Continuous Variables

构建新概率分布

定义（可测函数和随机变量）

- 设 $(\Omega, \mathfrak{F})$ 和 $(\Gamma, \mathfrak{G})$ 是两个可测空间（即具有 σ-代数 $\mathfrak{F}$ 和 $\mathfrak{G}$ 的空间）。一个函数 $X : \Omega \to \Gamma$ 称为可测的，如果对于所有 $G \in \mathfrak{G}$，有 $X^{-1}(G) \in \mathfrak{F}$。
- 如果在 $(\Omega, \mathfrak{F})$ 上还有一个概率测度 P，那么 X 被称为随机变量。

解释：可测空间是由一组可能结果（如样本空间 Ω）和一个 σ-代数（如事件集合 $\mathfrak{F}$）组成，σ-代数定义了哪些事件可以被测量。函数 X 从样本空间 Ω 映射到目标空间 Γ，如果它能将目标空间中的可测集合 G 映射回样本空间中的可测集合 $X^{-1}(G)$，则称为可测的。当样本空间上有概率测度 P 时，X 成为随机变量，表示具有概率分布的变量。

定义（分布测度）

- 设 $X : \Omega \to \Gamma$ 是一个随机变量。那么 X 的分布测度（或法则）P_X 对于任何 $G \subset \Gamma$ 定义为：

$$P_X(G) = P(X^{-1}(G)) = P(\{\omega | X(\omega) \in G\}) \tag{1}$$

解释：分布测度 P_X 描述了随机变量 X 取值的概率分布。对于目标空间 Γ 中的任何子集 G，$P_X(G)$ 是 X 落在 G 中的概率。这通过概率测度 P 和逆映射 $X^{-1}(G)$（即所有使得 $X(\omega) \in G$ 的 ω 集合）来定义。换句话说，$P_X(G)$ 告诉我们随机变量 X 落在某个集合 G 的概率是多少。

总结

- 随机变量是可测函数，当样本空间上有概率测度时，它可以描述变量的概率分布。
- 分布测度 P_X 是随机变量 X 的概率分布的数学表示，用于量化 X 取值在特定区域的概率。

解释：这些定义为连续随机变量的概率建模奠定了基础。通过可测性和概率测度，随机变量和其分布测度可以用于描述复杂系统中的不确定性，特别是在机器学习和统计建模中。

二项分布（Binomial Distribution）

（累积伯努利实验的统计）

1. 问题描述与背景

- 有一枚不公平硬币，出现正面（heads）的概率为 f。该硬币被抛掷 N 次。问：正面次数 r 的概率分布是什么？

解释：这个例子模拟了 N 次独立伯努利实验（每次抛硬币），目标是确定在这些实验中出现正好 r 次正面的概率。二项分布是描述这种重复独立试验中成功次数的离散概率分布，广泛用于统计学和机器学习。

2. 模型与随机变量

- 每个抛掷的结果用随机变量 X_i 表示：
  $X_i := \begin{cases} 1 & \text{if } i\text{-th toss is heads} \\ 0 & \text{else} \end{cases}$
- 总正面次数 R 定义为所有 X_i 的和：
  $R := \sum_{i=1}^{N} X_i$
- R 是一个随机变量，取值范围为 $[0, \ldots, N]$，即 $\Gamma \subset \mathbb{N}$。

解释：X_i 是伯努利随机变量，表示第 i 次抛掷的结果（1 为正面，0 为反面）。R 是这些独立实验的累积和，反映了总共出现正面的次数。R 的取值是离散的，从 0（无正面）到 N（全为正面）。

3. 概率空间与结构

- **原始空间**：$\Omega = \{0, 1\}^N$（可数有限集合，表示每次抛掷的结果组合）。
- **σ-代数**：2^Ω（幂集，所有可能的子集）。
- **随机变量 R**：$R = \sum_{i=1}^{N} X_i \in [0, \ldots, N]$，取值范围是自然数子集 $\Gamma \subset \mathbb{N}$。
- **分布（测度）/法则**：R 的分布是二项分布。

解释：这些定义构成了概率空间的数学框架。Ω 包含所有可能的抛掷结果组合，σ-代数确保我们能定义概率，R 的分布测度描述了其取值的概率分布。

4. 二项分布的概率公式

- $R = r$ 的概率（即正面次数为 r 的概率）可以通过所有可能的抛掷组合计算：
  $P(R = r) = \sum_{\omega \in \{X | R = r\}} \prod_{i=1}^{N} P(X_i) = \sum_{\omega \in \{X | R = r\}} f^r \cdot (1 - f)^{N - r} := P(r | f, N)$
- 更简洁的表达为：
  $P(r | f, N) = \binom{N}{r} \cdot f^r \cdot (1 - f)^{N - r}$
  其中，$\binom{N}{r} = \frac{N!}{(N - r)! \cdot r!}$ 是组合数（选择 r 个正面的方式）。

解释：这个公式计算了在 N 次抛掷中出现正好 r 次正面的概率。$f^r \cdot (1 - f)^{N - r}$ 是每个特定序列的概率（r 次正面，$N - r$ 次反面），$\binom{N}{r}$ 考虑了所有可能出现 r 次正面的组合方式。最终结果 $P(r | f, N)$ 是二项分布的概率质量函数（PMF），依赖于正面概率 f 和抛掷次数 N。

5. 分布测度与可视化

- 二项分布的分布测度 P_R（或法则）描述了随机变量 R 的概率分布。对于给定的 f 和 N，$P(r | f, N)$ 给出了 $R = r$ 的概率。

解释：分布测度 P_R 是随机变量 R 的概率分布的数学表示。图示展示了具体参数（如 $f = 1/3$、$N = 10$）下二项分布的形状，反映了概率随 r 变化的趋势。这类分布通常呈现钟形或偏态，取决于 f 和 N 的值。

6. 符号约定与注意事项

- **符号简化**：课程中常简化为 $P(r)$ 代替 $P(R = r)$，但需要注意 $P(X) \neq P(Y)$（不同随机变量的概率分布可能不同）。
- **常见应用**：这种符号简化在概率建模中很常见，便于表达，但需明确上下文以避免歧义。

解释：这种符号缩写是为了简化表达，但在严格的数学推导中，应注意区分不同随机变量的分布。课程后续可能会频繁使用这种简写，需要理解其含义。

7. 总结

- 二项分布是描述重复独立伯努利实验中成功次数的离散概率分布，公式为 $P(r|f,N)=\binom{N}{r}\cdot f^r\cdot(1-f)^{N-r}$。
- 它通过累积 N 次实验的随机变量 R 构建，适用于统计学和机器学习中的计数问题（如成功/失败的次数）。

解释：二项分布是概率论和统计学中的核心工具，通过硬币抛掷的例子直观展示了其定义、计算和应用。理解其公式和分布特性有助于分析类似随机实验的结果。

连续空间的处理

连续空间与可测性

- 在可数空间 Ω 中，2^Ω 是一个 σ-代数（所有子集都可以被测量）。
- 但在连续空间中，如 $\Omega=\mathbb{R}^d$（欧几里得空间），并非所有集合都是可测的。

解释：可数空间（如离散集合）中的所有子集都可以形成 σ-代数，因此易于定义概率。但连续空间（如实数空间）更复杂，因为无限多的点和集合可能导致某些集合不可测（无法分配概率），需要特殊结构（如 σ-代数）来限制可测集合。

拓扑空间的引入

- 然而，$\mathbb{R}^d$ 是一个拓扑空间（topological space），需要用拓扑结构来定义可测集合。

解释：拓扑空间是一种数学结构，用于描述空间中的“近邻”或“连续性”关系。引入拓扑是为了在连续空间中定义“开集”（open sets），从而构建 σ-代数（如 Borel σ-代数），确保概率测度可以应用。

拓扑的定义（Topology）

- 设 Ω 是一个空间，τ 是一组集合。我们称 τ 是 Ω 上的拓扑，如果满足以下条件：
  - $\Omega\in\tau$，且 $\emptyset\in\tau$（整个空间和空集是开集）。
  - 如果 $A_1,A_2,\cdots\in\tau$，则 $\bigcup_{i=1}^{\infty}A_i\in\tau$（任意并集仍为开集）。
  - 如果 $A_1,A_2,\ldots,A_n\in\tau$（有限个），则 $\bigcap_{i=1}^{n}A_i\in\tau$（有限交集仍为开集）。

解释：拓扑定义了一组“开集”，这些开集满足特定的规则：包含整个空间和空集，任意并集和有限交集仍然是开集。这为连续空间提供了结构，用于定义距离、连续性和可测性。

欧几里得空间中的典型拓扑

- 在欧几里得向量空间 $\mathbb{R}^d$ 中，典型的拓扑（规范拓扑）是由所有满足以下条件的集合 U 组成：
  对于任意 $x\in U$，存在 $\epsilon>0$，使得 $\|y-x\|<\epsilon\implies y\in U$（即 U 是以 x 为中心的开球）。

解释：在 $\mathbb{R}^d$ 中，开集通常是围绕某个点的开球（例如，以某个点为中心，半径为 ϵ 的区域）。这个拓扑定义了空间中的连续性和邻近关系，是构建可测空间（如 Borel σ-代数）的基础。

注意事项

- **符号说明**：拓扑中的“任意并集”可能包括不可数个集合，但“有限交集”限制为有限个集合（可能符号表达不够严谨）。

- **复杂性**：连续空间的复杂性源于无限多点和不可测集合的存在，因此需要拓扑和 σ-代数来处理概率建模。

总结

- 连续空间（如 $\mathbb{R}^d$）不像可数空间那样简单，所有集合都不是可测的。
- 通过引入拓扑空间和开集的概念，可以在连续空间中定义 σ-代数，从而支持概率测度的定义和随机变量的建模。

解释：这个内容为处理连续随机变量奠定了基础。通过拓扑和可测性理论，解决连续空间中概率建模的复杂性，为后续的连续变量分布（如高斯分布）提供了数学框架。

定义（Borel 代数）

让 (Ω, τ) 是一个拓扑空间。Borel σ-代数是由 τ 生成的 σ-代数，也就是通过以下方式构造：

- 取 τ 中的所有元素（即开集），
- 加上 τ 中元素的无限交（infinite intersections），
- 以及在 Ω 中的所有补集（complements）。

解释：
Borel σ-代数是拓扑空间中一个非常重要的概念，它为定义概率空间提供了基础。在拓扑空间 (Ω, τ) 中，τ 是一个开集的集合（拓扑结构）。Borel σ-代数通过扩展 τ，包括所有开集、它们的无限交以及在整个空间 Ω 中的补集，从而形成一个足够丰富的集合体系，用于定义测度（如概率测度）。

在本次讲座中，我们几乎完全关注在离散空间或欧几里得空间（Euclidean spaces）上定义的（随机）变量。对于后者情况，σ-代数不会被特别提及，但默认假设它是 Borel σ-代数。

解释：

- 离散空间和欧几里得空间（如 $\mathbb{R}^n$）是概率论和机器学习中常见的空间。
- 在欧几里得空间中，Borel σ-代数是标准的选择，因为它包含所有开集、闭集以及它们的布尔运算结果（如并、交、补），从而适合定义连续随机变量的概率分布。

连续函数与可测性

考虑 $(\Omega, \mathfrak{F})$ 和 $(\Gamma, \mathfrak{G})$，如果 $\mathfrak{F}$ 和 $\mathfrak{G}$ 都是 Borel σ-代数，那么任何连续函数 X 都是可测的（measurable），因此可以用来定义随机变量。这是因为，对于连续函数，其预图像（pre-images） of 开集是开集。

解释：

- 可测性（measurability）是随机变量的一个核心属性，它确保我们可以对随机变量定义概率。
- 如果 $(\Omega, \mathfrak{F})$ 和 $(\Gamma, \mathfrak{G})$ 都是 Borel σ-代数，且 $X : \Omega \to \Gamma$ 是一个连续函数，那么对于 Γ 中的任意 Borel 集 B，$X^{-1}(B)$（即 X 的预图像）在 Ω 中也是 Borel 集。因此，X 是可测的，可以作为随机变量的基础。
- 连续函数的可测性依赖于这样一个事实：连续函数会“映射”开集到开集，这与 Borel σ-代数的定义一致。

后续内容

现在我们可以在连续空间上定义（Borel）σ-代数，从而定义概率分布测度。这些测度可能有点复杂（unwieldy）。

解释：

- 在连续空间（如 $\mathbb{R}$ 或 $\mathbb{R}^n$）上，Borel σ-代数允许我们定义概率分布，例如高斯分布或其他连续分布。
- 然而，这些测度的数学表达和计算可能较为复杂，因此被称为“有点复杂”。

定义（概率密度函数 (PDFs)）

让 $\mathfrak{B}$ 是在 $\mathbb{R}^d$ 上的 Borel σ-代数。如果概率测度 P 在 $(\mathbb{R}^d, \mathfrak{B})$ 上有一个密度 p，那么 p 是一个非负的（Borel）可测函数，满足对于所有 $B \in \mathfrak{B}$ 有：

$$P(B) = \int_B p(x)\,dx =: \int_B p(x_1, \ldots, x_d)\,dx_1 \ldots dx_d \tag{2}$$

密度 $P(B)$ 可以写成对 B 的积分，对于所有的 B。

解释：

- 概率密度函数（PDF，Probability Density Function）是描述连续随机变量概率分布的一种方式，特别适用于 $\mathbb{R}^d$（d 维欧几里得空间）上的连续分布。
- $\mathfrak{B}$ 是 $\mathbb{R}^d$ 上的 Borel σ-代数，它为概率测度的定义提供了数学基础。
- 如果概率测度 P 有一个密度 p，那么 p 必须是非负的（$p(x) \geq 0$），并且是 Borel 可测的（即在 Borel σ-代数下可测）。
- 对于任意 Borel 集 B（例如一个区间或更复杂的区域），$P(B)$（即 B 内的概率）可以通过对密度函数 $p(x)$ 在 B 上的积分来计算。积分的形式在单变量 ($d = 1$) 时是 $\int_B p(x)\,dx$，而在多变量 ($d > 1$) 时是 $\int_B p(x_1, \ldots, x_d)\,dx_1 \ldots dx_d$。
- 这种表示方法非常方便，特别是对于连续分布（如高斯分布），但并非所有概率测度都有密度函数。例如，含有点质量的测度（如狄拉克测度）没有密度，因为它们的概率集中在单个点上，无法用连续的密度函数表示。

定义（累积分布函数 (CDF)）

对于在 $(\mathbb{R}^d, \mathfrak{B})$ 上的概率测度 P，累积分布函数（CDF）是函数：

$$F(\mathbf{x}) = P\left(\bigcap_{i=1}^{d} (X_i < x_i)\right). \tag{3}$$

特别地，对于单变量情况 ($d = 1$)，我们有 $F(x) = P((-\infty, x])$。

如果 F 足够可微（sufficiently differentiable），那么 P 有一个密度，給定为：

$$p(\mathbf{x}) = \left.\frac{\partial^d F}{\partial x_1 \cdots \partial x_d}\right|_{\mathbf{x}}. \tag{4}$$

对于区间 $a \leq x < b$，有 $P(a \leq x < b) = F(b) - F(a) = \int_a^b f(x)\,dx$。

解释：

- **累积分布函数（CDF）** 是描述随机变量概率分布的另一种重要工具，特别是在连续变量的情况下。它表示随机变量 X 取值小于或等于某个值 $\mathbf{x}$ 的概率。
- 对于 d 维随机向量 $\mathbf{X} = (X_1, \ldots, X_d)$，CDF 定义为 $F(\mathbf{x}) = P(X_1 < x_1, \ldots, X_d < x_d)$，即所有分量 X_i 同时小于对应值 x_i 的概率。这是多变量情况下的定义。

- 在单变量 ($d = 1$) 情况下，CDF 简化为 $F(x) = P(X \leq x)$，表示随机变量 X 小于或等于 x 的概率，通常写作 $P((-\infty, x])$。
- 如果 CDF F 足够可微（例如，在单变量情况下是连续且可导的），那么概率测度 P 会有一个对应的概率密度函数（PDF）$p(\mathbf{x})$。对于 d 维情况，密度函数通过对 F 进行 d 次偏导数得到，公式为 $p(\mathbf{x}) = \frac{\partial^d F}{\partial x_1 \cdots \partial x_d}$，其中偏导是在点 $\mathbf{x}$ 处计算。
- 在单变量情况下，如果 $F(x)$ 可导，密度函数 $p(x) = F'(x)$，并且对于任意区间 $[a, b)$，概率 $P(a \leq x < b)$ 可以用 CDF 的差表示：$F(b) - F(a)$。根据概率密度函数的定义，这也等于 $\int_a^b p(x)\,dx$，即密度函数在区间上的积分。
- 这个关系表明，CDF 和 PDF 之间通过积分和微分相互联系。CDF 提供了累积的概率信息，而 PDF 提供了概率的“密度”或分布的细致描述。

概率密度满足概率论的法则

对于在 $(\mathbb{R}^d, \mathfrak{B})$ 上的概率密度 p，我们有：

$$P(E) \equiv 1 = \int_{\mathbb{R}^d} p(x)\,dx. \tag{5}$$

让 $X = (X_1, X_2) \in \mathbb{R}^2$ 是一个随机变量，密度为 p_X 在 $\mathbb{R}^2$ 上。那么，X_1 和 X_2 的边缘密度（marginal densities）由求和法则给出：

$$p_{X_1}(x_1) = \int_{\mathbb{R}} p_X(x_1, x_2)\,dx_2, \quad p_{X_2}(x_2) = \int_{\mathbb{R}} p_X(x_1, x_2)\,dx_1. \tag{6}$$

条件密度 $p(x_1|x_2)$（对于 $p(x_2) > 0$）由乘积法则给出：

$$p(x_1|x_2) = \frac{p(x_1, x_2)}{p(x_2)}. \tag{7}$$

此外，联合密度 $p(x_1, x_2)$ 可以通过条件密度表示为：

$$p(x_1|x_2) = \frac{p(x_1) \cdot p(x_2|x_1)}{\int p(x_1) \cdot p(x_2|x_1)\,dx_1}. \tag{8}$$

这也就是连续变量的贝叶斯公式。

解释：

- **概率密度的规范化**：概率密度函数 $p(x)$ 必须满足总概率为 1 的条件，即在整个空间 $\mathbb{R}^d$ 上的积分等于 1。这是因为概率测度的总和（或总体积）必须为 1，公式 $P(E) \equiv 1 = \int_{\mathbb{R}^d} p(x)\,dx$ 反映了这一点。积分是线性的，这使得密度函数能够遵循概率论的基本法则。
- **边缘密度（Marginal Densities）**：对于二维随机变量 $X = (X_1, X_2)$，其联合密度为 $p_X(x_1, x_2)$。边缘密度 $p_{X_1}(x_1)$ 是通过对 x_2 积分得到的，公式为 $p_{X_1}(x_1) = \int_{\mathbb{R}} p_X(x_1, x_2)\,dx_2$。类似地，$p_{X_2}(x_2) = \int_{\mathbb{R}} p_X(x_1, x_2)\,dx_1$。这反映了概率论中的“求和法则”或“边缘化”过程，用于从联合分布中提取单个变量的分布。
- **条件密度（Conditional Densities）**：条件密度 $p(x_1|x_2)$ 表示在给定 $X_2 = x_2$ 的条件下 $X_1 = x_1$ 的密度。对于 $p(x_2) > 0$，条件密度通过联合密度 $p(x_1, x_2)$ 除以边缘密度 $p(x_2)$ 计算，公式为 $p(x_1|x_2) = \frac{p(x_1, x_2)}{p(x_2)}$。这是概率论中的“乘积法则”或贝叶斯公式的基础。

- **联合密度的表示**：幻灯片最后给出了联合密度 $p(x_1, x_2)$ 的另一种形式，通过条件密度和边缘密度结合表示为 $p(x_1|x_2) = \frac{p(x_1)\cdot p(x_2|x_1)}{\int p(x_1)\cdot p(x_2|x_1)\, dx_1}$。这表明联合密度可以通过边缘密度和条件密度通过积分和标准化得到，体现了概率分布之间的关系。

概率密度函数的变量变换定理

双变量情况

让 X 是一个连续随机变量，概率密度函数（PDF）为 $p_X(x)$，定义域为 $c_1 < x < c_2$。让 $Y = u(X)$ 是一个单调可微函数，存在逆函数 $X = v(Y)$。那么 Y 的概率密度函数为：

$$p_Y(y) = p_X(v(y)) \cdot \left| \frac{dv(y)}{dy} \right| = p_X(v(y)) \cdot \left| \frac{du(x)}{dx} \right|^{-1}. \tag{9}$$

证明： 假设 $u'(x) > 0$，令 $d_1 = u(c_1) < y < u(c_2) = d_2$：

$$F_Y(y) = P(Y \leq y) = P(u(X) \leq y) = P(X \leq v(y)) = \int_{c_1}^{v(y)} p_X(x)\, dx. \tag{10}$$

因此，密度函数为：

$$p_Y(y) = \frac{dF_Y(y)}{dy} = p_X(v(y)) \cdot \frac{dv(y)}{dy} = p_X(v(y)) \cdot \left| \frac{dv(y)}{dy} \right|. \tag{11}$$

解释：

- **变量变换定理**：这个定理描述了当随机变量通过一个单调可微函数 $u(X)$ 变换时，其概率密度函数如何变化。X 是一个连续随机变量，具有密度 $p_X(x)$，定义域为 $c_1 < x < c_2$。$Y = u(X)$ 是 X 的单调可微变换，意味着 $u(x)$ 必须是单增或单减且可微，以便存在逆函数 $v(y)$，满足 $X = v(Y)$。
- **密度变换公式**：Y 的密度 $p_Y(y)$ 通过 X 的密度 $p_X(v(y))$ 和雅可比行列式的绝对值（即 $\left| \frac{dv(y)}{dy} \right|$ 或 $\left| \frac{du(x)}{dx} \right|^{-1}$）计算。雅可比行列式反映了变换如何“拉伸”或“压缩”概率密度。
  - 如果 $u(x)$ 是单增函数（$u'(x) > 0$），则 $\frac{dv(y)}{dy} = \frac{1}{u'(x)}$，公式简化为 $p_Y(y) = p_X(v(y)) \cdot \frac{1}{|u'(x)|}$。
  - 如果 $u(x)$ 是单减函数（$u'(x) < 0$），则需要取绝对值 $\left| \frac{dv(y)}{dy} \right|$，确保密度非负。
- **证明过程**：通过累积分布函数（CDF）$F_Y(y)$ 推导密度。$F_Y(y)$ 是 $Y \leq y$ 的概率，等价于 $u(X) \leq y$，再通过逆函数变为 $X \leq v(y)$。对 $F_Y(y)$ 求导得到密度 $p_Y(y)$，利用链式法则和单调性的性质，得出变换公式。
- 这个定理在概率论和统计学中非常重要，例如在随机变量变换、蒙特卡洛方法和统计推断中，用于计算变换后变量的分布。

一般多变量情况

让 $X = (X_1, \ldots, X_d)$ 有一个联合密度 p_X。让 $g : \mathbb{R}^d \to \mathbb{R}^d$ 是一个连续可微且满射（injective）的函数，具有非零雅可比行列式 J_g。那么，$Y = g(X)$ 的密度为：

$$p_Y(y) = \begin{cases} p_X(g^{-1}(y)) \cdot |J_{g^{-1}}(y)| & \text{if } y \text{ is in the range of } g, \\ 0 & \text{otherwise.} \end{cases} \tag{12}$$

雅可比行列式 J_g 是 $d \times d$ 矩阵，其中：

$$[J_g]_{ij} = \frac{\partial g_i(x)}{\partial x_j}. \tag{13}$$

解释：

- 这个定理推广了单变量情况到多变量情况。X 是一个 d 维随机向量，具有联合密度 p_X。变换 $Y = g(X)$ 是一个从 $\mathbb{R}^d$ 到 $\mathbb{R}^d$ 的连续可微且满射函数，意味着 g 是双射（一对一映射），且其雅可比行列式 J_g 非零（非奇异），以确保变换是可逆的。∏
- Y 的密度 $p_Y(y)$ 依赖于 X 的密度在逆变换 $g^{-1}(y)$ 处的值 $p_X(g^{-1}(y))$，并乘以逆变换 g^{-1} 的雅可比行列式绝对值 $|J_{g^{-1}}(y)|$。雅可比行列式 J_g 是 g 各分量 $g_i(x)$ 关于 x_j 的偏导数形成的矩阵，其行列式的绝对值反映了变换如何改变体积或密度。
- 如果 y 不在 g 的值域（range of g）内，$p_Y(y) = 0$，因为概率集中在 g 的映射范围内。
- 与单变量情况相比，多变量情况需要考虑更高维的雅可比矩阵，公式更复杂，但核心思想一致：通过变换后的密度调整雅可比行列式来保持概率测度的不变性。

联系与对比

- **共同点**：两者都基于变量变换的概率密度函数变化，核心思想是通过逆函数和雅可比行列式调整密度以保持概率的总和为 1。两者都假设变换是可微且可逆（单调可微或满射）。
- **区别**：前者处理单变量 ($d = 1$) 的简单情况，只需一阶导数（$\frac{du(x)}{dx}$ 或 $\frac{dv(y)}{dy}$）；后者处理多变量 ($d \geq 1$) 的情形，需要雅可比矩阵 J_g 和其行列式的绝对值。单变量情况是多变量情况的特殊情形，当 $d = 1$ 时，雅可比行列式简化为一阶导数的绝对值。
- **应用**：这两个定理在概率论和统计学中用于随机变量变换，单变量变换常用于简单分布（如线性变换），而多变量变换适用于更复杂的几何变换（如坐标变换或非线性映射）。

完整分析逻辑：推断戴眼镜概率的概率机器学习模型

1. 问题背景与目标

我们希望推断某人戴眼镜的概率，即确定一个概率值 π，表示该人戴眼镜的固有概率。假设我们有 5 次独立观察（observations），每次观察记录为二值变量 x_i（$x_i = 1$ 表示戴眼镜，$x_i = 0$ 表示不戴眼镜），用随机变量 X_1, X_2, X_3, X_4, X_5 表示。我们需要基于这些观察结果，更新对 π 的信念（从先验到后验）。

2. 构造 σ-代数：定义随机变量与图形模型（Step 1）

2.1 随机变量的定义

- 戴眼镜的概率用一个连续随机变量 Y 表示，Y 取值在区间 $[0, 1]$，我们用 π 简记，即 $Y = \pi$。
- 5 个观察结果用二值随机变量 X_1, X_2, X_3, X_4, X_5 表示，其中每个 X_i 取值 $\{0, 1\}$（0 表示不戴眼镜，1 表示戴眼镜）。

2.2 图形表示

- 使用概率图模型表示依赖关系：Y（即 π）是父节点，X_1 到 X_5 是子节点。这表明每个观察 X_i 依赖于 π，但 X_i 之间条件独立（给定 π 后，X_i 不直接相互影响）。
- 图形模型的结构为：

```
      Y (π)
↓   ↓   ↓   ↓   ↓
X₁, X₂, X₃, X₄, X₅
```

2.3 生成模型与联合概率

- π 的先验概率分布为 $p(\pi)$，这是一个定义在 $[0,1]$ 上的概率密度函数。
- 每个观察 X_i 的似然函数为 $p(x_i|\pi)$，表示给定 π 下观察到 x_i 的概率。联合概率分布为：

$$p(\pi, x_1, x_2, x_3, x_4, x_5) = p(\pi) \prod_{i=1}^{5} p(x_i|\pi) \tag{14}$$

其中，$p(x_i|\pi)$ 是 π 的函数，反映观察结果与概率之间的关系。

3. 定义概率空间：考虑条件独立性与贝叶斯更新（Step 2）

3.1 初始信念（无观察时的概率）

- 在没有观察结果时，我们仅依赖先验分布：

$$p(\pi|\text{"nothing"}) = p(\pi) \tag{15}$$

3.2 逐步更新后验概率

我们使用贝叶斯定理更新 π 的概率分布，基于观察结果 x_i。后验概率的计算公式为：

$$p(\pi|x) = \frac{p(x|\pi)p(\pi)}{\int p(x|\pi)p(\pi)d\pi} \tag{16}$$

其中，归一化常数 $Z = \int p(x|\pi)p(\pi)d\pi$ 确保后验概率密度函数的总和为 1。

- **一个观察 x_1 后的后验**：

$$p(\pi|x_1) = Z_1^{-1} p(x_1|\pi)p(\pi) \tag{17}$$

其中 $Z_1 = \int p(x_1|\pi)p(\pi)d\pi$。

- **两个观察 x_1, x_2 后的后验**：

由于 X_i 之间条件独立（给定 π 后），$p(x_2|x_1, \pi) = p(x_2|\pi)$，因此：

$$p(\pi|x_1, x_2) = Z_2^{-1} p(x_2|\pi)p(x_1|\pi)p(\pi) \tag{18}$$

其中 $Z_2 = \int p(x_2|\pi)p(x_1|\pi)p(\pi)d\pi$。

- **五个观察 x_1, x_2, x_3, x_4, x_5 后的后验**：

扩展到 5 个观察，结果为：

$$p(\pi|x_1, x_2, x_3, x_4, x_5) = Z_5^{-1} \left(\prod_{i=1}^{5} p(x_i|\pi) \right) p(\pi) \tag{19}$$

其中 $Z_5 = \int \left(\prod_{i=1}^{5} p(x_i|\pi) \right) p(\pi)d\pi$。

3.3 条件独立性的重要性

条件独立性（X_i 在给定 π 下相互独立）简化了计算，使联合似然分解为独立似然的乘积。这使得后验分布的更新更加高效。

4. 定义生成模型的解析形式：似然函数与辅助变量（Step 3）

4.1 似然函数的定义

为了进行具体计算，我们需要定义 $p(x_i|\pi)$ 的解析形式。假设每个 X_i 服从伯努利分布，概率 π 为成功（戴眼镜）的概率，则：

$$p(x_i|\pi) = \begin{cases} \pi & \text{for } x_i = 1 \\ 1-\pi & \text{for } x_i = 0 \end{cases} \tag{20}$$

这表示如果 $x_i = 1$，概率为 π；如果 $x_i = 0$，概率为 $1-\pi$。

4.2 简化后验计算

- 引入辅助随机变量 N 和 M：
- N：观察值为 1 的次数（取值 n）。
- M：观察值为 0 的次数（取值 m）。
- 总观察次数 $n + m = 5$。
- 五个观察后的联合似然为：

$$\prod_{i=1}^{5} p(x_i|\pi) = \pi^n (1-\pi)^m \tag{21}$$

因此，后验概率变为：

$$p(\pi|x_1, x_2, x_3, x_4, x_5) = Z_5^{-1} \pi^n (1-\pi)^m p(\pi) = p(\pi|n, m) \tag{22}$$

其中 Z_5 是归一化常数，确保后验为有效的概率密度函数。

4.3 解释

- 似然函数的形式表明，n 和 m 是关键统计量，分别表示成功（戴眼镜）和失败（不戴眼镜）的次数。
- 后验分布 $p(\pi|n, m)$ 依赖于观察数据（n 和 m）和先验 $p(\pi)$。

5. 选择计算上方便的先验：使用共轭先验（Step 4）

5.1 后验分布的形式

从上一步，我们知道后验分布为：

$$p(\pi|n, m) = Z_5^{-1} \pi^n (1-\pi)^m p(\pi) \tag{23}$$

为了简化计算，我们需要选择一个先验 $p(\pi)$，使其与似然函数的形状匹配，从而保持后验分布的计算简便。

5.2 共轭先验的选择

- 伯努利似然（$p(x_i|\pi) = \pi^{x_i}(1-\pi)^{1-x_i}$）的共轭先验是 Beta 分布。Beta 分布的概率密度函数为：

$$p(\pi) = Z^{-1}\pi^{a-1}(1-\pi)^{b-1} \tag{24}$$

其中 $a > 0$，$b > 0$ 是超参数，$Z = B(a,b)$ 是归一化常数，Beta 函数定义为：

$$B(a,b) = \int_0^1 \pi^{a-1}(1-\pi)^{b-1}d\pi \tag{25}$$

- Beta 分布定义在 $[0,1]$ 上，适合表示概率 π，且其形状由 a 和 b 控制：
- a 和 b 越大，分布越集中。
- $a = b = 1$ 对应均匀分布，表示无信息先验。

5.3 后验分布的简化

将 Beta 分布作为先验，代入后验公式：

- 先验 $p(\pi) = \frac{\pi^{a-1}(1-\pi)^{b-1}}{B(a,b)}$。
- 似然 $\pi^n(1-\pi)^m$。
- 后验为：

$$p(\pi|n,m) = Z_5^{-1}\pi^n(1-\pi)^m \cdot \frac{\pi^{a-1}(1-\pi)^{b-1}}{B(a,b)} \tag{26}$$

化简后：

$$p(\pi|n,m) \propto \pi^{n+a-1}(1-\pi)^{m+b-1} \tag{27}$$

这仍然是一个 Beta 分布，形式为 $Beta(n+a, m+b)$，归一化常数为 $B(n+a, m+b)$：

$$p(\pi|n,m) = \frac{\pi^{n+a-1}(1-\pi)^{m+b-1}}{B(n+a,m+b)} \tag{28}$$

这说明 Beta 分布是伯努利似然的共轭先验，后验保持 Beta 分布的形状，参数更新为 $a+n$ 和 $b+m$。

5.4 优势

- 使用共轭先验简化了计算，因为后验分布的形式与先验相同，避免了复杂的积分。
- 参数 a 和 b 可以解释为“伪计数”：a 相当于成功（戴眼镜）的先验次数，b 相当于失败（不戴眼镜）的先验次数。

6. 总结与应用

- **完整流程**：
  1. 定义问题：推断 π（戴眼镜的概率）。
  2. 构造随机变量：$Y = \pi$（连续，$[0,1]$），X_i（二值，0 或 1）。
  3. 建立图形模型：X_i 条件独立于 π。
  4. 定义似然：$p(x_i|\pi) = \pi^{x_i}(1-\pi)^{1-x_i}$（伯努利分布）。

5. 选择共轭先验：$p(\pi) \sim Beta(a, b)$。
6. 更新后验：观测 n 次 1 和 m 次 0 后，$p(\pi|n, m) \sim Beta(n + a, m + b)$。

- **实际应用**：
  1. 如果有 5 次观察，结果为 $x_1 = 1, x_2 = 0, x_3 = 1, x_4 = 0, x_5 = 1$，则 $n = 3$，$m = 2$。
  2. 假设先验为 $Beta(1, 1)$（均匀分布），后验为 $Beta(3 + 1, 2 + 1) = Beta(4, 3)$。
  3. 后验分布 $p(\pi|n = 3, m = 2)$ 反映了基于数据和先验的最新信念。
- **结果解释**：
  - 后验分布 $Beta(4, 3)$ 表示 π 的概率密度集中于 $[0, 1]$，峰值在 $\pi = \frac{4}{4+3} = 0.57$，表明根据观察和先验，估计戴眼镜的概率约为 57%