# Learning From Data
# Lecture 2 (Part 2): Generalized Linear Models

**Yang Li** **yangli@sz.tsinghua.edu.cn**

September 19, 2024

# Outline

Generalized Linear Models (GLM)

- Review: Exponential Families
- GLM Construction and Examples

# Summary: Linear models

What we've learned so far:

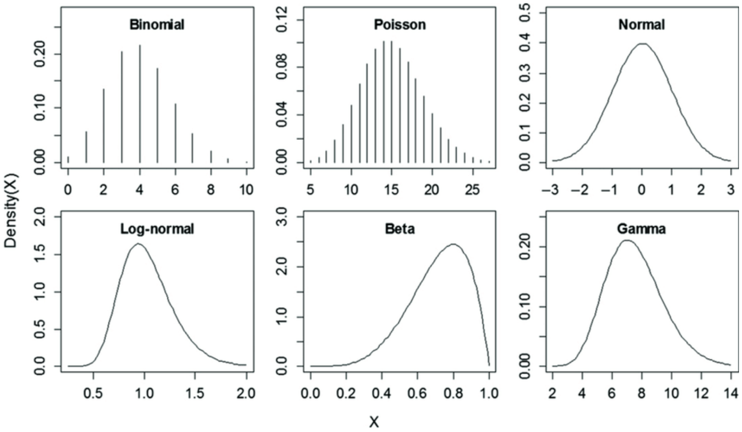
| Learning task | Model | $p(y\|x;\theta)$ |
|---|---|---|
| regression | Linear regression | $\mathcal{N}(h_\theta(x)\ ,\sigma^2)$ |
| binary classification | Logistic regression | Bernoulli( $h_\theta(x)$ ) |
| multi-class classification | Softmax regression | Multinomial($[h_\theta(x)]$ ) |

*Can we generalize the linear model to other distributions?*

**Generalized Linear Model (GLM)**: a recipe for constructing linear models in which $y|x;\theta$ is from an **exponential family**.

# Review: Exponential Family

# Exponential Family of Distributions




Examples of distribution classes in the exponential family.

# Exponential Family of Distributions

A class of distributions is in the **exponential family** if its density can be written in the *canonical form*:

$$p(y;\eta) = b(y)e^{\eta^T T(y) - a(\eta)}$$

- $y$: random variable
- $\eta$ : natural/canonical parameter (that depends on distribution parameter(s))
- $T(y)$: sufficient statistic of the distribution
- $b(y)$: a function of $y$
- $a(\eta)$ : log partition function (or "cumulant function")

# Exponential Family

**Log partition function** $a(\eta)$ is the log of a normalizing constant. i.e.

$$p(y;\eta) = b(y)e^{\eta^T T(y) - a(\eta)} = \frac{b(y)e^{\eta^T T(y)}}{e^{a(\eta)}}$$

Function $a(\eta)$ is chosen such that $\sum_y p(y;\eta) = 1$ (or $\int_y p(y;\eta)dy = 1$).

$$a(\eta) = \log\left(\sum_y b(y)e^{\eta^T T(y)}\right)$$

# Exponential Family Examples

## Gaussian Distribution (unit variance)

Probability density of a Gaussian distribution $\mathcal{N}(\mu, 1)$ over $y \in \mathbb{R}$:

$$p(y; \theta) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(y-\mu)^2}{2}\right)$$

- $\eta = \mu$
- $b(y) = \frac{1}{\sqrt{2\pi}} \exp(-y^2/2)$
- $T(y) = y$
- $a(\eta) = \frac{1}{2}\eta^2$

# Exponential Family Examples

Two parameter example:

## Gaussian Distribution

Probability density of a Gaussian distribution $\mathcal{N}(\mu, \sigma^2)$ over $y \in \mathbb{R}$:

$$p(y;\theta) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right)$$

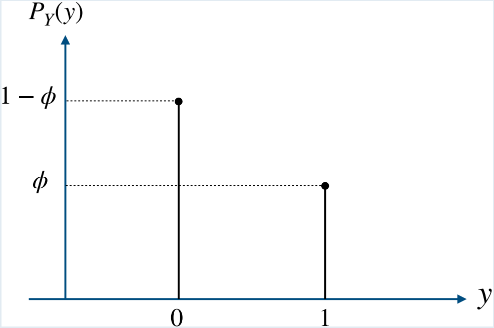
- $\eta = \begin{bmatrix} \frac{\mu}{\sigma^2} \\ -\frac{1}{2\sigma^2} \end{bmatrix}$
- $b(y) = \frac{1}{\sqrt{2\pi}}$
- $T(y) = \begin{bmatrix} y \\ y^2 \end{bmatrix}$
- $a(\eta) = \frac{\mu^2}{2\sigma^2} + \log\sigma$

# Exponential Family Examples

## Bernoulli Distribution

Bernoulli($\phi$): a distribution over $y \in \{0, 1\}$, such that

$$p(y; \phi) = \phi^y (1 - \phi)^{1-y}$$



## Bernoulli Distribution

Bernoulli($\phi$): a distribution over $y \in \{0, 1\}$, such that

$$p(y; \phi) = \phi^y (1 - \phi)^{1-y}$$

*How to write it in the form of $p(y; \eta) = b(y) e^{\eta^T T(y) - a(\eta)}$?*

# Exponential Family Examples

## Bernoulli Distribution

Bernoulli($\phi$): a distribution over $y \in \{0, 1\}$, such that

$$p(y; \phi) = \phi^y (1 - \phi)^{1-y}$$

- $\eta = \log\left(\frac{\phi}{1-\phi}\right)$
- $b(y) = 1$
- $T(y) = y$
- $a(\eta) = \log(1 + e^\eta)$

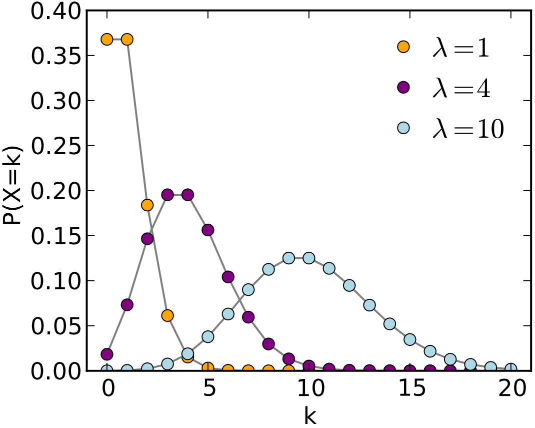
# Exponential Family Examples

**Poisson distribution:** Poisson($\lambda$)

Models the probability that an event occurring $y \in \mathbb{N}$ times in a fixed interval of time, *assuming events occur independently at a constant rate*

Probability density function of Poisson($\lambda$) over $y \in \mathcal{Y}$:

$$p(y; \lambda) = \frac{\lambda^y e^{-\lambda}}{y!}$$



# Exponential Family Examples

**Poisson distribution** Poisson($\lambda$)

Probability density function of Poisson($\lambda$) over $y \in \mathcal{Y}$:

$$p(y; \lambda) = \frac{\lambda^y e^{-\lambda}}{y!}$$

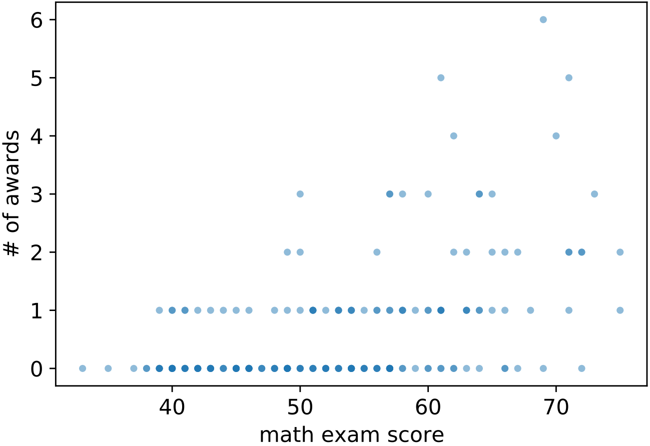
- $\eta =$
- $b(y) =$
- $T(y) =$
- $a(\eta) =$

# Generalized Linear Models

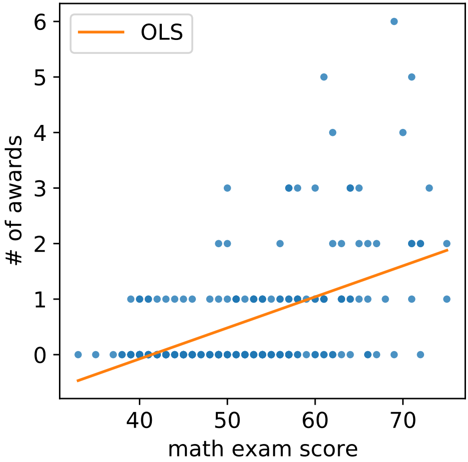
# Generalized Linear Models: Intuition

### Example 1: Award Prediction

Predict $y$, **the number of school awards** a student gets given $x$, the math exam score.



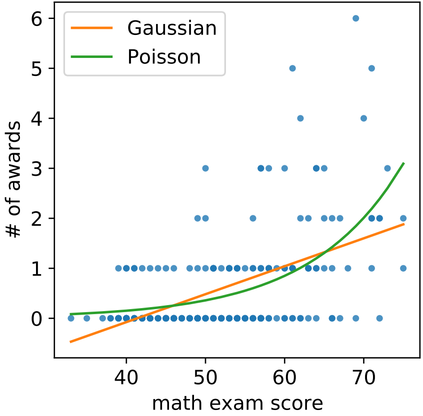
# Generalized Linear Models: Intuition




Problems with linear regression:

- Assumes $y|x;\theta$ has a Normal distribution.
- Assumes change in $x$ is proportional to change in $y$

# Generalized Linear Models: Intuition




Problems with linear regression:

- Assumes $y|x;\theta$ has a Normal distribution. ***Poisson** distribution is better for modeling occurrences*
- Assumes change in $x$ is proportional to change in $y$ *More realistic to be proportional to the **rate** of increase in y* (e.g. doubling or halving $y$)

# Generalized Linear Models : Intuition

**Generalized Linear Model (GLM)**: a recipe for constructing linear models in which $y|x;\theta$ is from an exponential family.

Design motivation of GLM

- We can select a distribution for **Response variables** $y$
- Allow (the **canonical link function** of $y$) to vary linearly with the input values $x$

e.g. $log(\lambda) = \theta^T x$

Nelder, John Ashworth, and Robert William Maclagan Wedderburn. 1972. Generalized Linear Models. Journal of the Royal Statistical Society. Series A (General) 135 (3): 37084.

# Generalized Linear Models: Construction

Formal GLM assumptions & design decisions:

1. $y|x;\theta \sim \text{ExponentialFamily}(\eta)$
   e.g. Gaussian, Poisson, Bernoulli, Multinomial, Beta ...
2. The hypothesis function $h(x)$ is $\mathbb{E}\left[T(y)|x\right]$
   e.g. When $T(y) = y$, $h(x) = \mathbb{E}\left[y|x\right]$
3. The natural parameter $\eta$ and the inputs $x$ are related linearly:

   **$\eta$ is a number:**

$$\eta = \theta^T x$$

   **$\eta$ is a vector:**

$$\eta_i = \theta_i^T x \quad \forall i = 1, \ldots, n \quad \text{or} \quad \eta = \Theta^T x$$

# Generalized Linear Models: Construction

Relate natural parameter $\eta$ to distribution mean $\mathbb{E}[T(y)|x]$ :

- **Canonical response function** $g$ gives the mean of the distribution

$$g(\eta) = \mathbb{E}[T(y)|x]$$

  a.k.a. the "mean function"

- $g^{-1}$ is called the **canonical link function**

$$\eta = g^{-1}(\mathbb{E}[T(y)|x])$$

# GLM example: ordinary least square

Apply GLM construction rules:

1. Let $y|x;\theta \sim N(\mu, 1)$

$$\eta = \mu, \; T(y) = y$$

2. Derive hypothesis function:

$$\begin{aligned} h_\theta(x) &= \mathbb{E}\left[T(y)|x;\theta\right] \\ &= \mathbb{E}\left[y|x;\theta\right] \\ &= \mu = \eta \end{aligned}$$

3. Adopt linear model $\eta = \theta^T x$:

$$h_\theta(x) = \eta = \theta^T x$$

Canonical response function: $\mu = g(\eta) = \eta$ (identity)
Canonical link function: $\eta = g^{-1}(\mu) = \mu$ (identity)

# GLM example: logistic regression

Apply GLM construction rules:

1. Let $y|x;\theta \sim \text{Bernoulli}(\phi)$

$$\eta = \log\left(\frac{\phi}{1-\phi}\right),\ T(y) = y$$

2. Derive hypothesis function:

$$\begin{aligned} h_\theta(x) &= \mathbb{E}\left[T(y)|x;\theta\right] \\ &= \mathbb{E}\left[y|x;\theta\right] \\ &= \phi = \frac{1}{1+e^{-\eta}} \end{aligned}$$

3. Adopt linear model $\eta = \theta^T x$:

$$h_\theta(x) = \frac{1}{1+e^{-\theta^T x}}$$

Canonical response function: $\phi = g(\eta) = \text{sigmoid}(\eta)$
Canonical link function : $\eta = g^{-1}(\phi) = \text{logit}(\phi)$

# GLM example: Poisson regression (Exercise)

**Example 1: Award Prediction**

Predict $y$, **the number of school awards** a student gets given $x$, the math exam score.

Use GLM to find the hypothesis function...

# GLM example: Poisson regression (Exercise)

Apply GLM construction rules:

1. Let $y|x;\theta \sim \text{Poisson}(\lambda)$

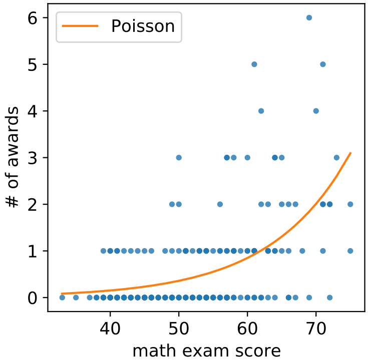
$$\eta =, \ T(y) =$$

2. Derive hypothesis function:

$$h_\theta(x) = \mathbb{E}\left[T(y)|x;\theta\right]$$

3. Adopt linear model $\eta = \theta^T x$:

$$h_\theta(x) = g(\eta) =$$




Canonical response function: $\lambda = g(\eta) =$
Canonical link function : $\eta = g^{-1}(\lambda) =$

# GLM example: Softmax regression

Probability mass function of a Multinomial distribution over $k$ outcomes

$$p(y;\phi) = \prod_{i=1}^{k} \phi_i^{\mathbf{1}\{y=i\}}$$

Derive the exponential family form of Multinomial$(\phi_1, .., \phi_k)$: Note: $\phi_k = 1 - \sum_{i=1}^{k-1} \phi_i$ is not a parameter

- $T(y) = \begin{bmatrix} \mathbf{1}\{y=1\} \\ \vdots \\ \mathbf{1}\{y=k-1\} \end{bmatrix}$

  $T(y)_i = \mathbf{1}\{y=i\} = \begin{cases} 0 & y \neq i \\ 1 & y = i \end{cases}$
- $a(\eta) = -\log(\phi_k) = \log \sum_{i=1}^{k} e^{\eta_i}$
- $\eta = \begin{bmatrix} \log\left(\frac{\phi_1}{\phi_k}\right) \\ \vdots \\ \log\left(\frac{\phi_{k-1}}{\phi_k}\right) \end{bmatrix}$
- $b(y) = 1$

# GLM example: Softmax regression

Apply GLM construction rules:

1. Let $y|x;\theta \sim \text{Multinomial}(\phi_1, \ldots, \phi_k)$, for all $i = 1 \ldots k-1$

$$\eta_i = \log\left(\frac{\phi_i}{\phi_k}\right), \; T(y) = \begin{bmatrix} \mathbf{1}\{y=1\} \\ \vdots \\ \mathbf{1}\{y=k-1\} \end{bmatrix}$$

Compute inverse: $\phi_i = \frac{e^{\eta_i}}{\sum_{j=1}^{k} e^{\eta_j}} \leftarrow$ *canonical response function*

2. Derive hypothesis function:

$$h_\theta(x) = \mathbb{E}\left[\begin{bmatrix} \mathbf{1}\{y=1\} \\ \vdots \\ \mathbf{1}\{y=k-1\} \end{bmatrix} \middle| x;\theta \right] = \begin{bmatrix} \phi_1 \\ \vdots \\ \phi_{k-1} \end{bmatrix}$$

$$\phi_i = \frac{e^{\eta_i}}{\sum_{j=1}^{k} e^{\eta_j}}$$

# GLM example: Softmax regression

**3.** Adopt linear model $\eta_i = \theta_i^T x$:

$$\phi_i = \frac{e^{\theta_i^T x}}{\sum_{j=1}^k e^{\theta_j^T x}} \text{ for all } i = 1 \dots k-1$$

$$h_\theta(x) = \frac{1}{\sum_{j=1}^k e^{\theta_j^T x}} \begin{bmatrix} e^{\theta_1^T x} \\ \vdots \\ e^{\theta_{k-1}^T x} \end{bmatrix}$$

Canonical response function: $\phi_i = g(\eta) = \dfrac{e^{\eta_i}}{\sum_{j=1}^k e^{\eta_j}}$

Canonical link function : $\eta_i = g^{-1}(\phi_i) = \log\left(\dfrac{\phi_i}{\phi_k}\right)$

# GLM Summary

**Sufficient statistic** $T(y)$

**Response function** $g(\eta)$

**Link function** $g^{-1}(\mathbb{E}[T(y);\eta])$

| Exponential Family | $\mathcal{Y}$ | $T(y)$ | $g(\eta)$ | $g^{-1}(\mathbb{E}[T(y);\eta])$ |
|---|---|---|---|---|
| $\mathcal{N}(\mu,1)$ | $\mathbb{R}$ | $y$ | $\eta$ | $\mu$ |
| Bernoulli($\phi$) | $\{0,1\}$ | $y$ | $\frac{1}{1+e^{-\eta}}$ | $\log\frac{\phi}{1-\phi}$ |
| Multinomial($\phi_1,\ldots,\phi_k$) | $\{1,\ldots,k\}$ | $\mathbf{1}\{y=i\}$ | $\frac{e^{\eta_i}}{\sum_{j=1}^k e^{\eta_j}}$ | $\eta_i = \log\left(\frac{\phi_i}{\phi_k}\right)$ |

GLM is effective for modelling different types of distributions over $y$