指数族分布、广义线性模型

前言

　　这一部分可以说是十分逆天的存在了，我们在以往的模型中，根据各种常见分布推出的回归模型，这一篇文章直接给统一了。是的，也就是标题当中的：指数族分布（但是这玩意推导死麻烦，还需要再去把以往的重新来一遍...）

　　指数族分布（Series family distribution）是一类分布，它包括：高斯分布、伯努利分布、二项分布、泊松分布、Beta 分布、Dirichlet 分布、伽马分布等一系列分布。你可以看到的是，我们的高斯分布（常规线性回归、批量梯度）与伯努利分布（逻辑回归、随机梯度），这些都属于指数族分布与其相关推论的特殊情况！所以我说这一部分是相当逆天的存在（我们在高中所学的那几个分布，实际上并没有深入，这里算是知识闭环了）

　　说了那么多，我们直接来看看怎么个事吧

指数族分布

先放公式：

$$f(y \mid \boldsymbol{\eta}) = h(y) \exp\left(\boldsymbol{\eta}^{\top} \mathbf{T}(y) - A(\boldsymbol{\eta})\right)$$

事实上，我们可以不进行该公式的推导。~~其实是我没看明白他怎么推的~~。

感兴趣可以自行询问（Prompt）： `{指数族分布公式推导，使用Latex公式进行推导}`

这里我的建议是，先熟悉公式的各个组成，我这里先列出AI的解释：

- y 是随机变量（响应变量）。
- η 是自然参数（natural parameter）。
- $T(y)$ 是充分统计量（sufficient statistic）。
- η^T 是将 η 映射到自然参数的函数。
- $A(\eta)$ 是累积量函数（cumulant function）或对数配分函数（log-partition function），确保分布积分为 1。
- $h(y)$ 是基测度（base measure），与参数无关。

　　你会发现AI说的，你是一点都看不懂...这些东西在之前的特殊推理情况中压根没见过。好吧，那就再来用人话解释一下，但是在开始之前，我需要先解释一些东西。首先，这个公式中是存在一些不需要太过于关注的参数，比如 $h(y)$，这个参数在实际的分布中一般为 1。所以我们可以先不关注它。之后我们就可以再看看其他参数了：

1. y：实际的数据点，也就是我们输入的特征值。
2. η：模型当中分布的核心参数，决定分布的"形状"（如均值、方差）
3. $T(y)$：数据 y 的一个函数。它压缩了 y 中关于 θ 的全部信息。知道 $T(y)$ 就足以推断 θ （比如在伯努利分布中，这个值就是： y 自己本身）。
4. η^T：是将 η 映射到自然参数的函数。这个值不是统一确定的，而是根据不同分布得到的不同 η^T （比如高斯分布中，这个值就是 $\eta^T=[\eta1,\eta2]^\top$ ）。它是把不同分布联系在一起的重要参数，好比是"中间参数"，把均值 μ 等参数转化为模型其他参数自然可求（见下图）
5. $A(\eta)$：确保概率分布总和/积分为 1，保证指数族分布最终归一。它的导数给出分布的重要特征（如均值 $A'(\eta)$、方差 $A''(\eta)$）。它的目的，就是确保模型归一

代码块

```
真实世界数据 (y) <---(由 μ 决定分布)---> 真实均值 (μ)
                         ^
                         |
               正则关联函数 (η = g⁻¹(μ))或者叫连接函数
                         |
                         v
                  自然参数 (η) <---(模型计算: η = b0 + b1*X1 + ...)--- 模型输
入 (X)
                         |
                         |
               正则响应函数 (μ = g(η))或者叫响应函数
                         |
                         v
                  模型预测均值 (μ_pred)
```

总结：概率 $p(y|\theta)$ 由基础结构 $h(y)$ 乘以一个指数项构成。该指数项的核心是：自然参数 $\eta(\theta)$ 与数据信息 $T(y)$ 的点积，再减去一个保证归一化的“调节器” $A(\theta)$。（十分精辟的总结了）

关键参数与公式推导

上边我提到，我没怎么看懂它的公式推导，但是还是要拿出来给大家放一下，万一各位能看明白呢（

$$f(y\mid\boldsymbol{\eta})=h(y)\exp\left(\boldsymbol{\eta}^\top\mathbf{T}(y)-A(\boldsymbol{\eta})\right)$$

放一下公式，不然来回看很麻烦

1. 自然参数 η 与均值 μ 的关系

对 $A(\boldsymbol{\eta})$ 求导可得均值 μ：

$$\mu\triangleq\mathbb{E}[y\mid\boldsymbol{\eta}]=\nabla A(\boldsymbol{\eta})$$

证明：

由归一化条件 $\int f(y \mid \boldsymbol{\eta}) dy = 1$：

$$\int h(y) \exp\left(\boldsymbol{\eta}^\top \mathbf{T}(y) - A(\boldsymbol{\eta})\right) dy = 1$$

两边对 $\boldsymbol{\eta}$ 求梯度：

$$\nabla \int h(y) \exp\left(\boldsymbol{\eta}^\top \mathbf{T}(y) - A(\boldsymbol{\eta})\right) dy = \nabla 1 = 0$$

交换积分与梯度顺序：

1. 使用乘积法则和链式法则：
$$\nabla \left[h(y) \exp(\cdot)\right] = h(y) \exp(\cdot) \cdot \nabla \left[\boldsymbol{\eta}^\top \mathbf{T}(y) - A(\boldsymbol{\eta})\right]$$
2. 计算中括号内的梯度：
$$\nabla \left[\boldsymbol{\eta}^\top \mathbf{T}(y) - A(\boldsymbol{\eta})\right] = \mathbf{T}(y) - \nabla A(\boldsymbol{\eta})$$
3. 因此：
$$\nabla \left[h(y) \exp(\cdot)\right] = h(y) \exp\left(\boldsymbol{\eta}^\top \mathbf{T}(y) - A(\boldsymbol{\eta})\right) \cdot (\mathbf{T}(y) - \nabla A(\boldsymbol{\eta}))$$
4. 即：
$$= (\mathbf{T}(y) - \nabla A(\boldsymbol{\eta})) \cdot f(y \mid \boldsymbol{\eta})$$
5. 得到关键等式，代回积分方程：
$$\int (\mathbf{T}(y) - \nabla A(\boldsymbol{\eta}))\, f(y \mid \boldsymbol{\eta}) dy = 0$$
6. 即：
$$\mathbb{E}[\mathbf{T}(y) \mid \boldsymbol{\eta}] - \nabla A(\boldsymbol{\eta}) = 0 \quad \Rightarrow \quad \boxed{\mu = \mathbb{E}[y \mid \boldsymbol{\eta}] = \nabla A(\boldsymbol{\eta})}$$

2. 正则响应函数（均值函数）

均值 μ 是自然参数 $\boldsymbol{\eta}$ 的函数：

$$\mu = g^{-1}(\boldsymbol{\eta}) = \nabla A(\boldsymbol{\eta})$$

其中 $g^{-1}(\cdot)$ 称为**正则响应函数**。

3. 正则关联函数（链接函数的反函数）

自然参数 $\boldsymbol{\eta}$ 是均值 μ 的函数：

$$\boldsymbol{\eta} = g(\mu)$$

其中 $g(\cdot)$ 称为 **正则关联函数，是响应函数的反函数**。

与真实数据 y 的关系

设观测数据集为 $\{y_i, \mathbf{x}_i\}_{i=1}^N$，在广义线性模型（GLM，也就是指在指数族分布下的线性模型）中：

1. **线性预测器**：$\boldsymbol{\eta}_i = \boldsymbol{\beta}^\top \mathbf{x}_i$
2. **均值映射**：$\mu_i = g^{-1}(\boldsymbol{\eta}_i) = \nabla A(\boldsymbol{\eta}_i)$
3. **数据生成**：$y_i \sim f(y_i \mid \boldsymbol{\eta}_i)$，即数据从指数族分布采样

哦，好吧，我貌似看懂了，我就来简单总结一下推导过程吧。$A(\boldsymbol{\eta})$ 这个参数的一阶导是均值 $\mu = A'(\eta)$（推导过程：对原式积分，再对自然参数 $\boldsymbol{\eta}$ 求梯度，交换积分与梯度顺序，最后再带入积分），而均值 μ 和 $\boldsymbol{\eta}$ 有一定的联系（伊塔映射到均值是响应函数 $g^{-1}(\eta)$，均值映射到伊塔是连接函数 $g(\mu)$，这两个函数互为反函数），而对于真实数据 y 则是不同的分布有不同的关系（线性预测器、均值映射、数据生成）

之后，我们又要开始从头推导线性模型相关的公式了，但是在开始之前，我们需要先将常用的几个回归所需要的分布，以指数族分布表示出来：

1. **高斯分布的指数族分布形式**
2. **伯努利分布的指数族分布形式**
3. **多项式分布的指数族分布形式**

话不多说，直接开始吧！

高斯分布的指数族分布形式

DeepSeek prompt： `根据给出的指数族分布公式，推导高斯分布的指数族分布形式，并请使用Latex表示，给出：自然参数、充分统计量、累积量函数、基测度、均值、方差、响应函数、连接函数`

这部分其实不难（比上边那个要简单），数学好点的应该都能看懂，最好是能手推，不过不强求，毕竟我们不是高斯（

AI：

我们的目标是将高斯分布表示为这种形式。所以我们将PDF原式尽可能地向指数族分布形式靠拢，高斯分布（正态分布）的 PDF 为：

$$f(y \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right)$$

其中 μ 是均值，σ^2 是方差（$\sigma^2 > 0$）。我们需要重写这个表达式以匹配指数族形式。

1. 展开平方项

展开 $(y-\mu)^2 = y^2 - 2y\mu + \mu^2$：

$$
\begin{aligned}
f(y \mid \mu, \sigma^2) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{y^2 - 2y\mu + \mu^2}{2\sigma^2}\right) \\
&= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{y^2}{2\sigma^2} + \frac{2y\mu}{2\sigma^2} - \frac{\mu^2}{2\sigma^2}\right)
\end{aligned}
$$

简化：

$$
f(y \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{\mu y}{\sigma^2} - \frac{y^2}{2\sigma^2} - \frac{\mu^2}{2\sigma^2}\right)
$$

2. 取对数再重组

将常数项 $\frac{1}{\sqrt{2\pi\sigma^2}}$ 写成指数形式：

$$
\frac{1}{\sqrt{2\pi\sigma^2}} = \exp\left(-\frac{1}{2}\log(2\pi\sigma^2)\right)
$$

代入：

$$
f(y \mid \mu, \sigma^2) = \exp\left(-\frac{1}{2}\log(2\pi\sigma^2)\right) \exp\left(\frac{\mu y}{\sigma^2} - \frac{y^2}{2\sigma^2} - \frac{\mu^2}{2\sigma^2}\right)
$$

合并指数：

$$
f(y \mid \mu, \sigma^2) = \exp\left(\frac{\mu y}{\sigma^2} - \frac{y^2}{2\sigma^2} - \frac{\mu^2}{2\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2)\right)
$$

3. 表示识别指数族参数

1. 推导 η^T 和 $\mathbf{T}(y)$

将上式的项分为三部分：与 y 相关的项、与参数相关的项（μ、σ）、常数项。目标是匹配指数族形式：

$$
f(y \mid \boldsymbol{\eta}) = h(y) \exp\left(\boldsymbol{\eta}^\top \mathbf{T}(y) - A(\boldsymbol{\eta})\right)
$$

这里：

- 充分统计量 $\mathbf{T}(y)$：选择为 $\mathbf{T}(y) = \begin{bmatrix} y \\ y^2 \end{bmatrix}$ （因为项 $\frac{\mu y}{\sigma^2}$ 和 $-\frac{y^2}{2\sigma^2}$ 依赖于 y ）。
- 自然参数 $\boldsymbol{\eta}$：定义为 $\boldsymbol{\eta} = \begin{bmatrix} \eta_1 \\ \eta_2 \end{bmatrix}$ ，其中：

$$
\eta_1 = \frac{\mu}{\sigma^2}, \quad \eta_2 = -\frac{1}{2\sigma^2}
$$

如果你将 η^T 和 $\mathbf{T}(y)$ 相积，那就是 $\frac{\mu y}{\sigma^2} - \frac{y^2}{2\sigma^2}$ 。注意： $\eta_2 < 0$ ，因为 $\sigma^2 > 0$ 。

- 用 $\boldsymbol{\eta}$ 表示 μ和σ^2 ：

$$\sigma^2 = -\frac{1}{2\eta_2}, \quad \mu = \eta_1 \sigma^2 = \eta_1 \left(-\frac{1}{2\eta_2}\right) = -\frac{\eta_1}{2\eta_2}$$

4. 定义 $A(\boldsymbol{\eta})$ 和 $h(y)$

1. 推导 $A(\boldsymbol{\eta})$

在指数族中，$A(\boldsymbol{\eta})$ 是累积量函数，确保 PDF 积分为 1。从上式提取与参数相关的项（即不依赖于 y 的常数项）：

$$-\frac{\mu^2}{2\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2)$$

我们上边已经计算出 μ和σ^2，所以用 $\boldsymbol{\eta}$ 表示常数项：

- 计算 $\frac{\mu^2}{2\sigma^2}$：

$$\frac{\mu^2}{2\sigma^2} = \frac{\left(-\frac{\eta_1}{2\eta_2}\right)^2}{2\left(-\frac{1}{2\eta_2}\right)} = \frac{\frac{\eta_1^2}{4\eta_2^2}}{-\frac{1}{\eta_2}} = \frac{\eta_1^2}{4\eta_2^2} \times (-\eta_2) = -\frac{\eta_1^2}{4\eta_2}$$

- 计算 $\frac{1}{2}\log(2\pi\sigma^2)$：

$$\sigma^2 = -\frac{1}{2\eta_2} \implies \log(2\pi\sigma^2) = \log\left(2\pi\left(-\frac{1}{2\eta_2}\right)\right) = \log\left(-\frac{\pi}{\eta_2}\right)$$

所以：

$$\frac{1}{2}\log(2\pi\sigma^2) = \frac{1}{2}\log\left(-\frac{\pi}{\eta_2}\right)$$

注意：$-\frac{\pi}{\eta_2} > 0$ 因为 $\eta_2 < 0$，所以对数定义良好。

- 因此：

$$\begin{aligned} -\frac{\mu^2}{2\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2) &= -\left(-\frac{\eta_1^2}{4\eta_2}\right) - \frac{1}{2}\log\left(-\frac{\pi}{\eta_2}\right) \\ &= \frac{\eta_1^2}{4\eta_2} - \frac{1}{2}\log\left(-\frac{\pi}{\eta_2}\right) \end{aligned}$$

所以最终常数项（也就是 $A(\boldsymbol{\eta})$ 累积量函数）为：

$$A(\boldsymbol{\eta}) = \frac{\eta_1^2}{4\eta_2} - \frac{1}{2}\log\left(-\frac{\pi}{\eta_2}\right)$$

2. 推导 $h(y)$

因为没有剩余项，因此外部剩余的 $h(y)$ 项为 1 （这里AI推导错了，还好我从头到尾看了一遍）：

$$h(y) = 1$$

最终参数整理

- 自然参数（Natural Parameter）：

$$\boldsymbol{\eta} = \begin{bmatrix} \eta_1 \\ \eta_2 \end{bmatrix} = \begin{bmatrix} \frac{\mu}{\sigma^2} \\ -\frac{1}{2\sigma^2} \end{bmatrix}$$

- 充分统计量（Sufficient Statistic）：

$$\mathbf{T}(y) = \begin{bmatrix} y \\ y^2 \end{bmatrix}$$

- 累积量函数（Cumulant Function / Log-Partition Function）：

$$A(\boldsymbol{\eta}) = \frac{\eta_1^2}{4\eta_2} + \frac{1}{2}\log\left(-\frac{\pi}{\eta_2}\right)$$ 注意变符号，这里AI错了

- 基测度（Base Measure）：

$$h(y) = 1$$

- 均值（Mean）：

$$\mu = A'(\eta) = -\frac{\eta_1}{2\eta_2}$$

- 方差（Variance）：

$$\sigma^2 = A''(\eta) = -\frac{1}{2\eta_2}$$

- 响应函数（Response Function）：

$$\boxed{\begin{aligned} \mu &= -\frac{\eta_1}{2\eta_2} \\ \sigma^2 &= -\frac{1}{2\eta_2} \end{aligned}}$$

- 连接函数（Link Function）：

$$\boxed{\begin{aligned} \eta_1 &= \frac{\mu}{\sigma^2} \\ \eta_2 &= -\frac{1}{2\sigma^2} \end{aligned}}$$

到此，我们就正确推导出了指数族分布下的高斯分布！

伯努利分布的指数族分布形式

DeepSeek prompt： **根据给出的指数族分布公式，推导伯努利分布的指数族分布形式，并请使用Latex表示，给出：自然参数、充分统计量、累积量函数、基测度、均值、方差、响应函数、连接函数**

AI：

以下是针对伯努利分布（Bernoulli distribution）的指数族分布形式推导。伯努利分布描述二元随机变量 $Y \in \{0,1\}$，其概率质量函数为：

$$P(Y=y \mid p)=p^{y}(1-p)^{1-y}, \quad y \in \{0,1\} \tag{1}$$

其中 p 是成功概率（即 $P(Y=1)=p$）。我们将此重写为指数族的一般形式：

$$f(y \mid \eta)=h(y)\exp\left(\eta^{\top}T(y)-A(\eta)\right) \tag{2}$$

其中 η 是自然参数，$T(y)$ 是充分统计量，$A(\eta)$ 是累积量函数（log-partition function），$h(y)$ 是基测度。

1. 重写伯努利分布的概率质量函数：

$$\begin{aligned}
P(Y=y \mid p) &= p^{y}(1-p)^{1-y} && (3)\\
&= \exp\left(\ln\left(p^{y}(1-p)^{1-y}\right)\right) && (4)\\
&= \exp\left(y\ln p+(1-y)\ln(1-p)\right) && (5)
\end{aligned}$$

展开指数部分：

$$\begin{aligned}
y\ln p+(1-y)\ln(1-p) &= y\ln p+\ln(1-p)-y\ln(1-p) && (6)\\
&= y\ln\left(\frac{p}{1-p}\right)+\ln(1-p) && (7)
\end{aligned}$$

因此：

$$P(Y=y \mid p)=\exp\left(y\ln\left(\frac{p}{1-p}\right)+\ln(1-p)\right)$$

2. 定义自然参数：

令自然参数 $\eta=\ln\left(\frac{p}{1-p}\right)$（即 logit 函数）。则：

$$\begin{aligned}
e^{\eta}=\frac{p}{1-p} \implies p &= \frac{e^{\eta}}{1+e^{\eta}}, && (8)\\
1-p &= \frac{1}{1+e^{\eta}} && (9)
\end{aligned}$$

代入 $\ln(1-p)$：

$$
\begin{aligned}
\ln(1-p) &= \ln\left(\frac{1}{1+e^\eta}\right) && (10) \\
&= -\ln(1+e^\eta) && (11)
\end{aligned}
$$

因此，概率质量函数变为：

$$P(Y=y \mid \eta) = \exp\left(\eta y - \ln(1+e^\eta)\right) \tag{12}$$

3. 匹配指数族形式：

比较一般形式：

$$f(y \mid \eta) = h(y) \exp\left(\eta^\top T(y) - A(\eta)\right) \tag{13}$$

可得：

- 充分统计量 $T(y) = y$ （标量）。
- 自然参数 $\eta = \ln\left(\frac{p}{1-p}\right)$ （标量）。
- 累积量函数 $A(\eta) = \ln(1+e^\eta)$。
- 基测度 $h(y)$：由于分布仅在 $y \in \{0,1\}$ 定义，且无额外因子，$h(y) = 1$ （更精确地，$h(y)$ 是计数测度在 $\{0,1\}$ 上的指示函数，但通常简写为 1）。

最终，伯努利分布的指数族形式为：

$$f(y \mid \eta) = 1 \cdot \exp\left(\eta y - \ln(1+e^\eta)\right), \quad y \in \{0,1\} \tag{14}$$

4. 各部分定义

以下是伯努利分布指数族形式的具体部分：

- **自然参数（Natural Parameter）**：

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

- **充分统计量（Sufficient Statistic）**：

$$T(y) = y$$

- **累积量函数（Cumulant Function / Log-Partition Function）**：

$$A(\eta) = \ln(1+e^\eta)$$

- **基测度（Base Measure）**：

$$h(y) = 1 \qquad \text{（对于 } y \in \{0,1\}\text{，否则为 0）}$$

（注：在实现中，基测度通常隐含支持集，因此常简写为 $h(y)=1$）

- **均值（Mean）**：

在指数族中，均值 $\mu=E[Y\mid\eta]=A'(\eta)$。计算一阶导数：

$$A'(\eta)=\frac{d}{d\eta}\ln(1+e^{\eta})=\frac{e^{\eta}}{1+e^{\eta}}=p=\mu$$

因此：

$$\mu=\frac{e^{\eta}}{1+e^{\eta}}$$

- **方差（Variance）**：

方差 $\mathrm{Var}(Y\mid\eta)=A''(\eta)$。计算二阶导数：

$$\begin{aligned}A''(\eta)&=\frac{d}{d\eta}\left(\frac{e^{\eta}}{1+e^{\eta}}\right)\\&=\frac{e^{\eta}(1+e^{\eta})-e^{\eta}\cdot e^{\eta}}{(1+e^{\eta})^2}\\&=\frac{e^{\eta}}{(1+e^{\eta})^2}\\&=\mu(1-\mu)\end{aligned}$$

因此：

$$\mathrm{Var}(Y)=\mu(1-\mu)$$

- **响应函数（Response Function）**：

响应函数将自然参数 η 映射到均值 μ，即 $\mu=g^{-1}(\eta)$。由均值表达式：

$$\mu=\frac{e^{\eta}}{1+e^{\eta}}$$

这称为 logistic 函数（或 sigmoid 函数）。因此：

$$g^{-1}(\eta)=\frac{e^{\eta}}{1+e^{\eta}}$$

- **连接函数（Link Function）**：

连接函数是响应函数的反函数，将均值 μ 映射到自然参数 η，即 $\eta=g(\mu)$。由自然参数定义：

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

这称为 logit 函数。因此：

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

5. 其他问题

关于我们怎么定义 logit 函数

在看AI推导的过程当中我们也需要思考，比如这个问题：**在 (8) 、(9) 推导过程中这个logit函数怎么来的？为什么要这么定义？是有什么规律或者要求么？**

然后你就可以带着你的问题去问AI：

My Question： `为什么要这么定义？有什么规律或者要求？ （图片）`

AI（Kimi）：

这个函数可以把（0，1）这个区间的值映射到 正负无限上，可以把 p 映射到整个实数范围，这在数学建模中非常有用，尤其是在需要线性表示的地方（例如线性回归或广义线性模型）。

- $\frac{p}{1-p}$ 表示事件发生的几率（odds），即事件发生概率与不发生概率的比值。
- $ln(\frac{p}{1-p})$ 是对数几率（log-odds），它将几率转换为对数尺度，使得模型更具线性性质。

My Question：也就是说，这个函数可以把非线性的二元关系转化为非二元的线性关系？

AI（Kimi）：

是的，您的理解是正确的。在统计学和机器学习中，这个转换过程确实可以把非线性的二元关系转化为非二元的线性关系...（下内容省略）

多项式分布的指数族分布形式

这个分布在之前的推导里是没有出现过的，所以我会先给出公式和相关案例让大家进行参考（这个部分AI给的推导太拉了，所以我又看了一边手动加了点详细的推导过程）。

DeepSeek prompt： `根据给出的指数族分布公式，推导多项式分布的指数族分布形式，并请使用Latex表示，给出：自然参数、充分统计量、累积量函数、基测度、均值、方差、响应函数、连接函数`

AI：

考虑 K 个类别的多项式分布（单次试验，即 $n=1$）。随机变量 Y 服从类别分布（Categorical distribution），其概率质量函数为：

$$P(Y=y \mid \mathbf{p})=\prod_{k=1}^{K} p_k^{y_k}=\exp\left(\sum_{k=1}^{K} y_k \ln p_k\right) \tag{1}$$

其中：

- $y=(y_1,\ldots,y_K)^\top$ 是 one-hot 向量（即 $y_k \in \{0,1\}$且$\sum_{k=1}^{K} y_k=1$） （什么是one-hot向量）

- $\mathbf{p}=(p_1,\ldots,p_K)^\top$ 满足 $p_k>0$ 和 $\sum_{k=1}^{K} p_k=1$

1. 场景：餐厅点餐

假设你正在一家餐厅点餐，餐厅的菜单上有3种不同的菜品：意大利面、汉堡和沙拉。你选择每种菜品的概率分别是：

- $p_1=0.4$ （选择意大利面的概率）
- $p_2=0.3$ （选择汉堡的概率）
- $p_3=0.3$ （选择沙拉的概率）

如果你选择了汉堡，对应的one-hot向量为 $Y=[0,1,0]$ 。

1. 概率质量函数计算

根据概率质量函数：

$$P(Y=[0,1,0]\mid\mathbf{p})=p_1^0\times p_2^1\times p_3^0=1\times0.3\times1=0.3$$

2. 指数族形式计算

根据指数族形式：

$$\begin{aligned}P(Y=[0,1,0]\mid\mathbf{p})&=\exp\left(0\cdot\ln p_1+1\cdot\ln p_2+0\cdot\ln p_3\right)\\&=\exp(\ln p_2)\\&=p_2=0.3\end{aligned}$$

两种方法计算的结果一致，验证了指数族形式的推导是正确的。这里大家应该对这个公式的具体计算方式就有所了解了，接下来我们来表示指数族分布。

2. 参数冗余处理

由于存在约束 $\sum_{k=1}^{K} p_k=1$ （所有菜品被选择概率之和为 1），我们以第 K 类为参考类，定义自然参数（和伯努利分布推导自然参数类似，也是对数，但是这里还有一个问题，我们**埋个雷**）：

$$\eta_k=\ln\left(\frac{p_k}{p_K}\right),\quad k=1,\ldots,K-1 \tag{2}$$

(2) 两边取反对数，移项解出 p_k ：

$$p_k=p_Ke^{\eta_k},\quad k=1,\ldots,K-1 \tag{3}$$

代入约束条件：

$$
\begin{aligned}
\sum_{k=1}^{K} p_k &= p_K + \sum_{k=1}^{K-1} p_K e^{\eta_k} \\
&= p_K \left(1 + \sum_{k=1}^{K-1} e^{\eta_k}\right) \\
&= 1
\end{aligned} \tag{4}
$$

(4) 得 p_K 、将 (4) p_K 带入 (3) 解得 p_k ：

$$
p_K = \frac{1}{1+\sum_{k=1}^{K-1} e^{\eta_k}}, \quad p_k = p_K e^{\eta_k} = \frac{e^{\eta_k}}{1+\sum_{j=1}^{K-1} e^{\eta_j}}, \quad k = 1, \ldots, K-1 \tag{5}
$$

3. 重写概率质量函数

将 $\ln p_k$ 用 η 表示：

$$
\ln p_k = \begin{cases} \eta_k + \ln p_K & k = 1, \ldots, K-1 \\ \ln p_K & k = K \end{cases} \tag{6}
$$

A. 这里可能有点不太好理解，为啥会有两种情况？实际上是上边 (3) 式得出的：

$$
\begin{aligned}
\ln p_k &= \ln p_K e^{\eta_k} \\
&= \ln e^{\eta_k} + \ln p_K \\
&= \eta_k + \ln p_K \text{，} k = 1, \ldots, K-1
\end{aligned}
$$

B. 这就出现了第一种情况，但是由于取值只能到 $K-1$ ，所以当取 K 时， $\eta_k = 0$ （η 是个 $k-1$ 的向量）

之后代入概率质量函数：

$$
\begin{aligned}
\text{公式}(1) \sum_{k=1}^{K} y_k \ln p_k &= \sum_{k=1}^{K-1} y_k (\eta_k + \ln p_K) + y_K \ln p_K \\
&= \sum_{k=1}^{K-1} y_k \eta_k + (\sum_{k=1}^{K-1} y_k \ln p_K + y_K \ln p_K) \\
&= \sum_{k=1}^{K-1} y_k \eta_k + \ln p_K \underbrace{\left(\sum_{k=1}^{K} y_k\right)}_{=1}
\end{aligned} \tag{7}
$$

代入 p_K 的表达式：

$$
\begin{aligned}
\ln p_K &= \ln\left(\frac{1}{1+\sum_{j=1}^{K-1} e^{\eta_j}}\right) \\
&= \ln\left(\frac{1+\sum_{j=1}^{K-1} e^{\eta_j}}{1+\sum_{j=1}^{K-1} e^{\eta_j}}\right) - \ln\left(1+\sum_{j=1}^{K-1} e^{\eta_j}\right) \\
&= -\ln\left(1+\sum_{j=1}^{K-1} e^{\eta_j}\right)
\end{aligned} \tag{8}
$$

最终将 (8) 和 (7) 的结论，带入 $\exp\left(\sum_{k=1}^{K} y_k \ln p_k\right)$ 得到指数族形式：

$$
P(Y=y \mid \eta) = \exp\left(\sum_{k=1}^{K-1} \eta_k y_k - \ln\left(1+\sum_{j=1}^{K-1} e^{\eta_j}\right)\right) \tag{9}
$$

4. 各部分定义

1. **自然参数（Natural Parameter）**：

$$
\boxed{\eta = \begin{pmatrix} \eta_1 \\ \vdots \\ \eta_{K-1} \end{pmatrix} = \begin{pmatrix} \ln \frac{p_1}{p_K} \\ \vdots \\ \ln \frac{p_{K-1}}{p_K} \end{pmatrix} \in \mathbb{R}^{K-1}\text{这个}\mathbb{R}\text{是自然数}}
$$

2. **充分统计量（Sufficient Statistic）**：

$$
\boxed{T(y) = \begin{pmatrix} y_1 \\ \vdots \\ y_{K-1} \end{pmatrix} \in \mathbb{R}^{K-1}}
$$

3. **累积量函数（Cumulant Function）**：

$$
\boxed{A(\eta) = \ln\left(1+\sum_{k=1}^{K-1} e^{\eta_k}\right)}
$$

4. **基测度（Base Measure）**：

$$
\boxed{h(y) = \begin{cases} 1 & \text{if } y \text{ is one-hot vector} \\ 0 & \text{otherwise} \end{cases}}
$$

5. **均值（Mean）**：

指数族性质：$\mathbb{E}[T(y)] = \nabla_\eta A(\eta)$

$$\frac{\partial A(\eta)}{\partial \eta_k} = \frac{e^{\eta_k}}{1+\sum_{j=1}^{K-1} e^{\eta_j}} = p_k, \quad k=1,\ldots,K-1$$

因此：

$$\boxed{\mathbb{E}[T(y)] = \nabla A(\eta) = \begin{pmatrix} p_1 \\ \vdots \\ p_{K-1} \end{pmatrix}}$$

其中 $p_k = \frac{e^{\eta_k}}{1+\sum_{j=1}^{K-1} e^{\eta_j}}$ 。

6. **方差-协方差矩阵（Variance-Covariance）**：

指数族性质： $\mathrm{Cov}(T(y)) = \nabla_\eta^2 A(\eta)$

$$\frac{\partial^2 A(\eta)}{\partial \eta_k \partial \eta_l} = \begin{cases} p_k(1-p_k) & k=l \\ -p_k p_l & k \neq l \end{cases}$$

因此：

$$\boxed{\mathrm{Cov}(T(y)) = \begin{pmatrix} p_1(1-p_1) & -p_1p_2 & \cdots & -p_1p_{K-1} \\ -p_2p_1 & p_2(1-p_2) & \cdots & -p_2p_{K-1} \\ \vdots & \vdots & \ddots & \vdots \\ -p_{K-1}p_1 & -p_{K-1}p_2 & \cdots & p_{K-1}(1-p_{K-1}) \end{pmatrix}}$$

7. **响应函数（Response Function）**：

将自然参数映射到均值：

$$\boxed{g^{-1}(\eta) = \begin{pmatrix} \frac{e^{\eta_1}}{1+\sum_{k=1}^{K-1} e^{\eta_k}} \\ \vdots \\ \frac{e^{\eta_{K-1}}}{1+\sum_{k=1}^{K-1} e^{\eta_k}} \end{pmatrix} = \begin{pmatrix} p_1 \\ \vdots \\ p_{K-1} \end{pmatrix}}$$

8. **连接函数（Link Function）**：

响应函数的反函数（将均值映射到自然参数）：

$$\boxed{g(\mu) = \begin{pmatrix} \ln \frac{\mu_1}{1-\sum_{j=1}^{K-1} \mu_j} \\ \vdots \\ \ln \frac{\mu_{K-1}}{1-\sum_{j=1}^{K-1} \mu_j} \end{pmatrix} = \begin{pmatrix} \ln \frac{\mu_1}{\mu_K} \\ \vdots \\ \ln \frac{\mu_{K-1}}{\mu_K} \end{pmatrix}}$$

其中 $\mu_k = \mathbb{E}[y_k] = p_k$ ，且 $\mu_K = 1 - \sum_{k=1}^{K-1} \mu_k$.

最终形式：

$$\boxed{f(y \mid \eta) = \exp\left(\sum_{k=1}^{K-1} \eta_k y_k - \ln\left(1 + \sum_{j=1}^{K-1} e^{\eta_j}\right)\right) \cdot \mathbf{1}_{\{y \in \{0,1\}^K,\, \sum_{k=1}^{K} y_k = 1\}}}$$

5. 其他问题

5.1 为什么我们的 (2) 取值是从 $1 \ldots K-1$ （雷处理）？

这个问题实际上我们应该可以有所察觉，如果所有选择的概率之和为1，那么只设定 $1 \ldots K-1$ 就可以表示全部的概率了，也就是说 p_K 的概率可以表示为 $1 - \sum_{i=1}^{K-1} p_k$ 。我们来看看 AI 怎么说的

AI（DeepSeek-R1）：

在多项式分布的指数族形式中，我们使用 $K-1$ 个类别（而不是 K 个）的原因源于概率分布的约束条件和参数冗余问题。以下是详细解释（使用 LaTeX 表示）：

i. 概率分布的约束条件

多项式分布的概率质量函数要求：

$\sum k = 1^K p_k = 1 \quad 且 \quad p_k > 0$

这意味着 K 个概率参数 $(p_1, \ldots, p_K)$ 不是独立的——只要确定了 $K-1$ 个参数，最后一个参数可通过约束导出：

$p_K = 1 - \sum k = 1^{K-1} p_k$

ii. 指数族参数化的要求

指数族分布的标准形式要求：

- 自然参数 η 的各个分量**相互独立**（1）
- 参数空间是开集（保证 $A(\eta)$ 可微）（2）

如果强行使用 K 个自然参数（例如 $\eta_k = \ln p_k$ ），会违反这些要求：

- 约束 $\sum_{k=1}^{K} p_k = 1$ 导致参数间线性相关（3）
- 协方差矩阵 $\nabla^2 A(\eta)$ 奇异（不满秩）（4）

好吧，实际上 AI 解释的有些难以理解了，我大致看了一下稍微总结的能看懂一点。首先是概率分布约束问题，这里**AI说的** $(p_1, \ldots, p_K)$ **不是独立的**，和矩阵的线性无关有联系，当我们表示了前 $K-1$ 个数后，实际上前边的数据都是线性无关的，而加入第 K 个数据后，整个数据就变成了线性有关，这个在矩阵的相关性质里有详细的解释，不懂的可以回去重修线性代数了。之后AI说了四个点（标绿的），1 和 2 不需要太清楚，~~因为我也不清楚~~，但是 3 和 4 好解释，3 就是我们说的

加入 K 会导致线性相关，而 4 则是站在需要用到数据的相关矩阵角度，**当一个矩阵的秩不是满秩，则这个矩阵必定不完全线性无关**。到此不懂的真的可以重修线代了。

5.2 为什么响应函数和连接函数使用梯度求的?

AI（DeepSeek-R1）：

在指数族分布中，自然参数 η 通常是向量形式（此例中为 $(K-1)$ 维向量），因此需要使用向量微积分而不是普通导数。而当自然参数是标量（如伯努利分布）时可用普通导数，但在多项式分布等多元情况下必须使用向量微积分。

其他

到此，我们就将**高斯分布的指数族分布形式、伯努利分布的指数族分布形式、多项式分布的指数族分布形式**推了一遍，但是我们还是有疑问，为什么我们一定要把这几个分布统一化？数学界是有什么怪癖么？自问自答，这个问题我想我应该站在实际的应用角度解释。

在统一数学建模的基础上，我们应该清楚的知道，不同的分布有自己不同的作用，高斯分布可以用来预测线性相关的数据，处理线性回归，伯努利分布可以有效地处理二分类问题，逻辑回归，而多项式分布可以将多元分类问题...但是指数族分布可不止这些：

- 泊松分布：对计数类数据建模
- 伽马和指数分布：这个用于对连续的、非负的随机变量进行建模，例如时间间隔
- 贝塔和狄利克雷分布：用于概率（参数）的分布的建模

想要了解更多的，你可以像上边每一个提示词那样，向 AI 发问，你也可以参考这个简单而易懂的视频。

广义线性模型

当我们再次回来看这个公式：

$$f(y \mid \boldsymbol{\eta}) = h(y) \exp\left(\boldsymbol{\eta}^\top \mathbf{T}(y) - A(\boldsymbol{\eta})\right)$$

你会发现好像也没有那么难理解了，甚至有点优雅~~（艹）~~，但是我们还不能止步，我们在之前的推导中（高斯分布），我们可是从最大似然推最小二乘再到批量梯度和正规方程的，我们以指数族分布形式大概率也是避免不了一顿算...所以还是老老实实脚踏实地的继续吧，毕竟各位都不想半途而废。

线性回归（高斯分布）

最大似然推导最小二乘法

考虑线性回归模型：

- 数据：n 个独立观测 $\{(x_i, y_i)\}_{i=1}^n$，其中 $x_i \in \mathbb{R}^d$ 为特征向量，$y_i \in \mathbb{R}$ 为响应。
- 模型：$y_i \sim \mathcal{N}(\mu_i, \sigma^2)$，其中均值 $\mu_i = w^\top x_i$（线性预测），方差 $\sigma^2 > 0$ 恒定（同方差）。
- 带入其指数族形式：

$$p(y_i|\eta_i) = h(y_i) \exp\left(\eta_i^\top T(y_i) - A(\eta_i)\right)$$

根据之前的推导，其中：

  - $h(y_i) = 1$（基测度）。
  - 充分统计量 $T(y_i) = \begin{bmatrix} y_i \\ y_i^2 \end{bmatrix}$。
  - 自然参数 $\eta_i = \begin{bmatrix} \eta_{1i} \\ \eta_{2i} \end{bmatrix} = \begin{bmatrix} \frac{\mu_i}{\sigma^2} \\ -\frac{1}{2\sigma^2} \end{bmatrix} = \begin{bmatrix} \frac{w^\top x_i}{\sigma^2} \\ -\frac{1}{2\sigma^2} \end{bmatrix}$（注意 $\eta_{2i} = \eta_2$ 恒定，因为 σ^2 常数恒定）。
  - 累积量函数：

$$A(\eta_i) = -\frac{\eta_{1i}^2}{4\eta_{2i}} + \frac{1}{2}\log\left(-\frac{\pi}{\eta_{2i}}\right)$$

对数似然函数

将指数族形式的高斯分布两边取对数（联合对数似然、n 个数据独立情况下）：

$$\begin{aligned}
\ell(w, \sigma^2) &= \sum_{i=1}^n \log p(y_i|\eta_i) \\
&= \sum_{i=1}^n \log\left(h(y_i) \exp\left(\eta_i^\top T(y_i) - A(\eta_i)\right)\right) \\
&= \sum_{i=1}^n \left[\eta_i^\top T(y_i) - A(\eta_i)\right]
\end{aligned}$$

代入 $\eta_i = \begin{bmatrix} \eta_{1i} \\ \eta_{2i} \end{bmatrix}$ 和 $T(y_i) = \begin{bmatrix} y_i \\ y_i^2 \end{bmatrix}$：

$$\ell(w, \sigma^2) = \sum_{i=1}^n \left[\eta_{1i} y_i + \eta_{2i} y_i^2 - A(\eta_i)\right]$$

其中：

$$\eta_{1i} = \frac{w^\top x_i}{\sigma^2}, \quad \eta_{2i} = -\frac{1}{2\sigma^2}, \quad A(\eta_i) = -\frac{\eta_{1i}^2}{4\eta_{2i}} + \frac{1}{2}\log\left(-\frac{\pi}{\eta_{2i}}\right)$$

简化 $A(\eta_i)$：

$$A(\eta_i) = -\frac{\left(\frac{w^\top x_i}{\sigma^2}\right)^2}{4\left(-\frac{1}{2\sigma^2}\right)} + \frac{1}{2}\log\left(-\frac{\pi}{-\frac{1}{2\sigma^2}}\right) = \frac{(w^\top x_i)^2}{2\sigma^2} + \frac{1}{2}\log(2\pi\sigma^2)$$

代入对数似然：

$$\ell(w,\sigma^2)=\sum_{i=1}^{n}\left[\left(\frac{w^\top x_i}{\sigma^2}\right)y_i+\left(-\frac{1}{2\sigma^2}\right)y_i^2-\left(\frac{(w^\top x_i)^2}{2\sigma^2}+\frac{1}{2}\log(2\pi\sigma^2)\right)\right]$$

整理项，得到对数似然函数：

$$\ell(w,\sigma^2)=\frac{1}{\sigma^2}\sum_{i=1}^{n}\left[(w^\top x_i)y_i-\frac{1}{2}(w^\top x_i)^2\right]-\frac{1}{2\sigma^2}\sum_{i=1}^{n}y_i^2-\frac{n}{2}\log(2\pi\sigma^2)$$

最大似然估计（MLE）与最小二乘法

最小二乘法关注估计权重 w （最小化残差平方和），而 σ^2 为 nuisance 参数（不感兴趣的参数）。

- 关于 w 的项（提取与 w 相关的部分）：

$$\ell_w(w)=\frac{1}{\sigma^2}\sum_{i=1}^{n}\left[(w^\top x_i)y_i-\frac{1}{2}(w^\top x_i)^2\right]+\text{const.}$$

其中 "const." 表示与 w 无关的项（包括 y_i^2 和 $\log(\sigma^2)$ 项）。

- 由于 $\sigma^2>0$，最大化 $\ell_w(w)$ 等价于最大化：

$$\sum_{i=1}^{n}\left[(w^\top x_i)y_i-\frac{1}{2}(w^\top x_i)^2\right]$$

- 等价于最小化：

$$-\sum_{i=1}^{n}\left[(w^\top x_i)y_i-\frac{1}{2}(w^\top x_i)^2\right]=\sum_{i=1}^{n}\left[\frac{1}{2}(w^\top x_i)^2-(w^\top x_i)y_i\right]$$

- 配方（吐槽配方，这东西我看半天没看出来怎么整的）：

$$\begin{aligned}\frac{1}{2}(w^\top x_i)^2-(w^\top x_i)y_i&=\frac{1}{2}\left[(w^\top x_i)^2-2(w^\top x_i)y_i\right]\\&=\frac{1}{2}\left[(w^\top x_i)^2-2(w^\top x_i)y_i+y_i^2-y_i^2\right]\\&=\frac{1}{2}\left[(w^\top x_i-y_i)^2-y_i^2\right]\\&=\frac{1}{2}(w^\top x_i-y_i)^2-\frac{1}{2}y_i^2\end{aligned}$$

因此：

$$\sum_{i=1}^{n}\left[\frac{1}{2}(w^\top x_i)^2-(w^\top x_i)y_i\right]=\frac{1}{2}\sum_{i=1}^{n}(w^\top x_i-y_i)^2-\frac{1}{2}\sum_{i=1}^{n}y_i^2$$

- 忽略与 w 无关的项（$-\frac{1}{2}\sum y_i^2$），最小化上式等价于最小化：

$$\sum_{i=1}^{n}(w^\top x_i-y_i)^2$$

这就是最小二乘法的目标函数（残差平方和）。

对数似然函数推正规方程（得分函数导出）

最大似然估计通过求解得分方程 $\nabla_w \ell = 0$ 得到（和之前的思路一模一样）。

- 对数似然关于 w 的梯度：

$$\nabla_w \ell = \sum_{i=1}^{n} \nabla_w \left[\eta_{1i} y_i - A(\eta_i)\right] \quad \text{(其他项与w无关)}$$

注：在指数族分布形式下，对数似然函数中与权重 w 相关的部分只涉及自然参数 η_{1i}，而 η_{2i} 的 σ^2 常数恒定，所以不需要考虑

其中：

$$\eta_{1i} = \frac{w^\top x_i}{\sigma^2}, \quad \frac{\partial \eta_{1i}}{\partial w} = \frac{x_i}{\sigma^2}, \quad \frac{\partial A(\eta_i)}{\partial w} = \frac{\partial A}{\partial \eta_{1i}} \frac{\partial \eta_{1i}}{\partial w} = \mu_i \frac{x_i}{\sigma^2} = (w^\top x_i) \frac{x_i}{\sigma^2}$$

因此：

$$\nabla_w \ell = \sum_{i=1}^{n} \left[y_i \frac{x_i}{\sigma^2} - (w^\top x_i) \frac{x_i}{\sigma^2} \right] = \frac{1}{\sigma^2} \sum_{i=1}^{n} (y_i - w^\top x_i) x_i$$

- 设梯度为零：

$$\nabla_w \ell = 0 \implies \sum_{i=1}^{n} (y_i - w^\top x_i) x_i = 0$$

这是最小二乘法的正规方程（Normal Equation）。解出 w 即为最小二乘估计。

对数似然函数推批量梯度

基于给定对数似然函数的参数求解

由上述得知的对数似然函数：

$$\ell_w(w) = \frac{1}{\sigma^2} \sum_{i=1}^{n} \left[(w^\top x_i) y_i - \frac{1}{2} (w^\top x_i)^2 \right] + \text{const.}$$

1. **计算梯度**

对权重向量 (w) 求梯度：

$$\nabla_w \ell_w(w) = \frac{1}{\sigma^2} \sum i = 1^n \left[y_i x_i - (w^\top x_i) x_i \right]$$

简化为残差形式：

$$\boxed{\nabla_w \ell_w(w) = \frac{1}{\sigma^2} \sum i = 1^n x_i (y_i - w^\top x_i)}$$

2. **闭式解（令梯度为零）**

设梯度等于零求解最优 (w):

$$\sum i=1^n x_i(y_i - w^\top x_i) = 0$$

展开后:

$$\sum i=1^n x_i y_i = \left(\sum_{i=1}^n x_i x_i^\top\right) w$$

矩阵形式（X 为设计矩阵，行为 $x_i^\top$ ）：

$$X^\top y = X^\top X w$$

解得:

$$\boxed{w = (X^\top X)^{-1} X^\top y}$$

3. 梯度上升/下降更新规则

迭代更新公式（学习率 α ）：

$$w_{\text{new}} = w_{\text{old}} \ \ (+ \, or \, -) \ \ \alpha \cdot \frac{1}{\sigma^2} \sum i=1^n x_i(y_i - w_{\text{old}}^\top x_i)$$

常数 (σ^2) 可吸收进学习率:

$$\boxed{w_{\text{new}} = w_{\text{old}} \ \ (+ \, or \, -) \ \ \alpha' \sum i=1^n x_i(y_i - w_{\text{old}}^\top x_i)}$$

其中 ($\alpha' = \alpha/\sigma^2$).

逻辑回归（伯努利分布）

恭喜你，如果你看到这里实际上就已经完结了，因为逻辑回归几个重要的公式已经在之前的表示就已经见过了（在**伯努利分布的指数族分布形式**部分），比如第 (8) 到 (9) 的公式:

$$e^\eta = \frac{p}{1-p} \implies p = \frac{e^\eta}{1+e^\eta} \quad (\text{sigmod函数})$$

$$1-p = \frac{1}{1+e^\eta}$$

这个公式就是我们所指的，当目标变量满足伯努利分布时，广义线性模型推导出了逻辑回归的假设函数 sigmod 函数（或者logistic 函数）。是剩下的内容，和 局部加权线性回归、逻辑回归、随机梯度与牛顿法 中的推导一样。

softmax 回归推导（多项式分布）

同理，我们在推导**多项式分布的指数族分布形式**就已得到**softmax 回归（公式（5）中的 p_k ）**：

$$p_K = \frac{1}{1+\sum_{k=1}^{K-1} e^{\eta_k}}, \quad p_k = p_K e^{\eta_k} = \frac{e^{\eta_k}}{1+\sum_{j=1}^{K-1} e^{\eta_j}}, \quad k = 1, \ldots, K-1 \tag{5}$$

这里的：

$$p_k = p_K e^{\eta_k} = \frac{e^{\eta_k}}{1+\sum_{j=1}^{K-1} e^{\eta_j}}, \quad k = 1, \ldots, K-1$$

就是我们要的结果！