Exponential Family: Canonical form (1D)

Canonical form 1. The exponential family represents a parametric class of probability distributions defined by their probability density function (pdf) or probability mass function (pmf):

$$f(\xi|\theta) := \frac{1}{Z(\theta)} \cdot h(\xi) \cdot e^{\theta \cdot \xi}, \tag{1}$$

where $\xi \in \mathbb{R}$ represents a value of random variable Ξ, $\theta \in \mathbb{R}$ is a parameter, $Z(\theta) \in \mathbb{R}$ represents a parameter-dependent normalization constant, and $h(\xi) \in \mathbb{R}$ is a parameter-independent scaling function, also called the *carrier measure*. In short notation, $Y \sim \operatorname{Exp}(\theta)$.

The equation (1) is the *canonical form* of the exponential family. The canonical form provides a standardized way to express all exponential and pre-exponential terms.

This family encompasses many common probability distributions. Any distribution whose pdf can be expressed in the form of (1) belongs to the exponential family.

Partition function. To hold normalization, the term called the *partition function* is introduced:

$$Z(\theta) := \int h(\xi) \cdot e^{\theta \cdot T(\xi)} \, \mathrm{d}\xi \,. \tag{2}$$

Corresponding logarithm $A(\theta) := \log Z(\theta)$ is called the *log partition function* or *cumulant function.*

Sufficient statistics. If the random variable Ξ does not have a linear relationship with the parameter θ, a function called *sufficient statistics* $T(\xi)$ is introduced to make the relationship linear:

$$f(\xi|\theta) := \frac{1}{Z(\theta)} \cdot h(\xi) \cdot e^{\theta \cdot T(\xi)}, \tag{3}$$

Technically, $T(\xi)$ is a new random variable ξ' for which (1) holds.

Canonical form 2. Equivalently to (3), the exponential family can be rewritten as a single exponential function when all pre-exponential terms are gathered:

$$f(\xi|\theta) := e^{\theta \cdot \xi - A(\theta) + C(\xi)} \tag{4}$$

where $A(\theta) := \log Z(\theta)$ is the log-partition (cumulant) function, and $C(\xi) := \log h(\xi)$ scales the distribution. Both forms are canonical as they are equivalent.

Fitting parameter θ. For a data points $x^* \sim \operatorname{Exp}(\theta)$, we can estimate θ by standard approaches, *e.g.* by maximizing the likelihood function:

$$\begin{aligned} \theta^* = \arg\max_\theta \ell(\theta) &= \arg\max_\theta \left\{ \log \prod_{x^* \in X^\ell} \mathbb{P}[x = x^*|\theta] \right\} \\ &= \arg\max_\theta \sum_{x^* \in X^\ell} \log \left\{ \frac{1}{Z(\theta)} \cdot h(x^*) \cdot e^{\theta \cdot T(x^*)} \right\} \\ &= \arg\max_\theta \sum_{x^* \in X^\ell} \{ -\log Z(\theta) + \log h(x^*) + \theta \cdot T(x^*) \} \to \max_\theta \,. \end{aligned} \tag{5}$$

the terms $\log h(x^*)$ are constant and can be ignored.

Modeling. While 1D exponential family can be used to model 1D densities, relationships between two variables x and y still can be modeled. If we assume that y has an exponential family distribution $y \sim \operatorname{Exp}(\theta)$, and joint distribution is in the form of $f_{X,Y}(x, y|\theta)$:

$$\begin{aligned} f_{X,Y}(x, y|\theta) &= \frac{f_{X,Y,\Theta}(x, y, \theta)}{f_\Theta(\theta)} = \frac{f_Y(y|x, \theta) \cdot f_{X,\Theta}(x, \theta)}{f_\Theta(\theta)} \\ &= \frac{f_Y(y|x, \theta) \cdot f_X(x) \cdot \cancel{f_\Theta(\theta)}}{\cancel{f_\Theta(\theta)}} = f_Y(y|x, \theta). \end{aligned} \tag{6}$$

If $f_Y(y|x, \theta)$ can be expressed as $f_Y(y|\theta(x))$, then $y \sim \operatorname{Exp}(\theta(x))$.

Assumptions. Distributions between X and Θ are independent (inputs do not depend on the parameter), the distribution of X is assumed uniform (data density is constant); as a result, the distribution of answers Y is conditioned both by the input X and the parameter θ of the exponential family.

Exponential family: Canonical form (nD)

Vector parameter $\boldsymbol{\theta} \in \mathbb{R}^m$. The scalar parameter $\theta \in \mathbb{R}$ combines with sufficient statistics $T(\xi) \in \mathbb{R}$ to produce a scalar value $\theta \cdot T(\xi) \in \mathbb{R}$ within the exponential function $e^{\theta \cdot T(\xi)}$.

Generalizing θ to a vector $\boldsymbol{\theta} \in \mathbb{R}^n$ requires only that the inner product $\langle \boldsymbol{\theta}, T(\xi) \rangle$ exists, where $T(\xi) \in \mathbb{R}^m$ maps the random variable Ξ into the same space $\mathbb{R}^m$ where $\boldsymbol{\theta}$ resides:

$$e^{\theta \cdot T(\xi)} \;:\to\; e^{\langle \boldsymbol{\theta}, T(\xi) \rangle}. \tag{7}$$

Parameters $\boldsymbol{\theta}$ are linear, *i.e.* they linearly transform the random vector $\boldsymbol{\xi}$ (or its sufficient statistics $T(\boldsymbol{\xi})$) to produce the scalar value.

Random vector $\boldsymbol{\xi} \in \mathbb{R}^k$. The generalization from scalar random variable Ξ to random vector $\boldsymbol{\xi} \in \mathbb{R}^k$ follows naturally through sufficient statistics $T : \mathbb{R}^k \to \mathbb{R}^m$ that ensures the inner product $\langle \boldsymbol{\theta}, T(\boldsymbol{\xi}) \rangle$ exists.

While dimensions of $\boldsymbol{\xi}$ and $\boldsymbol{\theta}$ need not match, the sufficient statistics T must create a valid inner product $\langle \boldsymbol{\theta}, T(\boldsymbol{\xi}) \rangle$.

Canonical form. A random vector $\boldsymbol{\xi} \in \mathbb{R}^k$ follows the exponential family distribution with parameter $\boldsymbol{\theta} \in \mathbb{R}^m$ when its pdf takes the form:

$$\begin{aligned} f(\boldsymbol{\xi}|\boldsymbol{\theta}) &:= \frac{1}{Z(\boldsymbol{\theta})} \cdot h(\boldsymbol{\xi}) \cdot e^{\langle \boldsymbol{\theta}, T(\boldsymbol{\xi}) \rangle} \\ &:= \exp\Big\{ \langle \boldsymbol{\theta}, T(\boldsymbol{\xi}) \rangle - A(\boldsymbol{\theta}) + C(\boldsymbol{\xi}) \Big\}. \end{aligned} \tag{8}$$

These equivalent canonical forms relate through $A(\boldsymbol{\theta}) = \log Z(\boldsymbol{\theta})$ and $C(\boldsymbol{\xi}) = \log h(\boldsymbol{\xi})$.

The remaining terms generalize naturally:

- Partition function:

$$Z(\boldsymbol{\theta}) := \int_{\mathrm{supp}(\boldsymbol{x})} h(\boldsymbol{\xi}) \cdot e^{\langle \boldsymbol{\theta}, T(\boldsymbol{\xi}) \rangle} \, \mathrm{d}\boldsymbol{\xi}, \tag{9}$$

  where $\mathrm{d}\boldsymbol{\xi} = \mathrm{d}\xi_1 \ldots \mathrm{d}\xi_k$ represents the differential volume element.

- Log partition function:

$$A(\boldsymbol{\theta}) := \log Z(\boldsymbol{\theta}). \tag{10}$$

- Carrier measure and its logarithm must be defined for vector argument:

$$h(\xi) \;:\to\; h(\boldsymbol{\xi}), \qquad C(\xi) \;:\to\; C(\boldsymbol{\xi}). \tag{11}$$

Modeling scalar response $y \in \mathbb{R}$. Given a joint distribution of k-dimensional inputs $\boldsymbol{x}$ and scalar responses y, we can model their relationship analogous to (6):

$$f_{X,Y}(\boldsymbol{x}, y|\boldsymbol{\theta}) = f_Y(y|\boldsymbol{x}, \boldsymbol{\theta}) \tag{12}$$

Note that y is a scalar random variable, with only the parameters being vectors: $y \sim \mathrm{Exp}(\boldsymbol{\theta})$.

For training data $(\boldsymbol{x}^*, y^*) \in (X, Y)^\ell$, we estimate $\boldsymbol{\theta}$ by maximizing:

$$\ell(\boldsymbol{\theta}) = \sum_{(\boldsymbol{x}^*, y^*) \in (X,Y)^\ell} \log f_Y(y = y^* | \boldsymbol{x} = \boldsymbol{x}^*, \boldsymbol{\theta}) \to \max_{\boldsymbol{\theta}}. \tag{13}$$

For prediction on new data $\boldsymbol{x}'$, we calculate the conditional expectation:

$$\hat{y}(\boldsymbol{x}') = \mathbb{E}[y | \boldsymbol{x} = \boldsymbol{x}', \boldsymbol{\theta} = \boldsymbol{\theta}^*]. \tag{14}$$

Modeling all responses $\boldsymbol{y} \in \mathbb{R}^\ell$. A straightforward approach collects all responses $y(\boldsymbol{x})$ for $\boldsymbol{x} \in X^\ell$ into a column vector $\boldsymbol{y} = \begin{pmatrix} y(\boldsymbol{x}_1) & \ldots & y(\boldsymbol{x}_\ell) \end{pmatrix}^\intercal \in \mathbb{R}^\ell$. The notation $\boldsymbol{y} \sim \mathrm{Exp}(\boldsymbol{\theta})$ indicates each training example $y \in \boldsymbol{y}$ shares a common parameter $\boldsymbol{\theta}$.

Modeling a single vector response $\boldsymbol{y} \in \mathbb{R}^m$. For vector-valued responses, each $\boldsymbol{y}$ represents multiple outputs for a single input $\boldsymbol{x} \in \mathbb{R}^k$. The joint distribution follows:

$$f_{X,Y}(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{\theta}) = f_Y(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta}) \tag{15}$$

Vector $\boldsymbol{y}$ can be represented as a vector of all answers $y(\boldsymbol{x})$:

$$\boldsymbol{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_m \end{pmatrix} = \begin{pmatrix} y(\boldsymbol{x}_1) \\ \vdots \\ y(\boldsymbol{x}_m) \end{pmatrix}.$$

Different responses $\boldsymbol{y}$ can be modeled with a shared vector $\boldsymbol{\theta}$ or individual parameters:

$$\boldsymbol{y}_1, \ldots, \boldsymbol{y}_\ell \sim \mathrm{Exp}(\boldsymbol{\theta}_1), \ldots, \mathrm{Exp}(\boldsymbol{\theta}_\ell). \tag{16}$$

Exponential family: Bernoulli distribution

Classical definition. Suppose we have a scenario with two outcomes: "success" and "failure," represented by a binary random variable $\xi \in \{0, 1\}$. The probability of "success" $\mathbb{P}[\xi = 1]$ is defined by a parameter $p \in (0..1)$.

In short, $\xi \sim \mathcal{B}(p)$ means that ξ follows the Bernoulli distribution with parameter p. The probability mass function (pmf) is:

$$\begin{aligned} p(\xi) &= \begin{cases} p & \text{if } \xi = 1 \\ 1-p & \text{if } \xi = 0 \end{cases} \\ &= p^{\xi}(1-p)^{1-\xi}. \end{aligned} \tag{17}$$

The Bernoulli distribution is perhaps the simplest member of the exponential family.

Problem statement. To express the Bernoulli distribution, we need to explicitly identify all components of the exponential family's pdf/pmf: parameter θ, sufficient statistic $T(\xi)$, partition function $Z(\theta)$, and scaling function $h(\xi)$.

Canonical form. Starting by taking the logarithm of the classical definition:

$$\begin{aligned} \log p(\xi) &= \log p^{\xi}(1-p)^{1-\xi} \\ &= \xi \cdot \log p + (1-\xi) \cdot \log(1-p) \\ &= \xi \cdot \log \frac{p}{1-p} + \log(1-p). \end{aligned} \tag{18}$$

Undoing the logarithm, we get:

$$\begin{aligned} p(\xi) &= \exp\left\{\xi \cdot \log \frac{p}{1-p} + \log(1-p)\right\} \\ &= e^{\xi \cdot \log \frac{p}{1-p}} \cdot e^{\log(1-p)} = e^{\xi \cdot \log \frac{p}{1-p}} \cdot (1-p). \end{aligned} \tag{19}$$

Our goal is to demonstrate that this us equivalent to the canonical form (8):

$$f(\xi|\theta) = \frac{1}{Z(\theta)} \cdot h(\xi) \cdot e^{\theta \cdot T(\xi)}$$

By comparing this with the canonical pmf (8), we can easily identify:

$$T(\xi) = \xi, \qquad \theta = \log \frac{p}{1-p}, \qquad \frac{1}{Z(\theta)} \cdot h(\xi) = 1-p. \tag{20}$$

Logit function. The parameter of the exponential family distribution $\operatorname{Exp}(\theta)$ depends on the parameter of the classical Bernoulli distribution $\mathcal{B}(p)$. This connection is established by the *logit function*:

$$\operatorname{logit} p := \log \frac{p}{1-p}. \tag{21}$$

In other words, the canonical parameter can be easily calculated as $\theta = \operatorname{logit} p$. Technically, the logit function maps the probability $p \in (0..1)$ to the arbitrary real number $\theta \in \mathbb{R}$ as the logarithm of $\frac{p}{1-p}$ can be any real number.

The term "logit" is a variation of "logarithm" as it comprises the logarithm function. You can think of it as a portmanteau of "logarithm" and "unit."

The relation of the probability of an event A to the probability of the complementary event is called the odds ratio:

$$\operatorname{odd} A := \frac{\mathbb{P}[A]}{\mathbb{P}\left[\overline{A}\right]} = \frac{\mathbb{P}[A]}{1 - \mathbb{P}[A]}.$$

The logit function is the logarithm of the odds ratio:

$$\operatorname{logit} p := \log \frac{p}{1-p} = \log \frac{\mathbb{P}[\text{«success»}]}{\mathbb{P}[\text{«failure»}]},$$

so it computes the ratio of the probability of success to the probability of failure.

Sigmoid function. Likewise, the classical probability $p \in (0..1)$ can be easily calculated from the canonical parameter $\theta \in \mathbb{R}$ by applying an inverse function to the logit function:

$$p = \operatorname{logit}^{-1} \theta = \sigma(\theta). \tag{22}$$

The commonly known sigmoid function $\sigma(x) = \frac{1}{1+e^{-x}}$ is the inverse of the logit function.

Inverse of logit. The inverse of the logit function is the sigmoid function:

$$\theta = \ln \frac{p}{1-p}$$

$$e^{\theta}(1-p) = p$$

$$p = \frac{e^{\theta}}{1+e^{\theta}} = \frac{1}{1+e^{-\theta}} =: \sigma(\theta)$$

Partition function. The pre-exponential term $\frac{1}{Z(\theta)} \cdot h(\xi)$ is equal to $1-p$, as we have shown. By applying $p = \sigma(\theta)$, we can see that the pre-exponential term depends only on the canonical parameter θ, not on the input ξ, so

$$h(\xi) = 1, \qquad \frac{1}{Z(\theta)} = 1 - \sigma(\theta). \tag{23}$$

Final form. The Bernoulli distribution in canonical exponential family form is:

$$f(\xi|\theta) = (1 - \sigma(\theta)) \cdot e^{\theta \cdot \xi}. \tag{24}$$

Exponential family: Normal distribution

Standard normal distribution. In the trivial case, the standard normal distribution $\mathcal{N}(\mu=0,\sigma=1)$ can be expressed as an exponential family distribution $\mathrm{Exp}(\theta)$:

$$f(\xi|\theta=-1/2)=\frac{1}{\sqrt{2\pi}}\cdot e^{-\xi^2/2}, \tag{25}$$

where the coefficient before ξ^2 is the canonical parameter $\theta=-\frac{1}{2}$, the sufficient statistics are $T(\xi)=\xi^2$, and the pre-exponential term is $\frac{1}{\sqrt{2\pi}}=\frac{1}{Z(\theta)}\cdot h(\xi)$.

Note. The choice of θ, $T(\xi)$, $Z(\theta)$, and $h(\xi)$ is not unique.

Non-standard normal distribution. Interestingly, the non-standard normal distribution $\mathcal{N}(\mu,\sigma)$ cannot be easily fitted into the exponential family. The pdf of the non-standard normal distribution is:

$$\begin{aligned}
f(\xi|\mu,\sigma)&=\frac{1}{\sigma\sqrt{2\pi}}\exp\left\{-\frac{(\xi-\mu)^2}{2\sigma^2}\right\}\\
&=\frac{1}{\sigma\sqrt{2\pi}}\exp\left\{-\frac{1}{2\sigma^2}\cdot\xi^2+\frac{\mu}{\sigma^2}\cdot\xi-\frac{\mu^2}{2\sigma^2}\right\}\\
&=\underbrace{\frac{1}{\sigma\sqrt{2\pi}}\exp\left\{-\frac{\mu^2}{2\sigma^2}\right\}}_{h(\xi)/Z(\boldsymbol{\theta})}\cdot\underbrace{\exp\left\{-\frac{1}{2\sigma^2}\cdot\xi^2+\frac{\mu}{\sigma^2}\cdot\xi\right\}}_{\langle\boldsymbol{\theta},T(\xi)\rangle}.
\end{aligned} \tag{26}$$

✳ The sufficient statistics and canonical parameter are both 2D vectors:

$$T(\xi)=\begin{pmatrix}\xi\\ \xi^2\end{pmatrix},\qquad \boldsymbol{\theta}\equiv\begin{pmatrix}\theta_1\\ \theta_2\end{pmatrix}=\begin{pmatrix}\mu/\sigma^2\\ -1/(2\sigma^2)\end{pmatrix} \tag{27}$$

✳ The parameters of the original distribution μ and σ can be expressed as:

$$\sigma=\sqrt{-\frac{1}{2\theta_2}}=\frac{1}{\sqrt{-2\theta_2}}>0,\qquad \mu=\theta_1\cdot\sigma^2=-\frac{\theta_1}{2\theta_2}. \tag{28}$$

Note. Since $\sigma^2>0$, the canonical parameter $\theta_2=-1/(2\sigma^2)<0$ must be negative; this constrains the parameter space.

✳ The partition function $Z(\theta_1=\mu/\sigma^2,\theta_2=-1/(2\sigma^2))$ depends on two parameters, and the scaling function $h(\xi)=1$ is a constant, so the pre-exponential term is:

$$\begin{aligned}
-\ln Z(\theta_1,\theta_2)&=-\ln\sigma-\ln\sqrt{2\pi}-\frac{\mu^2}{2\sigma^2}\\
&=\ln\sqrt{-2\theta_2}-\ln\sqrt{2\pi}+\frac{\theta_1^2}{4\theta_2}\\
&=\frac{\theta_1^2}{4\theta_2}+\ln\sqrt{\frac{-\theta_2}{\pi}}.
\end{aligned} \tag{29}$$

✳ $-\log Z$ comes from $\log\frac{1}{Z}$, also $h(\xi)=1$

✳ You can check the last by substituting $\theta_1=\frac{\mu}{\sigma^2}$ and $\theta_2=-\frac{1}{2\sigma^2}$ back into the pdf.

Finally, the canonical form of the non-standard normal distribution is:

$$f(\xi|\theta_1,\theta_2)=\exp\left\{\theta_1\cdot\xi-\theta_2\cdot\xi^2+\frac{\theta_1^2}{4\theta_2}+\ln\sqrt{\frac{-\theta_2}{\pi}}\right\}. \tag{30}$$

Multivariate normal distribution. Further generalization is relatively straightforward; the pdf of the multivariate normal distribution $\mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma})$ is:

$$f(\boldsymbol{\xi}|\boldsymbol{\mu},\boldsymbol{\Sigma})=\frac{1}{\sqrt{(2\pi)^k\det\boldsymbol{\Sigma}}}\exp\left\{-\frac{1}{2}(\boldsymbol{\xi}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\xi}-\boldsymbol{\mu})\right\}, \tag{31}$$

where $\boldsymbol{\xi}\in\mathbb{R}^k$ is a random vector, $\boldsymbol{\mu}\in\mathbb{R}^k$ is the mean vector, and $\boldsymbol{\Sigma}\in\mathbb{R}^{k\times k}$ is the covariance matrix, the sufficient statistics, canonical parameters and pre-exponential term are:

$$T(\boldsymbol{\xi})=\begin{pmatrix}\boldsymbol{\xi}\\ \boldsymbol{\xi}\boldsymbol{\xi}^T\end{pmatrix},\quad \boldsymbol{\theta}_1=\Sigma^{-1}\boldsymbol{\mu},\quad \boldsymbol{\theta}_2=-\frac{1}{2}\Sigma^{-1},\quad Z(\boldsymbol{\theta})=\sqrt{(2\pi)^k\det\Sigma}\exp\left\{\frac{1}{2}\boldsymbol{\mu}^T\Sigma^{-1}\boldsymbol{\mu}\right\} \tag{32}$$

Exponential family: Laplace distribution

Classical definition. The Laplace distribution arises naturally as the difference between two independent, identically distributed exponential variables. For this reason, it is also called the double exponential distribution.

The distribution has two parameters: μ is the location parameter (mean), and b is the scale parameter. Its pdf is similar to the normal distribution but has an absolute value in the exponent instead of a square:

$$f(y|\mu, b) = \frac{1}{2b} e^{-|y-\mu|\ /b}. \tag{33}$$

This distribution is useful for modeling data with sharp peaks and heavy tails compared to the normal distribution.

Special case. When $\mu = 0$, the Laplace distribution can be expressed in exponential family form:

$$\begin{aligned} f(y|b) &= \frac{1}{2b} \cdot e^{-|y|/b} \\ &= \underbrace{\frac{1}{2b}}_{1/Z(\theta)} \cdot e^{\overbrace{(1/b)\cdot(-|y|)}^{\theta\cdot T(y)}} \end{aligned} \tag{34}$$

The canonical parameter becomes $\theta = 1/b$, the sufficient statistics $T(y) = -|y|$, and the partition function $Z(\theta) = 2/\theta$.

General case. For $\mu \neq 0$, the Laplace distribution cannot be written as an exponential family distribution because $y, \mu \mapsto |y - \mu|$ cannot be represented as sufficient statistics $T(y)$, which by definition must be independent of distribution parameters.

The classical Laplace distribution parameters behave differently. The parameter b directly relates to the canonical parameter through $\theta = \frac{1}{b}$ in the exponential family form. However, the parameter μ does not correspond to any canonical parameter, making it impossible to express the doubly-parameterized Laplace distribution in exponential family form.

Trick 1: Shifting by μ. By shifting the distribution by μ and introducing a new variable $t := y - \mu$, the distribution of t follows the exponential family form:

$$f(t|\theta) = \frac{\theta}{2} \cdot e^{-|t|\ \cdot\theta}, \quad t := y - \mu. \tag{35}$$

Thus, while the general Laplace distribution itself lies outside the exponential family, the distribution of the shifted variable belongs to it.

Trick 2: Fixing μ. Alternatively, fixing μ to any constant value allows defining sufficient statistics $T(y) := -|y - \mu|$, which expresses the Laplace distribution in exponential family form:

$$f(y|\theta) = \frac{\theta}{2} \cdot e^{T(y)\cdot\theta}, \quad T(y) := -|y - \mu|. \tag{36}$$

Exponential family: Poisson distribution

Classical definition. The Poisson distribution models the number of events occurring within a fixed interval of time (or space). The distribution has a single parameter $\lambda > 0$ representing the average rate of event occurrences.

The pmf of the Poisson distribution is:

$$f(k|\lambda) = e^{-\lambda} \cdot \frac{\lambda^k}{k!}, \tag{37}$$

where $k \in \mathbb{N}_0$ represents the number of events occurring in the interval. This pmf can be rewritten in exponential family form.

$\mathbb{N}_0 := \{0\} \cup \mathbb{N}$ is the set of non-negative integers.

Solution. To express (37) as a one-dimensional exponential family distribution:

$$f(k|\theta) = h(k) \cdot \exp(\theta \cdot k - A(\theta)), \tag{38}$$

we combine all parameter-dependent terms (λ^k and $e^{-\lambda}$) from (37) into a single exponent, and gather all parameter-independent terms ($k!$) into the pre-exponential term:

$$\begin{aligned} f(k|\lambda) &= e^{-\lambda} \cdot \frac{\lambda^k}{k!} \\ &= \frac{1}{k!} \cdot \exp\{-\lambda + \ln \lambda^k\} \\ &= \frac{1}{k!} \cdot \exp\{k \cdot \ln \lambda - \lambda\}. \end{aligned} \tag{39}$$

Comparing terms with the canonical form yields the canonical parameter $\theta = \ln \lambda$, the log partition function $A(\theta) = e^\theta$, and the scaling function $h(k) = \frac{1}{k!}$.

The relationship between classical parameter λ and canonical parameter θ is given by $\lambda = e^\theta$ or equivalently $\theta = \ln \lambda$

The log partition function $A(\theta) = e^\theta$ follows from (39) and the relationship $\lambda = e^\theta$ (see the previous note).

Mean parameter. The expectation follows directly from the derivative of the log partition function:

$$\mu = A'(\theta) = e^\theta, \tag{40}$$

obtained through the formalism of the exponential family.

Classical approach. The same result emerges by directly calculating the expectation using the classical pmf:

$$\begin{aligned} \mathbb{E}[K|\lambda] &:= \sum_{k=0}^{\infty} k \cdot f(k|\lambda) \\ &= \sum_{k=0}^{\infty} k \cdot e^{-\lambda} \cdot \frac{\lambda^k}{k!} \\ &= e^{-\lambda} \cdot \lambda \cdot \sum_{k=1}^{\infty} \frac{\lambda^{k-1}}{(k-1)!} \\ &= e^{-\lambda} \cdot \lambda \cdot \sum_{t=0}^{\infty} \frac{\lambda^t}{t!} \\ &= e^{-\lambda} \cdot \lambda \cdot e^{\lambda} = \lambda. \end{aligned} \tag{41}$$

The exponential function expands as a Taylor series:

$$e^x = \sum_{t=0}^{\infty} \frac{x^t}{t!} = 1 + x + \frac{x^2}{2!} + \ldots$$

The summation index changes twice: first to factor out λ from λ^k, and then through the substitution $t := k - 1$.

As shown, since $\lambda = e^\theta$, the mean parameter $\mu = e^\theta = \lambda$. This demonstrates that the mean parameter of the exponential form directly corresponds to the classical expectation.

GLM: Cross-entropy and log-loss

Model. Logistic regression represents a special case of GLM where the binary response variable Y follows a Bernoulli distribution:

$$y_i \sim \mathcal{B}(p), \quad p := \mathbb{P}[y_i = 1] \tag{42}$$

Here, p represents the success probability in a single trial. The canonical form of the Bernoulli distribution is:

$$y_i \sim f(y|\theta) = \sigma(-\theta) \cdot e^{\theta \cdot y}, \quad \sigma(\theta) = \frac{1}{1+e^{-\theta}} \tag{43}$$

Starting from the general GLM form:

$$Y \sim f(y|\boldsymbol{\theta}) = \exp[\boldsymbol{\theta} \cdot T(y) - A(\boldsymbol{\theta}) + C(y)] \tag{44}$$

We can derive both cross-entropy and log-loss directly, assuming only the Bernoulli distribution of Y.

Link Function. The link function ψ connects the response variable's mean $\mu = \mathbb{E}[Y]$ to the distribution's canonical parameters $\boldsymbol{\theta}$:

$$\boldsymbol{\mu} = \psi(\boldsymbol{\theta}) \tag{45}$$

In GLM, we assume the canonical parameters are linear:

$$\theta_i = \boldsymbol{x}_i^\top \boldsymbol{\beta}, \quad \boldsymbol{\theta} = X\boldsymbol{\beta} \tag{46}$$

where $\boldsymbol{\beta}$ represents the linear coefficients corresponding to features in $\boldsymbol{x}$.

For the Bernoulli distribution, the link function takes the form:

$$\psi(\mu) = \log\frac{\mu}{1-\mu} = \operatorname{logit} \mu \tag{47}$$

Cross-entropy Loss. We begin with the log-likelihood function $l(\theta)$ for the Bernoulli-distributed response variable Y, assuming $\theta = \boldsymbol{x}^\top \boldsymbol{\beta}$:

$$\begin{aligned}
l(\theta) &= \log \prod_i f(y_i|\theta) \\
&= \log \prod_i \sigma(-\theta) \cdot e^{\theta \cdot y_i} \\
&= \sum_i \{\theta \cdot y_i + \log \sigma(-\theta)\} \\
&= \sum_i \left\{\theta \cdot y_i + \log \frac{1}{1+e^{-(-\theta)}}\right\} \\
&= \sum_i \left\{y_i \log\frac{\mu}{1-\mu} + \log\frac{1}{1+\frac{\mu}{1-\mu}}\right\} \\
&= \sum_i \left\{y_i \log\frac{\mu}{1-\mu} + \log\frac{1-\mu}{1-\mu+\mu}\right\} \\
&= \sum_i \{y_i \log \mu - y_i \log(1-\mu) + \log(1-\mu)\} \\
&= \sum_i \{y_i \log \mu + (1-y_i)\log(1-\mu)\} \\
&= \sum_i \{y_i \log p + (1-y_i)\log(1-p)\} \\
&= l(p(\boldsymbol{\beta})) \to \max_{\boldsymbol{\beta}}
\end{aligned} \tag{48}$$

Log-loss. The log-loss function $\ell(M)$ can be derived by taking the negative log-likelihood:

$$\begin{aligned}
-l(\theta) &= -\sum_i \left\{\theta \cdot y_i + \log\frac{e^{-\theta}}{1+e^{-\theta}}\right\} \\
&= \sum_i \begin{cases} -\log e^{\theta} + \log\frac{e^{-\theta}}{1+e^{-\theta}}, \text{if } y = 1 \\ -\log\frac{e^{-\theta}}{1+e^{-\theta}}, \text{if } y = 0 \end{cases} \\
&= \sum_i \begin{cases} \log(1+e^{-\theta}), \text{if } y = 1 \\ \log(1+e^{\theta}), \text{if } y = 0 \end{cases} \\
&= \sum_i \log\left(1+e^{\theta \cdot \operatorname{sgn} y_i}\right) \\
&= \sum_i \log\left(1+e^{\langle \boldsymbol{x}_i, \boldsymbol{\beta}\rangle \cdot \operatorname{sgn} y_i}\right) \\
&= \sum_i \log\left(1+e^{-M_i}\right) \\
&= \ell(M(\boldsymbol{\beta})) \to \min_{\boldsymbol{\beta}}
\end{aligned} \tag{49}$$

Making Predictions. To make a prediction:

$$\hat{p}(\boldsymbol{x}) = \mu(\boldsymbol{x}) = \psi(\theta = \boldsymbol{x} \cdot \boldsymbol{\beta}) = \frac{1}{1 + e^{\boldsymbol{x} \cdot \boldsymbol{\beta}}} \tag{50}$$