CEE 697M: Probabilistic Machine Learning

# M2 Linear Methods: Splines, GAMs and GLMs

**Jimi Oke**

UMass Amherst

College of Engineering

October 7, 2025

# Outline

# The exponential family

A probability distribution belongs to the exponential family if its density can be modeled as:

$$p(\boldsymbol{y}|\boldsymbol{\eta}) = \frac{1}{Z(\boldsymbol{\eta})} h(\boldsymbol{y}) \exp\left[\boldsymbol{\eta}^\top \mathcal{T}(\boldsymbol{y})\right] = h(\boldsymbol{y}) \exp\left[\boldsymbol{\eta}^\top \mathcal{T}(\boldsymbol{y}) - A(\boldsymbol{\eta})\right] \tag{1}$$

where:

- $Z(\boldsymbol{\eta})$ is the partition function (normalization constant)
- $h(\boldsymbol{y})$ is the base measure (scaling constant; typically 1)
- $\boldsymbol{\eta}$ are the natural/canonical parameters
- $\mathcal{T}(\boldsymbol{y})$ are the sufficient statistics
- $A(\boldsymbol{\eta}) = \ln Z(\boldsymbol{\eta})$ is the log-partition function

The log-likelihood is then given by:

$$\log p(\boldsymbol{y}|\boldsymbol{\eta}) = \log h(\boldsymbol{y}) + \boldsymbol{\eta}^\top \mathcal{T}(\boldsymbol{y}) - A(\boldsymbol{\eta}) + \text{const} \tag{2}$$

# Properties of exponential family

- Generalization: we define $\boldsymbol{\eta} = f(\phi)$, thus:

$$p(\boldsymbol{y}|\phi) = h(\boldsymbol{y}) \exp\left[f(\phi)^\top \mathcal{T}(\boldsymbol{y}) - A(f(\phi))\right] \tag{3}$$

- If $f(\phi)$ is nonlinear, then the model is in the curved exponential family
- If $\boldsymbol{\eta} = f(\phi) = \phi$, the model is in **canonical form**
- If $\mathcal{T}(\boldsymbol{y}) = \boldsymbol{y}$, the model is in the natural exponential family

$$p(\boldsymbol{y}|\boldsymbol{\eta}) = h(\boldsymbol{y}) \exp\left[\boldsymbol{\eta}^\top \boldsymbol{y} - A(\boldsymbol{\eta})\right] \tag{4}$$

# Bernoulli distribution in exponential family form (1/2)

The Bernoulli distribution is given by:

$$p(y|\mu) = \mu^y (1-\mu)^{1-y}, \quad y \in \{0,1\}, \quad 0 < \mu < 1 \tag{5}$$

where $\mu = \mathbb{E}(y)$ is the probability of success. Rewriting:

$$\begin{aligned} p(y|\mu) &= (1-\mu)\left(\frac{\mu}{1-\mu}\right)^y = (1-\mu)\exp\left[y\log\left(\frac{\mu}{1-\mu}\right)\right] \\ &= (1-\mu)\exp\left[y\log\left(\frac{\mu}{1-\mu}\right) - 0\right] \end{aligned}$$

Comparing to the exponential family form:

$$\begin{aligned} h(y) &= 1-\mu \quad \text{(base measure)} \\ \mathcal{T}(y) &= y \quad \text{(sufficient statistic)} \\ \eta &= \log\left(\frac{\mu}{1-\mu}\right) \quad \text{(natural parameter)} \\ A(\eta) &= 0 \quad \text{(log-partition function)} \end{aligned}$$

# Cumulant generating function

- Cumulants $\kappa_n(\boldsymbol{y})$ are functions of the central moments of a distribution
- For example, $\kappa_1(\boldsymbol{y}) = \mathbb{E}(\boldsymbol{y})$ and $\kappa_2(\boldsymbol{y}) = \mathbb{V}(\boldsymbol{y})$
- Higher order cumulants are polynomial functions of the central moments
- The cumulants of a distribution are defined by the cumulant generating function (CGF):
$$K_{\boldsymbol{y}}(t) = \log \mathbb{E}(\exp(t\boldsymbol{y})) \tag{6}$$
where $\mathbb{E}(\exp(t\boldsymbol{y}))$ is the moment generating function (MGF) of $\boldsymbol{x}$
- In the exponential family, the log-partition function $A(\boldsymbol{\eta})$ is the CGF of the sufficient statistics $\mathcal{T}(\boldsymbol{y})$
- Thus, the cumulants can be obtained by differentiating $A(\boldsymbol{\eta})$:

$$\begin{aligned}
\kappa_1(\mathcal{T}(\boldsymbol{y})) &= \mathbb{E}(\mathcal{T}(\boldsymbol{y})) = \nabla_{\boldsymbol{\eta}} A(\boldsymbol{\eta}) \\
\kappa_2(\mathcal{T}(\boldsymbol{y})) &= \mathrm{Cov}(\mathcal{T}(\boldsymbol{y})) = \nabla^2_{\boldsymbol{\eta}} A(\boldsymbol{\eta})
\end{aligned}$$

# Unique global maximum of the likelihood

From the CGF properties, we have:

$$\nabla^2_{\boldsymbol{\eta}} A(\boldsymbol{\eta}) = \mathrm{Cov}(\mathcal{T}(\boldsymbol{y})) > 0 \tag{7}$$

This implies that the log-partition function $A(\boldsymbol{\eta})$ is strictly convex. Thus, the log-likehood

$$\log p(\boldsymbol{y}|\boldsymbol{\eta}) = \log h(\boldsymbol{y}) + \boldsymbol{\eta}^\top \mathcal{T}(\boldsymbol{y}) - A(\boldsymbol{\eta}) + \mathrm{const} \tag{8}$$

is guaranteed to have a unique global maximum.

# The generalized linear model (GLM)

- Conventional linear regression models have the form:

$$p(y|\boldsymbol{x}, \boldsymbol{w}) \sim \mathcal{N}(y|\boldsymbol{x}^\top \boldsymbol{w}, \sigma^2) \tag{9}$$

  where
  - $y_n$ is a continuous response
  - $\boldsymbol{x}_n$ is a vector of quantitative and/or qualitative explanatory variables
- Generalized linear models (GLMs) were introduced to extend this framework to allow $y_n$ to be modeled by other exponential family distributions besides the normal/Gaussian, e.g.
  - exponential
  - binomial/multinomial (with fixed number of trials)
  - Poisson
- In the GLM framework:
  - The mean of $y_n$ is given by $\mu_n$
  - $\mu_n$ can be specified by a nonlinear function of $\boldsymbol{x}_n^\top \boldsymbol{w}$
  - Note that the simple linear regression is a special case of GLM in which $\mu_n = \boldsymbol{x}_n^\top \boldsymbol{w}$ and $y_n$ follows a Gaussian distribution

# GLM formulation

The GLM is a version of the exponential family distribution in which the natural parameters $\eta_n$ are a **linear function** of the output.[1] It is given by:

$$p(y_n|\boldsymbol{x}_n, \boldsymbol{w}, \sigma^2) = \exp\left[\frac{y_n\eta_n - A(\eta_n)}{\sigma^2} + \log h(y_n, \sigma^2)\right] \tag{10}$$

where:

- $\eta_n = \boldsymbol{w}^\top \boldsymbol{x}_n$ is the natural parameter (input)
- $y_n = \mathcal{T}(y_n)$ is the sufficient statistic
- $A(\eta_n)$ is the log-partition function (or log normalizer)
- $h(y_n, \sigma^2)$ is the base measure
- $\sigma^2$ is the **dispersion parameter** (typically known or set to 1)

[1] The technical name of the GLM form of the distribution is the **exponential dispersion model/family, often abbreviated as "EDM."**

# Link and mean functions

Recalling that the mean and variance of the sufficient statistics $\mathcal{T}(y_n) = y_n$ are given by the first and second derivatives of the log-partition function $A(\eta_n)$, we have:

$$\mathbb{E}(y_n|\boldsymbol{x}_n, \boldsymbol{w}, \sigma^2) = A'(\eta_n) = \ell^{-1}(\eta_n) \tag{11}$$

$$\mathbb{V}(y_n|\boldsymbol{x}_n, \boldsymbol{w}, \sigma^2) = A''(\eta_n)\sigma^2 \tag{12}$$

We define the **mean function** as

$$\mu_n = \ell^{-1}(\eta_n) \tag{13}$$

and the **link function** as its inverse:

$$g(\mu_n) = \ell(\mu_n) \tag{14}$$

The link function[2] is thus the inverse of the mean function, and its role is to map the mean output/response $\mu_n$ to the linear predictor $\eta_n = \boldsymbol{w}^\top \boldsymbol{x}_n$.

[2] In most textbooks, the link function is denoted as $g(\mu)$, while Murphy uses $\ell(\mu)$.

# Linear regression (1/2)

Linear regression has the form:

$$p(y_n|\boldsymbol{x}_n, \boldsymbol{w}, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(y_n - \boldsymbol{w}^\top \boldsymbol{x}_n)^2\right) \tag{15}$$

Taking logs:

$$\log p(y_n|\boldsymbol{x}_n, \boldsymbol{w}, \sigma^2) = -\frac{1}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}(y_n - \boldsymbol{w}^\top \boldsymbol{x}_n)^2 \tag{16}$$

Setting $\eta_n = \boldsymbol{w}^\top \boldsymbol{x}_n$, we can write in GLM form as:

$$\log p(y_n|\boldsymbol{x}_n, \boldsymbol{w}, \sigma^2) = \frac{y_n\eta_n - \eta_n^2/2}{\sigma^2} - \frac{1}{2}\left(\frac{y_n^2}{\sigma^2} + \log(2\pi\sigma^2)\right) \tag{17}$$

# Linear regression (2/2)

If we set:

$$A(\eta_n) = \eta_n^2/2 \tag{18}$$

$$h(y_n, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2} y_n^2\right) \tag{19}$$

then we can write:

$$\log p(y_n|\boldsymbol{x}_n, \boldsymbol{w}, \sigma^2) = \frac{y_n\eta_n - A(\eta_n)}{\sigma^2} + \log h(y_n, \sigma^2) \tag{20}$$

And thus, the cumulants are given by:

$$\mathbb{E}(y_n|\boldsymbol{x}_n, \boldsymbol{w}, \sigma^2) = A'(\eta_n) = \eta_n = \boldsymbol{w}^\top \boldsymbol{x}_n \tag{21}$$

$$\mathrm{Var}(y_n|\boldsymbol{x}_n, \boldsymbol{w}, \sigma^2) = A''(\eta_n)\sigma^2 = \sigma^2 \tag{22}$$

Thus, the mean function is $\mu_n = \ell^{-1}(\eta_n) = \eta_n$ and the link function is $g(\mu_n) = \ell(\mu_n) = \mu_n$ (identity link function).

# GLM components

A GLM can be considered as consisting of three parts:

- **Random component:** this is the probability distribution of the response variable $p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{w})$
- **Systematic component:** specifies the explanatory variables within the linear combination of their coefficients ($\boldsymbol{Xw}$)
- **Link function $\boldsymbol{g(\mu)}$:** defines the relationship between the random and systematic components:
  - Simple linear regression (**identity** link function):

$$g(\mu_n) = g(\mathbb{E}(y_n)) = \boldsymbol{w}^\top \boldsymbol{x}_n \tag{23}$$

  - Binary logistic regression (**logit** link function):

$$g(\mu_n) = g(p(\boldsymbol{x}_n)) = \text{logit}(p(\boldsymbol{x}_n)) = \ln\left(\frac{p(\boldsymbol{x}_n)}{1 - p(\boldsymbol{x}_n)}\right) = \boldsymbol{w}^\top \boldsymbol{x}_n \tag{24}$$

# Assumptions of GLM

- The observations of the response variable $\boldsymbol{y}$ are i.i.d.
- Response variable $y_n$ is typically exponentially distributed (not restricted to being normally distributed)
  - Implies that errors need not be normally distributed (but should be independent)
- Link function ($g(\mu_n)$) is linear with respect to the coefficients ($w_d$)
  - Relationship between response and explanatory variables does not have to be linear
  - Explanatory variables can be nonlinear transformations of original values (as in simple linear regression)
- Variance may not homogeneous (i.e. homoscedasticity is not a requirement)
- Parameters are estimated via MLE

# Commonly used GLM models and their components

| Model | Random component | Mean/output ($\mu_n$) | Link function |
|---|---|---|---|
| Linear regression | Gaussian | $\boldsymbol{w}^\top \boldsymbol{x}_n$ | Identity: $g(\mu_n) = \mu_n = \boldsymbol{w}^\top \boldsymbol{x}_n$ |
| Binary logistic regression | Bernoulli | $\boldsymbol{\sigma}(\boldsymbol{w}^\top \boldsymbol{x}_n)$ | Logit: $g(\mu_n) = \log\left(\frac{\mu_n}{1-\mu_n}\right)$ |
| Probit regression | Bernoulli | $\boldsymbol{\sigma}(\boldsymbol{w}^\top \boldsymbol{x}_n)$ | Probit: $g(\mu_n) = \Phi^{-1}(\mu_n)$ |
| Multinomial logit/logistic | Categorical | $\mathcal{S}(\boldsymbol{W}\boldsymbol{x}_n)$ | Multinomial logit: $g(\mu_{nc}) = \log\left(\frac{\mu_{nc}}{\mu_{nC}}\right)$ |
| Poisson regression | Poisson | $\exp(\boldsymbol{w}^\top \boldsymbol{x}_n)$ | Log: $g(\mu_n) = \log(\mu_n)$ |

Note that in all cases, the link function always results in:

$$g(\mu_n) = \boldsymbol{w}^\top \boldsymbol{x}_n = \eta_n \tag{25}$$

Its job is to "link" the response to the systematic component via a suitable transformation that results in a linear function of the $w$'s.

# Canonical and non-canonical link functions

Link functions can be classified as either canonical or non-canonical:

- Canonical link function: results in the canonical/natural parameters $\theta$ of the random component (i.e. $\eta = g(\mu) = \theta$)
    - **identity** (linear regression): $g(\mu) = \mu$; $\mu$ is the canonical parameter of the Gaussian distribution
    - **logit** (binary logistic regression): $g(\mu) = \log\left(\frac{\mu}{1-\mu}\right)$; the logit is the canonical parameter of the Bernoulli distribution
    - **log** (Poisson regression): $g(\mu) = \log(\mu)$; $\log(\mu)$ is the canonical parameter of the Poisson distribution
- Non-canonical link function: does not result in natural parameter of the random component/underlying distribution (i.e. $g(\mu_) \neq \mu_n$). Examples:
    - **probit** (binary probit regression): $g(\mu) = \Phi^{-1}(\mu)$; the probit is not the canonical parameter of the Bernoulli distribution
    - **complementary log-log** (binary regression): $g(\mu) = \log(-\log(1-\mu))$; the complementary log-log is not the canonical parameter of the Bernoulli distribution

# Canonical link function (another view)

Let $\theta$ be the natural parameter, $\eta$ the linear predictor, and $\mu$ the mean of the response variable. Recall:

$$\begin{aligned} \mu &= \mathbb{E}(y|\theta) = A'(\theta) && (26) \\ \eta &= \boldsymbol{w}^\top \boldsymbol{x} = g(\mu) = g(A'(\theta)) && (27) \end{aligned}$$

If the link function is canonical, then:

$$g(\mu) = (A')^{-1}(\mu) = \theta \tag{28}$$

If not:

$$g(\mu) = (A')^{-1}(\mu) \neq \theta \tag{29}$$

# Usage of canonical vs non-canonical link functions

- Canonical link functions are simpler to estimate and have desirable statistical properties.
- Non-canonical link functions may be used when the canonical link does not provide a good fit to the data or when interpretability of the model is a priority.
- In certain cases (e.g. probit regression), the non-canonical link (inverse CDF) allows for more efficient computation of the likelihood (e.g. Gibbs sampling)

# Fitting a GLM

The process of fitting a GLM involves the following steps:

- Specify the distribution of the response variable (e.g. Gaussian, Bernoulli, Poisson)
- Choose a link function (canonical or non-canonical)
- Estimate the model parameters (e.g. using MLE)
- Assess the model fit (e.g. using residuals, AIC/BIC)

# MLE of GLM parameters

The negative log-likelihood (ignoring constant terms) is given by

$$\mathrm{NLL}(\boldsymbol{w}) = -\log p(\mathcal{D}|\boldsymbol{w}) = -\sum_{n=1}^{N} \log p(y_n|\boldsymbol{x}_n, \boldsymbol{w}) = \sum_{n=1}^{N} \frac{A(\eta_n)}{\sigma^2} - \frac{y_n \eta_n}{\sigma^2} \qquad (30)$$

If we set $\mathcal{L}_n = \eta_n y_n - A(\eta_n)$, then the NLL can be written as:

$$\mathrm{NLL}(\boldsymbol{w}) = -\sum_{n=1}^{N} \frac{\mathcal{L}_n}{\sigma^2} \qquad (31)$$

where $\eta_n = \boldsymbol{w}^\top \boldsymbol{x}_n$.
The gradient of the NLL (for a single term) is then given by:

$$\boldsymbol{g}_n = \frac{y_n - \mu_n}{\sigma^2} \boldsymbol{x}_n \qquad (32)$$

where $\mu_n = A'(\eta_n) = \ell^{-1}(\eta_n)$ is the mean function.

# Summary

- In the exponential family, the log-partition function $A(\theta)$ is the cumulant generating function of the sufficient statistics $\mathcal{T}(\boldsymbol{y})$
- The log-partition function is strictly convex, thus the likelihood has a unique global maximum
- GLMs are a flexible class of models that extend linear regression to handle non-normal response variables
- Link functions $g(\mu)$ map the mean of the response variable to the linear predictor $\eta = \boldsymbol{w}^\top \boldsymbol{x}$
- Canonical link functions result in the natural parameters $\theta$ of the underlying distribution, while non-canonical link functions do not
- GLM parameters $\eta$ can be estimated via MLE using gradient-based optimization methods