第 2 章 机器学习的概率论基础

2.1 频率学派和贝叶斯学派

统计学家对于概率的认识也存在分歧，分为**频率学派**（Frequentism）和**贝叶斯学派**（Bayesian）两种主要流派。在频率学派的解释中，首先要定义随机实验，随机实验的所有可能结果的集合被称为该实验的样本空间。事件被定义为样本空间的一个子集。对于任何给定的事件，一次实验中只有发生或不发生两种可能。通过多次实验，我们可以观察并计算事件发生的频率。随着实验次数的增加，频率趋于稳定，这个稳定值就被视为事件的概率。因此，概率被定义为实验次数趋向无穷时频率的极限值。而按贝叶斯学派的观点，概率被理解为对事件发生的信念程度（The Degree of Belief）。它用来量化我们对事件发生的不确定性，这种不确定性的衡量基于我们所掌握的相关信息，而并不一定要求存在可重复的随机实验。这些信息不仅来源于我们对实际结果的观察，还包括了我们对事件发生可能性的先验知识。

机器学习中，频率学派和贝叶斯学派的差别主要表现在参数估计的方法及其结果上。我们以经典的投掷硬币为例来简单介绍他们的不同之处。假设有一枚没有经过特殊处理的正常硬币，每次投掷可能出现正面朝上或反面朝上两种结果之一。这个例子中，我们只需要估计一个参数，即正面朝上的概率（记为 μ），那么反面朝上的概率则为 $1-\mu$。假设投掷了三次硬币，其中两次正面朝上，一次反面朝上。采用最大似然估计，μ 的估计值为 $2/3$，因为在三次投掷中，有两次正面朝上。而贝叶斯学派的方法则不同，会先为参数 μ 设定一个先验分布，表示我们没有进行实验（即实际投掷硬币）之前对于硬币正面朝上发生可能的信念。对于一个正常的硬币，经验告诉我们正面向上的概率约为 $1/2$。贝叶斯方法会将观察到的实验结果用于更新先验分布，从而得到参数 μ 的后验分布（具体过程见2.2节）。如果采用贝塔先验分布（详见下文），μ 的后验点估计值会在 $1/2$ 和 $2/3$ 之间，即介于先验估计和仅考虑实验结果的频率估计这两种极端之间。

理论上，当硬币投掷次数足够多时，两种方法对参数 μ 的估计会收敛到相同的值。频率会趋于稳定，先验分布的影响则随着实验次数的增加不断减弱。但在实践中，基于成本和时间等因素，不可能进行过多的实验。因此，相较于频率学派，贝叶斯方法不仅能够结合先验信息和样本信息（实验结果），并且在先验分布选择合理的情况下，所得到的参数估计更为保守和稳健。贝叶斯方法的另一项特点是能够较为自然地描述特定单次事件的不确定性。例如，我们希望估计 2050 年北极冰盖完全消融的概率。类似事件或许在地球漫长的历史中出现过，但我们不可能在短时间内让同一个未来事件反复发生，因而无法通过反复观察该事件本身，从而获得其发生频率。然而，我们可以让研究北极冰盖消融的专家给出冰盖消融速度的估计（先验信息），然后结合近几年或未来几年实际消融速度的测量值（样本信息），利用贝叶斯公式得到后验估计。这个过程还可以不断迭代，将之前的后验估计作为下一阶段的先验估计，采集更多样本信息，不断修正先验以获得新的后验分布。这种方法让我们能够动态地调整对北极冰盖消融速度的估计，从而及时有效地采取必要的措施来预防和应对全球性环境变化可能带来的潜在危害。

统计学家伦纳德 · 吉米 · 萨维奇（Leonard Jimmie Savage）通过以下三个例子说明先验信息的重要性 [66]：

- 一位爱喝奶茶的女士声称她能够辨别出任何一杯奶茶在制作过程中是先倒入茶还是先倒入牛奶。在十次测试中，她每次都正确判断出奶和茶的添加顺序。
- 一位音乐家声称他仅凭一页乐谱就能够区分出是海顿还是莫扎特的作品。在十次测试中，他每次都做出了正确的判断。
- 一位喝醉的朋友声称他能预测一枚正常硬币被投掷后哪一面朝上。同样地，在十次测试中，他每次都给出了正确的预测。

采用经典统计学的假设检验方法来判断上述三位是否具备其所声称的能力时，我们一般先做出某个假设，然后通过观察实验结果来推断这个假设是否合理，即判断实验结果与假设是否有显著的差异。如果没有显著的差异，则我们没有足够的证据拒绝原假设；反之，则可以否定该假设。对于以上三个例子，我们一般将他们不具备所声称的能力（即他们每次都是猜的）作为原假设或零假设（Null Hypothesis），然后根据测试结果进行推断。如果这个假设是合理的，那么测试结果应服从十次、且每次猜中概率为 50% 的二项分布（详见2.2节），这十次测试中猜中的次数与对应的概率如图2.1所示。从图中可以看出，如果假设他们每次测试时都是猜的，十次都猜中的概率仅为 0.098% (0.5^{10})。英国统计学家罗纳德 · 艾尔默 · 费希尔（Ronald Aylmer Fisher）建议将假设检验中显著性水平定为 5%（当然也可以根据具体情况确定显著性水平）。这意味着，如果在原假设成立的条件下，观察到当前检验结果的概率小于所设定的显著性水平，我们应拒绝原假设。在上述三个例子中，如果仅靠猜测，出现相应测试结果的概率都非常小，并且均低于所设定的显著性水平。因此，我们有理由拒绝他们仅靠猜测获得这些结果的原假设，并认为这些测试结果为他们具备所声称的能力提供了统计学证据。

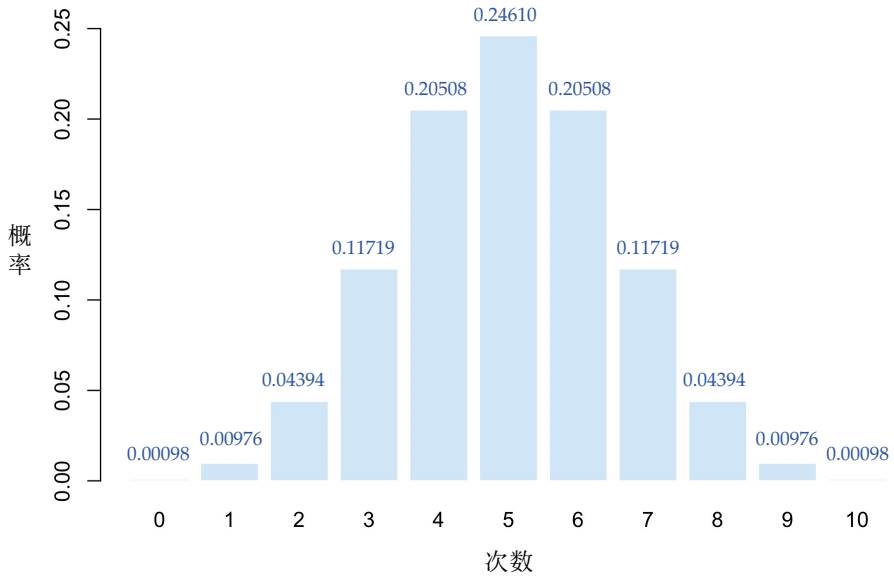


图 2.1: 十次测试中猜中次数与发生概率的对应关系。在十次测试中，每次测试的结果只有正确或错误两种可能，并且正确和错误发生的概率相同（均为 50%）。

我们使用经典的假设检验探讨了这三个例子，虽然它们在统计意义上无差别，但人们对这三个例子的解读却大相径庭。对于第二个音乐家的例子，我们没有理由怀疑该音乐家具备这样的能力。对于第三个例子，即一个醉酒的朋友能够准确预知硬币投掷的结果，大多数人会认为这是不可能的，仅仅是极度运气使然。这种强烈的先验知识让我们忽略测试的结果。而在第一个爱喝奶茶的女士的例子中，不同人有不同的看法，取决于他们的生活经验所带来的先验知识。在这三个统计意义上相同的情形中，先验信息显然不能被忽视。贝叶斯分析就是一种系统地引入先验信息并加以利用的统计方法。本书在基于统计的模型参数估计方面，将主要介绍贝叶斯学派的方法，并讨论其在机器学习中的应用。

2.2 伯努利分布与二项分布

我们先从最简单的只能取两种值的离散随机变量开始。仍然以投掷硬币为例，但这次所使用的硬币可能由于制造缺陷或人为干预等原因存在质量分布不均匀的情况，导致其两面的出现概率不一定相同。我们希望通过多次观察硬币的投掷结果，来估计硬币正面和反面朝上的概率。用随机变量 $x \in \{0, 1\}$ 表示硬币投掷的结果。当 $x = 1$ 时，表示正面朝上；而 $x = 0$ 时，则表示反面向上。设 μ 为正面朝上的概率，该随机变量 x 的概率函数（Probability Function）为：

$$\text{Bern}(x|\mu) = \mu^x(1-\mu)^{1-x} \tag{2.1}$$

这就是**伯努利分布**（Bernoulli Distribution）。显然，当 $x = 1$ 时（正面朝上），其概率为 μ，而当 $x = 0$ 时（反面朝上），其概率为 $1 - \mu$。伯努利分布的均值与方差分别为：

$$\mathbb{E}[x] = \mu \tag{2.2}$$

$$\text{Var}[x] = \mu(1-\mu) \tag{2.3}$$

伯努利分布的均值比较容易推导，可以直接代入离散随机变量期望的计算公式：

$$\begin{aligned}\mathbb{E}[x] &= 1 \times p(x=1|\mu) + 0 \times p(x=0|\mu) \\ &= 1 \times \mu + 0 \times (1-\mu) \\ &= \mu\end{aligned}$$

离散随机变量方差的定义为：

$$\text{Var}[x] = \mathbb{E}[(x - \mathbb{E}[x])^2] \tag{2.4}$$

根据期望的性质，以及 $\mathbb{E}[x]$ 是一个常数，可以得到计算方差的另一种一般相对简便的方法：

$$\begin{aligned}\text{Var}[x] &= \mathbb{E}[(x - \mathbb{E}[x])^2] \\ &= \mathbb{E}[x^2 - 2x\mathbb{E}[x] + \mathbb{E}[x]^2] \\ &= \mathbb{E}[x^2] - 2\mathbb{E}[x]\mathbb{E}[x] + \mathbb{E}[x]^2 \\ &= \mathbb{E}[x^2] - \mathbb{E}[x]^2\end{aligned}$$

对于伯努利分布，$\mathbb{E}[x^2]$ 计算如下：

$$\begin{aligned}\mathbb{E}[x^2] &= 1^2 \times p(x=1|\mu) + 0^2 \times p(x=0|\mu) \\ &= 1 \times \mu + 0 \times (1-\mu) \\ &= \mu\end{aligned}$$

使用计算方差的简便方法，伯努利分布的方差计算公式2.3推导如下：

$$\begin{aligned}\text{Var}[x] &= \mathbb{E}[x^2] - \mathbb{E}[x]^2 \\ &= \mu - \mu^2 \\ &= \mu(1-\mu)\end{aligned}$$

如果我们要计算在 N 次硬币投掷中，正面朝上的结果为 k（$k \leq N$）次的概率，就要使用**二项分布** (Binomial Distribution)。其概率函数为：

$$\text{Bin}(k|\mu, N) = \begin{pmatrix} N \\ k \end{pmatrix} \mu^k (1-\mu)^{N-k} \tag{2.5}$$

其中被称为**二项式系数**（Binomial Coefficient）的

$$\begin{pmatrix} N \\ k \end{pmatrix} = \frac{N \times (N-1) \times \cdots \times (N-k+1)}{k \times (k-1) \times \cdots \times 1} = \frac{N!}{k!(N-k)!} \tag{2.6}$$

表示从 N 个物品中随机选取 k 个，共有多少种取法。公式2.6中间那项的分子表示在考虑物品取出顺序的情况下，取出 k 个物品所有可能的取法。除以分母的 $k!$ 用于消除物品选择顺序所带来的重复计数（每一组 k 个物品的组合，存在 $k!$ 种不同的排列方式）。服从二项分布的随机变量 k 不考虑 N 次测试中结果出现的顺序，并且要求不同测试是相互独立的，即每次测试结果不受其他测试结果的影响。投掷硬币的例子显然满足独立性要求，因为前一次投掷的结果（无论是正面还是反面）不会影响到后一次投掷的结果。伯努利分布是二项分布的一种特例，即只进行一次测试（$N=1$）。

二项分布的均值和方差分别为：

$$\mathbb{E}[k] \equiv \sum_{k=0}^{N} k\,\text{Bin}(k|\mu, N) = N\mu \tag{2.7}$$

$$\text{Var}[k] \equiv \sum_{k=0}^{N} (k - \mathbb{E}[k])^2 \text{Bin}(k|\mu, N) = N\mu(1-\mu) \tag{2.8}$$

除了直接使用离散随机变量的均值和方差定义进行计算外，我们还可以利用以下均值和方差的性质方便地获得相同的结果。假设两个随机变量 x 和 z 是相互独立的，那么它们相加后的均值和方差分别为：

$$\mathbb{E}[x+z] = \mathbb{E}[x] + \mathbb{E}[z] \tag{2.9}$$

$$\text{Var}[x+z] = \text{Var}[x] + \text{Var}[z] \tag{2.10}$$

考虑到每次测试都是相互独立的，并且已知伯努利分布的均值方差分别为 μ 和 $\mu(1-\mu)$，可以

应用公式2.9和2.10推导出二项分布的均值和方差分别为 $N\mu$ 和 $N\mu(1-\mu)$。对于公式2.9，可以从均值的定义直接推导出来，并且无论随机变量 x 和 z 是否独立都成立。这一性质被称为期望的线性性质（Linearity of Expectation），即对于任意多个随机变量，它们的和的期望等于各自期望的和。至于公式2.10，从方差的定义出发，我们可以得到以下的推导过程：

$$
\begin{aligned}
\mathrm{Var}[x+z] &= \mathbb{E}[((x+z)-\mathbb{E}[x+z])^2] \\
&= \mathbb{E}[((x+z)-\mathbb{E}[x]-\mathbb{E}[z])^2] && \text{（期望的线性性质）} \\
&= \mathbb{E}[(x-\mathbb{E}[x]+z-\mathbb{E}[z])^2] \\
&= \mathbb{E}[(x-\mathbb{E}[x])^2+2(x-\mathbb{E}[x])(z-\mathbb{E}[z])+(z-\mathbb{E}[z])^2] && \text{（平方项展开）} \\
&= \mathbb{E}[(x-\mathbb{E}[x])^2]+2\mathbb{E}[(x-\mathbb{E}[x])(z-\mathbb{E}[z])]+\mathbb{E}[(z-\mathbb{E}[z])^2] \\
&= \mathrm{Var}[x]+2\mathrm{Cov}[x,z]+\mathrm{Var}[z] \\
&= \mathrm{Var}[x]+\mathrm{Var}[z] && \text{（设 } \mathrm{Cov}[x,z]=0\text{）}
\end{aligned}
$$

其中 $\mathrm{Cov}[x,z]$ 表示随机变量 x 和 z 之间的**协方差**（Covariance），用于衡量它们之间的相关程度。稍后我们将证明，当两个随机变量相互独立时，它们的协方差为零，则公式2.10得证。从协方差的定义 $\mathrm{Cov}[x,z]=\mathbb{E}[(x-\mathbb{E}[x])(z-\mathbb{E}[z])]$ 可以看出，它量化了两个随机变量偏离均值的程度及方向上的同步性。具体而言，如果两个随机变量的协方差为正值，那么当一个变量的值增加时，另一个变量的值通常也会按照一定的比例增加；如果协方差为负值，则一个变量的增加往往伴随着另一个变量按比例减少。这种按比例的增减变化体现了变量之间的线性关系。此外，对协方差进行标准化可以得到相关系数，用于衡量两个随机变量之间线性相关性的方向和强弱。相关系数的绝对值越大，表明线性相关性越强；绝对值越小，则表明线性相关性越弱。当相关系数接近于零时，说明两个随机变量之间的线性相关性较弱，但不意味着它们之间不存在其他形式的相关关系。如果两个随机变量相互独立，那么它们的协方差为零。其证明过程如下：

$$
\begin{aligned}
\mathrm{Cov}[x,z] &= \mathbb{E}[(x-\mathbb{E}[x])(z-\mathbb{E}[z])] && \text{（协方差定义）} \\
&= \mathbb{E}[xz-z\mathbb{E}[x]-x\mathbb{E}[z]+\mathbb{E}[x]\mathbb{E}[z]] \\
&= \mathbb{E}[xz]-\mathbb{E}[z]\mathbb{E}[x]-\mathbb{E}[x]\mathbb{E}[z]+\mathbb{E}[x]\mathbb{E}[z] && \text{（期望的线性性质）} \\
&= \mathbb{E}[x]\mathbb{E}[z]-\mathbb{E}[x]\mathbb{E}[z] && \text{（}x\text{ 和 }z\text{ 相互独立）} \\
&= 0
\end{aligned}
$$

伯努利分布和二项分布均依赖于参数 μ，现在我们开始讨论如何估计这一参数的值。假设对硬币进行了 N 次投掷，其结果记为 $\mathcal{D}=\{x_1,\ldots,x_N\}$，其中 $x_i\in\{0,1\}, i\in\{1,\ldots,N\}$。考虑到每次投掷的结果是相互独立的，我们可以认为每个 x_i 都是独立地从参数 μ 确定的分布 $p(x|\mu)$ 中抽样出来的，那么数据集 $\mathcal{D}$ 的似然函数可以表达成以下形式：

$$
p(\mathcal{D}|\mu)=\prod_{i=1}^{N}p(x_i|\mu)=\prod_{i=1}^{N}\mu^{x_i}(1-\mu)^{1-x_i} \tag{2.11}
$$

显然，每一次硬币投掷可以视为一次伯努利实验。因此，对于以上似然函数中的 $p(x_i|\mu)$，我们

可以用伯努利分布的概率函数（公式2.1）代入。按照频率学派的方法，我们只要找到使得上述似然函数最大化的参数 μ 的值。为了简化计算过程，通常对似然函数取对数。这一步骤将比较难处理的连乘运算转换为加和运算（因为对数函数是单调递增的，因此它不会改变参数最大化的极值点）。对数形式的似然函数为：

$$\ln p(\mathcal{D}|\mu) = \sum_{i=1}^{N} \ln p(x_i|\mu) = \sum_{i=1}^{N} \{x_i \ln \mu + (1 - x_i) \ln(1 - \mu)\} \tag{2.12}$$

上式对 μ 求导数，并让导数等于零来求其极值：

$$\begin{aligned}
\sum_{i=1}^{N} \left(\frac{x_i}{\mu} - \frac{1 - x_i}{1 - \mu} \right) = 0 \quad &\Rightarrow \\
\sum_{i=1}^{N} \left(\frac{x_i - \mu}{\mu(1 - \mu)} \right) = 0 \quad &\Rightarrow \\
\sum_{i=1}^{N} x_i - N\mu = 0 \quad &\Rightarrow \\
\mu = \frac{1}{N} \sum_{i=1}^{N} x_i &
\end{aligned}$$

以上对参数 μ 的**最大似然估计**（Maximum Likelihood Estimation）的结果就是在多次测试中硬币正面朝上的次数占总投掷次数 N 的比例。假设我们进行了 3 次硬币投掷测试，并且每次结果都是正面朝上。现在让我们根据目前的观测结果来预测之后的硬币投掷结果。根据最大似然估计，我们得到 $\mu = 1$，即认为之后所有投掷结果都是正面朝上。这显然与我们的常识相悖。从机器学习的角度来看，最大似然估计在这种极端情况下容易导致过拟合现象。

采用贝叶斯方法，先要为参数 μ 选择一个先验分布，然后结合似然函数计算 μ 的后验分布。在选择先验分布时，使用与后验分布相同形式的先验分布可以简化计算过程。对于给定的似然函数，如果后验分布与先验分布的形式相同，则先验分布与后验分布被称为**共轭分布**（Conjugate Distribution），而先验分布被称为似然函数的**共轭先验**（Conjugate Prior）。值得注意的是，共轭关系是指在给定的似然函数条件下，判断后验分布与先验分布是否具有相同的函数形式。对于多次投掷硬币的例子，其似然函数就是二项分布的概率函数（公式2.5），该概率函数其实就等于之前最大似然估计所使用的似然函数（公式2.11）乘以一个二项式系数。因为这个二项式系数是与 μ 无关的常数，在进行最大似然估计时忽略它并不影响参数 μ 的计算结果。在二项分布中，引入二项式系数的目的是为了实现概率的归一化，以确保所有可能结果的概率之和等于 1。

根据贝叶斯公式，我们可以推导出后验分布正比于先验分布和似然函数的乘积：

$$\underbrace{p(\mu|\mathcal{D})}_{\text{后验分布}} \propto \underbrace{p(\mathcal{D}|\mu)}_{\text{似然函数}} \underbrace{p(\mu)}_{\text{先验分布}} \tag{2.13}$$

上式右侧是随机变量 μ 和数据集 $\mathcal{D}$ 的联合概率分布 $p(\mathcal{D}, \mu)$ 的一种分解方式，而左侧可以看成该联合分布另一种分解方式 $p(\mu|\mathcal{D})p(\mathcal{D})$ 忽略 $p(\mathcal{D})$ 项后的结果。能够忽略这项的原因是，对于任何参数估计方法，数据集 $\mathcal{D}$ 是给定的，其出现概率 $p(\mathcal{D})$ 并不会影响参数估计的值。需要注

意的是，似然函数并不是概率分布（对参数 μ 求积分并不一定等于 1）。它表示的是在给定参数 μ 的情况下，数据集 $\mathcal{D}$ 的生成概率。

二项分布的概率函数（这里作为似然函数使用）的主要部分是 μ 和 $1-\mu$ 的指数形式。如果我们选择的先验分布也是呈 μ 和 $1-\mu$ 的指数形式，那么其后验分布（即先验分布与似然函数的乘积）也将是 μ 和 $1-\mu$ 的指数形式。这样的选择使得先验分布与后验分布的形式相同，即为共轭分布。为此，我们选择**贝塔分布**（Beta Distribution）作为参数 μ 的先验分布，其概率密度函数（Probability Density Function）的形式如下：

$$\text{Beta}(\mu|\alpha,\beta)=\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\mu^{\alpha-1}(1-\mu)^{\beta-1} \tag{2.14}$$

其中 $\Gamma(z)$ 表示伽玛函数（Gamma Function），而 α 和 β 是贝塔分布的两个超参数，它们的取值共同决定了贝塔分布的形态（见图2.3）。不同于伯努利分布和二项分布等离散随机分布，贝塔分布是一种连续型概率分布，其变量 μ 可以取无限多个不同的值。连续型概率分布可以理解为：连续随机变量 x 落入区间 $(x,x+\delta x)$ 的概率等于 $f(x)\delta x$（当 $\delta x\to 0$ 时），其中 $f(x)$ 是连续随机变量 x 的概率密度函数。因此，一般需要通过积分来计算连续随机变量落入某一区间的概率。贝塔分布的主要部分也是 μ 和 $1-\mu$ 的指数形式，而它的与 μ 无关的系数项作用与二项分布中的二项式系数类似，目的是为了概率的归一化，即使得（μ 的取值范围是 $[0,1]$）：

$$\int_0^1 \text{Beta}(\mu|\alpha,\beta)\,d\mu=1 \tag{2.15}$$

伽玛函数 $\Gamma(z)$ 是阶乘函数在实数和复数域上的扩展，如果 z 为正整数，则：

$$\Gamma(z)=(z-1)! \tag{2.16}$$

比较贝塔分布的系数项（公式2.14）和二项分布的二项式系数（公式2.6），我们可以看出这两者在形式上的相似性。具体而言，$\Gamma(\alpha+\beta)$ 对应于 $N!$，$\Gamma(\alpha)$ 对应于 $k!$，而 $\Gamma(\beta)$ 对应于 $(N-k)!$。这两种分布在整体形式上（包括系数和指数部分）的相似性表明，伽玛函数将阶乘推广到正实数及复数区域（但不包括非正整数极点），而贝塔分布则是定义在概率参数（$0\le\mu\le 1$）上的连续分布，经常作为二项分布参数的先验（贝塔分布与二项似然具有相容的函数形式）。

伽玛函数在概率论和组合数学中被广泛应用，对于实数 $z>0$，其定义为以下积分形式：

$$\Gamma(z)\equiv\int_0^\infty t^{z-1}e^{-t}\,dt \tag{2.17}$$

虽然这个积分形式看起来比较复杂，但通过观察伽玛函数在实数区间 $[0,6]$ 的函数图形（如图2.2）可以看出，它可以简单理解为对阶乘序列点的平滑曲线拟合，即一种对阶乘序列点（在大于零的实数域内）进行插值的解法。从阶乘的定义，可以推导出伽玛函数的一个重要递推公式：

$$\Gamma(z+1)=z\Gamma(z) \tag{2.18}$$

我们进一步讨论贝塔分布，图2.3显示了超参数 α 和 β 取四组不同值时的贝塔分布。从这些图形中可以看出，α 和 β 的值较大的贝塔分布比较小时的分布更窄且更尖锐。贝塔分布通常被作为先验分布来使用，它的两个超参数 α 和 β 可以理解为在进行后续实际测试之前，已观测到 $x=1$（硬币正面朝上）和 $x=0$（硬币反面朝上）的有效次数（Effective Number of Observations）。

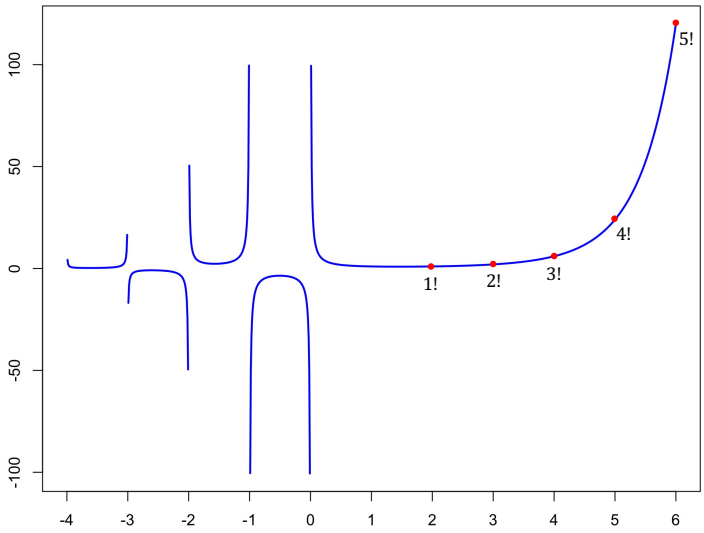


图 2.2: 伽玛函数在实数区间 $[-4, 6]$ 上的图形。

这种有效观测次数也可以基于主观先验信念来设定。随着有效观测次数的增加，我们对参数 μ 的不确定性逐步减少，从而使得其分布更加狭窄和尖锐。当 α 和 β 都为 1 时，贝塔分布退化为均匀分布（如图2.3右上角的子图）。值得注意的是，超参数 α 和 β 不仅可以取正整数，也可以取任意正实数。

贝塔分布的均值与方差分别为：

$$\mathbb{E}[\mu] = \frac{\alpha}{\alpha + \beta} \tag{2.19}$$

$$\text{Var}[\mu] = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)} \tag{2.20}$$

结合公式2.14和2.15（系数是归一项），我们可以得到：

$$\int_0^1 \mu^{\alpha-1}(1-\mu)^{\beta-1} d\mu = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)} \tag{2.21}$$

上式的积分结果为贝塔分布系数项的倒数。从连续随机变量均值的定义出发，我们可以通过以下过程推导出贝塔分布的均值（公式2.19）：

$$\begin{aligned}
\mathbb{E}[\mu] &= \int_0^1 \mu \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \mu^{\alpha-1}(1-\mu)^{\beta-1}\, d\mu && \text{（均值定义）} \\
&= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \int_0^1 \mu^{(\alpha+1)-1}(1-\mu)^{\beta-1}\, d\mu \\
&= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(\alpha+1)\Gamma(\beta)}{\Gamma(\alpha+\beta+1)} && \text{（利用公式2.21）} \\
&= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\alpha\Gamma(\alpha)\Gamma(\beta)}{(\alpha+\beta)\Gamma(\alpha+\beta)} && (\Gamma(z+1) = z\Gamma(z)) \\
&= \frac{\alpha}{\alpha+\beta}
\end{aligned}$$

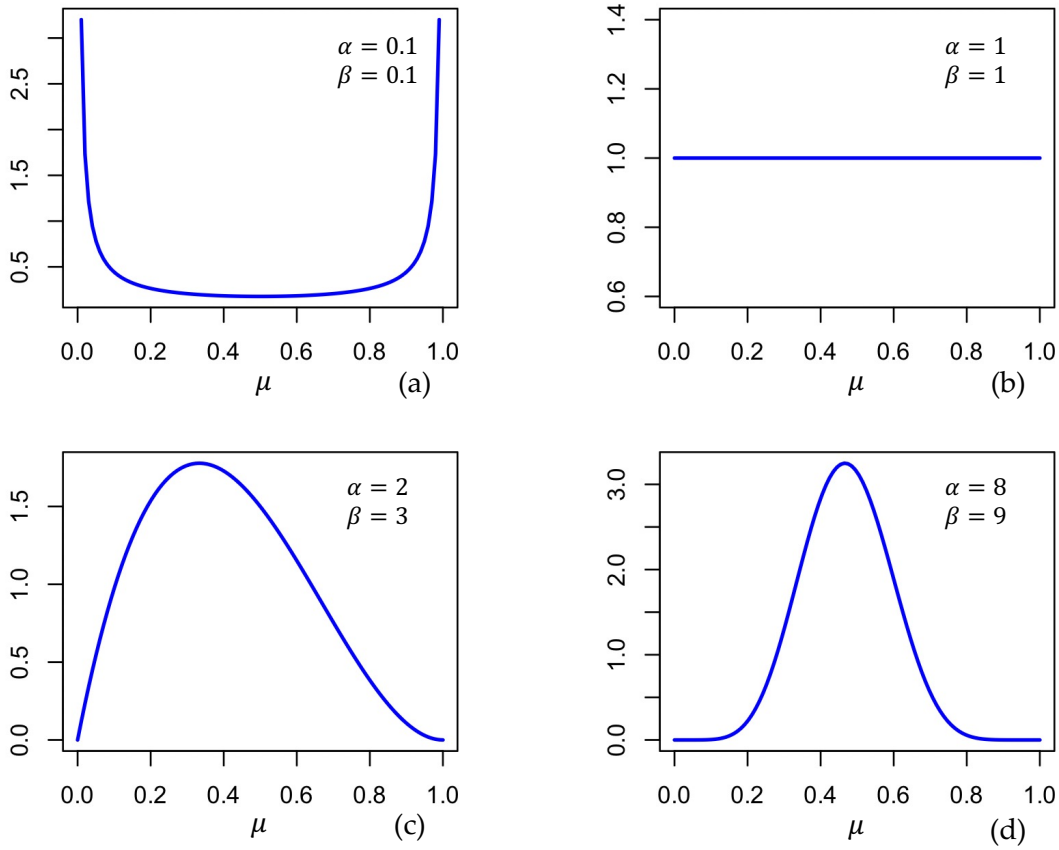


图 2.3: 超参数 α 和 β 取四种不同值时的贝塔分布概率密度函数图。

同理可得：

$$\begin{aligned}\mathbb{E}[\mu^2] &= \int_0^1 \mu^2 \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \mu^{\alpha-1}(1-\mu)^{\beta-1}\, d\mu \\ &= \frac{\alpha(\alpha+1)}{(\alpha+\beta)(\alpha+\beta+1)}\end{aligned}$$

已知 $\text{Var}[x] = \mathbb{E}[x^2] - \mathbb{E}[x]^2$，则贝塔分布的方差可推导如下：

$$\begin{aligned}\text{Var}[\mu] &= \frac{\alpha(\alpha+1)}{(\alpha+\beta)(\alpha+\beta+1)} - \left(\frac{\alpha}{\alpha+\beta}\right)^2 \\ &= \frac{\alpha(\alpha+1)(\alpha+\beta)}{(\alpha+\beta)^2(\alpha+\beta+1)} - \frac{\alpha^2(\alpha+\beta+1)}{(\alpha+\beta)^2(\alpha+\beta+1)} \\ &= \frac{(\alpha^3+\alpha^2\beta+\alpha^2+\alpha\beta) - (\alpha^3+\alpha^2\beta+\alpha^2)}{(\alpha+\beta)^2(\alpha+\beta+1)} \\ &= \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}\end{aligned}$$

将贝塔分布（公式2.14）作为二项分布中参数 μ 的先验分布，然后结合参数 μ 的似然函数（公式2.11），可以得到以下形式的后验分布：

$$p(\mu|k,l,\alpha,\beta) \propto \mu^{k+\alpha-1}(1-\mu)^{l+\beta-1} \tag{2.22}$$

其中 $k=\sum_{i=1}^{N} x_i$ 表示在 N 次硬币投掷测试中，正面朝上的次数，而 $l=N-k$ 表示反面朝上的次数。添加使概率归一化的系数项后，我们可以得到参数 μ 的后验分布：

$$p(\mu|k,l,\alpha,\beta)=\frac{\Gamma(k+\alpha+l+\beta)}{\Gamma(k+\alpha)\Gamma(l+\beta)}\mu^{k+\alpha-1}(1-\mu)^{l+\beta-1} \tag{2.23}$$

上式可以看作是由两个新的超参 $(k+\alpha)$ 和 $(l+\beta)$（对应于原先的 α 和 β）所确定的贝塔分布。因此，二项分布的参数 μ 的先验分布和后验分布都是贝塔分布，表明它们是共轭的。计算这个后验分布也非常简便，只需将新观测到的正面朝上的次数 k 加到超参数 α 上，并将新观测到的反面朝上的次数 l 加到 β 上。假设我们选择了超参数 α 和 β 均为 0.5 的贝塔分布作为先验分布，然后进行了三次硬币投掷测试，并且每次结果都是正面朝上，那么参数 μ 的后验分布为：

$$\text{Beta}(\mu|3.5,0.5)=\frac{\Gamma(4)}{\Gamma(3.5)\Gamma(0.5)}\mu^{2.5}(1-\mu)^{-0.5} \tag{2.24}$$

以上后验分布描述了参数 μ 取不同值时的概率。当我们利用它进行预测时，需要考虑 μ 所有可能的取值及其对应的概率。也就是说，需要使用 μ 的后验分布对预测对象求期望。具体而言，在给定数据集 $\mathcal{D}$（这里即为硬币投掷的观测结果）的情况下，结合先验分布和生成数据集的似然函数得到后验分布后，预测下一次投掷结果为正面朝上（$x=1$）的概率为：

$$p(x=1|\mathcal{D})=\int_0^1 \underbrace{p(x=1|\mu)}_{\text{预测对象}}\underbrace{p(\mu|\mathcal{D})}_{\text{后验分布}}\,d\mu=\int_0^1 \mu p(\mu|\mathcal{D})\,d\mu=\mathbb{E}[\mu|\mathcal{D}] \tag{2.25}$$

从上述公式可以看出，该概率实际上就是贝塔分布均值的计算公式2.19，即：

$$p(x=1|\mathcal{D})=\frac{k+\alpha}{k+\alpha+l+\beta} \tag{2.26}$$

延续之前的例子，利用贝叶斯方法预测下一次投掷结果为正面朝上的概率为：

$$p(x=1|\mathcal{D})=\frac{3+0.5}{3+0.5+0+0.5}=\frac{3.5}{4}=0.875$$

以上计算结果可以看出，对于 μ 的估计，贝叶斯方法与频率学派的最大似然估计结果有所不同。通过选择合适的先验分布，贝叶斯方法能够提供相对稳健的估计，并且不容易过拟合于数据集。

在使用贝叶斯方法时，后验分布可以作为下一阶段的先验分布。当获得更多数据后，我们可以结合新数据的似然函数得到新的后验分布。这一过程，可以不断持续迭代，被称为贝叶斯序贯分析，特别适用于在线学习场景。在线学习指的是模型以连续的方式从逐个到达的数据点中进行学习，而非从一次性提供的整个数据集中学习。这种方法特别适用于数据流不断产生且在获得全部数据之前就需要预测的情境。以公式2.24所示的贝塔分布作为下一阶段的先验分布，假设我们又进行了三次硬币投掷测试，其中两次正面朝上，一次反面朝上。根据这些新的测试结果，参数 μ 的后验分布更新为：

$$\text{Beta}(\mu|5.5,1.5)=\frac{\Gamma(7)}{\Gamma(5.5)\Gamma(1.5)}\mu^{4.5}(1-\mu)^{0.5}$$

预测下一次投掷结果为正面朝上的概率也调整为：

$$p(x=1|\mathcal{D})=\frac{2+3.5}{2+3.5+1+0.5}=\frac{5.5}{7}\approx 0.786$$

2.3 泊松分布和多项分布

泊松分布和多项分布可以看作二项分布在特定情况下的扩展。具体来说，泊松分布是当参数 μ 的值很小且观测次数 N 趋向于无穷大时，二项分布的一种极限形式。它主要用于描述在固定的时间或空间范围内独立事件发生的次数（例如单位时间内顾客到店的人数、网页被访问的次数、自然或者人为灾害的发生频率等），适用于事件发生率固定且事件之间相互独立的情况。多项分布则适用于结果可能性超过两种的情形，是二项分布在多种可能结果上的扩展。

我们首先讨论泊松分布。假设有一家古玩店，店主想准确掌握每天到店的顾客人数，因此要求员工每隔一分钟记录一下前一分钟内到店的顾客人数。由于古玩店的顾客通常较少，每分钟最多只有一位顾客到店。这家古玩店每日营业十小时，每天共计 $10 \times 60 = 600$ 次记录。假设各时间段到店人数相互独立，且一分钟内有顾客到店的概率为 0.01，则每日的到店人数 k 满足 $\text{Bin}(k|0.01, 600)$ 的二项分布，其中 $\mu = 0.01$，$N = 600$。当店主使用以上二项分布来计算每日到店人数 k 及其相应的概率时，他会发现这样的计算过程过于繁琐。具体计算公式如下：

$$\text{Bin}(k|0.01, 600) = \frac{600!}{k!(600-k)!}\mu^k(1-\mu)^{600-k}$$

这位店主希望能找到一个近似的分布来简化计算过程。根据二项分布的定义，并且设 $\lambda = N\mu$，我们可以推导出如下结果：

$$\begin{aligned}
\text{Bin}(k|\mu, N) &= \begin{pmatrix} N \\ k \end{pmatrix} \mu^k(1-\mu)^{N-k} \\
&= \frac{N \times (N-1) \times \cdots \times (N-k+1)}{k \times (k-1) \times \cdots \times 1}\mu^k(1-\mu)^{N-k} \\
&= \frac{\lambda^k}{k!} \cdot \frac{N}{N} \cdot \frac{N-1}{N} \cdots \frac{N-k+1}{N}\left(1-\frac{\lambda}{N}\right)^N\left(1-\frac{\lambda}{N}\right)^{-k} \\
&= \frac{\lambda^k}{k!} \cdot \frac{N}{N}\left(1-\frac{1}{N}\right)\cdots\left(1-\frac{k-1}{N}\right)\left(1-\frac{\lambda}{N}\right)^N\left(1-\frac{\lambda}{N}\right)^{-k}
\end{aligned}$$

当 $k \geq 0$，

$$\lim_{N\to\infty} \frac{N}{N}\left(1-\frac{1}{N}\right)\cdots\left(1-\frac{k-1}{N}\right)\left(1-\frac{\lambda}{N}\right)^{-k} = 1$$

根据指数函数的定义 $e^x = \lim\limits_{N\to\infty}\left(1+\frac{x}{N}\right)^N$，我们可以得到：

$$\lim_{N\to\infty}\left(1-\frac{\lambda}{N}\right)^N = e^{-\lambda}$$

结合以上等式，我们可得：

$$\lim_{N\to\infty}\text{Bin}(k|\mu, N) = \frac{\lambda^k e^{-\lambda}}{k!}$$

即当观测次数 N 趋向于无穷大，参数 μ 趋近于 0，且 $N\mu \to \lambda$ 时，二项分布趋于参数为 λ 的**泊松分布**（Poisson Distribution）：

$$\text{Poisson}(k|\lambda) = \frac{\lambda^k e^{-\lambda}}{k!} \tag{2.27}$$

从表2.1可以看出，对于 k 的不同取值，使用二项分布 $\text{Bin}(k|0.01, 600)$ 和泊松分布 $\text{Poisson}(k|6)$（其中 $\lambda = 600 \times 0.01 = 6$）计算得到的概率值非常相近。因此，泊松分布可以看作在 N 较大且 μ 较小时二项分布的近似分布。

表 2.1: 使用二项分布和泊松分布计算每日到店人数的概率对比

k	0	1	2	3	4	5
二项分布	0.00241	0.01458	0.04410	0.08878	0.13385	0.16116
泊松分布	0.00248	0.01487	0.04462	0.08924	0.13385	0.16062
k	6	7	8	9	10	11
二项分布	0.16143	0.13837	0.10360	0.06884	0.04109	0.02226
泊松分布	0.16062	0.13768	0.10326	0.06884	0.04130	0.02253

泊松分布的均值和方差分别为：

$$\mathbb{E}[k] \equiv \sum_{k=0}^{\infty} k\,\text{Poisson}(k|\lambda) = \lambda \tag{2.28}$$

$$\text{Var}[k] \equiv \sum_{k=0}^{\infty} (k - \mathbb{E}[k])^2 \text{Poisson}(k|\lambda) = \lambda \tag{2.29}$$

其中均值的推导如下：

$$\begin{aligned}
\mathbb{E}[k] &= \sum_{k=0}^{\infty} k\,\text{Poisson}(k|\lambda) \\
&= \sum_{k=1}^{\infty} k \frac{\lambda^k e^{-\lambda}}{k!} \\
&= \lambda e^{-\lambda} \sum_{k=1}^{\infty} \frac{\lambda^{k-1}}{(k-1)!} \\
&= \lambda e^{-\lambda} \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} \\
&= \lambda e^{-\lambda} e^{\lambda} && \left(\text{泰勒展开 } e^{\lambda} = \sum_{k=0}^{\infty} \frac{\lambda^k}{k!}\right) \\
&= \lambda
\end{aligned}$$

我们再次使用简便公式 $\text{Var}[x] = \mathbb{E}[x^2] - \mathbb{E}[x]^2$ 来计算泊松分布的方差。如果能够证明泊松分布的 $\mathbb{E}[k^2]$ 的形式为 $\lambda^2 + \lambda$，则其方差可推导如下：

$$\begin{aligned}
\text{Var}[k] &= \mathbb{E}[k^2] - \mathbb{E}[k]^2 \\
&= (\lambda^2 + \lambda) - \lambda^2 \\
&= \lambda
\end{aligned}$$

我们通过以下步骤可以推导出 $\mathbb{E}[k^2] = \lambda^2 + \lambda$：

$$
\begin{aligned}
\mathbb{E}[k^2] &= \mathbb{E}[k(k-1) + k] \\
&= \mathbb{E}[k(k-1)] + \mathbb{E}[k] && \text{（期望的线性性质）} \\
&= \sum_{k=0}^{\infty} k(k-1)\frac{\lambda^k e^{-\lambda}}{k!} + \lambda \\
&= \lambda^2 e^{-\lambda} \sum_{k=2}^{\infty} \frac{\lambda^{(k-2)}}{(k-2)!} + \lambda \\
&= \lambda^2 e^{-\lambda} \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} + \lambda \\
&= \lambda^2 e^{-\lambda} e^{\lambda} + \lambda && \text{（泰勒展开 } e^{\lambda} = \sum_{k=0}^{\infty} \frac{\lambda^k}{k!}\text{）} \\
&= \lambda^2 + \lambda
\end{aligned}
$$

从上述推导可以看出，泊松分布只有一个参数 λ，并且其均值和方差都等于 λ。参数 λ 的值可以解释为在固定时间（或空间）范围内独立事件发生的期望次数。

我们首先使用最大似然方法来估计泊松分布的参数 λ。假设我们对古玩店每日到店人数进行了 N 天的观察和统计，结果记为 $\mathcal{D} = \{k_1, \ldots, k_N\}$，其中 $k_i, i \in \{1, \ldots, N\}$ 表示第 i 天顾客的人数。对于数据集 $\mathcal{D}$，其似然函数为：

$$
p(\mathcal{D}|\lambda) = \prod_{i=1}^{N} p(k_i|\lambda) = \prod_{i=1}^{N} \left(\frac{\lambda^{k_i} e^{-\lambda}}{k_i!} \right) \tag{2.30}
$$

为了简化计算过程，我们将似然函数转换为其对数形式：

$$
\ln p(\mathcal{D}|\lambda) = \sum_{i=1}^{N} \ln \left(\frac{\lambda^{k_i} e^{-\lambda}}{k_i!} \right) = -N\lambda + \left(\sum_{i=1}^{N} k_i \right) \ln(\lambda) - \sum_{i=1}^{N} (\ln k_i!) \tag{2.31}
$$

将上式对 λ 求导数，并让导数等于零来求其极值：

$$
\begin{gathered}
-N + \left(\sum_{i=1}^{N} k_i \right) \frac{1}{\lambda} = 0 \quad \Rightarrow \\
\lambda = \frac{1}{N} \sum_{i=1}^{N} k_i
\end{gathered}
$$

仍以古玩店为例，假设进行了 10 天的观测。到店顾客人数分别为 3 人、4 人、5 人、6 人、7 人和 8 人的天数为 1 天、1 天、1 天、3 天、2 天和 2 天。使用最大似然估计，可以得到参数 λ 的估计值为：

$$
\frac{(3 \times 1) + (4 \times 1) + (5 \times 1) + (6 \times 3) + (7 \times 2) + (8 \times 2)}{10} = 6
$$

如果使用贝叶斯方法，首先需要为参数 λ 选择一个合适的先验分布。通常，我们优先考虑与似然函数共轭的先验分布。对于公式2.30所示的似然函数，其共轭先验应具有与指数项 λ^k 和

指数函数 $e^{-\lambda}$ 相乘的形式相似的结构。之前介绍过的伽玛函数（公式2.17）也具有类似的形式，通过引入适当的系数可以将其归一化为概率分布。根据伽玛函数的定义（公式2.17），我们可知：

$$\int_0^\infty t^{z-1}e^{-t}\,dt = \Gamma(z)$$

因此，归一化后的概率分布的系数项为 $1/\Gamma(z)$，它使得：

$$\int_0^\infty \frac{t^{z-1}e^{-t}}{\Gamma(z)}\,dt = 1$$

将上式中的 z 和 t 分别替换为 α 和 $\beta\lambda$，我们有：

$$\int_0^\infty \frac{(\beta\lambda)^{\alpha-1}e^{-\beta\lambda}}{\Gamma(\alpha)}\,d(\beta\lambda) = \int_0^\infty \frac{\beta^\alpha\lambda^{\alpha-1}e^{-\beta\lambda}}{\Gamma(\alpha)}\,d\lambda = 1$$

如此，我们就可以得到具有以下概率密度函数的**伽玛分布**（Gamma Distribution）：

$$\text{Gamma}(\lambda|\alpha,\beta) = \frac{\beta^\alpha\lambda^{\alpha-1}e^{-\beta\lambda}}{\Gamma(\alpha)} \tag{2.32}$$

其中超参 α 为形状参数（Shape Parameter），决定了分布曲线的形状；β 为逆尺度参数（Inverse Scale Parameter），决定了分布的尺度和宽度，即分布中值的分散或集中的程度。如图2.4所示，当 $\alpha \leq 1$ 时，伽玛分布呈长尾偏斜的形态；而当 $\alpha > 1$ 时，分布的偏斜程度随着 α 增大而减小。逆尺度参数之所以得名，是因为其值的大小与分布的宽度呈反比。超参 β 值越小，分布的标准差越大，其峰值相对较低且宽；β 值越大，则分布的标准差越小，其峰值更高且尖锐。值得注意的是，当 $\alpha = k+1$ 且 $\beta = 1$ 时，伽玛分布与泊松分布形式相同（原随机变量 λ 变成分布的超参数）。伽玛分布与泊松似然具有相容的函数形式，因此常用作泊松分布率参数的共轭先验。此外，当 $\alpha = 1$ 时，伽玛分布退化为只有一个逆尺度参数 β 的**指数分布**（Exponential Distribution）：

$$\text{Exp}(\lambda|\beta) = \beta e^{-\beta\lambda} \tag{2.33}$$

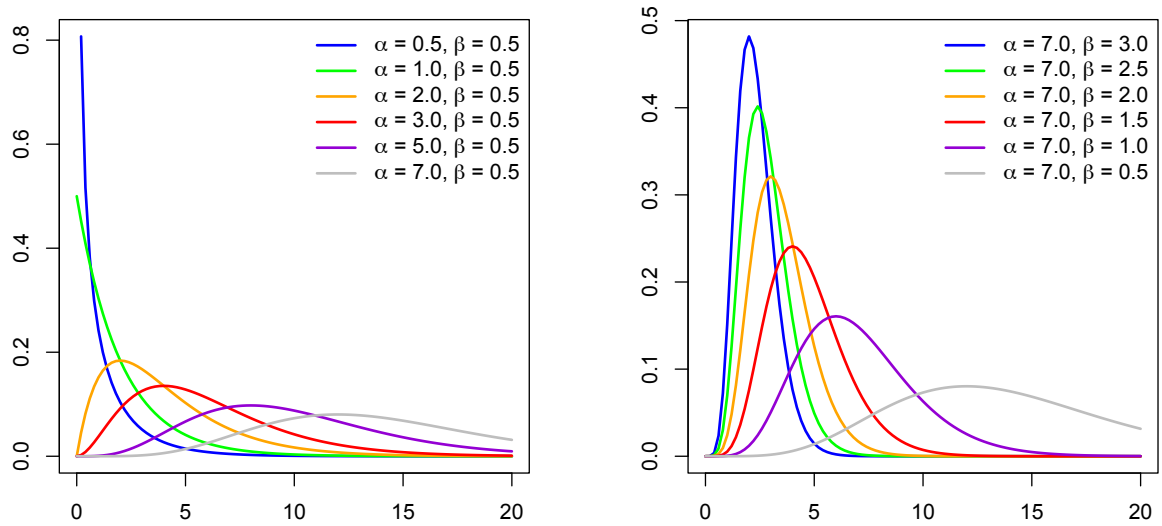


图 2.4: 伽玛分布在不同超参数设置下的概率密度函数图。左图中，参数 β 固定为 0.5，形状参数 α 决定了分布的形状。右图中，参数 α 固定为 7.0，逆尺度参数 β 决定了分布的尺度（即方差）。

伽玛分布的均值和方差分别为：

$$\mathbb{E}[\lambda] = \frac{\alpha}{\beta} \tag{2.34}$$

$$\text{Var}[\lambda] = \frac{\alpha}{\beta^2} \tag{2.35}$$

从连续随机变量均值的定义出发，我们通过以下过程可以推导出伽玛分布的均值：

$$\begin{aligned}
\mathbb{E}[\lambda] &= \int_0^\infty \lambda \frac{\beta^\alpha \lambda^{\alpha-1} e^{-\beta\lambda}}{\Gamma(\alpha)}\, d\lambda \\
&= \frac{\beta^\alpha}{\Gamma(\alpha)} \int_0^\infty \lambda^\alpha e^{-\beta\lambda}\, d\lambda \\
&= \frac{\beta^\alpha}{\Gamma(\alpha)} \int_0^\infty \left(\frac{\mu}{\beta}\right)^\alpha e^{-\mu} d\left(\frac{\mu}{\beta}\right) && \text{（积分部分进行换元 } \mu = \beta\lambda\text{）} \\
&= \frac{\beta^\alpha}{\Gamma(\alpha)} \cdot \frac{1}{\beta^{\alpha+1}} \int_0^\infty \mu^\alpha e^{-\mu}\, d\mu \\
&= \frac{1}{\beta\Gamma(\alpha)} \Gamma(\alpha+1) && \text{（积分项为 } \Gamma(\alpha+1) \text{ 的定义）} \\
&= \frac{\alpha\Gamma(\alpha)}{\beta\Gamma(\alpha)} && (\Gamma(\alpha+1) = \alpha\Gamma(\alpha)) \\
&= \frac{\alpha}{\beta}
\end{aligned}$$

为了计算方差，我们需要先求出 $\mathbb{E}[\lambda^2]$：

$$\begin{aligned}
\mathbb{E}[\lambda^2] &= \int_0^\infty \lambda^2 \frac{\beta^\alpha \lambda^{\alpha-1} e^{-\beta\lambda}}{\Gamma(\alpha)}\, d\lambda \\
&= \frac{\beta^\alpha}{\Gamma(\alpha)} \int_0^\infty \lambda^{\alpha+1} e^{-\beta\lambda}\, d\lambda \\
&= \frac{\beta^\alpha}{\Gamma(\alpha)} \int_0^\infty \left(\frac{\mu}{\beta}\right)^{\alpha+1} e^{-\mu}\, d\left(\frac{\mu}{\beta}\right) && \text{（积分部分进行换元 } \mu = \beta\lambda\text{）} \\
&= \frac{\beta^\alpha}{\Gamma(\alpha)} \cdot \frac{1}{\beta^{\alpha+2}} \int_0^\infty \mu^{(\alpha+2)-1} e^{-\mu}\, d\mu \\
&= \frac{1}{\beta^2\Gamma(\alpha)} \Gamma(\alpha+2) && \text{（积分项为 } \Gamma(\alpha+2) \text{ 的定义）} \\
&= \frac{\alpha(\alpha+1)\Gamma(\alpha)}{\beta^2\Gamma(\alpha)} && (\Gamma(\alpha+1) = \alpha\Gamma(\alpha)) \\
&= \frac{\alpha(\alpha+1)}{\beta^2}
\end{aligned}$$

然后，方差 $\text{Var}[\lambda]$ 可以计算如下：

$$\begin{aligned}
\text{Var}[\lambda] &= \mathbb{E}[\lambda^2] - \mathbb{E}[\lambda]^2 \\
&= \frac{\alpha(\alpha+1)}{\beta^2} - \left(\frac{\alpha}{\beta}\right)^2 \\
&= \frac{\alpha}{\beta^2}
\end{aligned}$$

将伽玛分布（公式2.32）作为泊松分布参数 λ 的先验分布，并通过贝叶斯方法结合参数 λ 的

似然函数（公式2.30），可以得到如下形式的后验分布：

$$p(\lambda|m,\alpha,\beta) \propto \lambda^{(m+\alpha-1)}e^{-(N+\beta)\lambda} \tag{2.36}$$

其中 $m=\sum_{i=1}^{N}k_i$ 表示在 N 天观测期间累计到店顾客的总人数。添加使概率归一化的系数项之后，可以得到如下参数 λ 的后验分布：

$$p(\lambda|m,\alpha,\beta) = \frac{(N+\beta)^{m+\alpha}\lambda^{m+\alpha-1}e^{-(N+\beta)\lambda}}{\Gamma(m+\alpha)} \tag{2.37}$$

上式可以看作是由两个新的超参 $(m+\alpha)$ 和 $(N+\beta)$（对应于原先的 α 和 β）所确定的伽玛分布。因此，伽玛分布是泊松分布唯一参数 λ 的似然函数的共轭先验。当进行预测时，需要使用 λ 的后验分布对预测对象求期望。举例来说，在给定数据集 $\mathcal{D}$（这里即为每日到店顾客人数的统计结果）的情况下，结合先验分布和生成数据的似然函数，预测明天到店顾客人数为 $(m+\alpha)/(N+\beta)$，即分布 Gamma$(\lambda|m+\alpha,N+\beta)$ 的均值。以上是贝叶斯方法的完整过程。首先结合先验分布与似然函数得到参数的后验分布；然后在进行预测时，使用该后验分布对预测对象求期望。

我们现在开始讨论多项分布。正如之前介绍的那样，多项分布是二项分布在多种可能结果上的扩展。接下来，我们以掷骰子为例来介绍多项分布。常见的骰子是一个立方体，有六个面。理想情况下，完全公正的骰子的重心应该在正中央，使得掷出每一面的概率相同。然而，在实际制作中，由于点数多的凹洞比较多，未经特殊调整的骰子可能会使点数为一的那一面比较重。我们可以通过多次观察投掷骰子的结果，来估计特定骰子每一面出现的概率。将每一面出现的概率记为 $\boldsymbol{\mu}=(\mu_1,\ldots,\mu_K)^\top$，其中 $K=6$ 且 μ_i 表示点数为 i 的那面出现的概率。注意，向量 $\boldsymbol{\mu}$ 需要满足 $\sum_{i=1}^{K}\mu_i=1$，即所有面出现概率之和应等于一（$\boldsymbol{\mu}$ 本身是骰子各面出现几率的概率分布）。假设我们进行了 N 次投掷，每一面出现次数分别为 $m_1,\ldots,m_K$，则这些出现次数服从**多项分布**（Multinomial Distribution），其概率函数为：

$$\text{Mult}(\boldsymbol{m}|\boldsymbol{\mu},N) = \begin{pmatrix} N \\ m_1\ldots m_K \end{pmatrix}\prod_{i=1}^{K}\mu_i^{m_i} \tag{2.38}$$

其中 $\boldsymbol{m}=(m_1,\ldots,m_K)^\top$，并且概率归一化的**多项式系数**

$$\begin{pmatrix} N \\ m_1\ldots m_K \end{pmatrix} = \frac{N!}{m_1!\ldots m_K!}$$

表示将 N 个物品分成 K 组，每组的数量分别为 $m_1,\ldots,m_K$ 时，共有多少种分法。

多项分布的均值、方差和协方差分别为：

$$\mathbb{E}[m_i] = N\mu_i \tag{2.39}$$

$$\text{Var}[m_i] = N\mu_i(1-\mu_i) \tag{2.40}$$

$$\text{Cov}[m_i,m_j] = -N\mu_i\mu_j \tag{2.41}$$

多项分布的均值和方差与二项分布（公式2.7和2.8）的形式相同，其原因也比较容易理解。尽管多项分布涉及多个随机变量，但当我们只考虑其中一个随机变量 m_i 时，其发生与不发生的概率仅与参数 μ_i 有关，即发生的概率为 μ_i，而不发生的概率为 $1-\mu_i$。随机变量 m_i 的边缘分布

（忽略其他变量）就是二项分布。关于两个随机变量 m_i 和 m_j $(i \neq j)$ 之间协方差的证明，可以从第2.2节推导的任意两个随机变量 x 和 z 之和的方差公式入手。已知：

$$\text{Var}[x+z] = \text{Var}[x] + 2\text{Cov}[x,z] + \text{Var}[z]$$

将变量 x 和 z 分别替换为 m_i 和 m_j，我们有：

$$\text{Var}[m_i+m_j] = \text{Var}[m_i] + 2\text{Cov}[m_i,m_j] + \text{Var}[m_j]$$

根据公式2.40，可得 $\text{Var}[m_i] = N\mu_i(1-\mu_i)$ 和 $\text{Var}[m_j] = N\mu_j(1-\mu_j)$。将 m_i+m_j 看作一个新的随机变量，而其他随机变量相加为另一个随机变量，那么随机变量 m_i+m_j 服从参数为 $\mu_i+\mu_j$ 和 N 的二项分布，其方差为 $\text{Var}[m_i+m_j] = N(\mu_i+\mu_j)(1-\mu_i-\mu_j)$。多项分布的协方差可以推导如下：

$$\begin{aligned}
\text{Var}[m_i+m_j] &= \text{Var}[m_i] + 2\text{Cov}[m_i,m_j] + \text{Var}[m_j] \quad \Rightarrow \\
2\text{Cov}[m_i,m_j] &= \text{Var}[m_i+m_j] - \text{Var}[m_i] - \text{Var}[m_j] \\
&= N(\mu_i+\mu_j)(1-\mu_i-\mu_j) - N\mu_i(1-\mu_i) - N\mu_j(1-\mu_j) \\
&= N(\mu_i - \mu_i^2 - \mu_i\mu_j + \mu_j - \mu_i\mu_j - \mu_j^2 - \mu_i + \mu_i^2 - \mu_j + \mu_j^2) \\
&= -2N\mu_i\mu_j \quad \Rightarrow \\
\text{Cov}[m_i,m_j] &= -N\mu_i\mu_j
\end{aligned}$$

随机变量 m_i 和 m_j 之间存在此消彼长的关系，导致它们的协方差为负值（呈负相关性）。并且协方差的绝对值大小与观测的总次数 N 和各自发生的概率 μ_i 和 μ_j 的乘积成正比。

我们仍然先使用最大似然方法来估计多项分布的参数。假设我们对一枚骰子进行了 N 次投掷测试，其观察结果记为数据集 $\mathcal{D} = \{x_1, \ldots, x_N\}$。通过对数据集进行简单统计，我们可以得到骰子的每一面出现的次数 $m_1, \ldots, m_K$。对于数据集 $\mathcal{D}$，其似然函数可以表示为：

$$p(\mathcal{D}|\boldsymbol{\mu}) = \prod_{i=1}^{K} \mu_i^{m_i} \tag{2.42}$$

除了将似然函数转换为对数形式以简化计算之外，我们还需要确保求解得到的参数 $\boldsymbol{\mu}$ 满足约束条件 $\sum_{i=1}^{K} \mu_i = 1$。我们采用拉格朗日乘数法（Lagrange Multiplier）来引入这项约束，从而得到以下优化目标：

$$\sum_{i=1}^{K} m_i \ln \mu_i + \lambda \left(\sum_{i=1}^{K} \mu_i - 1 \right) \tag{2.43}$$

将上式对 μ_i 求导数，并让导数等于零来求其极值。我们可得:

$$\mu_i = -\frac{m_i}{\lambda} \tag{2.44}$$

将以上公式2.44代入约束 $\sum_{i=1}^{K} \mu_i = 1$，同时考虑到 $\sum_{i=1}^{K} m_i = N$，我们可知 $\lambda = -N$。至此，我们可以推导出多项分布参数 $\boldsymbol{\mu}$ 的最大似然估计结果为：

$$\mu_i = \frac{m_i}{N} \tag{2.45}$$

在接下来的讨论中，我们探讨如何运用贝叶斯方法来估计多项分布的参数 $\boldsymbol{\mu}$。同样地，首先需要为参数 $\boldsymbol{\mu}$ 选择一个适当的先验分布。需要注意的是，与二项分布和泊松分布的情况不同，多项分布的参数 $\boldsymbol{\mu}$ 是一个包含多个值的向量，并且这些值的总和必须等于一，即它们构成一个离散的概率分布。此外，我们依然优先考虑与似然函数共轭的先验分布。对于公式2.42所示的似然函数，其共轭先验也应为参数 μ_i 的指数形式：

$$p(\boldsymbol{\mu}|\boldsymbol{\alpha}) \propto \prod_{i=1}^{K} \mu_i^{\alpha_i - 1} \tag{2.46}$$

其中 $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K)^\top$ 为先验分布的超参数，μ_i 满足 $0 \leq \mu_i \leq 1$ 和 $\sum_{i=1}^{K} \mu_i = 1$ 的约束。通过引入适当的系数，我们可以将上式2.46归一化为概率分布，即**狄利克雷分布** (Dirichlet Distribution)，其概率密度函数为：

$$\mathrm{Dir}(\boldsymbol{\mu}|\boldsymbol{\alpha}) = \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1)\cdots\Gamma(\alpha_K)} \prod_{i=1}^{K} \mu_i^{\alpha_i - 1} \tag{2.47}$$

其中 $\Gamma(\alpha)$ 为伽玛函数，并且 $\alpha_0 = \sum_{i=1}^{K} \alpha_i$。

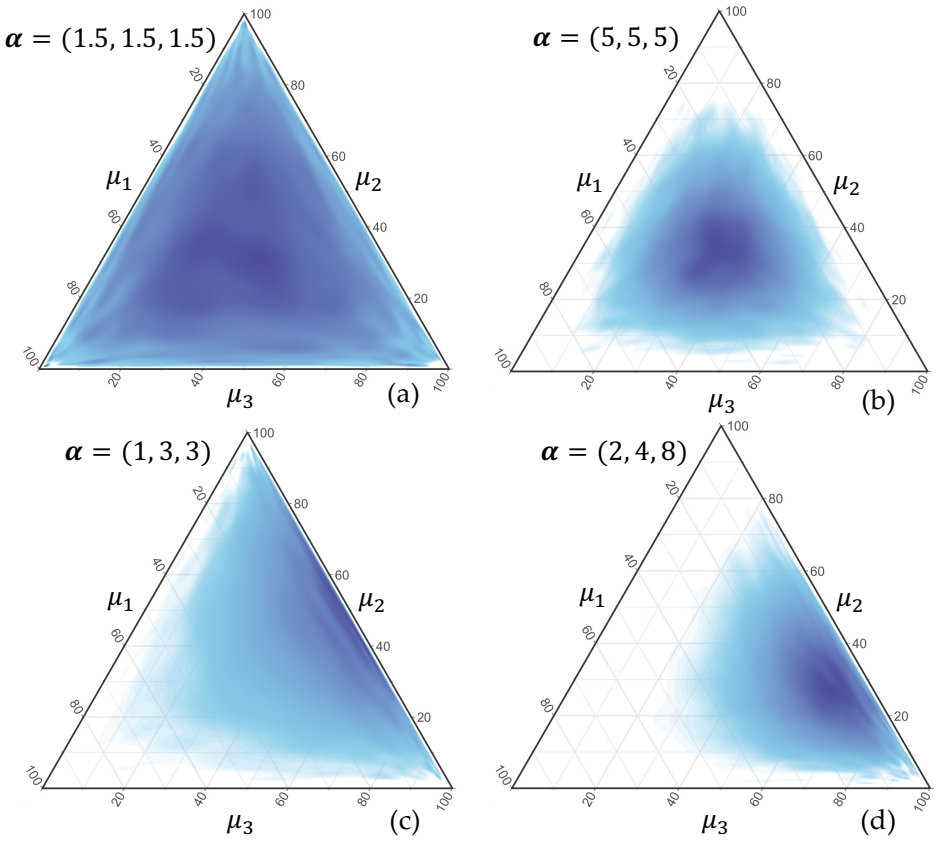


图 2.5: 超参数 $\boldsymbol{\alpha}$ 取四种不同值时的狄利克雷分布。颜色越深的区域表示概率值越高。

图2.5展示了包含三个超参、四种不同取值的狄利克雷分布的概率密度函数示意图。颜色的深浅代表概率值的高低，颜色越深表示对应区域的概率值越高。图中的三角形边缘上的刻度值

代表对应随机变量的取值（以百分数表示的概率），三角形所围的区域被称为单纯形（Simplex）。标准 D 维单纯形是在 $\mathbb{R}^{D+1}$ 空间中满足以下条件的一个子集：

$$\Delta^D = \left\{ (\mu_1, \ldots, \mu_{D+1}) \in \mathbb{R}^{D+1} \middle| \sum_{i=1}^{D+1} \mu_i = 1, \mu_i \geq 0 \right\} \tag{2.48}$$

显然，图2.5中每一个三角形所围的区域代表一个二维的单纯形，而区域中的每个点都对应一个概率分布。例如：区域的中心点代表概率分布 $\boldsymbol{\mu} = (1/3, 1/3, 1/3)$，三角形的三个顶点分别表示 $\boldsymbol{\mu} = (1, 0, 0)$（左下）、$\boldsymbol{\mu} = (0, 1, 0)$（顶部）和 $\boldsymbol{\mu} = (0, 0, 1)$（右下）。狄利克雷分布可以视作“分布的分布”，因为它定义了所有可能的离散分布发生的概率值。超参 $\boldsymbol{\alpha}$ 通常被称为浓度参数（Concentration Parameters），因为这些参数决定了分布的浓缩或集中程度。通过比较图2.5(a) 和 (b)，我们可以观察到，由于图2.5(b) 中 $\boldsymbol{\alpha}$ 向量的每一个分量都相对较大，所以分布表现得更为集中。此外，从图2.5(c) 中可以看出，分布主要集中在 μ_2 和 μ_3 取值较高的 $(0.3, 0.7)$ 区域，而颜色较深区域的 μ_1 值在 $(0, 0.3)$ 的低值区间。

狄利克雷分布的均值、方差和协方差分别为：

$$\mathbb{E}[\mu_i] = \frac{\alpha_i}{\alpha_0} \tag{2.49}$$

$$\text{Var}[\mu_i] = \frac{\alpha_i(\alpha_0 - \alpha_i)}{\alpha_0^2(\alpha_0 + 1)} \tag{2.50}$$

$$\text{Cov}[\mu_i, \mu_j] = \frac{-\alpha_i \alpha_j}{\alpha_0^2(\alpha_0 + 1)} \tag{2.51}$$

由于多项分布是二项分布在多个可能结果上的推广，因此它的边缘分布为二项分布。我们还知道贝塔分布可作为二项分布参数 μ 的先验分布，而狄利克雷分布可作为多项分布参数 $\boldsymbol{\mu}$ 的先验分布。由此类推，狄利克雷分布的边缘分布应该就是贝塔分布。只有两个随机变量的狄利克雷分布的概率密度函数与贝塔分布完全相同。考虑任意随机变量 μ_i，我们将随机变量 μ_i 之外所有变量合并成一个变量，并且根据所有 μ_j 相加必须等于一的约束，可知 $\sum_{j \neq i} \mu_j = 1 - \mu_i$，则随机变量 μ_i 的边缘分布的概率密度函数为：

$$p(\mu_i|\boldsymbol{\alpha}) = \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_i)\Gamma(\sum_{j \neq i} \alpha_j)} \mu_i^{\alpha_i - 1} (1 - \mu_i)^{\sum_{j \neq i} \alpha_j - 1}$$

将上式中的 α_i、$\sum_{j \neq i} \alpha_j$、α_0 和 μ_i 分别替换成 α、β、$\alpha + \beta$ 和 μ，则变成以下贝塔分布的概率密度函数：

$$\text{Beta}(\mu|\alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \mu^{\alpha - 1} (1 - \mu)^{\beta - 1}$$

因此狄利克雷分布的均值（公式2.49）和方差（公式2.50）可以使用贝塔分布的均值（公式2.19）和方差（公式2.20）公式简单计算得到。

我们来证明狄利克雷分布中任意两个随机变量 μ_i 和 μ_j $(i \neq j)$ 之间的协方差公式，从协方差的定义出发可以推导出以下等式（详见第2.2节）：

$$\text{Cov}[\mu_i, \mu_j] = \mathbb{E}[\mu_i \mu_j] - \mathbb{E}[\mu_i]\mathbb{E}[\mu_j]$$

已知 $\mathbb{E}[\mu_i] = \alpha_i/\alpha_0$ 和 $\mathbb{E}[\mu_j] = \alpha_j/\alpha_0$，$\mathbb{E}[\mu_i \mu_j]$ 可以应用连续型随机变量的期望公式进行求解，

具体过程如下（由于 $\sum_{k=1}^{K}\mu_k=1$，积分应在 $K-1$ 维单纯形上进行）：

$$\begin{aligned}
\mathbb{E}[\mu_i\mu_j] &= \int\cdots\int \mu_i\mu_j \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1)\cdots\Gamma(\alpha_K)}\prod_{k=1}^{K}\mu_k^{\alpha_k-1}\,d\mu_1\ldots\mu_{K-1} \\
&= \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_0+2)}\cdot\frac{\Gamma(\alpha_i+1)}{\Gamma(\alpha_i)}\cdot\frac{\Gamma(\alpha_j+1)}{\Gamma(\alpha_j)}\int\cdots\int\frac{\Gamma(\alpha_0+2)}{\Gamma(\alpha_1')\cdots\Gamma(\alpha_K')}\prod_{k=1}^{K}\mu_k^{\alpha_k'-1}\,d\mu_1\ldots\mu_{K-1} \\
&\quad\text{（如果 } k \text{ 等于 } i \text{ 或 } j\text{，则 } \alpha_k'=\alpha_k+1\text{；否则 } \alpha_k'=\alpha_k\text{）} \\
&= \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_0+2)}\cdot\frac{\Gamma(\alpha_i+1)}{\Gamma(\alpha_i)}\cdot\frac{\Gamma(\alpha_j+1)}{\Gamma(\alpha_j)}\quad\text{（参数为 } \boldsymbol{\alpha}' \text{ 的概率密度函数的积分结果等于 1）} \\
&= \frac{\Gamma(\alpha_0)}{\alpha_0(\alpha_0+1)\Gamma(\alpha_0)}\cdot\frac{\alpha_i\Gamma(\alpha_i)}{\Gamma(\alpha_i)}\cdot\frac{\alpha_j\Gamma(\alpha_j)}{\Gamma(\alpha_j)} \\
&= \frac{\alpha_i\alpha_j}{\alpha_0(\alpha_0+1)}
\end{aligned}$$

狄利克雷分布的协方差可以推导如下：

$$\begin{aligned}
\mathrm{Cov}[\mu_i,\mu_j] &= \mathbb{E}[\mu_i\mu_j]-\mathbb{E}[\mu_i]\mathbb{E}[\mu_j] \\
&= \frac{\alpha_i\alpha_j}{\alpha_0(\alpha_0+1)}-\frac{\alpha_i}{\alpha_0}\cdot\frac{\alpha_j}{\alpha_0} \\
&= \frac{\alpha_0\alpha_i\alpha_j-(\alpha_0+1)\alpha_i\alpha_j}{\alpha_0^2(\alpha_0+1)} \\
&= \frac{-\alpha_i\alpha_j}{\alpha_0^2(\alpha_0+1)}
\end{aligned}$$

将狄利克雷分布（公式2.47）作为多项分布参数 $\boldsymbol{\mu}$ 的先验分布，并结合参数 $\boldsymbol{\mu}$ 的似然函数（公式2.42），应用贝叶斯方法可以得到以下形式的后验分布：

$$p(\boldsymbol{\mu}|\mathcal{D},\boldsymbol{\alpha})=\mathrm{Dir}(\boldsymbol{\mu}|\boldsymbol{\alpha}+\boldsymbol{m})=\frac{\Gamma(\alpha_0+N)}{\Gamma(\alpha_1+m_1)\cdots\Gamma(\alpha_K+m_K)}\prod_{i=1}^{K}\mu_i^{\alpha_i+m_i-1} \tag{2.52}$$

该后验分布是参数为 $(\alpha_1+m_1,\ldots,\alpha_K+m_K)$ 的狄利克雷分布，并且与其先验分布呈共轭关系。后验分布的形式也比较直观，它是在先验分布参数 $\boldsymbol{\alpha}$ 的基础上，将 N 次新的观测结果中各类出现次数 m_i 加到对应的先验参数 α_i 上。在使用后验分布进行预测时，参数 μ_i（即事件 i 发生的概率）的期望为 $\mathbb{E}[\mu_i]=(\alpha_i+m_i)/(\alpha_0+N)$。

2.4 信息论

我们将在本节中介绍关于信息论（Information Theory）的基本概念和内容，这些知识对于理解和设计机器学习算法很有帮助。1948 年，美国数学家、电子工程师、计算机科学家和密码学家克劳德 · 艾尔伍德 · 香农（Claude Elwood Shannon）发表了其开创性论文《A Mathematical Theory of Communication》[70]，彻底改变了我们对信息的理解。信息论的核心观点是信息可以像质量或能量等物理量一样被度量。

信息通常以**比特**（Bit）为单位来度量，一个比特（类似一个可以表示 0 或 1 的二进制位）的

信息可以让我们在两个等可能的选项中做出抉择。为了理解为何如此，我们通过图2.6的例子来进行解释。假设我们从 A 出发，目标是到达 M。在 A 处时，有向左走（用 0 表示）前往 B 和向右走（用 1 表示）前往 C 两种选择。此时，当我们被告知应该向左走，则收到了一个比特信息，这一比特的信息让我们能够在两种不同走向中做出正确的选择。接着，我们来到了 B 处，又面临向左还是向右的抉择。另外一个比特信息能够让我们做出向右走到达 E 处的正确决定。注意 E 是 D、E、F 和 G 这四种可能的途径点之一，因此两个比特的信息使我们能够在面对四种不同的选择时做出正确的决定。同样，在到达 E 之后，我们需要第三个比特的信息（即向左走）才能成功抵达目标地 M。从 A 出发，经历三次向左或向右走的选择，可以到达八种（$8 = 2^3$）不同的终点。为了正确地到达目的地 M，需要在每个分叉点接收一个比特的信息以帮助选择，因此总共需要三个（$3 = \log_2 8$）比特的信息。更一般地，当我们面临 N 种等可能的选项时，需要 $\log_2 N$ 比特的信息量来帮助我们做出正确的抉择。

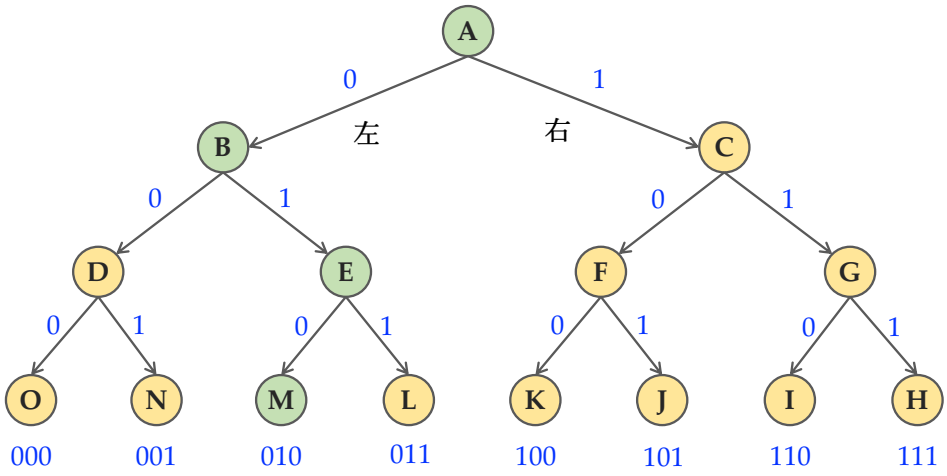


图 2.6: 从 A 出发，目的地为 M。在途径的每个分叉点，我们都面临两种选择：向左走（记为 0）或向右走（记为 1）。为了快速到达目的地 M，我们需要在每个分叉点接收指引信息以选择正确的前行方向。每次提供的指引信息会从两个选择中指定一个方向，相当于提供了一比特的信息。从 A 到 M 的整个过程需要三次方向指引（左走、右走、左走），因此总共需要三个比特的信息。

需要明确的是，虽然比特一词来源于二进制位（Binary Digit），但它们是截然不同的概念。一个二进制数是以二进制表示的具体数值，而比特则是信息量的度量单位。在信息论中，通常采用以 2 为底的对数。但也可以改用以自然常数 $e \approx 2.71828$ 为底的对数，此时，信息量的单位就变成了奈特或纳特（Nat）。

信息论主要关注信息的量化、存储和通信等方面。那么，上述关于信息度量的方法与信息的存储和通信有什么关联呢？为了解答这一问题，我们先将对数形式的信息量度量方法扩展至离散型随机变量。在通讯过程中，当发送者向接收者传输某一离散随机变量的特定值时，接收者获得了多少信息量呢？在图2.6所示的寻路例子中，所需信息量的计算公式 $\log_2 8$ 可以改写成：

$$\log_2 8 = \log_2 \frac{1}{1/8} = \log_2 \left(\frac{1}{8}\right)^{-1} = -\log_2 \frac{1}{8}$$

其中 $\frac{1}{8}$ 可以看作到达任一终点的概率（共有八种不同的终点）。若用随机变量 x 来描述，该随机变量有八个可能的结果，每个结果出现的概率均为 $\frac{1}{8}$。因此，确定出现某一特定结果（如例子中从 A 出发到达目的地 M）所需的信息量可计算为：

$$-\log_2 p(x) \quad \text{或者} \quad \log_2 p(x)^{-1} \tag{2.53}$$

其中 $p(x)$ 表示某一结果出现的概率。上式就是信息量的计算公式。根据该公式，知悉发生概率越低的结果或事件，所获得的信息量越大，反之则越小。因此，香农定义的信息量常常被形象地称为“惊讶度”（Degree of Surprise）。但这一定义本身并不令人惊讶。例如：某一事件必然会发生（$p(x)=1$），知悉该事件所能提供的信息量为零（$-\log_2 1 = 0$），因为它对于我们下一次预测该事件是否会发生毫无帮助。相反，在给定概率模型下，发生概率越低的结果越难预料，因此获知该结果时获得的自信息越大。进一步分析，所需的信息量与不确定性程度呈正比。不确定性越高或可能性越多，所需的信息量就越大，反之亦然。从这个角度来看，信息的获取实际上就是在降低不确定性，而这种不确定性降低的程度则对应着所获得信息的量。仍以图2.6为例，当我们受指引从 A 到达 B 后，可能抵达的终点数量从八个降到了四个，这表示不确定性从 $\frac{1}{8}$ 降到了 $\frac{1}{4}$，其降低程度正好等于指引我们从 A 到达 B 所获得的一比特信息，即 $-\log_2 \frac{1}{8} - (-\log_2 \frac{1}{4}) = 1$。

对于离散随机变量 x，其任一特定取值的信息量可以使用公式2.53来计算，而该离散变量 x 的平均信息量可以通过以下公式计算：

$$\mathrm{H}[x] = -\sum_x p(x)\log_2 p(x) \tag{2.54}$$

其中 $\mathrm{H}[x]$ 表示**信息熵**（Entropy），它是概率分布 $p(x)$ 对信息量的期望值。信息熵可以视为一个衡量随机变量整体不确定性的量度。那为何需要计算某个离散随机变量的信息熵呢？这个熵值有什么实际意义呢？其实通讯可以看作是发送者向接收者传递关于某个随机变量的信息的过程。以传输英文文本为例，英文文本一般由 26 个英文字母和少量标点符号组成。这些字符的出现并不是完全随机的，而是遵循一定的离散分布。从这个角度看，所有英文文本都可以视为从该离散分布中采样，并按照英语的语法和语义要求重新排序的结果。在实际的传输过程中，需要对这些字符进行编码，通常采用由 0 或 1 组成的二进制串来进行编码。根据信息论的原则，即发生概率低的事件携带的信息量较大，而发生概率高的事件携带的信息量较小。因此，我们为出现频率较少的字符分配较长的二进制编码（即具有较大的、以比特计量的信息量），而频率较高的字符则使用较短的编码。相对于等长编码，这样的编码策略能够有效地降低每个字符的平均编码长度，从而提高整体的编码效率。正如香农在 1948 年所阐述的，信息熵描述了存储或传输一个符号所需的平均比特数，它也是每个符号的平均编码长度通过编码优化可以达到的理论最小值。以下例子有助于理解信息熵的概念，并且了解如何在信息熵的指导下设计相应的编码。

假设我们需要传输的文本由 $\{a,b,c,d,e,f,g,h\}$ 八种字符组成，每个字符出现的概率分别为 $\{\frac{1}{2},\frac{1}{4},\frac{1}{8},\frac{1}{16},\frac{1}{64},\frac{1}{64},\frac{1}{64},\frac{1}{64}\}$[18]，这个离散概率分布的信息熵为：

$$\mathrm{H}[x] = -\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{4}\log_2\frac{1}{4} - \frac{1}{8}\log_2\frac{1}{8} - \frac{1}{16}\log_2\frac{1}{16} - \frac{4}{64}\log_2\frac{1}{64} = 2 \text{ 比特}$$

熵值为 2 比特的含义是，当使用二进制编码时，每个符号的平均编码长度的下界为 2，无法再

进一步缩短。在这个例子中，我们正好能够设计出平均长度为 2 比特的编码方式。对于字符集 $\{a, b, c, d, e, f, g, h\}$，我们分别采用 $\{1, 00, 011, 0100, 010100, 010101, 010110, 010111\}$ 进行编码。在这种情况下，我们可以如下计算每个字符的平均编码长度：

$$\text{平均编码长度} = \frac{1}{2} \times 1 + \frac{1}{4} \times 2 + \frac{1}{8} \times 3 + \frac{1}{16} \times 4 + \frac{4}{64} \times 6 = 2 \text{ 比特}$$

这种编码方式在理论上达到了最优，即字符的平均编码长度等于信息熵的值。在通讯过程中，一般先将字符转换为相应的编码，然后以串行方式进行传输。在接收端解码时，先需要将连续的串行编码进行切分，切分后的每个部分对应于一个字符。设计不当的不等长编码会导致在接收端解码时产生多种可能的切分，从而引发混淆。上述例子中所采用的不等长编码不会导致这样的混淆，因为每个编码都不是其他编码的前缀，所以不会出现解码时的切分歧义。

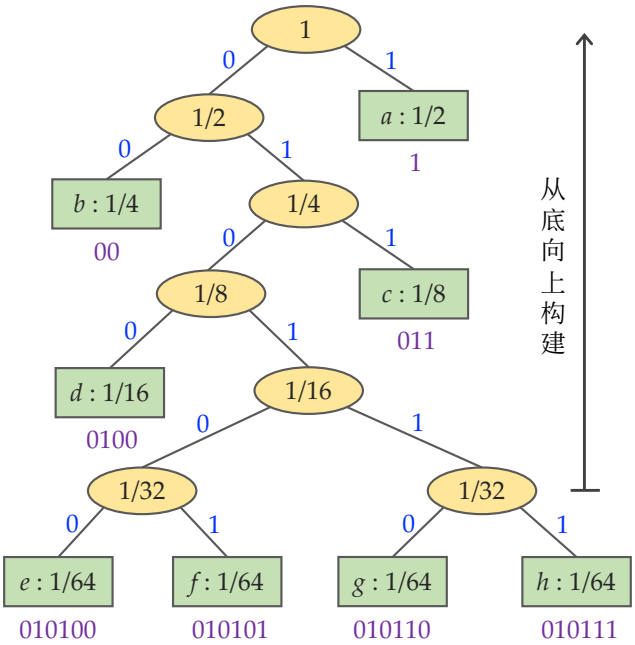


图 2.7: 霍夫曼编码构建过程的示例。获得的二进制编码以紫色标识。

可以采用霍夫曼编码算法来构建这种不等长编码，所得到的编码即为霍夫曼编码（Huffman Code）[17]。其构建过程如图2.7所示。首先，为每个字符创建一个结点（图中的绿色矩形），并将结点的值设为字符出现的概率。然后，每次挑选两个值最低的结点，并将其合并为一个新的结点（图中的黄色椭圆），新结点的值设为被合并结点的值之和。例如，我们首先选择出现概率都为 $\frac{1}{64}$ 的字符 e 和 f 这两个结点进行合并，合并之后的新结点的值设定为 $\frac{1}{32}$。重复上述过程，直到所有结点被合并为一个结点（图中最上方、值为 1 的黄色椭圆）。最终，我们会得到一棵如图2.7所示的二叉树，其中树的叶结点都为字符结点。将树的左分叉标记为 0，右分叉标记为 1。对于任一字符，从树的根结点到该字符叶结点的路径上，依次读出所经过的分叉标记，即可得到该字符的二进制编码。

如果采用等长编码的策略，因为有八种不同的字符，那么每个字符将需要三位二进制进行

编码。因此，霍夫曼编码在提升通信效率方面通常明显优于等长编码。假设我们想要传输 10 万个字符，等长编码需要发送 30 万个二进制位，而采用不等长的霍夫曼编码只需要发送 20 万个二进制位，所需传输的比特数减少了三分之一。类似方法也可以应用于文件压缩，与等长编码策略相比，能够节省三分之一的存储空间。但当字符集 $\{a, b, c, d, e, f, g, h\}$ 中的每个字符出现的概率相同（即都为 $\frac{1}{8}$）时，这个离散概率分布的信息熵为：

$$\mathrm{H}[x] = -\frac{1}{8}\log_2\frac{1}{8} \times 8 = 3 \text{ 比特}$$

在这种等概率分布的情况下，我们无法通过优化编码来减少字符的平均编码长度。因此，我们可以发现（如图2.8所示），如果分布在少数几个值附近呈尖锐的峰值，则其熵将相对较低；而概率在许多值上呈较均匀的分布，则其熵将相对较高（离散分布中，均匀分布的熵值最高）。信息熵的值越小，就越有可能通过优化编码来提高通信或存储的效率，反之则效果有限。

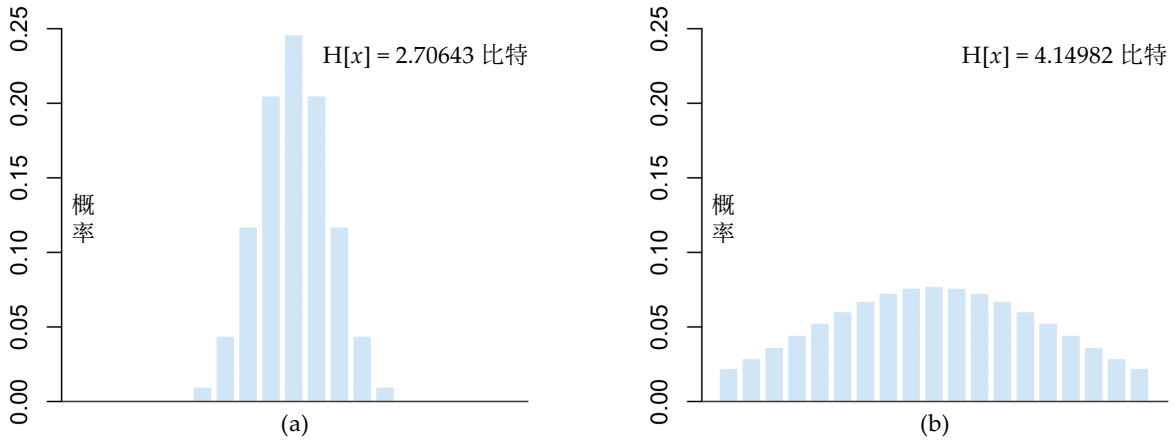


图 2.8: 两个不同概率分布的柱状图及其信息熵值。(a) 分布越集中于少数值，则其熵相对较低。(b) 分布较均匀分散于多个值，则其熵相对较高。

根据离散随机变量信息熵的定义，信息熵可以看作随机变量的概率分布对所有可能取值的信息量求期望值。将信息熵概念推广到连续型随机变量时，需要将公式2.54中的概率函数替换成概率密度函数，同时将连加操作替换为积分操作。对于连续型随机变量 x，其**微分熵**（Differential Entropy）或**连续熵**（Continuous Entropy）定义如下[1]：

$$\mathrm{H}[x] = \mathbb{E}[-\ln p(x)] = -\int p(x)\ln p(x)\,dx \tag{2.55}$$

这里的 $p(x)$ 表示连续型随机变量 x 的概率密度函数。为了便于后续将熵的概念关联到机器学习的相关内容，此处还将对数函数的底改成自然常数 e。连续随机变量的信息熵可被称为连续熵，相应地，离散随机变量的信息熵也可简称为**离散熵**。

[1]虽然最初是由香农提出将微分熵作为离散型随机变量的信息熵（简称为离散熵）到连续型随机变量的扩展，但严格来说，微分熵并非离散熵的精确对应。微分熵缺少离散熵在度量不确定性方面的一些有用的性质（例如可能会产生负值）。为了更严格地建立离散熵与连续情形之间的联系，埃德温 · 汤普森 · 杰恩斯（Edwin Thompson Jaynes）提出了基于离散点的极限密度（Limiting Density of Discrete Points）的方法，通过引入参考密度来定义连续情形下的熵。鉴于本书的主题和范围，我们不再进一步深入讨论并严格区分这些概念。

公式2.55描述的是一维随机变量的微分熵，很容易将其推广到多维随机变量的情形。以二维随机变量为例，设随机变量 x 和 y 的联合分布函数为 $p(x,y)$，那么这个联合分布的微分熵为：

$$\mathrm{H}[x,y] = -\iint p(x,y)\ln p(x,y)\,dx\,dy \tag{2.56}$$

给定随机变量 x 的值，另一个随机变量 y 的特定值能够提供的额外信息量为 $-\ln p(y|x)$。因此，在已知 x 值的前提下，确定 y 值所需要的额外平均信息量可以通过以下公式计算：

$$\mathrm{H}[y|x] = -\iint p(x,y)\ln p(y|x)\,dx\,dy \tag{2.57}$$

以上 $\mathrm{H}[y|x]$ 被称为**条件熵**（Conditional Entropy），它与随机变量 x 和 y 联合分布的微分熵之间的关系如下：

$$\mathrm{H}[x,y] = \mathrm{H}[y|x] + \mathrm{H}[x] \tag{2.58}$$

以上公式表明，确定 x 和 y 的值所需的信息量等于确定 x 值所需的信息量，再加上在已知 x 值的情况下，确定 y 值所需的额外信息量。公式2.58可以推导如下：

$$\begin{aligned}
\mathrm{H}[x,y] &= \mathrm{H}[y|x] + \mathrm{H}[x] \quad \Rightarrow \\
\mathrm{H}[x] &= \mathrm{H}[x,y] - \mathrm{H}[y|x] \\
&= -\iint p(x,y)\ln p(x,y)\,dx\,dy + \iint p(x,y)\ln p(y|x)\,dx\,dy \\
&= -\iint p(x,y)\left(\ln p(x,y) - \ln p(y|x)\right)\,dx\,dy \\
&= -\iint p(x,y)\ln\left(\frac{p(x,y)}{p(y|x)}\right)\,dx\,dy \\
&= -\iint p(x,y)\ln\left(\frac{p(y|x)p(x)}{p(y|x)}\right)\,dx\,dy \\
&= -\iint p(x,y)\ln p(x)\,dx\,dy \\
&= -\int p(x)\ln p(x)\,dx
\end{aligned}$$

以上我们介绍了信息论的一些基本概念，下面我们讨论如何将这些概念和思想应用于机器学习。在机器学习应用中，往往需要构造一个概率分布 $q(x)$ 来逼近一个未知的目标分布 $p(x)$。根据信息论原理，对于一个给定的连续分布 $p(x)$，其随机变量 x 取特定值所能提供的信息量为 $-\ln p(x)$，并且该随机变量 x 的平均信息量或微分熵为 $-\int p(x)\ln p(x)\,dx$。类比离散随机变量信息熵的意义，适当量化后的微分熵的值相当于对随机变量 x 的每个可能状态进行编码时，理论上的最短平均编码长度。当用近似分布 $q(x)$ 来替代目标分布 $p(x)$ 时，随机变量 x 取特定值的估计信息量变为 $-\ln q(x)$，而相应的“微分熵”变为 $-\int p(x)\ln q(x)\,dx$（应称为交叉熵）。在计算近似微分熵时，我们仍然使用目标分布 $p(x)$ 来对估计信息量求期望。这是因为随机变量 x 的实际取值分布并不会因为我们用分布 $q(x)$ 来估计相应的信息量而发生改变。近似微分熵的值相当于使用 $q(x)$，而非实际分布 $p(x)$ 来为随机变量 x 设计“最优”编码方案的平均编码长度。近似微分熵一定大于使用目标分布 $p(x)$ 计算得到的微分熵（稍后会给出严格的证明），这是因为微

分熵是最优编码长度的理论下界，而使用近似分布 $q(x)$ 进行编码设计必然导致平均编码长度的增加。最优编码设计的基本原则是为出现频率较高的状态分配较短的编码，而为出现频率较低的状态分配较长的编码。当使用与实际概率分布不一致的近似分布来设计编码方案时，就可能会出现所分配编码的长度与状态出现的概率不相匹配的情况。例如：只有两个状态的分布 $p(x)$，每个状态发生概率均为 0.5，其信息熵为 1。如果使用另一个分布 $q(x)$ 来近似 $p(x)$，其中 $q(x)$ 两个状态发生概率的估计值分别为 0.51 和 0.49，其近似信息熵约等于 1.00029，略大于 1。

上述近似微分熵与实际微分熵之间的差值被称为 **Kullback–Leibler 散度**（Kullback–Leibler Divergence）或**相对熵**（Relative Entropy）[46]，简称为 KL 散度。它表示当使用近似分布 $q(x)$ 为随机变量 x 的每个状态所设计的最优编码方案，相较于使用实际分布 $p(x)$ 所得到的最优编码方案需要额外增加的平均信息量（以奈特为单位），其表达式如下所示：

$$D_{\mathrm{KL}}(p||q) = -\int p(x)\ln q(x)\,dx - \left(-\int p(x)\ln p(x)\,dx\right) = -\int p(x)\ln\left(\frac{q(x)}{p(x)}\right)\,dx \tag{2.59}$$

KL 散度始终为非负值（即 $D_{\mathrm{KL}}(p||q) \geq 0$），且当两个分布 $q(x)$ 和 $p(x)$ 完全一致时，其值为零。因此，KL 散度可用于衡量两个概率分布之间的距离。但值得注意的是，KL 散度是非对称的，即 $D_{\mathrm{KL}}(p||q) \neq D_{\mathrm{KL}}(q||p)$，并且它也不满足三角不等式（即 $D_{\mathrm{KL}}(p||q) \leq D_{\mathrm{KL}}(p||g) + D_{\mathrm{KL}}(g||q)$ 不一定成立）。因此，KL 散度不能作为概率分布所构成的度量空间（Metric Space）中，计算任意两个分布之间距离的度量函数（Metric）。这是它通常被称为 “散度” 而非 “距离” 的原因。

由于 KL 散度可以用于衡量两个概率分布之间的距离，当我们试图构建一个概率分布 $q(x)$ 来逼近未知的目标分布 $p(x)$ 时，可以通过最小化 $q(x)$ 与 $p(x)$ 之间的 KL 散度来实现这一目标。假设分布 $q(x)$ 由一组参数 $\boldsymbol{\theta}$ 确定，则该分布可以表示为 $q(x|\boldsymbol{\theta})$。我们通过调整参数 $\boldsymbol{\theta}$ 来优化分布 $q(x)$，使其更好地逼近目标分布 $p(x)$。虽然目标分布 $p(x)$ 是未知的，但一般可以从中采样得到包含 N 个样本的训练集 $\mathcal{D} = \{x_i\}_{i=1}^N$。对于分布 $p(x)$ 求期望可以通过对训练集中的所有样本求平均值来近似，则 KL 散度公式2.59可以改写为以下形式：

$$D_{\mathrm{KL}}(p||q) \simeq \frac{1}{N}\sum_{i=1}^{N}\{-\ln q(x_i|\boldsymbol{\theta}) + \ln p(x_i)\}$$

上式右边的第二项 $\ln p(x_i)$ 并不依赖于参数 $\boldsymbol{\theta}$，在优化过程中可以忽略。第一项则是对于训练集使用近似分布 $q(x_i|\boldsymbol{\theta})$ 计算的负对数似然函数（即生成训练集的近似概率）。因此，最小化 KL 散度实际上等同于最大化似然函数。

当随机变量 x 和 y 是相互独立的，它们的联合概率分布 $p(x,y)$ 可以分解为两个边缘分布（Marginal Distribution）的乘积，即 $p(x,y) = p(x)p(y)$。我们可以通过 $p(x,y)$ 与 $p(x)p(y)$ 之间的 KL 散度来判断随机变量 x 和 y 是否独立。如果此 KL 散度为零，则表明 x 和 y 是相互独立的；如果散度大于零，则意味着它们不是独立的。这种联合分布 $p(x,y)$ 与其边缘分布乘积 $p(x)p(y)$ 之间的 KL 散度被称为随机变量 x 和 y 之间的**互信息**（Mutual Information）：

$$\mathrm{I}[x,y] \equiv D_{\mathrm{KL}}(p(x,y)||p(x)p(y)) = -\iint p(x,y)\ln\left(\frac{p(x)p(y)}{p(x,y)}\right)\,dx\,dy \tag{2.60}$$

我们可以推导出以下互信息与条件熵之间的关联：

$$\mathrm{I}[x,y] = \mathrm{H}[x] - \mathrm{H}[x|y] = \mathrm{H}[y] - \mathrm{H}[y|x] \tag{2.61}$$

因此，互信息可以理解为，知晓 y 的值后对 x 的不确定性减少的程度，或者知晓 x 的值后对 y 的不确定性减少的程度。公式2.61可以推导如下：

$$\begin{aligned}
\mathrm{H}[x] - \mathrm{H}[x|y] &= -\int p(x)\ln p(x)\,dx - \left(-\iint p(x,y)\ln p(x|y)\,dx\,dy\right) \\
&= -\iint p(x,y)\ln p(x)\,dx\,dy + \iint p(x,y)\ln p(x|y)\,dx\,dy \\
&= -\iint p(x,y)\ln\left(\frac{p(x)}{p(x|y)}\right)\,dx\,dy \\
&= -\iint p(x,y)\ln\left(\frac{p(x)p(y)}{p(x|y)p(y)}\right)\,dx\,dy \\
&= -\iint p(x,y)\ln\left(\frac{p(x)p(y)}{p(x,y)}\right)\,dx\,dy \\
&= \mathrm{I}[x,y]
\end{aligned}$$

在这一节的最后部分，我们将证明 KL 散度总是大于或等于零的，即 $D_{\mathrm{KL}}(p||q) \geq 0$。为了完成这一证明，我们需要引入凸函数的概念。如图2.9(a) 所示，凸函数任意两点之间的连线始终位于函数曲线的上方。任意介于 a 和 b 两点之间的点可以表示为 $\lambda a + (1-\lambda)b$，其中 $0 \leq \lambda \leq 1$。该点在函数上的值为 $f(\lambda a + (1-\lambda)b)$，而相应 $f(a)$ 和 $f(b)$ 两点之间连线上的点可以表示为 $\lambda f(a) + (1-\lambda)f(b)$。如果函数满足以下不等式，则该函数被称为**凸函数**（Convex Function）：

$$f(\lambda a + (1-\lambda)b) \leq \lambda f(a) + (1-\lambda)f(b) \tag{2.62}$$

对于定义在区间上的二阶可导函数，凸性等价于其二阶导数处处非负。当 $a \neq b$，如果公式2.62中的等式仅在 $\lambda = 0$ 或 $\lambda = 1$ 时成立，那么该函数被称为严格凸函数。若将公式2.62中的小于等于号 $(\leq)$ 替换为大于等于号 $(\geq)$，则满足这个新不等式的函数被定义为**凹函数**（Concave Function），示例可见图2.9(b)。显而易见，如果 $f(x)$ 是凸函数，那么 $-f(x)$ 是凹函数，反之亦然。

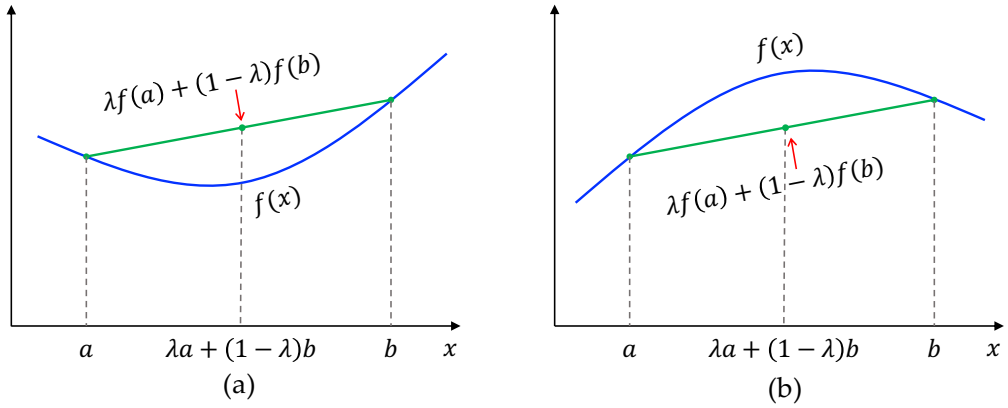


图 2.9: 凸函数和凹函数的特性对比。(a) 凸函数任意两点之间的连线始终位于函数曲线的上方。(b) 凹函数任意两点之间的连线始终位于函数曲线的下方。

利用数学归纳法（Mathematical Induction），我们可以证明凸函数满足以下的不等式：

$$f\left(\sum_{i=1}^{N}\lambda_i x_i\right) \leq \sum_{i=1}^{N}\lambda_i f(x_i) \tag{2.63}$$

其中 $\lambda_i \geq 0$，并且 $\sum_{i=1}^{N}\lambda_i = 1$。上式被称为**琴生不等式**（Jensen's Inequality），其证明过程也比较简单直观。在此，我们仅给出证明的基本思路。公式2.62作为应用数学归纳法的奠基步骤，即在两个 x_i 的情况下，公式2.63是成立的。然后，假设公式在 $N-1$ 个 x_i 的情况下成立，可以通过递推步骤证明当有 N 个 x_i 时，公式2.63仍然成立。这个结论是显然的，因为仅在 $N-1$ 个 x_i 的基础上通过调整权重 λ_i 来添加一个新的点。若将 λ_i 看作随机变量 x 在不同 x_i 值上的离散概率分布，那么公式2.63可以改写成：

$$f(\mathbb{E}[x]) \leq \mathbb{E}[f(x)] \tag{2.64}$$

其中 $\mathbb{E}[\cdot]$ 表示求期望。对于连续变量，琴生不等式的形式如下：

$$f\left(\int xp(x)\,dx\right) \leq \int f(x)p(x)\,dx \tag{2.65}$$

将上式应用到 KL 散度的公式2.59，我们可以得到：

$$\begin{aligned}
D_{\mathrm{KL}}(p||q) &= -\int p(x)\ln\left(\frac{q(x)}{p(x)}\right)\,dx \\
&\geq -\ln\left(\int \frac{q(x)}{p(x)}p(x)\,dx\right) && (-\ln(x)\text{ 为凸函数}) \\
&= -\ln\left(\int q(x)\,dx\right) \\
&= 0 && \left(\int q(x)\,dx = 1\right)
\end{aligned}$$

至此，$D_{\mathrm{KL}}(p||q) \geq 0$ 得证，并且当两个分布 $p(x)$ 和 $q(x)$ 完全相同时，KL 散度的值为零。因此，KL 散度可以作为衡量两个概率分布之间差异的指标。

2.5 高斯分布

高斯分布（Gaussian Distribution），又称为**正态分布**（Normal Distribution），它是最常用的连续型随机变量的概率分布之一。根据中心极限定理（Central Limit Theorem），在满足某些较为宽松的条件下，一组独立随机变量的和（其本身也是一个随机变量）会随着随机变量数量的增加而趋近于高斯分布。举例来说，假设有 N 个随机变量 $x_1,\ldots,x_N$，并且这些变量在区间 $[0,1]$ 上服从均匀分布。当 N 足够大时，样本均值 $(x_1+\cdots+x_N)/N$ 将趋向于高斯分布（如图2.10）。

对于单变量 x 的情况，一维高斯分布可以写成以下形式[2]：

$$\mathcal{N}(x|\mu,\sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}}\exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right) \tag{2.66}$$

[2]本书采用 e^x 或 $\exp(x)$ 两种方式表示自然常数 e 的指数函数，具体选择取决于表达的简洁性和可读性。当指数部分较为简单时，使用 e^x 形式；而在指数部分较为复杂时，则采用 $\exp(x)$ 形式。

其中 μ 表示均值，σ^2 为方差（σ 称为标准差），$\exp(\cdot)$ 为自然常数 e 的指数函数。对于一个 D 维向量 $\boldsymbol{x}$，多维高斯分布（又称多元高斯分布）的形式为：

$$\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^{\top}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right) \tag{2.67}$$

其中 $\boldsymbol{\mu}$ 表示 D 维的均值向量，$\boldsymbol{\Sigma}$ 则是 $D \times D$ 的协方差矩阵。$|\boldsymbol{\Sigma}|$ 表示矩阵 $\boldsymbol{\Sigma}$ 的行列式（Determinant），是一个标量。一维和多维高斯分布如图2.11所示。

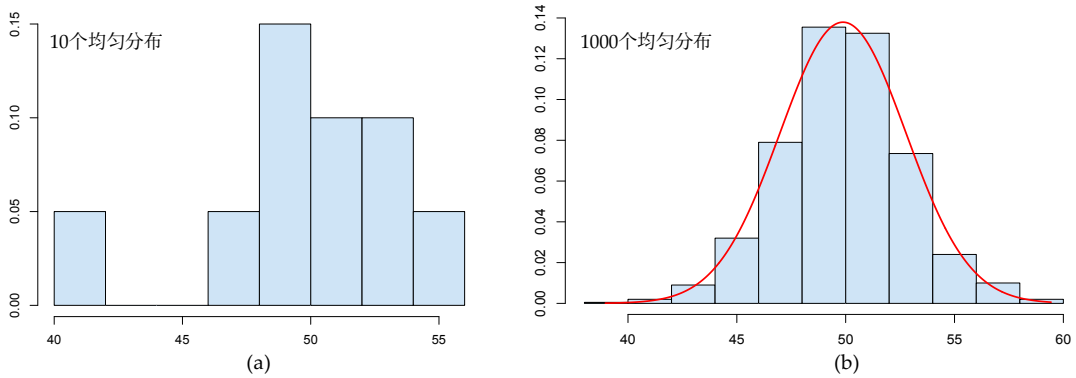


图 2.10: 独立随机变量之和随着随机变量数量增加趋近于高斯分布的示例（随机生成 100 组）。(a) 十个在 $[0,1]$ 区间服从均匀分布的随机变量之和。(b) 千个在 $[0,1]$ 区间服从均匀分布的随机变量之和。

在2.4信息论一节讨论编码问题时，我们已经了解到，在离散分布中，均匀分布的熵值是最高的。而在具有给定均值和方差的连续分布中，高斯分布具有最大的微分熵。接下来，我们简要证明这一结论。我们的目标是寻找某种形式的概率分布 $p(x)$，使其能够最大化微分熵：

$$\mathrm{H}[x] = -\int_{-\infty}^{\infty} p(x)\ln p(x)\,dx \tag{2.68}$$

并且满足以下三个约束条件：

$$\int_{-\infty}^{\infty} p(x)\,dx = 1 \tag{2.69}$$

$$\int_{-\infty}^{\infty} xp(x)\,dx = \mu \tag{2.70}$$

$$\int_{-\infty}^{\infty} (\mu-x)^2 p(x)\,dx = \sigma^2 \tag{2.71}$$

其中第一个约束条件是要求概率分布在所有可能取值上的积分为 1，即满足概率的归一化条件，而第二个和第三个约束则对应于均值和方差的定义。采用拉格朗日乘数法，我们可以得到如下的最大化目标函数：

$$-\int_{-\infty}^{\infty} p(x)\ln p(x)\,dx + \lambda_0\left(\int_{-\infty}^{\infty} p(x)\,dx - 1\right)$$
$$+\lambda_1\left(\int_{-\infty}^{\infty} xp(x)\,dx - \mu\right) + \lambda_2\left(\int_{-\infty}^{\infty} (\mu-x)^2 p(x)\,dx - \sigma^2\right)$$

其中 λ_0、λ_1 和 λ_2 分别是与三个约束条件对应的拉格朗日乘数。为了求得使上述目标函数最大

化的 $p(x)$，通常的做法是将目标函数对 $p(x)$ 求导，并将导数设为零，以解出相应的 $p(x)$。由于与 $p(x)$ 无关的项的导数为零，因此可以将这些项舍去，得到如下形式：

$$\int_{-\infty}^{\infty} -p(x)\ln p(x) + \lambda_0 p(x) + \lambda_1 x p(x) + \lambda_2(\mu - x)^2 p(x)\, dx$$

将上式在积分内的部分对 $p(x)$ 求导并设为零（此过程实际需要使用变分法），可以得到：

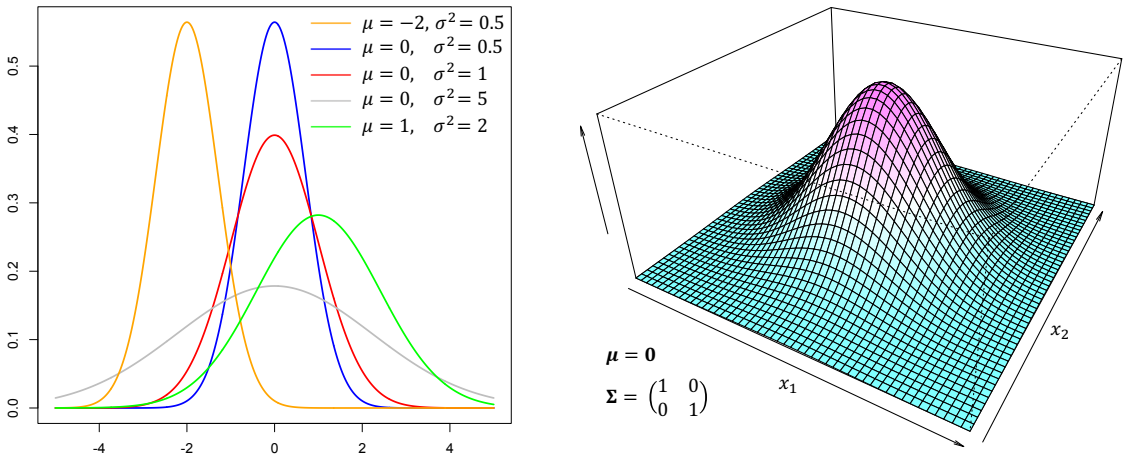
$$-\ln p(x) - 1 + \lambda_0 + \lambda_1 x + \lambda_2(\mu - x)^2 = 0 \quad \Rightarrow$$
$$p(x) = \exp\left(-1 + \lambda_0 + \lambda_1 x + \lambda_2(\mu - x)^2\right)$$

使微分熵最大的分布通常在均值附近呈对称性。对称分布意味着随机变量在均值两侧的表现是无偏的，即没有倾向于任何一侧。如果分布偏向均值的某一侧，其不确定性就会降低，从而导致微分熵的减少。对指数中的二次式配方后，其中心为 $\mu - \lambda_1/(2\lambda_2)$。根据均值必须等于 μ 的约束，可得 $\lambda_1 = 0$。然后，我们将 $p(x) = \exp\left(-1 + \lambda_0 + \lambda_2(\mu - x)^2\right)$ 代入公式2.69和2.71，即可求解出 $\lambda_0 = 1 + \ln(2\pi\sigma^2)^{-1/2}$ 和 $\lambda_2 = -1/(2\sigma^2)$。最后，将求得的 λ_0 和 λ_2 代回 $p(x)$ 的表达式中，就可以得到高斯分布的概率密度函数（公式2.66）。

如果对单变量高斯分布计算其微分熵，我们可以得到如下形式：

$$\mathrm{H}[x] = \mathbb{E}[-\ln p(x)] = \frac{1}{2}\left(1 + \ln(2\pi\sigma^2)\right) \tag{2.72}$$

从上式可以看出，随着高斯分布的方差增大（即分布变得更宽），其微分熵也随之增大。我们还观察到，当 $\sigma^2 < \frac{1}{2\pi e}$ 时，其微分熵 $\mathrm{H}[x] < 0$。与离散熵总是为非负不同，微分熵可能出现负值。



(a) (b)

图 2.11: 高斯分布的示例。(a) 一维高斯分布。(b) 二维高斯分布。

从一维高斯分布的概率密度函数（公式2.66）可以看出，高斯分布的实质是以数据点与均值 μ 之间的欧氏距离（Euclidean Distance）来确定其概率值。由于其指数部分存在负号，这意味着数据点距离均值越近，其对应的概率值就越大，反之则越小。对于多维高斯分布（公式2.67），让我们来分析其指数的二次项部分：

$$\Delta^2 = (\boldsymbol{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}) \tag{2.73}$$

其中 Δ 被称为从 $\boldsymbol{x}$ 到 $\boldsymbol{\mu}$ 的马氏距离（Mahalanobis Distance）。当协方差矩阵 $\boldsymbol{\Sigma}$ 是一个单位矩阵（即对角线元素为 1，其余元素为 0 的方阵）时，则马氏距离退化为欧氏距离。

协方差矩阵 $\boldsymbol{\Sigma}$ 是一个实对称矩阵。因此，$D \times D$ 的协方差矩阵有 D 个实数特征值 λ_i 及其特征向量 $\boldsymbol{u}_i$：

$$\boldsymbol{\Sigma}\boldsymbol{u}_i = \lambda_i \boldsymbol{u}_i, \quad i \in \{1, \ldots, D\} \tag{2.74}$$

其中 $\boldsymbol{u}_i$ 是两两正交且标准化的向量：

$$\boldsymbol{u}_i^\top \boldsymbol{u}_j = \begin{cases} 1, & \text{当 } i = j \\ 0, & \text{当 } i \neq j \end{cases} \tag{2.75}$$

根据公式2.74，我们有：

$$\boldsymbol{\Sigma} \begin{bmatrix} \boldsymbol{u}_1 & \cdots & \boldsymbol{u}_D \end{bmatrix} = \begin{bmatrix} \boldsymbol{u}_1 & \cdots & \boldsymbol{u}_D \end{bmatrix} \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_D \end{bmatrix} \Rightarrow$$

$$\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{u}_1 & \cdots & \boldsymbol{u}_D \end{bmatrix} \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_D \end{bmatrix} \begin{bmatrix} \boldsymbol{u}_1 & \cdots & \boldsymbol{u}_D \end{bmatrix}^{-1}$$

由于 $\boldsymbol{u}_i$ 是两两正交且标准化的向量，则由它们所组成的矩阵的逆矩阵就是其转置矩阵。因此，协方差矩阵 $\boldsymbol{\Sigma}$ 可以表示为：

$$\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{u}_1 & \cdots & \boldsymbol{u}_D \end{bmatrix} \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_D \end{bmatrix} \begin{bmatrix} \boldsymbol{u}_1^\top \\ \cdots \\ \boldsymbol{u}_D^\top \end{bmatrix} \Rightarrow$$

$$\boldsymbol{\Sigma} = \sum_{i=1}^{D} \lambda_i \boldsymbol{u}_i \boldsymbol{u}_i^\top$$

同样由于特征向量 $\boldsymbol{u}_i$ 两两正交的性质，协方差矩阵的逆矩阵 $\boldsymbol{\Sigma}^{-1}$ 可以表示为：

$$\boldsymbol{\Sigma}^{-1} = \sum_{i=1}^{D} \frac{1}{\lambda_i} \boldsymbol{u}_i \boldsymbol{u}_i^\top \tag{2.76}$$

将公式2.76代入公式2.73，多维高斯分布概率密度函数的指数部分的二次项可表示为：

$$\Delta^2 = \sum_{i=1}^{D} \frac{y_i^2}{\lambda_i} \tag{2.77}$$

其中 $y_i = \boldsymbol{u}_i^\top (\boldsymbol{x} - \boldsymbol{\mu})$。我们可以将 y_i 理解为由标准正交向量 $\boldsymbol{u}_i$ 所定义的新坐标系，这些向量相对于原始的 x_i 坐标进行了平移（减去均值向量 $\boldsymbol{\mu}$）和旋转（乘以特征向量 $\boldsymbol{u}_i$）。

从以上分析可以看出，协方差矩阵（描述了各维度的方差以及各维度之间的相关性）所起的作用是在计算多维高斯分布的概率时，数据点到均值的距离会先转换到相互正交的各特征向量方向上，然后使用对应特征值的倒数对各方向的距离进行加权（如公式2.77所示）。使用特征

值（协方差矩阵的特征值就表示数据在对应特征向量方向上的方差）的倒数进行加权是为了在计算数据点离均值的距离时，将不同方向上的尺度（即方差）调整到统一的尺度。

与之前几节一样，先使用最大似然估计方法来估计高斯分布的参数。我们先从简单的一维高斯分布开始，然后推广到多维高斯分布的情况。假设给定数据集 $\mathcal{D} = \{x_1, \dots, x_N\}$，其中每个样本 x_i 都是从同一个一维高斯分布 $\mathcal{N}(\mu, \sigma^2)$ 中独立采样的，其对数似然函数可以表示为：

$$\ln p(\mathcal{D}|\mu, \sigma^2) = \sum_{i=1}^{N} \ln \mathcal{N}(x_i|\mu, \sigma^2) = -\frac{N}{2}\ln(2\pi) - \frac{N}{2}\ln(\sigma^2) - \frac{1}{2}\sum_{i=1}^{N}\frac{(x_i-\mu)^2}{\sigma^2} \tag{2.78}$$

将上述对数似然函数分别对均值 μ 和方差 σ^2 求偏导，并令偏导为零，我们可以得到以下结果：

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i \tag{2.79}$$

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)^2 \tag{2.80}$$

公式2.79和2.80分别为高斯分布均值和方差的最大似然估计。我们先使用公式2.79计算出 μ 的值，然后再利用它来计算方差 σ^2。将其推广至多维高斯分布时，均值 $\boldsymbol{\mu}$ 和方差 $\boldsymbol{\Sigma}$ 的估计分别为：

$$\boldsymbol{\mu} = \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{x}_i \tag{2.81}$$

$$\boldsymbol{\Sigma} = \frac{1}{N}\sum_{i=1}^{N}(\boldsymbol{x}_i-\boldsymbol{\mu})(\boldsymbol{x}_i-\boldsymbol{\mu})^\top \tag{2.82}$$

其中 $\{\boldsymbol{x}_i\}_{i=1}^N$ 表示从同一多维高斯分布 $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ 中独立抽取的 N 个样本。

最大似然估计得到的参数值实际上也是随机变量，对于不同的样本数据集可能会得到不同的估计值。我们希望这些估计值的期望等于未知参数的真实值，即应当是无偏的。无偏估计需要满足 $\mathbb{E}[\hat{\theta}] = \theta$ 的条件，其中 $\hat{\theta}$ 表示参数 θ 的估计。无论随机变量服从何种分布，公式2.79和2.81是均值的无偏估计。对于方差的估计，当均值也未知时，则公式2.80和2.82是有偏的。我们以一维高斯分布的方差估计为例，来说明其并非为无偏估计。对公式2.80的右侧求期望（由于原公式中的均值 μ 也是估计值，所以改用 $\hat{\mu}$ 表示），可以得到以下结果：

$$\begin{aligned}
\mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}(x_i-\hat{\mu})^2\right] &= \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}x_i^2 - \hat{\mu}^2\right] \\
&= \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}x_i^2\right] - \mathbb{E}\left[\hat{\mu}^2\right] && \text{（期望的线性性质）} \\
&= \mathbb{E}\left[x_i^2\right] - \left(\mathrm{Var}[\hat{\mu}] + \mathbb{E}[\hat{\mu}]^2\right) && \left(\mathrm{Var}[x] = \mathbb{E}[x^2] - \mathbb{E}[x]^2\right) \\
&= \sigma^2 + \mu^2 - \left(\frac{\sigma^2}{N} + \mu^2\right) && \left(\mathrm{Var}[\hat{\mu}] = \mathrm{Var}\left[\frac{1}{N}\sum_{i=1}^{N}x_i\right]\right. \\
&= \frac{N-1}{N}\sigma^2 && \left. = \frac{1}{N^2}\sum_{i=1}^{N}\mathrm{Var}[x_i] = \frac{\sigma^2}{N}\right)
\end{aligned}$$

从以上推导中可以看出，最大似然估计得到的方差的期望是真实方差的 $(N-1)/N$ 倍。为了得到无偏估计，我们可以将最大似然估计乘以修正因子 $N/(N-1)$，这一方法称为无偏化。因此，公式2.80和2.82可以改写为无偏化后的样本方差计算公式：

$$\sigma^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i-\mu)^2 \tag{2.83}$$

$$\boldsymbol{\Sigma} = \frac{1}{N-1}\sum_{i=1}^{N}(\boldsymbol{x}_i-\boldsymbol{\mu})(\boldsymbol{x}_i-\boldsymbol{\mu})^\top \tag{2.84}$$

显然，用以上两个公式计算得到的方差估计的期望就等于真实的方差。

接下来，我们讨论如何使用贝叶斯方法来估计高斯分布的参数。我们同样先从较为简单的一维高斯分布开始入手。高斯分布包含均值 μ 和方差 σ^2 两类参数。因此，我们将分别讨论三种情况：已知方差时估计均值，已知均值时估计方差，以及均值与方差均未知的情形。

首先讨论方差 σ^2 已知，均值 μ 未知的情形。采用贝叶斯方法，需要为均值 μ 选择一个合适的先验分布。给定数据集 $\mathcal{D}=\{x_1,\dots,x_N\}$，其中每个样本 x_i 都是从同一个高斯分布 $\mathcal{N}(\mu,\sigma^2)$ 中独立抽取出来的，其似然函数的形式为：

$$p(\mathcal{D}|\mu,\sigma^2) = \prod_{i=1}^{N}\mathcal{N}(x_i|\mu,\sigma^2) = \frac{1}{(2\pi\sigma^2)^{N/2}}\exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{N}(x_i-\mu)^2\right) \tag{2.85}$$

从以上似然函数可以看出，它是关于均值 μ 的二次项的指数形式。如果 μ 的先验分布 $p(\mu)$ 也为高斯分布，那么似然函数与先验分布的乘积依然保持 μ 的二次项的指数形式，其后验分布也将是一个高斯分布。在这种情况下，均值 μ 的先验分布与后验分布是共轭的。假设均值 μ 取以下形式的先验分布：

$$p(\mu) = \mathcal{N}(\mu|\mu_0,\sigma_0^2) \tag{2.86}$$

其中 μ_0 是作为先验的高斯分布的均值，而 σ_0^2 是方差。均值 μ 的后验分布正比于：

$$p(\mu|\mathcal{D}) \propto p(\mathcal{D}|\mu,\sigma^2)\mathcal{N}(\mu|\mu_0,\sigma_0^2) \tag{2.87}$$

由于后验分布也是一个高斯分布，我们可以通过配方法（Completing the Square）进行求解。在具体介绍该方法之前，让我们回顾一下高斯分布的概率密度函数（公式2.66）。如以下公式及其注释所示，它可以看成由一个归一化项和一个指数项所组成，其中指数项可进一步分解为方差的倒数项和均值的二次项：

$$\mathcal{N}(x|\mu,\sigma^2) = \underbrace{\frac{1}{(2\pi\sigma^2)^{1/2}}}_{\text{归一化项}}\exp\underbrace{\left(-\frac{1}{2}\overbrace{\sigma^{-2}}^{\text{方差的倒数项}}\overbrace{(x-\mu)^2}^{\text{均值的二次项}}\right)}_{\text{指数项}} \tag{2.88}$$

其中归一化项是为了确保概率密度函数的积分总和为一，因此可以等确定了指数项之后再求得。高斯分布的关键部分在其指数项，特别是形如 $(x-\mu)^2$ 的二次项。对于某个未知参数值的高斯分布，只要将其指数项通过配方法配出类似 $(x-\mu)^2$ 这样的完全平方形式，就可以方便地确定

该分布的均值与方差。采用配方法求解公式2.87所示的后验高斯分布的均值与方差的过程如下：

$$
\begin{aligned}
p(\mu|\mathcal{D}) &\propto p(\mathcal{D}|\mu,\sigma^2)\mathcal{N}(\mu|\mu_0,\sigma_0^2) \\
&= \frac{1}{(2\pi\sigma^2)^{N/2}}\exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{N}(x_i-\mu)^2\right)\frac{1}{(2\pi\sigma_0^2)^{1/2}}\exp\left(-\frac{1}{2\sigma_0^2}(\mu-\mu_0)^2\right) \\
&= \frac{1}{(2\pi\sigma^2)^{N/2}(2\pi\sigma_0^2)^{1/2}}\exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{N}(x_i-\mu)^2-\frac{1}{2\sigma_0^2}(\mu-\mu_0)^2\right)
\end{aligned}
$$

根据上述分析，我们只需要关注指数项。因此，从上式中提取出指数部分，并对其进一步推导：

$$
\begin{aligned}
&-\frac{1}{2\sigma^2}\sum_{i=1}^{N}(x_i-\mu)^2-\frac{1}{2\sigma_0^2}(\mu-\mu_0)^2 \\
=&-\frac{1}{2\sigma^2\sigma_0^2}\left(\sigma_0^2\sum_{i=1}^{N}(x_i-\mu)^2+\sigma^2(\mu-\mu_0)^2\right) \\
=&-\frac{1}{2\sigma^2\sigma_0^2}\left(\sigma_0^2\sum_{i=1}^{N}(x_i^2-2x_i\mu+\mu^2)+\sigma^2(\mu^2-2\mu_0\mu+\mu_0^2)\right) \\
=&-\frac{1}{2\sigma^2\sigma_0^2}\left(\sigma_0^2\sum_{i=1}^{N}x_i^2-2N\sigma_0^2\hat{\mu}\mu+N\sigma_0^2\mu^2+\sigma^2\mu^2-2\sigma^2\mu_0\mu+\sigma^2\mu_0^2\right)\quad\left(\hat{\mu}=\frac{1}{N}\sum_{i=1}^{N}x_i\right) \\
=&-\frac{1}{2\sigma^2\sigma_0^2}\left((N\sigma_0^2+\sigma^2)\mu^2-2(\sigma^2\mu_0+N\sigma_0^2\hat{\mu})\mu+\text{const}\right)\quad\text{（const 表示与 }\mu\text{ 无关的常数项）} \\
=&-\frac{1}{2}\frac{N\sigma_0^2+\sigma^2}{\sigma^2\sigma_0^2}\left(\mu^2-2\frac{\sigma^2\mu_0+N\sigma_0^2\hat{\mu}}{N\sigma_0^2+\sigma^2}\mu+\text{const}\right) \\
=&-\frac{1}{2}\left(\frac{1}{\sigma_0^2}+\frac{N}{\sigma^2}\right)\left(\left(\mu-\frac{\sigma^2\mu_0+N\sigma_0^2\hat{\mu}}{N\sigma_0^2+\sigma^2}\right)^2+\text{const}\right)
\end{aligned}
$$

从以上推导的最后一步可以看出，均值 μ 的后验分布为：

$$
p(\mu|\mathcal{D})=\mathcal{N}(\mu|\mu_N,\sigma_N^2) \tag{2.89}
$$

其中：

$$
\mu_N=\frac{\sigma^2}{N\sigma_0^2+\sigma^2}\mu_0+\frac{N\sigma_0^2}{N\sigma_0^2+\sigma^2}\hat{\mu} \tag{2.90}
$$

$$
\frac{1}{\sigma_N^2}=\frac{1}{\sigma_0^2}+\frac{N}{\sigma^2} \tag{2.91}
$$

公式2.90表明后验分布的均值 μ_N 实际上是先验分布的均值 μ_0 和最大似然估计 $\hat{\mu}$ 的加权和，即两者之间的折衷。在没有可观察样本（$N=0$）的情况下，μ_N 等于先验分布的均值 μ_0。当样本数量 N 趋向于无穷大时，后验分布的均值 μ_N 逐渐逼近最大似然估计 $\hat{\mu}$。公式2.91表明，随着样本数量 N 的增加，后验分布的方差 σ_N^2 将逐渐减小，使得分布变得越来越集中。当 N 趋向于无穷大时，σ_N^2 趋向于零，即分布将变得极其狭窄且集中于其唯一的峰值 $\hat{\mu}$。

在多维高斯分布的情况下，当协方差矩阵 $\boldsymbol{\Sigma}$ 已知，但均值 $\boldsymbol{\mu}$ 未知时，其结果也是显然的。仍然选择高斯分布作为均值 $\boldsymbol{\mu}$ 的先验分布，只需将公式2.90和2.91中的方差 σ^2 改成协方差矩阵

$\boldsymbol{\Sigma}$，就可以得到后验分布的均值和协方差矩阵：

$$\boldsymbol{\mu}_N = \boldsymbol{\Sigma}_N \left(\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0 + N\boldsymbol{\Sigma}^{-1}\hat{\boldsymbol{\mu}}\right) \tag{2.92}$$

$$\boldsymbol{\Sigma}_N = \left(\boldsymbol{\Sigma}_0^{-1} + N\boldsymbol{\Sigma}^{-1}\right)^{-1} \tag{2.93}$$

其中 $\boldsymbol{\mu}_0$ 和 $\boldsymbol{\Sigma}_0$ 分别是均值 $\boldsymbol{\mu}$ 的先验分布的均值和协方差矩阵。

接着，我们来讨论均值 μ 已知，而方差 σ^2 未知的情形。在这种情形中，方差 σ^2 用其倒数 $\lambda = 1/\sigma^2$ 比较方便处理，这里的 λ 称为高斯分布的精度 (Precision)。给定数据集 $\mathcal{D} = \{x_1, \ldots, x_N\}$，用精度 λ 表示的似然函数形式如下：

$$p(\mathcal{D}|\mu, \lambda^{-1}) = \prod_{i=1}^{N} \mathcal{N}(x_i|\mu, \lambda^{-1}) = \left(\frac{\lambda}{2\pi}\right)^{N/2} \exp\left(-\frac{\lambda}{2}\sum_{i=1}^{N}(x_i - \mu)^2\right) \tag{2.94}$$

从以上似然函数的形式可以看出，与其共轭的先验分布需要是 λ 幂次和以 λ 的线性函数为指数的指数函数的乘积形式。我们在第2.3节中介绍过的伽玛分布正好符合这样的要求：

$$\text{Gamma}(\lambda|\alpha, \beta) = \frac{\beta^\alpha \lambda^{\alpha-1} e^{-\beta\lambda}}{\Gamma(\alpha)} \tag{2.95}$$

假设选择 $\text{Gamma}(\lambda|\alpha_0, \beta_0)$ 作为精度 λ 的先验分布，则其后验分布 $p(\lambda|\mathcal{D})$ 将正比于：

$$p(\lambda|\mathcal{D}) \propto \lambda^{\alpha_0-1}\lambda^{N/2} \exp\left(-\beta_0\lambda - \frac{\lambda}{2}\sum_{i=1}^{N}(x_i - \mu)^2\right) \tag{2.96}$$

由于后验与先验分布是共轭的，后验分布仍然是一个伽玛分布 $\text{Gamma}(\lambda|\alpha_N, \beta_N)$，其参数可以计算为：

$$\alpha_N = \alpha_0 + \frac{N}{2} \tag{2.97}$$

$$\beta_N = \beta_0 + \frac{1}{2}\sum_{i=1}^{N}(x_i - \mu)^2 = \beta_0 + \frac{N}{2}\hat{\sigma}^2 \tag{2.98}$$

其中 $\hat{\sigma}^2$ 是采用最大似然估计所得到的方差值（公式2.80）。在求解后验分布的过程中，我们并不需要考虑公式2.96中先验分布和似然函数的归一化项和常数项，因为它们最终会在伽玛分布（公式2.95）的归一化项中体现。

对于多维高斯分布，若均值 $\boldsymbol{\mu}$ 已知，而协方差矩阵 $\boldsymbol{\Sigma}$ 未知时，我们需要使用**威沙特分布**（Wishart Distribution）来作为精度矩阵（$\boldsymbol{\Lambda} = \boldsymbol{\Sigma}^{-1}$）的先验分布，该分布在此情况下是相应似然函数的共轭先验。威沙特分布是伽玛分布在多维上的扩展，其概率密度函数的定义如下：

$$\mathcal{W}(\boldsymbol{\Lambda}|\mathbf{W}, \eta) = \left(2^{\eta D/2}|\mathbf{W}|^{\eta/2}\Gamma_D\left(\frac{\eta}{2}\right)\right)^{-1} |\boldsymbol{\Lambda}|^{(\eta-D-1)/2} \exp\left(-\frac{1}{2}\text{tr}\left(\mathbf{W}^{-1}\boldsymbol{\Lambda}\right)\right) \tag{2.99}$$

其中 $\boldsymbol{\Lambda}$ 是一个 $D \times D$ 的半正定对称矩阵，η 是分布的自由度，$\mathbf{W}$ 则是一个 $D \times D$ 的正定尺度矩阵。$|\mathbf{W}|$ 表示矩阵 $\mathbf{W}$ 的行列式，$\text{tr}(\cdot)$ 表示矩阵的迹，而 Γ_D 是多维伽玛函数，其定义如下：

$$\Gamma_D\left(\frac{\eta}{2}\right) = \pi^{D(D-1)/4} \prod_{j=1}^{D} \Gamma\left(\frac{\eta}{2} + \frac{1-j}{2}\right) \tag{2.100}$$

给定包含 N 个样本的数据集 $\mathcal{D} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$，并且选择精度矩阵的先验分布为 $\mathcal{W}(\boldsymbol{\Lambda}|\mathbf{W}_0, \eta_0)$，

则后验分布仍为威沙特分布 $\mathcal{W}(\mathbf{\Lambda}|\mathbf{W}_N, \eta_N)$，其参数可以按以下方式计算：

$$\mathbf{W}_N^{-1} = \mathbf{W}_0^{-1} + \sum_{i=1}^{N}(\boldsymbol{x}_i - \boldsymbol{\mu})(\boldsymbol{x}_i - \boldsymbol{\mu})^\top \tag{2.101}$$

$$\eta_N = \eta_0 + N \tag{2.102}$$

最后，我们来讨论均值 μ 和方差 σ^2 都未知的情形。我们仍然优先考虑与似然函数共轭的先验分布。此时，我们不能再假设均值 μ 或方差 σ^2 中的任何一个已知，并将其作为常数对待，而需要将似然函数看成同时依赖于均值 μ 和方差 σ^2（使用 λ^{-1} 表示以便于处理），然后为其找到合适的先验分布。

$$\begin{aligned}
p(\mathcal{D}|\mu, \lambda^{-1}) &= \prod_{i=1}^{N}\mathcal{N}(x_i|\mu, \lambda^{-1}) \\
&= \prod_{i=1}^{N}\left(\frac{\lambda}{2\pi}\right)^{1/2}\exp\left(-\frac{\lambda}{2}(x_i-\mu)^2\right) \\
&\propto \left(\lambda^{1/2}\exp\left(-\frac{\lambda\mu^2}{2}\right)\right)^N \exp\left(\lambda\mu\sum_{i=1}^{N}x_i - \frac{\lambda}{2}\sum_{i=1}^{N}x_i^2\right) \\
&= \left(\lambda^{1/2}\exp\left(-\frac{\lambda\mu^2}{2}\right)\right)^\nu \exp\left(c\lambda\mu - d\lambda\right) \quad \left(\nu = N, c = \sum_{i=1}^{N}x_i, d = \frac{1}{2}\sum_{i=1}^{N}x_i^2\right) \\
&= \exp\left(-\frac{\nu\lambda}{2}\left(\mu - \frac{c}{\nu}\right)^2\right)\lambda^{\nu/2}\exp\left(-\left(d - \frac{c^2}{2\nu}\right)\lambda\right) \\
&\propto \underbrace{\left(\frac{\nu\lambda}{2\pi}\right)^2\exp\left(-\frac{\nu\lambda}{2}(\mu-\upsilon)^2\right)}_{\mathcal{N}(\mu|\upsilon,(\nu\lambda)^{-1})}\underbrace{\lambda^{\alpha-1}\exp\left(-\beta\lambda\right)}_{\text{Gamma}(\lambda|\alpha,\beta)} \quad \left(\upsilon = \frac{c}{\nu}, \alpha = 1 + \frac{\nu}{2}, \beta = d - \frac{c^2}{2\nu}\right)
\end{aligned}$$

从以上推导的最后一步可以看出，似然函数可以表示为高斯分布 $\mathcal{N}(\mu|\upsilon, (\nu\lambda)^{-1})$ 与伽玛分布 $\text{Gamma}(\lambda|\alpha, \beta)$ 的乘积形式。因此，如果选择的先验分布与此形式相同，则这种先验分布将是该似然函数的共轭先验。这样的先验分布被称为**高斯-伽玛分布**（Gaussian-Gamma Distribution）：

$$\text{NormalGamma}(\mu, \lambda|\upsilon, \nu, \alpha, \beta) = \mathcal{N}(\mu|\upsilon, (\nu\lambda)^{-1})\text{Gamma}(\lambda|\alpha, \beta) \tag{2.103}$$

值得注意的是，均值 μ 与精度 λ 的联合先验分布 $p(\mu, \lambda)$ 可以分解为 $p(\mu|\lambda)p(\lambda)$，其中 $p(\lambda)$ 对应于伽玛分布 $\text{Gamma}(\lambda|\alpha, \beta)$，而 $p(\mu|\lambda)$ 对应于高斯分布 $\mathcal{N}(\mu|\upsilon, (\nu\lambda)^{-1})$。由于均值 μ 的条件先验分布 $p(\mu|\lambda)$ 依赖于参数 λ，所以高斯-伽玛分布并非关于均值 μ 和精度 λ 两个独立分布的乘积。在似然函数中，均值 μ 与精度 λ 是相互关联的（无法完全分离），因此即使我们选择了均值 μ 与精度 λ 相互独立的先验分布，其后验分布仍会出现均值 μ 与精度 λ 之间依赖或耦合的情况。给定包含 N 个样本的数据集 $\mathcal{D} = \{x_1, \ldots, x_N\}$，并选择 $\text{NormalGamma}(\mu, \lambda|\upsilon_0, \nu_0, \alpha_0, \beta_0)$ 作为先验分布，则后验分布仍为高斯-伽玛分布 $\text{NormalGamma}(\mu, \lambda|\upsilon_N, \nu_N, \alpha_N, \beta_N)$。先验结合

似然函数后，后验分布的参数可以按以下方式计算：

$$\upsilon_N = \frac{\nu_0}{\nu_0 + N}\upsilon_0 + \frac{N}{\nu_0 + N}\hat{\mu} \tag{2.104}$$

$$\nu_N = \nu_0 + N \tag{2.105}$$

$$\alpha_N = \alpha_0 + \frac{N}{2} \tag{2.106}$$

$$\beta_N = \beta_0 + \frac{1}{2}\left(N\hat{\sigma^2} + \frac{\nu_0 N(\hat{\mu} - \upsilon_0)^2}{\nu_0 + N}\right) \tag{2.107}$$

其中 $\hat{\mu}$ 是使用最大似然估计所得到的均值，而 $\hat{\sigma}^2$ 是方差的最大似然估计，其计算公式分别为：

$$\hat{\mu} = \frac{1}{N}\sum_{i=1}^{N} x_i \tag{2.108}$$

$$\hat{\sigma}^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \hat{\mu})^2 \tag{2.109}$$

对于多维高斯分布的情况，如果均值 $\boldsymbol{\mu}$ 和精度矩阵 $\boldsymbol{\Lambda}$ 都未知，则需要使用**高斯-威沙特分布**（Gaussian-Wishart Distribution）作为先验分布。给定包含 N 个样本的数据集 $\mathcal{D} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$，并且采用高斯-威沙特分布 $\mathcal{N}(\boldsymbol{\mu}|\boldsymbol{\upsilon}_0, (\nu_0\boldsymbol{\Lambda})^{-1})\mathcal{W}(\boldsymbol{\Lambda}|\mathbf{W}_0, \eta_0)$ 作为先验分布。由于分布之间的共轭性质，后验分布仍然为高斯-威沙特分布 NormalWishart$(\boldsymbol{\mu}, \boldsymbol{\Lambda}|\boldsymbol{\upsilon}_N, \nu_N, \mathbf{W}_N, \eta_N)$，其参数可按以下方式计算：

$$\boldsymbol{\upsilon}_N = \frac{\nu_0}{\nu_0 + N}\boldsymbol{\upsilon}_0 + \frac{N}{\nu_0 + N}\hat{\boldsymbol{\mu}} \tag{2.110}$$

$$\nu_N = \nu_0 + N \tag{2.111}$$

$$\mathbf{W}_N^{-1} = \mathbf{W}_0^{-1} + \sum_{i=1}^{N}(\boldsymbol{x}_i - \hat{\boldsymbol{\mu}})(\boldsymbol{x}_i - \hat{\boldsymbol{\mu}})^\top + \frac{N\nu_0}{\nu_0 + N}(\hat{\boldsymbol{\mu}} - \boldsymbol{\upsilon}_0)(\hat{\boldsymbol{\mu}} - \boldsymbol{\upsilon}_0)^\top \tag{2.112}$$

$$\eta_N = \eta_0 + N \tag{2.113}$$

其中 $\hat{\boldsymbol{\mu}}$ 是使用最大似然估计所得到的均值向量（公式2.81）。

若两组变量服从联合高斯分布，则在已知一组变量的情况下，另一组变量的条件概率分布依然为高斯分布。同样地，任意一组变量的边缘分布也是高斯分布。借助之前所介绍的配方法，容易推导出以下推论。给定多维高斯分布 $\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$，将随机变量 $\boldsymbol{x}$ 任意划分成 $\boldsymbol{x}_1$ 和 $\boldsymbol{x}_2$ 两组：

$$\boldsymbol{x} = \begin{bmatrix}\boldsymbol{x}_1 \\ \boldsymbol{x}_2\end{bmatrix}, \quad \boldsymbol{\mu} = \begin{bmatrix}\boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2\end{bmatrix}, \quad \boldsymbol{\Lambda} = \boldsymbol{\Sigma}^{-1}$$

$$\boldsymbol{\Sigma} = \begin{bmatrix}\boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22}\end{bmatrix}, \quad \boldsymbol{\Lambda} = \begin{bmatrix}\boldsymbol{\Lambda}_{11} & \boldsymbol{\Lambda}_{12} \\ \boldsymbol{\Lambda}_{21} & \boldsymbol{\Lambda}_{22}\end{bmatrix}$$

则给定 $\boldsymbol{x}_2$ 时，$\boldsymbol{x}_1$ 的条件概率分布为：

$$p(\boldsymbol{x}_1|\boldsymbol{x}_2) = \mathcal{N}(\boldsymbol{x}_1|\boldsymbol{\mu}_{1|2}, \boldsymbol{\Lambda}_{11}^{-1}) \tag{2.114}$$

$$\boldsymbol{\mu}_{1|2} = \boldsymbol{\mu}_1 - \boldsymbol{\Lambda}_{11}^{-1}\boldsymbol{\Lambda}_{12}(\boldsymbol{x}_2 - \boldsymbol{\mu}_2) \tag{2.115}$$

随机变量 $\boldsymbol{x}_1$ 的边缘分布为：

$$p(\boldsymbol{x}_1) = \mathcal{N}(\boldsymbol{x}_1|\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11}) \tag{2.116}$$

尽管高斯分布是最常用的概率分布之一，但它也存在一些局限性。首先，多维高斯分布的协方差矩阵含有 $D(D+1)/2$ 个参数。当矩阵维度 D 较大时，处理如求逆这样的大矩阵运算会产生较大的计算代价，并且需要更多的数据用于估计与 D^2 呈正比增长的参数。减少参数量的方法是采用某种限定形式的协方差矩阵。例如，我们可以考虑对角矩阵 $\boldsymbol{\Sigma} = \text{diag}(\sigma_i^2)$，这将使参数量从 $D(D+1)/2$ 降到 D。如图2.12(b) 所示，此时二维高斯分布的概率等高线（呈椭圆形）的长轴和短轴将与坐标轴对齐。我们可以进一步将其设为各向同性的协方差矩阵（Isotropic Covariance）$\boldsymbol{\Sigma} = \sigma^2\mathbf{I}$，从而将参数量减少到 1。如图2.12(c) 所示，此时二维高斯分布的概率等高线将呈现为同心圆。这些方法虽然减少了协方差矩阵的参数量，并且加快了求逆等运算的速度，但同时也限制了高斯模型捕捉数据间相关性的能力。

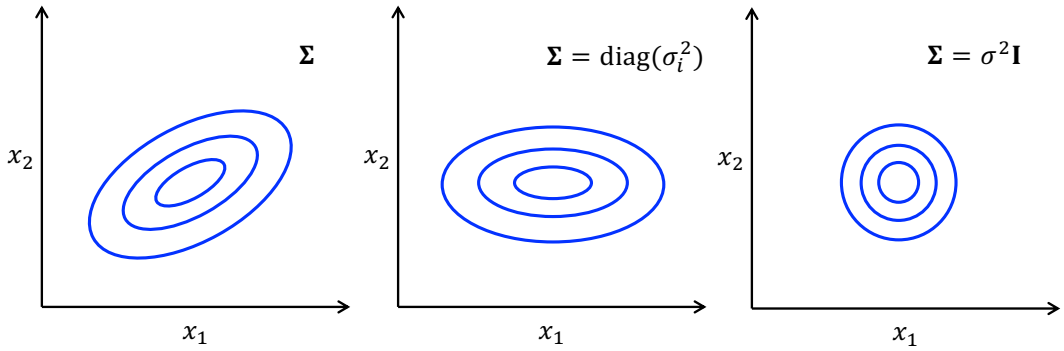


图 2.12: 二维高斯分布的概率等高线示意图，其中协方差矩阵为：(a) 一般形式；(b) 对角形式，椭圆形等高线的长短轴与坐标轴对齐；(c) 各向同性（参数 σ^2 乘以单位矩阵），等高线为同心圆。

高斯分布的另一个局限性是它本质上是单峰的（即在其均值处取得唯一最大值），从而不能较好地拟合多峰分布。一方面，从协方差矩阵参数过多的角度来看，高斯分布可能过于灵活（有过多的参数可供调整）。另一方面，由于其本质上的单峰性质，其在能够表达的概率分布的范围上又显得过于受限。一种解决方案是采用**混合高斯模型**（Mixture of Gaussians），即将数据表示为 K 个高斯分布的叠加形式：

$$p(\boldsymbol{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \tag{2.117}$$

其中每个 $\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ 称为混合高斯模型的子成分，而 π_k 为混合系数，需要满足以下条件：

$$\sum_{k=1}^{K} \pi_k = 1, \quad 0 \leq \pi_k \leq 1 \tag{2.118}$$

如图2.13(a) 所示，单一高斯分布（灰色曲线）无法有效地拟合由三个高斯分布（蓝色、绿色和橙色曲线）生成的数据，因为该单一高斯分布的均值附近并非（唯一的）数据密集区域。相比之

下，图2.13(b) 中的混合高斯模型（红色曲线）通过增加足够多的子成分，并调整它们的均值、协方差及混合系数，几乎可以以任意精度逼近任何复杂的分布。由于样本所属的混合成分未被观测到，边缘对数似然中的各成分参数相互耦合，通常无法直接求得最大似然估计的解析解。我们将在第10章中介绍如何使用期望最大化算法（Expectation Maximization）来估计混合高斯模型的参数。

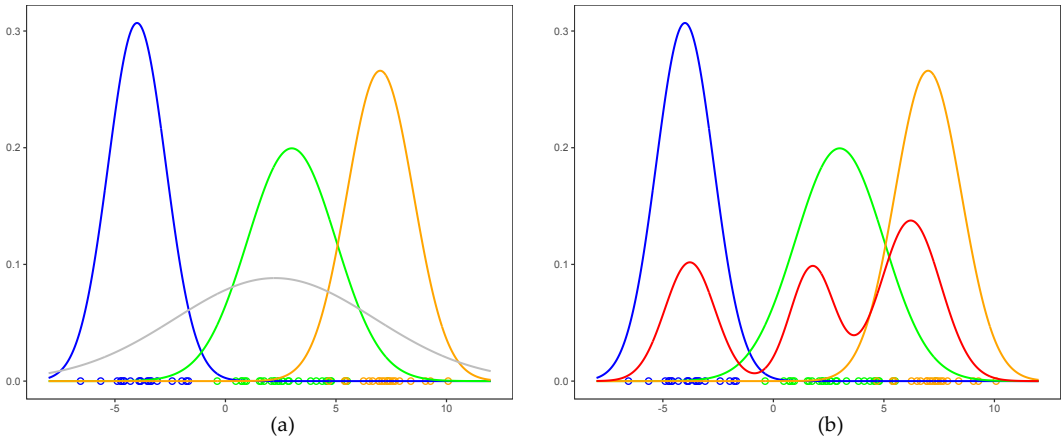


图 2.13: 混合高斯模型示例。(a) 一个高斯分布（灰色曲线）拟合由三个高斯分布（蓝色、绿色和橙色曲线）生成的数据。(b) 由三个子成分组成的混合高斯模型（红色曲线）拟合相同的数据。

最后需要注意的是，高斯分布对少数异常值比较敏感。如图2.14所示，在数据集中仅增加三个异常值之后，采用高斯分布拟合所得到的曲线（蓝色）发生了显著变化，明显受到异常值的影响而出现了强烈的扭曲。为了增强对少数异常值的鲁棒性，可以选择使用**学生氏分布**（Student's t-Distribution）。通过对比图2.14(a) 和 (b) 中的绿色曲线可以看出，学生氏分布的拟合结果几乎不受异常值的影响。接下来，我们将讨论什么是学生氏分布，以及为何它相较于高斯分布对于异常值具有更强的鲁棒性。

之前我们讨论高斯分布均值 μ 已知，而精度 λ 未知的情形时，采用伽玛分布 $\text{Gamma}(\lambda|\alpha,\beta)$ 作为似然函数的共轭先验。假设随机变量 x 服从高斯分布 $\mathcal{N}(x|\mu,\lambda^{-1})$，并且精度 λ 的先验分布为 $\text{Gamma}(\lambda|\alpha,\beta)$。通过对精度 λ 进行积分（即考虑 λ 的所有可能取值及概率），可以得到如下关于随机变量 x 的边缘分布：

$$\begin{aligned} p(x|\mu,\alpha,\beta) &= \int_0^\infty \mathcal{N}(x|\mu,\lambda^{-1})\text{Gamma}(\lambda|\alpha,\beta)\, d\lambda \\ &= \int_0^\infty \left(\frac{\lambda}{2\pi}\right)^{1/2} \exp\left(-\frac{\lambda}{2}(x-\mu)^2\right) \frac{\beta^\alpha \lambda^{\alpha-1} e^{-\beta\lambda}}{\Gamma(\alpha)}\, d\lambda \\ &= \frac{\beta^\alpha}{\Gamma(\alpha)} \left(\frac{1}{2\pi}\right)^{1/2} \left(\beta + \frac{(x-\mu)^2}{2}\right)^{-\alpha-1/2} \Gamma(\alpha+1/2) \end{aligned}$$

将以上最后一步中的变量替换为 $\eta = 2\alpha$ 和 $\tau = \alpha/\beta$，就可以得到学生氏分布的概率密度函数：

$$\text{St}(x|\mu,\tau,\eta) = \frac{\Gamma(\eta/2+1/2)}{\Gamma(\eta/2)} \left(\frac{\tau}{\pi\eta}\right)^{1/2} \left(1 + \frac{\tau(x-\mu)^2}{\eta}\right)^{-(\eta/2+1/2)} \tag{2.119}$$

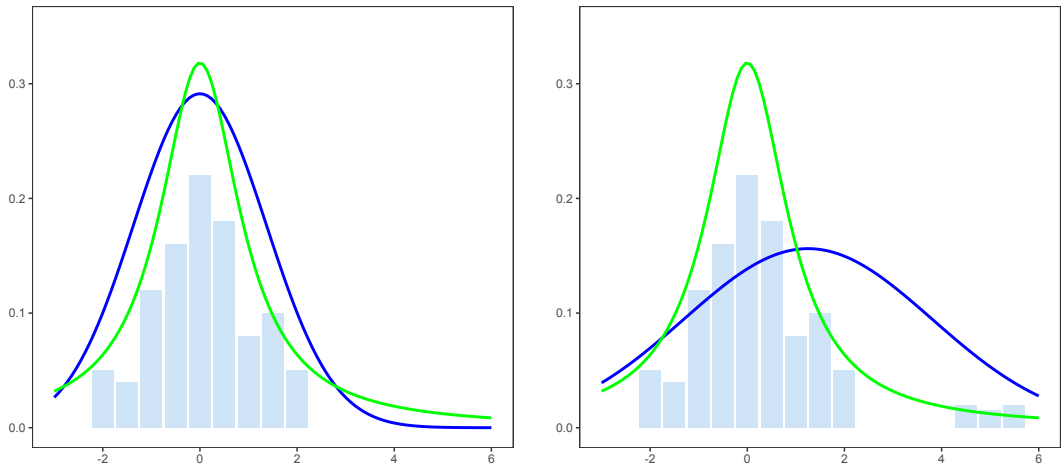


图 2.14: 学生氏分布较高斯分布更具鲁棒性的示意图。(a) 使用高斯分布（蓝色曲线）和学生氏分布（绿色曲线）分别拟合柱状图所示的数据集。(b) 在数据集中增加了三个异常值（位于右侧）的拟合结果。高斯分布（蓝色曲线）受到异常值的显著影响，导致其曲线形态发生了较大变化，而学生氏分布（绿色曲线）几乎不受影响（自由度为 1）。

其中参数 τ 称为精度，而参数 η 为分布的自由度（Degrees of Freedom）。自由度对分布的影响如图2.15(a) 所示。当自由度 $\eta = 1$ 时，学生氏分布退化为柯西分布（Cauchy Distribution）；而当自由度趋近于无穷大 $\eta \to \infty$ 时，则学生氏分布 $\mathrm{St}(x|\mu, \tau, \eta)$ 退化为均值为 μ、精度为 τ 的高斯分布 $\mathcal{N}(x|\mu, \tau^{-1})$。一维学生氏分布的均值（严格应称为位置参数）和方差分别为：

$$\mathbb{E}[x] = \mu \qquad \text{当} \ \ \eta > 1 \tag{2.120}$$

$$\mathrm{Var}[x] = \frac{\eta}{\eta - 2}\tau^{-1} \qquad \text{当} \ \ \eta > 2 \tag{2.121}$$

从推导出学生氏分布的过程可以看出，它是由具有相同均值但不同精度的无穷多个高斯分布叠加所得到的。这可以理解为在混合高斯模型中混合了无限多个高斯分布，从而产生了一个比高斯分布具有更重“长尾”（Heavy Tail）的分布。如图2.15(a) 所示，自由度为 1 或 0.1 的学生氏分布的两端相较于高斯分布（即自由度趋近无穷大的学生氏分布）分配了更多的概率密度。在实际应用中，异常值（出现在低概率区域的数据点）通常由于数据标注错误或者数据本身具有重尾特性而产生，因而具备更重“长尾”的学生氏分布在面对异常值时较高斯分布更加鲁棒。

多维学生氏分布的概率密度函数的形式如下：

$$\mathrm{St}(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Lambda}, \eta) = \frac{\Gamma(\eta/2 + D/2)}{\Gamma(\eta/2)} \frac{|\boldsymbol{\Lambda}|^{1/2}}{(\pi\eta)^{D/2}} \left(1 + \frac{(\boldsymbol{x} - \boldsymbol{\mu})^\top \boldsymbol{\Lambda} (\boldsymbol{x} - \boldsymbol{\mu})}{\eta}\right)^{-(\eta/2 + D/2)} \tag{2.122}$$

其中 D 表示 $\boldsymbol{x}$ 的维度，$\boldsymbol{\mu}$ 为均值向量，$\boldsymbol{\Lambda}$ 是精度矩阵，η 代表自由度。多维学生氏分布的均值和协方差分别为：

$$\mathbb{E}[\boldsymbol{x}] = \boldsymbol{\mu} \qquad \text{当} \ \ \eta > 1 \tag{2.123}$$

$$\mathrm{Var}[\boldsymbol{x}] = \frac{\eta}{\eta - 2}|\boldsymbol{\Lambda}|^{-1} \qquad \text{当} \ \ \eta > 2 \tag{2.124}$$

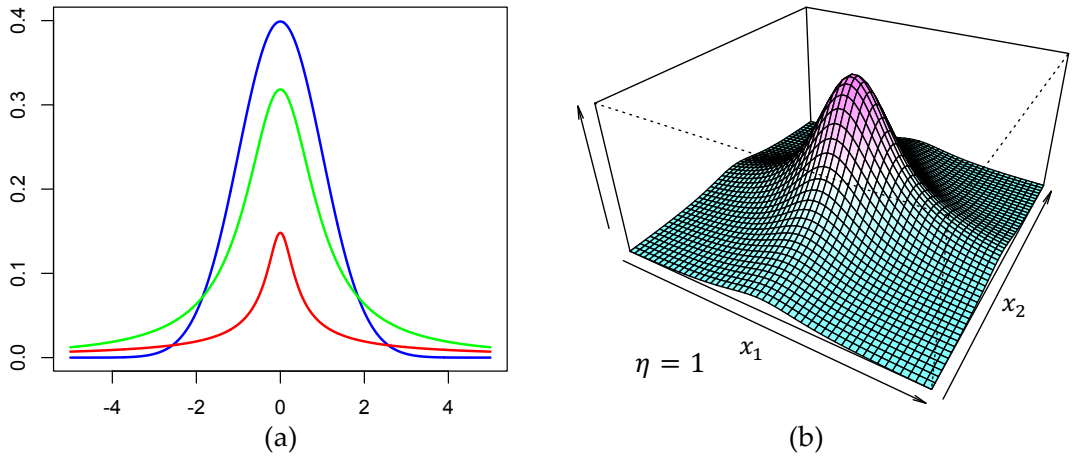


图 2.15: 学生氏分布的示例。(a) 一维学生氏分布。当自由度趋近于无穷大时（$\eta \to \infty$），退化为高斯分布。(b) 二维学生氏分布。

自由度为 1 的二维学生氏分布如图2.15(b) 所示。

2.6 指数族分布

本章中所讨论的离散或连续的概率分布（除了高斯混合模型和学生氏分布外）都是**指数族分布**（Exponential Family of Distributions）的特例。指数族分布的成员可以表示成统一的形式，并且具有许多重要的共同特性。对于指数族分布，采用最大似然方法进行参数估计时，具有统一且简便的表达式；而使用贝叶斯方法时，其似然函数必然存在相应的共轭先验。

随机变量 $\boldsymbol{x}$ 的指数族分布可表示为以下形式：

$$p(\boldsymbol{x}|\boldsymbol{\eta}) = h(\boldsymbol{x})g(\boldsymbol{\eta})\exp(\boldsymbol{\eta}^\top \boldsymbol{t}(\boldsymbol{x})) \tag{2.125}$$

其中 $\boldsymbol{x}$ 可以是标量或向量，它既可以是离散型随机变量，也可以是连续型随机变量。参数 $\boldsymbol{\eta}$ 被称为自然参数（Natural Parameter），它是分布原参数 $\boldsymbol{\theta}$ 的某种变换 $\boldsymbol{\eta}(\boldsymbol{\theta})$。例如，对于伯努利分布和二项分布的原参数 μ，在表示为指数族分布的统一形式之后，其自然参数为 $\eta = \ln(\mu/(1-\mu))$。再如，对于泊松分布的原参数 λ，其自然参数为 $\eta = \ln\lambda$。函数 $g(\boldsymbol{\eta})$ 可以理解为确保分布归一化所增加的系数项，即需要满足以下条件：

$$g(\boldsymbol{\eta}) \int h(\boldsymbol{x})\exp(\boldsymbol{\eta}^\top \boldsymbol{t}(\boldsymbol{x}))\, d\boldsymbol{x} = 1 \tag{2.126}$$

如果上式中的 $\boldsymbol{x}$ 是离散型随机变量，则用求和替换积分。$\boldsymbol{t}(\boldsymbol{x})$ 为**充分统计量**（Sufficient Statistic），它是从样本中计算得出的，包含所有关于分布参数的统计信息。假设我们利用样本对分布的参数进行估计，如果有一个统计量能够提取出样本中所有与参数相关的信息，而不会丢失任何重要信息（增加其他样本信息将变得冗余），那么这个统计量就是充分统计量。充分统计量的优点在于能够将大量样本数据压缩为一个统计量，无需继续存储样本数据，同时保留与参数相关的所有信息。从公式2.125可以看出，指数族分布的表示形式由三部分组成：仅与随机变量 $\boldsymbol{x}$ 相关

的 $h(\boldsymbol{x})$，仅与自然参数 $\boldsymbol{\eta}$ 相关的 $g(\boldsymbol{\eta})$，以及指数为自然参数 $\boldsymbol{\eta}$ 与充分统计量 $\boldsymbol{t}(\boldsymbol{x})$ 点积的部分。

我们以之前章节中介绍过的一些分布为例来说明它们是指数族分布的成员。首先来讨论伯努利分布，我们知道伯努利分布的概率函数为：

$$\text{Bern}(x|\mu) = \mu^x(1-\mu)^{1-x}$$

对上式右侧取对数后，再进行指数变换，我们有：

$$\begin{aligned}\text{Bern}(x|\mu) &= \exp(x\ln\mu + (1-x)\ln(1-\mu)) \\ &= \exp(x\ln\mu + \ln(1-\mu) - x\ln(1-\mu)) \\ &= (1-\mu)\exp\left(\ln\left(\frac{\mu}{1-\mu}\right)x\right)\end{aligned}$$

将上式与公式2.125对比之后可知：

$$\eta = \ln\left(\frac{\mu}{1-\mu}\right)$$

事件发生的概率 μ 与其不发生的概率 $1-\mu$ 之比称为几率，对该比值取对数后则称为**对数几率**（Log Odds）。从上式中解出 μ（即将 η 映射回 μ），可以得到以下对数几率的反函数：

$$\sigma(\eta) = \frac{1}{1+\exp(-\eta)} \tag{2.127}$$

该函数被称为 **Sigmoid 函数**。借助此函数，伯努利分布可以表示为以下指数族分布的标准形式：

$$p(x|\eta) = \sigma(-\eta)\exp(\eta x) \tag{2.128}$$

其中：

$$\begin{aligned}h(x) &= 1 \\ g(\eta) &= \sigma(-\eta) \\ t(x) &= x\end{aligned}$$

接着，我们讨论多项分布，其概率函数为：

$$\text{Mult}(\boldsymbol{m}|\boldsymbol{\mu}, N) = \begin{pmatrix} N \\ m_1 \dots m_K \end{pmatrix} \prod_{i=1}^{K} \mu_i^{m_i} = \frac{N!}{\prod_{i=1}^{K} m_i!} \exp\left(\sum_{i=1}^{K} m_i \ln\mu_i\right)$$

将上式与公式2.125进行对比，并且将 $\boldsymbol{\eta}$ 表示为 $(\ln\mu_1, \dots, \ln\mu_K)^\top$，则多项分布可以表示为以下指数族分布的标准形式：

$$\text{Mult}(\boldsymbol{m}|\boldsymbol{\mu}, N) = \frac{N!}{\prod_{i=1}^{K} m_i!} \exp\left(\boldsymbol{\eta}^\top \boldsymbol{x}\right)$$

其中：

$$\begin{aligned}h(\boldsymbol{x}) &= \frac{N!}{\prod_{i=1}^{K} m_i!} \\ g(\boldsymbol{\eta}) &= 1 \\ t(\boldsymbol{x}) &= (m_1, \dots, m_K)^\top\end{aligned}$$

值得注意的是，这种表示中的自然参数并非相互独立，而是需要满足以下约束：

$$\sum_{i=1}^{K} e^{\eta_i} = 1 \tag{2.129}$$

因此，上述形式可以看作一种受约束的指数族表示，主要用于直观地说明多项分布与指数族分布之间的联系。严格来说，为了得到自然参数之间不存在约束的指数族标准形式，应采用 $K-1$ 维自然参数，因为多项分布只有 $K-1$ 个自由参数。令：

$$\eta_i = \ln \frac{\mu_i}{\mu_K}, \quad i = 1, \ldots, K-1 \tag{2.130}$$

相应地，通过反向变换，可以将 μ_i 表示为：

$$\mu_i = \frac{e^{\eta_i}}{1 + \sum_{j=1}^{K-1} e^{\eta_j}}, \quad i = 1, \ldots, K-1 \tag{2.131}$$

$$\mu_K = \frac{1}{1 + \sum_{j=1}^{K-1} e^{\eta_j}} \tag{2.132}$$

此时，在多项分布的指数族标准形式中，$g(\boldsymbol{\eta})$ 和 $t(\boldsymbol{x})$ 分别为：

$$g(\boldsymbol{\eta}) = \left(1 + \sum_{i=1}^{K-1} e^{\eta_i}\right)^{-N} \tag{2.133}$$

$$t(\boldsymbol{x}) = (m_1, \ldots, m_{K-1})^\top \tag{2.134}$$

最后，我们来讨论高斯分布。对于一维高斯分布，我们有：

$$\begin{aligned}\mathcal{N}(x|\mu, \sigma^2) &= \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right) \\ &= \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}x^2 + \frac{\mu}{\sigma^2}x - \frac{1}{2\sigma^2}\mu^2\right)\end{aligned}$$

经过简单的变换，可以将上式转化为指数族分布的标准表示形式，其中：

$$\boldsymbol{\eta} = \begin{bmatrix} \frac{\mu}{\sigma^2} \\ -\frac{1}{2\sigma^2} \end{bmatrix}$$

$$h(\boldsymbol{x}) = \frac{1}{\sqrt{2\pi}}$$

$$g(\boldsymbol{\eta}) = \sqrt{-2\eta_2} \exp\left(\frac{\eta_1^2}{4\eta_2}\right)$$

$$t(\boldsymbol{x}) = \begin{bmatrix} x \\ x^2 \end{bmatrix}$$

表2.2列出了本章所讨论的属于指数族分布的各类离散或连续分布的标准表示形式。

由于指数族分布具有如公式2.125所示的标准表示形式，因此可以借助该标准形式来探讨使用最大似然方法对参数 $\boldsymbol{\eta}$ 进行估计时的通解形式。同时将公式2.126两边对参数 $\boldsymbol{\eta}$ 求导，我们有：

$$\nabla g(\boldsymbol{\eta}) \int h(\boldsymbol{x}) \exp(\boldsymbol{\eta}^\top \boldsymbol{t}(\boldsymbol{x}))\, d\boldsymbol{x} + g(\boldsymbol{\eta}) \int h(\boldsymbol{x}) \exp(\boldsymbol{\eta}^\top \boldsymbol{t}(\boldsymbol{x})) \boldsymbol{t}(\boldsymbol{x})\, d\boldsymbol{x} = 0$$

上式可以改写为：

$$\frac{\nabla g(\boldsymbol{\eta})}{g(\boldsymbol{\eta})} \underbrace{g(\boldsymbol{\eta}) \int h(\boldsymbol{x}) \exp(\boldsymbol{\eta}^\top \boldsymbol{t}(\boldsymbol{x}))\, d\boldsymbol{x}}_{=1\text{（公式2.126）}} + g(\boldsymbol{\eta}) \int h(\boldsymbol{x}) \exp(\boldsymbol{\eta}^\top \boldsymbol{t}(\boldsymbol{x})) \boldsymbol{t}(\boldsymbol{x})\, d\boldsymbol{x} = 0$$

经简单整理，我们可以得到：

$$-\frac{\nabla g(\boldsymbol{\eta})}{g(\boldsymbol{\eta})} = g(\boldsymbol{\eta}) \int h(\boldsymbol{x}) \exp(\boldsymbol{\eta}^\top \boldsymbol{t}(\boldsymbol{x})) \boldsymbol{t}(\boldsymbol{x})\, d\boldsymbol{x} = \mathbb{E}[\boldsymbol{t}(\boldsymbol{x})]$$

由于对函数 $\ln(x)$ 求导等于 $1/x$，上式最终可以写成：

$$-\nabla \ln g(\boldsymbol{\eta}) = \mathbb{E}[\boldsymbol{t}(\boldsymbol{x})] \tag{2.135}$$

表 2.2: 指数族分布

分布	原参数 $\boldsymbol{\theta}$	自然参数 $\boldsymbol{\eta}$	$\boldsymbol{\theta}(\boldsymbol{\eta})$	$h(\boldsymbol{x})$	$\boldsymbol{t}(\boldsymbol{x})$	$g(\boldsymbol{\eta})$
伯努利分布	μ	$\ln\left(\frac{\mu}{1-\mu}\right)$	$\frac{1}{1+\exp(-\eta)}$	1	x	$\frac{1}{1+\exp(\eta)}$
二项分布	μ	$\ln\left(\frac{\mu}{1-\mu}\right)$	$\frac{1}{1+\exp(-\eta)}$	$\begin{pmatrix} N \\ k \end{pmatrix}$	k	$\frac{1}{(1+\exp(\eta))^N}$
泊松分布	λ	$\ln\lambda$	$\exp(\eta)$	$\frac{1}{k!}$	k	$\exp(-\exp(\eta))$
指数分布	β	$-\beta$	$-\eta$	1	λ	$-\eta$
多项分布	$\mu_1, \ldots, \mu_{K-1}$	$\begin{bmatrix} \ln\frac{\mu_1}{\mu_K} \\ \vdots \\ \ln\frac{\mu_{K-1}}{\mu_K} \end{bmatrix}$	$\begin{bmatrix} \exp(\eta_1) \\ \vdots \\ \exp(\eta_{K-1}) \end{bmatrix}$	$\frac{N!}{\prod_{i=1}^K m_i!}$	$\begin{bmatrix} m_1 \\ \vdots \\ m_{K-1} \end{bmatrix}$	$\left(1+\sum_{i=1}^{K-1} e^{\eta_i}\right)^{-N}$
伽玛分布	α, β	$\begin{bmatrix} \alpha-1 \\ -\beta \end{bmatrix}$	$\begin{bmatrix} \eta_1+1 \\ -\eta_2 \end{bmatrix}$	1	$\begin{bmatrix} \ln\lambda \\ \lambda \end{bmatrix}$	$\frac{(-\eta_2)^{\eta_1+1}}{\Gamma(\eta_1+1)}$
贝塔分布	α, β	$\begin{bmatrix} \alpha \\ \beta \end{bmatrix}$	$\begin{bmatrix} \eta_1 \\ \eta_2 \end{bmatrix}$	$\frac{1}{\mu(1-\mu)}$	$\begin{bmatrix} \ln\mu \\ \ln(1-\mu) \end{bmatrix}$	$\frac{\Gamma(\eta_1+\eta_2)}{\Gamma(\eta_1)\Gamma(\eta_2)}$
一维高斯分布	μ, σ^2	$\begin{bmatrix} \frac{\mu}{\sigma^2} \\ -\frac{1}{2\sigma^2} \end{bmatrix}$	$\begin{bmatrix} -\frac{\eta_1}{2\eta_2} \\ -\frac{1}{2\eta_2} \end{bmatrix}$	$\frac{1}{\sqrt{2\pi}}$	$\begin{bmatrix} x \\ x^2 \end{bmatrix}$	$\sqrt{-2\eta_2}\exp\left(\frac{\eta_1^2}{4\eta_2}\right)$
多维高斯分布	$\boldsymbol{\mu}, \boldsymbol{\Sigma}$	$\begin{bmatrix} \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} \\ -\frac{1}{2}\boldsymbol{\Sigma}^{-1} \end{bmatrix}$	$\begin{bmatrix} -\frac{1}{2}\boldsymbol{\eta}_2^{-1}\boldsymbol{\eta}_1 \\ -\frac{1}{2}\boldsymbol{\eta}_2^{-1} \end{bmatrix}$	$(2\pi)^{-\frac{D}{2}}$	$\begin{bmatrix} \boldsymbol{x} \\ \boldsymbol{x}\boldsymbol{x}^\top \end{bmatrix}$	$\sqrt{\lvert -2\boldsymbol{\eta}_2\rvert}\exp\left(\frac{1}{4}\boldsymbol{\eta}_1^\top\boldsymbol{\eta}_2^{-1}\boldsymbol{\eta}_1\right)$
狄利克雷分布	$\alpha_1, \ldots, \alpha_K$	$\begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_K \end{bmatrix}$	$\begin{bmatrix} \eta_1 \\ \vdots \\ \eta_K \end{bmatrix}$	$\frac{1}{\prod_{i=1}^K \mu_i}$	$\begin{bmatrix} \ln\mu_1 \\ \vdots \\ \ln\mu_K \end{bmatrix}$	$\Gamma\left(\sum_{i=1}^K \eta_i\right)\prod_{i=1}^K \frac{1}{\Gamma(\eta_i)}$
威沙特分布	$\mathbf{W}, \varphi$	$\begin{bmatrix} -\frac{1}{2}\mathbf{W}^{-1} \\ \frac{\varphi-D-1}{2} \end{bmatrix}$	$\begin{bmatrix} -\frac{1}{2}\boldsymbol{\eta}_1^{-1} \\ 2\eta_2+D+1 \end{bmatrix}$	1	$\begin{bmatrix} \boldsymbol{\Lambda} \\ \ln\lvert\boldsymbol{\Lambda}\rvert \end{bmatrix}$	$\lvert-\boldsymbol{\eta}_1\rvert^{\frac{\varphi}{2}}\left(\Gamma_D\left(\frac{\varphi}{2}\right)\right)^{-1}$
高斯-伽玛分布	$\upsilon, \nu, \alpha, \beta$	$\begin{bmatrix} \nu\upsilon \\ -\frac{\nu}{2} \\ \alpha-\frac{1}{2} \\ -\beta-\frac{\nu\upsilon^2}{2} \end{bmatrix}$	$\begin{bmatrix} -\frac{\eta_1}{2\eta_2} \\ -2\eta_2 \\ \frac{1}{2}+\eta_3 \\ -\eta_4+\frac{\eta_1^2}{4\eta_2} \end{bmatrix}$	$\frac{1}{\sqrt{2\pi}}$	$\begin{bmatrix} \lambda\mu \\ \lambda\mu^2 \\ \ln\lambda \\ \lambda \end{bmatrix}$	$\sqrt{-2\eta_2}\left(\Gamma\left(\eta_3+\frac{1}{2}\right)\right)^{-1}$ $\left(-\eta_4+\frac{\eta_1^2}{4\eta_2}\right)^{\eta_3+\frac{1}{2}}$

注：随机变量 $\boldsymbol{x}$ 的符号视具体分布而定（表格2.3）。为了避免混淆，威沙特分布的原参数 η 改用 φ 表示。

同样地，对公式2.126求二阶导数，我们可以得到：

$$-\nabla^2 \ln g(\boldsymbol{\eta}) = \mathbb{E}[\boldsymbol{t}(\boldsymbol{x})\boldsymbol{t}(\boldsymbol{x})^\top] - \mathbb{E}[\boldsymbol{t}(\boldsymbol{x})]\mathbb{E}[\boldsymbol{t}(\boldsymbol{x})^\top] = \mathrm{Cov}[\boldsymbol{t}(\boldsymbol{x})] \tag{2.136}$$

从公式2.135和2.136可以看出，只要能对指数族分布进行归一化，对数配分函数（即 $-\ln g(\boldsymbol{\eta})$）的一阶导数给出充分统计量的期望，二阶导数得到其协方差，而更高阶导数给出更高阶累积量。这些统计量可使用样本数据进行估算，进而得到分布参数 $\boldsymbol{\eta}$ 的估计值。

表 2.3: 分布似然函数的共轭先验

似然函数		参数	共轭先验	
伯努利分布	$\mathrm{Bern}(x\|\mu)$	μ	贝塔分布	$\mathrm{Beta}(\mu\|\alpha,\beta)$
二项分布	$\mathrm{Bin}(k\|\mu,N)$	μ	贝塔分布	$\mathrm{Beta}(\mu\|\alpha,\beta)$
泊松分布	$\mathrm{Poisson}(k\|\lambda)$	λ	伽玛分布	$\mathrm{Gamma}(\lambda\|\alpha,\beta)$
指数分布	$\mathrm{Exp}(\lambda\|\beta)$	β	伽玛分布	$\mathrm{Gamma}(\beta\|\alpha_0,\beta_0)$
多项分布	$\mathrm{Mult}(\boldsymbol{m}\|\boldsymbol{\mu},N)$	$\boldsymbol{\mu}$	狄利克雷分布	$\mathrm{Dir}(\boldsymbol{\mu}\|\boldsymbol{\alpha})$
一维高斯分布	$\mathcal{N}(x\|\mu,\sigma^2)$	μ	一维高斯分布	$\mathcal{N}(\mu\|\mu_0,\sigma_0^2)$
一维高斯分布	$\mathcal{N}(x\|\mu,\lambda^{-1})$	λ	伽玛分布	$\mathrm{Gamma}(\lambda\|\alpha,\beta)$
一维高斯分布	$\mathcal{N}(x\|\mu,\lambda^{-1})$	μ,λ	高斯-伽玛分布	$\mathrm{NormalGamma}(\mu,\lambda\|\upsilon,\nu,\alpha,\beta)$
多维高斯分布	$\mathcal{N}(\boldsymbol{x}\|\boldsymbol{\mu},\boldsymbol{\Sigma})$	$\boldsymbol{\mu}$	多维高斯分布	$\mathcal{N}(\boldsymbol{\mu}\|\boldsymbol{\mu}_0,\boldsymbol{\Sigma}_0)$
多维高斯分布	$\mathcal{N}(\boldsymbol{x}\|\boldsymbol{\mu},\boldsymbol{\Lambda}^{-1})$	$\boldsymbol{\Lambda}$	威沙特分布	$\mathcal{W}(\boldsymbol{\Lambda}\|\mathbf{W},\eta)$
多维高斯分布	$\mathcal{N}(\boldsymbol{x}\|\boldsymbol{\mu},\boldsymbol{\Lambda}^{-1})$	$\boldsymbol{\mu},\boldsymbol{\Lambda}$	高斯-威沙特分布	$\mathrm{NormalWishart}(\boldsymbol{\mu},\boldsymbol{\Lambda}\|\boldsymbol{\upsilon},\nu,\mathbf{W},\eta)$

在使用贝叶斯方法对指数族分布的参数 $\boldsymbol{\eta}$ 进行估计时，必然存在如下形式的共轭先验：

$$p(\boldsymbol{\eta}|\boldsymbol{\chi},\nu) = f(\boldsymbol{\chi},\nu)g(\boldsymbol{\eta})^\nu \exp(\nu\boldsymbol{\eta}^\top\boldsymbol{\chi}) \tag{2.137}$$

其中 $f(\boldsymbol{\chi},\nu)$ 是归一化系数，ν 可以理解为在实际观测之前先验分布中假设的伪观测次数，$\boldsymbol{\chi}$ 是这些伪观测的充分统计量（相当于公式2.125中的 $\boldsymbol{t}(\boldsymbol{x})$），而 $g(\boldsymbol{\eta})$ 的含义则与公式2.125中出现的 $g(\boldsymbol{\eta})$ 相同。假设给定数据集 $\mathcal{D} = \{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_N\}$，指数族分布的似然函数为：

$$p(\mathcal{D}|\boldsymbol{\eta}) = \left(\prod_{i=1}^N h(\boldsymbol{x}_i)\right) g(\boldsymbol{\eta})^N \exp\left(\boldsymbol{\eta}^\top \sum_{i=1}^N \boldsymbol{t}(\boldsymbol{x}_i)\right) \tag{2.138}$$

将似然函数（公式2.138）与先验分布（公式2.137）结合，我们可以得到：

$$\begin{aligned}
p(\boldsymbol{\eta}|\mathcal{D},\boldsymbol{\chi},\nu) &\propto p(\mathcal{D}|\boldsymbol{\eta})p(\boldsymbol{\eta}|\boldsymbol{\chi},\nu) \\
&= \left(\prod_{i=1}^N h(\boldsymbol{x}_i)\right) g(\boldsymbol{\eta})^N \exp\left(\boldsymbol{\eta}^\top \sum_{i=1}^N \boldsymbol{t}(\boldsymbol{x}_i)\right) f(\boldsymbol{\chi},\nu)g(\boldsymbol{\eta})^\nu \exp(\nu\boldsymbol{\eta}^\top\boldsymbol{\chi}) \\
&\propto g(\boldsymbol{\eta})^N \exp\left(\boldsymbol{\eta}^\top \sum_{i=1}^N \boldsymbol{t}(\boldsymbol{x}_i)\right) g(\boldsymbol{\eta})^\nu \exp(\nu\boldsymbol{\eta}^\top\boldsymbol{\chi}) \text{（除去与 } \boldsymbol{\eta} \text{ 无关的项）} \\
&\propto g(\boldsymbol{\eta})^{\nu+N} \exp\left(\boldsymbol{\eta}^\top \left(\nu\boldsymbol{\chi} + \sum_{i=1}^N \boldsymbol{t}(\boldsymbol{x}_i)\right)\right)
\end{aligned}$$

从上述推导的最后一步可以看出，指数族分布的后验分布具有以下形式：

$$p(\boldsymbol{\eta}|\mathcal{D},\boldsymbol{\chi},\nu)=p\left(\boldsymbol{\eta}\middle|\frac{\nu\boldsymbol{\chi}+\sum_{i=1}^{N}\boldsymbol{t}(\boldsymbol{x}_i)}{\nu+N},\nu+N\right) \tag{2.139}$$

该后验分布显然与先验分布（公式2.137）共轭。

表2.3中列出了常用的共轭先验及其对应的似然函数。我们需要明确先验分布与参数所描述的概率分布之间的关系。先验分布是针对参数的似然函数而言的，而并非对参数所描述或确定的概率分布而言的。当一个概率分布涉及多个参数时，每个参数都可能有其对应的不同的先验分布。也就是说，只有在确定分布以及分布的某个或某些参数的情况下，我们才能确定其共轭先验的形式（如果存在）。

主要符号表

数学符号

$\boldsymbol{x}$	向量
y	标量
D	向量维度
x_j	向量的第 j 个分量（向量分量为标量）
f	输入到输出的某种映射
$\mathbf{X}^\top$	矩阵转置
$\mathbf{X}^{-1}$	逆矩阵
$\|\mathbf{X}\|$	矩阵的行列式
$\mathrm{tr}(\mathbf{X})$	矩阵的迹（矩阵所有特征值之和，也等于主对角线元素的总和）
$\ln$	以自然常数 e 为底的对数
e^x 或 $\exp(x)$	自然常数 e 的指数函数
$f'(x)$	一阶导数
$f''(x)$	二阶导数
$\nabla f(\boldsymbol{x})$	一阶梯度（梯度向量）
$\nabla^2 f(\boldsymbol{x})$	二阶梯度（海森矩阵）
$\frac{\partial f(x)}{\partial x}$	一阶偏导
$\frac{\partial^2 f(x)}{\partial x^2}$	二阶偏导

机器学习

$\mathcal{D}$	数据集
N	样本数量
$\boldsymbol{x}_i$	第 i 个训练样本
y_i	第 i 个训练样本对应的输出
$\kappa(\boldsymbol{x}, \boldsymbol{x}')$	核函数
$\mathcal{L}(\cdot)$	损失函数
$\mathcal{J}(\cdot)$	目标函数或代价函数

概率统计

$\mathbb{E}[x]$	均值或期望
$\text{Var}[x]$	方差
$\text{Cov}[x, z]$	随机变量 x 和 z 之间的协方差
$\Gamma(z)$	伽玛函数
$\mathcal{N}(x\|\mu, \sigma^2)$	高斯分布，其中 μ 为均值，σ 为标准差
$\mathcal{N}(x\|\mu, \lambda^{-1})$	高斯分布，其中 μ 为均值，λ 为精度
$\text{Bern}(x\|\mu)$	伯努利分布
$\text{Bin}(k\|\mu, N)$	二项分布
$\text{Beta}(\mu\|\alpha, \beta)$	贝塔分布
$\text{Poisson}(k\|\lambda)$	泊松分布
$\text{Exp}(\lambda\|\beta)$	指数分布
$\text{Gamma}(\lambda\|\alpha, \beta)$	伽玛分布，其中 α 为形状参数，β 为逆尺度参数
$\text{Mult}(\boldsymbol{m}\|\boldsymbol{\mu}, N)$	多项分布
$\text{Dir}(\boldsymbol{\mu}\|\boldsymbol{\alpha})$	狄利克雷分布
$\mathcal{N}(\boldsymbol{x}\|\boldsymbol{\mu}, \boldsymbol{\Sigma})$	多维高斯分布，其中 $\boldsymbol{\mu}$ 为均值向量，$\boldsymbol{\Sigma}$ 为协方差矩阵
$\mathcal{N}(\boldsymbol{x}\|\boldsymbol{\mu}, \boldsymbol{\Lambda})$	多维高斯分布，其中 $\boldsymbol{\mu}$ 为均值向量，$\boldsymbol{\Lambda}$ 为精度矩阵
$\mathcal{W}(\boldsymbol{\Lambda}\|\mathbf{W}, \eta)$	威沙特分布，其中 $\mathbf{W}$ 为尺度矩阵，η 为自由度
$\text{NormalGamma}(\mu, \lambda\|\upsilon, \nu, \alpha, \beta)$	高斯-伽玛分布
$\text{NormalWishart}(\boldsymbol{\mu}, \boldsymbol{\Lambda}\|\boldsymbol{\upsilon}, \nu, \mathbf{W}, \eta)$	高斯-威沙特分布
$\text{St}(x\|\mu, \tau, \eta)$	学生氏分布，其中 μ 为均值，τ 为精度，η 为自由度
$\text{St}(\boldsymbol{x}\|\boldsymbol{\mu}, \boldsymbol{\Lambda}, \eta)$	多维学生氏分布，其中 $\boldsymbol{\mu}$ 为均值向量，$\boldsymbol{\Lambda}$ 为精度矩阵
$\text{Laplace}(x\|\mu, \tau)$	拉普拉斯分布，其中 μ 为位置参数，τ 为尺度参数
$\chi^2(x\|\eta)$	卡方分布，其中 η 为自由度
$\text{U}(x\|a, b)$	连续型均匀分布，其中 a 为下界、b 为上界 $(a < b)$
$\text{H}[x]$	信息熵，其中 x 为随机变量
$\text{H}[y\|x]$	条件熵，其中 x 和 y 为随机变量
$\text{H}[p, q]$	分布 p 和 q 之间的交叉熵
$D_{\text{KL}}(p\|\|q)$	分布 p 和 q 之间的 KL 散度

参考文献

[1] Daniel J. Amit, Hanoch Gutfreund, and Haim Sompolinsky. “Storing infinite numbers of patterns in a spin-glass model of neural networks”. In: *Physical Review Letters* 55.14 (1985), pp. 1530–1533.

[2] Mikel Artetxe et al. “On the role of bidirectionality in language model pre-training”. In: *Findings of the Conference on Empirical Methods in Natural Language Processing* (2022), pp. 3973–3985.

[3] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. “Layer normalization”. In: *arXiv: 1607.06450* (2016).

[4] Yoshua Bengio, Patrice Simard, and Paolo Frasconi. “Learning long-term dependencies with gradient descent is difficult”. In: *IEEE Transactions on Neural Networks* 5.2 (1994), pp. 157–166.

[5] Jame O. Berger. *Statistical decision theory and Bayesian analysis (second edition)*. Springer, 1980.

[6] James Bergstra and Yoshua Bengio. “Random search for hyper-parameter optimization”. In: *Journal of Machine Learning Research* 13.2 (2012).

[7] Jeremy Bernstein and Laker Newhouse. “Old optimizer, new norm: An anthology”. In: *Proceedings of the 16th Annual Workshop on Optimization for Machine Learning* (2024), pp. 1–19.

[8] J Besag. “On spatio-temporal models and Markov fields”. In: *Transactions of the 7th Prague Conference on Information Theory, Statistical Decision Functions and Random Processes* (1974), pp. 47–75.

[9] Julian Besag. “On the statistical analysis of dirty pictures”. In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 48.3 (1986), pp. 259–279.

[10] Christopher M. Bishop. *Pattern recognition and machine learning*. Springer, 2006.

[11] Bernhard E. Boser, Isabelle M. Guyon, and Vladimir N. Vapnik. “A training algorithm for optimal margin classifiers”. In: *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*. 1992, pp. 144–152.

[12] George E.P. Box and Norman Richard Draper. *Empirical model-building and response surfaces*. John Wiley & Sons, 1987.

[13] Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge University Press, 2004.

[14] Tom Brown et al. “Language models are few-shot learners”. In: *Proceedings of the Conference on Neural Information Processing Systems* 33 (2020), pp. 1877–1901.

[15] Kyunghyun Cho et al. “Learning phrase representations using RNN encoder–decoder for statistical machine translation”. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing* (2014), pp. 1724–1734.

[16] Peter Clifford. “Markov random fields in statistics”. In: *Disorder in Physical Systems: A Volume in Honour of John M. Hammersley* (1990), pp. 19–32.

[17] Thomas H. Cormen et al. *Introduction to algorithms (third edition)*. The MIT Press, 2022.

[18] Thomas M. Cover and Joy A. Thomas. *Elements of information theory*. Wiley, 1991.

[19] George Cybenko. “Approximation by superpositions of a sigmoidal function”. In: *Mathematics of control, signals and systems* 2.4 (1989), pp. 303–314.

[20] Morris H. DeGroot and Mark J. Schevish. *Probability and statistics (forth edition)*. China Machine Press, 2012.

[21] Jacob Devlin et al. “BERT: Pre-training of deep bidirectional transformers for language understanding”. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (2019), pp. 4171–4186.

[22] Alexey Dosovitskiy et al. “An image is worth 16×16 words: Transformers for image recognition at scale”. In: *Proceedings of the International Conference on Learning Representations* (2021).

[23] John Duchi, Elad Hazan, and Yoram Singer. “Adaptive subgradient methods for online learning and stochastic optimization”. In: *Journal of Machine Learning Research* 12.7 (2011).

[24] Walter D. Fisher. “On grouping for maximum homogeneity”. In: *Journal of the American Statistical Association* 53.284 (1958), pp. 789–798.

[25] Roger Fletcher. *Practical methods of optimization (second edition)*. Wiley, 1987.

[26] Yoav Freund and Robert E. Schapire. “A decision-theoretic generalization of on-line learning and an application to boosting”. In: *Journal of Computer and System Sciences* 55.1 (1997), pp. 119–139.

[27] Stuart Geman and Donald Geman. “Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images”. In: *IEEE Transactions on pattern analysis and machine intelligence* 6 (1984), pp. 721–741.

[28] Felix A. Gers, Nicol N. Schraudolph, and Jürgen Schmidhuber. “Learning precise timing with LSTM recurrent networks”. In: *Journal of Machine Learning Research* 3 (2002), pp. 115–143.

[29] Xavier Glorot and Yoshua Bengio. “Understanding the difficulty of training deep feedforward neural networks”. In: *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics* (2010), pp. 249–256.

[30] Ian J Goodfellow et al. “Generative adversarial nets”. In: *Proceedings of the Conference on Neural Information Processing Systems* 27 (2014).

[31] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning: data mining, inference, and prediction (second edition)*. Springer, 2009.

[32] Kaiming He and Jian Sun. “Convolutional neural networks at constrained time cost”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2015), pp. 5353–5360.

[33] Kaiming He et al. “Deep residual learning for image recognition”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2016), pp. 770–778.

[34] Kaiming He et al. “Delving deep into rectifiers: Surpassing human-level performance on imagenet classification”. In: *Proceedings of the IEEE International Conference on Computer Vision* (2015), pp. 1026–1034.

[35] Nicholas J. Higham. *Functions of matrices: Theory and computation*. SIAM, 2008.

[36] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. “Distilling the knowledge in a neural network”. In: *arXiv: 1503.02531* (2015).

[37] Geoffrey E. Hinton. “Training products of experts by minimizing contrastive divergence”. In: *Neural Computation* 14.8 (2002), pp. 1771–1800.

[38] John J. Hopfield. “Neural networks and physical systems with emergent collective computational abilities”. In: *Proceedings of the National Academy of Sciences* 79.8 (1982), pp. 2554–2558.

[39] Sergey Ioffe and Christian Szegedy. “Batch normalization: Accelerating deep network training by reducing internal covariate shift”. In: *Proceedings of the International Conference on Machine Learning* (2015), pp. 448–456.

[40] Arieh Iserles. *A first course in the numerical analysis of differential equations*. 44. Cambridge university press, 2009.

[41] Max Jaderberg et al. “Population based training of neural networks”. In: *arXiv: 1711.09846* (2017).

[42] Kevin Jamieson and Ameet Talwalkar. “Non-stochastic best arm identification and hyperparameter optimization”. In: *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics* (2016), pp. 240–248.

[43] Keller Jordan et al. *Muon: An optimizer for hidden layers in neural networks*. 2024. URL: https://kellerjordan.github.io/posts/muon/.

[44] Diederik P. Kingma and Jimmy L. Ba. “Adam: A method for stochastic optimization”. In: *Proceedings of the International Conference on Learning Representations* (2015).

[45] Josef Kittler and Janos Föglein. “Contextual classification of multispectral pixel data”. In: *Image and vision computing* 2.1 (1984), pp. 13–29.

[46] Solomon Kullback and Richard A. Leibler. “On information and sufficiency”. In: *The Annals of Mathematical Statistics* 22.1 (1951), pp. 79–86.

[47] Yann LeCun et al. “Gradient-based learning applied to document recognition”. In: *Proceedings of the IEEE* 86.11 (2002), pp. 2278–2324.

[48] Mike Lewis et al. “BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension”. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (2020), pp. 7871–7880.

[49] Yaron Lipman et al. “Flow matching guide and code”. In: *arXiv: 2412.06264* (2024).

[50] Ilya Loshchilov and Frank Hutter. “Decoupled weight decay regularization”. In: *arXiv: 1711.05101* (2017).

[51] Calvin Luo. “Understanding diffusion models: A unified perspective”. In: *arXiv: 2208.11970* (2022).

[52] Marvin Minsky and Seymour Papert. *Perceptrons: An introduction to computational geometry*. The MIT Press, 1969.

[53] Kevin P. Murphy. *Machine learning: A probabilistic perspective*. The MIT Press, 2012.

[54] Brady Neal et al. “A modern take on the bias-variance tradeoff in neural networks”. In: *arXiv: 1810.08591* (2018).

[55] Long Ouyang et al. “Training language models to follow instructions with human feedback”. In: *Proceedings of the Conference on Neural Information Processing Systems* 35 (2022), pp. 27730–27744.

[56] Judea Pearl. *Probabilistic reasoning in intelligent systems*. Morgan kaufmann, 2014.

[57] Matthew E. Peters et al. “Deep Contextualized Word Representations”. In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (2018), pp. 2227–2237.

[58] John C. Platt. “Fast training of support vector machines using sequential minimal optimization”. In: *Advances in Kernel Methods* (1999), pp. 185–208.

[59] Alec Radford, Luke Metz, and Soumith Chintala. “Unsupervised representation learning with deep convolutional generative adversarial networks”. In: *Proceedings of the International Conference on Learning Representations* (2016).

[60] Alec Radford et al. “Improving language understanding by generative pre-training”. In: *OpenAI’s Technical Report* (2018).

[61] Colin Raffel et al. “Exploring the limits of transfer learning with a unified text-to-text transformer”. In: *Journal of Machine Learning Research* 21.140 (2020), pp. 1–67.

[62] Scott Reed et al. “Generative adversarial text to image synthesis”. In: *Proceedings of the International Conference on Machine Learning* (2016), pp. 1060–1069.

[63] Edmund T. Rolls. *Cerebral cortex: principles of operation*. Oxford University Press, 2016.

[64] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. “U-Net: Convolutional networks for biomedical image segmentation”. In: *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention* (2015), pp. 234–241.

[65] David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. “Learning representations by back-propagating errors”. In: *Nature* 323.6088 (1986), pp. 533–536.

[66] Leonard J. Savage. *The subjective basis of statistical practice*. University of Michigan, 1961.

[67] Andrew M Saxe, James L McClelland, and Surya Ganguli. “Exact solutions to the nonlinear dynamics of learning in deep linear neural networks”. In: *Proceedings of the International Conference on Learning Representations* (2014).

[68] Robert E. Schapire. “The strength of weak learnability”. In: *Machine learning* 5.2 (1990), pp. 197–227.

[69] Rico Sennrich, Barry Haddow, and Alexandra Birch. “Neural machine translation of rare words with subword units”. In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics* (2016), pp. 1715–1725.

[70] Claude Elwood Shannon. “A mathematical theory of communication”. In: *The Bell System Technical Journal* 27.3 (1948), pp. 379–423.

[71] Jasper Snoek, Hugo Larochelle, and Ryan P Adams. “Practical Bayesian optimization of machine learning algorithms”. In: *Proceedings of the Conference on Neural Information Processing Systems* 25 (2012).

[72] Yang Song et al. “Score-based generative modeling through stochastic differential equations”. In: *Proceedings of the International Conference on Learning Representations* (2021).

[73] Gilbert Strang. *Introduction to linear algebra*. Wellesley-Cambridge Press, 2016.

[74] Jianlin Su et al. “Roformer: Enhanced transformer with rotary position embedding”. In: *Neurocomputing* 568 (2024), p. 127063.

[75] Tijmen Tieleman and Geoffrey E. Hinton. “Lecture 6.5-RMSprop: Divide the gradient by a running average of its recent magnitude”. In: *COURSERA: Neural Networks for Machine Learning* 4.2 (2012), pp. 26–26.

[76] Antonio Torralba, Phillip Isola, and William Freeman. *Foundations of computer vision*. The MIT Press, 2024.

[77] Ashish Vaswani et al. “Attention is all you need”. In: *Proceedings of the Conference on Neural Information Processing Systems* 30 (2017).

[78] Xiaoqing Zheng, Hanyang Chen, and Tianyu Xu. “Deep learning for Chinese word segmentation and POS tagging”. In: *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing* (2013), pp. 647–657.

[79] Ji Zhu et al. “Multi-class AdaBoost”. In: *Statistics and its Interface* 2.3 (2009), pp. 349–360.

[80] Hui Zou and Trevor Hastie. “Regularization and variable selection via the elastic net”. In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 67.2 (2005), pp. 301–320.