MATH 7339 - Machine Learning and Statistical Learning Theory 2 –He Wang

Section Generalized Linear Models

1. Linear Models Components
2. Exponential Family
3. Construction of GLM
4. Maximum Likelihood Estimation

Review: Linear Regression and Logistic Regression

So far, we've seen two canonical settings for regression. Let $\vec{X} \in \mathbb{R}^d$ be a vector of predictors.

In **linear regression**, we observe $y \in \mathbb{R}$, and assume a linear model:

$Y = \vec{\beta}^T \vec{X} + \epsilon$ with $\epsilon \sim Normal(0, \sigma^2)$ Then, $E(Y|\vec{X}) = \vec{\beta}^T \vec{X}$

In **logistic regression**, we observe $y \in \{0,1\}$, and we assume a logistic model

$$\log\left(\frac{P(Y = 1|\vec{X})}{1 - P(Y = 1|\vec{X})}\right) = \vec{\beta}^T \vec{X}$$

In both settings, we are assuming that a **transformation(link)** $g(u)$ of the conditional expectation $E(Y|\vec{X})$ is a linear function of $\vec{X}$, i.e.,

$$g\left(E(Y|\vec{X})\right) = \vec{\beta}^T \vec{X}$$

In linear regression, this transformation was the identity transformation

$$g(u) = u;$$

In logistic regression, it was the logit transformation

$$g(u) = \log\left(\frac{u}{1-u}\right)$$

Different transformations might be appropriate for different types of data.

For a third data type, it is entirely possible that transformation neither is really appropriate. We think of another transformation.

❑ Components of a linear regression model

The two components (that we are going to relax) are

1. Random component: the response variable

$$Y|\vec{X} \sim Normal(\mu, \sigma^2)$$

is **continuous** and **normally** distributed with mean $\mu = \mu(\vec{X}) = E(Y|\vec{X})$.

2. Link(transformation): between the random and covariates $\vec{X}$.

$$\mu(\vec{X}) = \vec{\theta}^T \vec{X}$$

❑ Generalization (first view)

A generalized linear model (GLM) generalizes normal linear regression models in the following directions.

1. Random component:

$$Y|\vec{X} \sim \text{some distribution}$$

In principle, we could specify any distribution. But mathematics of GLM only works nicely for exponential family of distributions.

2. Link(transformation): between the random and covariates $\vec{X}$:

$$g\left(\mu(\vec{X})\right) = \vec{\beta}^T\vec{X}$$

where g called link function and $\mu(\vec{X}) = E\left(Y\middle|\vec{X}\right)$.

The random component specifies a distribution for the outcome variable (conditional on $\vec{X}$).

In the case of linear regression, we assume that

$$Y|\vec{X} \sim Normal(\mu, \sigma^2)$$

for some mean μ and variance σ^2

In the case of logistic regression, we assume that

$$Y|\vec{X} \sim Bernoulli(p)$$

for some probability p.

In a generalized model, we are allowed to assume that $Y|\vec{X} \sim$ a distribution from exponential family.

➢ Exponential Family

Exponential family comprises a set of flexible distribution ranging both continuous and discrete random variables. The members of this family have many important properties which merits discussing them in some general format. Most of the commonly used statistical distributions are members of the **exponential family** of distributions.

- ***Gaussian:*** $\mathbb{R}^p$
- ***Bernoulli: binary*** $\{0, 1\}$
- ***Binomial: counts of success/failure***
- ***Multinomial: categorical***
- ***Poisson:*** $\mathbb{N}^+$
- ***Exponential:*** $\mathbb{R}^+$
- ***Gamma:*** $\mathbb{R}^+$
- ***Laplace:*** $\mathbb{R}^+$
- ***Beta:*** $(0, 1)$
- *Von mises: sphere*
- *Dirichlet:* Δ *(Simplex)*
- *Weibull:* $\mathbb{R}^+$
- *Weishart: symmetric positive-definite matrices*

A number of common distributions are exponential families, but **only** when certain parameters are fixed and known. For example:

- Binomial (with fixed number of trials)
- Multinomial (with fixed number of trials)
- Negative binomial (with fixed number of failures)

Examples of common distributions that are **not** exponential families

- **Student's t**,
- most **mixture distributions**,
- and even the family of **uniform distributions** when the bounds are not fixed.

Check Wikipedia for each of the distributions.

https://en.wikipedia.org/wiki/Exponential_family

Definition(Exponential Family):

A pdf/pmf of a distribution in d-parameters **exponential family** densities is in the form

$$p(\vec{y};\vec{\eta}) = \frac{1}{Z(\vec{\eta})} h(\vec{y}) \exp[\vec{\eta}^T T(\vec{y})]$$

$$= h(\vec{y}) \exp[\vec{\eta}^T T(\vec{y}) - A(\vec{\eta})]$$

Here,

- $\vec{\eta} \in \mathbb{R}^d$ is the **natural parameter** of the distribution.
- $T(\vec{y}) \in \mathbb{R}^d$ is a vector of **sufficient statistics**. In many cases, $T(\vec{y}) = \vec{y}$, then the distribution is said to be in canonical form.
- $h(\vec{y})$ is the is the "**underlying/base measure**", in many cases, $h(\vec{y}) = 1$.
- $A(\vec{\eta}) = \log \mathrm{Z}(\vec{\eta})$ is called the **log partition function/log normalizer.** $A(\vec{\eta})$ is the normalization constant, to make sure the total probability is 1.

Hence,

$$A(\vec{\eta}) := \log \int h(y) \exp\left(\vec{\eta}^T T(\vec{y})\right) dy$$

Other formats of pdf/pmf of exponential family:

$$p(\vec{y}; \vec{\eta}) = h(\vec{y}) \exp[\vec{\eta}^T T(\vec{y}) - A(\vec{\eta})]$$

$$= \exp[\vec{\eta}^T T(\vec{y}) - A(\vec{\eta}) + C(\vec{y})]$$

where $C(\vec{y}) = \log h(\vec{y})$

Sometimes, for GLM construction, we also introduce an extra scale parameter ϕ, called the **dispersion** parameter, to control the shape of $p(\vec{y}\ )$

$$p(\vec{y}; \vec{\eta}, \phi) = \exp\left[\frac{\vec{\eta}^T T(\vec{y}) - A(\vec{\eta})}{\phi} + C(\vec{y}, \phi)\right]$$

This format is better for constructing generalized linear models.

In general, the parameter $\vec{\eta}$ is not the mean of the distribution. We can view $\vec{\eta}$ as a function of the mean $\vec{\mu} = E(\vec{y})$ and write $\vec{\eta} = g(\vec{\mu})$, which is called the **link function.** The inverse $\vec{\mu} = g^{-1}(\vec{\eta})$ is called the **response function**.

A fixed choice of A, h and ϕ defines a family (or set) of distributions that is parameterized by $\vec{\eta}$; as we vary $\vec{\eta}$, we then get different distributions within this family.

An even more general form is

$$p(\vec{y}; \vec{\eta}) = h(\vec{y}) \exp\left[[f(\vec{\eta})]^T T(\vec{y}) - A\big(f(\vec{\eta})\big)\right]$$

Example: (Bernoulli)

The pdf function of Bernoulli distribution:

$$p(y;\mu) = Bern(y;\mu) = \mu^{y}(1-\mu)^{1-y} \text{ for } y \in \{0,1\}$$

We can write it as

$$p(y;\mu) = \exp(y\log(\mu) + (1-y)\log(1-\mu))$$

$$= \exp\left(y\log\left(\frac{\mu}{1-\mu}\right) + \log(1-\mu)\right)$$

Compare to $p(y;\eta) = \ h(\vec{y})\exp[\vec{\eta}^{T}T(\vec{y}) - A(\vec{\eta})]$, we have

$$T(y) = y \ \text{ and } h(\vec{y}) = 1$$

Canonical Link function: $\eta = \log\left(\frac{\mu}{1-\mu}\right) \implies \mu = \frac{1}{1+e^{-\eta}}$

$$A(\vec{\eta}) = -\log(1-\mu) = \log(1+e^{\eta})$$

Example: (Categorical/Multinomial)

Categorical has a vector of parameters ϕ_k where k goes from 1 to K.

$$p(y;\vec{\phi}) = \phi_1^{\mathbb{I}(\mathrm{y}=1)}\phi_2^{\mathbb{I}(\mathrm{y}=2)} \dots \phi_K^{\mathbb{I}(\mathrm{y}=\mathrm{K})} = \phi_1^{\mathbb{I}(\mathrm{y}=1)}\phi_2^{\mathbb{I}(\mathrm{y}=2)} \dots \phi_K^{\mathbb{I}(\mathrm{y}=\mathrm{K})}$$

$$= \exp\left(\sum_{i=1}^{K-1} \mathbb{I}(\mathrm{y}=\mathrm{i})\log(\phi_i) + \left(1 - \sum_{i=1}^{K-1}\big(\mathbb{I}(\mathrm{y}=\mathrm{i})\big)\right)\log(\phi_k)\right)$$

$$= \exp\left(\sum_{i=1}^{K-1} \mathbb{I}(\mathrm{y}=\mathrm{i})\log(\phi_i/\phi_k) \ +\log(\phi_k)\right)$$

To express the multinomial as an exponential family distribution, define

$T(i) = \vec{e}_i \in \mathbb{R}^{k-1}$ and $T(k) = \vec{0}$ So, $\mathbb{I}(\mathrm{y}=\mathrm{i}) = \mathrm{T}(\mathrm{y})_\mathrm{i}$

$p(y;\vec{\phi}) = h(\vec{y})\exp[\vec{\eta}^T T(\vec{y}) - A(\vec{\eta})]$, where

$$h(\vec{y}) = 1; A(\vec{\eta}) = -\log(\phi_K) \qquad \vec{\eta} = \begin{bmatrix} \log(\phi_1/\phi_k) \\ \log(\phi_2/\phi_k) \\ \vdots \\ \log(\phi_{k-1}/\phi_k) \end{bmatrix}$$

Example: (Binomial)

The **binomial distribution ($\boldsymbol{Bi(p,n)}$)** is frequently used to model the number of successes in a sample of size n independent sequences yes-no experiments. The pmf is

$$p(x; n, \mu) = \binom{n}{x} \mu^x (1-\mu)^{n-x}$$

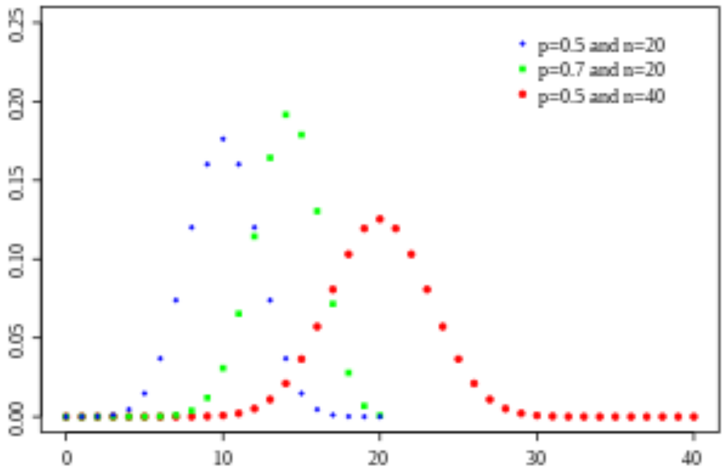
$$= \exp\left(x \log \frac{\phi}{1-\phi} + n \log(1-\phi) - \log \binom{n}{x}\right)$$

Thus: $$= \exp[\vec{\eta}^T T(\vec{y}) - A(\vec{\eta}) + C(\vec{y})]$$

Here, $T(\vec{y}) = \mathrm{x}$ $\quad \vec{\eta} = \log \frac{\phi}{1-\phi}$

$$A(\vec{\eta}) = -n \log(1-\phi) = n \log\left(1 + e^{\phi}\right)$$

$$C(\vec{y}) = -\log \binom{n}{x}$$



Example: (Normal)

The pdf function for normal distributation $Normal(\mu, \sigma^2)$ is

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

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

Compare to $p(y; \eta) = \ h(\vec{y}) \exp[\vec{\eta}^T T(\vec{y}) - A(\vec{\eta})]$, we have

(1) If we treat both (μ, σ^2) as **two parameters**, we need to define

$$h(y) = \sqrt{2\pi}\,\sigma$$

$$\vec{\eta} = \begin{bmatrix} \eta_1 \\ \\ \eta_2 \end{bmatrix} = \begin{bmatrix} \frac{\mu}{\sigma^2} \\ \\ -\frac{1}{2\sigma^2} \end{bmatrix} \qquad T(y) = \begin{bmatrix} y \\ \\ y^2 \end{bmatrix} \qquad A(\vec{\eta}) = (-2\eta_2)^{1/2} \exp\left(\frac{\eta_1^2}{4\eta_2}\right)$$

Example: (Normal with known σ^2)

(2) When σ^2 is known (treat as constant), denote $\vec{\theta} = \begin{bmatrix} \mu \\ \sigma^2 \end{bmatrix}$ it becomes a **one-parameter** exponential family on

$\eta = \frac{\mu}{\sigma^2}$, so $\mu = \sigma^2 \eta$ $\qquad$ $T(y) = y$

$$A(\eta) = \frac{1}{2\sigma^2}\mu^2 = \frac{\sigma^2\eta^2}{2} \qquad h(y) = \frac{1}{\sqrt{2\pi}\,\sigma}\exp\left(-\frac{x^2}{2\sigma^2}\right)$$

In format of $p(y;\eta,\phi) = \exp\left[\frac{\vec{\eta}^T T(\vec{y}) - A(\vec{\eta})}{\phi} + C(\vec{y},\phi)\right]$

$\phi = \sigma^2$, and $\eta = \mu$, and $A(\eta) = \frac{\eta^2}{2}$ $\qquad$ $C(y,\phi) = -\frac{1}{2}\left(\frac{y^2}{\phi} + \log(2\pi\phi)\right)$

Canonical Link function is identity.

Example: (Poisson)

Poisson is a discrete distribution defined to express the **number events** that occur in a **unit** of time or space. This distribution, which is similar to Gaussian distribution but for count data, is given by

$$p(y|\lambda) = \frac{\lambda^y e^{-\lambda}}{y!} = \frac{1}{y!}\exp(y \log\lambda - \lambda)$$

$$E(Y) = \lambda, \mathrm{Var(Y)} = \lambda$$

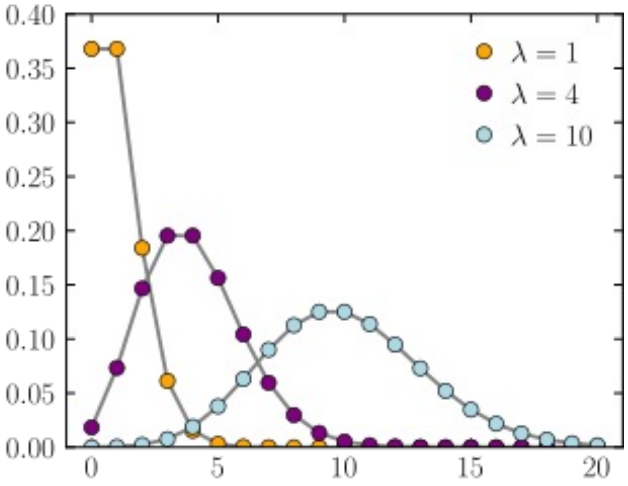
Compare to exponential family,

$$\eta = \log\lambda$$

$$T(y) = y$$

$$h(y) = \frac{1}{y!}$$

$$A(\eta) = \lambda = e^{\eta}$$



Example: (Exponential Distribution)

The exponential distribution is a distribution that models the **independent arrival** time. Its distribution (the probability density function, pdf) is given as

$$P(y;\lambda) = \lambda e^{-\lambda y} \text{ for y} \geq 0$$

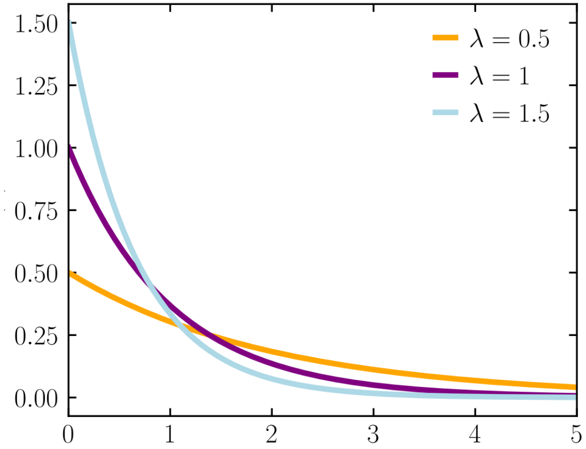
$$E(Y) = \frac{1}{\lambda}, \mathrm{Var(Y)} = \frac{1}{\lambda^2}$$

Compare to exponential family,

$$\eta = \lambda$$

$$T(y) = -y$$

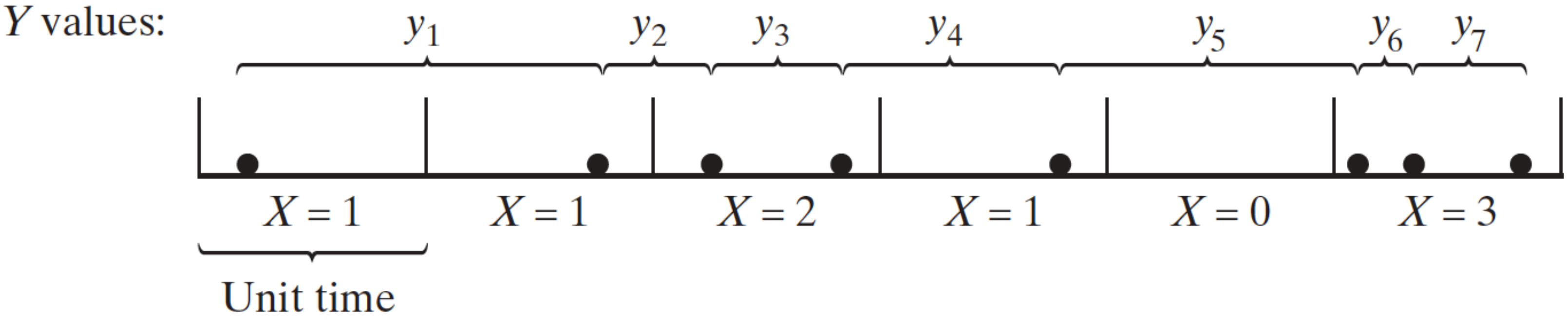
$$h(y) = \mathbb{I}(y \geq 0)$$



$$Z(\lambda) = \frac{1}{\lambda}$$ So, $A(\lambda) = \log Z(\lambda) = -\log \lambda$

Exponential v.s. Poisson

Exponential Distribution Y



Poisson Distribution

Example: Laplace Distribution (double exponential distribution)

The Laplace distribution $(Laplace(\mu, b))$ has been used in speech recognition and in JPEG image compression.

$$f(x|\mu, b) = \frac{1}{2b} \exp\left(-\frac{|x-\mu|}{b}\right)$$

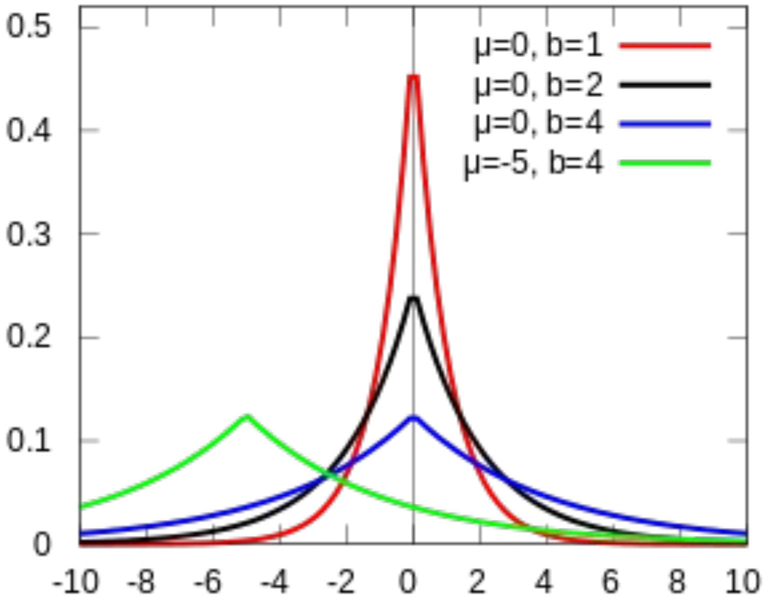
$$E(X) = \mu$$

$$Var(X) = 2b^2$$

With known μ

$$\eta = -\frac{1}{b} \qquad T(y) = |x-\mu|$$

In general, it is not.



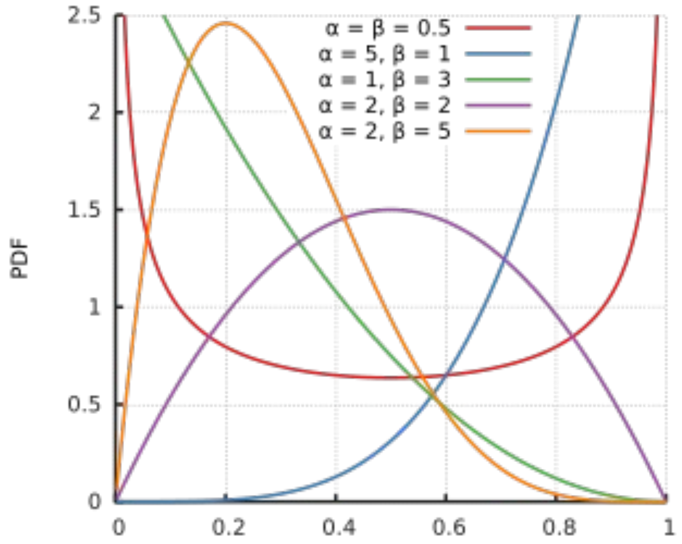
Example: Beta Distribution

Beta Distribution ($Beta(\alpha,\beta)$) is often used as prior on Binomial distributions (it is a conjugate prior).

$$f(x;\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\alpha)} x^{\alpha-1}(1-x)^{\beta-1}$$

$$= \exp\Big((\alpha-1)\log \mathrm{x} + (\beta-1)\log{(1-\mathrm{x})} + \log\big(\mathrm{B}(\alpha,\beta)\big)\Big)$$

where $\Gamma(z) = \int_0^\infty t^{z-1}e^{-t}dt$ is the Gamma function.

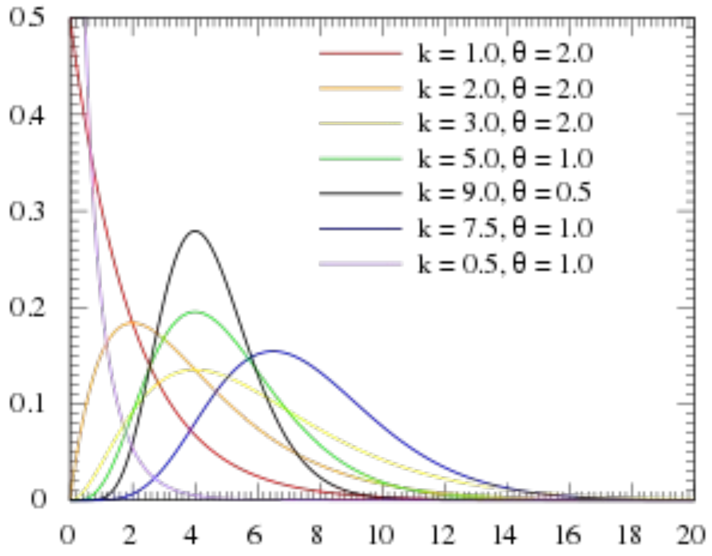


Example: Gamma Distribution

Gamma Distribution ($Gamma(k,\theta)$) is popular as a prior on coefficients. Obtained from integral over waiting times in Poisson distribution

$$p(x;k,\theta)=\frac{x^{k-1}e^{-\frac{x}{\theta}}}{\theta^k\Gamma(k)} = \exp\left((k-1)\log x - \frac{x}{\theta} - k\log\theta - \log\Gamma(k)\right)$$

Here, $\Gamma(k) = k!$ Is the gamma function.



➢ Moments of Exponential Family

In the family of exponential distributions, the $A(\vec{\eta})$ function is in fact the **Moment generating function of** $T(Y)$.

That is gradient $\nabla_{\vec{\eta}}\big(A(\vec{\eta})\big) = E\big(T(\vec{y})\big)$, Hessian matrix $H\big(A(\vec{\eta})\big) = Cov\big(T(\vec{y})\big)$

We show this by derivatizing this term:

$$A(\vec{\eta}) \colon = \log \int h(y) \exp\big(\vec{\eta}^T T(\vec{y})\big)\, dy$$

Let us compute the one dimensional case:

$$\frac{d(A(\eta))}{d\,\eta} = \frac{\frac{d}{d\eta} \int h(y) \exp\big(\vec{\eta}^T T(\vec{y})\big)\, dy}{\int h(y) \exp\big(\vec{\eta}^T T(\vec{y})\big)\, dy} = \frac{\int T(\vec{y}) h(y) \exp\big(\vec{\eta}^T T(\vec{y})\big)\, dy}{\int h(y) \exp\big(\vec{\eta}^T T(\vec{y})\big)\, dy}$$

$$= \frac{\int T(\vec{y}) h(y) \exp\big(\vec{\eta}^T T(\vec{y})\big)\, dy}{\exp(A(\eta))} = \int T(\vec{y}) h(y) \exp(\vec{\eta}^T T(\vec{y}) - A(\eta))\, dy = E\big(T(\vec{y})\big)$$

Similarly, for the second derivative,

$$\frac{d^2(A(\eta))}{d\eta^2} = Var\left(T(\vec{y})\right)$$

Remark: Here our calculation is for one dimension η. In general, when $\vec{\eta} \in \mathbb{R}^{\boldsymbol{d}}$, we only need to change differential $\frac{d(A(\eta))}{d\,\eta}$ to gradient $\nabla_{\vec{\eta}}\left(A(\vec{\eta})\right)$.

$\mathrm{A}(\eta)$ is a convex function, since (co)variance matrix is positive semi-definite.

Example: (Bernouli)

$$A(\vec{\eta}) = \log(1 + e^{\eta})$$

So,

$$\frac{d(A(\eta))}{d\,\eta} = \cdots = \frac{1}{1 + e^{-\eta}} = \mu = E(Y)$$

➢ Construction of Generalized Linear Models (GLM)

1. Random component:

$Y|\vec{X}$ $\sim$ some exponential family distribution (with parameter $\vec{\eta}$)

Our goal is to estimate the **expectation** of this distribution $\mu(\vec{X}) \coloneqq E(Y|\vec{X})$.

2. Linear assumption (systematic(non-random) component):

Assume there is a linear predictor $\vec{\xi} = \vec{\beta}^T\vec{X}$

3. Link: between the random $\mu(\vec{X})$ and covariates $\vec{X}$:

$$g\left(\mu(\vec{X})\right) = \vec{\xi} = \vec{\beta}^T\vec{X}$$

where g called link function, and $\mu(\vec{X}) = E(Y|\vec{X}) = \Psi^{-1}(\vec{\eta})$

Since our goal is $\mu(\vec{X}) = g^{-1}(\vec{\xi}) = g^{-1}(\vec{\beta}^T\vec{X})$, we need the link function to be **invertible,** in addition, we also need g is **monotonic**.

General relationship between the variables in a generalized linear model:

$$\vec{\beta}^T \vec{X} = \vec{\xi}$$

Exponential Family

$$\begin{tikzcd}
\vec{\beta} \arrow[dr] & & & & \\
& \vec{\xi} \arrow[r, shift left, "g^{-1}"] & \vec{\mu} \arrow[l, shift left, "g"] \arrow[r, shift left, "\Psi"] & \vec{\eta} \arrow[l, shift left, "\Psi^{-1}"] \arrow[r] & Y|\vec{X} \\
\vec{X} \arrow[ur] & & & &
\end{tikzcd}$$

link function

Usually we assume $\vec{\xi} = \vec{\eta}$ and $g = \Psi$, which is the **canonical link function.**

If $T = id$, since $\nabla_{\vec{\eta}}\left(A(\vec{\eta})\right) = \mathrm{E}\left(\mathrm{Y}\middle|\vec{X}\right) = \vec{\mu}$, then the **canonical link function** is

$$\vec{\mu} = \nabla_{\vec{\eta}}\left(A(\vec{\eta})\right)$$

Canonical link examples:

Normal: Identity function: $g(\mu) = \mu$

Binomial: **logit** function: $g(\mu) = \log \frac{\mu}{1-\mu}$

Poisson: log function: $g(\mu) = \log(\mu)$

Gamma: $g(\mu) = -\frac{1}{\mu}$

Negative binomial: $g(\mu) = \log\left[\frac{\mu}{k\left(1+\frac{\mu}{k}\right)}\right]$

Non-Canonical link examples for binary classification:

We need invertible functions from $\mathbb{R}$ to $[0,1]$, here are some popular examples

Probit model:

$$g^{-1}(\xi) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\xi} \exp\left(-\frac{u^2}{2}\right) du$$

Log-log model

$$g^{-1}(\xi) = \exp\left(-\exp(-\xi)\right)$$

Complementary Log-log model

$$g^{-1}(\xi) = 1 - \exp\left(-\exp(\xi)\right)$$

Recover linear regression

$$Y|\vec{X} \sim Normal(\mu, \sigma^2)$$

Recall $p(y;\eta,\phi) = \exp\left[\frac{\vec{\eta}^T T(\vec{y}) - A(\vec{\eta})}{\phi} + C(\vec{y},\phi)\right]$

$$T(y) = y;\ \ \phi = \sigma^2;\ \eta = \mu;\ A(\eta) = \frac{\eta^2}{2}\ ;$$

$$C(y,\phi) = -\frac{1}{2}\left(\frac{y^2}{\phi} + \log(2\pi\phi)\right)$$

Canonical Link function: $g = \Psi = \mathrm{id}$

□ Summary of Generalized Linear Models (GLM)

- $Y|\vec{X}$ has a distribution function (Usually $T = \text{identity}$):

$$p(y|\vec{x};\eta,\phi) = \exp\left[\frac{\vec{\eta}^{T}T(\vec{y}) - A(\vec{\eta})}{\phi} + C(\vec{y},\phi)\right]$$

- Gradient $\nabla_{\vec{\eta}}\big(A(\vec{\eta})\big) = E\big(T(\vec{y})\big) =: \vec{\mu}$, ⇨ $\vec{\mu} = \Psi^{-1}(\vec{\eta})$

$$\begin{tikzcd}
\vec{\beta}^{T}\vec{x} \arrow[r, shift left, "g^{-1}"] & \vec{\mu} \arrow[l, shift left, "g"] \arrow[r, shift left, "\Psi"] & \vec{\eta} \arrow[l, shift left, "\Psi^{-1}"]
\end{tikzcd}$$

- If $g = \Psi$, it is the canonical link. But there are other non-canonical choices.

$$\vec{\eta} = \Psi\left(g^{-1}(\vec{\beta}^{T}\vec{x})\right)$$

- Optimization via **MLE** $\hat{\vec{\beta}} = \underset{\vec{\beta}}{\operatorname{argmin}}\, P(\vec{y}|X,\vec{\eta})$

- **GLM model for prediction:** $\vec{\mu} = g^{-1}(\vec{\beta}^{T}\vec{x})$.

➢ Maximum Likelihood estimates for $\vec{\eta}$

Suppose the i.i.d. data set $D = \{\vec{y}^{(i)}\}_{i=1}^{N}$ is observed from a distribution with exponential family pdf/pmf

$$p(\vec{y};\vec{\eta},\phi) = \exp\left[\frac{\vec{\eta}^T T(\vec{y}) - A(\vec{\eta})}{\phi} + C(\vec{y},\phi)\right]$$

The **likelihood** is given by (fix ϕ)

$$L(\vec{\eta}) = \prod_{i=1}^{N} p\left(\vec{y}^{i};\vec{\eta}\,,\phi\right) = \prod_{i=1}^{N} \exp\left[\frac{\vec{\eta}^T\, T\left(\vec{y}^{(i)}\right) - A(\vec{\eta})}{\phi} + C\left(\vec{y}^{(i)},\phi\right)\right]$$

The **log-likelihood** is given by

$$l(\vec{\eta}) = \log L(\vec{\eta}) = \frac{1}{\phi}\left(\sum_{i=1}^{N} \vec{\eta}^{\,T}\, T\left(\vec{y}^{(i)}\right) - NA(\vec{\eta})\right) + \sum_{i=1}^{N} C\left(\vec{y}^{(i)},\phi\right)$$

$A(\vec{\eta})$ is convex. In most exponential family models, $A(\vec{\eta})$ is strictly convex, l( ) has a unique maximum at $\hat{\vec{\eta}}$, *solved by*

$$\nabla_{\vec{\eta}}\left(l(\vec{\eta})\right) = 0$$

$$\nabla_{\vec{\eta}}\left(l(\vec{\eta})\right) = \frac{1}{\phi}\left(\sum_{i=1}^{N} T\left(\vec{y}^{(i)}\right) - N\nabla_{\vec{\eta}}\left(A(\vec{\eta})\right)\right) = 0$$

Theorem: The **unique solution** $\hat{\vec{\eta}}$ that maximizes $l(\vec{\eta})$ and $\mathrm{L}(\vec{\eta})$ is when

$$\nabla_{\vec{\eta}}\left(A(\vec{\eta})\right)\Big|_{\vec{\eta}=\hat{\vec{\eta}}} = \frac{1}{N}\sum_{i=1}^{N} T\left(\vec{y}^{(i)}\right)$$

➢ Maximum Likelihood estimates for GLM parameters

Suppose the i.i.d. data set $D = \left\{(\vec{x}^{(i)}, \vec{y}^{(i)})\right\}_{i=1}^{N}$, where $Y|X$ is from a distribution with exponential family pdf/pmf.

$$p(\vec{y} \mid \vec{x}; \vec{\eta}, \phi) = \exp\left[\frac{\vec{\eta}^T T(\vec{y}) - A(\vec{\eta})}{\phi} + C(\vec{y}, \phi)\right]$$

The **likelihood** is given by (fix ϕ)

$$L(\vec{\eta}) = \prod_{i=1}^{N} p\left(\vec{y}^{(i)}; \vec{\eta}^{(i)}, \phi\right) = \prod_{i=1}^{N} \exp\left[\frac{\vec{\eta}^{(i)T} T\left(\vec{y}^{(i)}\right) - A\left(\vec{\eta}^{(i)}\right)}{\phi} + C(\vec{y}^{(i)}, \phi)\right]$$

The **log-likelihood** is given by

$$l(\vec{\eta}) = \log L(\vec{\eta}) = \frac{1}{\phi}\left(\sum_{i=1}^{N} {\vec{\eta}^{(i)}}^T T\left(\vec{y}^{(i)}\right) - A\left(\vec{\eta}^{(i)}\right)\right) + \sum_{i=1}^{N} C(\vec{y}^{(i)}, \phi)$$

Link function $g\left(\mu(\vec{X})\right) = \vec{\xi} = \vec{\beta}^T \vec{x}$ and $\mu(\vec{X}) = \Psi(\vec{\eta})$

GLM Model: $\mu(\vec{x}) = g^{-1}(\vec{\beta}^T \vec{x})$

Replacing $\vec{\eta} = \Psi\left(g^{-1}(\vec{\xi})\right) = \Psi\left(g^{-1}(\vec{\beta}^T \vec{x})\right)$, the **log-likelihood** is given by

$$l(\vec{\beta}) = \frac{1}{\phi}\left(\sum_{i=1}^{N} \Psi\left(g^{-1}(\vec{\beta}^T \vec{x}^{(i)})\right)^T T(\vec{y}^{(i)}) - A\left(\Psi\left(g^{-1}(\vec{\beta}^T \vec{x}^{(i)})\right)\right)\right) + \sum_{i=1}^{N} C(\vec{y}^{(i)}, \phi)$$

Now, it is an **optimization** question. We can use **Gradient Descent** or **Newton's method** to find **argmax** $l(\vec{\beta})$ of the log-likelihood.

Example(Gaussian)

Consider the GLM for independent Gaussian observations $y^{(i)} \sim N(\mu^{(i)}, \sigma^2)$ for $i = 1, \ldots, N$ with fixed σ^2

Recall that the natural parameter $\vec{\eta} = \mu = \vec{\beta}^T \vec{x} = \vec{x}^T \vec{\beta}$

The **log-likelihood** is given by

$$l(\vec{\beta}) = \frac{1}{\phi}\left(\sum_{i=1}^{N} (\vec{\eta}^T T(\vec{y}^{(i)}) - A(\vec{\eta}))\right) + \sum_{i=1}^{N} C(\vec{y}^{(i)}, \phi)$$

$$= \frac{1}{\sigma^2}\left(\sum_{i=1}^{N} \vec{\beta}^T \vec{x}^{(i)}(\vec{y}^{(i)}) - \frac{\vec{\beta}^T \vec{x}^{(i)} \vec{x}^{(i)^T} \vec{\beta}}{2}\right) + \sum_{i=1}^{N} C(\vec{y}^{(i)}, \phi)$$

$$= \frac{1}{\sigma^2}\left(\vec{\beta}^T X^T \vec{y} - \frac{\vec{\beta}^T X^T X \vec{\beta}}{2}\right) + \sum_{i=1}^{N} C(\vec{y}^{(i)}, \phi)$$

$T(y) = y$
$\phi = \sigma^2,$
$\eta = \mu,$
$A(\eta) = \frac{\eta^2}{2}$

Maximizing this function with respect to$\vec{\beta}$, we have the OLS:

$$X^T X \vec{\beta} = X^T \vec{y}$$

Example(Logistics Regression)

Consider the GLM for independent Bernoulli observations $y^{(i)} \sim Ber\left(\mu^{(i)}\right)$ for $i = 1, \dots, N$

Recall that $T(y) = y$ and $h(\vec{y}) = 1$ $\phi = 1.$

$$\eta = \log\left(\frac{\mu}{1-\mu}\right) \implies \mu = \frac{1}{1+e^{-\eta}}$$

$$A(\vec{\eta}) = -\log(1-\mu) = \log(1+e^{\eta})$$

Canonical Link function: logit function: $\vec{\xi} = \eta = g(\mu) = \log\frac{\mu}{1-\mu}$

$$\vec{\xi} = \vec{\beta}^T\vec{x} = \vec{x}^T\vec{\beta}$$

Logistics Model:

$$\mu = \frac{1}{1+e^{-\vec{x}^T\vec{\beta}}}$$

The **log-likelihood** is given by

$$l(\vec{\beta}) = \frac{1}{\phi}\left(\sum_{i=1}^{N}(\vec{\eta}^{T}T(\vec{y}^{(i)}) - A(\vec{\eta}))\right) + \sum_{i=1}^{N} C(\vec{y}^{(i)}, \phi)$$

$$= \sum_{i=1}^{N}(\vec{\beta}^{T}\vec{x}^{(i)}(\vec{y}^{(i)}) - \log(1 + e^{\vec{\beta}^{T}\vec{x}})$$

$$= \vec{\beta}^{T}X^{T}\vec{y} + \sum_{i=1}^{N} - \log\left(1 + e^{\vec{\beta}^{T}\vec{x}^{(i)}}\right)$$

To maximize the log-likelihood, the gradient of $l(\vec{\beta})$ is

$$\nabla_{\vec{\beta}}\, l(\vec{\beta}) = X^{T}\vec{y} - X^{T}h_{\vec{\beta}}(X)$$

where, $[X^{T}h(X)]_{j} = \sum_{i=1}^{N} \frac{e^{\vec{\beta}^{T}\vec{x}^{(i)}}}{\left(1 + e^{\vec{\beta}^{T}\vec{x}^{(i)}}\right)} [\vec{x}^{(i)}]_{j}$

Example(Probit Regression, not canonical link)

Consider the GLM for independent Bernoulli observations $y^{(i)} \sim Ber\left(\mu^{(i)}\right)$ for $i = 1, \dots, N$

With pdf as $p(y|x;\eta) = h(\vec{y}) \exp[\vec{\eta}^T T(\vec{y}) - A(\vec{\eta})]$, where

$T(y) = y$ and $h(\vec{y}) = 1$ $\eta = \log\left(\frac{\mu}{1-\mu}\right) = \Psi(\mu) \implies \mu = \frac{1}{1+e^{-\eta}}$

$A(\vec{\eta}) = -\log(1-\mu) = \log(1+e^{\eta})$

Probit model by using a non-canonical link function:

$$\begin{tikzcd}
\vec{\beta} \arrow[dr] & & & & \\
& \vec{\xi} \arrow[r, shift left, "g^{-1}"] & \vec{\mu} \arrow[l, shift left, "g"] \arrow[r, shift left, "\Psi"] & \vec{\eta} \arrow[l, shift left, "\Psi^{-1}"] \arrow[r] & Y|\vec{X} \\
\vec{X} \arrow[ur] & & & &
\end{tikzcd}$$

$$\mu = g^{-1}(\xi) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\xi} \exp\left(-\frac{u^2}{2}\right) du$$

This the cdf function of Normal(0,1)

The probability is

$$p(y|x;\eta) = \exp[\vec{\eta}^T T(\vec{y}) - A(\vec{\eta})] = \mu^y\ (1-\mu)^{1-y} \text{ for } y \in \{0,1\},$$

In **canonical link**, we have $\vec{\eta} = \vec{\xi} = \vec{\beta}^T\vec{x} = \vec{x}^T\vec{\beta}$

In **non-canonical link**, we have $\vec{\eta} = \Psi\left(g^{-1}(\vec{\xi})\right) = \Psi\left(g^{-1}(\vec{\beta}^T\vec{x})\right)$

Probit model

$$\mu = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\vec{\beta}^T\vec{x}} \exp\left(-\frac{u^2}{2}\right) du$$

The **likelihood** is given by

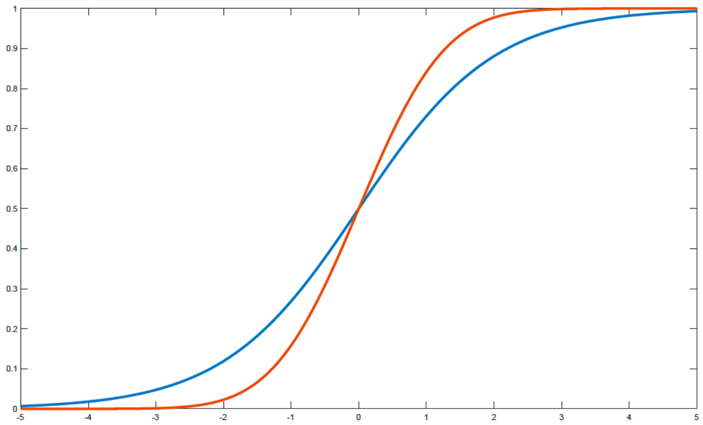
$$L(\vec{\beta}) = \prod_{i=1}^{N} p(\vec{y}^{(i)};\vec{\eta}^{(i)}) = \prod_{i=1}^{N} g^{-1}(\vec{\beta}^T\vec{x}^{(i)})^{y^{(i)}} \left(1 - g^{-1}(\vec{\beta}^T\vec{x}^{(i)})\right)^{1-y^{(i)}}$$

The **log-likelihood** is given by

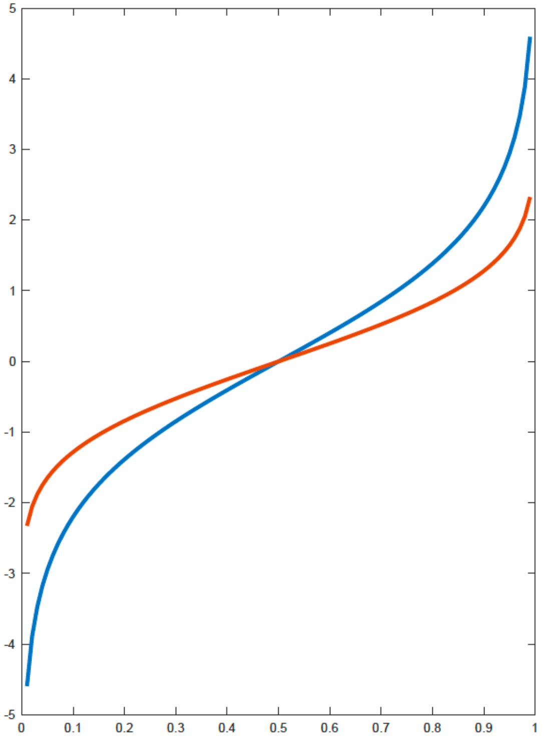
$$l(\vec{\beta}) = \sum_{i=1}^{N} y^{(i)} \log g^{-1}\left(\vec{\beta}^T \vec{x}^{(i)}\right) + \left(1 - y^{(i)}\right) \log\left(1 - g^{-1}\left(\vec{\beta}^T \vec{x}^{(i)}\right)\right)$$

This log-likelihood function is **globally concave** in $\vec{\beta}$, and therefore standard numerical algorithms for optimization will converge rapidly to the **unique** maximum.

Logistics v.s. Probit



Sigmoid v.s. Gaussian cdf



Logit link v.s. probit link

Logistics Model: $$\mu(\vec{x}) = \frac{1}{1 + e^{-\vec{x}^T\vec{\beta}}}$$

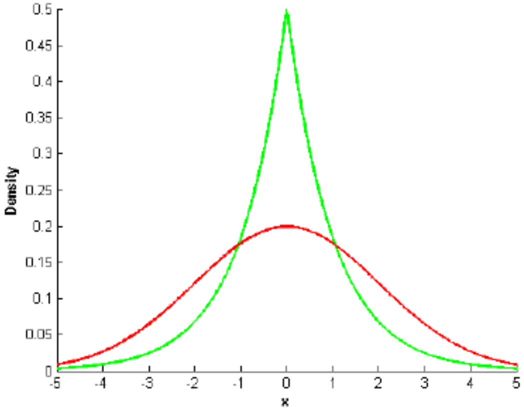
Probit model $$\mu(\vec{x}) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\vec{\beta}^T\vec{x}} \exp\left(-\frac{u^2}{2}\right) du$$

Example(practice, not exponential family)

Suppose $Y|\vec{X}$ is a distribution with pdf $p(y;\mu) = \frac{1}{2}\exp(-|y-\mu|)$

Q: Do it belongs to exponential family?

Q: Find the MLE for the GLM.



Historically the GLM was first developed for the exponential family but was later extended to the non-exponential family and even to the case where the distribution is not completely known.

Example: (Poisson Model)

Consider the GLM for independent Poisson observations

$$y^{(i)} \sim Poi\left(\mu^{(i)}\right) \text{ for } i = 1, \ldots, N$$

Recall $p(y|\lambda) = \dfrac{\lambda^y e^{-\lambda}}{y!} = \dfrac{1}{y!}\exp(y \log \lambda - \lambda)$

with $\eta = \log \lambda;$ $T(y) = y$ $h(y) = \dfrac{1}{y!}$ $A(\eta) = \lambda = e^{\eta}$

So, the **canonical link function is $\boldsymbol{g(\lambda) = \log \lambda}$** when $\eta = \vec{\xi} = \vec{\beta}^T \vec{x} = \vec{x}^T \vec{\beta}$

$E(Y) = \lambda$

The **Poisson model** is $\lambda(\vec{x}) = e^{\vec{\beta}^T \vec{x}}$

MLE:

$$l(\vec{\beta}) = \sum_{i=1}^{N} \vec{\eta}^{(i)^T}\left(\vec{y}^{(i)}\right) - A\left(\vec{\eta}^{(i)}\right) = \sum_{i=1}^{N} \vec{x}^{(i)^T}\vec{\beta}\left(\vec{y}^{(i)}\right) - \exp\left(\vec{\beta}^T\vec{x}^{(i)}\right)$$

Optimization: find **argmax** $l(\vec{\beta})$

Poisson distribution usually gives a good model for numbers of visitors.

Build a model to estimate the number y of customers arriving in your store (or number of page-views on your website) in any given hour, based on certain features $\vec{x}$ such as store promotions, recent advertising, weather, day-of-week, etc.

Example: (Disease Occurring Rate, Exponential)

In the early stages of a disease epidemic, the rate at which new cases occur increases exponentially through time.

Thus, if μ_i is the expected number of new cases on day t_i, it might be appropriate a model of the form:

$$\mu_i = \gamma \exp(\delta t_i)$$

Take the log of both sides,

$$\log(\mu_i) = \log(\gamma) + \delta t_i$$

$$= \theta_0 + \theta_1 t_i$$

Furthermore, since the outcome is a count, the Poisson distribution seems reasonable. Thus, this model fits into the GLM framework with a Poisson outcome distribution, a log link, and a linear predictor of $\theta_0 + \theta_1 t_i$

Example(Prey Capture Rate, Gamma Distribution):

The rate of capture of prey, y_i, by a hunting animal increases as the density of prey, x_i, increases, but will eventually level off as the predator has as much food as it can eat

A suitable model for this situation might be

$$\mu_i = \frac{\alpha x_i}{h + x_i}$$

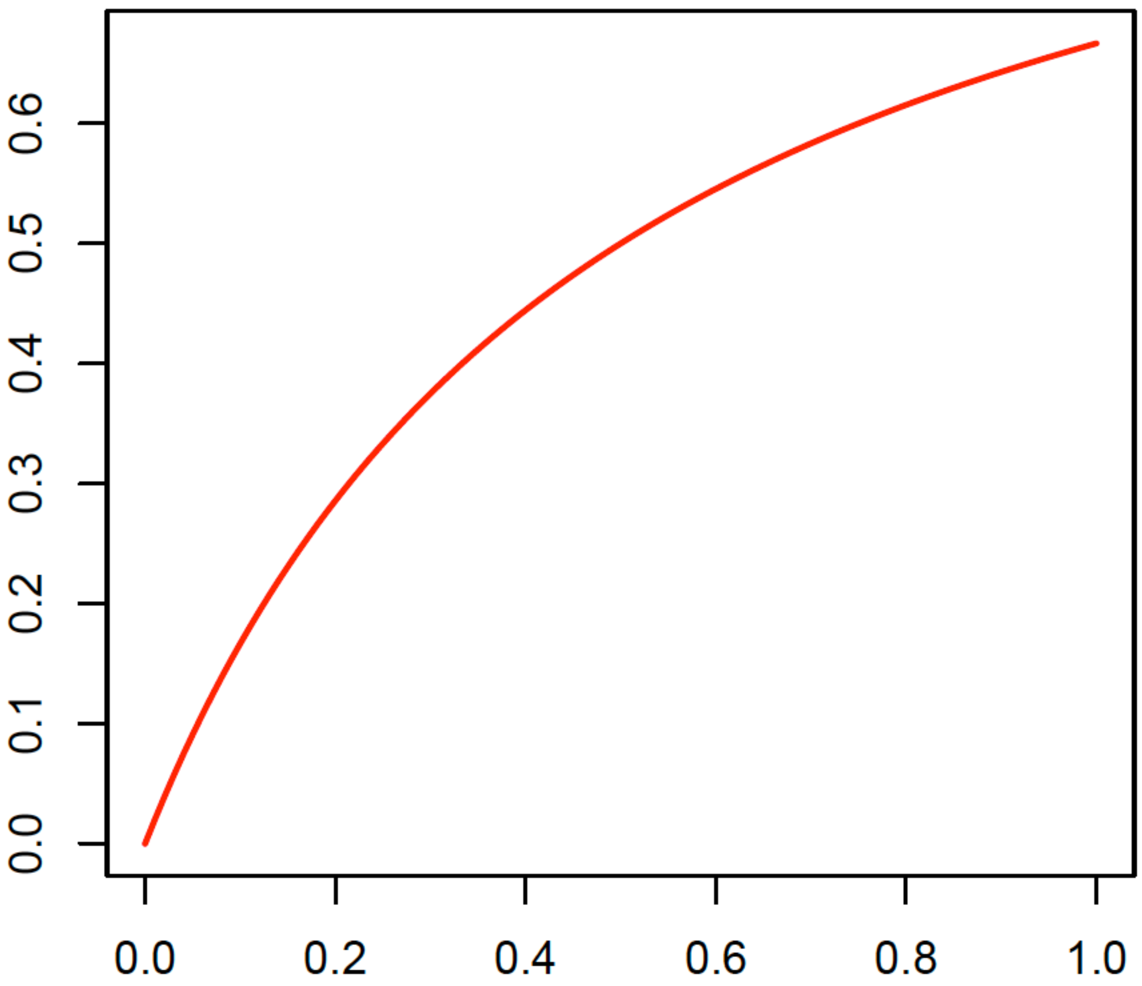
where α represents the maximum capture rate, and h represents the prey density at which the capture rate is half the maximum rate.

This model is not linear, but taking the reciprocal of both sides,

$$\frac{1}{\mu_i} = \frac{h + x_i}{\alpha x_i} = \theta_0 + \theta_1 \frac{1}{x_i}$$

Because the variability in prey capture likely increases with the mean, we might use a GLM with a reciprocal link and a gamma distribution.

Prey Capture Rate



- **Multivariate Gaussian**

The probability density function of the multivariate Gaussian distribution $N(\vec{x}, \vec{\mu}, \Sigma)$ is

$$p(\vec{x}|\vec{\mu}, \Sigma\ ) = \frac{1}{\sqrt{(2\pi)^d|\Sigma|}} \exp\left(-\frac{1}{2}(\vec{x}-\vec{\mu})^T\Sigma^{-1}(\vec{x}-\vec{\mu})\right)$$

where **covariance** matrix Σ is an symmetric positive definite matrix.

$$\Sigma = \mathrm{cov}(\vec{x}, \vec{x}) = \mathrm{E}[(\vec{x}-\vec{\mu})(\vec{x}-\vec{\mu})^{\mathrm{T}}]$$

The dependence of the multivariate Gaussian density on $\vec{x}$ is entirely through the value of the quadratic form

$$\Delta^2 := (\vec{x}-\vec{\mu})^T\Sigma^{-1}(\vec{x}-\vec{\mu})$$

The value Δ is Mahalanobis distance, and can be seen as a generalization of the Z score $z = \frac{x-\mu}{\sigma}$

- ❑ **Marginalization**

Suppose $\vec{x}$ has a multivariate Gaussian distribution $N(\vec{x}, \vec{\mu}, \Sigma)$

Let us partition the vector into two components:

$$\vec{x} = \begin{bmatrix} \overrightarrow{x_1} \\ \overrightarrow{x_2} \end{bmatrix}$$

We partition the mean vector and covariance matrix in the same way:

$$\vec{\mu} = \begin{bmatrix} \overrightarrow{\mu_1} \\ \overrightarrow{\mu_2} \end{bmatrix} \qquad \Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}$$

The **marginal distribution** of the sub-vector $\overrightarrow{x_1}$ has a simple form

$$N(\overrightarrow{x_1}, \overrightarrow{\mu_1}, \Sigma_{11})$$

❑ Conditioning

Suppose now that we learn the exact value of the sub-vector $\overrightarrow{x_2}$

We may then condition our prior distribution on this observation, giving a posterior distribution over the remaining variables.

- The posterior distribution $p(\overrightarrow{x_1}|\vec{x}_2, \vec{\mu}, \Sigma)$ is a Gaussian distribution, denoted as

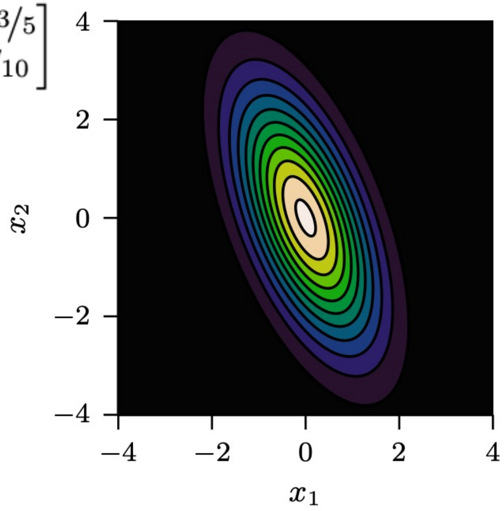
$$N(\overrightarrow{x_1}, \overrightarrow{\mu_{1|2}}, \Sigma_{11|2})$$

$$\overrightarrow{\mu_{1|2}} = \vec{\mu}_1 + \Sigma_{12}\Sigma_{22}^{-1}(\vec{x}_2 - \vec{\mu}_2)$$

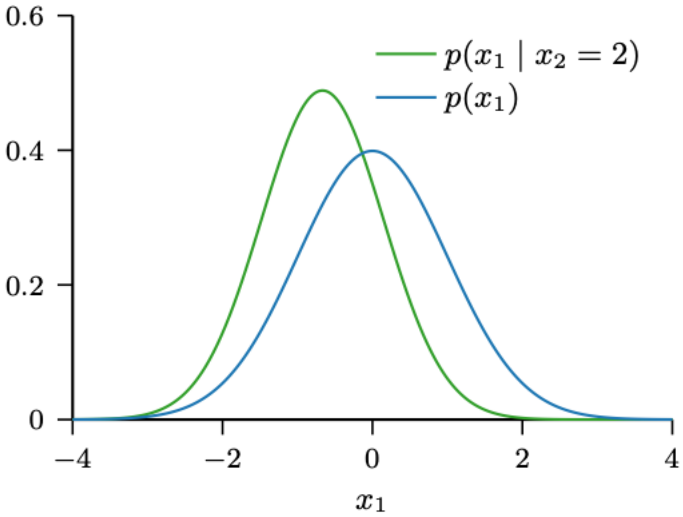
$$\Sigma_{11|2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$$

$$\mathbf{A} = \begin{bmatrix} 1/5 & -3/5 \\ 1/2 & 3/10 \end{bmatrix}$$

$$\vec{b} = \begin{bmatrix} 1 \\ -1 \end{bmatrix}$$



(a) $p(\mathbf{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma})$

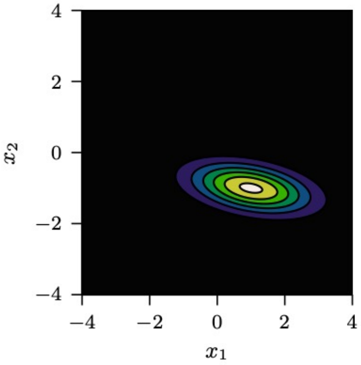


(b) $p(x_1 \mid x_2, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \mathcal{N}\left(x_1; -2/3, (2/3)^2\right)$

Other important operations of normal random variables include

Pointwise multiplication ; Convolutions $\vec{x} * \vec{y}$; Affine transformations $A\vec{x} + \vec{b}$

They all Gaussian.



(b) $p(\mathbf{y} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{A}, \mathbf{b})$

$A\vec{x} + \vec{b}$

$N(\vec{x}, A\vec{\mu} + \vec{\mathrm{b}}, \mathrm{A}\Sigma A^{-1})$

References:

Machine Learning: a Probabilistic Perspective- Kevin P. Murphy

Pattern Recognition and Machine Learning - Christopher M. Bishop

More Textbooks:

1. P. McCullagh; John A. Nelder. Generalized Linear Models, Second Edition Chapman and Hall

2. Charles E. McCulloch, Shayle R. Searle Generalized, Linear, and Mixed Models Wiley, New York

3. *Paul Roback and Julie Legler,* Beyond Multiple Linear Regression: Applied Generalized Linear Models and Multilevel Models in R. (**Chapter 4 Poisson Regression**)

MATLAB: Fit generalized linear regression model

https://www.mathworks.com/help/stats/glmfit.html

https://www.mathworks.com/help/stats/generalized-linear-regression.html

Example:

https://www.mathworks.com/help/stats/fitting-data-with-generalized-linear-models.html