

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。

## 5 Bayesian statistics

### 5.1 Introduction

We have now seen a variety of different probability models, and we have discussed how to fit them to data, i.e., we have discussed how to compute MAP parameter estimates  $\hat{\theta} = \operatorname{argmax} p(\theta|D)$ , using a variety of different priors. We have also discussed how to compute the full posterior  $p(\theta|D)$ , as well as the posterior predictive density,  $p(x|D)$ , for certain special cases (and in later chapters, we will discuss algorithms for the general case). Using the posterior distribution to summarize everything we know about a set of unknown variables is at the core of Bayesian statistics. In this chapter, we discuss this approach to statistics in more detail. In Chapter 6, we discuss an alternative approach to statistics known as frequentist or classical statistics.

好的，我先将你给的英文原文逐句翻译成中文，保证没有遗漏，然后再用通俗易懂的方式详细解释里面的数学概念和逻辑。

## 逐句翻译

### 5 贝叶斯统计

#### 5.1 引言

我们已经看过多种不同的概率模型，并且讨论了如何将它们拟合到数据上。换句话说，我们已经讨论了如何计算 MAP 参数估计

$$\hat{\theta} = \operatorname{arg max} p(\theta|D)$$

使用各种不同的先验分布。

我们还讨论了如何计算完整的后验分布  $p(\theta|D)$ ，以及在某些特殊情况下如何计算后验预测密度  $p(x|D)$ （在后面的章节中，我们会讨论适用于一般情况的算法）。

使用**后验分布**来总结我们对一组未知变量的全部认知，是贝叶斯统计的核心思想。在本章中，我们会更详细地讨论这种统计方法。

在第 6 章，我们将讨论另一种统计方法，称为**频率派或经典统计**。

## 通俗解释（数学基础友好版）

### 1. 背景

这里的主题是**贝叶斯统计** (Bayesian statistics)。它是一种用概率来描述“我们对未知事物的信心”的方法。

- “未知变量”就是你不知道的量，比如某个实验的真实概率、模型参数等等。
  - 在贝叶斯方法中，我们不会只给出一个数字作为答案，而是给出一个“分布”，这个分布表示我们对各种可能值的信心程度。
- 

## 2. 关键词解释

- **概率模型**：一种用概率规则来描述数据是如何产生的方式。例如：抛硬币可以用“硬币正面朝上的概率”来描述，这个概率就是模型的参数。
  - **拟合模型到数据**：就是根据你手上的数据，去推测模型的参数。
  - **MAP (最大后验估计)**：Maximum A Posteriori，最大后验概率估计。
    - 数学公式： $\hat{\theta} = \arg \max p(\theta|D)$
    - 含义：找到最有可能的参数值 ( $\theta$ )  
这里的  $p(\theta|D)$  叫做“后验概率”，表示在看到数据  $D$  后，参数  $\theta$  取某个值的可能性。
    - “使用不同的先验”就是在计算时对  $\theta$  的初始信念不同。
  - **后验分布  $p(\theta|D)$** ：在结合数据之后，我们对参数可能值的全部信念分布。
  - **后验预测分布  $p(x|D)$** ：预测新的数据  $x$  会长什么样的分布（是基于后验分布算出来的）。
- 

## 3. 贝叶斯思想

贝叶斯统计的核心是：

1. 一开始，我们对未知的量有一个**先验** (prior) 信念。
  2. 观察数据后，用贝叶斯公式更新这个信念，得到**后验分布**。
  3. 这个后验分布就包含了我们对未知量的全部认知。
  4. 如果要预测新数据，就用**后验预测分布**。
- 

## 4. 对比下一章的内容

作者说第 6 章会讲**频率派统计** (Frequentist statistics)：

- 贝叶斯：用概率来表示“我们对未知数的信念”。
  - 频率派：把概率理解为“长期频率”，参数是固定的，只是我们不知道。
- 

如果你愿意，我可以帮你画一个简单的流程图，把“先验 → 后验 → 预测”整个贝叶斯流程可视化，让你更容易记住。这样你以后遇到 MAP、后验分布、预测分布这些词就不会乱了。

你要我画吗？

### 贝叶斯统计与先前的高斯模型有何区别

你可以把“贝叶斯统计”和“高斯模型”理解成两个**不在同一层级**的概念：

- **贝叶斯统计**：是一种**思维方式 / 推理框架**，规定了你如何处理未知量、数据和概率。
- **高斯模型 (Gaussian model)**：是一种**具体的概率模型**，规定了数据在数学上服从什么分布，比如“均值  $\mu$ , 方差  $\sigma^2$  的正态分布”。

下面我给你分层讲清楚，并且结合例子解释它们的区别和联系。

## 1. 层级区别

	贝叶斯统计（方法论）	高斯模型（模型类型）
定义	一套用概率分布表示 <b>不确定性的</b> 统计方法，核心是用 <b>先验 + 数据 → 后验</b>	一种假设数据服从 <b>正态分布</b> 的概率模型
作用	告诉你 <b>怎么推断</b> 未知参数或未来数据	告诉你 <b>数据长什么样</b> （概率分布的形状）
关键元素	先验分布、后验分布、贝叶斯公式、预测分布	均值 $\mu$ 、方差 $\sigma^2$ 、正态分布公式
关系	框架可以用在任何模型上（高斯、二项、泊松...）	可以被放在贝叶斯框架中做推断

## 2. 举个例子对比

假设我们要估计某种测量值的平均高度（单位：cm），并且认为测量误差是正态分布的。

### 高斯模型（仅模型假设）

- 假设数据  $x$  服从  $N(\mu, \sigma^2)$
- 我们的目标是估计  $\mu$  和  $\sigma^2$ 。
- 如果只用“高斯模型”+最大似然法（频率派思想），我们会：
  1. 假设  $\mu$  和  $\sigma^2$  是固定未知的数。
  2. 找到能让观测数据概率最大的  $\mu$ 、 $\sigma^2$ （最大似然估计 MLE）。

### 贝叶斯统计 + 高斯模型

- 同样假设  $x \sim N(\mu, \sigma^2)$ ，但我们：

1. **先验**: 在看到数据之前, 对  $\mu$  有一个信念, 比如  $\mu$  服从  $N(170, 10^2)$  (我们觉得平均值大概在 170cm 左右, 但有一定不确定性)。
2. **观察数据**: 把测量的高度数据放进去。
3. **后验分布**: 通过贝叶斯公式更新, 得到  $\mu$  的后验分布 (它告诉我们数据之后对  $\mu$  的信念变化)。
4. **预测分布**: 用后验分布去预测新人的身高分布。

**区别在于:**

- 高斯模型只规定数据服从什么分布, 但不规定**怎么推断参数**。
- 贝叶斯统计规定了一个推断框架, 而高斯模型只是其中的一个具体实例。

### 3. 总结成一句话

高斯模型是一个具体的“数据长什么样”的假设, 贝叶斯统计是一种“如何根据数据更新认知”的方法。贝叶斯统计可以用在高斯模型上, 也可以用在别的模型上。

如果你愿意, 我可以帮你画一个**贝叶斯统计 + 高斯模型的推断流程图**, 让你直观看到“先验 → 数据 → 后验 → 预测”的关系。这样比单纯文字解释更容易记住。

你想让我画吗?



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 5.1 (a) A bimodal distribution in which the mode is very untypical of the distribution. The thin blue vertical line is the mean, which is arguably a better summary of the distribution, since it is near the majority of the probability mass. Figure generated by bimodalDemo. (b) A skewed distribution in which the mode is quite different from the mean. Figure generated by gammaPlotDemo.

Figure 5.2 Example of the transformation of a density under a nonlinear transform. Note how the mode of the transformed distribution is not the transform of the

original mode. Based on Exercise 1.4 of (Bishop 2006b).

Figure generated by bayesChangeOfVar.

## 5.2 Summarizing posterior distributions

The posterior  $p(\theta|D)$  summarizes everything we know about the unknown quantities  $\theta$ . In this section, we discuss some simple quantities that can be derived from a probability distribution, such as a posterior. These summary statistics are often easier to understand and visualize than the full joint.

### 5.2.1 MAP estimation

We can easily compute a point estimate of an unknown quantity by computing the posterior mean, median or mode. In Section 5.7, we discuss how to use decision theory to choose between these methods. Typically the posterior mean or median is the most appropriate choice for a realvalued quantity, and the vector of posterior marginals is the best choice for a discrete quantity.

However, the posterior mode, aka the MAP estimate, is the most popular choice because it reduces to an optimization problem, for which efficient algorithms often exist. Furthermore, MAP estimation can be interpreted in non-Bayesian terms, by thinking of the log prior as a regularizer (see Section 6.5 for more details).

Although this approach is computationally appealing, it is important to point out that there are various drawbacks to MAP estimation, which we briefly discuss below. This will provide motivation for the more thoroughly Bayesian approach which we will study later in this chapter (and elsewhere in this book).

#### 5.2.1.1 No measure of uncertainty

The most obvious drawback of MAP estimation, and indeed of any other point estimate such as the posterior mean or median, is that it does not provide any measure of uncertainty. In many applications, it is important to know how much one can trust a given estimate. We can derive such confidence measures from the posterior, as we discuss in Section 5.2.2.

#### 5.2.1.2 Plugging in the MAP estimate can result in overfitting

In machine learning, we often care more about predictive accuracy than in interpreting the parameters of our models. However, if we don't model the uncertainty in our parameters, then our predictive distribution will be overconfident. We saw several examples of this in Chapter 3, and we will see more examples later. Overconfidence in predictions is particularly problematic in situations where

we may be risk averse; see Section 5.7 for details.

#### 5.2.1.3 The mode is an untypical point

Choosing the mode as a summary of a posterior distribution is often a very poor choice, since the mode is usually quite untypical of the distribution, unlike the mean or median. This is illustrated in Figure 5.1(a) for a 1d continuous space. The basic problem is that the mode is a point of measure zero, whereas the mean and median take the volume of the space into account. Another example is shown in Figure 5.1(b): here the mode is 0, but the mean is non-zero. Such skewed distributions often arise when inferring variance parameters, especially in hierarchical models. In such cases the MAP estimate (and hence the MLE) is obviously a very bad estimate.

How should we summarize a posterior if the mode is not a good choice? The answer is to use decision theory, which we discuss in Section 5.7. The basic idea is to specify a loss function, where  $L(\theta, \hat{\theta})$  is the loss you incur if the truth is  $\theta$  and your estimate is  $\hat{\theta}$ . If we use 0-1 loss,  $L(\theta, \hat{\theta}) = I(\theta = \hat{\theta})$ , then the optimal estimate is the posterior mode. 0-1 loss means you only get “points” if you make no errors, otherwise you get nothing: there is no “partial credit” under this loss function! For continuous-valued quantities, we often prefer to use squared error loss,  $L(\theta, \hat{\theta}) = (\theta - \hat{\theta})^2$ ; the corresponding optimal estimator is then the posterior mean, as we show in Section 5.7. Or we can use a more robust loss function,  $L(\theta, \hat{\theta}) = |\theta - \hat{\theta}|$ , which gives rise to the posterior median.

#### 5.2.1.4 MAP estimation is not invariant to reparameterization

A more subtle problem with MAP estimation is that the result we get depends on how we parameterize the probability distribution. Changing from one representation to another equivalent representation changes the result, which is not very desirable, since the units of measurement are arbitrary (e.g., when measuring distance, we can use centimetres or inches). To understand the problem, suppose we compute the posterior for  $x$ . If we define  $y = f(x)$ , the distribution for  $y$  is given by Equation 2.87, which we repeat here for convenience:

$$py(y) = px(x) |dx dy| \quad (5.1)$$

The  $|dx dy|$  term is called the Jacobian, and it measures the change in size of a unit volume passed through  $f$ . Let  $\hat{x} = \operatorname{argmax}_x px(x)$  be the MAP estimate for  $x$ . In general it is not the case that  $\hat{y} = \operatorname{argmax}_y py(y)$  is given by  $f(\hat{x})$ . For example, let  $x \sim N(6, 1)$  and  $y = f(x)$ , where

$$f(x) = \frac{1}{1 + \exp(-x + 5)} \quad (5.2)$$

We can derive the distribution of  $y$  using Monte Carlo simulation (see Section 2.7.1). The result is shown in Figure 5.2. We see that the original Gaussian has become "squashed" by the sigmoid nonlinearity. In particular, we see that the mode of the transformed distribution is not equal to the transform of the original mode.

To see how this problem arises in the context of MAP estimation, consider the following example, due to Michael Jordan. The Bernoulli distribution is typically parameterized by its mean  $\mu$ , so  $p(y=1|\mu) = \mu$ , where  $y \in \{0, 1\}$ . Suppose we have a uniform prior on the unit interval:  $p(\mu) = 1 \text{ I}(0 \leq \mu \leq 1)$ . If there is no data, the MAP estimate is just the mode of the prior, which can be anywhere between 0 and 1. We will now show that different parameterizations can pick different points in this interval arbitrarily. First let  $\theta = \sqrt{\mu}$  so  $\mu = \theta^2$ . The new prior is

$$p(\theta) = p(\mu) d\mu d\theta = 2\theta \quad (5.3)$$

for  $\theta \in [0, 1]$  so the new mode is

$$\hat{\theta}^{\text{MAP}} = \arg \max_{\theta \in [0, 1]} 2\theta = 1 \quad (5.4)$$

Now let  $\varphi = 1 - \sqrt{1 - \mu}$ . The new prior is

$$p(\varphi) = p(\mu) d\mu d\varphi = 2(1 - \varphi) \quad (5.5)$$

for  $\varphi \in [0, 1]$ , so the new mode is

$$\hat{\varphi}^{\text{MAP}} = \arg \max_{\varphi \in [0, 1]} 2 - 2\varphi = 0 \quad (5.6)$$

Thus the MAP estimate depends on the parameterization. The MLE does not suffer from this since the likelihood is a function, not a probability density. Bayesian inference does not suffer from this problem either, since the change of measure is taken into account when integrating over the parameter space. One solution to the problem is to optimize the following objective function:

$$\hat{\theta}^{\text{MAP}} = \operatorname{argmax}_{\theta} \theta p(D|\theta) p(\theta) |I(\theta)|^{-1/2} \quad (5.7)$$

Here  $I(\theta)$  is the Fisher information matrix associated with  $p(x|\theta)$  (see Section 6.2.2). This estimate is parameterization independent, for reasons explained in (Jermyn 2005; Druihet and Marin 2007). Unfortunately, optimizing Equation 5.7 is often difficult, which minimizes the appeal of the whole approach.

好的，我先逐句翻译，然后再用基础且详尽的方式帮你解释，包括两张图的含义。

## 逐句翻译

### 图 5.1

(a) 一个双峰分布 (bimodal distribution)，其中众数 (mode) 非常不典型。细蓝色竖线

表示**均值** (mean) , 它或许是更好的分布摘要, 因为它更接近大多数概率质量所在的位置。图由 `bimodalDemo` 生成。

(b) 一个偏斜分布 (skewed distribution) , 其中众数与均值差异很大。图由 `gammaPlotDemo` 生成。

## 图 5.2

密度在非线性变换下的一个例子。注意, **变换后的分布的众数不等于原众数经过变换的结果**。基于 (Bishop 2006b) 习题 1.4。图由 `bayesChangeOfVar` 生成。

---

## 5.2 后验分布的摘要

后验分布  $p(\theta|D)$  总结了我们对未知量  $\theta$  的全部已知信息。本节我们讨论一些可以从概率分布 (如后验分布) 中推导出来的简单统计量。这些摘要统计量通常比完整的联合分布更容易理解和可视化。

---

### 5.2.1 MAP 估计

我们可以通过计算后验分布的**均值、中位数或众数**来获得未知量的一个点估计。在 5.7 节中, 我们会用决策理论 (decision theory) 来选择这些方法中的哪一个。

通常:

- 对实数值变量, 后验均值或中位数是最合适的;
- 对离散变量, 后验边缘分布向量 (vector of posterior marginals) 是最佳选择;
- **后验众数** (MAP 估计) 是最常见的选择, 因为它等价于一个优化问题, 而优化通常有高效算法。

此外, MAP 估计还能用非贝叶斯的角度来解释: 把对数先验看作正则项 (regularizer) (详见 6.5 节)。

尽管这种方法计算方便, 但它有一些缺点, 我们会在下面简要讨论。这也为我们采用更彻底的贝叶斯方法提供了动机。

---

#### 5.2.1.1 无法衡量不确定性

MAP 估计 (以及后验均值、中位数等其他点估计) 的一个明显缺点是: 它不能告诉我们估计的不确定性。在很多应用中, 知道我们能多大程度上信任某个估计是很重要的。这类置信度信息可以从后验分布中得到 (详见 5.2.2 节)。

---

#### 5.2.1.2 代入 MAP 估计可能导致过拟合

在机器学习中, 我们通常更关心预测准确性, 而不是模型参数的解释。但如果不对参数的不确定性建模, 那么预测分布往往**过于自信**。我们在第 3 章已经见过一些例子, 后面还会看到更多。过度自信的预测在风险规避场景中特别危险 (详见 5.7 节)。

### 5.2.1.3 众数是一个不典型的点

用众数作为后验分布的摘要，常常是一个很差的选择，因为众数通常并不能代表分布的典型值，而均值或中位数则更具代表性。

图 5.1(a) 展示了一维连续空间中的这种情况：众数是一个“测度为零”的点，而均值和中位数则考虑了空间体积的影响。

另一个例子是图 5.1(b)：这里众数是 0，但均值却不为 0。这种偏斜分布在推断方差参数时很常见，尤其是在分层模型中。这种情况下，MAP（以及 MLE）显然是糟糕的估计。

如果众数不是好选择，该如何总结后验分布？答案是用**决策理论**（详见 5.7 节）：

- 定义损失函数  $L(\theta, \hat{\theta})$ ，表示当真实值是  $\theta$ ，估计值是  $\hat{\theta}$  时的损失。
- 如果用 0-1 损失  $L(\theta, \hat{\theta}) = I(\theta = \hat{\theta})$ ，最优估计是后验众数（完全正确才得分，没有部分分）。
- 对连续变量，常用平方误差损失  $L(\theta, \hat{\theta}) = (\theta - \hat{\theta})^2$ ，最优估计是后验均值。
- 也可以用更稳健的绝对误差损失  $L(\theta, \hat{\theta}) = |\theta - \hat{\theta}|$ ，最优估计是后验中位数。

### 5.2.1.4 MAP 估计不具备重参数化不变性

一个更隐蔽的问题是，MAP 结果依赖于概率分布的参数化方式。换一种等价的表示，结果可能会变，这很不理想，因为计量单位是任意的（如距离用厘米还是英寸）。

假设我们计算  $x$  的后验分布。如果定义  $y = f(x)$ ，则  $y$  的分布为：

$$p_y(y) = p_x(x) \left| \frac{dx}{dy} \right| \quad (5.1)$$

这里  $|dx/dy|$  叫雅可比 (Jacobian)，它衡量通过  $f$  变换时单位体积的大小变化。

令  $\hat{x} = \arg \max_x p_x(x)$  是  $x$  的 MAP 估计，一般情况下， $\hat{y} = \arg \max_y p_y(y)$  并不等于  $f(\hat{x})$ 。

例如： $x \sim N(6, 1)$ ,  $y = f(x)$ , 其中

$$f(x) = \frac{1}{1 + \exp(-x + 5)} \quad (5.2)$$

用蒙特卡罗方法可得  $y$  的分布（见图 5.2）。可以看到，高斯分布经过 Sigmoid 非线性变换后被“压缩”，且变换后的众数  $\neq$  原众数的变换值。

### Michael Jordan 的例子：

伯努利分布常用均值  $\mu$  参数化： $p(y=1|\mu) = \mu$ ,  $y \in \{0, 1\}$ 。假设  $\mu$  的先验在  $[0, 1]$  上均

匀:  $p_\mu(\mu) = 1 \cdot I(0 \leq \mu \leq 1)$ 。

无数据时, MAP 是先验的众数, 可以是 0 到 1 之间任意值。

1. 令  $\theta = \sqrt{\mu}$ , 则  $\mu = \theta^2$ , 新先验为:

$$p_\theta(\theta) = p_\mu(\mu) \left| \frac{d\mu}{d\theta} \right| = 2\theta \quad (5.3)$$

$\theta \in [0, 1]$ , 众数是:

$$\hat{\theta}_{MAP} = \arg \max_{\theta \in [0, 1]} 2\theta = 1 \quad (5.4)$$

2. 令  $\phi = 1 - \sqrt{1 - \mu}$ , 新先验为:

$$p_\phi(\phi) = p_\mu(\mu) \left| \frac{d\mu}{d\phi} \right| = 2(1 - \phi) \quad (5.5)$$

$\phi \in [0, 1]$ , 众数是:

$$\hat{\phi}_{MAP} = \arg \max_{\phi \in [0, 1]} 2 - 2\phi = 0 \quad (5.6)$$

可见 MAP 依赖参数化方式。MLE 不受此影响, 因为似然是函数而不是概率密度。完整的贝叶斯推断也不受影响, 因为积分时考虑了测度变换。

一个解决方法是优化:

$$\hat{\theta} = \arg \max_{\theta} p(D|\theta)p(\theta)|I(\theta)|^{-1/2} \quad (5.7)$$

其中  $I(\theta)$  是 Fisher 信息矩阵 (见 6.2.2 节)。该估计对参数化不敏感, 但优化困难, 降低了其实用性。

## 基础且详尽解释

- 图 5.1(a): 双峰分布有两个山峰 (两个高概率区)。众数 (最高点位置) 可能在概率质量很少的地方, 所以不代表整体分布, 均值更接近多数数据所在。
- 图 5.1(b): 偏斜分布 (如长尾分布) 中, 众数可能在一边, 而均值偏移很多, 这在方差推断等场景很常见。

### 3. MAP 的问题:

- 没有告诉你不确定性 → 你只得到一个点, 不知道它是否可靠。
- 容易过拟合 → 忽略参数不确定性时, 预测太自信。
- 众数不典型 → 特别是分布不对称或多峰时, 众数可能很不代表整体。
- 不具备重参数化不变性 → 换个变量定义, 结果可能变了 (图 5.2 演示)。

### 4. 决策理论的替代方法:

- 用不同损失函数选均值、中位数或众数, 来更合理地总结后验分布。

5. **贝叶斯的优点：**完整的后验分布考虑了所有可能的参数值和它们的概率，自然避免了很多 MAP 的问题。

如果你愿意，我可以帮你画一张\*\*“MAP 缺点总结图 + 图 5.1 & 图 5.2 的直观解释图”\*\*，让你更直观地理解众数和均值的差别，以及重参数化问题。这样会比单看文字更容易掌握。

你要我画吗？

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。

### 5.2.2 Credible intervals

In addition to point estimates, we often want a measure of confidence. A standard measure of confidence in some (scalar) quantity  $\theta$  is the “width” of its posterior distribution. This can be measured using a  $100(1 - \alpha)\%$  credible interval, which is a (contiguous) region  $C = (l, u)$  (standing for lower and upper) which contains  $1 - \alpha$  of the posterior probability mass, i.e.,

$$C_{\alpha}(D) = (l, u): P(l \leq \theta \leq u | D) = 1 - \alpha \quad (5.8)$$

There may be many such intervals, so we choose one such that there is  $(1 - \alpha)/2$  mass in each tail; this is called a central interval.

If the posterior has a known functional form, we can compute the posterior central interval using  $l = F^{-1}(\alpha/2)$  and  $u = F^{-1}(1 - \alpha/2)$ , where  $F$  is the cdf of the posterior.

For example, if the posterior is Gaussian,  $p(\theta | D) = N(0, 1)$ , and  $\alpha = 0.05$ , then we have  $l = \Phi(\alpha/2) = -1.96$ , and  $u = \Phi(1 - \alpha/2) = 1.96$ , where  $\Phi$  denotes the cdf of the Gaussian.

This is illustrated in Figure 2.3(c). This justifies the common practice of quoting a credible interval in the form of  $\mu \pm 2\sigma$ , where  $\mu$  represents the posterior mean,  $\sigma$  represents the posterior standard deviation, and 2 is a good approximation to 1.96. Of course, the posterior is not always Gaussian. For example, in our coin example, if we use a uniform prior and we observe  $N_1 = 47$  heads out of  $N = 100$  trials, then the posterior is a beta distribution,  $p(\theta | D) = \text{Beta}(48, 54)$ . We find the 95% posterior credible interval is  $(0.3749, 0.5673)$  (see `betaCredibleInt` for the one line of Matlab code we used to compute this). If we don't know the functional form, but we can draw samples from

the posterior, then we can use a Monte Carlo approximation to the posterior quantiles: we simply sort the  $S$  samples, and find the one that occurs at location  $\alpha/S$  along the sorted list. As  $S \rightarrow \infty$ , this converges to the true quantile. See `mcQuantileDemo` for a demo. People often confuse Bayesian credible intervals with frequentist confidence intervals. However, they are not the same thing, as we discuss in Section 6.6.1. In general, credible intervals are usually what people want to compute, but confidence intervals are usually what they actually compute, because most people are taught frequentist statistics but not Bayesian statistics. Fortunately, the mechanics of computing a credible interval is just as easy as computing a confidence interval (see e.g., `betaCredibleInt` for how to do it in Matlab).

好的，我会先逐句翻译你提供的英文内容，确保不遗漏任何细节，然后再用**基础且详尽**的方式解释这些概念，并结合例子帮你理解。

---

## 逐句翻译

### 5.2.2 可信区间 (Credible intervals)

除了点估计 (point estimates) 之外，我们通常还希望有一个**置信度的度量**。

对于某个 (标量) 变量  $\theta$ ，一种常见的置信度度量是它的后验分布的“宽度”。

这种宽度可以用  $100(1 - \alpha)\%$  **可信区间** (credible interval) 来衡量，这个区间是一个 (连续的) 区域

$$C = (l, u) \quad (\text{表示下界和上界})$$

它包含了后验概率质量的  $1 - \alpha$ ，即

$$C_\alpha(D) = (l, u) : P(l \leq \theta \leq u | D) = 1 - \alpha \quad (5.8)$$

可能会有多个这样的区间，所以我们选择一个使得左右两侧尾部 (tails) 各有  $(1 - \alpha)/2$  的概率质量，这称为**中心区间** (central interval)。

如果后验分布的函数形式已知，我们可以用公式

$$l = F^{-1}(\alpha/2), \quad u = F^{-1}(1 - \alpha/2)$$

来计算中心区间，其中  $F$  是后验分布的累积分布函数 (CDF)。

例如，如果后验是高斯分布  $p(\theta | D) = N(0, 1)$ ，且  $\alpha = 0.05$ ，那么：

$$l = \Phi^{-1}(\alpha/2) = -1.96, \quad u = \Phi^{-1}(1 - \alpha/2) = 1.96$$

这里  $\Phi$  表示高斯分布的 CDF。这在图 2.3(c) 中有示意。

这解释了为什么常见的做法是用  $\mu \pm 2\sigma$  来表示可信区间，其中：

- $\mu$  是后验均值，
- $\sigma$  是后验标准差，
- 数字 2 是 1.96 的近似值。

当然，后验分布并不总是高斯分布。

例如，在我们的抛硬币例子中，如果使用均匀先验，并且观察到  $N_1 = 47$  次正面、总共  $N = 100$  次试验，那么后验分布是 Beta 分布：

$$p(\theta | D) = \text{Beta}(48, 54)$$

我们得到的 95% 后验可信区间是  $(0.3749, 0.5673)$  (见 `betaCredibleInt` 这一行 Matlab 代码的计算)。

如果我们不知道分布的函数形式，但可以从后验分布中采样，那么可以用蒙特卡罗方法近似后验分位数：

- 把采样得到的  $S$  个样本排序，
- 找到排序后处在  $\alpha/S$  位置的那个值。

当  $S \rightarrow \infty$  时，这个近似会收敛到真实的分位数。

(见 `mcQuantileDemo` 示例。)

人们常常混淆贝叶斯可信区间 (credible interval) 与频率派置信区间 (confidence interval)。然而，它们并不是同一回事 (详见 6.6.1 节)。

一般来说，可信区间才是人们想要计算的东西，但置信区间却是人们实际计算的东西，因为大多数人学过的是频率派统计，而不是贝叶斯统计。

幸运的是，计算可信区间的方法与计算置信区间一样容易 (比如在 Matlab 中用 `betaCredibleInt` 计算)。

## 基础且详尽解释

### 1. 可信区间是什么？

- **点估计**：只给你一个值，比如“后验均值是 0.5”。
- **可信区间**：告诉你一个范围，比如“有 95% 的概率  $\theta$  在 0.38 到 0.57 之间”。  
这个概率是基于后验分布算出来的。

它和“高”或“窄”有关：

- 如果后验分布很窄，可信区间就短，表示我们对参数很确定。
- 如果分布很宽，可信区间就长，表示不确定性高。

### 2. 中心区间的含义

中心区间保证：

- 左边有  $(1 - \alpha)/2$  的概率质量，
- 右边也有  $(1 - \alpha)/2$  的概率质量。

比如：

- 95% 可信区间 ( $\alpha = 0.05$ ) → 左右各有 2.5% 的概率在区间外。
- 

### 3. 高斯分布的例子

- $p(\theta | D) = N(0, 1)$ ,  $\alpha = 0.05$
  - 中心区间是  $(-1.96, 1.96)$
  - 这个 1.96 来自高斯分布的分位数公式  $\Phi^{-1}(0.975)$ 。
  - 常见简写  $\mu \pm 2\sigma$  其实就是这个意思，只是把 1.96 近似成 2。
- 

### 4. 非高斯分布的例子

- 抛硬币 100 次，正面 47 次，均匀先验 → 后验是 Beta(48, 54)。
  - 算出来的 95% 可信区间是  $(0.3749, 0.5673)$  → 表示参数（硬币正面概率）在这个范围的可能性是 95%。
- 

### 5. 用采样来算（蒙特卡罗方法）

如果后验分布没公式可写，可以：

1. 从后验分布采样很多次（比如  $S = 10000$  个样本）。
2. 排序。
3. 找到处于 2.5% 和 97.5% 位置的样本值，就是 95% 可信区间。

样本越多，结果越准。

---

### 6. 和置信区间的区别

- **可信区间（贝叶斯）**：直接解释为“参数在这个区间的概率是 95%”。
- **置信区间（频率派）**：解释更复杂，不直接表示参数在区间内的概率，而是基于长期重复实验的比例。

多数人想要的是**可信区间**，但往往计算的是**置信区间**，因为传统统计课教的是频率派方法。

---

如果你愿意，我可以帮你画一张**可信区间 vs 置信区间的直观对比图**，用概率曲线直接展示两者区别，这样你的印象会特别深。

你要我画吗？



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 5.3 (a) Central interval and (b) HPD region for a Beta(3,9) posterior. The CI is (0.06, 0.52) and the HPD is (0.04, 0.48). Based on Figure 3.6 of (Hoff 2009). Figure generated by betaHPD.

Figure 5.4 (a) Central interval and (b) HPD region for a hypothetical multimodal posterior. Based on Figure 2.2 of (Gelman et al. 2004). Figure generated by postDensityIntervals.

### 5.2.2.1 Highest posterior density regions

A problem with central intervals is that there might be points outside the CI which have higher probability density. This is illustrated in Figure 5.3(a), where we see that points outside the left-most CI boundary have higher density than those just inside the right-most CI boundary. This motivates an alternative quantity known as the highest posterior density or HPD region. This is defined as the (set of) most probable points that in total constitute  $100(1 - \alpha)\%$  of the probability mass. More formally, we find the threshold  $p^*$  on the pdf such that

$$1 - \alpha = \theta : p(\theta|D) > p^* \int p(\theta|D)d\theta \quad (5.9)$$

and then define the HPD as

$$C\alpha(D) = \{\theta : p(\theta|D) \geq p^*\} \quad (5.10)$$

In 1d, the HPD region is sometimes called a highest density interval or HDI. For example, Figure 5.3(b) shows the 95% HDI of a Beta(3, 9) distribution, which is (0.04, 0.48). We see that this is narrower than the CI, even though it still contains 95% of the mass; furthermore, every point inside of it has higher density than every point outside of it. For a unimodal distribution, the HDI will be the narrowest interval around the mode containing 95% of the mass. To see this, imagine “water filling” in reverse, where we lower the level until 95% of the mass is

revealed, and only 5% is submerged. This gives a simple algorithm for computing HDIs in the 1d case: simply search over points such that the interval contains 95% of the mass and has minimal width. This can be done by 1d numerical optimization if we know the inverse CDF of the distribution, or by search over the sorted data points if we have a bag of samples (see betaHPD for a demo). If the posterior is multimodal, the HDI may not even be a connected region: see Figure 5.4(b) for an example. However, summarizing multimodal posteriors is always difficult.

好的，我先逐句翻译，然后再用你的数学基础能理解的方式，结合图解释清楚。

## 逐句翻译

### 图 5.3

(a) Beta(3,9) 后验分布的**中心区间** (Central Interval) 和

(b) **最高后验密度区域** (HPD region)。

CI (中心区间) 是 (0.06, 0.52), HPD 是 (0.04, 0.48)。

基于 (Hoff 2009) 图 3.6, 图由 `betaHPD` 生成。

### 图 5.4

(a) 一个假想的多峰 (multimodal) 后验分布的中心区间,

(b) 对应的 HPD 区域。

基于 (Gelman et al. 2004) 图 2.2, 图由 `postDensityIntervals` 生成。

### 5.2.2.1 最高后验密度区域 (Highest posterior density regions)

#### 中心区间 (CI) 的问题

中心区间可能会包含一些概率密度较低的点，同时在区间外可能有更高密度的点。

例如图 5.3(a): 在左侧区间外的某些点的密度，比右侧区间边界内的点还高。

这就引出了另一种指标: **最高后验密度区域 (HPD region)**。

定义: 由一组概率最高的点组成, 总共包含  $100(1 - \alpha)\%$  的概率质量。

形式化定义:

1. 找到概率密度函数 (pdf) 上的一个阈值  $p^*$ , 使得:

$$1 - \alpha = \int_{\theta: p(\theta|D) > p^*} p(\theta | D) d\theta \quad (5.9)$$

2. 定义 HPD 为:

$$C_\alpha(D) = \{\theta : p(\theta | D) \geq p^*\} \quad (5.10)$$

在一维情况下，HPD 区域有时称为**最高密度区间** (HDI, Highest Density Interval)。

例如：图 5.3(b) 显示了 Beta(3,9) 分布的 95% HDI：区间是 (0.04, 0.48)。

它比中心区间更窄，但仍包含 95% 的概率质量，并且区间内的所有点密度都比区间外所有点高。

对于单峰分布 (unimodal distribution)，HDI 是围绕众数的最窄区间，包含 95% 的概率质量。

形象理解：像“倒灌水”一样，把水位从高处慢慢降低，直到露出 95% 的面积（剩下 5% 被水覆盖）。

计算 HDI 的方法 (1D 情况)：

- 搜索所有包含 95% 概率质量的区间，选择最窄的一个。
- 如果已知分布的反 CDF，可以用一维数值优化计算。
- 如果有样本，可以先排序再找最窄区间 (`betaHPD` 示例)。

如果后验分布是多峰的，HDI 可能不是一个连续区间（见图 5.4(b)）。

但多峰分布本身的摘要就比较困难。

## 基础且详尽解释

### 1. 中心区间 (CI) 的特点

- 它是根据概率分布的累计概率来取的两端（左右尾部对称各去掉  $\alpha/2$  的概率质量）。
- 好处：计算方便，直观。
- 问题：可能把一些高密度区域排除在外，而把低密度区域包含进来。

图 5.3(a) 就是个例子：

- 左边区间外的密度，比右边区间内的密度还高 → 这说明 CI 并不总是把“最有可能”的区域选出来。

### 2. HPD (最高后验密度区域) 的特点

- 按照**密度最高优先**的原则选点，直到总概率质量达到 95%（或其他设定值）。
- 保证区间内每个点的密度  $\geq$  区间外的任意一点。
- 对单峰分布，它就是围绕众数的最窄区间。
- 对多峰分布，它可能变成多个分离的区间（图 5.4(b)）。

图 5.3(b)：

- HPD 比 CI 短 ( $0.48 - 0.04 < 0.52 - 0.06$ )。
- 里面全是“高概率”的点，更符合“可信区域”的直觉。

### 3. 多峰分布的情况

图 5.4(a): 中心区间可能跨过低密度谷地, 把两个峰之间的低概率区域也包括进来。

图 5.4(b): HPD 会直接避开这种低概率区域, 只保留各个高密度峰顶周围的部分, 即可能是多个区间。

### 4. 形象类比

- CI: 像是用尺子量长度, 两边截掉同样的尾巴, 不管里面是高山还是低谷。
- HPD: 像是给山涂色, 从山顶开始往下涂, 直到总面积够 95%, 这样涂出来的一定是“山高”的地方。

我建议我帮你画一张CI vs HPD 对比示意图, 把“高密度区域优先”与“等尾切法”直观放在一起, 你会一眼看出区别。

要我帮你画吗?



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 5.5 (a) Exact posteriors  $p(\theta_i|D_i)$ . (b) Monte Carlo approximation to  $p(\delta|D)$ . We use kernel density estimation to get a smooth plot. The vertical lines enclose the 95% central interval. Figure generated by amazonSellerDemo,

#### 5.2.3 Inference for a difference in proportions

Sometimes we have multiple parameters, and we are interested in computing the posterior distribution of some function of these parameters. For example, suppose you are about to buy something from Amazon.com, and there are two sellers offering it for the same price. Seller 1 has 90 positive reviews and 10 negative reviews. Seller 2 has 2 positive reviews and 0 negative reviews. Who should you

buy from?

On the face of it, you should pick seller 2, but we cannot be very confident that seller 2 is better since it has had so few reviews. In this section, we sketch a Bayesian analysis of this problem. Similar methodology can be used to compare rates or proportions across groups for a variety of other settings. Let  $\theta_1$  and  $\theta_2$  be the unknown reliabilities of the two sellers. Since we don't know much about them, we'll endow them both with uniform priors,  $\theta_i \sim \text{Beta}(1, 1)$ . The posteriors are  $p(\theta_1|D_1) = \text{Beta}(91, 11)$  and  $p(\theta_2|D_2) = \text{Beta}(3, 1)$ . We want to compute  $p(\theta_1 > \theta_2|D)$ . For convenience, let us define  $\delta = \theta_1 - \theta_2$  as the difference in the rates. (Alternatively we might want to work in terms of the log-odds ratio.) We can compute the desired quantity using numerical integration:

$$p(\delta > 0|D) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} I(\theta_1 > \theta_2) \text{Beta}(\theta_1|y_1 + 1, N_1 - y_1 + 1) \text{Beta}(\theta_2|y_2 + 1, N_2 - y_2 + 1) d\theta_1 d\theta_2 \quad (5.11)$$

We find  $p(\delta > 0|D) = 0.710$ , which means you are better off buying from seller 1! See `amazonSellerDemo` for the code. (It is also possible to solve the integral analytically (Cook 2005).) A simpler way to solve the problem is to approximate the posterior  $p(\delta|D)$  by Monte Carlo sampling. This is easy, since  $\theta_1$  and  $\theta_2$  are independent in the posterior, and both have beta distributions, which can be sampled from using standard methods. The distributions  $p(\theta_i|D_i)$  are shown in Figure 5.5(a), and a MC approximation to  $p(\delta|D)$ , together with a 95% HPD, is shown Figure 5.5(b). An MC approximation to  $p(\delta > 0|D)$  is obtained by counting the fraction of samples where  $\theta_1 > \theta_2$ ; this turns out to be 0.718, which is very close to the exact value. (See `amazonSellerDemo` for the code.)

好的，我先逐句、无遗漏地翻译，然后再用基础且详尽的方式解释图和公式里在做什么，以及它们意味着什么。

## 逐句翻译

### 图 5.5

- (a) 精确的后验分布  $p(\theta_i | D_i)$ 。
- (b)  $p(\delta | D)$  的蒙特卡罗近似。我们使用核密度估计来得到一条平滑的曲线。两条竖线围住的是 95% 中心区间。图由 `amazonSellerDemo` 生成。

### 5.2.3 比例之差的推断 (Inference for a difference in proportions)

有时我们有多个参数，并且我们关心的是这些参数某个函数的后验分布。例如，假设你要在 Amazon.com 上买东西，有两个卖家同价：

卖家 1 有 90 条好评、10 条差评；卖家 2 有 2 条好评、0 条差评。你该选谁？

表面看应该选卖家 2，但由于卖家 2 的评价太少，我们很难对“卖家 2 更好”这件事有足够的信心。本节我们给出这个问题的一个贝叶斯分析。类似的方法可用于在许多场景下对不同组的率或比例进行比较。

设  $\theta_1$  和  $\theta_2$  是两个卖家的未知可靠度。因为我们对他们不了解，给二者都赋予均匀先验：  
 $\theta_i \sim \text{Beta}(1, 1)$ 。

后验为  $p(\theta_1 | D_1) = \text{Beta}(91, 11)$  和  $p(\theta_2 | D_2) = \text{Beta}(3, 1)$ 。

我们想计算  $p(\theta_1 > \theta_2 | D)$ 。为方便，令  $\delta = \theta_1 - \theta_2$  表示两者比例之差。（或者，我们也可以改用**对数优势比** (log-odds ratio) 来表述。）

我们可以用数值积分来计算所需量：

$$p(\delta > 0 | D) = \int_0^1 \int_0^1 \mathbf{I}(\theta_1 > \theta_2) \text{Beta}(\theta_1 | y_1 + 1, N_1 - y_1 + 1) \text{Beta}(\theta_2 | y_2 + 1, N_2 - y_2 + 1) d\theta_1 d\theta_2$$

计算得  $p(\delta > 0 | D) = 0.710$ ，这意味着你买**卖家 1**更好！见 `amazonSellerDemo` 的代码。（该积分也可以解析求解 (Cook 2005)。）

一个更简单的办法是用**蒙特卡罗采样**来近似  $p(\delta | D)$ 。这很容易，因为在后验中  $\theta_1$  和  $\theta_2$  独立，且都服从 Beta 分布，可以用标准方法采样。

$p(\theta_i | D_i)$  如图 5.5(a) 所示；而  $p(\delta | D)$  的 MC 近似及其 95% HPD（最高后验密度区间）见图 5.5(b)。

$p(\delta > 0 | D)$  的 MC 近似通过统计样本中  $\theta_1 > \theta_2$  的比例得到；结果为 0.718，与精确值非常接近。（见 `amazonSellerDemo` 代码。）

## 通俗解释与图解读（数学基础友好版）

### 1) 我们在比较什么？

- $\theta_1$ : 卖家 1 “好评率”的真实值（我们不知道）。
- $\theta_2$ : 卖家 2 “好评率”的真实值。
- 数据：卖家 1 共  $N_1 = 100$  次评价，其中好评  $y_1 = 90$ ；卖家 2 共  $N_2 = 2$  次评价，好评  $y_2 = 2$ 。

**关键点：**卖家 2 虽然目前是“2/2 全正”，但样本太少，不确定性巨大；卖家 1 有大量评价，不确定性小。

### 2) 为什么是 Beta 后验？

对于“成败/好坏”这类二项数据，若把好评率  $\theta$  作为参数，选**均匀先验**  $\text{Beta}(1, 1)$ ，在看到  $y$  次好评、 $N - y$  次差评后，后验是

$$\theta | D \sim \text{Beta}(y + 1, N - y + 1).$$

所以：

- 卖家 1: Beta(91, 11), 均值  $\approx 91/(91 + 11) = 0.892$ , 而且分布很“窄”(确定性高)。
- 卖家 2: Beta(3, 1), 均值  $= 3/4 = 0.75$ , 但分布很“宽”(不确定性高)。

图 5.5(a) 就画出了这两个精确后验的形状：红色那条（卖家 1）在 0.9 附近窄而高；绿色虚线（卖家 2）更扁更宽，表示可能性分散。

### 3) 我们真正想要的量： $\delta = \theta_1 - \theta_2$

- $\delta > 0$  表示卖家 1 的真实好评率更高。
- 我们要的是  $p(\delta > 0 | D)$ ——**基于数据，卖家 1 更好的概率。**

#### 两种求法

- **积分法 (式 5.11)**：在  $\theta_1 > \theta_2$  的区域，对两个后验密度的乘积积分。结果 0.710 (71%)。
- **蒙特卡罗法**：从两个 Beta 后验各采样很多次，做差得到一堆  $\delta$  值；
  - $\hat{p}(\delta > 0)$  就是这些样本里  $\delta > 0$  的比例；得到 0.718，贴近精确值。
  - 这些  $\delta$  样本还可画出  $p(\delta | D)$  的近似密度。

图 5.5(b)：黑色曲线是用\*\*核密度估计 (KDE) \*\*把  $\delta$  的直方图“抹平”得到的平滑曲线；两条蓝色竖线围住的是 **95% 中心区间**（注：正文随后一句说“95% HPD”，而图注写“95% 中心区间”，两者并非完全一致，但此图用的是“中心区间”的画法）。

### 4) 结果怎么解读？

- $p(\delta > 0 | D) \approx 0.71 \sim 0.72$ : **有大约 7 成把握**卖家 1 的真实口碑更高。
- 为什么不是“肯定选卖家 2”？因为 2 条评价太少，尽管目前“全正”，但后验分布很宽，存在不少“真实口碑并不高”的可能；这正是贝叶斯方法对不确定性建模的体现。
- 95% 的中心区间（蓝线之间）显示了  $\delta$  的不确定范围：它可能略小于 0 (卖家 2 更好一点)，也可能明显大于 0 (卖家 1 更好很多) ——但总体而言， $\delta > 0$  的概率占优。

### 5) 一句话 takeaway

- **大量数据** (卖家 1) → 更窄的后验、更可靠的判断；
- **数据稀少** (卖家 2) → 后验很宽、不要被表面“100% 好评”迷惑；
- 通过比较  $\delta$  的后验，我们能直接得到“谁更好”的概率结论，而不仅是一个点估计。

### 小词条：图里提到的技术名词

- **核密度估计 (KDE)**：把随机样本用“平滑的小钟形核”叠加起来，得到一条连续、平滑的密度曲线；比直方图更美观也更容易读。
- **中心区间 (central interval)**：把左右两端各切掉  $\alpha/2$  的概率质量，留下中间  $1 - \alpha$  的部分。

- **HPD/HDI:** 把密度最高的点优先纳入，直到累计概率达到  $1 - \alpha$ ；比中心区间更能“装进最有可能的区域”。本图的注解与正文对 HPD/CI 的说法略有出入，但核心思想如上。

如果你愿意，我可以把你给的数据跑一版**可视化小实验**：随机采样、画出  $\theta_1, \theta_2, \delta$  的密度，并计算中心区间与 HPD 的数值，这能帮助你把“样本多少→不确定性大小→决策概率”这条链条彻底吃透。



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 7.8 (a) Training error (dotted blue) and test error (solid red) for a degree 14 polynomial fit by ridge regression, plotted vs  $\log(\lambda)$ . Data was generated from noise with variance  $\sigma^2 = 4$  (training set has size  $N = 21$ ). Note: Models are ordered from complex (small regularizer) on the left to simple (large regularizer) on the right. The stars correspond to the values used to plot the functions in Figure 7.7. (b) Estimate of performance using training set. Dotted blue: 5-fold cross-validation estimate of future MSE. Solid black: negative log marginal likelihood,  $-\log p(D|\lambda)$ . Both curves have been vertically rescaled to [0,1] to make them comparable. Figure generated by linregPolyVsRegDemo.

### 5.3 Bayesian model selection

In Figure 1.18, we saw that using too high a degree polynomial results in overfitting, and using too low a degree results in underfitting. Similarly, in Figure 7.8(a), we saw that using too small a regularization parameter results in overfitting, and too large a value results in underfitting. In general, when faced with a set of models (i.e., families of parametric distributions) of different complexity, how should we choose the best one? This is

called the model selection problem. One approach is to use cross-validation to estimate the generalization error of all the candidate models, and then to pick the model that seems the best. However, this requires fitting each model K times, where K is the number of CV folds. A more efficient approach is to compute the posterior over models,

$$p(m|D) = p(D|m)p(m) \quad m \in M \quad p(m,D) \quad (5.12)$$

From this, we can easily compute the MAP model,  $\hat{m} = \operatorname{argmax}_m p(m|D)$ . This is called Bayesian model selection. If we use a uniform prior over models,  $p(m) \propto 1$ , this amounts to picking the model which maximizes

$$p(D|m) = p(D|\theta)p(\theta|m)d\theta \quad (5.13)$$

This quantity is called the marginal likelihood, the integrated likelihood, or the evidence for model m. The details on how to perform this integral will be discussed in Section 5.3.2. But first we give an intuitive interpretation of what this quantity means.

好的，我先逐句翻译，然后再用基础且详尽的方式解释，让你能完全理解这张图和相关概念。

## 逐句翻译

### 图 7.8

(a) **训练误差**（蓝色虚线）和**测试误差**（红色实线），对应 14 阶多项式的岭回归（ridge regression）拟合，横轴为  $\log(\lambda)$ 。

数据是方差  $\sigma^2 = 4$  的噪声生成的（训练集大小  $N = 21$ ）。

注意：模型从左到右的顺序是——左边复杂（正则化系数小），右边简单（正则化系数大）。

图中的星号对应图 7.7 中绘制函数时用到的  $\lambda$  值。

(b) 使用训练集来估计性能。

蓝色虚线：5 折交叉验证（5-fold cross-validation）估计的未来 MSE。

黑色实线：负对数边际似然  $(-\log p(D | \lambda))$ 。

为了方便比较，这两条曲线都被垂直缩放到  $[0, 1]$  范围内。

图由 `linregPolyVsRegDemo` 生成。

## 5.3 贝叶斯模型选择 (Bayesian model selection)

在图 1.18 中，我们看到多项式阶数太高会导致过拟合，阶数太低会导致欠拟合。

同样，在图 7.8(a) 中，我们看到正则化参数太小会导致过拟合，太大则会欠拟合。

一般来说，当面对一组不同复杂度的模型（即一组不同参数化分布族）时，我们该如何选择最优模型？

这就是**模型选择问题**。

一种方法是：

- 用交叉验证估计每个候选模型的泛化误差 (generalization error) ,
- 再选出看起来最好的模型。

但这需要对每个模型拟合  $K$  次 ( $K$  是交叉验证的折数) , 计算量大。

一种更高效的方法是：直接计算模型的后验分布：

$$p(m | D) = \frac{p(D | m)p(m)}{\sum_{m' \in M} p(D, m')} \quad (5.12)$$

从中我们可以很容易求出 **MAP 模型**：

$$\hat{m} = \arg \max p(m | D)$$

这叫做**贝叶斯模型选择** (Bayesian model selection) 。

如果我们对模型使用均匀先验  $p(m) \propto 1$ , 那么选择 MAP 模型等价于选择能最大化：

$$p(D | m) = \int p(D | \theta)p(\theta | m) d\theta \quad (5.13)$$

的模型。

这个量叫做：

- **边际似然** (marginal likelihood) ,
- **积分似然** (integrated likelihood) ,
- 或**模型  $m$  的证据** (evidence) 。

至于如何计算这个积分，我们会在 5.3.2 节详细讨论。

但在那之前，我们先给出它的直观解释。

## 基础且详尽解释

### 1. 图 7.8(a) —— 复杂度与误差的关系

- **横轴**:  $\log(\lambda)$ , 其中  $\lambda$  是岭回归的正则化系数。
  - 左边  $\lambda$  很小 → 正则化弱 → 模型很复杂 → 容易过拟合。
  - 右边  $\lambda$  很大 → 正则化强 → 模型很简单 → 容易欠拟合。
- **蓝色虚线 (train MSE)** : 训练误差，随着  $\lambda$  增大而上升，因为模型被限制得越来越简单。
- **红色实线 (test MSE)** : 测试误差，先下降 (抑制过拟合) , 后上升 (欠拟合) 。

- 最佳测试误差出现在中间某个  $\lambda$  值，这就是最佳正则化强度。
- 

## 2. 图 7.8(b) —— 两种性能评估方法

- 蓝色虚线 (CV estimate of MSE)**：用 5 折交叉验证直接估计未来测试误差，找到最优  $\lambda$ 。
  - 黑色实线 (negative log marginal likelihood)**：用贝叶斯方法，通过最大化边际似然来选  $\lambda$ 。
  - 两条曲线形状相似，说明**边际似然最大化**可以替代交叉验证进行模型选择，而且只需训练一次模型（效率高）。
- 

## 3. 模型选择问题

- 欠拟合**：模型太简单，无法捕捉数据模式。
- 过拟合**：模型太复杂，把噪声也当成模式去拟合。
- 我们希望在不同复杂度模型中找到“刚好”的那个。

### 方法 1：交叉验证

- 思路：反复划分数据 → 训练 → 测试 → 平均误差。
- 缺点：计算量大，因为每个候选模型要训练  $K$  次。

### 方法 2：贝叶斯模型选择

- 思路：直接计算每个模型的**后验概率**  $p(m | D)$ ，选概率最大的那个。
  - 如果对模型均匀先验 → 等价于比较它们的**边际似然**  $p(D | m)$ 。
  - 优点：一次训练就能比较模型，效率高。
- 

## 4. 边际似然的意义

- 边际似然是“模型整体对数据的解释能力”，综合考虑：
    - 拟合度** (likelihood 大 → 数据符合模型的预测)
    - 复杂度惩罚** (integral 自动惩罚参数太多的模型)
  - 因为积分把所有可能的参数值都考虑了，而不是只看最优参数。
- 

如果你愿意，我可以帮你画一个**交叉验证 vs 边际似然模型选择**的流程对比图，让你一眼就看懂两者的计算流程和差异。

这样你能更直观地理解为什么贝叶斯方法在图 7.8(b) 中表现接近交叉验证，但计算量更小。

你要我画吗？



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 5.6 A schematic illustration of the Bayesian Occam's razor. The broad (green) curve corresponds to a complex model, the narrow (blue) curve to a simple model, and the middle (red) curve is just right. Based on Figure 3.13 of (Bishop 2006a). See also (Murray and Ghahramani 2005, Figure 2) for a similar plot produced on real data.

Figure 5.7 (a-c) We plot polynomials of degrees 1, 2 and 3 fit to  $N = 5$  data points using empirical Bayes. The solid green curve is the true function, the dashed red curve is the prediction (dotted blue lines represent  $\pm \sigma$  around the mean). (d) We plot the posterior over models,  $p(d|D)$ , assuming a uniform prior  $p(d) \propto 1$ . Based on a figure by Zoubin Ghahramani. Figure generated by linregEbModelSelVsN.

Figure 5.8 Same as Figure 5.7 except now  $N = 30$ . Figure generated by linregEbModelSelVsN.

### 5.3.1 Bayesian Occam's razor

One might think that using  $p(D|m)$  to select models would always favor the model with the most parameters. This is true if we use  $p(D|\theta_m)$  to select models, where  $\theta_m$  is the MLE or MAP estimate of the parameters for model  $m$ , because models with more parameters will fit the data better, and hence achieve higher likelihood. However, if we integrate out the parameters, rather than maximizing them, we are automatically protected from overfitting: models with more parameters do not necessarily have higher marginal likelihood. This is called the Bayesian Occam's razor effect (MacKay 1995b; Murray and Ghahramani 2005), named after the principle known as Occam's razor, which says one should pick the simplest model that adequately explains the data. One way to understand the Bayesian Occam's razor is to notice that the marginal likelihood can be rewritten as follows, based

on the chain rule of probability (Equation 2.5):

$$p(D) = p(y_1)p(y_2|y_1)p(y_3|y_1:y_2)\dots p(y_N|y_1:N-1) \quad (5.14)$$

where we have dropped the conditioning on  $x$  for brevity. This is similar to a leave-one-out cross-validation estimate (Section 1.4.8) of the likelihood, since we predict each future point given all the previous ones. (Of course, the order of the data does not matter in the above expression.) If a model is too complex, it will overfit the “early” examples and will then predict the remaining ones poorly. Another way to understand the Bayesian Occam’s razor effect is to note that probabilities must sum to one. Hence  $\sum_m p(D|m) = 1$ , where the sum is over all possible data sets. Complex models, which can predict many things, must spread their probability mass thinly, and hence will not obtain as large a probability for any given data set as simpler models. This is sometimes called the conservation of probability mass principle, and is illustrated in Figure 5.6. On the horizontal axis we plot all possible data sets in order of increasing complexity (measured in some abstract sense). On the vertical axis we plot the predictions of 3 possible models: a simple one, M1; a medium one, M2; and a complex one, M3. We also indicate the actually observed data D0 by a vertical line. Model 1 is too simple and assigns low probability to D0. Model 3 also assigns D0 relatively low probability, because it can predict many data sets, and hence it spreads its probability quite widely and thinly. Model 2 is “just right”: it predicts the observed data with a reasonable degree of confidence, but does not predict too many other things. Hence model 2 is the most probable model.

As a concrete example of the Bayesian Occam’s razor, consider the data in Figure 5.7. We plot polynomials of degrees 1, 2 and 3 fit to  $N=5$  data points. It also shows the posterior over models, where we use a Gaussian prior (see Section 7.6 for details). There is not enough data to justify a complex model, so the MAP model is  $d=1$ . Figure 5.8 shows what happens when  $N=30$ . Now it is clear that  $d=2$  is the right model (the data was in fact generated from a quadratic).

As another example, Figure 7.8(c) plots  $\log p(D|\lambda)$  vs  $\log(\lambda)$ , for the polynomial ridge regression model, where  $\lambda$  ranges over the same set of values used in the CV experiment. We see that the maximum evidence occurs at roughly the same point as the minimum of the test MSE, which also corresponds to the point chosen by CV.

When using the Bayesian approach, we are not restricted

to evaluating the evidence at a finite grid of values. Instead, we can use numerical optimization to find  $\lambda^*$  = $\operatorname{argmax}_\lambda p(D|\lambda)$ . This technique is called empirical Bayes or type II maximum likelihood (see Section 5.6 for details). An example is shown in Figure 7.8(b): we see that the curve has a similar shape to the CV estimate, but it can be computed more efficiently.

已思考 10s

好的，我先逐句完整翻译，再给出基础且详尽的解释，并结合三张图说明“贝叶斯奥卡姆剃刀”的直觉与做法。

---

## 逐句翻译（无遗漏）

### 图 5.6

贝叶斯奥卡姆剃刀的示意图。宽的（绿色）曲线对应一个**复杂模型**，窄的（蓝色）曲线对应一个**简单模型**，而中间的（红色）曲线是**刚刚好的模型**。改编自 (Bishop 2006a) 的图 3.13。参见 (Murray and Ghahramani 2005, 图 2)，在那里用真实数据画出了类似的图。

### 图 5.7

(a–c) 我们对  $N = 5$  个数据点，用经验贝叶斯 (empirical Bayes) 分别拟合 1、2、3 次多项式。实线绿色曲线是真实函数；虚线红色曲线是模型的预测（点状蓝线表示**均值上下  $\pm \sigma$** ）。

(d) 我们绘制了模型后验  $p(d | D)$ （假设模型度数的先验为均匀  $p(d) \propto 1$ ）。基于 Zoubin Ghahramani 的一幅图。图由 `linregEbModelSelVsN` 生成。

### 图 5.8

与图 5.7 相同，但现在  $N = 30$ 。图由 `linregEbModelSelVsN` 生成。

---

### 5.3.1 贝叶斯奥卡姆剃刀

有人可能会认为，用  $p(D | m)$  来选择模型，总会偏向**参数最多的模型**。如果我们用  $p(D | \hat{\theta}_m)$  来选 ( $\hat{\theta}_m$  是模型  $m$  的 MLE 或 MAP)，这确实会发生，因为参数越多的模型往往**拟合得更好**，从而获得更高的似然。

然而，**如果我们把参数积分掉**（而不是把它们最大化），就会**自动避免过拟合**：参数更多的模型不一定有更高的**边际似然** (marginal likelihood)。这被称为**贝叶斯奥卡姆剃刀效应** (MacKay 1995b; Murray and Ghahramani 2005)。这个名字来自“奥卡姆剃刀”原则：在能充分解释数据的前提下，应该选择**最简单的模型**。

理解贝叶斯奥卡姆剃刀的一种方法，是注意到根据概率的链式法则（式 2.5），边际似然可以重写为

$$p(D) = p(y_1)p(y_2 | y_1)p(y_3 | y_{1:2}) \cdots p(y_N | y_{1:N-1}) \quad (5.14)$$

这里为简洁起见省略了对  $x$  的条件。这与“留一交叉验证”（第 1.4.8 节）对似然的估计很相似，因为我们在给定之前所有点的条件下预测每一个后续点。（当然，上式中数据的顺序并不影响结果。）

如果一个模型过于复杂，它会对“前面的”数据过拟合，从而**较差地**预测剩余的数据。

理解贝叶斯奥卡姆剃刀的另一种方法是：概率必须求和为 1。因而  $\sum_D p(D | m) = 1$ ，这里的求和在所有可能的数据集上进行。**复杂模型**能预测很多种情况，就必须把它的概率质量摊得很薄，于是对任何**特定**数据集所分配到的概率都不会太大；相反，**简单模型**只预测少数情况，因此能在这些情况上分配到更大的概率。这有时称为**概率质量守恒原则**，如图 5.6 所示。横轴按某种抽象的“复杂度”把所有可能的数据集排序，纵轴画出三种模型的预测：简单的  $M_1$ 、中等的  $M_2$ 、复杂的  $M_3$ 。竖直虚线表示真实观测到的数据  $D_0$ 。模型 1 太简单，对  $D_0$  给的概率低；模型 3 也给  $D_0$  相对较低的概率，因为它把概率铺得很开，能预测太多东西；模型 2“刚刚好”：它对观测数据有相当的置信度，但又不会预测太多其他情形。因此模型 2 是最可能的模型。

作为贝叶斯奥卡姆剃刀的一个具体示例，考虑图 5.7 的数据。我们对  $N = 5$  个点拟合 1、2、3 次多项式，并给出模型后验（先验取高斯分布，细节见 7.6 节）。数据量不足以支撑复杂模型，因此 MAP 模型为  $d = 1$ 。图 5.8 显示了  $N = 30$  时会发生什么：这时很清楚  $d = 2$  是正确的模型（数据实际上由一个二次函数生成）。

另一个例子：图 7.8(c) 绘制了多项式岭回归模型的  $\log p(D | \lambda)$  随  $\log \lambda$  的变化，其中  $\lambda$  取与 CV 实验相同的一组值。可以看到，**证据最大的点大致落在测试 MSE 最小的位置**，也与 CV 选择的点一致。

采用贝叶斯方法时，我们不必局限于在一组离散的  $\lambda$  上评估证据；我们可以用数值优化直接找到

$$\lambda^* = \arg \max_{\lambda} p(D | \lambda).$$

这种技术称为**经验贝叶斯** (empirical Bayes) 或**II 型极大似然** (type II maximum likelihood，见 5.6 节)。图 7.8(b) 给出了一个例子：可以看到这条曲线与 CV 的估计形状相似，但计算更高效。

## 基础且详尽解释

### 1) “边际似然”为何自动惩罚复杂模型？

- **最大化似然**  $p(D | \theta)$ ：只看\*\*最会“取巧”\*\*的那组参数，参数越多越容易把噪声也拟合上 → 偏向复杂模型。
- **积分**  $p(D | m) = \int p(D | \theta)p(\theta | m) d\theta$ ：把**所有可能参数**按先验权重平均，
  - 若模型太灵活，只有很窄的一小片参数能很好地解释数据，其它大部分参数解释得并不好，被“平均”后证据就被稀释了；
  - 若模型太简单，无论怎样平均都解释不好数据，证据也低；

- **恰当复杂度**的模型在相当一片参数区域都能合理解释数据 → 平均后证据最高。  
这就是“奥卡姆剃刀”的贝叶斯版：在能解释数据的前提下，越简单越好，但要刚  
刚好（图 5.6 中红色曲线）。

## 2) 与“留一交叉验证”的联系 (式 5.14)

把  $p(D)$  写成逐点的条件概率连乘，相当于“用前面的数据预测下一个点”的连锁过程。

- 过度复杂的模型会过拟合前面的小样本，导致后面点的条件概率变小，连乘后得到的  $p(D)$  也小；
- 合理的模型对“未来点”的预测较稳健，连乘起来更大。  
因此，最大证据与**泛化能力一致**，常常与 CV 的最优点吻合（文中也用图 7.8 说明了这一点）。

## 3) 看懂三组图

### 图 5.6 (示意图)

- 横轴：可能的数据集按复杂度排序；纵轴：各模型对这些数据集赋予的概率。
- 简单模型  $M_1$ ：只“相信”很少的简单数据 → 对  $D_0$  概率低。
- 复杂模型  $M_3$ ：什么都能解释一点 → 概率铺得很薄 → 对  $D_0$  也不高。
- 中等模型  $M_2$ ：对  $D_0$  概率最高 → **被选中**。

### 图 5.7 ( $N = 5$ )

- (a)(b)(c)：用经验贝叶斯分别拟合 1、2、3 次多项式：
  - 绿色实线是真实函数；红色虚线为预测均值；蓝色点线是均值  $\pm \sigma$ 。
  - 数据太少，二次/三次的灵活性用不上，证据反而低。
- (d)：模型后验  $p(d | D)$  的柱状图（先验均匀）。最高的是  $d = 1$ ：**MAP 模型是一阶**。

### 图 5.8 ( $N = 30$ )

- 数据足够了，二次模型能显著更好地解释数据：
- (d) 可见  $d = 2$  的后验概率几乎为 1；而  $d = 1$  与  $d = 3$  的证据都低。
- 这体现了：**数据越多，越能支持恰当的复杂度**（这里真实数据来自二次函数）。

## 4) 经验贝叶斯 (EB) 如何选 $\lambda$ ？

- 不是网格穷举，而是直接优化证据  $p(D | \lambda)$  找到  $\lambda^*$ 。
- 通常  $\lambda^*$  与交叉验证选到的  $\lambda$  很接近，但 EB 只需要一次训练/一次优化，**更高效**。
- 小提示：EB 对超参数（如  $\lambda$ ）做点估计，不是完全贝叶斯；但在实践中效果和效率都很好。

## 一句话总结

- **最大化证据** = “既考虑拟合度，又自动惩罚不必要的复杂度”。
- 数据少时选简单模型（图 5.7）；数据多时选真正的生成模型（图 5.8）。

- 在正则化模型里，用\*\*证据最大化（经验贝叶斯）\*\*选超参数，往往与 CV 的最佳点一致，但计算更省。

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。

### 5.3.2 Computing the marginal likelihood (evidence)

When discussing parameter inference for a fixed model, we often wrote

$$p(\theta|D,m) \propto p(\theta|m)p(D|\theta,m) \quad (5.15)$$

thus ignoring the normalization constant  $p(D|m)$ . This is valid since  $p(D|m)$  is constant wrt  $\theta$ . However, when comparing models, we need to know how to compute the marginal likelihood,  $p(D|m)$ . In general, this can be quite hard, since we have to integrate over all possible parameter values, but when we have a conjugate prior, it is easy to compute, as we now show.

Let  $p(\theta) = q(\theta)/Z_0$  be our prior, where  $q(\theta)$  is an unnormalized distribution, and  $Z_0$  is the normalization constant of the prior. Let  $p(D|\theta) = q(D|\theta)/Z$  be the likelihood, where  $Z$  contains any constant factors in the likelihood. Finally let  $p(\theta|D) = q(\theta|D)/Z_N$  be our posterior, where  $q(\theta|D) = q(D|\theta)q(\theta)$  is the unnormalized posterior, and  $Z_N$  is the normalization constant of the posterior. We have

$$p(\theta|D) = p(D|\theta)p(\theta) / p(D) \quad (5.16)$$

$$q(\theta|D) Z_N = q(D|\theta)q(\theta) / Z_0 p(D) \quad (5.17)$$

$$p(D) = Z_N Z_0 Z \quad (5.18)$$

So assuming the relevant normalization constants are tractable, we have an easy way to compute the marginal likelihood. We give some examples below.

好的，我会先完整翻译这段内容，然后再给你做一个基础且详尽的解释，并且会用通俗语言帮助你理解公式背后的逻辑。

## 逐句完整翻译（无遗漏）

### 5.3.2 计算边际似然（证据）

当我们讨论已知模型下的参数推断时，我们经常写成

$$p(\theta | D, m) \propto p(\theta | m)p(D | \theta, m) \quad (5.15)$$

因此忽略了归一化常数  $p(D \mid m)$ 。

这是合理的，因为  $p(D \mid m)$  对  $\theta$  来说是常数。

然而，当我们**比较不同模型**时，我们就需要知道如何计算这个**边际似然**  $p(D \mid m)$ 。

一般来说，这可能很难，因为我们必须对**所有可能的参数值**进行积分；

但是，当我们有\*\*共轭先验 (conjugate prior) \*\*时，这个计算就很容易，下面我们将展示。

设

$$p(\theta) = \frac{q(\theta)}{Z_0}$$

是我们的先验分布，其中  $q(\theta)$  是\*\*未归一化 (unnormalized)\*\* 的分布， $Z_0$  是先验的归一化常数。

设

$$p(D \mid \theta) = \frac{q(D \mid \theta)}{Z}$$

是似然函数，其中  $Z$  包含似然中的所有常数因子。

最后，设

$$p(\theta \mid D) = \frac{q(\theta \mid D)}{Z_N}$$

是我们的后验分布，其中

$$q(\theta \mid D) = q(D \mid \theta)q(\theta)$$

是未归一化的后验， $Z_N$  是后验的归一化常数。

我们有：

$$p(\theta \mid D) = \frac{p(D \mid \theta)p(\theta)}{p(D)} \quad (5.16)$$

$$\frac{q(\theta \mid D)}{Z_N} = \frac{q(D \mid \theta)q(\theta)}{Z Z_0 p(D)} \quad (5.17)$$

$$p(D) = \frac{Z_N}{Z_0 Z} \quad (5.18)$$

因此，如果这些相关的归一化常数是**可求的**，我们就有一种简单的方法来计算**边际似然**。

下面我们会给出一些例子。

## 基础且详尽的解释

## 1) 边际似然是什么?

- 边际似然 (marginal likelihood) , 又叫**模型证据 (evidence)** , 记作:

$$p(D \mid m) = \int p(D \mid \theta, m) p(\theta \mid m) d\theta$$

- 它的意思是: “在给定模型  $m$  下, 所有可能的参数值按照先验的权重加权平均后, 这个模型生成数据  $D$  的概率”。
  - 在做**参数推断** (固定模型) 时, 我们关心的是  $p(\theta \mid D, m)$ , 归一化常数  $p(D \mid m)$  对  $\theta$  来说是常数, 可以忽略。
  - 但在**模型比较**时, 必须计算  $p(D \mid m)$ , 因为它在不同模型间会变 (这就是用它做模型选择的依据) 。
- 

## 2) 文中符号含义

- $q(\cdot)$  → “未归一化”的概率函数 (即没有除以归一化常数的版本) 。
- $Z_0$  → 先验的归一化常数 (让  $p(\theta)$  的积分为 1) 。
- $Z$  → 似然的归一化常数 (例如在某些分布中会有常数项) 。
- $Z_N$  → 后验的归一化常数 (确保  $p(\theta \mid D)$  积分为 1) 。

这三者的关系是:

$$p(D) = \frac{\text{后验的归一化常数}}{\text{先验常数} \times \text{似然常数}}$$

即:

$$p(D) = \frac{Z_N}{Z_0 Z}$$


---

## 3) 为什么共轭先验能让计算变简单?

- 共轭先验**的好处是: 先验  $\times$  似然  $\rightarrow$  后验的形式仍在同一个分布族里。
  - 这样, 先验常数  $Z_0$ 、似然常数  $Z$ 、后验常数  $Z_N$  都能用已知公式直接写出来。
  - 所以只要把它们代入  $p(D) = Z_N / (Z_0 Z)$ , 就能算出边际似然, 不需要做复杂积分。
- 

## 4) 直观类比

你可以把它想成:

- 先验常数  $Z_0$** : 先验这个“调味料”的总量。
- 似然常数  $Z$** : 模型对数据的“基本度量单位”。
- 后验常数  $Z_N$** : 结合数据之后调出的“新配方”的总量。
- 边际似然  $p(D)$  就是“结合后得到的新配方”相对于“原料和单位”的比例。

如果这些常数可以用公式求出，就能很容易得到  $p(D)$ 。

## 5) 小结

- 在参数推断时，边际似然是可以忽略的常数；
- 在模型选择时，它是核心指标；
- 公式 (5.18) 提供了一个在有共轭先验时非常简洁的计算方法；
- 共轭先验的优势就是归一化常数容易求，积分不再麻烦。

如果你愿意，我可以帮你画一张示意图，展示  $Z_0, Z, Z_N$  如何通过“先验  $\times$  似然  $\rightarrow$  后验”联系起来，并且如何得到边际似然  $p(D)$ 。这样会比公式更直观。

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。

### 5.3.2.1 Beta-binomial model

Let us apply the above result to the Beta-binomial model.

Since we know  $p(\theta|D) = \text{Beta}(\theta|a, b)$ , where  $a = a + N_1$  and  $b = b + N_0$ , we know the normalization constant of the posterior is  $B(a, b)$ . Hence

$$\begin{aligned} p(\theta|D) &= p(D|\theta)p(\theta)/p(D) \quad (5.19) \\ &= 1/p(D) \int B(a, b) \theta^{a-1} (1-\theta)^{b-1} N(N_1 \theta + N_0 (1-\theta))^{N_0} \quad (5.20) \end{aligned}$$

$$= N(N_1 + N_0) \int B(a, b) \theta^{a-1} (1-\theta)^{b-1} \theta^{N_1} (1-\theta)^{N_0} \quad (5.21)$$

So

$$1/B(a + N_1, b + N_0) = N(N_1 + N_0) \int B(a, b) \quad (5.22)$$

$$p(D) = N(N_1 + N_0) B(a + N_1, b + N_0) / B(a, b) \quad (5.23)$$

The marginal likelihood for the Beta-Bernoulli model is the same as above, except it is missing the  $N(N_1)$  term.

### 5.3.2.2 Dirichlet-multinoulli model

By the same reasoning as the Beta-Bernoulli case, one can show that the marginal likelihood for the Dirichlet-multinoulli model is given by

$$p(D) = B(N + \alpha) / B(\alpha) \quad (5.24)$$

where

$$B(\alpha) = \prod_{k=1}^K \Gamma(\alpha_k) / \Gamma(\sum \alpha_k) \quad (5.25)$$

Hence we can rewrite the above result in the following form, which is what is usually presented in the literature:

$$p(D) = \Gamma(\sum \alpha_k) / \Gamma(N + \sum \alpha_k) \prod_{k=1}^K \Gamma(N_k + \alpha_k) / \Gamma(\alpha_k) \quad (5.26)$$

We will see many applications of this equation later.

### 5.3.2.3 Gaussian-Gaussian-Wishart model

Consider the case of an MVN with a conjugate NIW prior.

Let  $Z_0$  be the normalizer for the prior,  $Z_N$  be normalizer for the posterior, and let  $Z_l = (2\pi)^{ND/2}$  be the normalizer for the likelihood. Then it is easy to see that

$$p(D) = Z_N Z_0 Z_l \quad (5.27)$$

$$= 1/\pi^{ND/2} 1/2^{ND/2} + 2\pi^{-N/2} |S_N|$$

$$-vN/2(v_0+N)D/2\Gamma(vN/2) + 2\pi^{-N/2} |S_0|$$

$$-v_0/2v_0D/2\Gamma(v_0/2) \quad (5.28)$$

$$= 1/\pi^{ND/2} \kappa_0^{-1} \kappa_N^{-1} D/2 |S_0| v_0/2 |S_N| vN/2 \Gamma(vN/2)$$

$$\Gamma(v_0/2) \quad (5.29)$$

This equation will prove useful later.

好的，我们一步步来，我会先**完整翻译**，然后用**通俗语言做详细解释**，并且尽量帮你绕开太多高深的数学，让你能抓住核心。

## 逐句完整翻译（无遗漏）

### 5.3.2.1 Beta-二项式模型

让我们把上面的方法应用到 Beta-二项式 (Beta-binomial) 模型。

因为我们知道

$$p(\theta | D) = \text{Beta}(\theta | a', b')$$

其中

$$a' = a + N_1, \quad b' = b + N_0$$

所以我们知道**后验分布的归一化常数是**

$$B(a', b')$$

因此：

$$p(\theta | D) = \frac{p(D | \theta) p(\theta)}{p(D)} \quad (5.19)$$

$$= \frac{1}{p(D)} \cdot \frac{1}{B(a, b)} \theta^{a-1} (1-\theta)^{b-1} \cdot \binom{N}{N_1} \theta^{N_1} (1-\theta)^{N_0} \quad (5.20)$$

$$= \binom{N}{N_1} \cdot \frac{1}{p(D)} \cdot \frac{1}{B(a, b)} \theta^{a+N_1-1} (1-\theta)^{b+N_0-1} \quad (5.21)$$

所以：

$$\frac{1}{B(a+N_1, b+N_0)} = \binom{N}{N_1} \cdot \frac{1}{p(D)} \cdot \frac{1}{B(a, b)} \quad (5.22)$$

$$p(D) = \binom{N}{N_1} \cdot \frac{B(a + N_1, b + N_0)}{B(a, b)} \quad (5.23)$$

对于 Beta-Bernoulli 模型，边际似然的形式与上式相同，只是缺少了组合项  $\binom{N}{N_1}$ 。

---

### 5.3.2.2 Dirichlet-多项式 (multinoulli) 模型

用与 Beta-Bernoulli 情况相同的推理，可以得到 Dirichlet-多项式模型的边际似然为：

$$p(D) = \frac{B(N + \alpha)}{B(\alpha)} \quad (5.24)$$

其中：

$$B(\alpha) = \frac{\prod_{k=1}^K \Gamma(\alpha_k)}{\Gamma(\sum_k \alpha_k)} \quad (5.25)$$

因此，我们可以把上式改写成文献中常见的形式：

$$p(D) = \frac{\Gamma(\sum_k \alpha_k)}{\Gamma(N + \sum_k \alpha_k)} \cdot \frac{\prod_k \Gamma(N_k + \alpha_k)}{\prod_k \Gamma(\alpha_k)} \quad (5.26)$$

我们将在后面看到这个公式的许多应用。

---

### 5.3.2.3 高斯-高斯-Wishart 模型

考虑一个多元高斯分布 (MVN) 结合共轭的 NIW (Normal-Inverse-Wishart) 先验的情况。

设  $Z_0$  是先验的归一化常数， $Z_N$  是后验的归一化常数， $Z_l = (2\pi)^{ND/2}$  是似然的归一化常数。

那么很容易得到：

$$p(D) = \frac{Z_N}{Z_0 Z_l} \quad (5.27)$$

$$= \frac{1}{\pi^{ND/2}} \cdot \frac{1}{2^{ND/2}} \cdot \frac{(2\pi)^{\kappa_N, D/2}}{|S_N|^{\nu_N/2}} \cdot \frac{\Gamma_D(\nu_N/2)}{(2\pi)^{\kappa_0, D/2} |S_0|^{\nu_0/2} \Gamma_D(\nu_0/2)} \quad (5.28)$$

$$= \frac{1}{\pi^{ND/2}} \cdot \left( \frac{\kappa_0}{\kappa_N} \right)^{D/2} \cdot \frac{|S_0|^{\nu_0/2}}{|S_N|^{\nu_N/2}} \cdot \frac{\Gamma_D(\nu_N/2)}{\Gamma_D(\nu_0/2)} \quad (5.29)$$

这个公式在后面会非常有用。

---

## 详细解释 (基础版)

### 1. 这些模型的共同点

这三个例子 (Beta-binomial、Dirichlet-multinoulli、Gaussian-Gaussian-Wishart) 都是  
\*\*似然 + 共轭先验\*\*的组合。

- 因为用了**共轭先验**, 后验的形式是已知的;
  - 后验的归一化常数  $Z_N$ 、先验常数  $Z_0$ 、似然常数  $Z_l$  都有公式;
  - 所以边际似然  $p(D)$  可以直接用比值公式算出来。
- 

## 2. Beta-binomial 模型

- **场景:**  $\theta$  表示硬币正面朝上的概率;  $N_1$  是观测到的正面次数,  $N_0$  是反面次数;
- **先验:**  $\theta \sim \text{Beta}(a, b)$
- **似然:** 二项分布  $\binom{N}{N_1} \theta^{N_1} (1 - \theta)^{N_0}$
- **后验:**  $\theta | D \sim \text{Beta}(a + N_1, b + N_0)$

利用 Beta 分布的归一化常数  $B(\cdot, \cdot)$  直接写出  $p(D)$ 。

如果是 Beta-Bernoulli (一次试验的情况), 只是去掉了组合项  $\binom{N}{N_1}$ 。

---

## 3. Dirichlet-multinoulli 模型

- **场景:** 多分类问题, 比如  $K$  种颜色的球;
- **先验:** 类别概率向量  $\theta \sim \text{Dirichlet}(\alpha)$
- **似然:** 多项分布 (multinomial)
- **后验:**  $\theta | D \sim \text{Dirichlet}(\alpha + N)$

因为 Dirichlet 的归一化常数用多元 Beta 函数  $B(\alpha)$  表示, 所以  $p(D)$  的公式可以直接写出来。

---

## 4. Gaussian-Gaussian-Wishart 模型

- **场景:** 数据服从多元高斯分布 (MVN), 但均值和协方差未知;
  - **先验:** Normal-Inverse-Wishart (NIW) 分布 (多元高斯的均值和协方差的共轭先验)
  - **结论:** 边际似然是归一化常数的比值, 涉及:
    - $\kappa_0, \kappa_N$  (先验和后验的均值精度参数)
    - $|S_0|, |S_N|$  (先验和后验的协方差矩阵行列式)
    - $v_0, v_N$  (自由度)
    - 多元 Gamma 函数  $\Gamma_D$
- 

## 5. 总结思路

不管是哪种模型:

1. 写出先验 (带  $Z_0$ )
2. 写出似然 (带  $Z_l$ )
3. 写出后验 (带  $Z_N$ )
4. 用公式

$$p(D) = \frac{Z_N}{Z_0 Z_l}$$

直接算出边际似然。

---

如果你愿意，我可以帮你画一个**三种模型的“先验-似然-后验-边际似然”关系图**，这样你能一眼看出它们的结构和公式来源，比看公式容易很多。  
你要我画这个图吗？这样你会更快记住这些公式的来源。

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。

#### 5.3.2.4 BIC approximation to log marginal likelihood

In general, computing the integral in Equation 5.13 can be quite difficult. One simple but popular approximation is known as the Bayesian information criterion or BIC, which has the following form (Schwarz 1978):

$$\text{BIC} = \log p(D|\hat{\theta}) - \text{dof}(\hat{\theta}) 2 \log N \approx \log p(D) \quad (5.30)$$

where  $\text{dof}(\hat{\theta})$  is the number of degrees of freedom in the model, and  $\hat{\theta}$  is the MLE for the model.<sup>2</sup> We see that this has the form of a penalized log likelihood, where the penalty term depends on the model's complexity. See Section 8.4.2 for the derivation of the BIC score. As an example, consider linear regression. As we show in

Section 7.3, the MLE is given by  $\hat{w} = (X^T X)^{-1} X^T Y$  and  $\hat{\sigma}^2 = \text{RSS}/N$ , where  $\text{RSS} = \sum_{i=1}^N (y_i - \hat{w}^T m_i)^2$ . The corresponding log likelihood is given by

$$\log p(D|\hat{\theta}) = -\frac{N}{2} \log(2\pi\hat{\sigma}^2) - \frac{N}{2} \quad (5.31)$$

Hence the BIC score is as follows (dropping constant terms)

$$\text{BIC} = -\frac{N}{2} \log(\hat{\sigma}^2) - D \frac{N}{2} \log(N) \quad (5.32)$$

where  $D$  is the number of variables in the model. In the statistics literature, it is common to use an alternative definition of BIC, which we call the BIC cost (since we want to minimize it):

$$\text{BIC-cost} = -2 \log p(D|\hat{\theta}) + \text{dof}(\hat{\theta}) \log N \approx -2 \log p(D) \quad (5.33)$$

In the context of linear regression, this becomes

$$\text{BIC-cost} = N \log(\hat{\sigma}^2) + D \log(N) \quad (5.34)$$

The BIC method is very closely related to the minimum description length or MDL principle, which characterizes the score for a model in terms of how well it fits the data, minus how complex the model is to define. See (Hansen and Yu 2001) for details. There is a very similar expression to BIC/ MDL called the Akaike information criterion or AIC, defined as

$$\text{AIC}(m, D) = \log p(D | \hat{\theta}_{\text{MLE}}) - \text{dof}(m) \quad (5.35)$$

This is derived from a frequentist framework, and cannot be interpreted as an approximation to the marginal likelihood. Nevertheless, the form of this expression is very similar to BIC. We see that the penalty for AIC is less than for BIC. This causes AIC to pick more complex models. However, this can result in better predictive accuracy. See e.g., (Clarke et al. 2009, sec 10.2) for further discussion on such information criteria.

好，我们先逐句无遗漏翻译，然后再用基础且详细的方式解释，尽量用生活化的比喻帮你理解 BIC、AIC 这些东西的意义。

## 逐句翻译

### 5.3.2.4 BIC 对数边际似然的近似

在一般情况下，直接计算公式 (5.13) 中的积分是非常困难的。

一个简单但很流行的近似方法叫做**贝叶斯信息准则** (Bayesian Information Criterion, BIC) ，它的形式如下 (Schwarz 1978) :

$$\text{BIC} = \log p(D | \hat{\theta}) - \frac{\text{dof}(\hat{\theta})}{2} \log N \approx \log p(D) \quad (5.30)$$

其中：

- $\text{dof}(\hat{\theta})$  表示模型的自由度 (参数个数)
- $\hat{\theta}$  是该模型的最大似然估计 (MLE)

我们可以看到，这个公式的结构是一个**带惩罚项的对数似然**，其中惩罚项与模型的复杂度有关。

BIC 公式的推导见第 8.4.2 节。

**举例：**考虑线性回归。

如第 7.3 节所示，MLE 给出：

$$\hat{w} = (X^T X)^{-1} X^T y$$

以及：

$$\hat{\theta}^2 = \frac{\text{RSS}}{N}$$

其中：

$$\text{RSS} = \sum_{i=1}^N (y_i - \hat{w}_{\text{mle}}^T x_i)^2$$

对应的对数似然为：

$$\log p(D | \hat{\theta}) = -\frac{N}{2} \log(2\pi\hat{\theta}^2) - \frac{N}{2} \quad (5.31)$$

因此，BIC 分数为（去掉常数项）：

$$\text{BIC} = -\frac{N}{2} \log(\hat{\theta}^2) - \frac{D}{2} \log(N) \quad (5.32)$$

其中  $D$  是模型中的变量个数。

在统计学文献中，人们常常使用 BIC 的另一种定义，我们称它为 **BIC 成本**（因为我们希望最小化它）：

$$\text{BIC-cost} = -2 \log p(D | \hat{\theta}) + \text{dof}(\hat{\theta}) \log N \approx -2 \log p(D) \quad (5.33)$$

**在线性回归**的情况下，这变为：

$$\text{BIC-cost} = N \log(\hat{\theta}^2) + D \log(N) \quad (5.34)$$

**BIC 方法与最小描述长度原则** (MDL, Minimum Description Length) 密切相关，MDL 用模型对数据的拟合程度减去描述该模型的复杂度来评估模型的好坏。详情见 (Hansen 和 Yu 2001)。

还有一个与 BIC/MDL 很相似的指标，叫做**赤池信息准则** (Akaike Information Criterion, AIC)，定义为：

$$\text{AIC}(m, D) = \log p(D | \hat{\theta}_{\text{MLE}}) - \text{dof}(m) \quad (5.35)$$

AIC 来源于频率学派的推导，不能被解释为边际似然的近似值。

尽管如此，AIC 的形式与 BIC 非常相似。我们可以看到，AIC 的惩罚项比 BIC 更小，这会导致 AIC 更倾向于选择**更复杂的模型**。

然而，这种选择可能会带来更好的预测精度。更多关于这些信息准则的讨论可见 (Clarke 等人 2009, 第 10.2 节)。

## 详细解释（基础版）

我们先用**生活化比喻**帮你理解：

想象你要请人来修理电脑，候选有两类人：

- A：工具少（模型简单），出错少，但可能修不好复杂问题；
- B：工具多（模型复杂），能修更多问题，但有可能“过度维修”（过拟合）

BIC 和 AIC 就是用来帮你选择是请 A 还是 B 的规则。

---

### 1. BIC 的本质

- 第一项  $\log p(D | \hat{\theta})$ ：模型对数据的拟合程度（越大越好）
- 第二项  $-\frac{\text{dof}}{2} \log N$ ：惩罚模型复杂度（参数越多，惩罚越大）
- 这种惩罚是随样本量  $N$  增长而增加的，所以数据越多，BIC 对复杂模型越严格。

换句话说：

- BIC 更“保守”，它会更倾向于选择简单模型，除非复杂模型的拟合好得多。
- 

### 2. BIC 成本

BIC 成本只是 BIC 的另一种写法（符号不同），常用于**最小化**的场景（越小越好）。

---

### 3. AIC

- AIC 和 BIC 很像，但惩罚项小：
    - AIC 惩罚： $-\text{dof}(m)$
    - BIC 惩罚： $-\frac{\text{dof}(m)}{2} \log N$ （通常更大）
  - 结果：
    - AIC 更容易选复杂模型（因为惩罚小）
    - 复杂模型有时预测更准（尤其样本不多时）
- 

### 4. MDL 原则

- MDL 认为：一个好的模型应该让“数据 + 模型的描述”的总长度最短
  - 如果模型太复杂，描述它本身的“文字”就会很长
  - 如果模型太简单，解释数据时需要额外很多“文字”去补充误差
  - BIC 的思路和这个几乎一样
- 

### 5. 公式记忆窍门

以线性回归为例：

$$\text{BIC} = -\frac{N}{2} \log(\hat{\sigma}^2) - \frac{D}{2} \log N$$

- 第一项：拟合误差 ( $\hat{\sigma}^2$  越小越好)
- 第二项：惩罚项 (D 越多惩罚越大)

如果你愿意，我可以帮你画一个 **BIC vs AIC 对比图**，用直观的曲线展示为什么 BIC 更倾向简单模型、AIC 更倾向复杂模型，这样你一眼就能看懂它们的区别。

你要我画这个图吗？这样你的印象会更深。

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。

### 5.3.2.5 Effect of the prior

Sometimes it is not clear how to set the prior. When we are performing posterior inference, the details of the prior may not matter too much, since the likelihood often overwhelms the prior anyway. But when computing the marginal likelihood, the prior plays a much more important role, since we are averaging the likelihood over all possible parameter settings, as weighted by the prior. In Figures 5.7 and 5.8, where we demonstrated model selection for linear regression, we used a prior of the form  $p(w) = N(0, \alpha^{-1} I)$ . Here  $\alpha$  is a tuning parameter that controls how strong the prior is. This parameter can have a large effect, as we discuss in Section 7.5. Intuitively, if  $\alpha$  is large, the weights are “forced” to be small, so we need to use a complex model with many small parameters (e.g., a high degree polynomial) to fit the data. Conversely, if  $\alpha$  is small, we will favor simpler models, since each parameter is “allowed” to vary in magnitude by a lot. If the prior is unknown, the correct Bayesian procedure is to put a prior on the prior. That is, we should put a prior on the hyper-parameter  $\alpha$  as well as the parameters  $w$ . To compute the marginal likelihood, we should integrate out all unknowns, i.e., we should compute

$$p(D|m) = p(D|w)p(w|\alpha, m)p(\alpha|m)dwd\alpha \quad (5.36)$$

Of course, this requires specifying the hyper-prior. Fortunately, the higher up we go in the Bayesian hierarchy, the less sensitive are the results to the prior

settings. So we can usually make the hyper-prior uninformative. A computational shortcut is to optimize  $\alpha$  rather than integrating it out. That is, we use

$$p(D|m) \approx p(D|w)p(w|\hat{\alpha}, m)dw \quad (5.37)$$

where

$$\hat{\alpha} = \operatorname{argmax} \alpha \ p(D|\alpha, m) = \operatorname{argmax} \alpha \ p(D|w)p(w|\alpha, m)dw \quad (5.38)$$

This approach is called empirical Bayes (EB), and is discussed in more detail in Section 5.6. This is the method used in Figures 5.7 and 5.8.

好的，我们先逐句完整翻译，然后再用通俗易懂、数学基础友好的方式解释，保证你能明白为什么先验 (prior) 会影响边际似然，以及“经验贝叶斯”是怎么回事。

---

## 逐句翻译

### 5.3.2.5 先验的影响

有时我们并不清楚该如何设定先验分布 (prior)。

在进行后验推断 (posterior inference) 时，先验的细节可能并不那么重要，因为似然函数 (likelihood) 往往会“压倒”先验的影响。

但是，当我们计算**边际似然** (marginal likelihood) 时，先验就会发挥更加重要的作用，因为在这个过程中，我们是将**似然在所有可能的参数取值上做加权平均**，而加权的权重正是由先验分布给出的。

在图 5.7 和图 5.8 中，我们展示了线性回归的模型选择，并使用了这样一种先验：

$$p(w) = N(0, \alpha^{-1} I)$$

这里  $\alpha$  是一个调节参数，用来控制先验的强度。

这个参数可能会带来很大的影响（我们会在 7.5 节中进一步讨论）。

直观来说：

- 如果  $\alpha$  很大，权重  $w$  会被“强迫”变小，因此为了拟合数据，我们需要使用一个**复杂的模型**，该模型中有很多很小的参数（例如高阶多项式）。
- 相反，如果  $\alpha$  很小，那么每个参数  $w$  的幅度就被允许有较大变化，这样会更倾向于选择**简单模型**。

如果先验未知，正确的贝叶斯方法是**对先验再加一个先验**。也就是说，我们不仅要给  $w$  设定先验，还要给超参数  $\alpha$  设定一个先验 (hyper-prior)。

为了计算边际似然，我们应该把所有未知量都积分掉：

$$p(D|m) = \iint p(D|w)p(w|\alpha, m)p(\alpha|m)dw d\alpha \quad (5.36)$$

当然，这需要先指定超先验（hyper-prior）。

幸运的是，越往贝叶斯层级的高层走，结果对先验设置的敏感度就越低，因此我们通常可以把超先验设为非信息性的（uninformative）。

一个计算上的捷径是优化  $\alpha$  而不是把它积分掉。也就是说，我们使用：

$$p(D|m) \approx \int p(D|w) p(w|\hat{\alpha}, m) dw \quad (5.37)$$

其中：

$$\hat{\alpha} = \arg \max_{\alpha} p(D|\alpha, m) = \arg \max_{\alpha} \int p(D|w) p(w|\alpha, m) dw \quad (5.38)$$

这种方法叫做**经验贝叶斯**（Empirical Bayes, EB），我们会在 5.6 节中详细讨论。

图 5.7 和图 5.8 中用的就是这种方法。

## 详细解释（基础+直观版）

### 1. 先验为什么平时不太重要，但在边际似然中很重要？

- 在后验推断时（比如计算  $p(w|D)$ ），似然  $p(D|w)$  通常会因为数据量大而占主导作用，先验只是稍微调整结果。
- 但在**边际似然**  $p(D|m)$  中，我们是：

$$p(D|m) = \int p(D|w) p(w) dw$$

这就像你要计算模型“总体表现”，不是只看最佳参数，而是看所有可能参数的加权平均，而权重就是先验。

所以先验在这里起到很大作用。

### 2. $\alpha$ 的作用（直观解释）

- 先验  $p(w) = N(0, \alpha^{-1} I)$  表示：参数  $w$  平均值为 0，方差为  $1/\alpha$ 。
- $\alpha$  控制了“我们先验上相信参数多大”：
  - **大**  $\alpha \rightarrow$  方差  $1/\alpha$  很小  $\rightarrow$  参数被压得很小  $\rightarrow$  要解释数据只能用**更多的参数**  $\rightarrow$  倾向复杂模型。
  - **小**  $\alpha \rightarrow$  方差大  $\rightarrow$  每个参数可以随意取较大值  $\rightarrow$  简单模型就可能拟合数据。

### 3. 如果 $\alpha$ 不知道该怎么选？

- **标准贝叶斯**：给  $\alpha$  再加一个先验（hyper-prior），并把它一起积分掉（公式 5.36）。
- **经验贝叶斯（EB）**：不积分，直接找一个最优  $\hat{\alpha}$ （最大化边际似然，公式 5.38）。

## 4. 经验贝叶斯的意义

- EB 是一种折中方法：计算量比真正的完全贝叶斯少，但能自动调整先验强度。
- 缺点：不是完全遵循贝叶斯原则，因为它在某一层用了优化而不是积分。

## 5. 生活化比喻

想象你在做衣服：

- $w$  = 衣服的尺寸（模型参数）
- $\alpha$  = 量尺的紧度（先验强度）
- 大  $\alpha \rightarrow$  量尺收得紧  $\rightarrow$  做出来的衣服偏小  $\rightarrow$  需要多加一些补丁（复杂模型）
- 小  $\alpha \rightarrow$  量尺放松  $\rightarrow$  尺寸随意  $\rightarrow$  有时简单结构就够

经验贝叶斯就像是：你不提前规定量尺多紧，而是先做几件衣服，看哪种紧度让成衣最合身，再用那个紧度做最终成品。

我可以帮你画一个图解，用曲线展示  $\alpha$  大小对模型复杂度和边际似然的影响，这样你能直观看出为什么先验在边际似然里影响大、在后验里影响小。

你要我帮你画这个图吗？这样你会更容易记住这个知识点。



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 5.9 (a) Log marginal likelihood for the coins example. (b) BIC approximation. Figure generated by coinsModelSelDemo.

### 5.3.3 Bayes factors

Suppose our prior on models is uniform,  $p(m) \propto 1$ . Then model selection is equivalent to picking the model with the highest marginal likelihood. Now suppose we just have two models we are considering, call them the null

hypothesis, M<sub>0</sub>, and the alternative hypothesis, M<sub>1</sub>. Define the Bayes factor as the ratio of marginal likelihoods:

$$BF_{1,0} = p(D|M_1) p(D|M_0) = p(M_1|D) p(M_0|D) / p(M_1) p(M_0) \quad (5.39)$$

(This is like a likelihood ratio, except we integrate out the parameters, which allows us to compare models of different complexity.) If  $BF_{1,0} > 1$  then we prefer model 1, otherwise we prefer model 0. Of course, it might be that  $BF_{1,0}$  is only slightly greater than 1. In that case, we are not very confident that model 1 is better. Jeffreys (1961) proposed a scale of evidence for interpreting the magnitude of a Bayes factor, which is shown in Table 5.1. This is a Bayesian alternative to the frequentist concept of a p-value.<sup>3</sup> Alternatively, we can just convert the Bayes factor to a posterior over models. If  $p(M_1) = p(M_0) = 0.5$ , we have

$$p(M_0|D) = BF_{0,1} / (1 + BF_{0,1}) = 1 / (1 + BF_{1,0}) \quad (5.40)$$

### 5.3.3.1 Example: Testing if a coin is fair

Suppose we observe some coin tosses, and want to decide if the data was generated by a fair coin,  $\theta = 0.5$ , or a potentially biased coin, where  $\theta$  could be any value in [0, 1]. Let us denote the first model by M<sub>0</sub> and the second model by M<sub>1</sub>. The marginal likelihood under M<sub>0</sub> is simply

$$p(D|M_0) = 1 / N! \quad (5.41)$$

where N is the number of coin tosses. The marginal likelihood under M<sub>1</sub>, using a Beta prior, is

$$p(D|M_1) = p(D|\theta)p(\theta)d\theta = B(\alpha_1 + N_1, \alpha_0 + N_0) / B(\alpha_1, \alpha_0) \quad (5.42)$$

We plot  $\log p(D|M_1)$  vs the number of heads N<sub>1</sub> in Figure 5.9(a), assuming N = 5 and  $\alpha_1 = \alpha_0 = 1$ . (The shape of the curve is not very sensitive to  $\alpha_1$  and  $\alpha_0$ , as long as  $\alpha_0 = \alpha_1$ .) If we observe 2 or 3 heads, the unbiased coin hypothesis M<sub>0</sub> is more likely than M<sub>1</sub>, since M<sub>0</sub> is a simpler model (it has no free parameters) — it would be a suspicious coincidence if the coin were biased but happened to produce almost exactly 50/50 heads/tails. However, as the counts become more extreme, we favor the biased coin hypothesis. Note that, if we plot the log Bayes factor,  $\log BF_{1,0}$ , it will have exactly the same shape, since  $\log p(D|M_0)$  is a constant. See also Exercise 3.18.

In Figure 5.9(b) shows the BIC approximation to  $\log p(D|M_1)$  for our biased coin example from Section 5.3.3.1. We see that the curve has approximately the same shape as the exact log marginal likelihood, which is all that matters for model selection purposes, since the absolute scale is irrelevant. In particular, it favors the simpler model

unless the data is overwhelmingly in support of the more complex model.

好的，我会先完整翻译你给的内容，然后再结合图 5.9 (a) 和 (b) 给你一个基础且详细的解释，让你即使数学基础不强也能明白贝叶斯因子 (Bayes factor) 在做什么，以及它和 BIC 的关系。

## 逐句翻译

### 图 5.9

(a) 硬币 (coins) 示例的对数边际似然 (log marginal likelihood)。

(b) BIC 近似值。

该图由 `coinsModelSelDemo` 生成。

### 5.3.3 贝叶斯因子 (Bayes factors)

假设我们对模型的先验是均匀分布，即  $p(m) \propto 1$ 。

那么模型选择就等价于选择边际似然 (marginal likelihood) 最大的模型。

现在假设我们只比较两个模型：

- 零假设  $M_0$
- 备择假设  $M_1$

定义贝叶斯因子 (Bayes factor) 为边际似然的比值：

$$BF_{1,0} = \frac{p(D|M_1)}{p(D|M_0)} = \frac{\frac{p(M_1|D)}{p(M_1)}}{\frac{p(M_0|D)}{p(M_0)}} \quad (5.39)$$

(这就像一个似然比检验，只不过这里我们把参数积分掉了，因此可以比较不同复杂度的模型。)

如果  $BF_{1,0} > 1$ ，那么我们更偏向模型 1，否则更偏向模型 0。

当然，如果  $BF_{1,0}$  只是略大于 1，那就说明我们对“模型 1 更好”的信心并不高。

Jeffreys (1961) 提出了一个证据强度的刻度 (evidence scale) 来解释贝叶斯因子的大小，这个表格在表 5.1 中展示。这是贝叶斯版本的 p 值概念。

另一种做法是直接把贝叶斯因子转化为模型的后验概率。如果

$$p(M_1) = p(M_0) = 0.5$$

那么：

$$p(M_0|D) = \frac{BF_{0,1}}{1+BF_{0,1}} = \frac{1}{BF_{1,0}+1} \quad (5.40)$$


---

### 5.3.3.1 示例：检验一枚硬币是否公平

假设我们观察到一些抛硬币的结果，想判断数据是否来自：

- **公平硬币** ( $\theta = 0.5$ )
- **可能有偏的硬币** ( $\theta$  可以是  $[0,1]$  中的任意值)

我们把前者称为模型  $M_0$ ，后者称为模型  $M_1$ 。

在  $M_0$  下的边际似然是：

$$p(D|M_0) = \frac{1}{2^N} \quad (5.41)$$

其中  $N$  是抛硬币的次数。

在  $M_1$  下，假设使用 Beta 先验，边际似然是：

$$p(D|M_1) = \int p(D|\theta) p(\theta) d\theta = \frac{B(\alpha_1 + N_1, \alpha_0 + N_0)}{B(\alpha_1, \alpha_0)} \quad (5.42)$$

这里：

- $N_1$  = 正面次数
- $N_0$  = 反面次数
- $\alpha_1, \alpha_0$  是 Beta 先验的参数

我们绘制了  $\log p(D|M_1)$  与正面次数  $N_1$  的关系（见图 5.9(a)），假设  $N = 5$  且  $\alpha_1 = \alpha_0 = 1$ 。

（曲线的形状对  $\alpha_1$  和  $\alpha_0$  的精确值不敏感，只要它们相等即可。）

如果我们观察到 2 次或 3 次正面，公平硬币假设  $M_0$  比  $M_1$  更可能成立，因为  $M_0$  更简单（它没有自由参数）——如果硬币是有偏的，却刚好抛出了接近 50/50 的结果，那会显得很可疑。

然而，如果结果非常极端（比如几乎全是正面或全是反面），我们就更倾向于有偏硬币的假设。

注意，如果我们绘制对数贝叶斯因子  $\log BF_{1,0}$ ，它的形状和  $\log p(D|M_1)$  完全一样，因为  $\log p(D|M_0)$  是一个常数。

---

在图 5.9(b) 中，我们给出了前面硬币示例中  $\log p(D|M_1)$  的 BIC 近似。

我们看到曲线形状和精确的对数边际似然非常接近，这对模型选择来说已经足够，因为我们只关心曲线的相对变化，而不关心绝对数值。

特别地，这种方法倾向于选择更简单的模型，除非数据有压倒性证据支持更复杂的模型。

---

## 详细解释（基础版）

### 1. 贝叶斯因子是什么？

- 它是两个模型边际似然的比值：

$$BF_{1,0} = \frac{\text{模型1对数据的支持度}}{\text{模型0对数据的支持度}}$$

- 如果比值  $> 1 \rightarrow$  数据更支持模型 1；反之更支持模型 0。

### 2. 为什么它比直接看似然更好？

- 似然只在最佳参数处比较，而贝叶斯因子会把所有可能的参数取值考虑进去（积分），因此能比较不同复杂度的模型。

### 3. 硬币例子怎么用？

- $M_0$ : 硬币固定是公平的（无参数）
- $M_1$ : 硬币可能有偏 ( $\theta$  是参数)
- 如果结果接近一半一半，简单的  $M_0$  更合理；如果结果极端，复杂的  $M_1$  更合理。

### 4. 图 5.9 (a) 和 (b)

- (a) 是精确计算的  $\log p(D|M_1)$ ，中间（2-3 个正面）最低，表示此时公平硬币更可能。
- (b) 是 BIC 近似，形状几乎一样  $\rightarrow$  所以可以用它代替复杂的积分计算。

### 5. 总结

- 贝叶斯因子给了我们一种“数值化”比较模型的方式。
  - BIC 是它的近似版本，计算更快，但结果趋势相同。
- 

如果你愿意，我可以帮你把**贝叶斯因子的大小和 Jeffreys 的证据强度表**画出来，这样你一眼就能看出不同 BF 值代表的信心等级。这样你读硬币例子时就会很快判断“这个证据强还是弱”。

你要我画这个图吗？

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。

5.3.4 Jeffreys-Lindley paradox

Problems can arise when we use improper priors (i.e.,

priors that do not integrate to 1) for model selection/hypothesis testing, even though such priors may be acceptable for other purposes. For example, consider testing the hypotheses  $M_0 : \theta \in \Theta_0$  vs  $M_1 : \theta \in \Theta_1$ . To define the marginal density on  $\theta$ , we use the following mixture model

$$p(\theta) = p(\theta|M_0)p(M_0) + p(\theta|M_1)p(M_1) \quad (5.43)$$

This is only meaningful if  $p(\theta|M_0)$  and  $p(\theta|M_1)$  are proper (normalized) density functions. In this case, the posterior is given by

$$p(M_0|D) = p(M_0)p(D|M_0) p(M_0)p(D|M_0) + p(M_1)p(D|M_1) \quad (5.44)$$

$$= p(M_0) \int_{\Theta_0} p(D|\theta)p(\theta|M_0)d\theta p(M_0) \int_{\Theta_0} p(D|\theta)p(\theta|M_0)d\theta + p(M_1) \int_{\Theta_1} p(D|\theta)p(\theta|M_1)d\theta \quad (5.45)$$

Now suppose we use improper priors,  $p(\theta|M_0) \propto c_0$  and  $p(\theta|M_1) \propto c_1$ . Then

$$p(M_0|D) = p(M_0)c_0 \int_{\Theta_0} p(D|\theta)d\theta p(M_0)c_0 \int_{\Theta_0} p(D|\theta)d\theta + p(M_1)c_1 \int_{\Theta_1} p(D|\theta)d\theta \quad (5.46)$$

$$= p(M_0)c_0 \quad 0 p(M_0)c_0 \quad 0 + p(M_1)c_1 \quad 1 \quad (5.47)$$

where  $i = \int_{\Theta_i} p(D|\theta)d\theta$  is the integrated or marginal likelihood for model  $i$ . Now let  $p(M_0) = p(M_1) = 1/2$ . Hence

$$p(M_0|D) = c_0 \quad 0 c_0 \quad 0 + c_1 \quad 1 = 0 \quad 0 + (c_1/c_0) \quad 1 \quad (5.48)$$

Thus we can change the posterior arbitrarily by choosing  $c_1$  and  $c_0$  as we please. Note that using proper, but very vague, priors can cause similar problems. In particular, the Bayes factor will always favor the simpler model, since the probability of the observed data under a complex model with a very diffuse prior will be very small. This is called the Jeffreys-Lindley paradox.

Thus it is important to use proper priors when performing model selection. Note, however, that, if  $M_0$  and  $M_1$  share the same prior over a subset of the parameters, this part of the prior can be improper, since the corresponding normalization constant will cancel out.

好的，我会先给你一个完整、无遗漏的中文翻译，然后再做一个基础且详细的解释，用直观类比帮你理解 Jeffreys-Lindley 悖论的含义。

## 逐句翻译

### 5.3.4 Jeffreys–Lindley 悖论

在进行模型选择或假设检验时，如果我们使用**不适当先验**（improper priors，意思是该先验分布的积分不等于 1），可能会出现问题，尽管这种先验在其他用途下可能是可接受的。

例如，考虑检验如下假设：

- $M_0 : \theta \in \Theta_0$
- $M_1 : \theta \in \Theta_1$

为了定义  $\theta$  的边际密度，我们使用如下的混合模型：

$$p(\theta) = p(\theta|M_0)p(M_0) + p(\theta|M_1)p(M_1) \quad (5.43)$$

只有当  $p(\theta|M_0)$  和  $p(\theta|M_1)$  是**适定的** (proper, 即已经归一化的) 概率密度函数时，这个式子才有意义。

在这种情况下，后验概率为：

$$p(M_0|D) = \frac{p(M_0)p(D|M_0)}{p(M_0)p(D|M_0) + p(M_1)p(D|M_1)} \quad (5.44)$$

进一步写为：

$$= \frac{p(M_0) \int_{\Theta_0} p(D|\theta)p(\theta|M_0) d\theta}{p(M_0) \int_{\Theta_0} p(D|\theta)p(\theta|M_0) d\theta + p(M_1) \int_{\Theta_1} p(D|\theta)p(\theta|M_1) d\theta} \quad (5.45)$$


---

现在假设我们使用不**适定先验**：

$$p(\theta|M_0) \propto c_0, \quad p(\theta|M_1) \propto c_1$$

那么：

$$p(M_0|D) = \frac{p(M_0) c_0 \int_{\Theta_0} p(D|\theta) d\theta}{p(M_0) c_0 \int_{\Theta_0} p(D|\theta) d\theta + p(M_1) c_1 \int_{\Theta_1} p(D|\theta) d\theta} \quad (5.46)$$

记：

$$m_i = \int_{\Theta_i} p(D|\theta) d\theta$$

它是模型  $i$  的**积分似然或边际似然**，于是：

$$p(M_0|D) = \frac{p(M_0) c_0 m_0}{p(M_0) c_0 m_0 + p(M_1) c_1 m_1} \quad (5.47)$$


---

如果再假设：

$$p(M_0) = p(M_1) = \frac{1}{2}$$

则：

$$p(M_0|D) = \frac{c_0 m_0}{c_0 m_0 + c_1 m_1} = \frac{m_0}{m_0 + (c_1/c_0)m_1}$$
贝叶斯统计翻译解释  
(5.48)

因此，我们可以通过任意选择  $c_1$  和  $c_0$  来随意改变后验概率。

注意，即使是适定的但非常宽泛 (vague) 的先验，也可能导致类似的问题。尤其是，**贝叶斯因子会总是偏向更简单的模型**，因为在复杂模型下，带有非常分散先验的情况下，观测数据的概率会非常小。

这种现象被称为 Jeffreys–Lindley 悖论。

因此，在进行模型选择时，**使用适定先验非常重要**。

不过需要注意的是，如果  $M_0$  和  $M_1$  在一部分参数上的先验是相同的，那么这部分先验可以是不适当的，因为对应的归一化常数会在比较时相互抵消。

---

## 详细解释（基础版）

### 1. 什么是“不适当先验”

- 正常的概率分布必须满足“积分等于 1”。
- 不适当先验 (improper prior) 就是积分是无穷大 (或者没有定义) 的分布，例如：

$$p(\theta) \propto 1 \quad (\theta \in \mathbb{R})$$

它的积分无限大，所以不是严格意义上的概率分布。

### 2. 为什么它在模型比较里有问题

- 在模型比较中，我们需要计算**边际似然**：

$$p(D|M) = \int p(D|\theta) p(\theta|M) d\theta$$

- 如果  $p(\theta|M)$  不适当，它其实相当于乘了一个任意的常数  $c$ 。
- 这个常数会直接影响**模型的后验概率**，而且你可以随便调它 → 导致结论完全不可信。

### 3. 公式里的问题

- 在式 (5.48) 里，后验概率中出现了  $c_1/c_0$  这个比值。
- 由于  $c_0, c_1$  可以随意选择，后验概率就完全取决于你**主观设的数值**，而不是数据本身。

### 4. Jeffreys–Lindley 悖论的核心

- 即使你不用不适当先验，而是用**非常宽泛的先验**（看似“无信息”），复杂模型的边际似然还是会被压得很低，因为大部分参数取值下，模型预测的概率很小。
- 结果是：**贝叶斯因子会总是倾向简单模型**，哪怕数据可能真的支持复杂模型。

### 5. 怎么避免

- 选用**适当** (proper) 先验。

- 如果两个模型有一部分参数的先验完全相同，这部分可以不适用，因为它们会相互抵消，不影响比较。

如果你愿意，我可以帮你画一个Jeffreys–Lindley 悖论的直观示意图，用一个“公平硬币 vs 偏硬币”的例子演示为什么过于宽泛的先验会让复杂模型输给简单模型，即使它可能是对的。这样你能直观感受到“先验越宽 → 复杂模型越难赢”。

你要我画这个图吗？

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。

#### 5.4 Priors

The most controversial aspect of Bayesian statistics is its reliance on priors. Bayesians argue this is unavoidable, since nobody is a tabula rasa or blank slate: all inference must be done conditional on certain assumptions about the world. Nevertheless, one might be interested in minimizing the impact of one's prior assumptions. We briefly discuss some ways to do this below.

##### 5.4.1 Uninformative priors

If we don't have strong beliefs about what  $\theta$  should be, it is common to use an uninformative or non-informative prior, and to “let the data speak for itself”.

The issue of designing uninformative priors is actually somewhat tricky. As an example of the difficulty, consider a Bernoulli parameter,  $\theta \in [0, 1]$ . One might think that the most uninformative prior would be the uniform distribution, Beta(1, 1). But the posterior mean in this case is  $E[\theta|D] = N_1 + 1 / (N_1 + N_0 + 2)$ , whereas the MLE is  $N_1 / (N_1 + N_0)$ . Hence one could argue that the prior wasn't completely uninformative after all.

Clearly by decreasing the magnitude of the pseudo counts, we can lessen the impact of the prior. By the above argument, the most non-informative prior is

$$\lim_{c \rightarrow 0} \text{Beta}(c, c) = \text{Beta}(0, 0) \quad (5.49)$$

which is a mixture of two equal point masses at 0 and 1 (see (Zhu and Lu 2004)). This is also called the Haldane prior. Note that the Haldane prior is an improper prior, meaning it does not integrate to 1. However, as long as we see at least one head and at least one tail, the posterior will be proper. In Section 5.4.2.1 we will argue

that the “right” uninformative prior is in fact Beta(1 2 , 1 2 ). Clearly the difference in practice between these three priors is very likely negligible. In general, it is advisable to perform some kind of sensitivity analysis, in which one checks how much one’s conclusions or predictions change in response to change in the modeling assumptions, which includes the choice of prior, but also the choice of likelihood and any kind of data preprocessing. If the conclusions are relatively insensitive to the modeling assumptions, one can have more confidence in the results.

好，我会先**完整、逐句翻译**原文（无遗漏），再做一个**基础且详尽**的解释，让你即使数学基础不强也能明白。

---

## 逐句翻译

### 5.4 先验 (Priors)

贝叶斯统计中最具争议的部分，就是它对**先验分布**的依赖。

贝叶斯派认为这是不可避免的，因为没有人是一块白板 (tabula rasa，意为“空白的石板”)：所有推断都必须建立在对世界某种假设的条件之上。

尽管如此，人们可能还是希望**尽量减少**先验假设的影响。

下面我们会简要讨论一些实现这一目标的方法。

---

#### 5.4.1 无信息先验 (Uninformative priors)

如果我们对参数  $\theta$  没有强烈的信念（先验认知），通常会使用**无信息先验** (uninformative prior) 或**非信息先验** (non-informative prior)，并“让数据自己说话”。

设计无信息先验实际上有点棘手。举个难点的例子：

考虑一个伯努利 (Bernoulli) 分布的参数  $\theta \in [0, 1]$ 。

有人可能会觉得，最无信息的先验就是均匀分布，即  $Beta(1, 1)$ 。

但这种情况下，后验均值是：

$$E[\theta|D] = \frac{N_1 + 1}{N_1 + N_0 + 2}$$

而最大似然估计 (MLE) 是：

$$\frac{N_1}{N_1 + N_0}$$

因此可以说，这个先验并不是完全“无信息”的，因为它推高了数值。

---

显然，通过减少伪计数（pseudo counts）的大小，我们可以减弱先验的影响。

根据上述推理，最“无信息”的先验是：

$$\lim_{c \rightarrow 0} \text{Beta}(c, c) = \text{Beta}(0, 0) \quad (5.49)$$

它是 0 和 1 处各有一半概率质量的混合分布（参见 Zhu 和 Lu, 2004）。

这种先验也被称为 **Haldane 先验**。

注意，Haldane 先验是不适当先验（improper prior），也就是说它的积分不等于 1。

不过，只要我们至少观察到一次正面（head）和一次反面（tail），它的后验就是适当的（proper）。

---

在 5.4.2.1 节，我们会论证真正“正确”的无信息先验其实是  $\text{Beta}(\frac{1}{2}, \frac{1}{2})$ 。

显然，这三种先验（ $\text{Beta}(1,1)$ 、 $\text{Beta}(0,0)$ 、 $\text{Beta}(1/2,1/2)$ ）在实际应用中的差别很可能微不足道。

总体来说，建议进行**敏感性分析**（sensitivity analysis）：

检查在改变建模假设（包括先验的选择、似然函数的选择、数据预处理方式等）后，你的结论或预测变化有多大。

如果结论对这些假设变化**不敏感**，那么我们就可以对结果更有信心。

---

## 详细解释（基础版）

### 1. 什么是“先验”

- 在贝叶斯统计里，我们在看到数据之前，对参数可能的取值有一个**预先的信念**，用概率分布表示，这就是**先验分布**。
- 比如抛硬币前，你可能相信它大致是公平的，那就给  $\theta$ （正面概率）一个接近 0.5 的先验分布。

### 2. 无信息先验的目的

- 如果我们不想让“主观假设”影响结论，就想用一个尽量“平等对待所有可能值”的先验，让数据主导推断。

### 3. 为什么“均匀分布”不是真正完全无信息

- 均匀分布  $\text{Beta}(1, 1)$  看起来好像“对所有  $\theta$  都一样”，但它还是会对后验值有轻微的偏移（加上了 1 个伪计数）。
- 比如观测到  $N_1$  次正面、 $N_0$  次反面时，均匀先验会让后验均值稍微偏向 0.5。

### 4. Haldane 先验 $\text{Beta}(0, 0)$

- 这是数学意义上的“极限情况”，它相当于只在 0 和 1 上有概率质量。

- 它非常“激进”，因为它表示：在你看到数据之前，你只相信硬币要么永远正面，要么永远反面。
- 它是不适当的（improper），但如果数据中正反面都至少出现一次，就能得到一个正常的后验分布。

## 5. 另一种推荐的无信息先验

- 文中提到  $\text{Beta}(1/2, 1/2)$  被认为是理论上的“正确”无信息先验（来自 Jeffreys 先验的推导）。
- 三者 ( $\text{Beta}(1,1)$ 、 $\text{Beta}(0,0)$ 、 $\text{Beta}(1/2,1/2)$ ) 在实际计算中差别通常很小。

## 6. 敏感性分析的重要性

- 不光先验会影响结果，似然函数的选择、数据怎么预处理，也都会影响结果。
- 所以要尝试改变这些假设，看结论是否变化很大。
- 如果变化不大 → 结果更可信。

如果你愿意，我可以帮你画一张**比较三种无信息先验形状的图** ( $\text{Beta}(1,1)$ 、 $\text{Beta}(0,0)$ 、 $\text{Beta}(1/2,1/2)$ )，让你直观理解它们对后验的影响。

这样可以用“抛硬币”这个简单例子让你看到差别很小，但原理不同。

你要我画吗？

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。

### 5.4.2 Jeffreys priors

Harold Jeffreys designed a general purpose technique for creating non-informative priors. The result is known as the Jeffreys prior. The key observation is that if  $p(\varphi)$  is non-informative, then any re-parameterization of the prior, such as  $\theta = h(\varphi)$  for some function  $h$ , should also be non-informative. Now, by the change of variables formula,

$$p\theta(\theta) = p\varphi(\varphi) \frac{d\varphi}{d\theta} d\theta \quad (5.50)$$

so the prior will in general change. However, let us pick

$$p\varphi(\varphi) \propto I(\varphi)^{-1/2} \quad (5.51)$$

where  $I(\varphi)$  is the Fisher information:

$$I(\varphi) = -E \frac{d}{d\varphi} \log p(X|\varphi) \frac{d\varphi}{d\theta} \quad (5.52)$$

This is a measure of curvature of the expected negative log likelihood and hence a measure of stability of the MLE (see Section 6.2.2). Now

$$\frac{d}{d\theta} \log p(x|\theta) = \frac{d}{d\theta} \log p(x|\varphi) \frac{d\varphi}{d\theta} \quad (5.53)$$

Squaring and taking expectations over  $x$ , we have

$$I(\theta) = -E d \log p(X|\theta) d\theta^2 = I(\varphi) d\varphi d\theta^2 \quad (5.54)$$

$$I(\theta) d\theta^2 = I(\varphi) d\varphi d\theta^2 \quad (5.55)$$

so we find the transformed prior is

$$p_\theta(\theta) = p_\varphi(\varphi) d\varphi d\theta \propto (I(\varphi))^{1/2} d\varphi d\theta = I(\theta)^{1/2} \quad (5.56)$$

So  $p_\theta(\theta)$  and  $p_\varphi(\varphi)$  are the same. Some examples will make this clearer.

好的，我会先逐句翻译原文（不遗漏），再做一个基础且详尽的解释，确保数学基础不强也能理解。

---

## 逐句翻译

### 5.4.2 Jeffreys 先验 (Jeffreys priors)

Harold Jeffreys 设计了一种通用的方法来构造无信息先验。

这种方法的结果被称为 Jeffreys 先验。

关键观察是：如果  $p(\varphi)$  是无信息的，那么它的任何重新参数化 (re-parameterization)，例如  $\theta = h(\varphi)$  (其中  $h$  是某个函数)，也应该是无信息的。

---

根据变量替换公式 (change of variables formula)：

$$p_\theta(\theta) = p_\varphi(\varphi) \cdot \left| \frac{d\varphi}{d\theta} \right| \quad (5.50)$$

因此，一般情况下，先验会发生变化。

但是，我们选择：

$$p_\varphi(\varphi) \propto \sqrt{I(\varphi)} \quad (5.51)$$

其中  $I(\varphi)$  是 Fisher 信息量 (Fisher information)：

$$I(\varphi) = -E \left[ \left( \frac{d}{d\varphi} \log p(X|\varphi) \right)^2 \right] \quad (5.52)$$

这衡量了期望负对数似然的曲率，因此也是 MLE (最大似然估计) 的稳定性的度量（见第 6.2.2 节）。

---

接下来：

$$\frac{d}{d\theta} \log p(x|\theta) = \frac{d}{d\varphi} \log p(x|\varphi) \cdot \frac{d\varphi}{d\theta} \quad (5.53)$$

平方并对  $x$  取期望，我们有：

$$I(\theta) = -E \left[ \left( \frac{d}{d\theta} \log p(X|\theta) \right)^2 \right] = I(\varphi) \cdot \left( \frac{d\varphi}{d\theta} \right)^2 \quad (5.54)$$

因此：

$$\sqrt{I(\theta)} = \sqrt{I(\varphi)} \cdot \left| \frac{d\varphi}{d\theta} \right| \quad (5.55)$$

所以我们得到变换后的先验为：

$$p_\theta(\theta) = p_\varphi(\varphi) \cdot \left| \frac{d\varphi}{d\theta} \right| \propto \sqrt{I(\varphi)} \cdot \left| \frac{d\varphi}{d\theta} \right| = \sqrt{I(\theta)} \quad (5.56)$$

由此可见， $p_\theta(\theta)$  和  $p_\varphi(\varphi)$  形式上是相同的。

接下来我们会通过一些例子来说明这一点。

---

## 详细解释（基础版）

### 1. 背景

Jeffreys 想解决的问题是：

- 我们在选先验时，有时候会换一种方式表示参数（比如用概率  $p$  或者用 log-odds 表示）。
- 如果先验是“无信息”的，不管用哪种参数表示，它都应该是“无信息”的。
- 但是普通的均匀先验并不具备这个特性——换个参数形式，它就可能变得有偏了。

### 2. 关键数学思想

- 变量替换公式**告诉我们：当参数换成另一种形式时，概率密度要乘以雅可比 (Jacobian)  $\left| \frac{d\varphi}{d\theta} \right|$ 。
- 这意味着“看起来均匀”的先验，换个参数后不一定均匀。

### 3. Fisher 信息量的作用

- Jeffreys 的方法是：选择先验为

$$p(\varphi) \propto \sqrt{I(\varphi)}$$

- Fisher 信息量  $I(\varphi)$  衡量了数据告诉我们关于参数的信息多少（也可以理解为似然曲线的陡峭程度）。
- 取它的平方根，就得到了一个在不同参数化下不变的先验。

### 4. 公式推导思路

- 先写出在  $\theta$  表示下的导数关系（链式法则，公式 5.53）。

- 再平方并取期望，得到 Fisher 信息量的变换公式 (5.54)。
- 然后取平方根 (5.55)，代回变量替换公式 (5.50)。
- 结果就是 Jeffreys 先验的形式在参数变换下保持一致 (5.56)。

## 5. 直观意义

- Jeffreys 先验是根据模型的统计结构自动生成的，不会依赖你选择哪种参数化方式。
- 它通常会比简单的均匀分布更“公平”，因为它考虑了不同参数值对数据拟合的敏感程度。

如果你愿意，我可以给你画一个伯努利分布的 Jeffreys 先验和均匀先验的对比图，让你直观看到它们的区别，并解释为什么 Jeffreys 先验在 0 和 1 处更高。

你要我帮你画吗？

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。

### 5.4.2.1 Example: Jeffreys prior for the Bernoulli and multinoulli

Suppose  $X \sim \text{Ber}(\theta)$ . The log likelihood for a single sample is

$$\log p(X|\theta) = X \log \theta + (1 - X) \log(1 - \theta) \quad (5.57)$$

The score function is just the gradient of the log-likelihood:

$$s(\theta) = d \frac{d\theta}{d\theta} \log p(X|\theta) = X \theta - 1 - X 1 - \theta \quad (5.58)$$

The observed information is the second derivative of the log-likelihood:

$$J(\theta) = - \frac{d^2}{d\theta^2} \log p(X|\theta) = -s'(\theta|X) = X \theta^2 + 1 - X (1 - \theta)^2 \quad (5.59)$$

The Fisher information is the expected information:

$$I(\theta) = E[J(\theta|X)|X \sim \theta] = \theta \theta^2 + 1 - \theta (1 - \theta)^2 = 1 \theta(1 - \theta) \quad (5.60)$$

Hence Jeffreys' prior is

$$p(\theta) \propto \theta^{-1/2} (1 - \theta)^{-1/2} = 1 \theta(1 - \theta)^{-1} \propto \text{Beta}(1/2, 1/2) \quad (5.61)$$

Now consider a multinoulli random variable with K states.

One can show that the Jeffreys' prior is given by

$$p(\theta) \propto \text{Dir}(1/2, \dots, 1/2) \quad (5.62)$$

Note that this is different from the more obvious choices of  $\text{Dir}(1/K, \dots, 1/K)$  or  $\text{Dir}(1, \dots, 1)$ .

### 5.4.2.2 Example: Jeffreys prior for location and scale

### parameters

One can show that the Jeffreys prior for a location parameter, such as the Gaussian mean, is  $p(\mu) \propto 1$ . Thus is an example of a translation invariant prior, which satisfies the property that the probability mass assigned to any interval,  $[A, B]$  is the same as that assigned to any other shifted interval of the same width, such as  $[A - c, B - c]$ .

That is,

$$B - c A - c p(\mu) d\mu = (A - c) - (B - c) = (A - B) = B A p(\mu) d\mu \quad (5.63)$$

This can be achieved using  $p(\mu) \propto 1$ , which we can approximate by using a Gaussian with infinite variance,  $p(\mu) = N(\mu | 0, \infty)$ . Note that this is an improper prior, since it does not integrate to 1. Using improper priors is fine as long as the posterior is proper, which will be the case provided we have seen  $N \geq 1$  data points, since we can "nail down" the location as soon as we have seen a single data point.

Similarly, one can show that the Jeffreys prior for a scale parameter, such as the Gaussian variance, is  $p(\sigma^2) \propto 1/\sigma^2$ . This is an example of a scale invariant prior, which satisfies the property that the probability mass assigned to any interval  $[A, B]$  is the same as that assigned to any other interval  $[A/c, B/c]$  which is scaled in size by some constant factor  $c > 0$ . (For example, if we change units from meters to feet we do not want that to affect our inferences.) This can be achieved by using

$$p(s) \propto 1/s \quad (5.64)$$

To see this, note that

$$B/c A/c p(s) ds = [\log s] B/c A/c = \log(B/c) - \log(A/c) \quad (5.65)$$

$$= \log(B) - \log(A) = B A p(s) ds \quad (5.66)$$

We can approximate this using a degenerate Gamma distribution (Section 2.4.4),  $p(s) = Ga(s | 0, 0)$ . The prior  $p(s) \propto 1/s$  is also improper, but the posterior is proper as soon as we have seen  $N \geq 2$  data points (since we need at least two data points to estimate a variance).

好的，我会先逐句完整翻译，然后用通俗+详细的方式解释每个部分，让你在数学基础不强的情况下也能理解。

## 逐句翻译

### 5.4.2.1 例子：伯努利分布与多项分布 (multinoulli) 的 Jeffreys 先验

假设  $X \sim Ber(\theta)$ 。

单个样本的对数似然为：

$$\log p(X|\theta) = X \log \theta + (1 - X) \log(1 - \theta) \quad (5.57)$$

\*\*得分函数 (score function) \*\*就是对数似然的梯度:

$$s(\theta) = \frac{d}{d\theta} \log p(X|\theta) = \frac{X}{\theta} - \frac{1-X}{1-\theta} \quad (5.58)$$

\*\*观察信息量 (observed information) \*\*是对数似然的二阶导数:

$$J(\theta) = -\frac{d^2}{d\theta^2} \log p(X|\theta) = -s'(\theta|X) = \frac{X}{\theta^2} + \frac{1-X}{(1-\theta)^2} \quad (5.59)$$

Fisher 信息量是期望的观察信息量:

$$I(\theta) = E[J(\theta|X) | X \sim \theta] = \frac{\theta}{\theta^2} + \frac{1-\theta}{(1-\theta)^2} = \frac{1}{\theta(1-\theta)} \quad (5.60)$$

因此, Jeffreys 先验为:

$$p(\theta) \propto \theta^{-\frac{1}{2}} (1-\theta)^{-\frac{1}{2}} = \frac{1}{\sqrt{\theta(1-\theta)}} \propto \text{Beta}(\frac{1}{2}, \frac{1}{2}) \quad (5.61)$$

现在考虑有  $K$  个状态的多项 (multinoulli) 随机变量, 可以证明 Jeffreys 先验为:

$$p(\theta) \propto \text{Dir}(\frac{1}{2}, \dots, \frac{1}{2}) \quad (5.62)$$

注意, 这与更显而易见的  $\text{Dir}(\frac{1}{K}, \dots, \frac{1}{K})$  或  $\text{Dir}(1, \dots, 1)$  是不同的。

#### 5.4.2.2 例子：位置参数与尺度参数的 Jeffreys 先验

可以证明, 对于**位置参数** (如高斯分布的均值), Jeffreys 先验为:

$$p(\mu) \propto 1$$

这是一个**平移不变先验** (translation invariant prior), 满足这样的性质:

任何区间  $[A, B]$  分配的概率质量, 与其平移后的区间  $[A - c, B - c]$  (宽度相同) 分配的概率质量相同, 即:

$$\int_{A-c}^{B-c} p(\mu) d\mu = (A - c) - (B - c) = (A - B) = \int_A^B p(\mu) d\mu \quad (5.63)$$

这可以用  $p(\mu) \propto 1$  实现, 我们可以用方差无限大的高斯分布来近似它:

$$p(\mu) = N(\mu | 0, \infty)$$

注意, 这是一个**不适当先验** (improper prior), 因为它不能积分为 1。

使用不适当先验是可以的, 只要后验是适当的。当我们有  $N \geq 1$  个数据点时, 后验就是适当的, 因为只要有一个数据点, 我们就能确定位置参数。

同样可以证明，对于**尺度参数**（如高斯方差），Jeffreys 先验为：

$$p(\sigma^2) \propto \frac{1}{\sigma^2}$$

这是一个**尺度不变先验** (scale invariant prior)，满足这样的性质：

区间  $[A, B]$  的概率质量，与按某个常数  $c > 0$  缩放后的区间  $[A/c, B/c]$  的概率质量相同  
(例如，我们把单位从米改成英尺，不希望推断结果受到影响)。

这可以用：

$$p(s) \propto \frac{1}{s} \quad (5.64)$$

来实现。

为验证这一点，注意：

$$\int_{A/c}^{B/c} p(s) ds = [\log s]_{A/c}^{B/c} = \log(B/c) - \log(A/c) \quad (5.65)$$

$$= \log(B) - \log(A) = \int_A^B p(s) ds \quad (5.66)$$

我们可以用一个退化的 Gamma 分布（第 2.4.4 节）来近似它：

$$p(s) = \text{Ga}(s | 0, 0)$$

先验  $p(s) \propto 1/s$  也是不适当的，但只要  $N \geq 2$  个数据点，后验就是适当的（因为至少需要两个数据点来估计方差）。

---

## 详细解释（通俗版）

### 1. 伯努利分布的 Jeffreys 先验

- 伯努利分布只有一个参数  $\theta$ ，表示成功的概率。
- Jeffreys 方法会计算 Fisher 信息量  $I(\theta)$ ，它衡量“数据对参数的敏感度”。
- 对伯努利来说， $I(\theta) = \frac{1}{\theta(1-\theta)}$ 。
- 取平方根并归一化后，得到 Jeffreys 先验是  $\text{Beta}(\frac{1}{2}, \frac{1}{2})$ ，这意味着在 0 和 1 附近概率更高，而在中间（0.5 附近）概率较低。
- 这种先验对极端值（接近 0 或 1）更开放，而不像均匀先验那样偏好中间值。

### 2. 多项分布的 Jeffreys 先验

- 类似地，对于 K 分类的多项分布，Jeffreys 先验是  $\text{Dirichlet}(\frac{1}{2}, \dots, \frac{1}{2})$ ，而不是均匀的  $\text{Dirichlet}(1, \dots, 1)$ 。

### 3. 位置参数的 Jeffreys 先验

- 位置参数（如均值）改变时，整个分布只是左右平移，推断不应受影响。
- Jeffreys 先验是常数 ( $p(\mu) \propto 1$ )，相当于完全不限制位置。

- 这种先验是“平移不变”的。

#### 4. 尺度参数的 Jeffreys 先验

- 尺度参数（如方差、标准差）改变时，是分布的“伸缩”。
- Jeffreys 先验  $p(s) \propto 1/s$  是“尺度不变”的：换计量单位不会影响推断。
- 这确保了无论用米还是英尺测量，模型结论都一致。

如果你愿意，我可以帮你画**三种先验的曲线对比图**（均匀先验、Beta(1/2,1/2)、以及 Beta(1,1)），让你直观看到 Jeffreys 先验在伯努利分布中的特点。这样你会更容易记住它为什么在 0 和 1 处更高。

你想让我画吗？

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。

##### 5.4.3 Robust priors

In many cases, we are not very confident in our prior, so we want to make sure it does not have an undue influence on the result. This can be done by using robust priors (Insua and Ruggeri 2000), which typically have heavy tails, which avoids forcing things to be too close to the prior mean. Let us consider an example from (Berger 1985, p7). Suppose  $x \sim N(\theta, 1)$ . We observe that  $x = 5$  and we want to estimate  $\theta$ . The MLE is of course  $\hat{\theta} = 5$ , which seems reasonable. The posterior mean under a uniform prior is also  $\theta = 5$ . But now suppose we know that the prior median is 0, and the prior quantiles are at -1 and 1, so  $p(\theta \leq -1) = p(-1 < \theta \leq 0) = p(0 < \theta \leq 1) = p(1 < \theta) = 0.25$ . Let us also assume the prior is smooth and unimodal. It is easy to show that a Gaussian prior of the form  $N(\theta|0, 2.192)$  satisfies these prior constraints. But in this case the posterior mean is given by 3.43, which doesn't seem very satisfactory. Now suppose we use as a Cauchy prior  $T(\theta|0, 1, 1)$ . This also satisfies the prior constraints of our example. But this time we find (using numerical method integration: see robustPriorDemo for the code) that the posterior mean is about 4.6, which seems much more reasonable.

好的，我会先逐句完整翻译，然后做**非常基础且详细的解释**，帮你理解为什么“稳健先验 (Robust prior) ”能解决问题。

## 逐句翻译

### 5.4.3 稳健先验 (Robust priors)

在许多情况下，我们对自己的先验并不是很有信心，因此希望确保它不会对结果产生过度的影响。

这可以通过使用**稳健先验** (Insua 和 Ruggeri, 2000) 来实现。

稳健先验通常具有**厚尾** (heavy tails)，这样可以避免把参数值过于强制地拉到先验均值附近。

让我们来看一个来自 Berger (1985, p.7) 的例子：

假设  $x \sim N(\theta, 1)$ 。

我们观测到  $x = 5$ ，并且想估计  $\theta$ 。

最大似然估计 (MLE) 显然是：

$$\hat{\theta} = 5$$

这看起来很合理。

在**均匀先验**下的后验均值也是：

$$\theta = 5$$

但是现在假设我们知道以下先验信息：

- 先验的**中位数**是 0。
- 先验的分位点 (quantiles) 分别在 -1 和 1：

$$p(\theta \leq -1) = 0.25, \quad p(-1 < \theta \leq 0) = 0.25, \quad p(0 < \theta \leq 1) = 0.25, \quad p(1 < \theta) = 0.25$$

- 先验是**光滑的** (smooth) 且**单峰的** (unimodal)。

容易证明，一个高斯先验  $N(\theta|0, 2.192)$  满足这些先验约束。

但是在这种情况下，**后验均值**变成了：

$$3.43$$

这看起来并不令人满意（因为数据是 5，结果被拉得太靠近 0）。

现在假设我们改用一个**柯西分布** (Cauchy prior)：

$$T(\theta|0, 1, 1)$$

它同样满足上面的先验约束。

但这次我们用数值积分方法（见 `robustPriorDemo` 的代码）计算后，发现**后验均值**大约是：

$$4.6$$

这看起来合理得多。

## 详细解释（通俗版）

### 1. 什么是稳健先验 (Robust prior)

- 在贝叶斯统计中，先验是你对参数的**事先信念**。
  - 如果先验太窄（比如正态分布方差很小），它会**强行拉着结果向先验均值靠拢**。
  - 稳健先验的想法是：
    - 先验要宽松一些。
    - 特别是在远离均值的地方，**尾部概率要大**（厚尾，heavy tail）。
    - 这样，即使数据落在远离先验均值的位置，先验也不会把结果强行拉回。
- 

### 2. 为什么例子里的高斯先验会出问题

- 数据： $x = 5$ （远离先验中位数 0）。
  - 先验： $N(0, 2.192)$ 
    - 虽然它满足分位点约束，但它的尾部下降很快。
    - 在  $\theta = 5$  处，先验概率非常小。
  - 后验  $\propto$  似然  $\times$  先验
    - 似然在  $\theta = 5$  时很大（因为观测值就是 5）。
    - 但是先验在那里的值太小，所以乘积后，峰值位置被**往回拉**。
  - 结果：后验均值 3.43（明显比 5 小很多）。
- 

### 3. 为什么 Cauchy (柯西分布) 先验效果好

- 柯西分布的尾部**很厚**，衰减速度比高斯慢得多：

$$\text{Gaussian tail} \sim e^{-x^2}, \quad \text{Cauchy tail} \sim \frac{1}{x^2}$$

- 这意味着在  $\theta = 5$  时，先验值不会小得离谱。
  - 数据的影响力不会被先验严重削弱。
  - 结果：后验均值 4.6，离观测值更接近，更符合直觉。
- 

### 4. 类比理解

想象你在拉一根弹簧：

- 高斯先验就像一个**很硬的弹簧**，你把它拉到远处，它会强力拉回到中心。
  - 柯西先验就像一个**很松的弹簧**，你拉远一点，它的拉力变得很弱，让你可以保持远离中心的位置。
-

## 5. 稳健先验的实际意义

- 在不确定先验是否正确时，可以用厚尾先验（比如 Cauchy、Student-t）来避免数据被先验“绑架”。
- 适用于：
  - 数据可能会远离先验均值的情况。
  - 想减少先验对结果的影响时。

如果你愿意，我可以帮你画一个高斯先验 vs Cauchy先验的尾部比较图，让你直观感受到“厚尾”的效果，以及为什么它让结果更接近数据本身。这样会非常容易记住这个概念。你需要我画吗？

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。

### 5.4.4 Mixtures of conjugate priors

Robust priors are useful, but can be computationally expensive to use. Conjugate priors simplify the computation, but are often not robust, and not flexible enough to encode our prior knowledge. However, it turns out that a mixture of conjugate priors is also conjugate (Exercise 5.1), and can approximate any kind of prior (Dallal and Hall 1983; Diaconis and Ylvisaker 1985). Thus such priors provide a good compromise between computational convenience and flexibility. For example, suppose we are modeling coin tosses, and we think the coin is either fair, or is biased towards heads. This cannot be represented by a beta distribution. However, we can model it using a mixture of two beta distributions. For example, we might use

$$p(\theta) = 0.5\text{Beta}(\theta|20, 20) + 0.5\text{Beta}(\theta|30, 10) \quad (5.67)$$

If  $\theta$  comes from the first distribution, the coin is fair, but if it comes from the second, it is biased towards heads. We can represent a mixture by introducing a latent indicator variable  $z$ , where  $z = k$  means that  $\theta$  comes from mixture component  $k$ . The prior has the form

$$p(\theta) = k p(z = k)p(\theta|z = k) \quad (5.68)$$

where each  $p(\theta|z = k)$  is conjugate, and  $p(z = k)$  are called the (prior) mixing weights. One can show (Exercise 5.1) that the posterior can also be written as a mixture of

conjugate distributions as follows:

$$p(\theta|D) = k p(z=k|D)p(\theta|D, z=k) \quad (5.69)$$

where  $p(Z=k|D)$  are the posterior mixing weights given by

$$p(Z=k|D) = p(Z=k)p(D|Z=k) / \sum_k p(Z=k)p(D|Z=k) \quad (5.70)$$

Here the quantity  $p(D|Z=k)$  is the marginal likelihood for mixture component  $k$  (see Section 5.3.2.1).

好，我会先**完整翻译**这段 5.4.4，然后给你一个**基础且详细的解释**，尤其帮助你理解“共轭先验混合”为什么既方便计算又更灵活。

---

## 逐句翻译

### 5.4.4 共轭先验的混合 (Mixtures of conjugate priors)

稳健先验 (Robust priors) 很有用，但在计算上可能开销较大。

共轭先验 (Conjugate priors) 可以简化计算，但通常不够稳健，而且也不够灵活，无法完全表达我们想要的先验知识。

然而，事实证明，**共轭先验的混合**仍然是共轭的（参见练习 5.1），并且它可以逼近任何类型的先验分布 (Dallal 和 Hall 1983; Diaconis 和 Ylvisaker 1985)。

因此，这类先验在计算便利性与灵活性之间提供了一个很好的折中方案。

举个例子：假设我们在建模抛硬币的过程，我们认为硬币要么是公平的，要么是偏向正面 (heads)。

这种信念无法用单一的 Beta 分布表示。

但是，我们可以用**两个 Beta 分布的混合**来建模。

例如，我们可以这样写：

$$p(\theta) = 0.5 \cdot \text{Beta}(\theta|20, 20) + 0.5 \cdot \text{Beta}(\theta|30, 10) \quad (5.67)$$

如果  $\theta$  来自第一个分布，说明硬币是公平的；

如果来自第二个分布，说明它偏向正面。

我们可以通过引入一个**潜在指示变量** (latent indicator variable)  $z$  来表示混合，

其中  $z = k$  表示  $\theta$  来自混合成分  $k$ 。

先验形式为：

$$p(\theta) = \sum_k p(z=k) p(\theta|z=k) \quad (5.68)$$

其中每个  $p(\theta|z=k)$  都是共轭分布，

而  $p(z=k)$  称为 (先验) 混合权重 (mixing weights)。

可以证明 (练习 5.1) 后验分布也可以写成共轭分布的混合形式：

$$p(\theta|D) = \sum_k p(z=k|D)p(\theta|D, z=k) \quad (5.69)$$

其中  $p(z=k|D)$  是后验混合权重，公式为：

$$p(z=k|D) = \frac{p(z=k)p(D|z=k)}{\sum_k p(z=k)p(D|z=k)} \quad (5.70)$$

这里  $p(D|z=k)$  是混合成分  $k$  的边际似然（见 5.3.2.1 节）。

---

## 基础且详尽的解释

### 1. 背景：为什么要混合先验

- **稳健先验 (Robust prior)**：尾部厚，不会强行拉结果，但计算复杂。
  - **共轭先验 (Conjugate prior)**：计算非常方便（有解析公式），但形状固定，不够灵活。
  - **混合共轭先验 (Mixture of conjugate priors)**：
    - 用多个共轭先验加权组合。
    - 形状更灵活，可以近似任何先验分布。
    - 仍保留共轭的计算便利性。
- 

### 2. 核心思想

假设我们想对  $\theta$ （抛硬币正面概率）建模：

- 如果我们确信硬币**要么是公平的** ( $\theta \approx 0.5$ )，**要么强烈偏向正面** ( $\theta$  接近 0.75)，那么单个 Beta 分布很难同时表达这两种可能性。
  - 解决方法：**用两个 Beta 分布的混合**：
    - 一个 Beta(20,20) → 集中在 0.5 附近（公平）
    - 一个 Beta(30,10) → 偏向 0.75（偏向正面）
  - 混合权重（例如各 0.5）表示我们事先对这两种情况的信任程度相等。
- 

### 3. 数学形式

#### 1. 引入潜在变量 $z$

- $z$  表示选择哪个 Beta 分布。
- $p(z=1) = 0.5, p(z=2) = 0.5$ 。
- 先验写成：

$$p(\theta) = \sum_k p(z=k)p(\theta|z=k)$$

## 2. 后验计算

- 由于每个  $p(\theta|z = k)$  都是共轭先验，所以它们各自的后验形式容易计算。
- 总体的后验就是这些**单独后验的加权和**，权重是  $p(z = k|D)$ 。

## 4. 好处

- 计算方便**: 每个分量都是共轭分布，计算后验有解析解。
- 灵活**: 可以用多个分量模拟复杂形状的先验（双峰、多峰、偏斜等）。
- 解释性好**: 每个分量代表一种假设（公平硬币 / 偏向硬币）。

## 5. 形象类比

想象你在猜硬币：

- 共轭先验 = 你只考虑“一种可能的硬币” → 太死板。
- 稳健先验 = 你考虑“各种可能性” → 但计算很麻烦。
- 混合共轭先验 = 你同时考虑几种具体可能性（几种 Beta 分布），并给它们不同权重  
→ 灵活 + 易算。

如果你愿意，我可以帮你画一个**混合 Beta 分布的形状图**，让你直观看到它是如何比单个 Beta 分布更灵活的，这样会更容易理解。

你要我帮你画吗？



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 5.10 A mixture of two Beta distributions. Figure generated by mixBetaDemo.

#### 5.4.4.1 Example

Suppose we use the mixture prior

$$p(\theta) = 0.5\text{Beta}(\theta|a_1, b_1) + 0.5\text{Beta}(\theta|a_2, b_2) \quad (5.71)$$

where  $a_1 = b_1 = 20$  and  $a_2 = b_2 = 10$ . and we observe  $N_1$  heads and  $N_0$  tails. The posterior becomes

$$\begin{aligned} p(\theta|D) &= p(Z=1|D)\text{Beta}(\theta|a_1 + N_1, b_1 + N_0) + p(Z \\ &= 2|D)\text{Beta}(\theta|a_2 + N_1, b_2 + N_0) \end{aligned} \quad (5.72)$$

If  $N_1 = 20$  heads and  $N_0 = 10$  tails, then, using Equation 5.23, the posterior becomes

$$p(\theta|D) = 0.346 \text{Beta}(\theta|40, 30) + 0.654 \text{Beta}(\theta|50, 20) \quad (5.73)$$

See Figure 5.10 for an illustration.

#### 5.4.4.2 Application: Finding conserved regions in DNA and protein sequences

We mentioned that Dirichlet-multinomial models are widely used in biosequence analysis. Let us give a simple example to illustrate some of the machinery that has developed. Specifically, consider the sequence logo discussed in Section 2.3.2.1. Now suppose we want to find locations which represent coding regions of the genome. Such locations often have the same letter across all sequences, because of evolutionary pressure. So we need to find columns which are “pure”, or nearly so, in the sense that they are mostly all As, mostly all Ts, mostly all Cs, or mostly all Gs. One approach is to look for low-entropy columns; these will be ones whose distribution is nearly deterministic (pure). But suppose we want to associate a confidence measure with our estimates of purity. This can be useful if we believe adjacent locations are conserved together. In this case, we can let  $Z_1 = 1$  if location  $t$  is conserved, and let  $Z_t = 0$  otherwise. We can then add a dependence between adjacent  $Z_t$  variables using a Markov chain; see Chapter 17 for details. In any case, we need to define a likelihood model,  $p(N_t|Z_t)$ , where  $N_t$  is the vector of (A,C,G,T) counts for column  $t$ . It is natural to make this be a multinomial distribution with parameter  $\theta_t$ . Since each column has a different distribution, we will want to integrate out  $\theta_t$  and thus compute the marginal likelihood

$$p(N_t|Z_t) = p(N_t|\theta_t)p(\theta_t|Z_t)d\theta_t \quad (5.74)$$

But what prior should we use for  $\theta_t$ ? When  $Z_t = 0$  we can use a uniform prior,  $p(\theta|Z_t = 0) = \text{Dir}(1, 1, 1, 1)$ , but what should we use if  $Z_t = 1$ ? After all, if the column is conserved, it could be a (nearly) pure column of As, Cs, Gs, or Ts. A natural approach is to use a mixture of Dirichlet priors, each one of which is “tilted” towards the appropriate corner of the 4-dimensional simplex, e.g.,

$$p(\theta | Z_t = 1) = 1/4 \text{Dir}(\theta | (10, 1, 1, 1)) + \dots + 1/4 \text{Dir}(\theta | (1, 1, 1, 10)) \quad (5.75)$$

Since this is conjugate, we can easily compute  $p(N_t | Z_t)$ .

See (Brown et al. 1993) for an application of these ideas to a real bio-sequence problem.

已思考 26s

下面先逐句完整翻译，再给你基础且详尽的解释。

## 逐句翻译（无遗漏）

图 5.10

两个 Beta 分布的混合。图由 `mixBetaDemo` 生成。

### 5.4.4.1 例子

假设我们使用如下混合先验

$$p(\theta) = 0.5 \text{Beta}(\theta | a_1, b_1) + 0.5 \text{Beta}(\theta | a_2, b_2) \quad (5.71)$$

其中  $a_1 = b_1 = 20$ , 且  $a_2 = b_2 = 10$ 。并且我们观测到  $N_1$  次正面、 $N_0$  次反面。则后验为

$$p(\theta | D) = p(Z = 1 | D) \text{Beta}(\theta | a_1 + N_1, b_1 + N_0) + p(Z = 2 | D) \text{Beta}(\theta | a_2 + N_1, b_2 + N_0)$$

如果  $N_1 = 20$  (正面) 且  $N_0 = 10$  (反面), 那么利用等式 (5.23), 后验变为

$$p(\theta | D) = 0.346 \text{Beta}(\theta | 40, 30) + 0.654 \text{Beta}(\theta | 50, 20) \quad (5.73)$$

见图 5.10 的示意。

### 5.4.4.2 应用：在 DNA 和蛋白质序列中寻找保守区域

我们提到 Dirichlet-多项式 (Dirichlet-multinomial) 模型在生物序列分析中被广泛使用。

下面给一个简单例子来说明已经发展出的一些方法。具体地，考虑第 2.3.2.1 节讨论的序列 logo。现在假设我们希望找到代表基因组编码区的位置。由于进化压力，这些位置常常在所有序列中都出现同一个字母。因此我们需要找到那些“纯” (pure) 或接近纯的列——意思是它们几乎全是 A, 或几乎全是 T, 或几乎全是 C, 或几乎全是 G。

一种做法是寻找熵很低的列；这类列的分布接近确定（纯）。不过，假如我们想给“纯度”的估计配一个置信度，这在我们相信相邻位置会一起被保守时就很有用。在这种情况下，我们可以令  $Z_t = 1$  表示位置  $t$  是保守的，令  $Z_t = 0$  表示不是。随后可以用马尔可夫链在相邻的  $Z_t$  之间加入依赖；细节见第 17 章。

无论如何，我们需要定义一个似然模型  $p(N_t | Z_t)$ ，其中  $N_t$  是列  $t$  的 (A,C,G,T) 计数向量。自然地，可令其为参数  $\theta_t$  的多项式分布。由于每一列的分布不同，我们希望把  $\theta_t$  积分掉，从而计算边际似然

$$p(N_t | Z_t) = \int p(N_t | \theta_t) p(\theta_t | Z_t) d\theta_t \quad (5.74)$$

但  $\theta_t$  的先验该怎么选？当  $Z_t = 0$  时可以用均匀先验  $p(\theta | Z_t = 0) = \text{Dir}(1, 1, 1, 1)$ ；当  $Z_t = 1$  时该用什么？毕竟，若该列是保守的，它可能是（近乎）全 A、全 C、全 G 或全 T。一个自然的做法是使用 Dirichlet 先验的混合，每个先验都朝 4 维单纯形的一个角“倾斜”，例如

$$p(\theta | Z_t = 1) = \frac{1}{4} \text{Dir}(\theta | (10, 1, 1, 1)) + \dots + \frac{1}{4} \text{Dir}(\theta | (1, 1, 1, 10)) \quad (5.75)$$

由于这是共轭的，我们可以容易地计算  $p(N_t | Z_t)$ 。关于这些思想在真实生物序列问题上的应用，见 (Brown 等, 1993)。

---

## 通俗解释（一步步来）

### 1) 为什么用“共轭先验的混合”

- **单个共轭先验**: 计算方便，但形状死板，表达不了“要么公平、要么偏向正面”这种**多峰/分叉**的信念。
- **稳健先验**: 形状灵活，但计算常常重。
- **混合共轭先验**: 把多个共轭先验按权重相加  $\Rightarrow$  仍然“好算”，又能做出多峰、偏斜等复杂形状——两全其美。

### 2) 硬币例子怎么读

- 先验是两条 Beta:
  - Beta(20,20): 集中在 0.5 (公平硬币)
  - Beta(10,10): 更宽、更分散 (文中参数写  $a_2 = b_2 = 10$ ; 在图示中也常见用一条偏向正面的分布来表达“偏向 heads”的可能性)
- 观测  $N_1 = 20, N_0 = 10$  (正面占 2/3)。
- **更新后**:
  - 每个分量都按 Beta-二项式共轭规则更新: 形参  $a \rightarrow a + N_1, b \rightarrow b + N_0$ 。
  - 同时更新**混合权重**  $p(Z = k | D)$  (用等式 5.70, 权重  $\propto$  先验权重  $\times$  该分量的边际似然)。
- 结果 (5.73) : 后验是两条新 Beta 的**加权和**，权重分别为 0.346 和 0.654。直觉上，因为数据更“偏正面”，所以“偏正面”的那一条分量权重更大。
- 图 5.10 里: 红色虚线是先验的混合形状，蓝色实线是观测数据后得到的后验混合；你会看到峰从先验的形状向**数据支持的区域移动并变尖**。

小结：混合先验 = “我们事先认为有两种可能的世界”，后验会根据数据把权重自动**更偏向更能解释数据的那一种**。

### 3) DNA/蛋白序列的应用（找“保守列”）

- 每一列  $t$  有一个 4 维概率  $\theta_t = (p_A, p_C, p_G, p_T)$ , 计数  $N_t = (n_A, n_C, n_G, n_T)$  服从多项式分布。
- 若该列不保守 ( $Z_t = 0$ ) , 就用均匀 Dirichlet:  $\text{Dir}(1,1,1,1)$ 。
- 若该列保守 ( $Z_t = 1$ ) , 它可能几乎全 A、或几乎全 C、或 G、或 T —— 于是用 4 个偏向角落的 Dirichlet 的平均 (式 5.75) , 比如  $\text{Dir}(10,1,1,1)$  偏向 A,  $\text{Dir}(1,10,1,1)$  偏向 C, 等等。
- 因为 Dirichlet 与多项式是共轭的,  $\theta_t$  可以被积分掉 (式 5.74) , 闭式得到边际似然  $p(N_t | Z_t)$ 。
- 有了  $p(N_t | Z_t)$  (以及可选的相邻  $Z$  间马尔可夫依赖) , 就能算出每列是保守的概率  $p(Z_t = 1 | \text{数据})$ , 从而标出保守区域。

#### 4) 你需要记住的三句话

1. 多个共轭先验的混合本身仍是共轭, 后验也是“混合”的形式 (式 5.69) 。
2. 后验中的每个分量按原来的共轭规则更新, 分量权重按“先验权重  $\times$  分量边际似然”再归一化得到 (式 5.70) 。
3. 这种做法既好算又灵活: 能表达“要么这样要么那样”的不确定结构 (多峰) , 非常适合像“公平或偏置硬币”“纯或不纯的列”这类问题。

如果你想, 我可以把上述硬币例子的权重计算步骤用数字演示出来 (含边际似然的数值) , 你就能看到 0.346 与 0.654 是怎么来的, 以及后验是怎样“更信偏正面”的那一分量。



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 5.11 (a) Results of fitting the model using the data from (Johnson and Albert 1999, p24). First row: Number of cancer incidents  $x_i$  in 20 cities in Missouri. Second row: population size  $N_i$ . The largest city (number 15) has a population of  $N_{15} = 53637$  and  $x_{15} = 54$  incidents, but we truncate the vertical axes of the first two rows so that the differences between the other cities are visible. Third row: MLE  $\hat{\theta}_i$ . The red line is the pooled MLE. Fourth row: posterior mean  $E[\theta_i | D]$ . The red line is  $E[a/(a+b) | D]$ , the population-level mean. (b) Posterior 95% credible intervals

on the cancer rates. Figure generated by cancerRatesEb

## 5.5 Hierarchical Bayes

A key requirement for computing the posterior  $p(\theta|D)$  is the specification of a prior  $p(\theta|\eta)$ , where  $\eta$  are the hyperparameters. What if we don't know how to set  $\eta$ ? In some cases, we can use uninformative priors, we we discussed above. A more Bayesian approach is to put a prior on our priors! In terms of graphical models (Chapter 10), we can represent the situation as follows:

$$\eta \rightarrow \theta \rightarrow D \quad (5.76)$$

This is an example of a hierarchical Bayesian model, also called a multi-level model, since there are multiple levels of unknown quantities. We give a simple example below, and we will see many others later in the book.

### 5.5.1 Example: modeling related cancer rates

Consider the problem of predicting cancer rates in various cities (this example is from (Johnson and Albert 1999, p24)). In particular, suppose we measure the number of people in various cities,  $N_i$ , and the number of people who died of cancer in these cities,  $x_i$ . We assume  $x_i \sim \text{Bin}(N_i, \theta_i)$ , and we want to estimate the cancer rates  $\theta_i$ .

One approach is to estimate them all separately, but this will suffer from the sparse data problem (underestimation of the rate of cancer due to small  $N_i$ ). Another approach is to assume all the  $\theta_i$  are the same; this is called parameter tying. The resulting pooled MLE is just  $\hat{\theta} = \sum i x_i / \sum i N_i$ . But the assumption that all the cities have the same rate is a rather strong one. A compromise approach is to assume that the  $\theta_i$  are similar, but that there may be city-specific variations. This can be modeled by assuming the  $\theta_i$  are drawn from some common distribution, say  $\theta_i \sim \text{Beta}(a, b)$ . The full joint distribution can be written as

$$p(D, \theta, \eta|N) = p(\eta) \prod_{i=1}^N \text{Bin}(x_i|N_i, \theta_i) \text{Beta}(\theta_i|\eta) \quad (5.77)$$

where  $\eta = (a, b)$ . Note that it is crucial that we infer  $\eta = (a, b)$  from the data; if we just clamp it to a constant, the  $\theta_i$  will be conditionally independent, and there will be no information flow between them. By contrast, by treating  $\eta$  as an unknown (hidden variable), we allow the data-poor cities to borrow statistical strength from data-rich ones.

Suppose we compute the joint posterior  $p(\eta, \theta|D)$ . From this we can get the posterior marginals  $p(\theta_i|D)$ . In Figure 5.11(a), we plot the posterior means,  $E[\theta_i|D]$ , as blue bars, as well as the population level mean,  $E[a/(a+b)|D]$ , shown as a red line (this represents the average of the  $\theta_i$ 's). We see that the posterior mean is shrunk towards the pooled estimate more strongly for cities with small

sample sizes  $N_i$ . For example, city 1 and city 20 both have a 0 observed cancer incidence rate, but city 20 has a smaller population, so its rate is shrunk more towards the population-level estimate (i.e., it is closer to the horizontal red line) than city 1. Figure 5.11(b) shows the 95% posterior credible intervals for  $\theta_i$ . We see that city 15, which has a very large population (53,637 people), has small posterior uncertainty. Consequently this city has the largest impact on the posterior estimate of  $\eta$ , which in turn will impact the estimate of the cancer rates for other cities. Cities 10 and 19, which have the highest MLE, also have the highest posterior uncertainty, reflecting the fact that such a high estimate is in conflict with the prior (which is estimated from all the other cities).

In the above example, we have one parameter per city, modeling the probability the response is on. By making the Bernoulli rate parameter be a function of covariates,  $\theta_i = \text{sigm}(w^T i x)$ , we can model multiple correlated logistic regression tasks. This is called multi-task learning, and will be discussed in more detail in Section 9.5.

我先给你逐句无遗漏翻译，然后再用基础且详尽的方式解释图和文字内容，让你能看懂每一个细节，即使数学基础不强也能跟上。

## 逐句翻译

图 5.11

(a) 使用 (Johnson 和 Albert 1999, p24) 中数据拟合模型的结果。

第一行：密苏里州 20 个城市的癌症病例数  $x_i$ 。

第二行：城市人口数  $N_i$ 。人口最多的城市（编号 15）人口为  $N_{15} = 53,637$ ，癌症病例  $x_{15} = 54$ ，但为了让其他城市之间的差异看得更清楚，我们把前两行的纵轴做了截断。

第三行：最大似然估计（MLE） $\hat{\theta}_i$ 。红线表示“合并”后的 MLE。

第四行：后验均值  $E[\theta_i | D]$ 。红线表示  $E[\frac{a}{a+b} | D]$ ，即总体（population-level）均值。

(b) 癌症发病率的 95% 后验可信区间。图中 \* 表示中位数。

图由 `cancerRatesEb` 生成。

## 5.5 分层贝叶斯 (Hierarchical Bayes)

在计算后验  $p(\theta | D)$  时，一个关键需求是我们必须指定先验  $p(\theta | \eta)$ ，其中  $\eta$  是超参数（hyper-parameters）。那如果我们不知道该怎么设  $\eta$  呢？在某些情况下，我们可以使用

**无信息先验** (uninformative prior) , 前面已经讨论过。但更“贝叶斯化”的方法是：**给我们的先验再加一个先验！**

在图模型（第 10 章）的形式下，可以表示为：

$$\eta \rightarrow \theta \rightarrow D \quad (5.76)$$

这是一个分层贝叶斯模型 (hierarchical Bayesian model) 的例子，也叫多层次模型 (multi-level model) , 因为它有多个层次的未知量。我们下面给一个简单例子，书中后面还会看到更多。

---

### 5.5.1 例子：建模相关的癌症发病率

考虑预测多个城市癌症发病率的问题（例子来自 Johnson 和 Albert 1999, p24）。

假设我们知道每个城市的人口数  $N_i$ , 以及该城市死于癌症的人数  $x_i$ 。我们假设：

$$x_i \sim \text{Bin}(N_i, \theta_i)$$

并希望估计癌症发病率  $\theta_i$ 。

- **方法 1：分别估计每个城市的  $\theta_i$**

问题：小城市样本量少，估计会很不稳定（稀疏数据问题）。

- **方法 2：假设所有  $\theta_i$  完全相同** (参数绑定 parameter tying)

这样得到的合并 MLE 是：

$$\hat{\theta} = \frac{\sum_i x_i}{\sum_i N_i}$$

问题：假设太强，不现实。

- **折中方法：**假设  $\theta_i$  彼此相似，但存在城市特有的差异。

做法：假设  $\theta_i$  来自同一个分布，比如

$$\theta_i \sim \text{Beta}(a, b)$$

完整联合分布写作：

$$p(D, \theta, \eta | N) = p(\eta) \prod_{i=1}^N \text{Bin}(x_i | N_i, \theta_i) \text{Beta}(\theta_i | \eta) \quad (5.77)$$

其中  $\eta = (a, b)$ 。

关键：**必须从数据中推断**  $\eta = (a, b)$ 。如果把它们固定为常数， $\theta_i$  在条件上相互独立，不会相互“传递信息”。

但如果把  $\eta$  当作未知（隐藏变量）来推断，数据少的城市可以**借用**数据多的城市的统计信息 (statistical strength) 。

假设我们算出了联合后验  $p(\eta, \theta | D)$ ，就能得到每个城市的边际后验：

$$p(\theta_i | D)$$

在图 5.11(a) 中，我们画出了：

- 蓝色柱子：后验均值  $E[\theta_i | D]$
- 红线：总体均值  $E[a/(a+b) | D]$ ，表示所有  $\theta_i$  的平均值。

我们看到，小样本的城市，其后验均值会**更强地被拉向** (shrunk towards) 总体估计值。

例如：

- 城市 1 和城市 20 都观测到癌症率为 0 (没有病例)。
- 但城市 20 人口更少，所以它的估计值**更靠近红线**，也就是总体均值。

图 5.11(b) 画出了每个城市的  $\theta_i$  的 95% **后验可信区间**：

- 城市 15 人口最多 (53,637)，因此不确定性最小，对  $\eta$  的后验影响最大，也就影响了其他城市的癌症率估计。
- 城市 10 和城市 19 虽然 MLE 最高，但它们的后验不确定性也最高，因为这么高的估计值与从所有其他城市估计出的先验分布冲突较大。

在上面的例子中，我们为每个城市建了一个参数，表示该城市发生癌症的概率。

如果把伯努利分布的参数改成与协变量的函数：

$$\theta_i = \sigma(w^\top x_i)$$

我们就能建模多个相关的逻辑回归任务，这叫**多任务学习 (multi-task learning)**，详见第 9.5 节。

---

## 通俗解释

### 1. 图 (a) 左边 4 行的含义

- 第一行：每个城市的癌症病例数（但把纵轴截断到 5，看得更清楚）。
- 第二行：每个城市的人口数（纵轴截断到 2000，让差异可见）。
- 第三行：用每个城市单独的数据做最大似然估计 (MLE)，再乘以 1000，红线是“合并所有城市数据”得到的一个整体 MLE。
- 第四行：分层贝叶斯估计得到的后验均值（乘以 1000），红线是总体均值。

#### 关键观察：

小城市的估计值更接近红线（总体均值）→ 这就是“借力”大城市数据的效果（统计学叫“收缩” shrinkage）。

### 2. 图 (b) 右边的含义

- 横轴：癌症发病率 ( $\times 10^{-3}$ )
- 每一行的星号是中位数，横线是 95% 可信区间。
- 大城市（样本量多）可信区间短（估计准），小城市可信区间长（估计不准）。
- 数据极端的城市（MLE 高或低）可信区间也会长，因为和整体推断出的先验冲突。

### 3. 分层模型的优势

- 单独估计：小样本会导致结果极端（比如小城市可能恰好没病例 → 估计为 0）。
- 合并估计：忽略城市差异，过于简单。
- 分层模型：既考虑城市差异，又让小样本向整体均值收缩，估计更稳健。

### 4. 为什么叫“借力”

- 把  $\eta = (a, b)$  当作未知，从所有城市的数据共同推断出来。
- 小城市数据少时，“信不过”它自己的 MLE，就让它往整体推的方向走。
- 这种思想不仅用在癌症率上，任何“多个相关比例/概率”的问题都可以用。

如果你愿意，我可以帮你用一个简单的数字例子算出这种“收缩”是如何发生的，比如用 2 个城市数据对比单独估计 vs 分层估计，你会看到小城市的比例被拉向总体均值的过程。这样你就能直观看懂图 5.11 的红线是怎么起作用的。



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。

#### 5.6 Empirical Bayes

In hierarchical Bayesian models, we need to compute the posterior on multiple levels of latent variables. For example, in a two-level model, we need to compute  $p(\eta, \theta|D) \propto p(D|\theta)p(\theta|\eta)p(\eta)$  (5.78)

In some cases, we can analytically marginalize out  $\theta$ ; this leaves us with the simpler problem of just computing  $p(\eta|D)$ . As a computational shortcut, we can approximate the posterior on the hyper-parameters with a point-estimate,  $p(\eta|D) \approx \delta^{\hat{\eta}}(\eta)$ , where  $\hat{\eta} = \text{argmax } p(\eta|D)$ . Since  $\eta$  is typically much smaller than  $\theta$  in dimensionality, it is less prone to overfitting, so we can safely use a uniform prior on  $\eta$ . Then the estimate becomes

$$\hat{\eta} = \text{argmax } p(D|\eta) = \text{argmax } p(D|\theta)p(\theta|\eta)d\theta \quad (5.79)$$

where the quantity inside the brackets is the marginal or integrated likelihood, sometimes called the evidence. This overall approach is called empirical Bayes (EB) or type-II maximum likelihood. In machine learning, it is sometimes called the evidence procedure. Empirical Bayes violates the principle that the prior should be chosen

independently of the data. However, we can just view it as a computationally cheap approximation to inference in a hierarchical Bayesian model, just as we viewed MAP estimation as an approximation to inference in the one level model  $\theta \rightarrow D$ . In fact, we can construct a hierarchy in which the more integrals one performs, the “more Bayesian” one becomes:

Note that EB can be shown to have good frequentist properties (see e.g., (Carlin and Louis 1996; Efron 2010)), so it is widely used by non-Bayesians. For example, the popular James-Stein estimator, discussed in Section 6.3.3.2, can be derived using EB.

已思考 22s

下面先把原文逐句完整翻译，再给你做一份基础且详尽的解释；我也把图里的方法对照表翻成中文放在开头，方便你对照理解。

## 图中方法与定义（译文）

方法	定义
最大似然 (Maximum likelihood)	$\hat{\theta} = \arg \max_{\theta} p(D   \theta)$
MAP 估计	$\hat{\theta} = \arg \max_{\theta} p(D   \theta) p(\theta   \eta)$
ML-II (经验贝叶斯, Empirical Bayes)	$\hat{\eta} = \arg \max_{\eta} \int p(D   \theta) p(\theta   \eta) d\theta = \arg \max_{\eta} p(D   \eta)$
MAP-II	$\hat{\eta} = \arg \max_{\eta} \int p(D   \theta) p(\theta   \eta) p(\eta) d\theta = \arg \max_{\eta} p(D   \eta) p(\eta)$
完全贝叶斯 (Full Bayes)	$p(\theta, \eta   D) \propto p(D   \theta) p(\theta   \eta) p(\eta)$

## 5.6 经验贝叶斯（原文翻译）

在层级贝叶斯模型中，我们需要为多个层次的隐变量计算后验分布。

例如，在一个两层模型里，我们需要计算

$$p(\eta, \theta | D) \propto p(D | \theta) p(\theta | \eta) p(\eta) \quad (5.78)$$

在某些情况下，我们可以把  $\theta$  积分掉（边缘化）；这样问题就简化为只需计算  $p(\eta | D)$ 。

作为一种计算上的捷径，我们可以用点估计来近似超参数的后验：

$$p(\eta | D) \approx \delta_{\hat{\eta}}(\eta), \quad \hat{\eta} = \arg \max_{\eta} p(\eta | D)$$

由于  $\eta$  的维度通常比  $\theta$  小得多，它更不容易过拟合，因此我们可以对  $\eta$  安全地使用一个均匀先验。

于是估计变为

$$\hat{\eta} = \arg \max_{\eta} p(D \mid \eta) = \arg \max_{\eta} \int p(D \mid \theta) p(\theta \mid \eta) d\theta \quad (5.79)$$

其中方括号里的量就是**边际（或积分）似然**，有时也叫**证据**（evidence）。

这种整体方法称为**经验贝叶斯（EB）或二型最大似然（type-II ML）**；在机器学习中也叫**证据程序（evidence procedure）**。

经验贝叶斯违背了“先验应当独立于数据选择”的原则。不过，我们可以把它看作是对层级贝叶斯推断的一种**计算便宜的近似**，正如我们把 MAP 估计看作单层模型  $\theta \rightarrow D$  上贝叶斯推断的近似一样。事实上，我们可以构造一个层级：**做的积分越多，就越“贝叶斯”**（见上表）。

需要注意的是，EB 被证明具有良好的**频率学派性质**（例如参见 Carlin & Louis, 1996; Efron, 2010），所以非贝叶斯学者也广泛使用它。举例来说，第 6.3.3.2 节讨论的**James-Stein 估计量**就可以用 EB 推导出来。

## 通俗解释（一步步来）

### 1) 层级贝叶斯里都有谁？

- $\theta$ : 底层参数（比如每个城市的癌症率、回归系数等）
- $\eta$ : 超参数（控制先验的形状，如 Beta 的  $a, b$ , 或岭回归的  $\lambda$ ）
- $D$ : 数据

完全贝叶斯会对**两层都做推断**：先给  $\eta$  一个先验  $p(\eta)$ ，再根据数据更新得到  $p(\theta, \eta \mid D)$

。

### 2) 经验贝叶斯在做什么？

- 先把  $\theta$  **积分掉**，得到只关于  $\eta$  的“证据”  $p(D \mid \eta)$ 。
- 然后直接**最大化证据**来选一个最好的  $\hat{\eta}$ （这一步就叫 ML-II / EB）。
- 选好  $\hat{\eta}$  后，再在固定  $\hat{\eta}$  下做常规的贝叶斯或 MAP 推断  $\theta$ 。

直观地： $\hat{\eta}$  是让“模型在整合所有可能  $\theta$  后，对当前数据最可信”的那组超参数。

### 3) 与其他方法的区别（配合上表记忆）

- **MLE**: 只选  $\hat{\theta}$ ，不考虑先验。
- **MAP**: 有固定的  $\eta$ ，选  $\hat{\theta}$  来最大化“似然  $\times$  先验”。
- **ML-II/EB**: 先选  $\hat{\eta} = \arg \max_{\eta} p(D \mid \eta)$ ，再据此估计  $\theta$ 。
- **MAP-II**: 像 EB，但给  $\eta$  也配一个先验  $p(\eta)$ ，最大化  $p(D \mid \eta)p(\eta)$ 。
- **Full Bayes**: 不给任何量做点估计，直接得到  $p(\theta, \eta \mid D)$  的完整后验。

理解口诀：做点估计越多 → 越便宜越近似；积分越多 → 越“贝叶斯”越全面。

#### 4) 什么时候 EB 很香？

- $\eta$  的维度很小（一个或几个超参数），而  $\theta$  很大（成百上千个参数）。
- 典型例子：
  - 岭回归/高斯过程里选择正则化强度  $\lambda$ （最大化  $\log p(D | \lambda)$ ）。
  - 主题模型、层级计数模型里选择 Dirichlet 超参数。
  - 多城市/多组别的分层 Beta-Binomial 模型里选择  $a, b$ 。

优点：比交叉验证便宜很多；把“调参”变成了有概率意义的证据最大化。

#### 5) EB 的利与弊

##### 优点

- 计算高效；
- 自动化地设定超参数；
- 在很多场景有不错的频率学派性质（如收缩估计、James–Stein）。

##### 局限

- 把  $\eta$  当点估计会低估不确定性（过于自信）；
- 某种意义上“用数据两次”（先选  $\hat{\eta}$  再估  $\theta$ ），在样本很小或模型错配时可能偏差更大；
- 对于多峰证据面，可能选到局部极值。

#### 6) 一个极简心智图

- 证据 (evidence) :  $p(D | \eta) = \int p(D | \theta) p(\theta | \eta) d\theta$
- EB/ML-II:  $\hat{\eta} = \arg \max_{\eta} p(D | \eta)$
- 之后：用  $\hat{\eta}$  做 MAP / Bayes 推断  $\theta$

如果你想，我可以举一个小例子（比如岭回归选  $\lambda$ ，或Beta–Binomial 选  $a, b$ ）把  $\log p(D | \eta)$  的计算和最大化步骤写成具体数字，让你看到 EB 是怎么“省去交叉验证”的。

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。

##### 5.6.1 Example: beta-binomial model

Let us return to the cancer rates model. We can analytically integrate out the  $\theta_i$ 's, and write down the marginal likelihood directly, as follows:

$$p(D|a, b) = \prod_i \text{Bin}(x_i|N_i, \theta_i) \text{Beta}(\theta_i|a, b) d\theta_i \quad (5.80)$$

$$= i B(a + xi, b + Ni - xi) B(a, b) \quad (5.81)$$

Various ways of maximizing this wrt a and b are discussed in (Minka 2000e).

Having estimated a and b, we can plug in the hyper-parameters to compute the posterior  $p(\theta_i | a, b, D)$  in the usual way, using conjugate analysis. The net result is that the posterior mean of each  $\theta_i$  is a weighted average of its local MLE and the prior means, which depends on  $\eta = (a, b)$ ; but since  $\eta$  is estimated based on all the data, each  $\theta_i$  is influenced by all the data.

### 5.6.2 Example: Gaussian-Gaussian model

We now study another example that is analogous to the cancer rates example, except the data is real-valued. We will use a Gaussian likelihood and a Gaussian prior. This will allow us to write down the solution analytically. In particular, suppose we have data from multiple related groups. For example,  $x_{ij}$  could be the test score for student  $i$  in school  $j$ , for  $j = 1: D$  and  $i = 1: N_j$ . We want to estimate the mean score for each school,  $\theta_j$ . However, since the sample size,  $N_j$ , may be small for some schools, we can regularize the problem by using a hierarchical Bayesian model, where we assume  $\theta_j$  come from a common prior,  $N(\mu, \tau^2)$ . The joint distribution has the following form:

$$p(\theta, D | \eta, \sigma^2) = \prod_{j=1}^D \prod_{i=1}^{N_j} N(x_{ij} | \theta_j, \sigma^2) \quad (5.82)$$

where we assume  $\sigma^2$  is known for simplicity. (We relax this assumption in Exercise 24.4.) We explain how to estimate  $\eta$  below. Once we have estimated  $\eta = (\mu, \tau)$ , we can compute the posteriors over the  $\theta_j$ 's. To do that, it simplifies matters to rewrite the joint distribution in the following form, exploiting the fact that  $N_j$  Gaussian measurements with values  $x_{ij}$  and variance  $\sigma^2$  are equivalent to one measurement of value  $x_j = \frac{1}{N_j} \sum_{i=1}^{N_j} x_{ij}$  with variance  $\sigma^2_j = \sigma^2/N_j$ . This yields

$$p(\theta, D | \eta, \sigma^2) = \prod_{j=1}^D N(\theta_j | \mu, \tau^2) N(x_j | \theta_j, \sigma^2_j) \quad (5.83)$$

From this, it follows from the results of Section 4.4.1 that the posteriors are given by

$$p(\theta_j | D, \mu, \tau^2) = N(\theta_j | \hat{\mu}_j, \hat{\tau}^2_j) \quad (5.84)$$

$$\hat{\mu}_j = \sigma^2_j \bar{x}_j + \hat{\tau}^2_j \mu$$

where  $\hat{\mu}_j = \bar{x}_j$  and  $\hat{\tau}^2_j$  will be defined below. The quantity  $0 \leq \hat{\tau}^2_j \leq 1$  controls the degree of shrinkage towards the overall mean,  $\mu$ . If the data is reliable for group  $j$  (e.g., because the sample size  $N_j$  is large), then  $\sigma^2_j$  will be small relative to  $\tau^2$ ; hence  $\hat{\tau}^2_j$  will be small, and we will put more weight on  $x_j$  when we estimate  $\theta_j$ . However, groups with small sample sizes will get regularized (shrunk)

towards the overall mean  $\mu$ ) more heavily. We will see an example of this below. If  $\sigma_j = \sigma$  for all groups  $j$ , the posterior mean becomes

$$\hat{\theta}_j = \hat{B}x + (1 - \hat{B})\bar{x} = x + (1 - \hat{B})(\bar{x} - x) \quad (5.86)$$

This has exactly the same form as the James Stein estimator discussed in Section 6.3.3.2.

好的，我会逐句完整翻译你给的原文，然后用**非常基础且详尽的方式解释**里面的数学逻辑和直观意义。

我们分成两个小节来看 (Beta–Binomial 和 Gaussian–Gaussian) 。

---

## 原文翻译

### 5.6.1 例子：Beta–Binomial 模型

让我们回到癌症发病率模型。我们可以把每个  $\theta_i$  积分掉（解析地边缘化），并直接写出边际似然 (marginal likelihood) 如下：

$$p(D | a, b) = \prod_i \int \text{Bin}(x_i | N_i, \theta_i) \text{Beta}(\theta_i | a, b) d\theta_i \quad (5.80)$$

$$= \prod_i \frac{B(a + x_i, b + N_i - x_i)}{B(a, b)} \quad (5.81)$$

这里  $B(\cdot, \cdot)$  是 Beta 函数。如何对  $a$  和  $b$  最大化边际似然的各种方法，可参见 Minka (2000e)。

在估计出  $a$  和  $b$  之后，我们可以把这些超参数代入，用共轭分析 (conjugate analysis) 来按常规方法计算后验  $p(\theta_i | a, b, D)$ 。最终结果是：每个  $\theta_i$  的后验均值是它的**局部最大似然估计** (local MLE) 和**先验均值**的加权平均，这个权重取决于  $\eta = (a, b)$ ；而因为  $\eta$  是用**所有数据**估计的，所以每个  $\theta_i$  都会受到全部数据的影响。

---

### 5.6.2 例子：Gaussian–Gaussian 模型

现在我们看另一个例子，它和癌症发病率例子类似，只是数据是实数值。我们会用高斯似然 (Gaussian likelihood) 和高斯先验 (Gaussian prior)。这样我们可以直接写出解析解。

假设我们有多个相关群体 (groups) 的数据。例如， $x_{ij}$  可能是第  $j$  所学校第  $i$  个学生的考试分数，其中  $j = 1 : D$ ,  $i = 1 : N_j$ 。我们要估计每所学校的平均分  $\theta_j$ 。但是因为某些学校的样本量  $N_j$  很小，我们可以用**层级贝叶斯模型**来正则化 (regularize) 问题：假设  $\theta_j$  来自一个公共的先验  $N(\mu, \tau^2)$ 。

联合分布为：

$$p(\theta, D \mid \eta, \sigma^2) = \prod_{j=1}^D N(\theta_j \mid \mu, \tau^2) \prod_{i=1}^{N_j} N(x_{ij} \mid \theta_j, \sigma^2) \quad (5.82)$$

这里为了简单起见，我们假设  $\sigma^2$  已知（在练习 24.4 中放宽这一假设）。

下面我们解释如何估计  $\eta$ 。一旦我们估计出  $\eta = (\mu, \tau)$ ，就可以计算每个  $\theta_j$  的后验。为了方便推导，我们注意到：

$N_j$  个方差为  $\sigma^2$  的高斯测量，其均值是

$$x_j = \frac{1}{N_j} \sum_{i=1}^{N_j} x_{ij}$$

方差为

$$\sigma_j^2 = \frac{\sigma^2}{N_j}$$

所以可以等价看作一次测量。于是联合分布可写成：

$$p(\theta, D \mid \hat{\eta}, \sigma^2) = \prod_{j=1}^D N(\theta_j \mid \hat{\mu}, \hat{\tau}^2) N(x_j \mid \theta_j, \sigma_j^2) \quad (5.83)$$

由第 4.4.1 节的结果可得：

$$p(\theta_j \mid D, \hat{\mu}, \hat{\tau}^2) = N(\theta_j \mid \hat{B}_j \hat{\mu} + (1 - \hat{B}_j)x_j, (1 - \hat{B}_j)\sigma_j^2) \quad (5.84)$$

$$\hat{B}_j = \frac{\sigma_j^2}{\sigma_j^2 + \hat{\tau}^2} \quad (5.85)$$

其中  $\hat{\mu} = \bar{x}$ （总体均值），而  $\hat{\tau}^2$  会在后面定义。

$\hat{B}_j$  控制了\*\*向总体均值  $\mu$  收缩（shrinkage）\*\*的程度：

- 如果某个群体的数据很可靠（例如  $N_j$  大），则  $\sigma_j^2$  相对  $\hat{\tau}^2$  很小， $\hat{B}_j$  会小，此时估计  $\theta_j$  时会更多依赖  $x_j$ 。
- 如果样本很少，则会被更强烈地向总体均值  $\mu$  收缩。

如果所有群体的  $\sigma_j = \sigma$ ，那么后验均值为：

$$\hat{\theta}_j = \hat{B}\bar{x} + (1 - \hat{B})x_j = \bar{x} + (1 - \hat{B})(x_j - \bar{x}) \quad (5.86)$$

这与第 6.3.3.2 节讨论的 James–Stein 估计量形式完全相同。

## 基础详解与直观理解

### 1) Beta–Binomial 模型（癌症率例子）

- 每个地区  $i$  有一个真实癌症率  $\theta_i$ （概率）。
- 我们在该地区看了  $N_i$  人，发现  $x_i$  人得病，数据服从二项分布  $\text{Bin}(x_i \mid N_i, \theta_i)$ 。

- 我们的先验是  $\theta_i \sim \text{Beta}(a, b)$ , 且所有地区共享同一组  $(a, b)$ 。

### EB 的做法:

1. 把所有  $\theta_i$  积分掉, 得到只关于  $a, b$  的边际似然 (式 5.80–5.81)。
2. 最大化这个边际似然, 找到最佳  $\hat{a}, \hat{b}$ 。
3. 再用共轭更新公式得到每个  $\theta_i$  的后验。

### 效果:

- 每个  $\theta_i$  的估计值是“本地比例”和“整体均值”的加权平均。
  - 因为整体均值是用所有地区的数据估出来的, 所以即便某地区数据很少, 也会被全国情况“拉一把”, 避免极端值。
- 

## 2) Gaussian–Gaussian 模型 (学校分数例子)

- $x_{ij}$ : 第  $j$  所学校第  $i$  个学生分数。
- $\theta_j$ : 学校  $j$  的真实平均分。
- 假设  $\theta_j \sim N(\mu, \tau^2)$ , 所有学校共享  $\mu, \tau$ 。
- 单个学生成绩测量噪声方差  $\sigma^2$ 。

### EB 流程:

1. 把每所学校的平均成绩  $x_j$ , 方差是  $\sigma_j^2 = \sigma^2/N_j$ 。
2. 估计  $\mu, \tau$  (经验贝叶斯)。
3. 后验均值  $E[\theta_j | D]$  是:

$$\hat{B}_j\mu + (1 - \hat{B}_j)x_j$$

其中  $\hat{B}_j$  越小, 越信任学校自己的数据;  $\hat{B}_j$  越大, 越收缩到总体均值。

### 直觉类比:

- 数据多且稳定  $\rightarrow$  “我相信你学校自己的均值”  $\rightarrow$  少收缩。
  - 数据少且噪声大  $\rightarrow$  “我不太信你这个学校的均值”  $\rightarrow$  多收缩到全体均值。
- 

## 3) 和 James–Stein 的关系

- James–Stein 估计是一个著名的“收缩估计”方法: 在多维均值估计中, 把各个样本均值向总体均值收缩能降低总体误差。
  - Gaussian–Gaussian 模型用 EB 得到的公式, 恰好和 James–Stein 的形式一样。
- 

如果你愿意, 我可以帮你画一张图, 把\*\*EB 在 Beta–Binomial 和 Gaussian–Gaussian 例子中的“局部均值 + 收缩”\*\*过程可视化, 让你一眼看到 EB 是如何利用全部数据改进小样本群体估计的。这样会比公式更直观。

你要我帮你画吗?

“EB 的做法：“中的1和3是如何实现的

好，我给你拆开讲 EB 的做法 里 第 1 步（积分掉  $\theta_i$ ）和 第 3 步（计算后验）是怎么一步一步做到的，用最基础的思路走一遍。

## 第 1 步：积分掉 $\theta_i$

目标是从

$$p(D | a, b) = \prod_i \underbrace{\text{Bin}(x_i | N_i, \theta_i)}_{\text{似然}} \underbrace{\text{Beta}(\theta_i | a, b)}_{\text{先验}} d\theta_i$$

得到式 (5.81)

$$= \prod_i \frac{B(a + x_i, b + N_i - x_i)}{B(a, b)}$$

怎么做：

### 1. 写出二项分布和 Beta 分布公式

- 二项分布：

$$\text{Bin}(x_i | N_i, \theta_i) = \binom{N_i}{x_i} \theta_i^{x_i} (1 - \theta_i)^{N_i - x_i}$$

- Beta 分布：

$$\text{Beta}(\theta_i | a, b) = \frac{\theta_i^{a-1} (1 - \theta_i)^{b-1}}{B(a, b)}$$

### 2. 相乘：

二项  $\times$  Beta =

$$\binom{N_i}{x_i} \frac{\theta_i^{x_i+a-1} (1 - \theta_i)^{N_i-x_i+b-1}}{B(a, b)}$$

### 3. 对 $\theta_i$ 积分（从 0 到 1）：

$$\int_0^1 \theta_i^{x_i+a-1} (1 - \theta_i)^{N_i-x_i+b-1} d\theta_i$$

这个积分正好就是 Beta 函数  $B(a + x_i, b + N_i - x_i)$  的定义。

### 4. 把常数项提出来，得到：

$$\frac{\binom{N_i}{x_i} B(a + x_i, b + N_i - x_i)}{B(a, b)}$$

5. 因为最大化边际似然时组合数  $\binom{N_i}{x_i}$  与  $a, b$  无关, 可以忽略掉, 剩下的就是书里的式子 (5.81)。

所以积分的结果用 Beta 函数公式直接化简了, 这就是“解析积分掉  $\theta_i$ ”的含义。

---

## 第 3 步：计算后验

我们已经得到了最佳的  $\hat{a}, \hat{b}$ , 接下来要算每个  $\theta_i$  的后验:

$$p(\theta_i | x_i, N_i, \hat{a}, \hat{b}) \propto \text{Bin}(x_i | N_i, \theta_i) \text{Beta}(\theta_i | \hat{a}, \hat{b})$$

**怎么做:**

1. **写出公式** (跟上面第 1 步相似) :

$$p(\theta_i | \dots) \propto \theta_i^{x_i + \hat{a} - 1} (1 - \theta_i)^{N_i - x_i + \hat{b} - 1}$$

2. **认出来这就是 Beta 分布的形式:**

所以:

$$p(\theta_i | \dots) = \text{Beta}(\theta_i | \hat{a} + x_i, \hat{b} + N_i - x_i)$$

3. **后验均值直接用 Beta 分布的均值公式:**

$$\text{E}[\theta_i | \dots] = \frac{\hat{a} + x_i}{\hat{a} + \hat{b} + N_i}$$

4. **解释:**

这个均值等于:

$$\frac{N_i}{\hat{a} + \hat{b} + N_i} \times \frac{x_i}{N_i} + \frac{\hat{a} + \hat{b}}{\hat{a} + \hat{b} + N_i} \times \frac{\hat{a}}{\hat{a} + \hat{b}}$$

前半部分是**本地样本比例**, 后半部分是**先验均值**, 两个加权平均, 权重取决于样本量  $N_i$ 。

---

## 小结

- 第 1 步就是用 Beta-Binomial 共轭性, 把积分直接算成 Beta 函数的比值。
  - 第 3 步就是用共轭更新公式直接得到 Beta 分布后验, 再用均值公式得出结果。
- 

如果你愿意, 我可以帮你画一个**从局部比例到全局均值的收缩曲线图**, 你会看到  $N_i$  越小, 结果越被拉向全局均值, 这样 EB 的直观就特别清楚。

你要我帮你画吗?



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 5.12 (a) MLE parameters (top) and corresponding shrunken estimates (bottom). (b) We plot the true parameters (blue), the posterior mean estimate (green), and the MLEs (red) for 5 of the players. Figure generated by shrinkageDemoBaseball.

#### 5.6.2.1 Example: predicting baseball scores

We now give an example of shrinkage applied to baseball batting averages, from (Efron and Morris 1975). We observe the number of hits for  $D = 18$  players during the first  $T = 45$  games. Call the number of hits  $b_i$ . We assume  $b_j \sim \text{Bin}(T, \theta_j)$ , where  $\theta_j$  is the “true” batting average for player  $j$ . The goal is to estimate the  $\theta_j$ . The MLE is of course  $\hat{\theta}_j = x_j$ , where  $x_j = b_j/T$  is the empirical batting average. However, we can use an EB approach to do better. To apply the Gaussian shrinkage approach described above, we require that the likelihood be Gaussian,  $x_j \sim N(\theta_j, \sigma^2)$  for known  $\sigma^2$ . (We drop the  $i$  subscript since we assume  $N_j = 1$ , since  $x_j$  already represents the average for player  $j$ .)

However, in this example we have a binomial likelihood. While this has the right mean,  $E[x_j] = \theta_j$ , the variance is not constant:

$$\text{var}[x_j] = 1/T \quad \text{var}[b_j] = T\theta_j(1 - \theta_j)/T^2 \quad (5.87)$$

So let us apply a variance stabilizing transform<sup>5</sup> to  $x_j$  to better match the Gaussian assumption:

$$y_j = f(y_j) = \sqrt{T} \arcsin(2y_j - 1) \quad (5.88)$$

Now we have approximately  $y_j \sim N(f(\theta_j), 1) = N(\mu_j, 1)$ . We use Gaussian shrinkage to estimate the  $\mu_j$  using Equation 5.86 with  $\sigma^2 = 1$ , and we then transform back to get

$$\hat{\theta}_j = 0.5(\sin(\hat{\mu}_j / \sqrt{T}) + 1) \quad (5.89)$$

The results are shown in Figure 5.12(a-b). In (a), we plot the MLE  $\hat{\theta}_j$  and the posterior mean  $\theta_j$ . We see that all the estimates have shrunk towards the global mean, 0.265. In (b), we plot the true value  $\theta_j$ , the MLE  $\hat{\theta}_j$  and the posterior mean  $\theta_j$ . (The “true” values of  $\theta_j$  are estimated

from a large number of independent games.) We see that, on average, the shrunken estimate is much closer to the true parameters than the MLE is. Specifically, the mean squared error, defined by  $MSE = \frac{1}{N} \sum_{j=1}^N (\hat{\theta}_j - \theta_j)^2$ , is over three times smaller using the shrinkage estimates  $\hat{\theta}_j$  than using the MLEs  $\hat{\theta}_j$ .

已思考 8s

下面先逐句完整翻译原文（无遗漏），然后给出基础且详尽的解释，帮你把做法与图都看明白。

---

## 逐句翻译

图 5.12

- (a) 最大似然估计 (MLE) 参数 (上方) 以及相应的收缩估计 (下方)。
- (b) 对 5 名球员绘制真实参数 (蓝色)、后验均值估计 (绿色) 以及 MLE (红色)。

图由 `shrinkageDemoBaseball` 生成。

### 5.6.2.1 例子：预测棒球击球率

下面给出一个把收缩方法用于棒球击球率 (batting averages) 的例子，来自 (Efron and Morris, 1975)。

我们观察到  $D = 18$  位球员在前  $T = 45$  场比赛中的安打 (hit) 次数，记为  $b_i$ 。

假设  $b_j \sim \text{Bin}(T, \theta_j)$ ，其中  $\theta_j$  是第  $j$  位球员的“真实”击球率。

目标是估计各个  $\theta_j$ 。

MLE 显然是  $\hat{\theta}_j = x_j$ ，其中  $x_j = b_j/T$  为经验击球率。

不过我们可以用\*\*经验贝叶斯 (EB) \*\*做得更好。

为了应用上节描述的高斯收缩方法，我们需要似然是高斯的：

$x_j \sim N(\theta_j, \sigma^2)$ ，且  $\sigma^2$  已知。（这里把下标  $i$  省略了，因为我们设  $N_j = 1$ —— $x_j$  已经是第  $j$  位球员的“平均值”。）

但在本例中，真实的似然是二项式。它的均值虽对 ( $E[x_j] = \theta_j$ )，但方差并非常数：

$$\text{var}[x_j] = \frac{1}{T^2} \text{var}[b_j] = \frac{T\theta_j(1-\theta_j)}{T^2} \quad (5.87)$$

因此我们对  $x_j$  施加一个方差稳定变换，使其更接近高斯假设：

$$y_j = f(x_j) = \sqrt{T} \arcsin(2x_j - 1) \quad (5.88)$$

于是近似有  $y_j \sim N(f(\theta_j), 1) = N(\mu_j, 1)$ 。

我们用式 (5.86) 且取  $\sigma^2 = 1$  对  $\mu_j$  做高斯收缩估计，然后再反变换得到

$$\hat{\theta}_j = \frac{1}{2}(\sin(\hat{\mu}_j/\sqrt{T}) + 1) \quad (5.89)$$

结果见图 5.12(a-b)。在 (a) 中，我们绘制了 MLE  $\hat{\theta}_j$  与后验均值  $\theta_j$ 。可以看到所有估计都向全局均值 0.265 收缩。

在 (b) 中，我们绘制了真实值  $\theta_j$ 、MLE  $\hat{\theta}_j$  与后验均值  $\hat{\theta}_j$  (“真实”的  $\theta_j$  是用大量独立比赛估出来的）。

可以看到，平均而言，收缩后的估计比 MLE 更接近真实参数。具体地，用

$$\text{MSE} = \frac{1}{N} \sum_{j=1}^D (\theta_j - \hat{\theta}_j)^2$$

衡量，收缩估计的均方误差比 MLE 小 3 倍以上。

---

## 基础且详尽的解释

### 1) 问题设定

- 每名球员前 45 场的安打次数  $b_j$ ，视为  $b_j \sim \text{Bin}(T, \theta_j)$ 。
- 经验击球率  $x_j = b_j/T$  就是 MLE。小样本（45 场其实不大）下，MLE 波动大：有人一开始“爆种”（很高），有人“手冷”（很低）。

### 2) 为什么需要“收缩”

- 我们相信不同球员的真实击球率  $\theta_j$  彼此相关：都来自同一个总体分布（上一小节里是  $\theta_j \sim N(\mu, \tau^2)$  的想法）。
- 收缩 (shrinkage)**：把个体的估计值往总体均值  $\mu$  适度拉近。
- 直觉：样本少/噪声大  $\rightarrow$  更不可靠  $\rightarrow$  往总体均值拉得更多；样本多/噪声小  $\rightarrow$  拉得更少。

### 3) 为什么要“方差稳定变换”

- 若直接对  $x_j$  用高斯模型，噪声方差却依赖  $\theta_j$ （上式 5.87），不满足“方差相等”的前提。
- 采用经典的 arcsin 变换（等价于常见的  $2\sqrt{T} \arcsin \sqrt{x}$  形式），把比例  $x_j$  变成  $y_j \approx N(f(\theta_j), 1)$ ，从而方差约为常数 1。
- 这样就可以用高斯收缩的解析公式（上一节式 (5.86) 的 James–Stein 形式）来估计  $\mu_j = f(\theta_j)$ 。

### 4) 收缩公式怎么用

- 在高斯模型下： $\hat{\mu}_j = \hat{B}\bar{y} + (1 - \hat{B})y_j = \bar{y} + (1 - \hat{B})(y_j - \bar{y})$ ，其中  $\bar{y}$  是所有球员的平均， $\hat{B} \in [0, 1]$  是收缩系数（由噪声方差与总体方差比值确定）。
- 有了  $\hat{\mu}_j$  后，通过式 (5.89) 反变换回  $\hat{\theta}_j$ 。

### 5) 图 5.12 如何读

- (a) 把每名球员的 MLE（上端点）连到收缩后的估计（下端点），所有线条都向中间的全局均值 0.265 合拢，这就是“收缩”。

- (b) 蓝：真实  $\theta_j$ （用很多比赛估出的“真相”）；红：只用 45 场做的 MLE；绿：收缩后的估计。

你会发现绿色一般比红色更贴近蓝色 → 这就是整体误差 (MSE) 更小的原因。

## 6) 直观比喻

- 把 MLE 看作“只听一耳朵风”的结论；
- 收缩估计等于“参考了所有人的平均水平”再修正你的判断；
- 对样本少、波动大的个体，这样能显著降低过度乐观/悲观的偏差。

## 要点总结

- 先对二项比例做**方差稳定变换**，把问题近似成**已知方差的高斯**；
- 用**高斯收缩/James–Stein 形式**估计变换后的均值；
- **再反变换回击球率**；
- 结果：在小样本阶段，收缩估计比 MLE **更稳、更准**（更低的 MSE）。

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。

### 5.6.2.2 Estimating the hyper-parameters

In this section, we give an algorithm for estimating  $\eta$ .

Suppose initially that  $\sigma^2_j = \sigma^2$  is the same for all groups.

In this case, we can derive the EB estimate in closed form, as we now show. From Equation 4.126, we have

$$p(x_j|\mu, \tau^2, \sigma^2) = N(x_j|\theta_j, \sigma^2)N(\theta_j|\mu, \tau^2)d\theta_j = N(x_j|\mu, \tau^2 + \sigma^2) \quad (5.90)$$

Hence the marginal likelihood is

$$p(D|\mu, \tau^2, \sigma^2) = \prod_{j=1}^J N(x_j|\mu, \tau^2 + \sigma^2) \quad (5.91)$$

Thus we can estimate the hyper-parameters using the usual MLEs for a Gaussian. For  $\mu$ , we have

$$\hat{\mu} = \frac{1}{J} \sum_{j=1}^J x_j = \bar{x} \quad (5.92)$$

which is the overall mean. For the variance, we can use moment matching (which is equivalent to the MLE for a Gaussian): we simply equate the model variance to the empirical variance:

$$\hat{\tau}^2 = \frac{1}{J} \sum_{j=1}^J (x_j - \bar{x})^2 = s^2 \quad (5.93)$$

so  $\hat{\tau}^2 = s^2 - \sigma^2$ . Since we know  $\tau^2$  must be positive, it is common to use the following revised estimate:

$$\hat{\tau}^2 = \max\{0, s^2 - \sigma^2\} = (s^2 - \sigma^2)^+ \quad (5.94)$$

Hence the shrinkage factor is

$$\hat{B} = \sigma^2 / (\sigma^2 + \hat{\tau}^2) = \sigma^2 / (s^2 + (s^2 - \sigma^2)^+) \quad (5.95)$$

In the case where the  $\sigma_j^2$ 's are different, we can no longer derive a solution in closed form. Exercise 11.13 discusses how to use the EM algorithm to derive an EB estimate, and Exercise 24.4 discusses how to perform full Bayesian inference in this hierarchical model.

下面先逐句翻译原文（保证无遗漏），然后给出基础且详尽的解释，帮你理解每个公式的来源和意义。

---

## 逐句翻译

### 5.6.2.2 估计超参数 (hyper-parameters)

在本节中，我们给出一个估计  $\eta$  的算法。假设一开始所有组 (group) 都有相同的方差  $\sigma_j^2 = \sigma^2$ 。在这种情况下，我们可以推导出 EB (经验贝叶斯) 的闭式解，如下所示。

根据公式 (4.126)，我们有

$$p(x_j | \mu, \tau^2, \sigma^2) = N(x_j | \theta_j, \sigma^2) N(\theta_j | \mu, \tau^2) d\theta_j = N(x_j | \mu, \tau^2 + \sigma^2) \quad (5.90)$$

因此，**边际似然** (marginal likelihood) 为

$$p(D | \mu, \tau^2, \sigma^2) = \prod_{j=1}^D N(x_j | \mu, \tau^2 + \sigma^2) \quad (5.91)$$

因此我们可以用普通高斯分布的最大似然估计 (MLE) 来估计超参数。

对于  $\mu$ ，我们有

$$\hat{\mu} = \frac{1}{D} \sum_{j=1}^D x_j = \bar{x} \quad (5.92)$$

这就是总体均值。

对于方差，我们可以使用**矩匹配法** (moment matching)（在高斯分布下等价于 MLE）：我们直接令模型方差等于经验方差：

$$\tau^2 + \sigma^2 = \frac{1}{D} \sum_{j=1}^D (x_j - \bar{x})^2 = s^2 \quad (5.93)$$

所以

$$\hat{\tau}^2 = s^2 - \sigma^2$$

由于我们知道  $\tau^2$  必须是非负的，所以常见的修正估计为：

$$\hat{\tau}^2 = \max\{0, s^2 - \sigma^2\} = (s^2 - \sigma^2)_+ \quad (5.94)$$

因此收缩系数为：

$$\hat{B} = \frac{\sigma^2}{\sigma^2 + \hat{\tau}^2} = \frac{\sigma^2}{\sigma^2 + (s^2 - \sigma^2)_+} \quad (5.95)$$

在  $\sigma_j^2$  不相等的情况下，我们就不能再推导出闭式解。练习 11.13 讨论了如何用 EM 算法求 EB 估计，练习 24.4 讨论了如何在该分层模型中做完整的贝叶斯推断。

---

## 基础且详尽的解释

### 1) 问题背景

这里是在做\*\*经验贝叶斯 (EB) \*\*估计，用的是一个分层高斯模型：

- 第  $j$  组的真实参数： $\theta_j \sim N(\mu, \tau^2)$
  - 观测数据的均值： $x_j | \theta_j \sim N(\theta_j, \sigma^2)$
  - $\sigma^2$  是已知的观测噪声方差（例如由测量误差决定）
  - $\mu, \tau^2$  是未知的超参数 (hyper-parameters)，需要用 EB 从数据中估计
- 

### 2) 为什么能直接积分得到 (5.90)

- 我们有两个正态分布相乘并对  $\theta_j$  积分：

$$N(x_j | \theta_j, \sigma^2) \cdot N(\theta_j | \mu, \tau^2)$$

- 在贝叶斯推断中，这是一个“先验  $\times$  似然”的结构。
- 两个高斯相乘再对中间变量积分，结果还是一个高斯分布，方差相加、均值用先验均值。
- 所以积分结果是

$$x_j \sim N(\mu, \tau^2 + \sigma^2)$$

这就是 (5.90) 的来源。

---

### 3) 边际似然 (5.91)

- 所有  $x_j$  独立  $\rightarrow$  联合概率就是单个的乘积。
  - 这个公式告诉我们：如果直接用边际分布看数据，就等价于假设所有  $x_j$  来自一个均值为  $\mu$ ，方差为  $\tau^2 + \sigma^2$  的高斯分布。
- 

### 4) 估计 $\mu$ 和 $\tau^2$

- 因为是高斯分布，MLE 很简单：
  - 均值  $\hat{\mu} =$  所有  $x_j$  的平均值  $\bar{x}$  (公式 5.92)

- 方差  $\tau^2 + \sigma^2 =$  数据的经验方差  $s^2$  (公式 5.93)
  - 所以  $\hat{\tau}^2 = s^2 - \sigma^2$
  - 但方差不能是负的, 所以如果  $s^2 < \sigma^2$ , 就设  $\hat{\tau}^2 = 0$  (公式 5.94)
- 

## 5) 收缩系数 $\hat{B}$ (公式 5.95)

- 回忆前面的收缩公式:

$$\hat{\theta}_j = \bar{x} + (1 - B)(x_j - \bar{x})$$

其中

$$B = \frac{\sigma^2}{\sigma^2 + \tau^2}$$

- 用上面估出来的  $\hat{\tau}^2$  代进去, 就得到 (5.95)。

### 直观解释:

- 如果总体方差  $\tau^2$  很小 (大家水平接近), 则  $B$  很大,  $(1-B)$  很小  $\rightarrow$  个体估计大幅收缩到均值。
  - 如果总体方差很大 (大家差异大), 则  $B$  小  $\rightarrow$  个体估计更依赖自己的数据。
- 

## 6) 特殊情况

- 如果不同组的  $\sigma_j^2$  不一样, 公式就没法直接简化成闭式。
  - 这种情况下需要用 EM 算法来迭代估计 (练习 11.13), 或者用完全贝叶斯方法 (练习 24.4)。
- 

## 7) 小结

- 这个过程就是用数据的总体均值和总体方差来反推出先验分布的超参数, 再用这些超参数计算收缩系数, 从而得到改进后的个体估计。
  - 本节给出了  $\sigma^2$  相同情况下的闭式解, 这在实际中非常高效。
- 

如果你愿意, 我可以帮你画一张图, 展示  $\hat{\tau}^2$ 、 $\hat{B}$  与收缩效果之间的关系, 这样你就能直观地看到为什么方差越大收缩越少、方差越小收缩越多。

你要我帮你画吗?

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。

### 5.7 Bayesian decision theory

We have seen how probability theory can be used to represent and updates our beliefs about the state of the world. However, ultimately our goal is to convert our beliefs into actions. In this section, we discuss the optimal way to do this. We can formalize any given statistical decision problem as a game against nature (as opposed to a game against other strategic players, which is the topic of game theory, see e.g., (Shoham and Leyton-Brown 2009) for details). In this game, nature picks a state or parameter or label,  $y \in Y$ , unknown to us, and then generates an observation,  $x \in X$ , which we get to see. We then have to make a decision, that is, we have to choose an action  $a$  from some action space  $A$ . Finally we incur some loss,  $L(y, a)$ , which measures how compatible our action  $a$  is with nature's hidden state  $y$ . For example, we might use misclassification loss,  $L(y, a) = I(y = a)$ , or squared loss,  $L(y, a) = (y - a)^2$ . We will see some other examples below.

Our goal is to devise a decision procedure or policy,  $\delta : X \rightarrow A$ , which specifies the optimal action for each possible input. By optimal, we mean the action that minimizes the expected loss:

$$\delta(x) = \operatorname{argmin}_{a \in A} E[L(y, a)] \quad (5.96)$$

In economics, it is more common to talk of a utility function; this is just negative loss,  $U(y, a) = -L(y, a)$ . Thus the above rule becomes

$$\delta(x) = \operatorname{argmax}_{a \in A} E[U(y, a)] \quad (5.97)$$

This is called the maximum expected utility principle, and is the essence of what we mean by rational behavior. Note that there are two different interpretations of what we mean by "expected". In the Bayesian version, which we discuss below, we mean the expected value of  $y$  given the data we have seen so far. In the frequentist version, which we discuss in Section 6.3, we mean the expected value of  $y$  and  $x$  that we expect to see in the future. In the Bayesian approach to decision theory, the optimal action, having observed  $x$ , is defined as the action  $a$  that minimizes the posterior expected loss:

$$\rho(a|x) = E_p(y|x) [L(y, a)] = y L(y, a) p(y|x) \quad (5.98)$$

(If  $y$  is continuous (e.g., when we want to estimate a parameter vector), we should replace the sum with an

integral.) Hence the Bayes estimator, also called the Bayes decision rule, is given by

(If  $y$  is continuous (e.g., when we want to estimate a parameter vector), we should replace the sum with an integral.) Hence the Bayes estimator, also called the Bayes decision rule, is given by

$$\delta(x) = \operatorname{argmin}_{a \in A} \rho(a|x) \quad (5.99)$$

好的，我会先完整翻译你给的内容（逐句无遗漏），然后给出通俗易懂且数学基础友好的解释，帮你弄清楚公式背后的含义。

---

## 逐句翻译

### 5.7 贝叶斯决策理论 (Bayesian decision theory)

我们已经看到，概率论可以用来表示并更新我们对世界状态的信念。

然而，我们的最终目标是将信念转化为行动。

在本节中，我们讨论如何以最优的方式做到这一点。

我们可以将任何给定的统计决策问题形式化为一个与自然的博弈（而不是与其他有策略的玩家的博弈——后者是博弈论的主题，详细内容可参见 Shoham 和 Leyton-Brown 2009）。

在这个博弈中：

- 自然 (Nature) 首先选择一个状态、参数或标签  $y \in Y$ ，我们并不知道它的值；
- 然后它生成一个观测值  $x \in X$ ，这是我们能够看到的；
- 接着，我们必须做出一个决策，即从某个动作空间  $A$  中选择一个动作  $a$ ；
- 最后，我们会遭受某种损失  $L(y, a)$ ，这个损失衡量了我们的动作  $a$  与自然的隐藏状态  $y$  的契合程度。

例如：

- 如果是分类问题，可以使用**误分类损失**：

$$L(y, a) = I(y \neq a)$$

( $I$  是指示函数，若预测错了就损失 1，预测对了损失 0)

- 如果是回归问题，可以使用**平方损失**：

$$L(y, a) = (y - a)^2$$

我们将在下面看到其他例子。

---

我们的目标是设计一个**决策过程或策略**  $\delta : X \rightarrow A$ ，它会为每一个可能的输入  $x$  指定一个最优的动作。

这里的“最优”是指**最小化期望损失**的动作：

$$\delta(x) = \arg \min_{a \in A} E[L(y, a)] \quad (5.96)$$


---

在经济学中，更常用**效用函数** (utility function) 的说法；它只是损失函数的相反数：

$$U(y, a) = -L(y, a)$$

于是，上面的规则就变成了：

$$\delta(x) = \arg \max_{a \in A} E[U(y, a)] \quad (5.97)$$

这被称为**最大期望效用原则** (maximum expected utility principle)，也是我们所说的“理性行为”的核心。

注意，这里的“期望”有两种不同的解释：

1. **贝叶斯解释**：指的是在我们已经看到的数据条件下， $y$  的期望值；
  2. **频率学派解释** (见 6.3 节)：指的是我们期望在未来看到的  $y$  和  $x$  的期望值。
- 

在贝叶斯的决策理论中，当我们已经观察到  $x$  时，最优的动作定义为**使得后验期望损失最小**的动作：

$$\rho(a|x) = E_{p(y|x)}[L(y, a)] = \sum_y L(y, a) p(y|x) \quad (5.98)$$

(如果  $y$  是连续的，例如当我们要估计一个参数向量时，应将求和替换为积分。)

因此，**贝叶斯估计器** (Bayes estimator)，也叫**贝叶斯决策规则** (Bayes decision rule)，为：

$$\delta(x) = \arg \min_{a \in A} \rho(a|x) \quad (5.99)$$


---

## 通俗详解

你可以这样理解：

1. **自然是“出题人”**
  - 它先偷偷选一个“正确答案”  $y$  (可能是一个分类标签，也可能是一个数值)。
  - 然后它给我们一个提示 (观测值)  $x$ 。
2. **我们是“答题人”**
  - 我们看到了提示  $x$ ，要选择一个答案  $a$  (也就是采取一个动作)。
  - 如果我们选错或选得不够好，就会受到“惩罚” (损失  $L(y, a)$ )。

## 1. 期望损失 (Expected Loss)

我们不知道  $y$  的真实值，所以没法直接算损失。

但是我们有一个信念分布  $p(y|x)$  (由贝叶斯定理得到的后验概率)。

- 于是可以计算平均意义下的损失：

$$\rho(a|x) = \sum_y L(y, a) p(y|x)$$

这就是后验期望损失。

## 2. 决策规则

- 最优动作  $\delta(x) =$  让后验期望损失最小的动作。
- 换句话说：别选可能带来大损失的动作，要在“平均风险”最小的地方下手。

## 3. 损失和效用是“反义词”

- 如果你想“最大化好处”，可以用效用函数  $U = -L$ 。
- 最大化期望效用  $\leftrightarrow$  最小化期望损失，本质一样。

## 4. 具体例子

- 分类：**如果损失是“预测错就是 1，预测对就是 0”，最优动作就是选后验概率最大的类 (MAP 决策)。
- 回归：**如果损失是平方误差，最优动作就是选后验分布的均值。

如果你愿意，我可以帮你画一张**决策流程图**，直观展示“自然 → 产生数据 → 计算后验 → 选动作”的全过程，这样你会更容易记住公式 (5.96) 到 (5.99) 的逻辑关系。

你要我帮你画吗？



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。

### 5.7.1 Bayes estimators for common loss functions

In this section we show how to construct Bayes estimators for the loss functions most commonly arising in machine learning.

#### 5.7.1.1 MAP estimate minimizes 0-1 loss

The 0-1 loss is defined by

$$L(y, a) = I(y = a) = 0 \text{ if } a = y \quad 1 \text{ if } a \neq y \quad (5.100)$$

This is commonly used in classification problems where  $y$  is the true class label and  $a = \hat{y}$  is the estimate. For example, in the two class case, we can write the loss matrix as follows:

(In Section 5.7.2, we generalize this loss function so it penalizes the two kinds of errors on the off-diagonal differently.) The posterior expected loss is

$$p(a|x) = p(a = y|x) = 1 - p(y|x) \quad (5.101)$$

Hence the action that minimizes the expected loss is the posterior mode or MAP estimate

$$y^*(x) = \arg \max_{y \in Y} p(y|x) \quad (5.102)$$

好的，我先逐句完整翻译，然后再做一个基础且详细的解释，帮你把公式和表格都看懂。

## 逐句翻译

### 5.7.1 常见损失函数的贝叶斯估计器

在这一节中，我们将展示如何为机器学习中最常见的损失函数构造贝叶斯估计器。

#### 5.7.1.1 MAP 估计器最小化 0-1 损失

**0-1 损失 (0-1 loss)** 定义为：

$$L(y, a) = I(y \neq a) = \begin{cases} 0 & \text{如果 } a = y \\ 1 & \text{如果 } a \neq y \end{cases} \quad (5.100)$$

它通常用于分类问题，其中  $y$  是真实的类别标签，而  $a = \hat{y}$  是预测的标签。

例如，在二分类的情况下，我们可以把损失矩阵写成如下形式（你图中那个表格）：

	$\hat{y} = 1$	$\hat{y} = 0$
$y = 1$	0	1
$y = 0$	1	0

	$\hat{y} = 1$	$\hat{y} = 0$
$y = 0$	1	0

(在 5.7.2 节中，我们会推广这种损失函数，使它可以对矩阵中两个不同位置的错误施加不同的惩罚。)

---

**后验期望损失 (posterior expected loss)** 是：

$$\rho(a|x) = p(a \neq y|x) = 1 - p(y = a|x) \quad (5.101)$$

因此，使期望损失最小的动作就是**后验分布的众数**，也叫 MAP (Maximum A Posteriori) 估计：

$$y^*(x) = \arg \max_{y \in Y} p(y|x) \quad (5.102)$$


---

## 基础且详细的解释

### 1. 0-1 损失的含义

- 核心思想：预测对了损失是 0，预测错了损失是 1。
- 表格里的含义：
  - 行是**真实标签**  $y$
  - 列是**预测标签**  $\hat{y}$
  - 表格的数字是损失值
  - 例如：
    - $y = 1, \hat{y} = 1 \rightarrow$  预测正确  $\rightarrow$  损失 = 0
    - $y = 1, \hat{y} = 0 \rightarrow$  预测错误  $\rightarrow$  损失 = 1
    - $y = 0, \hat{y} = 1 \rightarrow$  预测错误  $\rightarrow$  损失 = 1
    - $y = 0, \hat{y} = 0 \rightarrow$  预测正确  $\rightarrow$  损失 = 0

这就是名字里“0-1”的由来。

---

### 2. 后验期望损失的推导

- 我们已经看到观测值  $x$
- 如果我们选择动作  $a$  (即预测标签  $\hat{y} = a$ )，损失的期望就是：

$$\rho(a|x) = p(\text{预测错} | x)$$

- 因为“预测错”这个事件的概率就是  $1 - p(\text{预测对} | x)$ ：

$$\rho(a|x) = 1 - p(y = a|x)$$

### 3. 最优决策 (MAP)

- 我们要让  $p(a|x)$  最小 → 也就是要让  $p(y = a|x)$  最大
- 所以最优动作  $y^*(x)$  是：

选择后验概率最大的类别

- 这就是 MAP 规则 (Maximum A Posteriori, 最大后验估计)。

### 4. 直观理解

- 你可以把它想成下注：
  - $p(y = 1|x)$  是这局下注 1 赢的概率
  - $p(y = 0|x)$  是下注 0 赢的概率
  - 如果错误惩罚都一样 (0-1损失)，那就永远押概率大的那个。

如果你愿意，我可以帮你画一张\*\*“0-1 损失 → MAP 决策”的流程图，把表格、公式 (5.101) 和 (5.102) 结合起来，你就能用一张图理解这个推导。

你需要我帮你画吗？



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 5.13 For some regions of input space, where the class posteriors are uncertain, we may prefer not to choose class 1 or 2; instead we may prefer the reject option. Based on Figure 1.26 of (Bishop 2006a).

#### 5.7.1.2 Reject option

In classification problems where  $p(y|x)$  is very uncertain, we may prefer to choose a reject action, in which we refuse to classify the example as any of the specified classes, and instead say "don't know". Such ambiguous cases can be handled by e.g., a human expert. See Figure 5.13 for an illustration. This is useful in risk averse domains such as medicine and finance. We can formalize the reject option as follows. Let choosing  $a = C + 1$  correspond to picking the reject action, and choosing  $a \in \{1, \dots, C\}$  correspond to picking one of the classes. Suppose we define the loss function as

$$L(y = j, a = i) = \begin{cases} 0 & \text{if } i = j \text{ and } i, j \in \{1, \dots, C\} \\ \lambda_r & \text{if } i = C + 1 \\ \lambda_s & \text{otherwise} \end{cases} \quad (5.103)$$

where  $\lambda_r$  is the cost of the reject action, and  $\lambda_s$  is the cost of a substitution error. In Exercise 5.3, you will show that the optimal action is to pick the reject action if the most probable class has a probability below  $1 - \lambda_r \lambda_s$ ; otherwise you should just pick the most probable class.

好的，我会先完整翻译原文，然后再做一个基础且详细的解释，帮你结合图 5.13 理解“拒绝选项 (reject option)”的数学含义和实际意义。

## 逐句翻译

### 图 5.13

对于输入空间中的某些区域，如果类别的后验概率很不确定，我们可能更倾向于不选择类别 1 或类别 2；相反，我们可能更倾向于选择**拒绝选项 (reject option)**。

拒绝选项的意思是拒绝将该样本分类到任何已定义的类别中，而是直接说“不知道”。

像这种不确定性较高的样本，可以交由人工专家来处理。

这在一些**风险敏感领域**（例如医疗、金融）中很有用。

图改编自 (Bishop 2006a) 的图 1.26。

### 5.7.1.2 拒绝选项 (Reject option)

在分类问题中，如果  $p(y|x)$  非常不确定，我们可能会更倾向于选择**拒绝动作 (reject action)**——即拒绝将该样本分类为任何指定的类别，而是回答“我不知道”。

这样模糊的情况可以交由人类专家来判断（见图 5.13）。

这种方法在**风险规避型领域**（如医疗和金融）中特别有用。

我们可以形式化地定义拒绝选项：

- 设选择  $a = C + 1$  表示选择**拒绝动作**

- 选择  $a \in \{1, \dots, C\}$  表示选择某个类别

定义损失函数如下：

$$L(y = j, a = i) = \begin{cases} 0 & \text{如果 } i = j \text{ 且 } i, j \in \{1, \dots, C\} \\ \lambda_r & \text{如果 } i = C + 1 \quad (\text{拒绝动作的代价}) \\ \lambda_s & \text{否则} \quad (\text{分类错误的代价}) \end{cases} \quad (5.103)$$

其中：

- $\lambda_r$  = 拒绝动作的成本
- $\lambda_s$  = 误分类的成本

在练习 5.3 中，你会证明：

**最优动作**的规则是：

- 如果最可能类别的概率  $< 1 - \frac{\lambda_r}{\lambda_s}$ ，就选择**拒绝动作**
  - 否则，就选择最可能的类别。
- 

## 基础且详细解释

### 1. 为什么需要拒绝选项？

- 在某些场景里，随便猜可能带来**很高风险**：
    - 医疗诊断：误诊会影响病人生命
    - 金融交易：错误预测可能导致巨额亏损
  - 如果模型对某个样本不确定（例如分类概率差不多），我们宁可说“我不知道”交给专家处理，也不要乱分类。
- 

### 2. 图 5.13 讲的是什么？

- 横轴  $x$ ：输入特征值
  - 纵轴：类别的**后验概率**  $p(y|x)$
  - 蓝线：预测是类别 1 的概率  $p(y = 1|x)$
  - 红线：预测是类别 2 的概率  $p(y = 2|x)$
  - 虚线 threshold (阈值)：决定拒绝的标准
  - Reject Region (拒绝区域)**：
    - 在这个区域里，两个类别的概率都低于阈值
    - 说明模型非常不确定 → 不分类，直接拒绝
- 

### 3. 损失函数 (5.103) 的含义

- 0 损失**：预测完全正确

- $\lambda r$  损失：选择拒绝动作（即“我不知道”）的代价
  - $\lambda s$  损失：分类错误的代价
  - 你可以理解为：
    - $\lambda s$  很大  $\rightarrow$  错误很贵  $\rightarrow$  模型会更倾向于拒绝
    - $\lambda r$  较小  $\rightarrow$  拒绝的代价低  $\rightarrow$  模型会更常说“我不知道”
- 

## 4. 最优策略公式

- 规则：
    - 如果最大类别的概率  $< 1 - \frac{\lambda_r}{\lambda_s}$   $\rightarrow$  拒绝
    - 否则  $\rightarrow$  选择最大概率类别 (MAP)
  - 直观理解：
    - $\frac{\lambda_r}{\lambda_s}$  越大，拒绝越“贵”，模型就更少拒绝
    - 这个公式是通过比较“拒绝的期望损失”和“分类的期望损失”推导出来的
- 

## 5. 实际案例

- 假设  $\lambda s = 1$  (误分类成本是 1) ,  $\lambda r = 0.2$  (拒绝成本是 0.2)
- 则阈值：

$$1 - \frac{\lambda_r}{\lambda_s} = 1 - 0.2 = 0.8$$

- 规则：
    - 如果最大类别概率  $< 0.8 \rightarrow$  拒绝
    - 否则  $\rightarrow$  选最大概率类别
- 

如果你愿意，我可以帮你画一个结合公式和图 5.13 的决策流程图，把“拒绝区域”用概率条件和阈值公式表示出来，这样你一眼就能看懂什么时候拒绝、什么时候分类。  
你要我帮你画吗？

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。

5.7.1.3 Posterior mean minimizes 2 (quadratic) loss  
For continuous parameters, a more appropriate loss function is squared error, 2 loss, or quadratic loss, defined

as

$$L(y, a) = (y - a)^2 \quad (5.104)$$

The posterior expected loss is given by

$$\rho(a|x) = E[(y - a)^2 | x] = E[y^2 | x] - 2aE[y|x] + a^2 \quad (5.105)$$

Hence the optimal estimate is the posterior mean:

$$\frac{\partial}{\partial a} \rho(a|x) = -2E[y|x] + 2a = 0 \Rightarrow \hat{y} = E[y|x] = \int y p(y|x) dy \quad (5.106)$$

This is often called the minimum mean squared error estimate or MMSE estimate. In a linear regression problem, we have

$$p(y|x, \theta) = N(y|xTw, \sigma^2) \quad (5.107)$$

In this case, the optimal estimate given some training data D is given by

$$E[y|x, D] = x^T E[w|D] \quad (5.108)$$

That is, we just plug-in the posterior mean parameter estimate. Note that this is the optimal thing to do no matter what prior we use for w.

好的，我会先给你完整逐句翻译，然后做一个详细、基础的解释，让你能在数学基础不强的情况下理解“后验均值最小化二次损失（MMSE估计）”的原理。

## 逐句翻译

### 5.7.1.3 后验均值最小化二次（平方）损失

对于连续参数，更合适的损失函数是**平方误差**（square error）、 $L_2$  **损失**或**二次损失**（quadratic loss），定义为

$$L(y, a) = (y - a)^2 \quad (5.104)$$

后验期望损失（posterior expected loss）为

$$\rho(a|x) = E[(y - a)^2 | x] = E[y^2 | x] - 2a E[y|x] + a^2 \quad (5.105)$$

因此，最优估计是**后验均值**：

$$\frac{\partial}{\partial a} \rho(a|x) = -2 E[y|x] + 2a = 0 \Rightarrow \hat{y} = E[y|x] = \int y p(y|x) dy \quad (5.106)$$

这种估计通常称为**最小均方误差估计**（MMSE estimate）。

在一个线性回归问题中：

$$p(y|x, \theta) = N(y | x^T w, \sigma^2) \quad (5.107)$$

在这种情况下，给定一些训练数据 D，最优估计为

$$E[y | x, D] = x^T E[w | D] \quad (5.108)$$

也就是说，我们只需代入**参数的后验均值估计**即可。

注意，无论我们对  $w$  使用什么先验，这样做都是最优的。

---

## 详细解释（基础版）

### 1. 背景

- 我们之前讨论过 **0-1 损失**（适合分类任务），对应的最优估计是**MAP**（后验概率最大值）。
- 如果目标变量  $y$  是**连续的**（比如预测房价、温度），0-1 损失就没意义了。
- 在这种情况下，常用的损失是**平方误差**：

$$(y - a)^2$$

这里：

- $y$  = 真值
  - $a$  = 我们的预测
  - 损失就是“预测和真实的差的平方”
- 

### 2. 为什么平方误差好用？

- 平方误差惩罚**大偏差**比小偏差更严重（因为平方增长很快）。
  - 它使数学推导变得很简单（因为它是二次函数）。
  - 在很多统计和机器学习模型里，这样的损失对应的最优解都有解析公式。
- 

### 3. 推导关键步骤

我们要找最优预测  $a$ ，即：

$$a^* = \arg \min_a \rho(a|x)$$

其中：

$$\rho(a|x) = E[(y - a)^2|x]$$

展开平方项：

$$(y - a)^2 = y^2 - 2ay + a^2$$

对  $x$  条件下取期望：

$$\rho(a|x) = E[y^2|x] - 2a E[y|x] + a^2$$

然后对  $a$  求导，设为 0：

$$-2 \mathbb{E}[y|x] + 2a = 0$$

解得：

$$a = \mathbb{E}[y|x]$$

**结论：**当损失是平方误差时，最优预测就是**后验均值** (posterior mean)。

---

## 4. 这就是 MMSE (最小均方误差) 估计

MMSE 全称 Minimum Mean Squared Error:

- Minimum: 最小化
- Mean Squared Error: 平均平方误差

它的意义是：

- 在所有可能预测值中，选取那个能使平均平方误差最小的预测值
  - 答案就是**预测分布的平均值**
- 

## 5. 在线性回归中的应用

线性回归假设：

$$y \sim N(x^T w, \sigma^2)$$

- 给定数据  $D$ ，参数  $w$  也有一个**后验分布**  $p(w|D)$
- 由于期望的线性性质：

$$\mathbb{E}[y|x, D] = x^T \mathbb{E}[w|D]$$

意思是：

- 先找到  $w$  的后验均值（根据数据和先验计算）
  - 然后用它来预测新的  $y$
- 

## 6. 直观理解

- 如果你把  $p(y|x)$  想象成“可能的  $y$  值的分布”，
  - 0-1 损失  $\rightarrow$  选概率最大的值 (MAP)
  - 平方误差  $\rightarrow$  选所有可能值的**加权平均** (MMSE)
  - 因为平方误差惩罚大偏差很厉害，所以平均值可以平衡各种可能性
- 

如果你愿意，我可以帮你画一个**对比图**，显示

- 当损失函数是 0-1 损失时  $\rightarrow$  选 MAP

- 当损失函数是平方误差时 → 选 后验均值  
这样更直观地看到两种方法的区别。  
你要我帮你画吗？

公式5.108怎么得到的

公式 (5.108)

$$E[y|x, D] = x^\top E[w|D]$$

其实是从线性回归的模型假设和期望的线性性质直接推出来的，我帮你分成3步详细说明。

---

## 1. 模型假设

公式 (5.107) 给的是线性回归的条件分布：

$$p(y | x, \theta) = N(y | x^\top w, \sigma^2)$$

意思是：

- 在给定输入  $x$  和参数  $w$  的情况下， $y$  服从均值为  $x^\top w$ 、方差为  $\sigma^2$  的高斯分布。
- 这里  $\theta$  其实就是  $(w, \sigma^2)$ 。

由高斯分布的性质可知：

$$E[y | x, w] = x^\top w$$

也就是说，如果  $w$  已知， $y$  的条件期望就是线性预测  $x^\top w$ 。

---

## 2. 后验分布的引入

在贝叶斯框架下， $w$  本身是未知的随机变量，给定数据集  $D$  后， $w$  有一个**后验分布**  
 $p(w | D)$ 。

我们想要求的是：

$$E[y | x, D]$$

根据概率的全期望公式 (law of total expectation) :

$$E[y | x, D] = E_{w|D}[ E[y | x, w] ]$$

这一步是关键：我们先对已知  $w$  时的期望取平均，再对  $w$  的后验分布取期望。

---

### 3. 代入条件期望

我们已经知道：

$$E[y | x, w] = x^T w$$

代入到上面的公式：

$$E[y | x, D] = E_{w|D}[x^T w]$$

因为  $x$  是固定向量，可以拿到期望符号外面：

$$E[y | x, D] = x^T E_{w|D}[w]$$

这就得到了 **公式 (5.108)**。

#### 核心原因：

- 线性回归的预测均值是  $x^T w$
- 全期望公式把“先给定  $w$  取期望”转成“对  $w$  的后验均值取线性组合”
- 线性性质保证了期望和内积可以互换顺序

如果你愿意，我可以帮你画一张**公式 (5.108) 的推导流程图**，用直观的图示解释“先对  $y$  取期望，再对  $w$  取期望”的过程，这样你会更容易记住。

你要我画吗？



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 5.14 (a-c). Plots of the  $L(y, a) = |y - a|^q$  vs  $|y - a|$  for  $q = 0.2$ ,  $q = 1$  and  $q = 2$ . Figure generated by lossFunctionFig.

5.7.1.4 Posterior median minimizes 1 (absolute) loss  
The 2 loss penalizes deviations from the truth quadratically, and thus is sensitive to outliers. A more robust alternative is the absolute or 1 loss,  $L(y, a) = |y - a|$  (see Figure 5.14). The optimal estimate is the posterior

median, i.e., a value  $a$  such that  $P(y < a|x) = P(y \geq a|x) = 0.5$ .

See Exercise 5.9 for a proof.

#### 5.7.1.5 Supervised learning

Consider a prediction function  $\delta : X \rightarrow Y$ , and suppose we have some cost function  $(y, \delta(x))$  which gives the cost of predicting  $y$  when the truth is  $y$ . We can define the loss incurred by taking action  $\delta$  (i.e., using this predictor) when the unknown state of nature is  $\theta$  (the parameters of the data generating mechanism) as follows:

$$L(\theta, \delta) = E(x, y) \sim p(x, y|\theta) [ (y, \delta(x))] = x y L(y, \delta(x)) p(x, y|\theta) \quad (5.109)$$

This is known as the generalization error. Our goal is to minimize the posterior expected loss, given by

$$p(\delta|D) = p(\theta|D)L(\theta, \delta)d\theta \quad (5.110)$$

This should be contrasted with the frequentist risk which is defined in Equation 6.47.

我先把你给的英文内容逐句翻译，然后用基础且详细的方式解释每一部分，包括图像 Figure 5.14 的含义。

## 原文翻译

Figure 5.14 (a-c).  $L(y, a) = |y - a|^q$  与  $|y - a|$  的关系图，分别对应  $q = 0.2$ 、 $q = 1$  和  $q = 2$  的情况。图由 `lossFunctionFig` 生成。

### 5.7.1.4 后验中位数最小化 1 (绝对值) 损失

平方损失 ( $L_2$  损失) 会以平方的方式惩罚与真实值的偏差，因此对异常值非常敏感。一个更稳健 (robust) 的替代方法是绝对值损失 ( $L_1$  损失)：

$$L(y, a) = |y - a|$$

(见图 5.14)

最优估计是后验分布的中位数，也就是说，一个值  $a$  满足：

$$P(y < a | x) = P(y \geq a | x) = 0.5$$

关于这个结论的证明可参考练习 5.9。

### 5.7.1.5 监督学习

考虑一个预测函数：

$$\delta : X \rightarrow Y$$

假设我们有一个代价函数  $c(y, \hat{y})$ , 表示当真实值为  $y$  时预测成  $\hat{y}$  的代价。

当真实的自然状态是  $\theta$  (即数据生成机制的参数) 时, 使用预测器  $\delta$  造成的损失可以定义为:

$$L(\theta, \delta) = E_{(x,y) \sim p(x,y|\theta)} [ c(y, \delta(x)) ] = \sum_x \sum_y L(y, \delta(x)) p(x, y | \theta) \quad (5.109)$$

这被称为**泛化误差** (generalization error)。

我们的目标是最小化**后验期望损失**:

$$\rho(\delta | D) = \int p(\theta | D) L(\theta, \delta) d\theta \quad (5.110)$$

这里  $D$  是观测到的数据集。

需要注意, 这与频率学派的风险 (frequentist risk, 公式 6.47) 不同。

---

## 详细解释 (数学基础友好版)

### 1. 图 Figure 5.14 的含义

图中画的是**不同  $q$  值的  $L_q$  损失函数**, 即:

$$L_q(y, a) = |y - a|^q$$

- (a)  $q = 0.2$ : 曲线很“平”, 除了在 0 附近比较尖锐。这种损失函数对大的偏差不敏感, 对小偏差敏感。
  - (b)  $q = 1$ : V 字形, 这就是**绝对值损失** ( $L_1$  损失)。
  - (c)  $q = 2$ : U 字形, 这就是**平方损失** ( $L_2$  损失)。
- 

### 2. $L_2$ 损失 (平方损失) 的问题

平方损失:

$$L(y, a) = (y - a)^2$$

偏差平方, 所以如果预测差错很大 (比如异常值), 惩罚会特别大 → 对异常值非常敏感。

---

### 3. $L_1$ 损失 (绝对值损失) 更稳健

绝对值损失:

$$L(y, a) = |y - a|$$

不管偏差多大, 都只按比例增加损失, 不会像平方那样爆炸 → 对异常值更稳健 (robust)。

---

## 4. 后验中位数是 $L_1$ 损失的最优解

当我们用  $L_1$  损失时，最小化后验期望损失：

$$\rho(a|x) = E_{y|x} [ |y - a| ]$$

数学推导可以证明，使这个期望最小的  $a$  是**后验分布的中位数**，即：

$$P(y < a | x) = P(y \geq a | x) = 0.5$$

意思是：预测值  $a$  把后验分布的概率分成一半一半。

---

## 5. 监督学习部分

这里是把“损失函数”的思想推广到整个监督学习框架：

- 我们有一个预测器  $\delta(x)$ ，输入  $x$ ，输出预测的  $y$ 。
- 如果数据由参数  $\theta$  的概率分布  $p(x, y|\theta)$  生成，那么预测器的平均损失就是：

$$L(\theta, \delta) = E_{(x,y) \sim p(x,y|\theta)} [ c(y, \delta(x)) ]$$

这叫**泛化误差**（它是真实分布下的平均损失）。

---

## 6. 贝叶斯版与频率学派版的区别

- **贝叶斯版**： $\theta$  是随机变量，有后验分布  $p(\theta|D)$ ，我们最小化的是：

$$\rho(\delta|D) = E_{\theta \sim p(\theta|D)} [ L(\theta, \delta) ]$$

- **频率学派版**： $\theta$  是固定的未知量，不对它求期望。
- 

如果你愿意，我可以帮你画一张  $L_1$ 、 $L_2$  和  $q=0.2$  的损失函数对比图 + 异常值影响演示，这样你会更直观地理解为什么  $L_1$  比  $L_2$  更稳健。

你需要我画吗？



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。

### 5.7.2 The false positive vs false negative tradeoff

In this section, we focus on binary decision problems, such as hypothesis testing, two-class classification, object/event detection, etc. There are two types of error we can make: a false positive (aka false alarm), which arises when we estimate  $\hat{y} = 1$  but the truth is  $y = 0$ ; or a false negative (aka missed detection), which arises when we estimate  $\hat{y} = 0$  but the truth is  $y = 1$ . The 0-1 loss treats these two kinds of errors equivalently. However, we can consider the following more general loss matrix:

where LFN is the cost of a false negative, and LFP is the cost of a false positive. The posterior expected loss for the two possible actions is given by

$$\rho(\hat{y} = 0|x) = LFN p(y = 1|x) \quad (5.111)$$

$$\rho(\hat{y} = 1|x) = LFP p(y = 0|x) \quad (5.112)$$

Hence we should pick class  $\hat{y} = 1$  if

$$\rho(\hat{y} = 0|x) > \rho(\hat{y} = 1|x) \quad (5.113)$$

$$p(y = 1|x) p(y = 0|x) > LFP LFN \quad (5.114)$$

If  $LFN = cLFP$ , it is easy to show (Exercise 5.10) that we should pick  $\hat{y} = 1$  iff  $p(y = 1|x)/p(y = 0|x) > \tau$ , where  $\tau = c/(1 + c)$  (see also (Muller et al. 2004)). For example, if a false negative costs twice as much as false positive, so  $c = 2$ , then we use a decision threshold of  $2/3$  before declaring a positive. Below we discuss ROC curves, which provide a way to study the FP-FN tradeoff without having to choose a specific threshold.

我先把你给的英文完整翻译，然后再用**基础且详细**的方式解释，包括你图里这个矩阵代表什么、公式是怎么来的、以及它在实际中的意义。

## 翻译

### 5.7.2 假阳性与假阴性的权衡

在这一节中，我们专注于**二分类决策问题**，比如假设检验、二类分类、目标/事件检测等。

我们可能会犯两种类型的错误：

1. **假阳性** (false positive, 简称 FP, 也叫误报)：当我们预测  $\hat{y} = 1$ ，但真实情况是  $y = 0$ 。
2. **假阴性** (false negative, 简称 FN, 也叫漏检)：当我们预测  $\hat{y} = 0$ ，但真实情况是  $y = 1$ 。

0-1 损失会把这两类错误看成**同样的代价**。然而，我们可以考虑如下**更一般的损失矩阵**：

	$\hat{y} = 1$	$\hat{y} = 0$
$y = 1$	0	$L_{FN}$
$y = 0$	$L_{FP}$	0

其中：

- $L_{FN}$  是假阴性的代价；
- $L_{FP}$  是假阳性的代价。

对于两种可能的决策，其后验期望损失为：

$$\rho(\hat{y} = 0 | x) = L_{FN} p(y = 1 | x) \quad (5.111)$$

$$\rho(\hat{y} = 1 | x) = L_{FP} p(y = 0 | x) \quad (5.112)$$

因此，我们应该选择  $\hat{y} = 1$  当且仅当：

$$\rho(\hat{y} = 0 | x) > \rho(\hat{y} = 1 | x) \quad (5.113)$$

即：

$$\frac{p(y = 1 | x)}{p(y = 0 | x)} > \frac{L_{FP}}{L_{FN}} \quad (5.114)$$

如果  $L_{FN} = cL_{FP}$ ，很容易证明（见习题 5.10）我们应选择  $\hat{y} = 1$  当且仅当：

$$\frac{p(y = 1 | x)}{p(y = 0 | x)} > \tau$$

其中：

$$\tau = \frac{c}{1 + c}$$

例如，如果假阴性的代价是假阳性的两倍，即  $c = 2$ ，那么在宣布“正例”之前，我们的决策阈值应该设为  $2/3$ 。

下面我们会讨论 ROC 曲线，它可以在不选择特定阈值的情况下，研究假阳性-假阴性的权衡。

## 详细解释（数学基础友好版）

### 1. 损失矩阵的含义

这个表格的行是实际情况  $y$ ，列是模型预测的类别  $\hat{y}$ ：

- **左上角 ( $y=1$ , 预测=1)**: 预测正确 → 损失 0。
- **右上角 ( $y=1$ , 预测=0)**: 假阴性 (FN)，损失是  $L_{FN}$ 。
- **左下角 ( $y=0$ , 预测=1)**: 假阳性 (FP)，损失是  $L_{FP}$ 。
- **右下角 ( $y=0$ , 预测=0)**: 预测正确 → 损失 0。

## 2. 后验期望损失公式

假设你要选择  $\hat{y} = 0$ , 那么:

- 只有在真实  $y = 1$  时才会有损失, 而且这个损失是  $L_{FN}$ ;
- 真实  $y = 1$  的概率就是  $p(y = 1 | x)$ ;
- 所以:

$$\rho(\hat{y} = 0 | x) = L_{FN} \times p(y = 1 | x)$$

同理, 如果选择  $\hat{y} = 1$ :

- 只有在真实  $y = 0$  时才会有损失  $L_{FP}$ ;
- 概率是  $p(y = 0 | x)$ ;
- 所以:

$$\rho(\hat{y} = 1 | x) = L_{FP} \times p(y = 0 | x)$$


---

## 3. 决策规则的推导

我们想选损失更小的那个动作, 所以:

$$\hat{y} = 1 \quad \text{当且仅当} \quad \rho(\hat{y} = 0 | x) > \rho(\hat{y} = 1 | x)$$

代入公式:

$$L_{FN} p(y = 1 | x) > L_{FP} p(y = 0 | x)$$

整理得到:

$$\frac{p(y = 1 | x)}{p(y = 0 | x)} > \frac{L_{FP}}{L_{FN}}$$


---

## 4. 特殊情况 $L_{FN} = cL_{FP}$

如果假阴性的代价是假阳性的  $c$  倍:

$$\frac{p(y = 1 | x)}{p(y = 0 | x)} > \frac{1}{c}$$

进一步转化为概率阈值形式:

$$p(y = 1 | x) > \frac{c}{1 + c}$$

例如:

- 如果  $c = 2$ , 即漏检的代价是误报的两倍, 那么:

$$p(y = 1 | x) > \frac{2}{3}$$

也就是说，我们更“谨慎”才判定正例，需要概率超过 66.7%。

---

## 5. 实际意义

这个方法可以：

- 在医疗检测中减少漏诊（高  $L_{FN}$ ）。
  - 在安全警报中减少误报（高  $L_{FP}$ ）。
  - 不同场景可以调整代价比  $c$  来改变分类阈值。
- 

如果你愿意，我可以帮你画一个ROC 曲线 + 不同阈值下 FP-FN 变化的示意图，这样你会直观地看到这个代价比是如何影响决策边界的。

你需要我画吗？

详细解释后验期望损失公式和决策规则的推导以及特殊情况，我有点不太理解

已思考 1m 15s

没问题！我们把三件事讲清楚：

1. 后验期望损失公式怎么来的；
  2. 最优决策规则怎么从它推出；
  3. “特殊情况”阈值到底是多少、为什么有两种写法（也是最容易混淆的地方）。
- 

### 1) 后验期望损失公式的来历

二分类、代价敏感的损失矩阵是：

		$\hat{y} = 1$	$\hat{y} = 0$
$y = 1$	0	$L_{FN}$	
	$L_{FP}$	0	

- 选择动作  $\hat{y} = a$  之后，在看到  $x$  的条件下，期望损失就是“在不同真值  $y$  下的损失  $\times$  该真值的后验概率”的和：

$$\rho(\hat{y} = a \mid x) = \sum_{y \in \{0,1\}} L(y, a) p(y \mid x).$$

- 代入上表即可得到两种动作的期望损失：

$$\rho(\hat{y} = 0 \mid x) = L_{FN} p(y = 1 \mid x) \quad (5.111)$$

$$\rho(\hat{y} = 1 \mid x) = L_{FP} p(y = 0 \mid x) \quad (5.112)$$

(因为只有“错”的那一格有非零损失。)

---

## 2) 最优决策规则的推导

选择期望损失更小的动作即可。

“预测 1”优于“预测 0”的条件是

$$\rho(\hat{y} = 0 \mid x) > \rho(\hat{y} = 1 \mid x).$$

把 (5.111)(5.112) 代入并化简：

$$L_{FN} p_1 > L_{FP} p_0 \iff \frac{p_1}{p_0} > \frac{L_{FP}}{L_{FN}}, \quad (*)$$

其中  $p_1 = p(y = 1 \mid x)$ ,  $p_0 = p(y = 0 \mid x) = 1 - p_1$ .

- 这就是后验几率 (odds) 阈值形式：

选  $\hat{y} = 1 \iff$  后验几率  $\frac{p_1}{p_0}$  大于代价比  $\frac{L_{FP}}{L_{FN}}$ .

再把它换成概率阈值形式。由  $p_0 = 1 - p_1$  可得

$$L_{FN} p_1 > L_{FP} (1 - p_1) \iff p_1 > \frac{L_{FP}}{L_{FN} + L_{FP}}. \quad (**)$$

所以：

$$\hat{y} = 1 \iff p(y = 1 \mid x) > \frac{L_{FP}}{L_{FN} + L_{FP}}$$

- 若  $L_{FN} = L_{FP}$ , 阈值 = 1/2, 就回到了 0-1 损失/MAP 的规则。
  - $L_{FN}$  越大 (漏检更贵), 阈值越低, 更容易判为正例;  
 $L_{FP}$  越大 (误报更贵), 阈值越高, 更谨慎地判正例。  
 这与直觉完全一致。
- 

## 3) “特殊情况”的两种写法为何容易混淆?

文中写“若  $L_{FN} = c L_{FP}$ ”, 我们用上面的概率阈值式 (\*\*) 代入：

$$p_1 > \frac{L_{FP}}{c L_{FP} + L_{FP}} = \frac{1}{1 + c}.$$

所以当  $L_{FN} = c L_{FP}$  (漏检是误报的  $c$  倍) 时：

$$p(y = 1 | x) > \frac{1}{1 + c}$$

举例:  $c = 2$  (漏检代价是误报的 2 倍), 阈值 =  $1/3$ 。

——这表示: \*\*更愿意报“正”\*\*以减少漏检 (阈值更低), 符合常识。

很多资料也写成**几率阈值**: 由 (\*) 得

$$\frac{p_1}{p_0} > \frac{L_{FP}}{L_{FN}} = \frac{1}{c}.$$

把它再换回概率就是刚才的  $1/(1 + c)$ 。

有时你会看到另一种写法: 阈值  $c/(1 + c)$ 。那实际上对应的是把

$c$  定义成  $c = \frac{L_{FP}}{L_{FN}}$  (而不是我们这里的  $c = \frac{L_{FN}}{L_{FP}}$ )。

两个定义刚好互为倒数, 所以得到的阈值形式不同, 但**本质完全等价**。

小心辨认  $c$  的定义就不会混乱。

## 小结 + 快速心法

- **期望损失**:  $\rho(\hat{y} = a|x) = \sum L(y, a)p(y|x)$ 。
- **最优决策**: 选期望损失更小的动作。
- **最终阈值**:

$$\hat{y} = 1 \Leftrightarrow \frac{p_1}{p_0} > \frac{L_{FP}}{L_{FN}} \Leftrightarrow p_1 > \frac{L_{FP}}{L_{FN} + L_{FP}}.$$

- **直觉**: 漏检更贵  $\rightarrow$  阈值降; 误报更贵  $\rightarrow$  阈值升。

如果你愿意, 我可以用一个**具体数值例子 (给定  $L_{FP}, L_{FN}$  和若干  $p(y = 1|x)$ )**, 列出每个样本该如何判决, 顺便画出 ROC 点, 帮助你把“代价–阈值–决策”的对应关系彻底吃透。



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。

Table 5.2 Quantities derivable from a confusion matrix.  $N_+$  is the true number of positives,  $\hat{N}_+$  is the “called” number of positives,  $N_-$  is the true number of negatives,  $\hat{N}_-$  is the “called” number of negatives.

Table 5.3 Estimating  $p(\hat{y}|y)$  from a confusion matrix.  
Abbreviations: FNR = false negative rate, FPR = false positive rate, TNR = true negative rate, TPR = true positive rate.

Table 5.4 Estimating  $p(y|\hat{y})$  from a confusion matrix.  
Abbreviations: FDP = false discovery probability, NPV = negative predictive value, PPV = positive predictive value,  
Figure 5.15 (a) ROC curves for two hypothetical classification systems. A is better than B. We plot the true positive rate (TPR) vs the false positive rate (FPR) as we vary the threshold  $\tau$ . We also indicate the equal error rate (EER) with the red and blue dots, and the area under the curve (AUC) for classifier B. (b) A precision-recall curve for two hypothetical classification systems. A is better than B. Figure generated by PRhand.

### 5.7.2.1 ROC curves and all that

Suppose we are solving a binary decision problem, such as classification, hypothesis testing, object detection, etc. Also, assume we have a labeled data set,  $D = \{(x_i, y_i)\}$ . Let  $\delta(x) = I(f(x) > \tau)$  be our decision rule, where  $f(x)$  is a measure of confidence that  $y = 1$  (this should be monotonically related to  $p(y = 1|x)$ , but does not need to be a probability), and  $\tau$  is some threshold parameter. For each given value of  $\tau$ , we can apply our decision rule and count the number of true positives, false positives, true negatives, and false negatives that occur, as shown in Table 5.2. This table of errors is called a confusion matrix. From this table, we can compute the true positive rate (TPR), also known as the sensitivity, recall or hit rate, by using  $TPR = TP/N_+ \approx p(\hat{y} = 1|y = 1)$ . We can also compute the false positive rate (FPR), also called the false alarm rate, or the type I error rate, by using  $FPR = FP/N_- \approx p(\hat{y} = 1|y = 0)$ . These and other definitions are summarized in Tables 5.3 and 5.4. We can combine these errors in any way we choose to compute a loss function. However, rather than computing the TPR and FPR for a fixed threshold  $\tau$ , we can run our detector for a set of thresholds, and then plot the TPR vs FPR as an implicit function of  $\tau$ . This is called a receiver operating

characteristic or ROC curve. See Figure 5.15(a) for an example. Any system can achieve the point on the bottom left, ( $FPR = 0, TPR = 0$ ), by setting  $\tau = 1$  and thus classifying everything as negative; similarly any system can achieve the point on the top right, ( $FPR = 1, TPR = 1$ ), by setting  $\tau = 0$  and thus classifying everything as positive. If a system is performing at chance level, then we can achieve any point on the diagonal line  $TPR = FPR$  by choosing an appropriate threshold. A system that perfectly separates the positives from negatives has a threshold that can achieve the top left corner, ( $FPR = 0, TPR = 1$ ); by varying the threshold such a system will “hug” the left axis and then the top axis, as shown in Figure 5.15(a). The quality of a ROC curve is often summarized as a single number using the area under the curve or AUC. Higher AUC scores are better; the maximum is obviously 1. Another summary statistic that is used is the equal error rate or EER, also called the cross over rate, defined as the value which satisfies  $FPR = FNR$ . Since  $FNR = 1 - TPR$ , we can compute the EER by drawing a line from the top left to the bottom right and seeing where it intersects the ROC curve (see points A and B in Figure 5.15(a)). Lower EER scores are better; the minimum is obviously 0.

已思考 17s

下面先把原文完整翻译，再给你基础且详尽的解释与一个小例子。

## 逐句翻译

**表 5.2** 由混淆矩阵可以推导的各个数量。

$N_+$  为真实正例的数量,  $\hat{N}_+$  为被“判成正例”的数量;

$N_-$  为真实负例的数量,  $\hat{N}_-$  为被“判成负例”的数量。

**表 5.3** 由混淆矩阵估计  $p(\hat{y} | y)$ 。

缩写:  $FNR = \text{false negative rate}$  (假阴性率) ,  $FPR = \text{false positive rate}$  (假阳性率) ,

$TNR = \text{true negative rate}$  (真阴性率) ,  $TPR = \text{true positive rate}$  (真正率) 。

**表 5.4** 由混淆矩阵估计  $p(y | \hat{y})$ 。

缩写:  $FDP = \text{false discovery probability}$  (错误发现概率) ,

$NPV = \text{negative predictive value}$  (阴性预测值) ,

$PPV = \text{positive predictive value}$  (阳性预测值/精准率) 。

## 图 5.15

(a) 两个假想分类系统的 ROC 曲线, A 优于 B。我们把阈值  $\tau$  改变时的真正率 (TPR) 与假

阳性率 (FPR) 画成关系曲线。图中标出了等错误率 (EER) 的红蓝点，以及分类器 B 的曲线下面积 (AUC)。

(b) 两个系统的精确率–召回率 (precision-recall) 曲线，A 优于 B。图由 PRhand 生成。

---

### 5.7.2.1 ROC 曲线及相关概念

设我们在做二元决策 (分类、假设检验、目标检测等)，并且有带标签的数据集  $D = \{(x_i, y_i)\}$ 。

令  $\delta(x) = \mathbf{1}[f(x) > \tau]$  为决策规则，其中  $f(x)$  是“属于正类”的置信度 (与  $p(y = 1 | x)$  单调相关，但不一定是概率)， $\tau$  是阈值。

对每一个给定的  $\tau$ ，应用这个规则，就能数出真正例 TP、假阳性 FP、真阴性 TN、假阴性 FN，如表 5.2 所示。这张误差表叫**混淆矩阵**。

由此可计算：

- **真正率 TPR** (又称灵敏度/召回/命中率)： $TPR = TP / N_+ \approx p(\hat{y} = 1 | y = 1)$ 。
- **假阳性率 FPR** (又称误报率/I型错误率)： $FPR = FP / N_- \approx p(\hat{y} = 1 | y = 0)$ 。  
这些以及其它定义汇总在表 5.3 和表 5.4。

与其在固定阈值下报一个 TPR/FPR，不如在一组阈值上运行检测器，把 TPR 对 FPR 作图，这就是**ROC 曲线** (receiver operating characteristic)，见图 5.15(a)。

- 任何系统把  $\tau = 1$  (全判负) 时到达左下角 ( $FPR = 0, TPR = 0$ )；把  $\tau = 0$  (全判正) 时到达右上角 ( $1, 1$ )。
- 若系统只是碰运气，则曲线在对角线  $TPR = FPR$  上。
- 若能完美区分正负，则可达左上角 ( $0, 1$ )，并且曲线先沿左边界、再沿上边界。

AUC (曲线下面积) 常用来把 ROC 质量概括为一个数：越大越好，最大为 1。

EER (等错误率/交叉点) 定义为满足  $FPR = FNR$  的点。因  $FNR = 1 - TPR$ ，可画一条从左上到右下的直线，与 ROC 的交点即为 EER (越小越好，最小 0)。

---

## 混淆矩阵与各指标：通俗解释

### 1) 混淆矩阵与基本计数

- TP：真实 1、预测 1
- FP：真实 0、预测 1 (误报)
- FN：真实 1、预测 0 (漏报)
- TN：真实 0、预测 0

行/列合计：

- $N_+ = TP + FN$  (真实正例数)
- $N_- = FP + TN$  (真实负例数)

- $\hat{N}_+ = TP + FP$  (被判为正的个数)

- $\hat{N}_- = FN + TN$  (被判为负的个数)

- 总数  $N = TP + FP + FN + TN$

## 2) 从“真值条件”看分类器 $p(\hat{y} | y)$ (表 5.3)

- TPR =  $TP/N_+ = p(\hat{y} = 1 | y = 1)$  (也叫 recall/sensitivity)
- FNR =  $FN/N_+ = p(\hat{y} = 0 | y = 1) = 1 - \text{TPR}$
- FPR =  $FP/N_- = p(\hat{y} = 1 | y = 0)$
- TNR =  $TN/N_- = p(\hat{y} = 0 | y = 0) = 1 - \text{FPR}$

## 3) 从“预测条件”看真实 $p(y | \hat{y})$ (表 5.4)

- PPV =  $TP/\hat{N}_+$  (precision, 阳性预测值)
- FDP =  $FP/\hat{N}_+ = 1 - \text{PPV}$  (错误发现概率)
- NPV =  $TN/\hat{N}_-$  (阴性预测值)

小心: ROC 用的是  $p(\hat{y} | y)$ , 而 PR 曲线更关注  $p(y | \hat{y})$  (尤其是 PPV/precision), 在 **正例极少时**更有判别力。

## ROC/PR 与阈值: 怎么画、怎么看

1. 选一个打分函数  $f(x)$ ; 扫一系列阈值  $\tau$ 。
2. 每个  $\tau$  得到一组  $(TP, FP, TN, FN)$ 。
3. 画 ROC: 点  $(\text{FPR}, \text{TPR}) = (FP/N_-, TP/N_+)$ 。
4. 画 PR: 点  $(\text{recall}, \text{precision}) = (\text{TPR}, \text{PPV})$ 。
5. AUC 越大越好; EER 越小越好。
6. 类不平衡时, ROC 较“乐观”; PR 曲线更反映“检到的有多准” (precision)。

## 一个小例子 (把名词落地)

设真正正例 100 个、负例 900 个。对某个阈值得到:

- $TP = 60, FN = 40, FP = 45, TN = 855$

计算:

- $\text{TPR} = 60/100 = 0.60$  (召回)
- $\text{FPR} = 45/900 = 0.05$
- $\text{PPV} = 60/(60 + 45) = 0.571$  (精确率)
- $\text{NPV} = 855/(855 + 40) = 0.955$

ROC 上是点  $(0.05, 0.60)$ ; PR 上是点  $(\text{recall} = 0.60, \text{precision} = 0.571)$ 。

## 和代价敏感决策的联系（提要）

在 ROC 平面上，给定“先验/代价”可画出一族**等期望损失直线**，最优运行点是该直线与 ROC 的切点；等价于在概率空间用

$$p(y = 1 | x) > \frac{L_{FP}}{L_{FN} + L_{FP}}$$

来确定阈值。ROC/PR 则是在**不固定阈值**时全面比较模型能力。

如果你想，我可以用你的一组真实或模拟的分数  $f(x)$  和标签，帮你**计算混淆矩阵**→**画 ROC/PR 曲线**→**算 AUC/EER**，把这些指标——对上。

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。

### 5.7.2.2 Precision recall curves

When trying to detect a rare event (such as retrieving a relevant document or finding a face in an image), the number of negatives is very large. Hence comparing TPR = TP/N+ to FPR = FP/N- is not very informative, since the FPR will be very small. Hence all the “action” in the ROC curve will occur on the extreme left. In such cases, it is common to plot the TPR versus the number of false positives, rather than vs the false positive rate. However, in some cases, the very notion of “negative” is not well-defined. For example, when detecting objects in images (see Section 1.2.1.3), if the detector works by classifying patches, then the number of patches examined — and hence the number of true negatives — is a parameter of the algorithm, not part of the problem definition. So we would like to use a measure that only talks about positives. The precision is defined as  $TP / N+ = p(y=1|y=1)$  and the recall is defined as  $TP/N+ = p(\hat{y}=1|y=1)$ . Precision measures what fraction of our detections are actually positive, and recall measures what fraction of the positives we actually detected. If  $\hat{y}_i \in \{0, 1\}$  is the predicted label, and  $y_i \in \{0, 1\}$  is the true label, we can estimate precision and recall using

$$P = \sum_i \hat{y}_i \sum_i \hat{y}_i, R = \sum_i \hat{y}_i \sum_i y_i \quad (5.115)$$

A precision recall curve is a plot of precision vs recall as we vary the threshold  $\tau$ . See Figure 5.15(b). Hugging the top right is the best one can do. This curve can be

summarized as a single number using the mean precision (averaging over recall values), which approximates the area under the curve. Alternatively, one can quote the precision for a fixed recall level, such as the precision of the first K = 10 entities recalled. This is called the average precision at K score. This measure is widely used when evaluating information retrieval systems.

好的，我会先完整翻译原文，然后用基础且详尽的方式解释每一个概念，并给出直观的例子帮助你理解。

---

## 逐句翻译

### 5.7.2.2 精确率–召回率曲线 (Precision–Recall curves)

当我们尝试检测一个**稀有事件**（例如检索到一个相关文档，或在图像中找到一张人脸）时，\*\*负例（negative）\*\*的数量会非常大。

因此，如果比较

$$\text{TPR} = \frac{\text{TP}}{N_+}$$

和

$$\text{FPR} = \frac{\text{FP}}{N_-}$$

就不太有信息量，因为 FPR 会非常小。于是 ROC 曲线上的所有“动作”都会集中在最左边。

在这种情况下，人们通常绘制 \*\*TPR 与假阳性数（false positives）\*\* 的关系，而不是 TPR 与假阳性率（FPR）的关系。

但是，有时“负例”的概念本身并不明确。

例如，在图像目标检测（见 1.2.1.3 节）中，如果检测器的工作方式是**对图像的局部区域（patch）分类**，那么被检测的 patch 数量——也就是真实负例的数量——其实是算法的一个参数，而不是问题定义的一部分。

因此，我们希望使用一个**只关注正例的指标**。

- **精确率（Precision）** 定义为

$$\frac{\text{TP}}{N_+} = p(y = 1 | \hat{y} = 1)$$

即：被判为正的样本中，真正为正的比例。

- **召回率（Recall）** 定义为

$$\frac{\text{TP}}{N_+} = p(\hat{y} = 1 \mid y = 1)$$

即：真实正例中，被检测出来的比例。

如果  $\hat{y}_i \in \{0, 1\}$  是预测标签， $y_i \in \{0, 1\}$  是真实标签，我们可以用以下公式估计精确率和召回率：

$$P = \frac{\sum_i y_i \hat{y}_i}{\sum_i \hat{y}_i}, \quad R = \frac{\sum_i y_i \hat{y}_i}{\sum_i y_i} \quad (5.115)$$

**精确率—召回率曲线**是：当我们改变阈值  $\tau$  时，把精确率与召回率画出来的曲线（见图 5.15(b)）。

在右上角“贴近”才是最好的表现。

这个曲线可以用一个数字来概括：

- **平均精确率 (mean precision)**：在不同召回率下的精确率取平均，这大致等于曲线下的面积。
- 或者，也可以在固定召回水平下报告精确率，例如前  $K = 10$  个被检出的实体的精确率，这称为 **平均精确率@K (average precision at K)**。

这个指标在**信息检索系统评估**中被广泛使用。

## 通俗解释 + 公式分解

### 1. 为什么不用 FPR (假阳性率)？

在稀有事件场景中（正例很少），即使有很多假阳性（FP），除以庞大的负例数  $N_-$  后，FPR 也会非常小，看起来“几乎没错”，但其实误报很多。

比如：假设有 100 万个样本，正例只有 100 个。即使你误报了 1000 个负例， $FPR = 1000 / 999900 \approx 0.001$ ，看起来很小，但其实 1000 个误报是个很严重的问题。

### 2. Precision (精确率)

$$\text{Precision} = \frac{\text{预测为正且正确的数量}}{\text{预测为正的总数量}}$$

它回答的问题是：“我说是正的，这些预测有多准？”

### 3. Recall (召回率)

$$\text{Recall} = \frac{\text{预测为正且正确的数量}}{\text{真实正例的总数量}}$$

它回答的问题是：“所有真实正例里，我找到了多少？”

### 4. Precision–Recall 曲线

- 横轴：Recall
- 纵轴：Precision

- 阈值  $\tau$  从高到低变化：
  - 高阈值 → 只选最有把握的 → Precision 高, Recall 低
  - 低阈值 → 更多样本被预测为正 → Recall 高, 但 Precision 可能下降
- 右上角是理想情况 (Precision=1, Recall=1)

## 5. 为什么 PR 曲线更适合稀有事件?

因为它只关注正例 (TP、FP、FN) , 不受庞大负例数量  $N_-$  的影响。

## 6. 平均精确率 (Mean Precision / AP)

- 曲线下的面积 (AUC 类似, 但针对 PR)
- 数值越高越好
- 信息检索里也常用 Precision@K (看前 K 个结果的精准度)

## 例子

假设有一个搜索系统, 返回了 10 个结果, 其中:

- 6 个是相关的 (TP)
- 4 个是无关的 (FP)
- 实际上总共有 12 个相关结果 ( $N_+ = 12$ )

那么:

- Precision =  $6 / 10 = 0.6$  (搜索结果里有 60% 是对的)
- Recall =  $6 / 12 = 0.5$  (所有相关结果里找到了 50%)

如果把阈值调低, 多返回一些结果, Recall 可能会上升 (比如找到 9/12) , 但 Precision 可能下降 (因为多了一些 FP) 。

如果你愿意, 我可以帮你画一个ROC vs PR 对比图, 用一个稀有事件的模拟数据, 让你直观看到为什么在这种情况下 PR 曲线更有用。这样你会很快掌握它的优势。



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。

Table 5.5 Illustration of the difference between macro- and micro-averaging.  $y$  is the true label, and  $\hat{y}$  is the called label. In this example, the macro-averaged precision is  $[10/(10 + 10) + 90/(10 + 90)]/2 =$

$(0.5+0.9)/2=0.7$ . The micro-averaged precision is  $100/(100 + 20) \approx 0.83$ . Based on Table 13.7 of (Manning et al. 2008).

### 5.7.2.3 F-scores

For a fixed threshold, one can compute a single precision and recall value. These are often combined into a single statistic called the Fscore, or F1 score, which is the harmonic mean of precision and recall:

$$F1 = 2P R / (P + R) = 2PR / (R + P) \quad (5.116)$$

Using Equation 5.115, we can write this as

$$F1 = 2 \frac{\sum_i y_i \hat{y}_i}{\sum_i y_i + \sum_i \hat{y}_i} \quad (5.117)$$

This is a widely used measure in information retrieval systems. To understand why we use the harmonic mean instead of the arithmetic mean,  $(P + R)/2$ , consider the following scenario. Suppose we recall all entries, so  $R = 1$ . The precision will be given by the prevalence,  $p(y = 1)$ . Suppose the prevalence is low, say  $p(y = 1) = 10 - 4$ . The arithmetic mean of  $P$  and  $R$  is given by  $(P + R)/2 = (10 - 4 + 1)/2 \approx 50\%$ . By contrast, the harmonic mean of this strategy is only  $2 \times 10 - 4 \times 1 / (1 + 10 - 4) \approx 0.2\%$ . In the multi-class case (e.g., for document classification problems), there are two ways to generalize F1 scores. The first is called macro-averaged F1, and is defined as  $\bar{F1} = \frac{1}{C} \sum_c F1(c)$ , where  $F1(c)$  is the F1 score obtained on the task of distinguishing class  $c$  from all the others. The other is called micro-averaged F1, and is defined as the F1 score where we pool all the counts from each class's contingency table. Table 5.5 gives a worked example that illustrates the difference. We see that the precision of class 1 is 0.5, and of class 2 is 0.9. The macro-averaged precision is therefore 0.7, whereas the micro-averaged precision is 0.83. The latter is much closer to the precision of class 2 than to the precision of class 1, since class 2 is five times larger than class 1. To give equal weight to each class, use macro-averaging.

好，我会先逐句翻译原文，再用基础且详细的方式解释，并结合你给的表格进行计算演示。

## 逐句翻译

表 5.5 宏平均 (macro-averaging) 与微平均 (micro-averaging) 的差异示例。

$y$  是真实标签， $\hat{y}$  是预测标签。

在这个例子中，宏平均精确率 (macro-averaged precision) 为：

$$\frac{\frac{10}{10+10} + \frac{90}{10+90}}{2} = \frac{0.5 + 0.9}{2} = 0.7$$

微平均精确率 (micro-averaged precision) 为:

$$\frac{100}{100 + 20} \approx 0.83$$

该示例基于 (Manning et al., 2008) 的表 13.7。

---

### 5.7.2.3 F-分数 (F-scores)

对于一个固定的阈值，我们可以计算出一个精确率 (precision) 和召回率 (recall) 值。

它们常被组合成一个统计量，称为 **F-score** 或 **F1 分数**，定义为精确率与召回率的调和平均数：

$$F_1 = \frac{2}{\frac{1}{P} + \frac{1}{R}} = \frac{2PR}{P + R} \quad (5.116)$$

利用公式 (5.115)，我们可以写成：

$$F_1 = \frac{2 \sum_{i=1}^N y_i \hat{y}_i}{\sum_{i=1}^N y_i + \sum_{i=1}^N \hat{y}_i} \quad (5.117)$$

这是信息检索系统中广泛使用的指标。

为了理解为什么我们用调和平均而不是算术平均  $(P + R)/2$ ，考虑以下情况：

假设我们召回了所有样本（即  $R = 1$ ），那么精确率就等于正例比例  $p(y = 1)$ 。

假设正例比例很低，比如  $p(y = 1) = 10^{-4}$ （万分之一）。

- 算术平均为：

$$\frac{P + R}{2} = \frac{10^{-4} + 1}{2} \approx 50\%$$

看起来很高，但其实没啥用。

- 调和平均为：

$$\frac{2 \times 10^{-4} \times 1}{1 + 10^{-4}} \approx 0.02\%$$

这个值才更能反映模型性能的真实水平。

---

在多类别任务（例如文档分类）中，F1 分数有两种推广方式：

1. **宏平均 F1 (macro-averaged F1) :**

$$\frac{1}{C} \sum_{c=1}^C F_1(c)$$

其中  $F_1(c)$  是将类别  $c$  与所有其它类别区分开来时的 F1 分数。

## 2. 微平均 F1 (micro-averaged F1) :

先将所有类别的混淆矩阵 (contingency table) 计数加总，然后计算整体的 F1 分数。

表 5.5 给出了一个例子来说明它们的区别。

- 类别 1 的精确率是 0.5
- 类别 2 的精确率是 0.9
- 因此宏平均精确率是 0.7
- 微平均精确率是 0.83

由于类别 2 的样本数是类别 1 的 5 倍，所以微平均结果更接近类别 2 的精确率。

如果希望**每个类别权重相等**，就用宏平均。

---

## 详细解释

### 1. 表格中的数字含义

每个小表格是一个混淆矩阵 (confusion matrix) :

- 行：预测标签 ( $\hat{y}$ )
  - 列：真实标签 ( $y$ )
  - 例：Class 1 中，预测 1 且真实 1 的有 10 个 (TP)，预测 1 但真实 0 的有 10 个 (FP)，等等。
- 

### 2. 宏平均 vs 微平均

#### • 宏平均 (Macro) :

对每个类别单独算精确率 (Precision)，然后取平均。**不管类别样本量多少，权重相同。**

例子 (表中数据) :

- Class 1:  $P_1 = \frac{10}{10+10} = 0.5$
- Class 2:  $P_2 = \frac{90}{90+10} = 0.9$
- 宏平均:  $\frac{0.5+0.9}{2} = 0.7$

#### • 微平均 (Micro) :

把所有类别的 TP、FP 总数加起来后再计算 Precision。**类别样本量大的权重大。**

Pooled 表:

- 总 TP = 100

- 总 FP = 20
  - 微平均精确率 =  $\frac{100}{100+20} \approx 0.833$
- 

### 3. F1 分数

- 结合 Precision 和 Recall
  - 用**调和平均**, 因为它能惩罚一边高、一边低的情况
  - 如果某一边很低, F1 也会很低 → 更真实地反映模型性能
  - 在多类别问题中, 宏/微两种方式也适用于 F1
- 

### 4. 为什么调和平均比算术平均好

如果 Recall = 1 (全找回), 但 Precision 极低 (比如 0.0001), 算术平均会被高的 Recall 拉上去, 看起来还不错, 但实际模型没用。

调和平均会被低值强烈拉低, 更符合实际。

---

如果你愿意, 我可以帮你用表 5.5 里的数据**算出宏平均 F1 和微平均 F1**, 让你一步步看到它们的差别。这样你会更直观地理解“权重差异”的影响。

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。

#### 5.7.2.4 False discovery rates

Suppose we are trying to discover a rare phenomenon using some kind of high throughput measurement device, such as a gene expression micro array, or a radio telescope. We will need to make many binary decisions of the form  $p(y_i = 1|D) > \tau$ , where  $D = \{x_i\}_N$   $i=1$  and  $N$  may be large. This is called multiple hypothesis testing. Note that the difference from standard binary classification is that we are classifying  $y_i$  based on all the data, not just based on  $x_i$ . So this is a simultaneous classification problem, where we might hope to do better than a series of individual classification problems. How should we set the threshold  $\tau$ ? A natural approach is to try to minimize the expected number of false positives. In the Bayesian approach, this can be computed as follows:

$$FD(\tau, D) = \sum_i (1 - p_i) \cdot \text{pr. error } I(p_i > \tau) \cdot \text{$\# $ discovery}$$

(5.118)

where  $p_i = p(y_i = 1|D)$  is your belief that this object exhibits the phenomenon in question. We then define the posterior expected false discovery rate as follows:

$$FDR(\tau, D) = FD(\tau, D)/N(\tau, D) \quad (5.119)$$

where  $N(\tau, D) = \sum_i I(p_i > \tau)$  is the number of discovered items. Given a desired FDR tolerance, say  $\alpha = 0.05$ , one can then adapt  $\tau$  to achieve this; this is called the direct posterior probability approach to controlling the FDR (Newton et al. 2004; Muller et al. 2004). In order to control the FDR it is very helpful to estimate the  $p_i$ 's jointly (e.g., using a hierarchical Bayesian model, as in Section 5.5), rather than independently. This allows the pooling of statistical strength, and thus lower FDR. See e.g., (Berry and Hochberg 1999) for more information.

好，我会先逐句翻译原文，再用基础且详细的方式解释，并且帮你把公式 (5.118) 和 (5.119) 的含义一点点拆开。

## 逐句翻译

### 5.7.2.4 假发现率 (False discovery rates)

假设我们正在用某种高通量测量设备（比如基因表达芯片或射电望远镜）去发现一种稀有现象。

我们需要做大量的二元（是/否）决策，形式为：

$$p(y_i = 1|D) > \tau$$

其中  $D = \{x_i\}_{i=1}^N$ ，并且  $N$  可能很大。

这叫做**多重假设检验** (multiple hypothesis testing)。

注意，这和标准的二分类任务不同：

在这里，我们是在利用**全部数据 D** 来对每个  $y_i$  做判断，而不是仅用它自己的  $x_i$ 。

因此，这是一个**同时分类问题** (simultaneous classification)，有希望比一系列独立分类问题做得更好。

## 阈值 $\tau$ 怎么设？

一种自然的方法是尽量最小化假阳性 (false positives) 的期望数。

在贝叶斯方法中，可以这样计算：

$$FD(\tau, D) = \sum_i \underbrace{(1 - p_i)}_{\text{出错的概率被判定为发现}} \underbrace{I(p_i > \tau)}_{\text{ }} \quad (5.118)$$

这里：

- $p_i = p(y_i = 1 | D)$  是你对第  $i$  个对象确实表现出该现象的置信度（后验概率）
  - $1 - p_i$  是它其实是假的概率（即错误概率）
  - $I(p_i > \tau)$  是指示函数（indicator function），如果  $p_i > \tau$  则值为 1，否则为 0，用来判断这个对象是否被标记为“发现”
- 

接着我们定义**后验期望假发现率**（posterior expected false discovery rate）：

$$FDR(\tau, D) = \frac{FD(\tau, D)}{N(\tau, D)} \quad (5.119)$$

其中：

$$N(\tau, D) = \sum_i I(p_i > \tau)$$

是被标记为“发现”的对象总数。

---

如果我们希望 FDR 控制在某个容忍水平，比如  $\alpha = 0.05$ ，那么我们就可以调节  $\tau$  来实现这一点。

这种方法叫做**直接后验概率法控制假发现率**（direct posterior probability approach to controlling the FDR）（Newton et al. 2004; Muller et al. 2004）。

为了更好地控制 FDR，最好是**联合地**（jointly）**估计**这些  $p_i$ （例如用第 5.5 节介绍的分层贝叶斯模型 hierarchical Bayesian model），而不是独立估计。

这样可以**汇聚统计力量**（pooling statistical strength），从而降低 FDR。

更多信息可参考 Berry and Hochberg (1999)。

---

## 基础且详细的解释

### 1. 这个问题要解决什么？

- 我们有很多个对象（如基因、天文信号等）
  - 每个对象都有可能是“真阳性”（确实有我们想找的现象）或“假阳性”（其实没有，只是看起来像有）
  - 我们需要设一个**阈值**  $\tau$ ，大于这个值的概率就认为是“发现”
  - 但我们要小心**假发现率**（FDR），即“我们认为发现的东西里面，多少是假的”
- 

### 2. 公式 (5.118) 的含义

- 对每个对象：

- $p_i$  = 它是真的的概率 (来自贝叶斯后验)
  - $1 - p_i$  = 它是假的的概率
  - $I(p_i > \tau) =$  如果它被我们列为“发现”，则记 1，否则 0
  - $(1 - p_i) \cdot I(p_i > \tau) =$  “它是假的”  $\times$  “它被我们列为发现”  $\rightarrow$  这是一个假发现的概率
  - 对所有对象求和  $\rightarrow$  得到**期望的假发现个数**  $FD(\tau, D)$
- 

### 3. 公式 (5.119) 的含义

- $N(\tau, D)$  = 总的发现数
  - **FDR** = **期望假发现个数 / 总发现数**  
 $\rightarrow$  发现里有多少比例是假的
- 

### 4. 如何用 FDR 控制 $\tau$

- 如果我们希望  $FDR \leq 0.05$  (假发现比例不超过 5%)
  - 就调节  $\tau$ , 直到计算出来的  $FDR(\tau, D)$  达到或低于 0.05
- 

### 5. 为什么要联合估计 $p_i$

- 如果单独估计, 每个对象只能用自己的数据, 信息有限
  - 如果联合建模 (比如分层贝叶斯), 可以让所有对象共享一些统计信息 (比如噪声水平、先验分布等)
  - 这样得到的  $p_i$  更准确, FDR 控制也更稳定
- 

如果你愿意, 我可以帮你画一个示意图, 展示 FDR 随  $\tau$  变化的趋势, 让你看到为什么调节  $\tau$  可以控制假发现比例。这样你会对公式的实际意义有更直观的感觉。

你要我画吗?

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。

#### 5.7.3 Other topics

In this section, we briefly mention a few other topics related to Bayesian decision theory. We do not have space to go into detail, but we include pointers to the relevant literature.

### 5.7.3.1 Contextual bandits

A one-armed bandit is a colloquial term for a slot machine, found in casinos around the world. The game is this: you insert some money, pull an arm, and wait for the machine to stop; if you're lucky, you win some money. Now imagine there is a bank of  $K$  such machines to choose from. Which one should you use? This is called a multi-armed bandit, and can be modeled using Bayesian decision theory: there are  $K$  possible actions, and each action has an unknown reward (payoff function)  $r_k$ . By maintaining a belief state,  $p(r_1|D) = k p(r_k|D)$ , one can devise an optimal policy; this can be compiled into a series of Gittins Indices (Gittins 1989). This optimally solves the exploration-exploitation tradeoff, which specifies how many times one should try each action before deciding to go with the winner. Now consider an extension where each arm, and the player, has an associated feature vector; call all these features  $x$ . This is called a contextual bandit (see e.g., (Sarkar 1991; Scott 2010; Li et al. 2011)). For example, the "arms" could represent ads or news articles which we want to show to the user, and the features could represent properties of these ads or articles, such as a bag of words, as well as properties of the user, such as demographics. If we assume a linear model for reward,  $r_k = \theta^T k x$ , we can maintain a distribution over the parameters of each arm,  $p(\theta_k|D)$ , where  $D$  is a series of tuples of the form  $(a, x, r)$ , which specifies which arm was pulled, what its features were, and what the resulting outcome was (e.g.,  $r = 1$  if the user clicked on the ad, and  $r = 0$  otherwise). We discuss ways to compute  $p(\theta_k|D)$  from linear and logistic regression models in later chapters. Given the posterior, we must decide what action to take. One common heuristic, known as UCB (which stands for "upper confidence bound") is to take the action which maximizes  $k^* = K \operatorname{argmax} k=1 \mu_k + \lambda \sigma_k$  (5.120)

where  $\mu_k = E[r_k|D]$ ,  $\sigma_k^2 = \operatorname{var}[r_k|D]$  and  $\lambda$  is a tuning parameter that trades off exploration and exploitation. The intuition is that we should pick actions about which we believe are good ( $\mu_k$  is large), and/or actions about which we are uncertain ( $\sigma_k$  is large). An even simpler method, known as Thompson sampling, is as follows. At each step, we pick action  $k$  with a probability that is equal to its probability of being the optimal action:

$$p_k = I(E[r|a, x, \theta] = \max_a E[r|a, x, \theta]) p(\theta|D) d\theta \quad (5.121)$$

We can approximate this by drawing a single sample from the posterior,  $\theta_t \sim p(\theta|D)$ , and then choosing  $k^*$

=argmax<sub>k</sub> E r|x, k, θ<sub>t</sub>. Despite its simplicity, this has been shown to work quite well (Chapelle and Li 2011).

好，我会先把原文逐句翻译，然后再用**简单且详细的方式**解释，包括一些直观例子帮你理解公式 (5.120) 和 (5.121)。

---

## 逐句翻译

### 5.7.3 其他主题

在本节中，我们简要提到一些与贝叶斯决策理论相关的其他主题。

我们没有空间深入细节，但会给出相关文献的参考。

---

#### 5.7.3.1 上下文老虎机 (Contextual bandits)

“单臂老虎机” (one-armed bandit) 是赌场中常见的一种老虎机的俗称。

玩法是：投币 → 拉下拉杆 → 等机器停下；如果运气好，就能赢钱。

现在想象有一排  $K$  台这样的机器可供选择。

你应该玩哪一台？

这就叫做**多臂老虎机问题** (multi-armed bandit)，可以用贝叶斯决策理论建模：

有  $K$  个可能的动作（选择哪台机器），每个动作都有一个未知的回报（收益函数） $r_k$ 。

通过维持一个信念状态 (belief state)：

$$p(r_{1:K}|D) = \prod_k p(r_k|D)$$

就可以设计一个最优策略，这个策略可以转化成一系列 **Gittins 指数** (Gittins 1989)。

它可以最优地解决**探索-利用权衡** (exploration-exploitation tradeoff)：

也就是应该尝试每个动作多少次，然后决定选哪一个。

---

现在考虑一个扩展：每个拉杆（机器）以及玩家，都有一个特征向量；

把所有这些特征叫做  $x$ 。

这种情况叫做**上下文老虎机** (contextual bandit)（见 Sarkar 1991; Scott 2010; Li et al. 2011）。

例如，“拉杆”可能是我们想展示给用户的广告或新闻文章；

特征可能包括广告或文章的属性（比如词袋 bag-of-words），以及用户的属性（比如人口统计信息）。

如果我们假设回报是一个线性模型：

$$r_k = \theta_k^T x$$

我们就可以为每个“拉杆”的参数维持一个分布  $p(\theta_k|D)$ ,

其中  $D$  是一系列元组  $(a, x, r)$ :

- $a$  = 选择了哪台拉杆 (哪个广告)
- $x$  = 它的特征
- $r$  = 结果 (比如用户是否点击广告, 点击  $r = 1$ , 没点  $r = 0$ )

后面章节会讲如何从线性回归或逻辑回归模型中计算  $p(\theta_k|D)$ 。

---

得到后验分布后, 我们必须决定下一步采取什么动作。

一个常见的启发式方法是 UCB (upper confidence bound, 上置信界) :

选择使下面这个值最大的动作:

$$k^* = \arg \max_{k=1}^K \mu_k + \lambda \sigma_k \quad (5.120)$$

其中:

- $\mu_k = E[r_k|D]$  = 预期回报的均值
- $\sigma_k^2 = \text{var}[r_k|D]$  = 预期回报的不确定性 (方差)
- $\lambda$  = 调节探索和利用的参数

直观理解:

- 选择我们**相信会好的**动作 ( $\mu_k$  大)
  - 以及/或者我们**还不确定的**动作 ( $\sigma_k$  大)
- 

一个更简单的方法是 Thompson sampling:

每一步, 我们以一个概率选择动作  $k$ , 这个概率等于它是最优动作的概率:

$$p_k = \int_a I(E[r|a, x, \theta] = \max_a E[r|a, x, \theta]) p(\theta|D) d\theta \quad (5.121)$$

我们可以近似计算:

- 从后验  $p(\theta|D)$  中采样一次  $\theta^t$
- 然后选择:

$$k^* = \arg \max_k E[r|x, k, \theta^t]$$

尽管这个方法很简单, 但已经被证明效果很好 (Chapelle and Li 2011)。

---

## 基础且详细解释

### 1. 多臂老虎机问题

想象你进了一个赌场，有 10 台老虎机 ( $K=10$ )，但你不知道哪台更容易赢钱。

- **探索 (exploration)** = 去试试不同的机器，收集数据
- **利用 (exploitation)** = 选已经觉得最好的机器来多玩几次

问题：要如何平衡探索和利用，才能赢得最多？

---

## 2. 上下文老虎机 (Contextual bandit)

在普通多臂老虎机中，每个动作的胜率是固定的。

在上下文老虎机中，胜率会**依赖上下文特征** (context features)。

例如：

- 用户是男性 25 岁，广告 A 胜率高；用户是女性 40 岁，广告 B 胜率高
  - 这时模型就要学会：根据上下文  $x$  (用户信息+广告信息) 预测回报  $r$
- 

## 3. UCB 方法 (公式 5.120)

- $\mu_k$ : 对第  $k$  个动作的平均预期回报
  - $\sigma_k$ : 对第  $k$  个动作的不确定性
  - $\lambda$ : 决定探索程度
    - $\lambda$  大  $\rightarrow$  多去探索不确定的选项
    - $\lambda$  小  $\rightarrow$  偏向已知高收益的选项
  - 选最大  $\mu_k + \lambda\sigma_k$   $\rightarrow$  既重视高平均值，也重视高不确定性
- 

## 4. Thompson Sampling (公式 5.121)

思路：

1. 假设每个动作的回报分布都不确定
2. 每一步，假装我们知道真实的参数（从后验里随机抽一次）
3. 根据这次抽到的参数，选择最好的动作
4. 下一次再随机抽一次，可能换一个动作

优点：

- 自然平衡探索和利用（因为有时会抽到参数让你去试不常选的动作）
  - 实现简单，效果好
- 

如果你需要的话，我可以帮你画一个UCB vs Thompson Sampling 的对比图，让你看到两种方法是怎么选择动作的。这样更容易直观理解探索-利用平衡。

你要我画吗？

	Surgery	No surgery
No cancer	20	0
Lung cancer	10	50
Breast cancer	10	60

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。

### 5.7.3.2 Utility theory

Suppose we are a doctor trying to decide whether to operate on a patient or not. We imagine there are 3 states of nature: the patient has no cancer, the patient has lung cancer, or the patient has breast cancer. Since the action and state space is discrete, we can represent the loss function  $L(\theta, a)$  as a loss matrix, such as the following: These numbers reflects the fact that not performing surgery when the patient has cancer is very bad (loss of 50 or 60, depending on the type of cancer), since the patient might die; not performing surgery when the patient does not have cancer incurs no loss (0); performing surgery when the patient does not have cancer is wasteful (loss of 20); and performing surgery when the patient does have cancer is painful but necessary (10). It is natural to ask where these numbers come from. Ultimately they represent the personal preferences or values of a fictitious doctor, and are somewhat arbitrary: just as some people prefer chocolate ice cream and others prefer vanilla, there is no such thing as the "right" loss/ utility function.

However, it can be shown (see e.g., (DeGroot 1970)) that any set of consistent preferences can be converted to a scalar loss/ utility function. Note that utility can be measured on an arbitrary scale, such as dollars, since it is only relative values that matter.

### 5.7.3.3 Sequential decision theory

So far, we have concentrated on one-shot decision problems, where we only have to make one decision and then the game ends. In Setion 10.6, we will generalize this to multi-stage or sequential decision problems. Such problems frequently arise in many business and engineering settings. This is closely related to the problem of reinforcement learning. However, further discussion of this point is beyond the scope of this book.

好，我会先逐句翻译原文，然后再用通俗且详细的方式解释这张表、数字的含义以及“效用理论”和“序贯决策理论”的核心概念。

---

## 逐句翻译

### 5.7.3.2 效用理论 (Utility theory)

假设我们是一名医生，需要决定是否对病人进行手术。

我们假设有 3 种自然状态：

1. 病人没有癌症
2. 病人有肺癌
3. 病人有乳腺癌

由于行动空间和状态空间都是离散的，我们可以把损失函数  $L(\theta, a)$  表示成一个损失矩阵（如下表）：

	手术 (Surgery)	不手术 (No surgery)
无癌症	20	0
肺癌	10	50
乳腺癌	10	60

这些数字反映了以下事实：

- 当病人有癌症但不手术时，损失很大（肺癌损失 50，乳腺癌损失 60），因为病人可能会死亡。
- 当病人没有癌症且不手术时，没有损失 (0)。
- 当病人没有癌症却手术了，是一种浪费（损失 20）。
- 当病人有癌症且手术了，虽然手术痛苦，但这是必要的（损失 10）。

一个很自然的问题是：这些数字是从哪里来的？

最终它们代表的是一个虚构医生的**个人偏好或价值观**，带有一定的任意性：

就像有些人喜欢巧克力冰淇淋，有些人喜欢香草口味，没有所谓“正确”的损失/效用函数。

不过，可以证明（见 DeGroot 1970），任何一组一致的偏好都可以转换成一个标量的损失/效用函数。

需要注意的是，效用可以用任意量纲来衡量（比如用美元），因为只有相对大小才重要。

---

### 5.7.3.3 序贯决策理论 (Sequential decision theory)

到目前为止，我们只集中在“一次性”的决策问题，也就是我们只需要做一次决策，然后游戏就结束了。

在第 10.6 节，我们会把这个推广到**多阶段或序贯决策问题**。

这种问题在很多商业和工程场景中经常出现。

它与强化学习问题密切相关。

然而，更深入的讨论超出了本书的范围。

---

## 详细解释

### 1. 状态空间和动作空间

- **状态空间 (state space)** : 可能的真实情况, 比如:
    - 病人没癌症
    - 病人有肺癌
    - 病人有乳腺癌
  - **动作空间 (action space)** : 你可以采取的决策, 比如:
    - 做手术
    - 不做手术
- 

### 2. 损失矩阵的含义

损失矩阵的每一格表示：在某个真实情况（行）下，采取某个动作（列）会带来多少损失（数字越大越糟）。

#### 表格解读

状态 →\动作 ↓	手术	不手术
无癌症	20 (不必要风险)	0 (最好)
肺癌	10 (痛苦但救命)	50 (高死亡风险)
乳腺癌	10 (痛苦但救命)	60 (高死亡风险)

- **无癌症 → 手术 (20)** : 病人其实没病, 却做了手术, 浪费医疗资源且有手术风险
  - **无癌症 → 不手术 (0)** : 没病不治, 完美
  - **有癌症 → 手术 (10)** : 虽然手术有创伤, 但可以延命, 是相对较小的损失
  - **有癌症 → 不手术 (50 或 60)** : 非常糟糕, 病人可能会死
- 

### 3. 为什么数字是这样的?

这些数字不是绝对的科学结论, 而是基于医生的主观判断:

- 手术痛苦的“代价”=10
- 不必要手术的“代价”=20
- 肺癌不手术的损失=50
- 乳腺癌不手术的损失=60 (可能认为更难治疗)

效用理论允许你用**任意单位**来表达，只要相对关系正确即可（比如可以换成美元、风险指数、痛苦分数等）。

---

#### 4. 序贯决策和一次性决策的区别

- **一次性决策**: 比如医生只做一次判断（手术 or 不手术），然后事情结束
- **序贯决策**: 需要多次决策，每次决策会影响后续信息和选择
  - 例子：先做一个检测（花钱/有风险），然后根据检测结果决定是否手术
  - 这类问题是**强化学习**里的核心内容