (红叉、黑星) 显著低于 1, 优势明显;
  - $N$  增大:  $w \rightarrow 1$ , MAP 渐近回到 MLE, 曲线靠近 1;
  - 先验过强:**  $\kappa_0 = 3$  (绿三角) 始终  $\geq 1$ , 因先验均值错得多, 偏差成本压过了方差收益。

这说明**先验强度需要调节** ( $\kappa_0$  不能太大也不能太小) :

- 太弱: 几乎就是 MLE, 没得到降方差的好处;
- 太强: 如果  $\theta_0$  有误, 就会引入过大偏差, MSE 变差。

#### 4) 把它与偏差-方差折中联系起来

- MSE = 方差 + 偏差<sup>2</sup>**。
- MAP 相比 MLE: 方差减少为  $w^2$  倍, 但偏差增为  $(1 - w)|\theta_0 - \theta^{*}|$ 。
- 小样本** ( $w$  小) : 方差减少很多, 偏差增加有限  $\Rightarrow$  **总体更好**;
- 大样本** ( $w$  接近 1) : 两者几乎一样;
- 强错先验:** 偏差<sup>2</sup>太大  $\Rightarrow$  **总体更差**。

#### 5) 实用建议 (直观)

- 当有**合理但不完美的先验** (如历史经验、物理范围、行业常识) 且样本不多时, 用**温和的收缩** ( $\kappa_0$  取小到中等) 常能显著降低 MSE。
- 可用**交叉验证/经验贝叶斯**等方法来“调”  $\kappa_0$  (或等价地调  $w$ )。
- 随着数据增多, 自动更依赖数据 ( $w \uparrow$ ), 先验影响自然减弱。

### 小结

- MAP/后验均值是一个**收缩估计器**:  $\tilde{x} = w\bar{x} + (1 - w)\theta_0$ 。
- 它以**偏换方**: 引入偏差  $(1 - w)(\theta_0 - \theta^{*})$ , 换来更小方差  $w^2\sigma^2/N$ 。
- 小样本**时合适的  $\kappa_0$  能显著降低 MSE; **大样本**时与 MLE 等价; **先验过强且均值错误**会伤害性能。
- 图 6.4 直观展示了这一切: 左图看分布 (偏 + 窄), 右图看相对 MSE (随  $N$  接近 1,  $\kappa_0$  过大变差)。



翻译并解释以下内容，翻译不能有遗漏，解释请基础且详尽，因为我的数学基础不太好。

Figure 6.5 Illustration of bias-variance tradeoff for ridge regression. We generate 100 data sets from the true function, shown in solid green. Left: we plot the regularized fit for 20 different data sets. We use linear regression with a Gaussian RBF expansion, with 25 centers evenly spread over the [0, 1] interval. Right: we plot the average of the fits, averaged over all 100 datasets. Top row: strongly regularized: we see that the individual fits are similar to each other (low variance), but the average is far from the truth (high bias). Bottom row: lightly regularized: we see that the individual fits are quite different from each other (high variance), but the average is close to the truth (low bias). Based on (Bishop 2006a) Figure 3.5. Figure generated by biasVarModelComplexity3.

6.4.4.2 Example: ridge regression Another important example of the bias variance tradeoff arises in ridge regression, which we discuss in Section 7.5. In brief, this corresponds to MAP estimation for linear regression under a Gaussian prior,  $p(w) = N(w|0, \lambda^{-1}I)$ . The zero-mean prior encourages the weights to be small, which reduces overfitting; the precision term,  $\lambda$ , controls the strength of this prior. Setting  $\lambda = 0$  results in the MLE; using  $\lambda > 0$  results in a biased estimate. To illustrate the effect on the variance, consider a simple example. Figure 6.5 on the left plots each individual fitted curve, and on the right plots the average fitted curve. We see that as we increase the

strength of the regularizer, the variance decreases, but the bias increases.

#### 6.4.4.3 Bias-variance tradeoff for classification

If we use 0-1 loss instead of squared error, the above analysis breaks down, since the frequentist risk is no longer expressible as squared bias plus variance. In fact, one can show (Exercise 7.2 of (Hastie et al. 2009)) that the bias and variance combine multiplicatively. If the estimate is on relative MSE the correct side of the decision boundary, then the bias is negative, and decreasing the variance will decrease the misclassification rate. But if the estimate is on the wrong side of the decision boundary, then the bias is positive, so it pays to increase the variance (Friedman 1997a). This little known fact illustrates that the bias-variance tradeoff is not very useful for classification. It is better to focus on expected loss (see below), not directly on bias and variance. We can approximate the expected loss using cross validation, as we discuss in Section 6.5.3.

已思考 23s

下面先完整翻译，再给出基础且详尽的解释（含直观读图与要点）。我会保留式子与术语；文末一段里原文出现了小排版瑕疵（“on relative MSE the correct side”），根据上下文应为“on the correct side”。

## 翻译

**图 6.5** 用岭回归 (ridge regression) 说明偏差-方差折中。我们从真实函数 (绿色实线所示) 生成 100 个数据集。**左列：**对 20 个不同数据集画出正则化后的拟合曲线。我们用**带高斯 RBF 展开的线性回归**，在区间  $[0, 1]$  上等距放置 25 个中心。**右列：**把 100 个数据集得到的拟合曲线做**平均**后绘制。**上排：****强正则化**；可见各次拟合彼此很相似 (**低方差**)，但平均曲线与真相相差很远 (**高偏差**)。**下排：****弱正则化**；可见各次拟合差异很大 (**高方差**)，但平均曲线更接近真相 (**低偏差**)。基于 (Bishop 2006a) 图 3.5。图由 `biasVarModelComplexity3` 生成。

#### 6.4.4.2 例子：岭回归

偏差-方差折中的另一个重要例子出现在岭回归（第 7.5 节将讨论）。简言之，这对应于在**高斯先验**  $p(\mathbf{w}) = N(\mathbf{w} \mid \mathbf{0}, \lambda^{-1} \mathbf{I})$  下做线性回归的 MAP 估计。零均值先验会鼓励权重变小，从而减少过拟合； $\lambda$  (**精度**) 控制先验强度。取  $\lambda = 0$  得到 MLE；用  $\lambda > 0$  会得到有

偏估计。为说明对方差的影响，考虑一个简单例子。图 6.5 左边画的是每个单独数据集的拟合曲线，右边画的是平均拟合曲线。可以看到，随着正则化强度增大，**方差下降**，但**偏差上升**。

#### 6.4.4.3 分类中的偏差-方差折中

如果我们使用 **0-1 损失** 而不是平方误差，上面的分析就失效了，因为此时频率学派风险不再能写成“偏差平方 + 方差”。事实上（见 Hastie 等 2009 的习题 7.2），偏差和方差会以**乘法**方式组合。若估计值位于**决策边界的正确一侧**（原文处有小排版瑕疵），则偏差为**负**，并且**降低方差会降低误分类率**；但如果估计值位于**错误一侧**，偏差为**正**，这时**增加方差反而有益** (Friedman 1997a)。这个不太为人所知的事实说明：对分类任务来说，“偏差-方差折中”并不太有用。更好的做法是关注**期望损失**（见下文），而不是直接盯着偏差和方差。我们可以用**交叉验证**来近似期望损失，第 6.5.3 节将讨论。

## 解释（基础且详尽）

### 1) 图 6.5 在说什么？

- 我们用很多**重复采样**的数据来观察不同正则强度下的拟合表现。
- **左列（单次拟合）**：看“波动性”——同一设置在不同数据集上的差异，就是**方差**。
- **右列（平均拟合）**：把波动“平均掉”后，看平均曲线与真相的偏离，就是**偏差**。
- **强正则（上排）**：模型被“勒住”，曲线很平；各次结果相似（低方差），但平均明显偏离真函数（高偏差）→ **欠拟合**。
- **弱正则（下排）**：模型自由度高，能紧贴每份数据；各次结果差很多（高方差），平均接近真函数（低偏差）→ **过拟合风险大**。

这正是“**以偏换方**”的可视化：加大正则 → 偏差↑、方差↓；减小正则 → 偏差↓、方差↑。最佳点通常在两者之间，需要调参找到。

### 2) 岭回归为何体现偏差-方差折中？

- 岭回归目标： $\min_{\mathbf{w}} \|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2 + \lambda \|\mathbf{w}\|_2^2$ 。
- 等价**贝叶斯**表述： $\mathbf{w} \sim N(\mathbf{0}, \lambda^{-1} \mathbf{I})$  的 MAP。
- $\lambda > 0$  让权重靠近 0（**收缩**），所以**方差下降**；但同时引入了**偏差**（预测被“拉回中庸”）。
- 图里使用了**RBF 基函数展开**（25 个等距中心），可看作把线性模型变成更灵活的非线性回归；这时正则尤为关键，否则极易过拟合。

直觉： $\lambda$  像弹簧的“拉力”。拉力大（强正则），曲线拉直，稳但偏；拉力小（弱正则），曲线灵活，准但抖。

### 3) 如何选择 $\lambda$ ?

- 没有放之四海而皆准的固定值；应通过**数据驱动的方法**（如**交叉验证、验证集、或边际似然/经验贝叶斯**）来选。
- 小样本/噪声大 → 往往需要**更强正则**；大样本/信噪比高 → 可用**更弱正则**。

### 4) 为什么“偏差-方差”在\*\*分类 (0-1 损失) \*\*下不太好用？

- 回归（平方损失）里有漂亮的分解： $MSE = \text{方差} + \text{偏差}^2$ 。
- 分类 (0-1 损失) 下，**风险不是它们的加法**。结论变成：
  - 若预测分布的**平均在正确一侧**（负偏差），减少方差会降低越界的机会 → **错分率** ↓；
  - 若平均在**错误一侧**（正偏差），**适度增大方差**有时反而能“把一部分样本抛回正确侧”，平均错分率**可能下降**。
- 因此，直接调“偏差/方差”不如直接最小化**期望 0-1 损失**（或其可替代的凸损失，如对数损失/hinge 损失），并用**交叉验证**来估计与选择。

### 5) 从工程角度的要点

- **回归**：把  $\lambda$  当作“旋钮”，在\*\*欠拟合（高偏差）与过拟合（高方差）\*\*之间寻优。
  - **分类**：别纠结偏差或方差本身，直接以**验证误差/交叉验证误差**为目标来选模型与超参。
  - RBF 展开或其他高维特征时，正则（如岭/L2、Lasso/L1）几乎是必需品。
- 

## 小结

- 图 6.5 直观展示：正则越强 → **低方差/高偏差**；正则越弱 → **高方差/低偏差**。
- 岭回归 = 线性回归 + L2 正则 = 高斯先验下的 MAP； $\lambda$  控强度，决定偏差-方差位置。
- 在**分类的 0-1 损失**下，偏差-方差“加法分解”不成立；更应关注**期望损失**，用**交叉验证**做选择。

翻译并解释以下内容，翻译不能有遗漏，解释请基础且详尽，因为我的数学基础不太好。

6.5 Empirical risk minimization

Frequentist decision theory suffers from the fundamental

problem that one cannot actually compute the risk function, since it relies on knowing the true data distribution. (By contrast, the Bayesian posterior expected loss can always be computed, since it conditions on the data rather than conditioning on  $\theta^*$ .) However, there is one setting which avoids this problem, and that is where the task is to predict observable quantities, as opposed to estimating hidden variables or parameters. That is, instead of looking at loss functions of the form  $L(\theta, \delta(D))$ , where  $\theta$  is the true but unknown parameter, and  $\delta(D)$  is our estimator, let us look at loss functions of the form  $L(y, \delta(x))$ , where  $y$  is the true but unknown response, and  $\delta(x)$  is our prediction given the input  $x$ . In this case, the frequentist risk becomes

$$R(p^*, \delta) = E(x,y) \sim p^* [L(y, \delta(x))] = \sum_{i=1}^N L(y_i, \delta(x_i)) p^*(x_i, y_i) \quad (6.47)$$

where  $p^*$  represents "nature's distribution". Of course, this distribution is unknown, but a simple approach is to use the empirical distribution, derived from some training data, to approximate  $p^*$ , i.e.,

$$p^*(x, y) \approx p_{\text{emp}}(x, y) = \frac{1}{N} \sum_{i=1}^N \delta(x_i, y_i) \quad (6.48)$$

We then define the empirical risk as follows:

$$R_{\text{emp}}(D, \delta) = R(p_{\text{emp}}, \delta) = \frac{1}{N} \sum_{i=1}^N L(y_i, \delta(x_i)) \quad (6.49)$$

In the case of 0-1 loss,  $L(y, \delta(x)) = I(y = \delta(x))$ , this becomes the misclassification rate. In the case of squared error loss,  $L(y, \delta(x)) = (y - \delta(x))^2$ , this becomes the mean squared error. We define the task of empirical risk minimization or ERM as finding a decision procedure (typically a classification rule) to minimize the empirical risk:

$$\delta_{\text{ERM}}(D) = \operatorname{argmin}_{\delta} R_{\text{emp}}(D, \delta) \quad (6.50)$$

In the unsupervised case, we eliminate all references to  $y$ , and replace  $L(y, \delta(x))$  with  $L(x, \delta(x))$ , where, for example,  $L(x, \delta(x)) = \|x - \delta(x)\|^2$ , which measures the reconstruction error. We can define the decision rule using  $\delta(x) = \text{decode}(\text{encode}(x))$ , as in vector quantization (Section 11.4.2.6) or PCA (section 12.2). Finally, we define the empirical risk as

$$R_{\text{emp}}(D, \delta) = \frac{1}{N} \sum_{i=1}^N L(x_i, \delta(x_i)) \quad (6.51)$$

Of course, we can always trivially minimize this risk by setting  $\delta(x) = x$ , so it is critical that the encoder-decoder go via some kind of bottleneck.

已思考 20s

下面先逐字翻译，再给出基础且详尽的解释（把符号讲清楚、举直观例子）。个别公式处原文有轻微排版小瑕疵（如  $R_{\text{emp}}(D, D)$  应为  $R_{\text{emp}}(D, \delta)$ ），我会在解释里说明，但翻译不遗漏任何内容。

---

## 翻译

### 6.5 经验风险最小化 (Empirical risk minimization)

频率学派决策理论的一个根本问题是：**风险函数实际上无法计算**，因为它依赖于已知真实数据分布这一点。（相比之下，**贝叶斯**的后验期望损失总是可计算的，因为它**以数据为条件**，而不是以  $\theta^{*}$  为条件。）不过有一种情形可以回避这个问题，那就是**任务是预测可观测量**，而不是估计隐变量或参数。也就是说，我们不再看形如  $L(\theta, \delta(D))$  的损失函数（其中  $\theta$  是真实但未知的参数， $\delta(D)$  是我们的估计器），而去看形如  $L(y, \delta(x))$  的损失函数，其中  $y$  是真实但未知的**响应**， $\delta(x)$  是给定输入  $x$  时我们的**预测**。在这种情况下，频率学派的风险变为

$$R(p^{*}, \delta) = E_{(x,y) \sim p^{*}} [L(y, \delta(x))] = \int \int L(y, \delta(x)) p^{*}(x, y) dx dy \quad (6.47)$$

其中  $p^{*}$  表示“自然的分布”。当然，这个分布是未知的，但一个简单方法是：**用来自训练数据的经验分布去近似  $p^{*}$** ，即

$$p^{*}(x, y) \approx p_{\text{emp}}(x, y) = \frac{1}{N} \sum_{i=1}^N \delta_{x_i}(x) \delta_{y_i}(y) \quad (6.48)$$

于是定义**经验风险**为

$$R_{\text{emp}}(D, \delta) = R(p_{\text{emp}}, \delta) = \frac{1}{N} \sum_{i=1}^N L(y_i, \delta(x_i)) \quad (6.49)$$

当采用**0-1 损失**  $L(y, \delta(x)) = \mathbf{1}\{y = \delta(x)\}$  时，它变成**误分类率**；当采用**平方误差**  $L(y, \delta(x)) = (y - \delta(x))^2$  时，它变成**均方误差**。

我们把**经验风险最小化 (ERM)** \*\*定义为：寻找一个决策过程（通常是分类规则）去最小化经验风险：

$$\delta_{\text{ERM}}(D) = \arg \min_{\delta} R_{\text{emp}}(D, \delta) \quad (6.50)$$

在**无监督**情形下，我们去掉所有关于  $y$  的东西，用  $L(x, \delta(x))$  代替  $L(y, \delta(x))$ ，例如  $L(x, \delta(x)) = \|x - \delta(x)\|_2^2$ ，它度量**重构误差**。我们可以把决策规则写成  $\delta(x) = \text{decode}(\text{encode}(x))$ ，如**矢量量化**（第 11.4.2.6 节）或**PCA**（第 12.2 节）。最后，经验风险定义为

$$R_{\text{emp}}(D, \delta) = \frac{1}{N} \sum_{i=1}^N L(x_i, \delta(x_i)) \quad (6.51)$$

当然，把  $\delta(x) \equiv x$  就能平凡地把这个风险最小化，所以编码—解码中必须有某种“瓶颈”。

## 解释（基础且详尽）

### 1) 为何需要 ERM?

- 频率学派的“真风险”  $R(\theta^{*}, \delta)$  要对真实分布求期望，但  $p^{*}$  未知  $\Rightarrow$  直接算不了。
- 若任务是预测可观测的  $y$  (而非估计参数  $\theta$ )，可以把风险写成对  $(x, y)$  的期望  $R(p^{*}, \delta)$ 。
- 用训练集的经验分布  $p_{\text{emp}}$  近似  $p^{*}$ ，就得到经验风险：训练集上平均损失。

$\delta_{x_i}(x) / \delta_{y_i}(y)$  是“狄拉克  $\delta$ ”/“点质量”的记号：把全部概率放在样本点  $(x_i, y_i)$  上。于是对它的积分，等价于对训练样本做平均。

### 2) ERM 在监督学习中长什么样？

- 经验风险：**  $R_{\text{emp}} = \frac{1}{N} \sum L(y_i, \delta(x_i))$ 。
- 0-1 损失  $\rightarrow$  训练误分类率；平方损失  $\rightarrow$  训练均方误差。
- ERM 目标：**  $\delta_{\text{ERM}} = \arg \min R_{\text{emp}}$ ，即“在训练集上出错最少/误差最小”。

直观例子：二分类， $\delta$  是阈值分类器。ERM 就是选那个阈值，使训练集中被分错的点最少。

### 3) 无监督版 ERM

- 没有标签  $y$ ，就用**重构损失** (如  $\|x - \delta(x)\|^2$ )。
- 写作  $\delta(x) = \text{decode}(\text{encode}(x))$ ：
  - PCA：encode=投影到低维主成分，decode=投回原空间；
  - 矢量量化 / k-means：encode=指派到某个码字/簇心，decode=输出该码字。
- 为什么要“瓶颈”：** 若允许  $\delta(x) = x$ ，训练误差为 0 (完美重构)，但**没有学到结构** (完全过拟合)。瓶颈=低维表示/有限码书/稀疏性/正则化，强迫模型**概括**。

### 4) 优缺点与注意点

- 优点：** 不需要知道  $p^{*}$ ；实现简单、直接。
- 缺点：** 只优化**训练误差**，可能**过拟合**。因此实际会配合
  - 正则化** (限制模型复杂度)，

- **结构性风险最小化 (SRM)** /容量控制,
- **交叉验证估计泛化误差** (6.5.3 节会讲) 。
- 许多经典算法都可视为 ERM (或其凸替代) :
  - 线性回归 (平方损失) 、逻辑回归/线性 SVM (对 0-1 损失的凸替代) 、k-means/PCA (无监督重构损失) 。

## 5) 把符号再串一下

- $R(p^*, \delta)$ : **真风险** (对真实分布的期望) 。
  - $p_{\text{emp}}$ : 经验分布 =  $\frac{1}{N} \sum \delta_{(x_i, y_i)}$ 。
  - $R_{\text{emp}}(D, \delta)$ : **经验风险** (训练平均损失) 。
  - $\delta_{\text{ERM}}(D)$ : 在数据  $D$  上学到的**决策规则** (依赖  $D$ ) 。
- 

## 小结

- ERM 用**训练数据的平均损失**近似“真实期望损失”，从而回避了“不知道  $p^*$ ”的问题。
- 监督：最小化  $\frac{1}{N} \sum L(y_i, \delta(x_i))$ ; 无监督：最小化  $\frac{1}{N} \sum L(x_i, \delta(x_i))$  (需**瓶颈**防止平凡解) 。
- ERM 简单有效，但必须搭配**容量控制/正则化与交叉验证**来保证**泛化**。

翻译并解释以下内容，翻译不能有遗漏，解释请基础且详尽，因为我的数学基础不太好。

### 6.5.1 Regularized risk minimization

Note that the empirical risk is equal to the Bayes risk if our prior about “nature’s distribution” is that it is exactly equal to the empirical distribution (Minka 2001b):

$$E [R(p^*, \delta) | p^* = p_{\text{emp}}] = R_{\text{emp}}(D, \delta) \quad (6.52)$$

Therefore minimizing the empirical risk will typically result in overfitting. It is therefore often necessary to add a complexity penalty to the objective function:

$$R(D, \delta) = R_{\text{emp}}(D, \delta) + \lambda C(\delta) \quad (6.53)$$

where  $C(\delta)$  measures the complexity of the prediction function  $\delta(x)$  and  $\lambda$  controls the strength of the complexity penalty. This approach is known as regularized risk minimization (RRM). Note that if the loss function is

negative log likelihood, and the regularizer is a negative log prior, this is equivalent to MAP estimation. The two key issues in RRM are: how do we measure complexity, and how do we pick  $\lambda$ . For a linear model, we can define the complexity of in terms of its degrees of freedom, discussed in Section 7.5.3. For more general models, we can use the VC dimension, discussed in Section 6.5.4. To pick  $\lambda$ , we can use the methods discussed in Section 6.5.2.

### 6.5.2 Structural risk minimization

The regularized risk minimization principle says that we should fit the model, for a given complexity penalty, by using

$$\hat{\delta} = \operatorname{argmin}_{\delta} [R_{\text{emp}}(D, \delta) + \lambda C(\delta)] \quad (6.54)$$

But how should we pick  $\lambda$ ? We cannot use the training set, since this will underestimate the true risk, a problem known as optimism of the training error. As an alternative, we can use the following rule, known as the structural risk minimization principle: (Vapnik 1998):

$$\hat{\lambda} = \operatorname{argmin}_{\lambda} \hat{R}(\delta) \quad (6.55)$$

where  $\hat{R}(\delta)$  is an estimate of the risk. There are two widely used estimates: cross validation and theoretical upper bounds on the risk. We discuss both of these below.

已思考 11s

下面先完整翻译，再给出基础且详尽的解释（含直观要点与小例子）。我会保留原式号，不做遗漏。

## 翻译

### 6.5.1 正则化风险最小化 (Regularized risk minimization)

注意：如果我们对“自然的分布”的先验恰好等于经验分布，那么经验风险就等于贝叶斯风险 (Minka 2001b)：

$$\mathbb{E}[R(p^*, \delta) | p^* = p_{\text{emp}}] = R_{\text{emp}}(D, \delta) \quad (6.52)$$

因此，仅仅最小化经验风险通常会导致过拟合。所以常常需要在目标函数里加入复杂度惩罚：

$$R_\lambda(D, \delta) = R_{\text{emp}}(D, \delta) + \lambda C(\delta) \quad (6.53)$$

其中  $C(\delta)$  度量预测函数  $\delta(x)$  的**复杂度**， $\lambda$  控制复杂度惩罚的强度。这种方法称为**正则化风险最小化 (RRM)**。注意：如果损失函数是**负对数似然**，而正则项是**负对数先验**，那么这与 MAP 估计是等价的。RRM 的两个关键问题是：**如何度量复杂度、如何选择  $\lambda$** 。在线性模型中，我们可以用其**自由度**来定义复杂度（7.5.3 节）。对更一般的模型，可以用**VC 维**（6.5.4 节）。选择  $\lambda$  的方法见 6.5.2 节。

### 6.5.2 结构风险最小化 (Structural risk minimization)

正则化风险最小化原则说：给定一个复杂度惩罚，按下式拟合模型

$$\hat{\delta}_\lambda = \arg \min_{\delta} [R_{\text{emp}}(D, \delta) + \lambda C(\delta)] \quad (6.54)$$

但我们该如何选  $\lambda$ ? **不能用训练集来选**，因为这会**低估真实风险**（称为训练误差的**乐观性问题**）。一种替代方法是下面的规则，称为**结构风险最小化原则** (Vapnik 1998)：

$$\hat{\lambda} = \arg \min_{\lambda} \hat{R}(\hat{\delta}_\lambda) \quad (6.55)$$

其中  $\hat{R}(\delta)$  是**风险的估计**。常用的两种估计方法是：**交叉验证与理论界**。我们将在下文讨论二者。

## 解释 (基础且详尽)

### 1) 为什么需要“正则化风险最小化 (RRM) ”?

- 经验风险  $R_{\text{emp}}$  是**训练集平均损失**。如果你“先验地”相信“真实分布  $p^*$  就是这份训练数据的经验分布”，那么“对真实风险的贝叶斯期望”就等于“经验风险”（式 6.52）。
- 这等于把训练样本当作**世界的全部**  $\Rightarrow$  最优解会**紧贴训练数据**  $\Rightarrow$  **过拟合**。
- 解决：在目标里加一项**复杂度惩罚**  $C(\delta)$ ，形成

$$\min_{\delta} R_{\text{emp}}(D, \delta) + \lambda C(\delta).$$

这就是把“拟合好数据”和“保持模型简单/平滑”两件事进行**权衡**。

### 和 MAP 的等价

- 若损失 = **负对数似然** ( $-\log p(D | \theta)$ )，正则项 = **负对数先验** ( $-\log p(\theta)$ )，那么

$$\underbrace{-\log p(D | \theta)}_{\text{经验风险}} + \underbrace{-\log p(\theta)}_{\lambda C(\delta)} = -\log p(\theta | D) + \text{常数},$$

最小化它  $\Leftrightarrow$  最大化后验  $\Rightarrow$  MAP。

常见例子：线性回归 + L2 正则（岭） $\Leftrightarrow$  高斯先验；L1 正则（Lasso） $\Leftrightarrow$  拉普拉斯先验。

## 复杂度怎么量？

- 线性/广义线性模型：**自由度**（有效参数个数、帽子矩阵迹等，见 7.5.3）。
- 更一般模型：**VC 维**（假设空间容量），见 6.5.4；也可用 Rademacher 复杂度、平滑度范数、网络权值范数等。

## 2) 为什么训练集不能用来选 $\lambda$ ？

- 任何把  $\lambda$  朝着更小训练误差的方向调的过程，都会系统性地**乐观**：你在训练上看到的误差比真实泛化误差更小。
- 因此，需用**独立的风险估计**来挑  $\lambda$ 。

## 3) 结构风险最小化 (SRM) 怎么做？

SRM 是一个**两层**过程：

1. **内层拟合**：对每个  $\lambda$ ，求  $\hat{\delta}_\lambda = \arg \min_\delta [R_{\text{emp}} + \lambda C]$  (式 6.54)；
2. **外层选择**：用**风险估计**  $\hat{R}(\cdot)$  挑出最好的  $\lambda$ ：  
 $\hat{\lambda} = \arg \min_\lambda \hat{R}(\hat{\delta}_\lambda)$  (式 6.55)。

### $\hat{R}(\cdot)$ 的两类常用估计

- **交叉验证 (CV)**：把数据划分成折，反复“训练在训练折、评估在验证折”，取平均验证损失作为  $\hat{R}$ 。
  - 回归：均方误差；分类：验证集的对数损失/错误率等。
  - 优点：数据驱动、少假设；缺点：计算量高。
- **理论上界**：如 VC 维/Rademacher 的**泛化误差上界**

$$\text{真风险} \leq \text{训练风险} + \text{容量惩罚}(\text{模型复杂度}, N, \delta),$$

以概率  $1 - \delta$  成立。选使上界最小的  $\lambda$ （更保守，但不依赖额外验证集）。

## 4) 直观小例子（回归）

- 目标： $\min_w \frac{1}{N} \sum_i (y_i - w^\top x_i)^2 + \lambda \|w\|_2^2$ 。
- $\lambda$  小  $\Rightarrow$  复杂度惩罚弱  $\Rightarrow$  更贴数据（**方差** $\uparrow$ 、过拟合风险 $\uparrow$ ）。
- $\lambda$  大  $\Rightarrow$  权重被收缩  $\Rightarrow$  更平滑（**偏差** $\uparrow$ 、但**方差** $\downarrow$ ）。
- **SRM**：对一串  $\lambda$  做训练；用 k 折 CV 求  $\hat{R}$ ；取  $\hat{\lambda}$ ；最后用  $\hat{\lambda}$  在全训练集上再拟合一次得到最终模型。

## 5) 要点与实践建议

- RRM = “经验风险 + 复杂度惩罚”的统一范式；SRM = 用独立风险估计来挑惩罚强度。
  - 复杂度度量：线性模型用**自由度**，一般模型用**VC 维/容量度量**或**正则范数**。
  - 选  $\lambda$ ：优先用**交叉验证**；资源有限或需保证型结论时，可参考**理论上界**。
  - 与贝叶斯一致性：把正则看作先验、 $\lambda$ 看作先验强度；经验上常用**经验贝叶斯/证据最大化**、或**CV**来定  $\lambda$ 。
- 

## 小结

- 仅最小化经验风险会过拟合；在目标里加入**复杂度惩罚**形成 RRM。当损失是 NLL、正则是负对数先验时， $RRM \equiv MAP$ 。
- SRM 用一个**外层**风险估计（交叉验证或理论上界）来选择  $\lambda$ ，避免训练误差的“乐观偏差”。
- 这两者构成现代机器学习中“**拟合 vs. 复杂度、偏差-方差**”权衡的标准做法。

翻译并解释以下内容，翻译不能有遗漏，解释请基础且详尽，因为我的数学基础不太好。

6.5.3 Estimating the risk using cross validation We can estimate the risk of some estimator using a validation set. If we don't have a separate validation set, we can use cross validation (CV), as we briefly discussed in Section 1.4.8. More precisely, CV is defined as follows. Let there be  $N = |D|$  data cases in the training set. Denote the data in the  $k$ 'th test fold by  $D_k$  and all the other data by  $D - k$ . (In stratified CV, these folds are chosen so the class proportions (if discrete labels are present) are roughly equal in each fold.) Let  $F$  be a learning algorithm or fitting function that takes a dataset and a model index  $m$  (this could a discrete index, such as the degree of a polynomial, or a continuous index, such as the strength of a regularizer) and returns a parameter vector:

$$\hat{\theta}^m = F(D, m) \quad (6.56)$$

Finally, let  $P$  be a prediction function that takes an input and a parameter vector and returns a prediction:

$$\hat{y} = P(x, \hat{\theta}) = f(x, \hat{\theta}) \quad (6.57)$$

Thus the combined fit-predict cycle is denoted as

$$f_m(x, D) = P(x, F(D, m)) \quad (6.58)$$

The K-fold CV estimate of the risk of  $f_m$  is defined by

$$R(m, D, K) = \frac{1}{N} \sum_{k=1}^K \sum_{i \in D_k} L(y_i, P(x_i, F(D - k, m))) \quad (6.59)$$

Note that we can call the fitting algorithm once per fold.

Let  $f_k m(x) = P(x, F(D - k, m))$  be the function that was trained on all the data except for the test data in fold  $k$ .

Then we can rewrite the CV estimate as

$$R(m, D, K) = \frac{1}{N} \sum_{k=1}^K \sum_{i \in D_k} L(y_i, f_k m(x_i)) = \frac{1}{N} \sum_{i=1}^N L(y_i, f_{k(i)} m(x_i)) \quad (6.60)$$

where  $k(i)$  is the fold in which  $i$  is used as test data. In other words, we predict  $y_i$  using a model that was trained on data that does not contain  $x_i$ . If  $K = N$ , the method is known as leave one out cross validation or LOOCV. In this case, the estimated risk becomes

$$R(m, D, N) = \frac{1}{N} \sum_{i=1}^N L(y_i, f_{-i} m(x_i)) \quad (6.61)$$

where  $f_{-i} m(x) = P(x, F(D - i, m))$ . This requires fitting the model  $N$  times, where for  $f_{-i} m$  we omit the  $i$ 'th training case. Fortunately, for some model classes and loss functions (namely linear models and quadratic loss), we can fit the model once, and analytically "remove" the effect of the  $i$ 'th training case. This is known as generalized cross validation or GCV.

已思考 17s

下面先完整翻译，再给出基础且详尽的解释（把符号都说清楚、步骤化说明）。原文里有几处显然的排版小瑕疵（例如“Of  $K = N$ ”应为“If  $K = N$ ”，以及个别下标空格），我在解释里会指出，但翻译不做遗漏。

## 翻译

### 6.5.3 用交叉验证估计风险

我们可以用一个**验证集**来估计某个估计器的风险。如果没有单独的验证集，就可以使用**交叉验证 (CV)**，我们在第 1.4.8 节做过简要讨论。更精确地，CV 定义如下。设训练集中有  $N = |D|$  条数据。把第  $k$  个**测试折**中的数据记为  $D_k$ ，其余所有数据记为  $D_{-k}$ 。（在**分层 CV** 中，这些折被选择为：如果存在离散标签，则每折中的类别比例大致相等。）设  $F$  是一个学习算法或拟合函数，它接收一个数据集和一个**模型索引**  $m$ （它可以是离散的，比如多项式次数；也可以是连续的，比如正则化强度），并返回一个参数向量：

$$\hat{\theta}_m = F(D, m) \quad (6.56)$$

最后，设  $P$  是一个**预测函数**，它接收一个输入和一个参数向量并返回一个预测：

$$\hat{y} = P(x, \hat{\theta}) = f(x, \hat{\theta}) \quad (6.57)$$

因此，合在一起的“拟合→预测”流程记为

$$f_m(x, D) = P(x, F(D, m)) \quad (6.58)$$

$K$ -折 CV 对  $f_m$  的**风险估计**定义为

$$R(m, D, K) = \frac{1}{N} \sum_{k=1}^K \sum_{i \in D_k} L(y_i, P(x_i, F(D_{-k}, m))) \quad (6.59)$$

注意我们可以**每折调用一次拟合算法**。令

$$f_m^k(x) = P(x, F(D_{-k}, m))$$

表示在“除去第  $k$  折测试数据之外的所有数据”上训练出来的函数。于是 CV 估计可重写为

$$R(m, D, K) = \frac{1}{N} \sum_{k=1}^K \sum_{i \in D_k} L(y_i, f_m^k(x_i)) = \frac{1}{N} \sum_{i=1}^N L(y_i, f_m^{k(i)}(x_i)) \quad (6.60)$$

其中  $k(i)$  表示样本  $i$  被放到哪一折作为测试数据。换言之，我们用**不包含**  $x_i$  的训练数据训练出的模型来预测  $y_i$ 。

如果  $K = N$ ，该方法称为**留一法交叉验证** (leave-one-out CV, LOOCV)。在这种情况下，所估计的风险变为

$$R(m, D, N) = \frac{1}{N} \sum_{i=1}^N L(y_i, f_m^{-i}(x_i)) \quad (6.61)$$

其中

$$f_m^{-i}(x) = P(x, F(D_{-i}, m)).$$

这需要**训练**  $N$  次模型，其中对  $f_m^{-i}$  我们省略第  $i$  个训练样本。幸运的是，对某些模型类别和损失函数（即**线性模型**与**二次损失**），我们可以只**训练一次模型**，然后**解析地“移除”第  $i$  个训练样本的影响**。这称为**广义交叉验证** (generalized cross validation, GCV)。

## 解释 (基础且详尽)

## 1) 要解决什么问题?

- **目标**: 估计某个算法/超参数设置  $m$  的**真实风险** (在未知真实分布下的期望损失)。
- **难点**: 真实分布不知道。
- **办法**: 用**看不见该样本的模型**来预测它的标签/数值, 然后对所有样本的\*\*“看不见”预测误差\*\*求平均  $\Rightarrow$  这就是交叉验证。

## 2) 记号速读

- $D = \{(x_i, y_i)\}_{i=1}^N$ : 训练数据。
- 把  $D$  分成  $K$  份 (“折”) : 每次用其中 1 折做**测试**  $D_k$ , 其余  $K - 1$  折做**训练**  $D_{-k}$ 。
- $F(D, m)$ : 在数据  $D$  和模型索引  $m$  (如多项式次数、正则强度等) 上**拟合**, 得到参数  $\hat{\theta}_m$ 。
- $P(x, \hat{\theta})$ : 用参数做**预测**, 得到  $\hat{y}$ 。
- $f_m(x, D) = P(x, F(D, m))$ : 把“训练+预测”合成一个函数。
- $L(y, \hat{y})$ : 损失, 如平方误差、对数损失、0–1 损失等。
- $R(m, D, K)$ : **K 折 CV 的风险估计** (式 6.59 / 6.60)。

## 3) 一步步操作版 (K 折 CV)

1. **分析**: 把样本平均 (或**分层**保持类比) 分成  $K$  份。

2. **循环**  $k = 1..K$ :

- 用  $D_{-k}$  训练模型:  $\hat{\theta}_m^{(k)} = F(D_{-k}, m)$ ;
- 用该模型在测试折  $D_k$  上做预测, 累计损失。

3. **平均**: 把  $K$  次测试损失加总后除以  $N$  (式 6.59)。

这等价于式 (6.60): 每个样本都恰好被“看不见”一次, 我们用**不含它的模型**来预测它。

## 4) 几种常见的 $K$

- **LOOCV ( $K = N$ )** : 训练  $N$  次, 训练集几乎与全量一样大; **偏差小**但计算贵, 方差可能偏大。
- $K = 5$  或  $10$ : 实践最常用, **计算量可控、方差较小**。
- **分层 CV**: 分类时每折的类别比例尽量一致, 降低方差。
- **时间序列/组别 CV**: 如果数据非 i.i.d. (如同一用户的多条记录), 应用**按时间或按组**的拆分, 避免泄露。

## 5) 模型选择与 CV (为什么式 6.55 需要“外层”)

- 对每个候选  $m$  (比如不同的  $\lambda$ 、多项式次数等), 都做一遍  $K$ -折 CV, 得到  $R(m, D, K)$ 。
- 选择  $\hat{m} = \arg \min_m R(m, D, K)$ 。这实现了第 6.5.2 节的**结构风险最小化思想**。

- **注意：**若还要报告一个最终的“泛化误差”，最好再用**独立测试集或嵌套 CV**，避免在同一份 CV 上“既选又评”。

## 6) LOOCV 与 GCV

- LOOCV 要训练  $N$  次，成本高。
- 在线性模型 + 二次损失下，可用**帽子矩阵** (hat matrix) 技巧：

$$\text{LOOCV 误差}_i \approx \frac{r_i}{1 - h_{ii}},$$

其中  $r_i$  是残差， $h_{ii}$  是杠杆值；把这些代入即可一次训练、快速近似 LOOCV。

- 这种一次训练、解析调整的做法就是 GCV；在岭回归等情形中非常常见。

## 7) 选什么损失 $L$ ？

- **回归：**平方误差、绝对误差、对数似然（高斯假设）等。
- **分类：**0-1 误差（不光滑）、更常用的是**对数损失/交叉熵**或 hinge 损失，更稳定。
- **排序/概率：**AUC、Brier 分数、NLL 等，按任务选。

## 8) 常见坑

- **数据泄露：**特征标准化、特征选择、PCA 等必须在**训练折**内拟合、再应用到**测试折**；不能用全数据算均值方差。
- **重复 CV：**为减少随机性，可重复多次随机分折取平均。
- **不平衡数据：**优先采用**分层**；用合适的指标（如 AUC、F1、PR-AUC）。
- **大规模高成本：**可用较小  $K$ 、早停、或近似（如 GCV、折内 warm-start）。

## 小结

- **交叉验证**用“**看不见该样本的模型去预测它**”的方式，估计真实风险。
- 公式 (6.59)/(6.60) 就是把各折的**测试损失**平均起来；LOOCV 是  $K = N$  的特例；GCV 在特定模型/损失下可“**一次训练近似 LOOCV**”。
- 在模型选择时，对每个超参数  $m$  计算 CV 风险，再选使其最小者；注意避免**数据泄露**，并在需要时用**嵌套 CV** 或**独立测试集**做最终评估。



已上传的图片

翻译并解释以下内容，翻译不能有遗漏，解释请基础且详尽，因为我的数学基础不太好。

Figure 6.6 (a) Mean squared error for 2 penalized degree 14 polynomial regression vs log regularizer. Same as in Figures 7.8, except now we have N =50 training points instead of 21. The stars correspond to the values used to plot the functions in Figure 7.7. (b) CV estimate. The vertical scale is truncated for clarity. The blue line corresponds to the value chosen by the one standard error rule. Figure generated by linregPolyVsRegDemo.

6.5.3.1 Example: using CV to pick  $\lambda$  for ridge regression  
As a concrete example, consider picking the strength of the 2 regularizer in penalized linear regression. We use the following rule:

$$\hat{\lambda} = \arg \min_{\lambda \in [\lambda_{\min}, \lambda_{\max}]} R(\lambda, D_{\text{train}}, K) \quad (6.62)$$

where  $[\lambda_{\min}, \lambda_{\max}]$  is a finite range of  $\lambda$  values that we search over, and  $R(\lambda, D_{\text{train}}, K)$  is the K-fold CV estimate of the risk of using  $\lambda$ , given by

$$R(\lambda, D_{\text{train}}, K) = \frac{1}{|D_{\text{train}}|} \sum_{k=1}^K \sum_{i \in D_k} L(y_i, f_k(\lambda)(x_i)) \quad (6.63)$$

where  $f_k(\lambda)(x) = x^T \hat{w} \lambda(D - k)$  is the prediction function trained on data excluding fold  $k$ , and  $\hat{w} \lambda(D) = \arg \min_w NLL(w, D) + \lambda \|w\|^2$  is the MAP estimate. Figure 6.6(b) gives an example of a CV estimate of the risk vs  $\log(\lambda)$ , where the loss function is squared error. When performing classification, we usually use 0-1 loss. In this case, we optimize a convex upper bound on the empirical risk to estimate  $w \lambda m$  but we optimize (the CV estimate of) the risk itself to estimate  $\lambda$ . We can handle the non-smooth 0-1 loss function when estimating  $\lambda$  because we are using brute-force search over the entire (one-dimensional) space. When we have more than one or two

tuning parameters, this approach becomes infeasible. In such cases, one can use empirical Bayes, which allows one to optimize large numbers of hyper-parameters using gradient-based optimizers instead of brute-force search. See Section 5.6 for details.

已思考 41s

下面先完整翻译，再给出基础且详尽的解释（尽量用直白语言讲清楚每个符号和做法）。我会把“ $\ell_2$  正则/ $\ell_2$  罚项”明确为  $\ell_2$  (L2) 正则。

---

## 翻译

### 图 6.6

- (a) 14 次多项式回归 + L2 正则 的均方误差 (MSE) 随  $\log \lambda$  变化的曲线。与图 7.8 相同，只是现在训练点数是  $N = 50$  (之前是 21)。图中的星号表示在图 7.7 里用来绘制函数的那些  $\lambda$  取值。
- (b) 交叉验证 (CV) 得到的风险估计。为了使图更清晰，纵轴做了截断。蓝色竖线表示用“一倍标准误差规则”选出的  $\lambda$ 。图由 `linregPolyVsRegDemo` 生成。

#### 6.5.3.1 例子：用 CV 为岭回归选择 $\lambda$

作为一个具体例子，考虑在带 L2 罚项 的线性回归里选择正则强度。我们使用下述规则：

$$\hat{\lambda} = \arg \min_{\lambda \in [\lambda_{\min}, \lambda_{\max}]} R(\lambda, D_{\text{train}}, K) \quad (6.62)$$

其中  $[\lambda_{\min}, \lambda_{\max}]$  是我们搜索的有限取值范围； $R(\lambda, D_{\text{train}}, K)$  是在给定  $\lambda$  时，**K 折交叉验证对风险的估计**：

$$R(\lambda, D_{\text{train}}, K) = \frac{1}{|D_{\text{train}}|} \sum_{k=1}^K \sum_{i \in D_k} L(y_i, f_{\lambda}^k(x_i)) \quad (6.63)$$

这里  $f_{\lambda}^k(x) = x^T \hat{w}_{\lambda}(D_{-k})$  是在不包含第  $k$  折测试数据的训练集上学到的预测函数；而

$$\hat{w}_{\lambda}(D) = \arg \min_w \text{NLL}(w, D) + \lambda \|w\|_2^2$$

是 MAP 估计 (最小化负对数似然 + L2 罚项)。图 6.6(b) 展示了当损失是**平方误差**时，CV 估计的风险与  $\log(\lambda)$  的关系。

进行分类时，我们通常用**0-1 损失**。这种情况下，训练权重  $w_{\lambda}$  时会最小化**经验风险的凸上界** (如对数损失或 hinge 损失)，但选择  $\lambda$  时仍最小化 (CV 估计的) **真实风险本身**。之所以能处理 0-1 的非光滑损失，是因为我们对  $\lambda$  做的是一维的穷举搜索。当需要同时调的

参数多于一两个时，这种做法就不可行了；这时可以用**经验贝叶斯**，用基于梯度的优化器来同时优化大量超参数，而不是穷举搜索。详见第 5.6 节。

## 解释（基础且详尽）

### 1) 图 6.6 在表达什么？

- 横轴是  $\log \lambda$ ,  $\lambda$  越大表示**正则化越强**（权重被更厉害地收缩到 0）。
- (a) 训练/测试 MSE:
  - 蓝色虚线：训练误差；红色实线：独立测试集误差。
  - 典型 U 形：
    - $\lambda$  很小  $\rightarrow$  模型太复杂  $\rightarrow$  训练误差小、测试误差变大（过拟合）。
    - $\lambda$  很大  $\rightarrow$  模型太简单  $\rightarrow$  训练/测试误差都大（欠拟合）。
    - 中间某个  $\lambda$  处测试误差最小。星号是拿去画其他图的取值。
- (b) CV 曲线：
  - 用  $K$  折（例子里是 5 折）CV 得到**平均验证 MSE 及标准误差条**（误差线）。
  - **一倍标准误差规则**（蓝线）：选择**最简单**（ $\lambda$  最大）的、其 CV 误差**不超过最小 CV 误差 + 1 个标准误差**的那一点。
  - 这样牺牲极少的验证误差，换取**更简单、更稳、方差更小**的模型。

### 2) 用 CV 选 $\lambda$ 的步骤（可直接照做）

1. 取一组候选  $\lambda$ （常在对数尺度均匀取点）。
2. 对每个  $\lambda$ ，做  **$K$  折 CV**：
  - 每次用  $D_{-k}$  训练  $w_\lambda$ ，在测试折  $D_k$  上算损失；
  - 把  $K$  次的测试损失平均，得到  $R(\lambda, D_{\text{train}}, K)$ 。
3. 选  $\lambda$ ：
  - **最小 CV 误差**；或
  - **一倍标准误差规则**：在“误差  $\leq$  最小值 + 1SE”的集合里选**最大的**  $\lambda$ （最简单模型）。
4. 用选出的  $\hat{\lambda}$  在**整份训练集**上重训，得到最终模型；最后再在**独立测试集**上评估一次。

小贴士：对所有  $\lambda$  使用**同一套分折**，能减小比较时的随机波动；标准化/特征选择等步骤要在**折内拟合**，避免数据泄露。

### 3) 为什么式子里有 $NLL + \lambda \|w\|_2^2$ (MAP) ?

- 这就是岭回归/高斯先验:  $w \sim N(0, \lambda^{-1} I)$ 。
- $\lambda = 0$  给出 MLE (无正则, 易过拟合) ;
- $\lambda > 0$  引入偏差但降低方差 ("以偏换方") , 通常泛化更好。

## 4) 分类时怎么用?

- 训练  $w \rightarrow$  用**凸替代损失** (对数损失/hinge) , 好优化;
  - 挑  $\lambda \rightarrow$  仍然可以直接以 0-1 错误率的 CV 均值为目标, 因为只是一维网格搜索。
  - 若要调的超参数很多 (例如多个罚项、核宽度等) , 网格/穷举不可承受  $\rightarrow$  用**经验贝叶斯/证据最大化**之类的**梯度法**统一优化。
- 

## 小结

- **目标:** 选让**泛化误差**最小的  $\lambda$ , 不用训练误差评估, 以免“乐观偏差”。
- **方法:** 对候选  $\lambda$  做 K 折 CV 得到  $R(\lambda)$ ; 按最小值或“一倍标准误差规则”选择。
- **直觉:**  $\lambda$  太小  $\rightarrow$  过拟合; 太大  $\rightarrow$  欠拟合; CV 帮你找到中间“甜蜜点”。
- **多参调优:** 用 CV 的网格/随机搜索可行于少量参数; 大量参数时用**经验贝叶斯/梯度方法**更高效。

翻译并解释以下内容, 翻译不能有遗漏, 解释请基础且详尽, 因为我的数学基础不太好。

6.5.3.2 The one standard error rule The above procedure estimates the risk, but does not give any measure of uncertainty. A standard frequentist measure of uncertainty of an estimate is the standard error of the mean, defined by

$$se = \hat{\sigma} / \sqrt{N} = \hat{\sigma} / \sqrt{2N} \quad (6.64)$$

where  $\hat{\sigma}^2$  is an estimate of the variance of the loss:

$$\hat{\sigma}^2 = \frac{1}{N} \sum_{i=1}^N (L_i - \bar{L})^2, \quad L_i = L(y_i, f_k(i)) \quad m(x_i) \quad \bar{L} = \frac{1}{N} \sum_{i=1}^N L_i \quad (6.65)$$

Note that  $\sigma$  measures the intrinsic variability of  $L_i$  across samples, whereas  $se$  measures our uncertainty about the mean  $\bar{L}$ .

Suppose we apply CV to a set of models and compute the mean and  $se$  of their estimated risks. A common heuristic for picking a model from these noisy estimates is to pick

the value which corresponds to the simplest model whose risk is no more than one standard error above the risk of the best model; this is called the one-standard error rule (Hastie et al. 2001, p216). For example, in Figure 6.6, we see that this heuristic does not choose the lowest point on the curve, but one that is slightly to its right, since that corresponds to a more heavily regularized model with essentially the same empirical performance.

已思考 18s

下面先完整翻译，再给出基础且详尽的解释（包含一步步怎么做、直观含义与常见做法的小提示）。原文第 (6.64) 式里有一个常见的排版小瑕疵：标准误应为  $\hat{\sigma}/\sqrt{N}$ （而不是  $\hat{\sigma}^2/N$ ）；我会在“解释”部分说明这一点，但在翻译里不作改写以免遗漏。

## 翻译

### 6.5.3.2 一倍标准误差规则

上面的过程给出了风险的估计，但未给出不确定性的度量。对一个估计量，不确定性的一个标准的频率学派度量是均值的标准误（standard error of the mean），定义为

$$se = \frac{\hat{\sigma}}{\sqrt{N}} = \frac{\hat{\sigma}^2}{N} \quad (6.64)$$

其中  $\hat{\sigma}^2$  是损失的方差的一个估计：

$$\hat{\sigma}^2 = \frac{1}{N} \sum_{i=1}^N (L_i - \bar{L})^2, \quad L_i = L(y_i, f_m^{k(i)}(x_i)), \quad \bar{L} = \frac{1}{N} \sum_{i=1}^N L_i \quad (6.65)$$

注意， $\sigma$  度量的是  $L_i$  在不同样本间的固有变异性，而 se 度量的是我们对均值  $\bar{L}$  的不确定性。

假设我们把交叉验证应用到一组模型上，并计算它们的风险估计的均值与标准误。从这些带噪声的估计中选模型的一个常见启发式是：选择那个对应于“最简单的模型”，并且其风险不超过“最佳模型的风险 + 一个标准误”的取值；这称为“一倍标准误差规则”（Hastie 等, 2001, 第 216 页）。例如，在图 6.6 中，我们可以看到，这个启发式并不会选择曲线上的最低点，而是选择其略偏右的一个点，因为那对应着一个更强正则化的模型，并且它与最低点的经验表现几乎相同。

# 解释 (基础且详尽)

## 1) 标准误到底是什么?

- 我们对每个样本  $i$  得到一个\*\*“看不见”预测\*\*的损失  $L_i$  (来自交叉验证: 用不含  $x_i$  的模型去预测  $y_i$ )。
- 这些  $L_i$  的**平均**  $\bar{L}$  就是 CV 的风险估计; 它本身仍有**随机波动**。
- $\hat{\sigma}^2$  是这些  $L_i$  的**样本方差**;
- **标准误** se 是**均值**  $\bar{L}$  的不确定性:

$$se = \frac{\hat{\sigma}}{\sqrt{N}}$$

注: 原文 (6.64) 的第二个等号是常见的排版错误; 正确应为  $\hat{\sigma}/\sqrt{N} = \sqrt{\hat{\sigma}^2/N}$ 。

- 区分:
  - $\hat{\sigma}$  (标准差) → “每个样本的损失  $L_i$  彼此差多大”;
  - se (标准误) → “我们的**平均损失**估计  $\bar{L}$  还会抖多大”。

(在 K 折 CV 中, 许多实现也会把**每一折的平均损失**当作 1 个观测, 再用折数  $K$  计算标准误:  $se = sd(\{\bar{L}^{(k)}\})/\sqrt{K}$ 。两种做法思想一致——都在给“均值”的不确定性估计一个量级。)

## 2) 一倍标准误差规则 (1-SE rule) 怎么用?

**目标:** 从“带误差条”的 CV 曲线里挑一个既**表现不差**又**更简单更稳**的模型。

**步骤:**

1. 对每个候选模型 (或超参数  $\lambda$ ) , 算出 CV **平均损失**  $\bar{L}(\lambda)$  与 **标准误**  $se(\lambda)$ 。
2. 找到**最小**的平均损失:  $\bar{L}_{min}$  (在  $\lambda^{*}$  处)。
3. 计算**阈值**:  $threshold = \bar{L}_{min} + se(\lambda^{*})$ 。
4. 在所有满足  $\bar{L}(\lambda) \leq threshold$  的候选里, 选**最简单**的那个 (比如在线性/岭回归里就是**最大的**  $\lambda$ )。

**直观:** 最低点附近往往都“半斤八两”; 在误差条范围内我们无法区分谁真更好。那就选**更简单/更正则**的那个, 以降低方差、提升稳健性与泛化。

## 3) 结合图 6.6 如何理解?

- CV 曲线 (MSE vs  $\log \lambda$ ) 有一个**最低点**; 但右边稍大一点的  $\lambda$  给出的 MSE 与最低点**几乎一样** (落在“最低点 + 1SE”的阈值以内)。
- 1-SE 规则因此会选**更靠右** ( $\lambda$  更大、模型更简单/更受约束) 的那个, 这就是图中的蓝色竖线。

## 4) 为什么要这么“保守”一点?

- CV 估计有噪声；选恰好最低常因偶然波动而偏向更复杂的模型（方差大，易过拟合）。
- 1-SE 规则牺牲极少的验证性能（不超过 1 个标准误），换来更简单且稳定的模型，通常更可靠，尤其在样本不多时。

## 5) 实践小提示

- 把候选超参数放在对数刻度上均匀取点；
  - 用同一套折分做整条曲线（减少相对比较的噪声）；
  - 先标准化特征，但要在折内估计均值方差以避免数据泄露；
  - 若曲线很平，1-SE 规则常比“取最低点”更稳健；
  - 对分类任务，可把  $L_i$  换成0-1 错误、对数损失或你关心的指标。
- 

## 小结

- 标准误给出了“CV 平均风险”的不确定性量级： $se = \hat{\sigma}/\sqrt{N}$ 。
- 一倍标准误差规则：选不劣于“最优 + 1SE”且最简单的模型；它在“同等表现”里偏向更稳健的那一个。
- 这条启发式在模型选择中非常常用，尤其适合小样本或曲线噪声较大的场景。

翻译并解释以下内容，翻译不能有遗漏，解释请基础且详尽，因为我的数学基础不太好。

### 6.5.3.3 CV for model selection in non-probabilistic unsupervised learning

If we are performing unsupervised learning, we must use a loss function such as  $L(x, \delta(x)) = \|x - \delta(x)\|_2^2$ , which measures reconstruction error. Here  $\delta(x)$  is some encode-decode scheme. However, as we discussed in Section 11.5.2, we cannot use CV to determine the complexity of  $\delta$ , since we will always get lower loss with a more complex model, even if evaluated on the test set. This is because more complex models will compress the data less, and induce less distortion. Consequently, we must either use probabilistic models, or invent other heuristics.

### 6.5.4 Upper bounding the risk using statistical learning theory

The principle problem with cross validation is that it is slow, since we have to fit the model multiple times. This motivates the desire to compute analytic approximations or bounds to the generalization error. This is the studied in the field of statistical learning theory (SLT). More precisely, SLT tries to bound the risk  $R(p^*, h)$  for any data distribution  $p^*$  and hypothesis  $h \in H$  in terms of the empirical risk  $\text{Remp}(D, h)$ , the sample size  $N = |D|$ , and the size of the hypothesis space  $H$ . Let us initially consider the case where the hypothesis space is finite, with size  $\dim(H) = |H|$ . In other words, we are selecting a model/hypothesis from a finite list, rather than optimizing real-valued parameters. Then we can prove the following.

**Theorem 6.5.1.** For any data distribution  $p^*$ , and any dataset  $D$  of size  $N$  drawn from  $p^*$ , the probability that our estimate of the error rate will be more than wrong, in the worst case, is upper bounded as follows:

$$P \max_{h \in H} |\text{Remp}(D, h) - R(p^*, h)| > \leq 2\dim(H)e^{-2N} \quad 2 \\ (6.66)$$

This bound tells us that the optimism of the training error increases with  $\dim(H)$  but decreases with  $N = |D|$ , as is to be expected. If the hypothesis space  $H$  is infinite (e.g., we have real-valued parameters), we cannot use  $\dim(H) = |H|$ . Instead, we can use a quantity called the Vapnik-Chervonenkis or VC dimension of the hypothesis class. See (Vapnik 1998) for details. Stepping back from all the theory, the key intuition behind statistical learning theory is quite simple. Suppose we find a model with low empirical risk. If the hypothesis space  $H$  is very big, relative to the data size, then it is quite likely that we just got "lucky" and were given a data set that is well-modeled by our chosen function by chance. However, this does not mean that such a function will have low generalization error. But if the hypothesis class is sufficiently constrained in size, and/or the training set is sufficiently large, then we are unlikely to get lucky in this way, so a low empirical risk is evidence of a low true risk.

Note that optimism of the training error does not necessarily increase with model complexity, but it does increase with the number of different models that are being searched over. The advantage of statistical learning

theory compared to CV is that the bounds on the risk are quicker to compute than using CV. The disadvantage is that it is hard to compute the VC dimension for many interesting models, and the upper bounds are usually very loose (although see (Kaariainen and Langford 2005)). One can extend statistical learning theory by taking computational complexity of the learner into account. This field is called computational learning theory or COLT. Most of this work focuses on the case where  $h$  is a binary classifier, and the loss function is 0-1 loss. If we observe a low empirical risk, and the hypothesis space is suitably "small", then we can say that our estimated function is probably approximately correct or PAC. A hypothesis space is said to be efficiently PAC-learnable if there is a polynomial time algorithm that can identify a function that is PAC. See (Kearns and Vazirani 1994) for details.

已思考 25s

下面先完整翻译，再做基础且详尽的解释（我会把原文里明显的排版小瑕疵指出来，但翻译不遗漏任何内容）。

## 翻译

### 6.5.3.3 在非概率式无监督学习中的模型选择：交叉验证 (CV)

如果我们做的是无监督学习，就必须使用一种损失函数，例如  $L(x, \delta(x)) = \|x - \delta(x)\|^2$ ，它度量重构误差。这里  $\delta(x)$  是某种编码-解码方案。然而，正如我们在第 11.5.2 节讨论的那样，不能用 CV 来确定  $\delta$  的复杂度，因为即便在测试集上评估，更复杂的模型也总会得到更小的损失。这是因为更复杂的模型对数据的压缩更少、从而引入的失真更小。因此，我们要么使用概率模型，要么发明其他启发式方法。

### 6.5.4 用统计学习理论给风险上界

交叉验证的主要问题是慢，因为我们必须把模型拟合多次。这就激发了对解析近似或泛化误差上界的需求，这一方向在统计学习理论 (SLT) 中被系统研究。更精确地说，SLT 试图用经验风险  $R_{\text{emp}}(D, h)$ 、样本量  $N = |D|$  和假设空间  $H$  的大小来上界任何数据分布  $p^{**}$  与任何假设  $h \in H$  的真实风险  $R(p^{**}, h)$ 。先考虑有限假设空间的情形，记  $\dim(H) = |H|$ 。换言之，我们是从一个有限列表中选择模型/假设，而不是优化实值参数。于是可以证明如下结论。

**定理 6.5.1.** 对任意数据分布  $p^{**}$ ，以及从  $p^{**}$  抽取的任意大小为  $N$  的数据集  $D$ ，我们关于错误率的估计在最坏情形下“超出真实值”超过某个量的概率，有如下上界：

$$P(\max_{h \in H} |R_{\text{emp}}(D, h) - R(p^*, h)| > \varepsilon) \leq 2 \dim(H) e^{-2N\varepsilon^2} \quad (6.66)$$

这个上界告诉我们：训练误差的乐观性随  $\dim(H)$  增大而增大，随  $N = |D|$  增大而减小，这与直觉一致。

如果假设空间  $H$  是无限的（例如我们有实值参数），就不能用  $\dim(H) = |H|$ 。此时可以用称为Vapnik–Chervonenkis 维 (VC 维) 的量来度量假设类。详见 (Vapnik 1998)。

**回到直觉层面，统计学习理论的核心思想很简单：假设我们找到一个经验风险很低的模型。**如果  $H$  相对数据量来说非常大，那么很可能只是我们“走运”，碰巧这份数据被我们挑的函数拟合得很好；但这并不意味着该函数的泛化误差也低。可如果假设类的规模足够受限，并且/或者训练集足够大，那么“走运”的机会就很小，于是低经验风险就能说明低真实风险。

注意：训练误差的乐观性不一定随着模型复杂度增加而增加，但会随着搜索的不同模型个数增加而增加。与 CV 相比，统计学习理论的优点是：风险上界比 CV 计算得更快；缺点是：对许多有趣模型难以计算 VC 维，而且这些上界通常很松（不过参见 Kaariainen 和 Langford 2005）。

统计学习理论可以扩展为把学习器的计算复杂度也考虑进去，这个领域称为计算学习理论 (COLT)。其中大部分工作聚焦于  $h$  为二分类器且损失是 0-1 损失的情形。若我们观察到低经验风险，且假设空间“足够小”，就可以说我们学得的函数是可能近似正确 (PAC) 的。若存在一个多项式时间的算法，能找到一个 PAC 的函数，则称该假设空间是有效 PAC 可学习 (efficiently PAC-learnable)。详见 (Kearns and Vazirani 1994)。

## 解释 (基础且详尽)

### 一、为什么无监督里“CV 选复杂度”会失灵？

- 这里的目标是重构： $\delta(x) = \text{decode}(\text{encode}(x))$ ，损失  $L = \|x - \delta(x)\|^2$ 。
- **模型越复杂**（更宽的自编码器、更多簇/码字、更多基函数），表示能力越强  $\Rightarrow$  压缩更少、失真更小  $\Rightarrow$  重构误差更低。
- 这在测试集上同样成立（因为我们只是在尽量“复制输入”），于是 CV 会偏好最复杂的模型，没法公平比较复杂度。
- 化解办法：
  - 用概率模型，比较对数似然/边际似然（天然带有“容量惩罚”，如 BIC/MDL）；
  - 或者强制瓶颈/正则化并配合其他启发式（如最小描述长度、稀疏性、早停等）。

### 二、式 (6.66) 的正确读法 (补上排版缺失的 $\varepsilon$ )

- 经典的Hoeffding + 联合界得到：

$$P(\max_{h \in H} |R_{\text{emp}}(D, h) - R(p^*, h)| > \varepsilon) \leq 2|H| e^{-2N\varepsilon^2}.$$

原文把  $\varepsilon$  漏掉了；理解时请按上式阅读。

- **含义：**想把“经验误差  $\approx$  真实误差”的置信区间弄紧，需要：
  - 更小的假设空间 ( $|H|$  小、或 VC 维小)，或
  - 更大的样本量  $N$ 。

### 三、为什么“模型复杂度”≠“搜索的模型个数”

- 训练误差的“乐观偏差”与你试了多少不同模型密切相关：试得越多，撞上“偶然贴合”的概率越高。
- 因此，即使单个模型并不复杂，只要你在巨大的搜索空间里反复试、选最低训练误差，也会严重过拟合选择过程。这正是 SLT 想量化的。

### 四、SLT/VC/PAC 与 CV 的取舍

- CV：数据驱动、几乎不用假设，但计算贵（要多次训练）。
- SLT 上界：计算快、给出保证型结论，但往往过松、且 VC 维难算；更多用于理论理解与容量控制的方向指引。
- 工程实践里常用的折中：正则化 + (小规模) CV；模型族很大时，配合早停、权重衰减、结构约束等手段降低有效容量。

### 五、无监督学习的实操建议

- 若可能：采用概率式方法，用\*\*（边际）对数似然\*\*或其信息准则（AIC/BIC/MDL）选复杂度；
- 自编码器/字典学习：必须有瓶颈（降维、稀疏、降秩）或显式正则，并用独立任务指标（如下游分类性能）或验证重构 + 正则项的综合指标来选超参；
- 避免仅凭重构误差的 CV 来决定宽度/层数/簇数。

### 小结

- 非概率式无监督：重构误差随模型复杂度单调下降  $\Rightarrow$  CV 无法选复杂度；需用概率模型或其他启发式。
- 统计学习理论给出：当假设空间“小”或样本“大”时，低经验风险  $\Rightarrow$  低真实风险；有限类的典型上界为  

$$P[\sup_h |R_{\text{emp}} - R| > \varepsilon] \leq 2|H| e^{-2N\varepsilon^2}.$$
- 优缺点：SLT 上界快但常很松、VC 维难算；CV 准但慢。实践中常把容量控制与 CV 结合使用，既控过拟合又兼顾可计算性。



翻译并解释以下内容，翻译不能有遗漏，解释请基础且详尽，因为我的数学基础不太好。

Figure 6.7 Illustration of various loss functions for binary classification. The horizontal axis is the margin  $y\eta$ , the vertical axis is the loss. The log loss uses log base 2. Figure generated by hingeLossPlot.

### 6.5.5 Surrogate loss functions

Minimizing the loss in the ERM/ RRM framework is not always easy. For example, we might want to optimize the AUC or F1 scores. Or more simply, we might just want to minimize the 0-1 loss, as is common in classification.

Unfortunately, the 0-1 risk is a very non-smooth objective and hence is hard to optimize. One alternative is to use maximum likelihood estimation instead, since log-likelihood is a smooth convex upper bound on the 0-1 risk, as we show below. To see this, consider binary logistic regression, and let  $y_i \in \{-1, +1\}$ . Suppose our decision function computes the log-odds ratio,

$$f(x_i) = \log p(y=1|x_i, w) - \log p(y=-1|x_i, w) = w^T x_i = \eta_i \quad (6.71)$$

Then the corresponding probability distribution on the output label is

$$p(y_i|x_i, w) = \text{sigm}(\eta_i) \quad (6.72)$$

Let us define the log-loss as as

$$L_{\text{LL}}(y, \eta) = -\log p(y|x, w) = \log(1 + e^{-y\eta}) \quad (6.73)$$

It is clear that minimizing the average log-loss is equivalent to maximizing the likelihood. Now consider

computing the most probable label, which is equivalent to using  $\hat{y} = -1$  if  $\eta_i < 0$  and  $\hat{y} = +1$  if  $\eta_i \geq 0$ . The 0-1 loss of our function becomes

$$L_{01}(y, \eta) = I(y = \hat{y}) = I(y\eta < 0) \quad (6.74)$$

Figure 6.7 plots these two loss functions. We see that the NLL is indeed an upper bound on the 0-1 loss. Log-loss is an example of a surrogate loss function. Another example is the hinge loss:

$$L_{\text{hinge}}(y, \eta) = \max(0, 1 - y\eta) \quad (6.75)$$

See Figure 6.7 for a plot. We see that the function looks like a door hinge, hence its name. This loss function forms the basis of a popular classification method known as support vector machines (SVM), which we will discuss in Section 14.5. The surrogate is usually chosen to be a convex upper bound, since convex functions are easy to minimize. See e.g., (Bartlett et al. 2006) for more information.

已思考 19s

下面先完整翻译，再给出基础且详尽的解释（把每个符号都讲清楚）。

## 翻译

图 6.7 二分类中若干损失函数的示意。横轴是间隔  $y\eta$ ，纵轴是损失。对数损失 (log loss) 使用的是以 2 为底的对数。图由 *hingeLossPlot* 生成。

### 6.5.5 代理 (替代) 损失函数 (Surrogate loss functions)

在 ERM/RRM 框架下直接最小化损失并不总是容易。例如，我们可能想直接优化 AUC 或 F1。更简单地说，在分类中我们可能就想最小化 0-1 损失。不幸的是，0-1 风险是一个非常不光滑的目标，因此难以优化。一个替代方案是改用极大似然估计，因为对数似然是 0-1 风险的一个光滑的凸上界 (如下所示)。为说明这一点，考虑二元逻辑回归，令  $y_i \in \{-1, +1\}$ 。假设我们的判别函数计算对数几率比：

$$f(x_i) = \log \frac{p(y=1 | x_i, w)}{p(y=-1 | x_i, w)} = w^\top x_i = \eta_i \quad (6.71)$$

于是输出标签的概率分布为

$$p(y_i | x_i, w) = \text{sigm}(y_i \eta_i) \quad (6.72)$$

定义对数损失为

$$L_{\text{nll}}(y, \eta) = -\log p(y | x, w) = \log(1 + e^{-y\eta}) \quad (6.73)$$

显然，最小化平均对数损失等价于最大化似然。再看“最可能标签”的预测：若  $\eta_i < 0$  则  $\hat{y} = -1$ ，若  $\eta_i \geq 0$  则  $\hat{y} = +1$ 。此时 0–1 损失为

$$L_{01}(y, \eta) = \mathbf{1}\{y = \hat{y}\} = \mathbf{1}\{y\eta < 0\} \quad (6.74)$$

图 6.7 画出了这两种损失。可以看到，NLL 确实是 0–1 损失的上界。对数损失是代理损失的一个例子。另一个例子是 hinge 损失：

$$L_{\text{hinge}}(y, \eta) = \max(0, 1 - y\eta) \quad (6.75)$$

见图 6.7。该函数形状像“门铰链”，故名。这个损失是著名分类方法 支持向量机 (SVM) 的基础，我们将于第 14.5 节讨论。通常选择凸的上界作为代理，因为凸函数易于最小化。更多信息见 (Bartlett et al., 2006)。

---

## 解释 (基础且详尽)

### 1) 关键符号与“间隔”

- $y \in \{-1, +1\}$ : 真实标签。
- $\eta = w^\top x$ : 模型打分 (线性时)。
- **间隔**  $y\eta$ : 若为正，模型把样本分对且离边界越远越自信；若为负，分错且绝对值越大错得越离谱。图 6.7 的横轴就是  $y\eta$ 。

### 2) 为什么 0–1 损失难优化？

- 0–1 损失:  $L_{01}(y, \eta) = \mathbf{1}\{y\eta < 0\}$ ；  
它是离散、非凸、处处不可导 (除了几乎处处为 0/1 的常数阶跃)。
- 用梯度方法几乎无从下手，也容易陷入指数时间的组合搜索。

### 3) 对数损失 (log loss / NLL) 为何是上界？

- 逻辑回归中  $p(y | x, w) = \sigma(y\eta)$ ,  $\sigma(z) = 1/(1 + e^{-z})$ 。
- NLL:  $\log(1 + e^{-y\eta})$  是光滑凸函数。
- 若用以 2 为底的对数,  $\log_2(1 + e^{-0}) = \log_2 2 = 1$ 。
  - 当  $y\eta \leq 0$  (错或临界),  $\text{NLL} \geq 1$ ;
  - 当  $y\eta > 0$  (分对),  $\text{NLL} > 0$  而 0–1 损失为 0。
- 因此整条曲线在图中都覆盖 (不低于) 黑色的 0–1 曲线——是其上界。
- 优点：可导、凸  $\Rightarrow$  能用梯度/二阶法稳定求解。

## 4) Hinge 损失与“间隔最大化”

- $L_{\text{hinge}}(y, \eta) = \max(0, 1 - y\eta)$ 。
- 若  $y\eta \geq 1$ : 损失 0 (不仅要分对, 还要有至少 1 的间隔)。
- 若  $0 < y\eta < 1$ : 分对但“靠近边界”, 有**正损失**, 鼓励把点推远;
- 若  $y\eta \leq 0$ : 分错, 损失线性增长。
- 这是 SVM 的核心: 最小化 hinge 损失 + 正则项 (如  $\|w\|_2^2$ ) , 得到**大间隔、稀疏支持向量的分类器**。
- 同样是 0-1 的**凸上界**, 但与 log loss 相比在大间隔处**更快地变为 0**, 更专注“推开边界”。

## 5) 为什么要用“代理损失” (surrogate) ?

- 许多真正关心的指标 (0-1、F1、AUC) 都**非凸或不可分解**, 直接优化困难。
- 选一个与目标“方向一致”的**凸上界** (log loss、hinge、指数损失等), **既可优化**, 又在理论上能保证**一致性/校准性** (risk 一致、分类一致)。
- 在 RRM 下再配合正则 (如  $\ell_2/\ell_1$ ) , 就完成了“可优的经验风险 + 容量控制”。

## 6) 何时选哪一种?

- log loss (**逻辑回归/最大熵**) : 输出**概率**, 对离群点相对**更温和**, 广泛用于可解释/概率需求的场景。
- hinge (SVM) : 关注**间隔**, 常在高维稀疏特征下表现很好; 不直接给概率, 可后处理 (Platt scaling)。
- 其它: **指数损失** (AdaBoost) 、**平方铰链**等, 都是 0-1 的凸上界。

## 小结

- 0-1 损失太“硬”, 直接最小化不可行; 用**代理损失** (凸上界) 替代是标准做法。
- 逻辑回归的 log loss 与 SVM 的 hinge loss 都是典型代理: 它们关于**间隔**  $y\eta$  单调下降, 且在图 6.7 中位于 0-1 曲线之上。
- 选择哪种代理取决于你是否需要**概率输出**、是否更在意**大间隔**、以及数据/特征的性质; 再配合正则与交叉验证来定强度与超参即可。

翻译并解释以下内容, 翻译不能有遗漏, 解释请基础且详尽, 因为我的数学基础不太好。

## 6.6 Pathologies of frequentist statistics

I believe that it would be very difficult to persuade an intelligent person that current [frequentist] statistical practice was sensible, but that there would be much less difficulty with an approach via likelihood and Bayes' theorem. — George Box, 1962.

Frequentist statistics exhibits various forms of weird and undesirable behaviors, known as pathologies. We give a few examples below, in order to caution the reader; these and other examples are explained in more detail in (Lindley 1972; Lindley and Phillips 1976; Lindley 1982; Berger 1985; Jaynes 2003; Minka 1999).

已思考 16s

下面先完整翻译，再做基础且详尽的解释。

## 翻译

### 6.6 频率学派统计的“病态”

我认为，要说服一个聪明人相信当下的【频率学派】统计实践是合理的非常困难；但若走似然与贝叶斯定理这条路，则说服起来困难会小得多。—— George Box, 1962

频率学派统计会出现各种古怪且不受欢迎的行为，被称为**病态** (pathologies)。下面我们会给出一些例子来提醒读者；这些例子及更多内容在 (Lindley 1972; Lindley and Phillips 1976; Lindley 1982; Berger 1985; Jaynes 2003; Minka 1999) 中有更详尽的解释。

## 解释（基础且详尽）

### 这段话在讲什么？

作者先引用统计大师 George Box 的一句话，表达一个观点：**传统的频率学派做法中，有不少让人难以接受或理解的现象；而基于似然与贝叶斯定理的做法往往更直观。**紧接着给出“病态 (pathology)”一词，指的是：按照频率学派的规则操作，会产生**反直觉、矛盾、或容易被误用**的结果。

小知识：Box 就是那位说“**所有模型都是错的，但有些是有用的**”的统计学家。

### 为什么会有“病态”？（通俗理解）

频率学派最核心的出发点是“假设数据是从某个固定但未知的分布中重复抽样得到的”。于

是很多结论——比如 p 值、显著性检验、置信区间——\*\*都与“如果反复做实验会怎样”\*\*相关，而不只与“我们这一次实际观察到的数据”相关。这会带来一些诡异之处：

### 1. 依赖抽样计划 / 停止规则 (optional stopping)

- 在频率学派里，p 值不仅取决于观察到的数据，还取决于**你打算何时停**（例如每收集10个样本看一次、直到  $p < 0.05$  就停）。不同的“停止计划”会给出不同的 p 值，即便**最终看到的数据一模一样**。
- 这让结果对实验者的“意图”敏感，容易被“p-hacking (挑着停)”利用。
- **贝叶斯**只基于实际数据的似然与先验，通常不受停止规则影响（只要似然相同）。

### 2. 置信区间的常见误解

- 95% 置信区间**不是**“参数有 95% 的概率在区间里”；正确含义是：**若重复做很多同样的实验并按同一方法做区间估计，约 95% 的区间会覆盖真值**。
- 这与我们直觉里“看到这一次的数据后，参数落在区间的概率有多大”并不相同，容易被误解或滥用。
- \*\*贝叶斯区间 (后验区间) \*\*则可直接说“参数以 95% 概率落在这里”，更贴近“这一回”的问题表述。

### 3. Lindley 悖论

- 当样本量很大时，一个极小的 p 值（频率学派会拒绝原假设）可能与**强烈支持原假设的贝叶斯因子**同时出现。
- 直观上，**微小而系统的偏差**在大样本下都会显著，但从贝叶斯角度看，这样的偏离在合理先验下并不一定足以推翻原假设。
- 这暴露了“显著性”与“实际证据强度”之间的脱节。

### 4. 多重比较 / 模型搜寻带来的乐观性

- 在同一数据上反复试很多模型（或很多特征、很多阈值）再报告最好的 p 值，**第一类错误率会被严重低估**；除非严格做校正 (Bonferroni、FDR 等)，否则极易“捡到”虚假的显著性。
- 频率学派框架可以修正，但实践中常被忽略，导致“可重复性危机”。

### 5. 违反似然原则 (likelihood principle)

- 频率学派的一些程序会让**未观察到的数据**（如果重复采样会看到什么）影响结论；而似然原则主张：**只应该由已观察到的数据的似然决定结论**。
- 这就是为什么在一些场景里，频率学派给出的答案会随“本可以发生但没发生的事”而改变。

## 作者接下来要做的事

这段只是**开场白**：提醒读者下面会列出几个典型“病态”的例子，帮助你在使用频率学派方

法时保持警惕。文末列出的参考书 (Lindley、Berger、Jaynes、Minka 等) 里有系统讨论与证明。

## 你应该带走的要点

- “病态”并不是说频率学派“不能用”，而是**容易被误解/误用**，在某些设定下会产生**违反直觉**的结果。
- 在需要对**一次性事件**给概率、或希望**只基于实际数据推断**时，**贝叶斯方法**通常更自然，也能避免部分悖论。
- 不论采用哪派方法，清楚**前提假设与适用范围**、并用**交叉验证/多重比较校正/预注册**等良好实践降低误用风险，才是关键。

翻译并解释以下内容，翻译不能有遗漏，解释请基础且详尽，因为我的数学基础不太好。

### 6.6.4 Why isn't everyone a Bayesian?

Given these fundamental flaws of frequentist statistics, and the fact that Bayesian methods do not have such flaws, an obvious question to ask is: "Why isn't everyone a Bayesian?" The (frequentist) statistician Bradley Efron wrote a paper with exactly this title (Efron 1986). His short paper is well worth reading for anyone interested in this topic. Below we quote his opening section:

The title is a reasonable question to ask on at least two counts. First of all, everyone used to be a Bayesian. Laplace wholeheartedly endorsed Bayes's formulation of the inference problem, and most 19th-century scientists followed suit. This included Gauss, whose statistical work is usually presented in frequentist terms. A second and more important point is the cogency of the Bayesian argument. Modern statisticians, following the lead of Savage and de Finetti, have advanced powerful theoretical arguments for preferring Bayesian inference. A byproduct of this work is a disturbing catalogue of inconsistencies in the frequentist point of view. Nevertheless, everyone is not a Bayesian. The current era (1986) is the first century in which statistics has been widely used for scientific reporting, and in fact, 20th-century statistics is mainly

non-Bayesian. However, Lindley (1975) predicts a change

for the 21st century.

Time will tell whether Lindley was right....

已思考 16s

下面先完整翻译，再给出基础且详尽的解释（尽量用通俗语言说明“为什么不是人人都做贝叶斯”）。

## 翻译

### 6.6.4 为什么不是所有人都是贝叶斯主义者？

既然频率学派统计有这些根本缺陷，而且贝叶斯方法没有这些缺陷，一个显而易见的问题是：“为什么不是所有人都是贝叶斯主义者？”（frequentist 的）统计学家 Bradley Efron 写过一篇题目就叫这个的论文（Efron 1986）。对这一主题感兴趣的人都值得读一读这篇短文。下面引用他开头的一段：

这个标题至少从两个方面看都是一个合理的问题。首先，以前**所有人本来就是贝叶斯的**。拉普拉斯全心全意地支持贝叶斯对推断问题的表述，19 世纪的大多数科学家也跟随其后。这其中包括高斯——他的统计工作通常被以频率学派的术语来呈现。

第二点、也是更重要的一点，是贝叶斯论证的**说服力**。当代统计学家——追随 Savage 和 de Finetti 的脚步——提出了许多有力的理论论证，来表明应当偏好贝叶斯推断。这些工作的一个副产品是：一份令人不安的清单，列举了频率学派观点中的各种不一致性。

尽管如此，**并不是所有人都是贝叶斯主义者**。当前这个时代（1986 年）是统计学被广泛用于科学报告的**第一个世纪**；事实上，20 世纪的统计学主要是**非贝叶斯的**。然而，Lindley (1975) 预测进入 21 世纪会发生改变。

时间会告诉我们 Lindley 是否说对了……

(注：上面原文里有少量拼写小误差，如 *everone* → *everyone*、*wholeheatedly* → *wholeheartedly*，不影响意思。)

## 解释（基础且详尽）

Efron 的段落先回顾：19 世纪很多大师（拉普拉斯、高斯）都以我们今天叫“贝叶斯”的思路做事；后来 20 世纪，科学界**实际用得最多的是非贝叶斯（频率学派）**。既然贝叶斯有很多优点，为什么现实里并不是“人人皆贝叶斯”？原因主要有这些（按易懂程度排列）：

### 1) 计算历史：贝叶斯曾经“算不动”

- 20世纪大半时间没有今天的计算力和软件。贝叶斯需要**积分/采样**（后验、边际似然），在复杂模型中几乎不可行。
- 频率学派方法常有**闭式解或近似**（正态近似、最小二乘、极大似然、渐近理论），**算得快、便于推广**。
- 直到1990年代MCMC、2000年代变分推断/高性能计算流行，贝叶斯才真正“落地”。很多学科仍沿用旧传统。

## 2) 先验的“主观性”与共识问题

- 贝叶斯需要**先验**。先验从哪来？不同专家可能给出不同先验，容易引发“主观/可操纵”的担忧。
- 虽然有**客观贝叶斯/参考先验、经验贝叶斯、层次模型**等缓解，但在**规范、审评、行业标准**里（如某些监管报告）仍常偏好不依赖先验的频率学派结论。

## 3) 频率学派的保证型性质很吸引人

- 无偏性、覆盖率、控制I类错误、最小方差、渐近最优等**频率学派保证**，在实验设计、假设检验、质量管控中**好沟通、好约束**。
- 例如“显著性检验/p值”在期刊与监管中根深蒂固；很多场合需要的是“长期错误率受控”的承诺，而非对单次事件的概率解读。

## 4) 教育与生态的“路径依赖”

- 20世纪的教材、课程、软件、审稿标准、行业培训**以频率学派为主**；研究者“用顺手了”。
- 论文写作与同行评审也形成了“默认语言”（p值、置信区间、显著性），迁移成本高。

## 5) 模型失配与稳健性顾虑

- 贝叶斯把**不确定性放在模型里**：若模型/先验错得离谱，后验也会“自信地错”。
- 频率学派的某些程序（如非参数检验、重抽样法、稳健估计）对模型失配更“迟钝”，在嘈杂数据下有时更可靠。
- 当然，现代贝叶斯也有**稳健先验、层次/非参数贝叶斯**来应对，但门槛较高。

## 6) “谁更好”取决于任务

- **需要概率、分层汇总、小样本、部分共享 (partial pooling)** → 贝叶斯常显著更好（比如多医院疗效对比、AB测试层次模型、稀疏事件率估计）。
- **需要极快、可扩展、合规的流水线** → 简单的频率学派法（正则化MLE、打分检验）往往更合用。
- 很多现代方法其实在中间：正则化 = MAP（先验）；交叉验证/结构风险与贝叶斯证据选择在精神上相通；“经验贝叶斯”就桥接了两者。

## 7) 交流与解读

- 受教育背景不同，“95% 置信区间”与“95% 贝叶斯后验区间”的**解读方式**不同；对非专业读者，频率学派表述（“若反复实验，有 95% 的区间会覆盖真值”）常被**误解**，但它是很多学科的既定语言。
  - 改换语言/范式需要共识与时间。
- 

## 今天的现实图景（简要）

- 越来越多领域**“贝叶斯化”**：层次模型、贝叶斯优化、贝叶斯深度学习、因果推断中的先验约束等。
  - 同时，频率学派工具仍**无处不在**：GLM、惩罚回归、重抽样、CV、显著性/多重比较校正、置信带等。
  - 许多团队采取**折衷**：当需要**全面不确定性与多层共享**时用贝叶斯；当需要**快速、稳健、易部署**时用频率学派；两者经常**互相校验**。
- 

## 你可以怎么选？

- **问任务**：要概率与决策（成本/效用）？样本小、组间需共享信息？→ 倾向贝叶斯。
  - **问资源**：算力/时间有限？需求是高速迭代与简洁报告？→ 倾向频率学派或近似贝叶斯（MAP/正则化）。
  - **双轨**：先用频率学派做一个基线 + 误差控制；再用贝叶斯做更丰富的推断与敏感性分析。
- 

**一句话总结：**不是人人都是贝叶斯，并不因为贝叶斯“不好”，而是**历史、计算、规范、沟通与任务需求**等多方面因素共同作用的结果。现在更现实的态度是：**两种思想都懂、根据问题选工具**。