Lecture 6: Generalized Linear Models

Tao LIN

SoE, Westlake University

October 15, 2025

Reading materials & Reference

Reading materials:

- Chapter 3, Stanford CS 229 Lecture Notes,
  `https://cs229.stanford.edu/notes2022fall/main_notes.pdf`
- Lecture 4, Stanford CS 231n, `http://cs231n.stanford.edu/schedule.html`

Table of Contents

Optimal classification for known generating model

What is the **optimal performance**, regardless of the finiteness of the training data?

Optimal classification for known generating model

What is the **optimal performance**, regardless of the finiteness of the training data?

Assume that we know the joint distribution $p(\mathbf{x}, y)$.

Optimal classification for known generating model

What is the **optimal performance**, regardless of the finiteness of the training data?

Assume that we know the joint distribution $p(\mathbf{x}, y)$.

- For a given input $\mathbf{x}$, *the probability that the "correct" label is* y is $p(y|\mathbf{x})$.

Optimal classification for known generating model

What is the **optimal performance**, regardless of the finiteness of the training data?

Assume that we know the joint distribution $p(\mathbf{x}, y)$.

- For a given input $\mathbf{x}$, *the probability that the "correct" label is* y is $p(y|\mathbf{x})$.
- Maximum A-Posteriori (MAP):

Optimal classification for known generating model

What is the **optimal performance**, regardless of the finiteness of the training data?

Assume that we know the joint distribution $p(\mathbf{x}, y)$.

- For a given input $\mathbf{x}$, *the probability that the "correct" label is* y is $p(y|\mathbf{x})$.
- Maximum A-Posteriori (MAP):
  If we want to maximize the probability of guessing the correct label,

Optimal classification for known generating model

What is the **optimal performance**, regardless of the finiteness of the training data?

Assume that we know the joint distribution $p(\mathbf{x}, y)$.

- For a given input $\mathbf{x}$, *the probability that the "correct" label is* y is $p(y|\mathbf{x})$.
- Maximum A-Posteriori (MAP):
  If we want to maximize the probability of guessing the correct label, then we should choose the decision rule

Optimal classification for known generating model

What is the **optimal performance**, regardless of the finiteness of the training data?

Assume that we know the joint distribution $p(\mathbf{x}, y)$.

- For a given input $\mathbf{x}$, *the probability that the "correct" label is* y is $p(y|\mathbf{x})$.
- Maximum A-Posteriori (MAP):
  If we want to maximize the probability of guessing the correct label, then we should choose the decision rule

$$\hat{y}(\mathbf{x}) = \arg\max_{y \in \mathcal{Y}} p(y|\mathbf{x}) \tag{1}$$

Optimal classification for known generating model

What is the **optimal performance**, regardless of the finiteness of the training data?

Assume that we know the joint distribution $p(\mathbf{x}, y)$.

- For a given input $\mathbf{x}$, *the probability that the "correct" label is* y is $p(y|\mathbf{x})$.
- Maximum A-Posteriori (MAP):
  If we want to maximize the probability of guessing the correct label, then we should choose the decision rule

$$\hat{y}(\mathbf{x}) = \arg\max_{y \in \mathcal{Y}} p(y|\mathbf{x}) \tag{1}$$

  This classifier is also called the *Bayes classifier*: $f^\star = \arg\min_f L_{\mathcal{D}}(f)$, where

$$f^\star(\mathbf{x}) \in \arg\max_{\{-1,1\}} \Pr(Y = y | X = \mathbf{x}) \tag{2}$$

Optimal classification for known generating model

What is the **optimal performance**, regardless of the finiteness of the training data?

Assume that we know the joint distribution $p(\mathbf{x}, y)$.

- For a given input $\mathbf{x}$, *the probability that the "correct" label is* y is $p(y|\mathbf{x})$.
- Maximum A-Posteriori (MAP):
  If we want to maximize the probability of guessing the correct label, then we should choose the decision rule

$$\hat{y}(\mathbf{x}) = \arg\max_{y \in \mathcal{Y}} p(y|\mathbf{x}) \tag{1}$$

This classifier is also called the *Bayes classifier*: $f^\star = \arg\min_f L_{\mathcal{D}}(f)$, where

$$f^\star(\mathbf{x}) \in \arg\max_{\{-1,1\}} \Pr(Y = y | X = \mathbf{x}) \tag{2}$$

- In practice, we do not know the joint distribution $p(\mathbf{x}, y)$

Optimal classification for known generating model

What is the **optimal performance**, regardless of the finiteness of the training data?

Assume that we know the joint distribution $p(\mathbf{x}, y)$.

- For a given input $\mathbf{x}$, *the probability that the "correct" label is* y is $p(y|\mathbf{x})$.
- Maximum A-Posteriori (MAP):
  If we want to maximize the probability of guessing the correct label, then we should choose the decision rule

$$\hat{y}(\mathbf{x}) = \arg\max_{y \in \mathcal{Y}} p(y|\mathbf{x}) \tag{1}$$

  This classifier is also called the *Bayes classifier*: $f^{\star} = \arg\min_f L_{\mathcal{D}}(f)$, where

$$f^{\star}(\mathbf{x}) \in \arg\max_{\{-1,1\}} \Pr(Y = y|X = \mathbf{x}) \tag{2}$$

- In practice, we do not know the joint distribution $p(\mathbf{x}, y)$
  $\Rightarrow$ Bayes classifier is an unattainable gold standard.

Optimal classification for known generating model

What is the **optimal performance**, regardless of the finiteness of the training data?

Assume that we know the joint distribution $p(\mathbf{x}, y)$.

- For a given input $\mathbf{x}$, *the probability that the "correct" label is* y is $p(y|\mathbf{x})$.
- Maximum A-Posteriori (MAP):
  If we want to maximize the probability of guessing the correct label, then we should choose the decision rule

$$\hat{y}(\mathbf{x}) = \arg\max_{y \in \mathcal{Y}} p(y|\mathbf{x}) \tag{1}$$

This classifier is also called the *Bayes classifier*: $f^\star = \arg\min_f L_{\mathcal{D}}(f)$, where

$$f^\star(\mathbf{x}) \in \arg\max_{\{-1,1\}} \Pr(Y = y | X = \mathbf{x}) \tag{2}$$

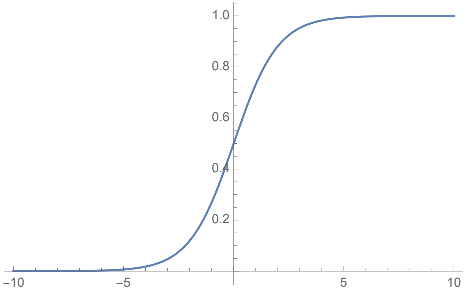
- In practice, we do not know the joint distribution $p(\mathbf{x}, y)$
  $\Rightarrow$ Bayes classifier is an unattainable gold standard.

The logistic function

Consider first of all the case of two classes.
The posterior probability for class $\mathcal{C}_1$:

$$p(\mathcal{C}_1|\mathbf{x}) = \frac{p(\mathbf{x}|\mathcal{C}_1)p(\mathcal{C}_1)}{p(\mathbf{x}|\mathcal{C}_1)p(\mathcal{C}_1) + p(\mathbf{x}|\mathcal{C}_2)p(\mathcal{C}_2)} \tag{3}$$

$$= \frac{1}{1+\exp(-\eta)} = \sigma(\eta) \tag{4}$$



Properties of the logistic function:

- $1 - \sigma(\eta) = \sigma(-\eta)$
- $\sigma'(\eta) = \sigma(\eta)\,(1-\sigma(\eta))$

The logistic function

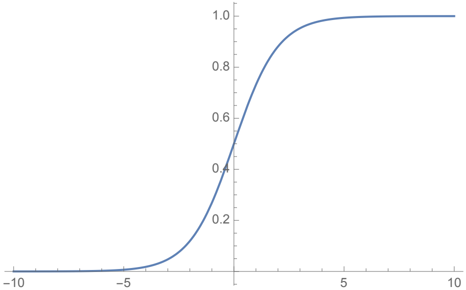
Consider first of all the case of two classes.
The posterior probability for class $\mathcal{C}_1$:

$$p(\mathcal{C}_1|\mathbf{x}) = \frac{p(\mathbf{x}|\mathcal{C}_1)p(\mathcal{C}_1)}{p(\mathbf{x}|\mathcal{C}_1)p(\mathcal{C}_1) + p(\mathbf{x}|\mathcal{C}_2)p(\mathcal{C}_2)} \tag{3}$$

$$= \frac{1}{1 + \exp(-\eta)} = \sigma(\eta) \tag{4}$$

where we have defined

$$\eta = \ln \frac{p(\mathbf{x}|\mathcal{C}_1)p(\mathcal{C}_1)}{p(\mathbf{x}|\mathcal{C}_2)p(\mathcal{C}_2)} \text{ and } \sigma(\eta) := \frac{e^\eta}{1 + e^\eta} \tag{5}$$



Properties of the logistic function:

- $1 - \sigma(\eta) = \sigma(-\eta)$
- $\sigma'(\eta) = \sigma(\eta)\,(1 - \sigma(\eta))$

The logistic function

Consider first of all the case of two classes.
The posterior probability for class $\mathcal{C}_1$:

$$p(\mathcal{C}_1|\mathbf{x}) = \frac{p(\mathbf{x}|\mathcal{C}_1)p(\mathcal{C}_1)}{p(\mathbf{x}|\mathcal{C}_1)p(\mathcal{C}_1) + p(\mathbf{x}|\mathcal{C}_2)p(\mathcal{C}_2)} \tag{3}$$

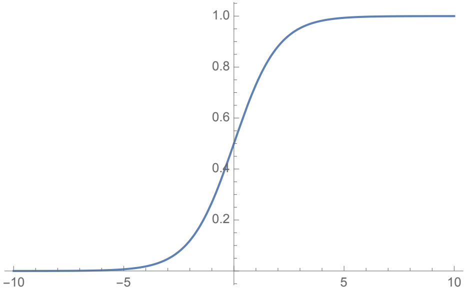
$$= \frac{1}{1+\exp(-\eta)} = \sigma(\eta) \tag{4}$$

where we have defined

$$\eta = \ln \frac{p(\mathbf{x}|\mathcal{C}_1)p(\mathcal{C}_1)}{p(\mathbf{x}|\mathcal{C}_2)p(\mathcal{C}_2)} \text{ and } \sigma(\eta) := \frac{e^\eta}{1+e^\eta} \tag{5}$$

For the case of $K > 2$ classes, we have

$$p(\mathcal{C}_k|\mathbf{x}) = \frac{p(\mathbf{x}|\mathcal{C}_k)p(\mathcal{C}_k)}{\sum_j p(\mathbf{x}|\mathcal{C}_j)p(\mathcal{C}_j)} = \frac{\exp(\eta_k)}{\sum_j \exp(\eta_j)} \tag{6}$$



Properties of the logistic function:

- $1 - \sigma(\eta) = \sigma(-\eta)$
- $\sigma'(\eta) = \sigma(\eta)\,(1-\sigma(\eta))$

Logistic Regression

Given a “new” feature vector $\mathbf{x}$, we predict the (posterior) probability of the two class labels given $\mathbf{x}$ by means of

$$p(1|\mathbf{x}) := \Pr\left[Y = 1|\mathbf{X} = \mathbf{x}\right] = \sigma\left(\mathbf{x}^\top \mathbf{w} + w_0\right) \tag{7}$$

$$p(0|\mathbf{x}) := \Pr\left[Y = 0|\mathbf{X} = \mathbf{x}\right] = 1 - \sigma\left(\mathbf{x}^\top \mathbf{w} + w_0\right) \ , \tag{8}$$

where we predict a real value (a probability) and not a label.

MLE is a method of estimating the parameters of a statistical model

The MLE finds the parameters $\mathbf{w}^\star$ under which $\{\mathbf{y}, \mathbf{X}\}$ are the most likely:

$$\mathbf{w}^\star = \arg\max_{\mathbf{w}} \left(\mathcal{L}(\mathbf{w}) := \prod_{n=1}^{N} p(\{\mathbf{x}_n, y_n\}|\mathbf{w}) \right) = \arg\min_{\mathbf{w}} \left[-\log \mathcal{L}(\mathbf{w}) \right] . \tag{9}$$

MLE is a method of estimating the parameters of a statistical model

The MLE finds the parameters $\mathbf{w}^\star$ under which $\{\mathbf{y}, \mathbf{X}\}$ are the most likely:

$$\mathbf{w}^\star = \arg\max_{\mathbf{w}} \left(\mathcal{L}(\mathbf{w}) := \prod_{n=1}^{N} p(\{\mathbf{x}_n, y_n\}|\mathbf{w}) \right) = \arg\min_{\mathbf{w}} \left[-\log \mathcal{L}(\mathbf{w}) \right] . \tag{9}$$

The likelihood of the data $\{\mathbf{y}, \mathbf{X}\}$ given the parameter $\mathbf{w}$, i.e., $p(\mathbf{y}, \mathbf{X}|\mathbf{w})$.

$$p(\mathbf{y}, \mathbf{X}|\mathbf{w}) = p(\mathbf{X}|\mathbf{w}) p(\mathbf{y}|\mathbf{X}, \mathbf{w}) = p(\mathbf{X}) p(\mathbf{y}|\mathbf{X}, \mathbf{w}) , \tag{10}$$

where $\mathbf{X}$ does not depend on $\mathbf{w}$.

MLE for Logistic Regression

For Logistic Regression, we have:

$$p(\mathbf{y}|\mathbf{X}, \mathbf{w}) = \prod_{n=1}^{N} p(y_n|\mathbf{x}_n) \tag{12}$$

MLE for Logistic Regression

For Logistic Regression, we have:

$$p(\mathbf{y}|\mathbf{X}, \mathbf{w}) = \prod_{n=1}^{N} p(y_n|\mathbf{x}_n) = \prod_{n:y_n=1} p(y_n = 1|\mathbf{x}_n) \prod_{n:y_n=0} p(y_n = 0|\mathbf{x}_n) \tag{11}$$

$$\tag{12}$$

MLE for Logistic Regression

For Logistic Regression, we have:

$$p(\mathbf{y}|\mathbf{X}, \mathbf{w}) = \prod_{n=1}^{N} p(y_n|\mathbf{x}_n) = \prod_{n:y_n=1} p(y_n = 1|\mathbf{x}_n) \prod_{n:y_n=0} p(y_n = 0|\mathbf{x}_n) \tag{11}$$

$$= \prod_{n=1}^{N} \sigma(\mathbf{x}_n^\top \mathbf{w})^{y_n} \left[1 - \sigma(\mathbf{x}_n^\top \mathbf{w})\right]^{1-y_n} \tag{12}$$

MLE for Logistic Regression

For Logistic Regression, we have:

$$p(\mathbf{y}|\mathbf{X},\mathbf{w}) = \prod_{n=1}^{N} p(y_n|\mathbf{x}_n) = \prod_{n:y_n=1} p(y_n = 1|\mathbf{x}_n) \prod_{n:y_n=0} p(y_n = 0|\mathbf{x}_n) \tag{11}$$

$$= \prod_{n=1}^{N} \sigma(\mathbf{x}_n^\top \mathbf{w})^{y_n} \left[1 - \sigma(\mathbf{x}_n^\top \mathbf{w})\right]^{1-y_n} \tag{12}$$

Minimizing $\mathcal{L}(\mathbf{w})$ through the property of stationary points.

$$\nabla\mathcal{L}(\mathbf{w}) = \frac{1}{N}\sum_{n=1}^{N} \mathbf{x}_n \left(\sigma(\mathbf{x}_n^\top \mathbf{w}) - y_n\right) = \frac{1}{N}\mathbf{X}^\top \left[\sigma(\mathbf{X}\mathbf{w}) - \mathbf{y}\right], \tag{13}$$

where $\mathbf{X} \in \mathbb{R}^{N\times d}$.

MLE for Logistic Regression

For Logistic Regression, we have:

$$p(\mathbf{y}|\mathbf{X}, \mathbf{w}) = \prod_{n=1}^{N} p(y_n|\mathbf{x}_n) = \prod_{n:y_n=1} p(y_n = 1|\mathbf{x}_n) \prod_{n:y_n=0} p(y_n = 0|\mathbf{x}_n) \tag{11}$$

$$= \prod_{n=1}^{N} \sigma(\mathbf{x}_n^\top \mathbf{w})^{y_n} \left[1 - \sigma(\mathbf{x}_n^\top \mathbf{w})\right]^{1-y_n} \tag{12}$$

Minimizing $\mathcal{L}(\mathbf{w})$ through the property of stationary points.

$$\nabla \mathcal{L}(\mathbf{w}) = \frac{1}{N} \sum_{n=1}^{N} \mathbf{x}_n \left(\sigma(\mathbf{x}_n^\top \mathbf{w}) - y_n\right) = \frac{1}{N} \mathbf{X}^\top \left[\sigma(\mathbf{X}\mathbf{w}) - \mathbf{y}\right] , \tag{13}$$

where $\mathbf{X} \in \mathbb{R}^{N \times d}$. It has no closed-form solution to $\nabla \mathcal{L}(\mathbf{w}) = 0$.

Last lecture:

- Logistic Regression

Last lecture:

- Logistic Regression

This lecture:

- Exponential Families
- Generalized Linear Models

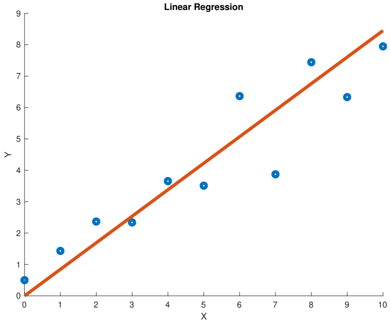
Table of Contents

Table of Contents

The Least-Squares can be defined in two different ways

- **Geometric way:**
  Minimizing the sum of the squares of the residuals:

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}} \frac{1}{2N} \sum_{n=1}^{N} (y_n - \mathbf{x}_n^\top \mathbf{w})^2 \tag{14}$$



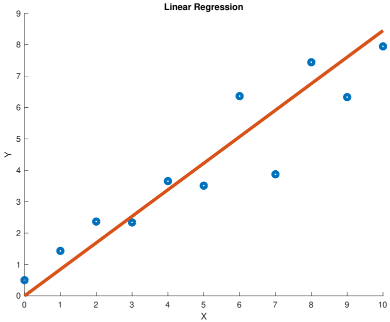
The Least-Squares can be defined in two different ways

- **Geometric way:**
  Minimizing the sum of the squares of the residuals:

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}} \frac{1}{2N} \sum_{n=1}^{N} (y_n - \mathbf{x}_n^\top \mathbf{w})^2 \tag{14}$$

- **Probabilistic way:**
  Assume the data follow a linear Gaussian model:

$$\mathbf{y} = \mathbf{x}^\top \mathbf{w} + \epsilon \text{ where } \epsilon \sim \mathcal{N}(0, \sigma^2) \tag{15}$$



The Least-Squares can be defined in two different ways

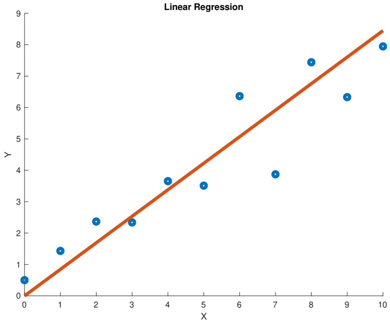
- **Geometric way:**
  Minimizing the sum of the squares of the residuals:

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}} \frac{1}{2N} \sum_{n=1}^{N} (y_n - \mathbf{x}_n^\top \mathbf{w})^2 \tag{14}$$

- **Probabilistic way:**
  Assume the data follow a linear Gaussian model:

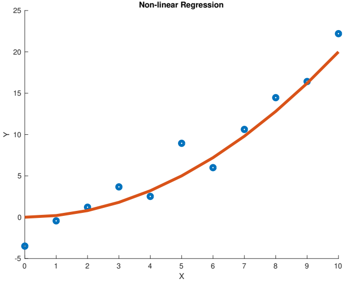
$$\mathbf{y} = \mathbf{x}^\top \mathbf{w} + \epsilon \text{ where } \epsilon \sim \mathcal{N}(0, \boldsymbol{\sigma}^2) \tag{15}$$

  Doing MLE recovers the LS estimator $\hat{\mathbf{w}}$.



How to get non-linear models?

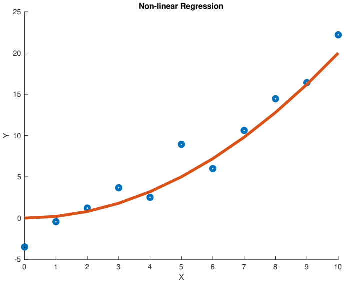
- **Features augmentations:**
  add non-linear features $(\mathbf{x}, \mathbf{x}^2, \mathbf{x}^3, \ldots)$



How to get non-linear models?

- **Features augmentations:**
  add non-linear features $(\mathbf{x}, \mathbf{x}^2, \mathbf{x}^3, \ldots)$

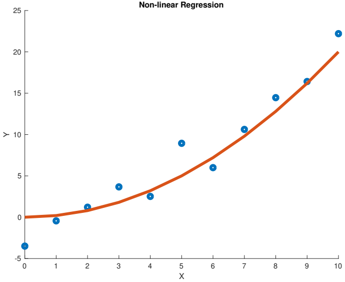
- **Different probabilistic models:**



How to get non-linear models?

- **Features augmentations:**
  add non-linear features $(\mathbf{x}, \mathbf{x}^2, \mathbf{x}^3, \ldots)$

- **Different probabilistic models:**
  - Least Squares: $y \sim \mathcal{N}\left(\mathbf{x}^\top \mathbf{w}, \boldsymbol{\sigma}^2\right)$

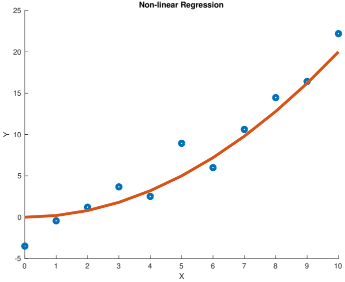


How to get non-linear models?

- **Features augmentations:**
  add non-linear features $(\mathbf{x}, \mathbf{x}^2, \mathbf{x}^3, \ldots)$

- **Different probabilistic models:**
  - Least Squares: $y \sim \mathcal{N}\left(\mathbf{x}^\top \mathbf{w}, \boldsymbol{\sigma}^2\right)$

    $\implies$ The linear model predicts the mean of a distribution μ (from which the data are sampled).



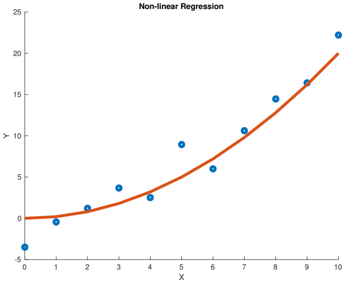
How to get non-linear models?

- **Features augmentations:**
  add non-linear features $(\mathbf{x}, \mathbf{x}^2, \mathbf{x}^3, \ldots)$

- **Different probabilistic models:**
  - Least Squares: $y \sim \mathcal{N}\left(\mathbf{x}^\top \mathbf{w}, \boldsymbol{\sigma}^2\right)$

    $\Longrightarrow$ The linear model predicts the mean of a distribution μ (from which the data are sampled).

  - Logistic Regression: $y \sim \mathcal{B}\left(\sigma(\mathbf{x}^\top \mathbf{w})\right)$



How to get non-linear models?

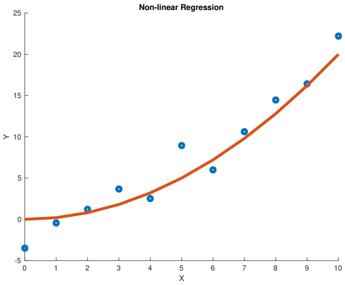
- **Features augmentations:**
  add non-linear features $(\mathbf{x}, \mathbf{x}^2, \mathbf{x}^3, \ldots)$

- **Different probabilistic models:**
  - Least Squares: $y \sim \mathcal{N}\left(\mathbf{x}^\top \mathbf{w}, \boldsymbol{\sigma}^2\right)$

    $\Longrightarrow$ The linear model predicts the mean of a distribution μ (from which the data are sampled).

  - Logistic Regression: $y \sim \mathcal{B}\left(\sigma(\mathbf{x}^\top \mathbf{w})\right)$

    $\Longrightarrow$ The linear model predicts another quantity $\eta := \mathbf{x}^\top \mathbf{w}$.



How to get non-linear models?

- **Features augmentations:**
  add non-linear features $(\mathbf{x}, \mathbf{x}^2, \mathbf{x}^3, \ldots)$

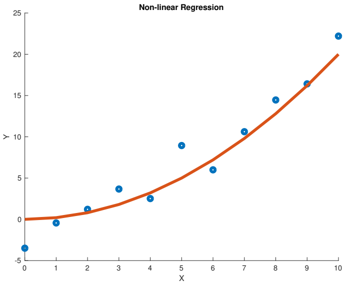
- **Different probabilistic models:**
  - Least Squares: $y \sim \mathcal{N}\left(\mathbf{x}^\top \mathbf{w}, \boldsymbol{\sigma}^2\right)$

    $\Longrightarrow$ The linear model predicts the mean of a distribution μ (from which the data are sampled).

  - Logistic Regression: $y \sim \mathcal{B}\left(\sigma(\mathbf{x}^\top \mathbf{w})\right)$

    $\Longrightarrow$ The linear model predicts another quantity $\eta := \mathbf{x}^\top \mathbf{w}$.

> $\Longrightarrow$ Generalized linear model
> $\Longrightarrow$ Exponential family



Recall the definition of Logistic Regression

Logistic Regression models the probability of the two classes $\{0, 1\}$ by

$$p(1|\eta) = \sigma(\eta) \text{ and } p(0|\eta) = 1 - \sigma(\eta)\,, \tag{16}$$

where $\eta = \mathbf{x}^\top \mathbf{w}$.

Recall the definition of Logistic Regression

Logistic Regression models the probability of the two classes $\{0, 1\}$ by

$$p(1|\eta) = \sigma(\eta) \text{ and } p(0|\eta) = 1 - \sigma(\eta)\,, \tag{16}$$

where $\eta = \mathbf{x}^\top \mathbf{w}$. This can be compactly written as (exercise: express it as the format of $\exp(\ldots)$)

$$p(y|\eta) = \frac{e^{\eta y}}{1 + e^{\eta}} \tag{17}$$

Recall the definition of Logistic Regression

Logistic Regression models the probability of the two classes $\{0, 1\}$ by

$$p(1|\eta) = \sigma(\eta) \text{ and } p(0|\eta) = 1 - \sigma(\eta)\,, \tag{16}$$

where $\eta = \mathbf{x}^\top \mathbf{w}$. This can be compactly written as (exercise: express it as the format of $\exp(\ldots)$)

$$p(y|\eta) = \frac{e^{\eta y}}{1 + e^\eta} = \exp\left(\eta y - \ln(1 + e^\eta)\right) \tag{17}$$

Recall the definition of Logistic Regression

Logistic Regression models the probability of the two classes $\{0, 1\}$ by

$$p(1|\eta) = \sigma(\eta) \text{ and } p(0|\eta) = 1 - \sigma(\eta)\,, \tag{16}$$

where $\eta = \mathbf{x}^\top \mathbf{w}$. This can be compactly written as (exercise: express it as the format of $\exp(\ldots)$)

$$p(y|\eta) = \frac{e^{\eta y}}{1 + e^{\eta}} = \exp\left(\eta y - \ln(1 + e^{\eta})\right) \tag{17}$$

- The linear model predicts $\sigma(\eta)$ which is not the mean of the distribution.

Recall the definition of Logistic Regression

Logistic Regression models the probability of the two classes $\{0,1\}$ by

$$p(1|\eta) = \sigma(\eta) \text{ and } p(0|\eta) = 1 - \sigma(\eta)\,, \tag{16}$$

where $\eta = \mathbf{x}^\top \mathbf{w}$. This can be compactly written as (exercise: express it as the format of $\exp(\ldots)$)

$$p(y|\eta) = \frac{e^{\eta y}}{1+e^{\eta}} = \exp\left(\eta y - \ln(1+e^{\eta})\right) \tag{17}$$

- The linear model predicts $\sigma(\eta)$ which is not the mean of the distribution.
- Rather η is related to the mean μ by the non-linear relation $\eta = \ln \frac{\mu}{1-\mu}$ or $\mu = \sigma(\eta)$.

Recall the definition of Logistic Regression

Logistic Regression models the probability of the two classes $\{0, 1\}$ by

$$p(1|\eta) = \sigma(\eta) \text{ and } p(0|\eta) = 1 - \sigma(\eta)\,, \tag{16}$$

where $\eta = \mathbf{x}^\top \mathbf{w}$. This can be compactly written as (exercise: express it as the format of $\exp(\ldots)$)

$$p(y|\eta) = \frac{e^{\eta y}}{1 + e^{\eta}} = \exp\left(\eta y - \ln(1 + e^{\eta})\right) \tag{17}$$

- The linear model predicts $\sigma(\eta)$ which is not the mean of the distribution.
- Rather η is related to the mean μ by the non-linear relation $\eta = \ln \frac{\mu}{1-\mu}$ or $\mu = \sigma(\eta)$.
- The relation between
    - η
    - μ

  makes possible to use linear model in this context.

Recall the definition of Logistic Regression

Logistic Regression models the probability of the two classes $\{0, 1\}$ by

$$p(1|\eta) = \sigma(\eta) \text{ and } p(0|\eta) = 1 - \sigma(\eta)\,, \tag{16}$$

where $\eta = \mathbf{x}^\top \mathbf{w}$. This can be compactly written as (exercise: express it as the format of $\exp(\ldots)$)

$$p(y|\eta) = \frac{e^{\eta y}}{1 + e^{\eta}} = \exp\left(\eta y - \ln(1 + e^{\eta})\right) \tag{17}$$

- The linear model predicts $\sigma(\eta)$ which is not the mean of the distribution.
- Rather η is related to the mean μ by the non-linear relation $\eta = \ln \frac{\mu}{1-\mu}$ or $\mu = \sigma(\eta)$.
- The relation between
  - η (the parameter predicted by the linear model)
  - μ

  makes possible to use linear model in this context.

Recall the definition of Logistic Regression

Logistic Regression models the probability of the two classes $\{0, 1\}$ by

$$p(1|\eta) = \sigma(\eta) \text{ and } p(0|\eta) = 1 - \sigma(\eta)\,, \tag{16}$$

where $\eta = \mathbf{x}^\top \mathbf{w}$. This can be compactly written as (exercise: express it as the format of $\exp(\ldots)$)

$$p(y|\eta) = \frac{e^{\eta y}}{1 + e^{\eta}} = \exp\left(\eta y - \ln(1 + e^{\eta})\right) \tag{17}$$

- The linear model predicts $\sigma(\eta)$ which is not the mean of the distribution.
- Rather η is related to the mean μ by the non-linear relation $\eta = \ln \frac{\mu}{1-\mu}$ or $\mu = \sigma(\eta)$.
- The relation between
  - η (the parameter predicted by the linear model)
  - μ (the distribution's mean)

  makes possible to use linear model in this context.

Recall the definition of Logistic Regression

Logistic Regression models the probability of the two classes $\{0, 1\}$ by

$$p(1|\eta) = \sigma(\eta) \text{ and } p(0|\eta) = 1 - \sigma(\eta)\,, \tag{16}$$

where $\eta = \mathbf{x}^\top \mathbf{w}$. This can be compactly written as (exercise: express it as the format of $\exp(\ldots)$)

$$p(y|\eta) = \frac{e^{\eta y}}{1 + e^{\eta}} = \exp\left(\eta y - \ln(1 + e^{\eta})\right) \tag{17}$$

- The linear model predicts $\sigma(\eta)$ which is not the mean of the distribution.
- Rather η is related to the mean μ by the non-linear relation $\eta = \ln \frac{\mu}{1-\mu}$ or $\mu = \sigma(\eta)$.
- The relation between
  - η (the parameter predicted by the linear model)
  - μ (the distribution's mean)

  makes possible to use linear model in this context.

It is called the **link function**.

A unified framework: exponential family

The distribution used in Logistic Regression can be written in a very specific form:

$$p(y|\eta) = \frac{e^{\eta y}}{1+e^{\eta}} = \exp\left(\eta y - \ln(1+e^{\eta})\right) \tag{18}$$

A unified framework: exponential family

The distribution used in Logistic Regression can be written in a very specific form:

$$p(y|\eta) = \frac{e^{\eta y}}{1 + e^{\eta}} = \exp\left(\eta y - \ln(1 + e^{\eta})\right) \tag{18}$$

Goals: a unified framework to generalize other forms of distributions.

A unified framework: exponential family

The distribution used in Logistic Regression can be written in a very specific form:

$$p(y|\eta) = \frac{e^{\eta y}}{1 + e^{\eta}} = \exp\left(\eta y - \ln(1 + e^{\eta})\right) \tag{18}$$

Goals: a unified framework to generalize other forms of distributions.

- The discussion on a class of distributions, known as *exponential families*.

A unified framework: exponential family

The distribution used in Logistic Regression can be written in a very specific form:

$$p(y|\eta) = \frac{e^{\eta y}}{1+e^{\eta}} = \exp\left(\eta y - \ln(1+e^{\eta})\right) \tag{18}$$

Goals: a unified framework to generalize other forms of distributions.

- The discussion on a class of distributions, known as *exponential families*.
- Many distributions (but not all) fit into this framework and that distributions in this family have many nice properties.

Table of Contents

Exponential family — definition

A distribution belongs to the exponential family if it can be written in the form

$$p(y|\boldsymbol{\eta}) = \underbrace{h(y)}_{\geq 0} \exp\left[\boldsymbol{\eta}^\top \boldsymbol{\phi}(y) - A(\boldsymbol{\eta})\right] \tag{19}$$

[1] Assume that we are given independent samples from this distribution. We do know $\boldsymbol{\phi}(y)$ and $h(y)$ but not $\boldsymbol{\eta}$. In order to optimally estimate $\boldsymbol{\eta}$ given these samples, all we need is the empirical average of the $\boldsymbol{\phi}(y)$.

Exponential family — definition

A distribution belongs to the exponential family if it can be written in the form

$$p(y|\boldsymbol{\eta}) = \underbrace{h(y)}_{\geq 0} \exp\left[\boldsymbol{\eta}^\top \boldsymbol{\phi}(y) - A(\boldsymbol{\eta})\right] \tag{19}$$

- $\boldsymbol{\eta}$: natural parameter of the distribution

[1] Assume that we are given independent samples from this distribution. We do know $\boldsymbol{\phi}(y)$ and $h(y)$ but not $\boldsymbol{\eta}$. In order to optimally estimate $\boldsymbol{\eta}$ given these samples, all we need is the empirical average of the $\boldsymbol{\phi}(y)$.

Exponential family — definition

A distribution belongs to the exponential family if it can be written in the form

$$p(y|\boldsymbol{\eta}) = \underbrace{h(y)}_{\geq 0} \exp\left[\boldsymbol{\eta}^\top \boldsymbol{\phi}(y) - A(\boldsymbol{\eta})\right] \tag{19}$$

- $\boldsymbol{\eta}$: natural parameter of the distribution (encodes the parameters of the distribution in a natural form)

[1] Assume that we are given independent samples from this distribution. We do know $\boldsymbol{\phi}(y)$ and $h(y)$ but not $\boldsymbol{\eta}$. In order to optimally estimate $\boldsymbol{\eta}$ given these samples, all we need is the empirical average of the $\boldsymbol{\phi}(y)$.

Exponential family — definition

A distribution belongs to the exponential family if it can be written in the form

$$p(y|\boldsymbol{\eta}) = \underbrace{h(y)}_{\geq 0} \exp\left[\boldsymbol{\eta}^\top \boldsymbol{\phi}(y) - A(\boldsymbol{\eta})\right] \tag{19}$$

- $\boldsymbol{\eta}$: natural parameter of the distribution (encodes the parameters of the distribution in a natural form)
- $\phi(y)$: sufficient statistics[1], containing all the relevant information

[1] Assume that we are given independent samples from this distribution. We do know $\phi(y)$ and $h(y)$ but not $\boldsymbol{\eta}$. In order to optimally estimate $\boldsymbol{\eta}$ given these samples, all we need is the empirical average of the $\phi(y)$.

Exponential family — definition

A distribution belongs to the exponential family if it can be written in the form

$$p(y|\boldsymbol{\eta}) = \underbrace{h(y)}_{\geq 0} \exp\left[\boldsymbol{\eta}^\top \boldsymbol{\phi}(y) - A(\boldsymbol{\eta})\right] \tag{19}$$

- $\boldsymbol{\eta}$: natural parameter of the distribution (encodes the parameters of the distribution in a natural form)
- $\phi(y)$: sufficient statistics[1], containing all the relevant information to estimate η.

[1] Assume that we are given independent samples from this distribution. We do know $\phi(y)$ and $h(y)$ but not $\boldsymbol{\eta}$. In order to optimally estimate $\boldsymbol{\eta}$ given these samples, all we need is the empirical average of the $\phi(y)$.

Exponential family — definition

A distribution belongs to the exponential family if it can be written in the form

$$p(y|\boldsymbol{\eta}) = \underbrace{h(y)}_{\geq 0} \exp\left[\boldsymbol{\eta}^\top \boldsymbol{\phi}(y) - A(\boldsymbol{\eta})\right] \tag{19}$$

- $\boldsymbol{\eta}$: natural parameter of the distribution (encodes the parameters of the distribution in a natural form)
- $\phi(y)$: sufficient statistics[1], containing all the relevant information to estimate η.
- $h(y)$: the base measure (a scaling factor independent of η)

[1] Assume that we are given independent samples from this distribution. We do know $\phi(y)$ and $h(y)$ but not $\boldsymbol{\eta}$. In order to optimally estimate $\boldsymbol{\eta}$ given these samples, all we need is the empirical average of the $\phi(y)$.

Exponential family — definition

A distribution belongs to the exponential family if it can be written in the form

$$p(y|\boldsymbol{\eta}) = \underbrace{h(y)}_{\geq 0} \exp\left[\boldsymbol{\eta}^\top \boldsymbol{\phi}(y) - A(\boldsymbol{\eta})\right] \tag{19}$$

- $\boldsymbol{\eta}$: natural parameter of the distribution (encodes the parameters of the distribution in a natural form)
- $\boldsymbol{\phi}(y)$: sufficient statistics[1], containing all the relevant information to estimate η.
- $h(y)$: the base measure (a scaling factor independent of η)
- $A(\boldsymbol{\eta})$: log-partition function, the quantity $e^{-A(\boldsymbol{\eta})}$ is used as a normalization constant:

[1] Assume that we are given independent samples from this distribution. We do know $\boldsymbol{\phi}(y)$ and $h(y)$ but not $\boldsymbol{\eta}$. In order to optimally estimate $\boldsymbol{\eta}$ given these samples, all we need is the empirical average of the $\boldsymbol{\phi}(y)$.

Exponential family — definition

A distribution belongs to the exponential family if it can be written in the form

$$p(y|\boldsymbol{\eta}) = \underbrace{h(y)}_{\geq 0} \exp\left[\boldsymbol{\eta}^\top \boldsymbol{\phi}(y) - A(\boldsymbol{\eta})\right] \tag{19}$$

- $\boldsymbol{\eta}$: natural parameter of the distribution (encodes the parameters of the distribution in a natural form)
- $\phi(y)$: sufficient statistics[1], containing all the relevant information to estimate η.
- $h(y)$: the base measure (a scaling factor independent of η)
- $A(\boldsymbol{\eta})$: log-partition function, the quantity $e^{-A(\boldsymbol{\eta})}$ is used as a normalization constant:

$$\text{(we need)} \int p(y|\boldsymbol{\eta})dy = \int h(y) \exp\left[\boldsymbol{\eta}^\top \boldsymbol{\phi}(y) - A(\boldsymbol{\eta})\right] dy = 1 \tag{20}$$

$$\tag{21}$$

[1] Assume that we are given independent samples from this distribution. We do know $\phi(y)$ and $h(y)$ but not $\boldsymbol{\eta}$. In order to optimally estimate $\boldsymbol{\eta}$ given these samples, all we need is the empirical average of the $\phi(y)$.

Exponential family — definition

A distribution belongs to the exponential family if it can be written in the form

$$p(y|\boldsymbol{\eta}) = \underbrace{h(y)}_{\geq 0} \exp\left[\boldsymbol{\eta}^\top \boldsymbol{\phi}(y) - A(\boldsymbol{\eta})\right] \tag{19}$$

- $\boldsymbol{\eta}$: natural parameter of the distribution (encodes the parameters of the distribution in a natural form)
- $\phi(y)$: sufficient statistics[1], containing all the relevant information to estimate η.
- $h(y)$: the base measure (a scaling factor independent of η)
- $A(\boldsymbol{\eta})$: log-partition function, the quantity $e^{-A(\boldsymbol{\eta})}$ is used as a normalization constant:

$$\text{(we need)} \int p(y|\boldsymbol{\eta})dy = \int h(y) \exp\left[\boldsymbol{\eta}^\top \boldsymbol{\phi}(y) - A(\boldsymbol{\eta})\right] dy = 1 \tag{20}$$

$$\Longrightarrow \text{(exercise)} \int h(y) \exp\left[\boldsymbol{\eta}^\top \boldsymbol{\phi}(y)\right] dy = \int h(y) \exp\left[A(\boldsymbol{\eta})\right] dy = \exp\left[A(\boldsymbol{\eta})\right] \tag{21}$$

[1] Assume that we are given independent samples from this distribution. We do know $\phi(y)$ and $h(y)$ but not $\boldsymbol{\eta}$. In order to optimally estimate $\boldsymbol{\eta}$ given these samples, all we need is the empirical average of the $\phi(y)$.

A distribution belongs to the exponential family if it can be written in the form

$$p(y|\boldsymbol{\eta}) = \underbrace{h(y)}_{\geq 0} \exp\left[\boldsymbol{\eta}^\top \boldsymbol{\phi}(y) - A(\boldsymbol{\eta})\right] \tag{22}$$

A distribution belongs to the exponential family if it can be written in the form

$$p(y|\boldsymbol{\eta}) = \underbrace{h(y)}_{\geq 0} \exp\left[\boldsymbol{\eta}^\top \boldsymbol{\phi}(y) - A(\boldsymbol{\eta})\right] \tag{22}$$

- A fixed choice of $\phi(y)$, $A(\boldsymbol{\eta})$ and $h(y)$ defines a family of distributions (parameterized by $\boldsymbol{\eta}$).

A distribution belongs to the exponential family if it can be written in the form

$$p(y|\boldsymbol{\eta}) = \underbrace{h(y)}_{\geq 0} \exp\left[\boldsymbol{\eta}^\top \boldsymbol{\phi}(y) - A(\boldsymbol{\eta})\right] \tag{22}$$

- A fixed choice of $\phi(y)$, $A(\boldsymbol{\eta})$ and $h(y)$ defines a family of distributions (parameterized by $\boldsymbol{\eta}$).
- As we vary $\boldsymbol{\eta}$, we then get different distribution within this family.

A distribution belongs to the exponential family if it can be written in the form

$$p(y|\boldsymbol{\eta}) = \underbrace{h(y)}_{\geq 0} \exp\left[\boldsymbol{\eta}^\top \boldsymbol{\phi}(y) - A(\boldsymbol{\eta})\right] \tag{22}$$

- A fixed choice of $\boldsymbol{\phi}(y)$, $A(\boldsymbol{\eta})$ and $h(y)$ defines a family of distributions (parameterized by $\boldsymbol{\eta}$).
- As we vary $\boldsymbol{\eta}$, we then get different distribution within this family.
- For some parameters $\boldsymbol{\eta}$, there exists some $h(y) \exp\left[\boldsymbol{\eta}^\top \boldsymbol{\phi}(y)\right]$ that cannot be normalized.

A distribution belongs to the exponential family if it can be written in the form

$$p(y|\boldsymbol{\eta}) = \underbrace{h(y)}_{\geq 0} \exp\left[\boldsymbol{\eta}^\top \boldsymbol{\phi}(y) - A(\boldsymbol{\eta})\right] \tag{22}$$

- A fixed choice of $\boldsymbol{\phi}(y)$, $A(\boldsymbol{\eta})$ and $h(y)$ defines a family of distributions (parameterized by $\boldsymbol{\eta}$).
- As we vary $\boldsymbol{\eta}$, we then get different distribution within this family.
- For some parameters $\boldsymbol{\eta}$, there exists some $h(y)\exp\left[\boldsymbol{\eta}^\top \boldsymbol{\phi}(y)\right]$ that cannot be normalized.

  For example, $h(y) = 1, \boldsymbol{\phi}(y) = y^2$ and $\boldsymbol{\eta} = 1$.

A distribution belongs to the exponential family if it can be written in the form

$$p(y|\boldsymbol{\eta}) = \underbrace{h(y)}_{\geq 0} \exp\left[\boldsymbol{\eta}^\top \boldsymbol{\phi}(y) - A(\boldsymbol{\eta})\right] \tag{22}$$

- A fixed choice of $\boldsymbol{\phi}(y)$, $A(\boldsymbol{\eta})$ and $h(y)$ defines a family of distributions (parameterized by $\boldsymbol{\eta}$).
- As we vary $\boldsymbol{\eta}$, we then get different distribution within this family.
- For some parameters $\boldsymbol{\eta}$, there exists some $h(y)\exp\left[\boldsymbol{\eta}^\top \boldsymbol{\phi}(y)\right]$ that cannot be normalized.

  For example, $h(y) = 1, \boldsymbol{\phi}(y) = y^2$ and $\boldsymbol{\eta} = 1$.

We will exclude such parameters by only looking at the set of parameters

$$\text{Natural parameter space } M := \left\{\boldsymbol{\eta} : \int_y h(y)\exp\left[\boldsymbol{\eta}^\top \boldsymbol{\phi}(y)\right] dy < \infty\right\} \tag{23}$$

Why?

Bernoulli distributions belong to the exponential family

Recall: A distribution belongs to the exponential family if it can be written in the form

$$p(y|\boldsymbol{\eta}) = h(y)\exp\left[\boldsymbol{\eta}^\top \boldsymbol{\phi}(y) - A(\boldsymbol{\eta})\right] \tag{24}$$

Bernoulli distributions belong to the exponential family

Recall: A distribution belongs to the exponential family if it can be written in the form

$$p(y|\boldsymbol{\eta}) = h(y) \exp\left[\boldsymbol{\eta}^\top \boldsymbol{\phi}(y) - A(\boldsymbol{\eta})\right] \tag{24}$$

The Bernoulli distribution is the binary random variable such that for $\mu \in [0, 1]$:

$$\Pr(Y = 1) = \mu \qquad \text{and} \qquad \Pr(Y = 0) = 1 - \mu \tag{25}$$

Bernoulli distributions belong to the exponential family

Recall: A distribution belongs to the exponential family if it can be written in the form

$$p(y|\boldsymbol{\eta}) = h(y) \exp\left[\boldsymbol{\eta}^\top \boldsymbol{\phi}(y) - A(\boldsymbol{\eta})\right] \tag{24}$$

The Bernoulli distribution is the binary random variable such that for $\mu \in [0, 1]$:

$$\Pr(Y = 1) = \mu \qquad \text{and} \qquad \Pr(Y = 0) = 1 - \mu \tag{25}$$

Claim: the Bernoulli distribution is a member of the exponential family.

$$p(y|\mu) = \mu^y (1-\mu)^{1-y}, \text{ where } \mu \in (0, 1) \tag{26}$$

$$= \exp\left\{(\ln \frac{\mu}{1-\mu})y + \ln(1-\mu)\right\} = \exp\left\{\eta\phi(y) - A(\eta)\right\}. \tag{27}$$

Bernoulli distributions belong to the exponential family

Recall: A distribution belongs to the exponential family if it can be written in the form

$$p(y|\boldsymbol{\eta}) = h(y)\exp\left[\boldsymbol{\eta}^\top \boldsymbol{\phi}(y) - A(\boldsymbol{\eta})\right] \tag{24}$$

The Bernoulli distribution is the binary random variable such that for $\mu \in [0,1]$:

$$\Pr(Y=1) = \mu \qquad \text{and} \qquad \Pr(Y=0) = 1-\mu \tag{25}$$

Claim: the Bernoulli distribution is a member of the exponential family.

$$p(y|\mu) = \mu^y(1-\mu)^{1-y}, \text{ where } \mu \in (0,1) \tag{26}$$

$$= \exp\left\{(\ln\frac{\mu}{1-\mu})y + \ln(1-\mu)\right\} = \exp\left\{\eta\phi(y) - A(\eta)\right\}. \tag{27}$$

where we can identify (exercise):

Bernoulli distributions belong to the exponential family

Recall: A distribution belongs to the exponential family if it can be written in the form

$$p(y|\boldsymbol{\eta}) = h(y) \exp\left[\boldsymbol{\eta}^\top \boldsymbol{\phi}(y) - A(\boldsymbol{\eta})\right] \tag{24}$$

The Bernoulli distribution is the binary random variable such that for $\mu \in [0, 1]$:

$$\Pr(Y = 1) = \mu \qquad \text{and} \qquad \Pr(Y = 0) = 1 - \mu \tag{25}$$

Claim: the Bernoulli distribution is a member of the exponential family.

$$p(y|\mu) = \mu^y (1-\mu)^{1-y}, \text{ where } \mu \in (0,1) \tag{26}$$

$$= \exp\left\{ (\ln \frac{\mu}{1-\mu}) y + \ln(1-\mu) \right\} = \exp\left\{\eta\phi(y) - A(\eta)\right\}. \tag{27}$$

where we can identify (exercise):

$$\phi(y) = y\,, \quad \eta = \ln \tfrac{\mu}{1-\mu}\,, \quad A(\eta) = -\ln(1-\mu) = \ln(1+e^\eta)\,, \quad h(y) = 1. \tag{28}$$

Bernoulli distributions belong to the exponential family

Recall: A distribution belongs to the exponential family if it can be written in the form

$$p(y|\boldsymbol{\eta}) = h(y) \exp\left[\boldsymbol{\eta}^\top \boldsymbol{\phi}(y) - A(\boldsymbol{\eta})\right] \tag{24}$$

The Bernoulli distribution is the binary random variable such that for $\mu \in [0, 1]$:

$$\Pr(Y = 1) = \mu \qquad \text{and} \qquad \Pr(Y = 0) = 1 - \mu \tag{25}$$

Claim: the Bernoulli distribution is a member of the exponential family.

$$p(y|\mu) = \mu^y (1-\mu)^{1-y}, \text{ where } \mu \in (0, 1) \tag{26}$$

$$= \exp\left\{(\ln \frac{\mu}{1-\mu})y + \ln(1-\mu)\right\} = \exp\left\{\eta\phi(y) - A(\eta)\right\}. \tag{27}$$

where we can identify (exercise):

$$\phi(y) = y\,, \quad \eta = \ln \tfrac{\mu}{1-\mu}\,, \quad A(\eta) = -\ln(1-\mu) = \ln(1+e^\eta)\,, \quad h(y) = 1. \tag{28}$$

$\Longrightarrow \eta = g(\mu) = \ln \frac{\mu}{1-\mu} \Leftrightarrow \mu = g^{-1}(\eta) = \frac{e^\eta}{1+e^\eta}.$

Bernoulli distributions belong to the exponential family

Recall: A distribution belongs to the exponential family if it can be written in the form

$$p(y|\boldsymbol{\eta}) = h(y) \exp\left[\boldsymbol{\eta}^\top \boldsymbol{\phi}(y) - A(\boldsymbol{\eta})\right] \tag{24}$$

The Bernoulli distribution is the binary random variable such that for $\mu \in [0, 1]$:

$$\Pr(Y = 1) = \mu \qquad \text{and} \qquad \Pr(Y = 0) = 1 - \mu \tag{25}$$

Claim: the Bernoulli distribution is a member of the exponential family.

$$p(y|\mu) = \mu^y (1-\mu)^{1-y}, \text{ where } \mu \in (0, 1) \tag{26}$$

$$= \exp\left\{(\ln \frac{\mu}{1-\mu})y + \ln(1-\mu)\right\} = \exp\left\{\eta\phi(y) - A(\eta)\right\}. \tag{27}$$

where we can identify (exercise):

$$\phi(y) = y\,, \quad \eta = \ln \tfrac{\mu}{1-\mu}\,, \quad A(\eta) = -\ln(1-\mu) = \ln(1+e^\eta)\,, \quad h(y) = 1. \tag{28}$$

$\Longrightarrow \eta = g(\mu) = \ln \frac{\mu}{1-\mu} \Leftrightarrow \mu = g^{-1}(\eta) = \frac{e^\eta}{1+e^\eta}$. Link function $g(\mu)$ links the mean of $\phi(y)$ to $\boldsymbol{\eta}$.

Gaussian distributions belong to the exponential family

Recall: A distribution belongs to the exponential family if it can be written in the form

$$p(y|\boldsymbol{\eta}) = h(y) \exp\left[\boldsymbol{\eta}^\top \phi(y) - A(\boldsymbol{\eta})\right] \tag{29}$$

Gaussian distributions belong to the exponential family

Recall: A distribution belongs to the exponential family if it can be written in the form

$$p(y|\boldsymbol{\eta}) = h(y) \exp\left[\boldsymbol{\eta}^\top \phi(y) - A(\boldsymbol{\eta})\right] \tag{29}$$

Claim: the Gaussian distribution with mean μ and variance σ^2 is also a member of the exponential family.

Gaussian distributions belong to the exponential family

Recall: A distribution belongs to the exponential family if it can be written in the form

$$p(y|\boldsymbol{\eta}) = h(y)\exp\left[\boldsymbol{\eta}^\top \boldsymbol{\phi}(y) - A(\boldsymbol{\eta})\right] \tag{29}$$

Claim: the Gaussian distribution with mean μ and variance σ^2 is also a member of the exponential family.

$$\begin{aligned} p(y|\mu) &= \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y-\mu)^2}{2\sigma^2}}, \qquad \mu \in \mathbb{R}, \sigma^2 \in \mathbb{R}^+ &(30)\\ &= \exp\left[\left(\mu/\sigma^2, -1/(2\sigma^2)\right)(y, y^2)^\top - \frac{\mu^2}{2\sigma^2} - \frac{1}{2}\ln(2\pi\sigma^2)\right]. &(31)\end{aligned}$$

Gaussian distributions belong to the exponential family

Recall: A distribution belongs to the exponential family if it can be written in the form

$$p(y|\boldsymbol{\eta}) = h(y) \exp\left[\boldsymbol{\eta}^\top \phi(y) - A(\boldsymbol{\eta})\right] \tag{29}$$

Claim: the Gaussian distribution with mean μ and variance σ^2 is also a member of the exponential family.

$$p(y|\mu) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y-\mu)^2}{2\sigma^2}}, \qquad \mu \in \mathbb{R}, \sigma^2 \in \mathbb{R}^+ \tag{30}$$

$$= \exp\left[\left(\mu/\sigma^2, -1/(2\sigma^2)\right)(y, y^2)^\top - \frac{\mu^2}{2\sigma^2} - \frac{1}{2}\ln(2\pi\sigma^2)\right]. \tag{31}$$

$$\phi(y) = (y, y^2)^\top, \ \boldsymbol{\eta} = (\eta_1 = \mu/\sigma^2, \eta_2 = -1/(2\sigma^2))^\top, \ A(\boldsymbol{\eta}) = -\frac{\eta_1^2}{4\eta_2} - \frac{1}{2}\ln(-\eta_2/\pi), \ h(y) = 1. \tag{32}$$

Gaussian distributions belong to the exponential family

Recall: A distribution belongs to the exponential family if it can be written in the form

$$p(y|\boldsymbol{\eta}) = h(y)\exp\left[\boldsymbol{\eta}^\top \boldsymbol{\phi}(y) - A(\boldsymbol{\eta})\right] \tag{29}$$

Claim: the Gaussian distribution with mean μ and variance σ^2 is also a member of the exponential family.

$$p(y|\mu) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y-\mu)^2}{2\sigma^2}}, \qquad \mu \in \mathbb{R}, \sigma^2 \in \mathbb{R}^+ \tag{30}$$

$$= \exp\left[\left(\mu/\sigma^2, -1/(2\sigma^2)\right)(y, y^2)^\top - \frac{\mu^2}{2\sigma^2} - \frac{1}{2}\ln(2\pi\sigma^2)\right]. \tag{31}$$

$$\boldsymbol{\phi}(y) = (y, y^2)^\top,\ \boldsymbol{\eta} = (\eta_1 = \mu/\sigma^2, \eta_2 = -1/(2\sigma^2))^\top,\ A(\boldsymbol{\eta}) = -\frac{\eta_1^2}{4\eta_2} - \frac{1}{2}\ln(-\eta_2/\pi),\ h(y) = 1. \tag{32}$$

Link function: $\eta_1 = \frac{\mu}{\sigma^2}, \eta_2 = -\frac{1}{2\sigma^2} \iff \mu = -\frac{\eta_1}{2\eta_2}, \sigma^2 = -\frac{1}{2\eta_2}$.

Poisson distributions belong to the exponential family

Recall: A distribution belongs to the exponential family if it can be written in the form

$$p(y|\boldsymbol{\eta}) = h(y) \exp\left[\boldsymbol{\eta}^\top \boldsymbol{\phi}(y) - A(\boldsymbol{\eta})\right] \tag{33}$$

Poisson distributions belong to the exponential family

Recall: A distribution belongs to the exponential family if it can be written in the form

$$p(y|\boldsymbol{\eta}) = h(y)\exp\left[\boldsymbol{\eta}^\top \boldsymbol{\phi}(y) - A(\boldsymbol{\eta})\right] \tag{33}$$

Claim: the Poisson distribution with mean μ belongs to the family: for $y \in \mathbb{N}$

Poisson distributions belong to the exponential family

Recall: A distribution belongs to the exponential family if it can be written in the form

$$p(y|\boldsymbol{\eta}) = h(y) \exp\left[\boldsymbol{\eta}^\top \boldsymbol{\phi}(y) - A(\boldsymbol{\eta})\right] \tag{33}$$

Claim: the Poisson distribution with mean μ belongs to the family: for $y \in \mathbb{N}$

$$p(y|\mu) = \frac{\mu^y e^{-\mu}}{y!} = \frac{1}{y!} e^{y \ln(\mu) - \mu} = h(y) e^{\boldsymbol{\eta}\boldsymbol{\phi}(y) - A(\boldsymbol{\eta})} \tag{34}$$

Poisson distributions belong to the exponential family

Recall: A distribution belongs to the exponential family if it can be written in the form

$$p(y|\boldsymbol{\eta}) = h(y) \exp\left[\boldsymbol{\eta}^\top \boldsymbol{\phi}(y) - A(\boldsymbol{\eta})\right] \tag{33}$$

Claim: the Poisson distribution with mean μ belongs to the family: for $y \in \mathbb{N}$

$$p(y|\mu) = \frac{\mu^y e^{-\mu}}{y!} = \frac{1}{y!} e^{y \ln(\mu) - \mu} = h(y) e^{\boldsymbol{\eta}\boldsymbol{\phi}(y) - A(\boldsymbol{\eta})} \tag{34}$$

We can identify (exercise):

$$h(y) = \frac{1}{y!}\,,\ \phi(y) = y\,,\ \text{ and } \eta = \ln(\mu) \tag{35}$$

Poisson distributions belong to the exponential family

Recall: A distribution belongs to the exponential family if it can be written in the form

$$p(y|\boldsymbol{\eta}) = h(y) \exp\left[\boldsymbol{\eta}^\top \boldsymbol{\phi}(y) - A(\boldsymbol{\eta})\right] \tag{33}$$

Claim: the Poisson distribution with mean μ belongs to the family: for $y \in \mathbb{N}$

$$p(y|\mu) = \frac{\mu^y e^{-\mu}}{y!} = \frac{1}{y!} e^{y \ln(\mu) - \mu} = h(y) e^{\boldsymbol{\eta}\boldsymbol{\phi}(y) - A(\boldsymbol{\eta})} \tag{34}$$

We can identify (exercise):

$$h(y) = \frac{1}{y!}\,,\ \phi(y) = y\,,\ \text{ and } \eta = \ln(\mu) \tag{35}$$

Link function (exercise):

$$\boldsymbol{\eta} = g(\mu) \tag{36}$$

Poisson distributions belong to the exponential family

Recall: A distribution belongs to the exponential family if it can be written in the form

$$p(y|\boldsymbol{\eta}) = h(y) \exp\left[\boldsymbol{\eta}^\top \boldsymbol{\phi}(y) - A(\boldsymbol{\eta})\right] \tag{33}$$

Claim: the Poisson distribution with mean μ belongs to the family: for $y \in \mathbb{N}$

$$p(y|\mu) = \frac{\mu^y e^{-\mu}}{y!} = \frac{1}{y!} e^{y \ln(\mu) - \mu} = h(y) e^{\boldsymbol{\eta} \boldsymbol{\phi}(y) - A(\boldsymbol{\eta})} \tag{34}$$

We can identify (exercise):

$$h(y) = \frac{1}{y!}\,,\ \phi(y) = y\,,\ \text{ and } \eta = \ln(\mu) \tag{35}$$

Link function (exercise):

$$\boldsymbol{\eta} = g(\mu) = \ln(\mu) \tag{36}$$

Poisson distributions belong to the exponential family

Recall: A distribution belongs to the exponential family if it can be written in the form

$$p(y|\boldsymbol{\eta}) = h(y) \exp\left[\boldsymbol{\eta}^\top \boldsymbol{\phi}(y) - A(\boldsymbol{\eta})\right] \tag{33}$$

Claim: the Poisson distribution with mean μ belongs to the family: for $y \in \mathbb{N}$

$$p(y|\mu) = \frac{\mu^y e^{-\mu}}{y!} = \frac{1}{y!} e^{y \ln(\mu) - \mu} = h(y) e^{\boldsymbol{\eta} \boldsymbol{\phi}(y) - A(\boldsymbol{\eta})} \tag{34}$$

We can identify (exercise):

$$h(y) = \frac{1}{y!}\,,\ \boldsymbol{\phi}(y) = y\,,\ \text{ and } \eta = \ln(\mu) \tag{35}$$

Link function (exercise):

$$\boldsymbol{\eta} = g(\mu) = \ln(\mu) \iff \mu = g^{-1}(\boldsymbol{\eta}) = e^{\boldsymbol{\eta}} \tag{36}$$

Basic properties

- Cumulant $A(\boldsymbol{\eta})$ is convex.

Basic properties

- Cumulant $A(\boldsymbol{\eta})$ is convex.
- $\nabla A(\boldsymbol{\eta}) = \mathbb{E}\left[\phi(y)\right]$

Basic properties

- Cumulant $A(\boldsymbol{\eta})$ is convex.
- $\nabla A(\boldsymbol{\eta}) = \mathbb{E}\left[\phi(y)\right]$
- $\nabla^2 A(\boldsymbol{\eta}) = \mathbb{E}\left[\phi(y)\phi(y)^\top\right] - \mathbb{E}\left[\phi(y)\right]\mathbb{E}\left[\phi(y)\right]^\top$

Basic properties

- Cumulant $A(\boldsymbol{\eta})$ is convex.

- $\nabla A(\boldsymbol{\eta}) = \mathbb{E}\left[\boldsymbol{\phi}(y)\right]$

- $\nabla^2 A(\boldsymbol{\eta}) = \mathbb{E}\left[\boldsymbol{\phi}(y)\boldsymbol{\phi}(y)^\top\right] - \mathbb{E}\left[\boldsymbol{\phi}(y)\right]\mathbb{E}\left[\boldsymbol{\phi}(y)\right]^\top$

- There is a $1-1$ relationship between the "mean" $\boldsymbol{\mu} := \mathbb{E}\left[\boldsymbol{\phi}(y)\right]$ and natural parameter $\boldsymbol{\eta}$, defined using a so-called *link function* $\mathbf{g}$:

Basic properties

- Cumulant $A(\boldsymbol{\eta})$ is convex.
- $\nabla A(\boldsymbol{\eta}) = \mathbb{E}\left[\phi(y)\right]$
- $\nabla^2 A(\boldsymbol{\eta}) = \mathbb{E}\left[\phi(y)\phi(y)^\top\right] - \mathbb{E}\left[\phi(y)\right]\mathbb{E}\left[\phi(y)\right]^\top$
- There is a $1-1$ relationship between the “mean” $\boldsymbol{\mu} := \mathbb{E}\left[\phi(y)\right]$ and natural parameter $\boldsymbol{\eta}$, defined using a so-called *link function* $\mathbf{g}$:

$$\boldsymbol{\eta} = \mathbf{g}\left(\boldsymbol{\mu} := \mathbb{E}\left[\phi(y)\right]\right) \Longleftrightarrow \boldsymbol{\mu} = \mathbf{g}^{-1}(\boldsymbol{\eta}) = \nabla A(\boldsymbol{\eta}) \tag{37}$$

Proof: convexity

- For $\boldsymbol{\eta}_1, \boldsymbol{\eta}_2$ two parameters, we define $\boldsymbol{\eta} = \lambda\boldsymbol{\eta}_1 + (1-\lambda)\boldsymbol{\eta}_2$. We want to show

$$A(\boldsymbol{\eta}) \leq \lambda A(\boldsymbol{\eta}_1) + (1-\lambda)A(\boldsymbol{\eta}_2) \tag{38}$$

Proof: convexity

- For $\boldsymbol{\eta}_1, \boldsymbol{\eta}_2$ two parameters, we define $\boldsymbol{\eta} = \lambda\boldsymbol{\eta}_1 + (1-\lambda)\boldsymbol{\eta}_2$. We want to show

$$A(\boldsymbol{\eta}) \leq \lambda A(\boldsymbol{\eta}_1) + (1-\lambda)A(\boldsymbol{\eta}_2) \tag{38}$$

- We have first

$$\exp A(\boldsymbol{\eta}) = \int h(y) \exp\left(\boldsymbol{\eta}^\top \boldsymbol{\phi}(y)\right) dy \tag{39}$$

(43)

Proof: convexity

- For $\boldsymbol{\eta}_1, \boldsymbol{\eta}_2$ two parameters, we define $\boldsymbol{\eta} = \lambda\boldsymbol{\eta}_1 + (1-\lambda)\boldsymbol{\eta}_2$. We want to show

$$A(\boldsymbol{\eta}) \leq \lambda A(\boldsymbol{\eta}_1) + (1-\lambda)A(\boldsymbol{\eta}_2) \tag{38}$$

- We have first

$$\exp A(\boldsymbol{\eta}) = \int h(y) \exp\left(\boldsymbol{\eta}^\top \boldsymbol{\phi}(y)\right) dy \tag{39}$$

$$= \int h(y) \exp\left((\lambda\boldsymbol{\eta}_1 + (1-\lambda)\boldsymbol{\eta}_2)^\top \boldsymbol{\phi}(y)\right) dy \tag{40}$$

(43)

Proof: convexity

- For $\boldsymbol{\eta}_1, \boldsymbol{\eta}_2$ two parameters, we define $\boldsymbol{\eta} = \lambda\boldsymbol{\eta}_1 + (1-\lambda)\boldsymbol{\eta}_2$. We want to show

$$A(\boldsymbol{\eta}) \leq \lambda A(\boldsymbol{\eta}_1) + (1-\lambda)A(\boldsymbol{\eta}_2) \tag{38}$$

- We have first

$$\exp A(\boldsymbol{\eta}) = \int h(y)\exp\left(\boldsymbol{\eta}^\top \boldsymbol{\phi}(y)\right) dy \tag{39}$$

$$= \int h(y)\exp\left(\left(\lambda\boldsymbol{\eta}_1 + (1-\lambda)\boldsymbol{\eta}_2\right)^\top \boldsymbol{\phi}(y)\right) dy \tag{40}$$

$$= \int \underbrace{\left[h(y)^\lambda \exp\left(\lambda\boldsymbol{\eta}_1^\top \boldsymbol{\phi}(y)\right)\right]}_{f(y)} \cdot \underbrace{\left[h(y)^{1-\lambda} \exp\left((1-\lambda)\boldsymbol{\eta}_2^\top \boldsymbol{\phi}(y)\right)\right]}_{g(y)} dy \tag{41}$$

$$\tag{43}$$

Proof: convexity

- For $\boldsymbol{\eta}_1, \boldsymbol{\eta}_2$ two parameters, we define $\boldsymbol{\eta} = \lambda\boldsymbol{\eta}_1 + (1-\lambda)\boldsymbol{\eta}_2$. We want to show

$$A(\boldsymbol{\eta}) \leq \lambda A(\boldsymbol{\eta}_1) + (1-\lambda)A(\boldsymbol{\eta}_2) \tag{38}$$

- We have first

$$\exp A(\boldsymbol{\eta}) = \int h(y) \exp\left(\boldsymbol{\eta}^\top \boldsymbol{\phi}(y)\right) dy \tag{39}$$

$$= \int h(y) \exp\left(\left(\lambda\boldsymbol{\eta}_1 + (1-\lambda)\boldsymbol{\eta}_2\right)^\top \boldsymbol{\phi}(y)\right) dy \tag{40}$$

$$= \int \underbrace{\left[h(y)^\lambda \exp\left(\lambda\boldsymbol{\eta}_1^\top \boldsymbol{\phi}(y)\right)\right]}_{f(y)} \cdot \underbrace{\left[h(y)^{1-\lambda} \exp\left((1-\lambda)\boldsymbol{\eta}_2^\top \boldsymbol{\phi}(y)\right)\right]}_{g(y)} dy \tag{41}$$

$$= \int f(y)g(y)dy \tag{42}$$

$$\tag{43}$$

Proof: convexity

- For $\boldsymbol{\eta}_1, \boldsymbol{\eta}_2$ two parameters, we define $\boldsymbol{\eta} = \lambda\boldsymbol{\eta}_1 + (1-\lambda)\boldsymbol{\eta}_2$. We want to show

$$A(\boldsymbol{\eta}) \leq \lambda A(\boldsymbol{\eta}_1) + (1-\lambda)A(\boldsymbol{\eta}_2) \tag{38}$$

- We have first

$$\exp A(\boldsymbol{\eta}) = \int h(y)\exp\left(\boldsymbol{\eta}^\top \boldsymbol{\phi}(y)\right)dy \tag{39}$$

$$= \int h(y)\exp\left(\left(\lambda\boldsymbol{\eta}_1 + (1-\lambda)\boldsymbol{\eta}_2\right)^\top \boldsymbol{\phi}(y)\right)dy \tag{40}$$

$$= \int \underbrace{\left[h(y)^\lambda \exp\left(\lambda\boldsymbol{\eta}_1^\top \boldsymbol{\phi}(y)\right)\right]}_{f(y)} \cdot \underbrace{\left[h(y)^{1-\lambda}\exp\left((1-\lambda)\boldsymbol{\eta}_2^\top \boldsymbol{\phi}(y)\right)\right]}_{g(y)} dy \tag{41}$$

$$= \int f(y)g(y)dy \tag{42}$$

$$= \|fg\|_1 \tag{43}$$

Proof: convexity

- For $\boldsymbol{\eta}_1, \boldsymbol{\eta}_2$ two parameters, we define $\boldsymbol{\eta} = \lambda\boldsymbol{\eta}_1 + (1-\lambda)\boldsymbol{\eta}_2$. We want to show
$$A(\boldsymbol{\eta}) \leq \lambda A(\boldsymbol{\eta}_1) + (1-\lambda)A(\boldsymbol{\eta}_2) \tag{38}$$
- We have first
$$\exp A(\boldsymbol{\eta}) = \int h(y)\exp\left(\boldsymbol{\eta}^\top \boldsymbol{\phi}(y)\right) dy \tag{39}$$
$$= \int h(y)\exp\left(\left(\lambda\boldsymbol{\eta}_1 + (1-\lambda)\boldsymbol{\eta}_2\right)^\top \boldsymbol{\phi}(y)\right) dy \tag{40}$$
$$= \int \underbrace{\left[h(y)^\lambda \exp\left(\lambda\boldsymbol{\eta}_1^\top \boldsymbol{\phi}(y)\right)\right]}_{f(y)} \cdot \underbrace{\left[h(y)^{1-\lambda}\exp\left((1-\lambda)\boldsymbol{\eta}_2^\top \boldsymbol{\phi}(y)\right)\right]}_{g(y)} dy \tag{41}$$
$$= \int f(y)g(y)dy \tag{42}$$
$$= \|fg\|_1 \tag{43}$$
- We will continue the proof with the Hoelder's inequality.

Proof: convexity

- We recall the **Hoelder's inequality:**

$$\|fg\|_1 \leq \|f\|_p \, \|g\|_q \tag{44}$$

for $p, q \in [1, +\infty]$ $s.t.$ $\frac{1}{p} + \frac{1}{q} = 1$, and $\|f\|_p = (\int |f(y)|^p \, dy)^{1/p}$.

Proof: convexity

- We recall the **Hoelder's inequality:**

$$\|fg\|_1 \leq \|f\|_p \|g\|_q \tag{44}$$

for $p, q \in [1, +\infty]$ $s.t.$ $\frac{1}{p} + \frac{1}{q} = 1$, and $\|f\|_p = (\int |f(y)|^p \, dy)^{1/p}$.

- We apply Hoelder's inequality to f and g for $p = 1/\lambda$ and $q = 1/(1-\lambda)$:

$$\|fg\|_1 \leq \|f\|_p \|g\|_q \quad \text{where} \quad 1/p + 1/q = \lambda + (1-\lambda) = 1 \tag{45}$$

Proof: convexity

- We recall the **Hoelder's inequality:**

$$\|fg\|_1 \leq \|f\|_p \, \|g\|_q \tag{44}$$

for $p, q \in [1, +\infty]$ $s.t.$ $\frac{1}{p} + \frac{1}{q} = 1$, and $\|f\|_p = (\int |f(y)|^p \, dy)^{1/p}$.

- We apply Hoelder's inequality to f and g for $p = 1/\lambda$ and $q = 1/(1-\lambda)$:

$$\|fg\|_1 \leq \|f\|_p \, \|g\|_q \quad \text{where} \quad 1/p + 1/q = \lambda + (1-\lambda) = 1 \tag{45}$$

- Note that

Proof: convexity

- We recall the **Hoelder's inequality:**

$$\|fg\|_1 \leq \|f\|_p \|g\|_q \tag{44}$$

  for $p, q \in [1, +\infty]$ $s.t.$ $\frac{1}{p} + \frac{1}{q} = 1$, and $\|f\|_p = (\int |f(y)|^p \, dy)^{1/p}$.
- We apply Hoelder's inequality to f and g for $p = 1/\lambda$ and $q = 1/(1-\lambda)$:

$$\|fg\|_1 \leq \|f\|_p \|g\|_q \quad \text{where} \quad 1/p + 1/q = \lambda + (1-\lambda) = 1 \tag{45}$$

- Note that

$$\|f\|_p = \left(\int f(y)^p dy\right)^{1/p} = \left(\int \left[h(y)^\lambda \exp\left(\lambda \boldsymbol{\eta}_1^\top \boldsymbol{\phi}(y)\right)\right]^{1/\lambda} dy\right)^\lambda = \left(\int h(y) \exp\left(\boldsymbol{\eta}_1^\top \boldsymbol{\phi}(y)\right) dy\right)^\lambda$$

Proof: convexity

- We recall the **Hoelder's inequality:**

$$\|fg\|_1 \leq \|f\|_p \, \|g\|_q \tag{44}$$

for $p, q \in [1, +\infty]$ $s.t.$ $\frac{1}{p} + \frac{1}{q} = 1$, and $\|f\|_p = (\int |f(y)|^p \, dy)^{1/p}$.

- We apply Hoelder's inequality to f and g for $p = 1/\lambda$ and $q = 1/(1-\lambda)$:

$$\|fg\|_1 \leq \|f\|_p \, \|g\|_q \quad \text{where} \quad 1/p + 1/q = \lambda + (1-\lambda) = 1 \tag{45}$$

- Note that

$$\|f\|_p = \left(\int f(y)^p dy \right)^{1/p} = \left(\int \left[h(y)^\lambda \exp\left(\lambda \boldsymbol{\eta}_1^\top \boldsymbol{\phi}(y) \right) \right]^{1/\lambda} dy \right)^\lambda = \left(\int h(y) \exp\left(\boldsymbol{\eta}_1^\top \boldsymbol{\phi}(y) \right) dy \right)^\lambda$$

$$\|g\|_q = \left(\int g(y)^q dy \right)^{1/q} = \left(\int \left[h(y)^{1-\lambda} \exp\left((1-\lambda) \boldsymbol{\eta}_2^\top \boldsymbol{\phi}(y) \right) \right]^{\frac{1}{1-\lambda}} dy \right)^{1-\lambda} = \left(\int h(y) \exp\left(\boldsymbol{\eta}_2^\top \boldsymbol{\phi}(y) \right) dy \right)^{1-\lambda}$$

Proof: convexity

- We recall the **Hoelder's inequality:**

$$\|fg\|_1 \leq \|f\|_p \, \|g\|_q \tag{44}$$

for $p, q \in [1, +\infty]$ $s.t.$ $\frac{1}{p} + \frac{1}{q} = 1$, and $\|f\|_p = (\int |f(y)|^p \, dy)^{1/p}$.

- We apply Hoelder's inequality to f and g for $p = 1/\lambda$ and $q = 1/(1-\lambda)$:

$$\|fg\|_1 \leq \|f\|_p \, \|g\|_q \quad \text{where} \quad 1/p + 1/q = \lambda + (1-\lambda) = 1 \tag{45}$$

- Note that

$$\|f\|_p = \left(\int f(y)^p dy \right)^{1/p} = \left(\int \left[h(y)^\lambda \exp\left(\lambda \boldsymbol{\eta}_1^\top \boldsymbol{\phi}(y) \right) \right]^{1/\lambda} dy \right)^\lambda = \left(\int h(y) \exp\left(\boldsymbol{\eta}_1^\top \boldsymbol{\phi}(y) \right) dy \right)^\lambda$$

$$\|g\|_q = \left(\int g(y)^q dy \right)^{1/q} = \left(\int \left[h(y)^{1-\lambda} \exp\left((1-\lambda) \boldsymbol{\eta}_2^\top \boldsymbol{\phi}(y) \right) \right]^{\frac{1}{1-\lambda}} dy \right)^{1-\lambda} = \left(\int h(y) \exp\left(\boldsymbol{\eta}_2^\top \boldsymbol{\phi}(y) \right) dy \right)^{1-\lambda}$$

- Therefore we have

$$\|f\|_p \, \|g\|_q = \left(\int h(y) \exp\left(\boldsymbol{\eta}_1^\top \boldsymbol{\phi}(y) \right) dy \right)^\lambda \left(\int h(y) \exp\left(\boldsymbol{\eta}_2^\top \boldsymbol{\phi}(y) \right) dy \right)^{1-\lambda} \tag{46}$$

$$= \exp\left(\lambda A(\boldsymbol{\eta}_1) \right) \exp\left((1-\lambda) A(\boldsymbol{\eta}_2) \right) \tag{47}$$

Summary of proof: convexity

We have

$$\exp A(\boldsymbol{\eta}) = \int h(y) \exp\left(\boldsymbol{\eta}^\top \boldsymbol{\phi}(y)\right) dy \tag{48}$$

$$= \int h(y) \exp\left(\left(\lambda\boldsymbol{\eta}_1 + (1-\lambda)\boldsymbol{\eta}_2\right)^\top \boldsymbol{\phi}(y)\right) dy \tag{49}$$

$$= \int \underbrace{\left[h(y)^\lambda \exp\left(\lambda\boldsymbol{\eta}_1^\top \boldsymbol{\phi}(y)\right)\right]}_{f(y)} \cdot \underbrace{\left[h(y)^{1-\lambda} \exp\left((1-\lambda)\boldsymbol{\eta}_2^\top \boldsymbol{\phi}(y)\right)\right]}_{g(y)} dy \tag{50}$$

$$\leq \left(\int h(y) \exp\left(\boldsymbol{\eta}_1^\top \boldsymbol{\phi}(y)\right) dy\right)^\lambda \left(\int h(y) \exp\left(\boldsymbol{\eta}_2^\top \boldsymbol{\phi}(y)\right) dy\right)^{1-\lambda} \tag{51}$$

$$= \exp\left(\lambda A(\boldsymbol{\eta}_1)\right) \exp\left((1-\lambda) A(\boldsymbol{\eta}_2)\right) \tag{52}$$

Derivate of $A(\boldsymbol{\eta})$ and moments: particular cases

- Bernoulli distribution:

$$A'(\boldsymbol{\eta}) = \frac{d}{d\boldsymbol{\eta}} \ln(1 + e^{\boldsymbol{\eta}}) = \frac{e^{\boldsymbol{\eta}}}{1 + e^{\boldsymbol{\eta}}} = \sigma(\boldsymbol{\eta}) = \mu \tag{53}$$

$$A''(\boldsymbol{\eta}) = \frac{d}{\boldsymbol{\eta}} \sigma(\boldsymbol{\eta}) = \sigma(\boldsymbol{\eta})\left(1 - \sigma(\boldsymbol{\eta})\right) = \mu(1 - \mu) \tag{54}$$

Derivate of $A(\boldsymbol{\eta})$ and moments: particular cases

- Bernoulli distribution:

$$A'(\boldsymbol{\eta}) = \frac{d}{d\boldsymbol{\eta}} \ln(1 + e^{\boldsymbol{\eta}}) = \frac{e^{\boldsymbol{\eta}}}{1 + e^{\boldsymbol{\eta}}} = \sigma(\boldsymbol{\eta}) = \mu \tag{53}$$

$$A''(\boldsymbol{\eta}) = \frac{d}{\boldsymbol{\eta}} \sigma(\boldsymbol{\eta}) = \sigma(\boldsymbol{\eta})\left(1 - \sigma(\boldsymbol{\eta})\right) = \mu(1 - \mu) \tag{54}$$

- Gaussian distribution:

$$\frac{\partial}{\partial \eta_1} A(\boldsymbol{\eta}) = \frac{\partial}{\partial \eta_1}\left(-\frac{\eta_1^2}{4\eta_2} - \frac{1}{2}\ln\left(-\eta_2/\pi\right)\right) = -\frac{\eta_1}{2\eta_2} = \mu \tag{55}$$

$$\frac{\partial}{\partial \eta_2} A(\boldsymbol{\eta}) = \frac{\partial}{\partial \eta_2}\left(-\frac{\eta_1^2}{4\eta_2} - \frac{1}{2}\ln\left(-\eta_2/\pi\right)\right) = \frac{\eta_1^2}{4\eta_2^2} - \frac{1}{2\eta_2} = \mu^2 + \sigma^2 \tag{56}$$

$$\frac{\partial^2}{\partial \eta_1^2} A(\boldsymbol{\eta}) = \frac{\partial}{\partial \eta_1}\left(-\frac{\eta_1}{2\eta_2}\right) = -\frac{1}{2\eta_2} = \sigma^2 \tag{57}$$

Derivate of $A(\boldsymbol{\eta})$ and moments: general cases

$$\nabla A(\boldsymbol{\eta}) = \nabla \left[\ln \left(\int h(y) \exp \left(\boldsymbol{\eta}^\top \boldsymbol{\phi}(y) \right) dy \right) \right] \tag{58}$$

(64)

Derivate of $A(\boldsymbol{\eta})$ and moments: general cases

$$\nabla A(\boldsymbol{\eta}) = \nabla \left[\ln \left(\int h(y) \exp\left(\boldsymbol{\eta}^\top \boldsymbol{\phi}(y) \right) dy \right) \right] \tag{58}$$

$$= \nabla \left[\left(\int h(y) \exp\left(\boldsymbol{\eta}^\top \boldsymbol{\phi}(y) \right) dy \right) \right] \cdot \left(\int h(y) \exp\left(\boldsymbol{\eta}^\top \boldsymbol{\phi}(y) \right) dy \right)^{-1} \tag{59}$$

(64)

Derivate of $A(\boldsymbol{\eta})$ and moments: general cases

$$\nabla A(\boldsymbol{\eta}) = \nabla \left[\ln \left(\int h(y) \exp\left(\boldsymbol{\eta}^\top \boldsymbol{\phi}(y) \right) dy \right) \right] \tag{58}$$

$$= \nabla \left[\left(\int h(y) \exp\left(\boldsymbol{\eta}^\top \boldsymbol{\phi}(y) \right) dy \right) \right] \cdot \left(\int h(y) \exp\left(\boldsymbol{\eta}^\top \boldsymbol{\phi}(y) \right) dy \right)^{-1} \tag{59}$$

$$= \nabla \left[\left(\int h(y) \exp\left(\boldsymbol{\eta}^\top \boldsymbol{\phi}(y) \right) dy \right) \right] \cdot \exp\left(-A(\boldsymbol{\eta}) \right) \tag{60}$$

(64)

Derivate of $A(\boldsymbol{\eta})$ and moments: general cases

$$\nabla A(\boldsymbol{\eta}) = \nabla \left[\ln \left(\int h(y) \exp\left(\boldsymbol{\eta}^\top \boldsymbol{\phi}(y) \right) dy \right) \right] \tag{58}$$

$$= \nabla \left[\left(\int h(y) \exp\left(\boldsymbol{\eta}^\top \boldsymbol{\phi}(y) \right) dy \right) \right] \cdot \left(\int h(y) \exp\left(\boldsymbol{\eta}^\top \boldsymbol{\phi}(y) \right) dy \right)^{-1} \tag{59}$$

$$= \nabla \left[\left(\int h(y) \exp\left(\boldsymbol{\eta}^\top \boldsymbol{\phi}(y) \right) dy \right) \right] \cdot \exp\left(-A(\boldsymbol{\eta}) \right) \tag{60}$$

$$= \int \nabla \left[h(y) \exp\left(\boldsymbol{\eta}^\top \boldsymbol{\phi}(y) \right) \right] dy \cdot \exp\left(-A(\boldsymbol{\eta}) \right) \tag{61}$$

(64)

Derivate of $A(\boldsymbol{\eta})$ and moments: general cases

$$\nabla A(\boldsymbol{\eta}) = \nabla \left[\ln \left(\int h(y) \exp\left(\boldsymbol{\eta}^\top \boldsymbol{\phi}(y) \right) dy \right) \right] \tag{58}$$

$$= \nabla \left[\left(\int h(y) \exp\left(\boldsymbol{\eta}^\top \boldsymbol{\phi}(y) \right) dy \right) \right] \cdot \left(\int h(y) \exp\left(\boldsymbol{\eta}^\top \boldsymbol{\phi}(y) \right) dy \right)^{-1} \tag{59}$$

$$= \nabla \left[\left(\int h(y) \exp\left(\boldsymbol{\eta}^\top \boldsymbol{\phi}(y) \right) dy \right) \right] \cdot \exp\left(-A(\boldsymbol{\eta})\right) \tag{60}$$

$$= \int \nabla \left[h(y) \exp\left(\boldsymbol{\eta}^\top \boldsymbol{\phi}(y) \right) \right] dy \cdot \exp\left(-A(\boldsymbol{\eta})\right) \tag{61}$$

$$= \int h(y) \exp\left(\boldsymbol{\eta}^\top \boldsymbol{\phi}(y) \right) \boldsymbol{\phi}(y) dy \cdot \exp\left(-A(\boldsymbol{\eta})\right) \tag{62}$$

$$\tag{64}$$

Derivate of $A(\boldsymbol{\eta})$ and moments: general cases

$$
\begin{aligned}
\nabla A(\boldsymbol{\eta}) &= \nabla \left[\ln \left(\int h(y) \exp\left(\boldsymbol{\eta}^\top \boldsymbol{\phi}(y)\right) dy\right)\right] && (58)\\
&= \nabla \left[\left(\int h(y) \exp\left(\boldsymbol{\eta}^\top \boldsymbol{\phi}(y)\right) dy\right)\right] \cdot \left(\int h(y) \exp\left(\boldsymbol{\eta}^\top \boldsymbol{\phi}(y)\right) dy\right)^{-1} && (59)\\
&= \nabla \left[\left(\int h(y) \exp\left(\boldsymbol{\eta}^\top \boldsymbol{\phi}(y)\right) dy\right)\right] \cdot \exp\left(-A(\boldsymbol{\eta})\right) && (60)\\
&= \int \nabla \left[h(y) \exp\left(\boldsymbol{\eta}^\top \boldsymbol{\phi}(y)\right)\right] dy \cdot \exp\left(-A(\boldsymbol{\eta})\right) && (61)\\
&= \int h(y) \exp\left(\boldsymbol{\eta}^\top \boldsymbol{\phi}(y)\right) \boldsymbol{\phi}(y) dy \cdot \exp\left(-A(\boldsymbol{\eta})\right) && (62)\\
&= \int h(y) \exp\left(\boldsymbol{\eta}^\top \boldsymbol{\phi}(y) - A(\boldsymbol{\eta})\right) \boldsymbol{\phi}(y) dy && (63)\\
& && (64)
\end{aligned}
$$

Derivate of $A(\boldsymbol{\eta})$ and moments: general cases

$$
\begin{aligned}
\nabla A(\boldsymbol{\eta}) &= \nabla \left[\ln \left(\int h(y) \exp \left(\boldsymbol{\eta}^\top \boldsymbol{\phi}(y) \right) dy \right) \right] && (58) \\
&= \nabla \left[\left(\int h(y) \exp \left(\boldsymbol{\eta}^\top \boldsymbol{\phi}(y) \right) dy \right) \right] \cdot \left(\int h(y) \exp \left(\boldsymbol{\eta}^\top \boldsymbol{\phi}(y) \right) dy \right)^{-1} && (59) \\
&= \nabla \left[\left(\int h(y) \exp \left(\boldsymbol{\eta}^\top \boldsymbol{\phi}(y) \right) dy \right) \right] \cdot \exp\left(-A(\boldsymbol{\eta})\right) && (60) \\
&= \int \nabla \left[h(y) \exp \left(\boldsymbol{\eta}^\top \boldsymbol{\phi}(y) \right) \right] dy \cdot \exp\left(-A(\boldsymbol{\eta})\right) && (61) \\
&= \int h(y) \exp \left(\boldsymbol{\eta}^\top \boldsymbol{\phi}(y) \right) \boldsymbol{\phi}(y) dy \cdot \exp\left(-A(\boldsymbol{\eta})\right) && (62) \\
&= \int h(y) \exp \left(\boldsymbol{\eta}^\top \boldsymbol{\phi}(y) - A(\boldsymbol{\eta}) \right) \boldsymbol{\phi}(y) dy && (63) \\
&= \int p(y|\boldsymbol{\eta}) \boldsymbol{\phi}(y) dy && (64)
\end{aligned}
$$

Derivate of $A(\boldsymbol{\eta})$ and moments: general cases

$$
\begin{aligned}
\nabla A(\boldsymbol{\eta}) &= \nabla \left[\ln \left(\int h(y) \exp\left(\boldsymbol{\eta}^\top \boldsymbol{\phi}(y) \right) dy \right) \right] && (58)\\
&= \nabla \left[\left(\int h(y) \exp\left(\boldsymbol{\eta}^\top \boldsymbol{\phi}(y) \right) dy \right) \right] \cdot \left(\int h(y) \exp\left(\boldsymbol{\eta}^\top \boldsymbol{\phi}(y) \right) dy \right)^{-1} && (59)\\
&= \nabla \left[\left(\int h(y) \exp\left(\boldsymbol{\eta}^\top \boldsymbol{\phi}(y) \right) dy \right) \right] \cdot \exp\left(-A(\boldsymbol{\eta}) \right) && (60)\\
&= \int \nabla \left[h(y) \exp\left(\boldsymbol{\eta}^\top \boldsymbol{\phi}(y) \right) \right] dy \cdot \exp\left(-A(\boldsymbol{\eta}) \right) && (61)\\
&= \int h(y) \exp\left(\boldsymbol{\eta}^\top \boldsymbol{\phi}(y) \right) \boldsymbol{\phi}(y) dy \cdot \exp\left(-A(\boldsymbol{\eta}) \right) && (62)\\
&= \int h(y) \exp\left(\boldsymbol{\eta}^\top \boldsymbol{\phi}(y) - A(\boldsymbol{\eta}) \right) \boldsymbol{\phi}(y) dy && (63)\\
&= \int p(y|\boldsymbol{\eta}) \boldsymbol{\phi}(y) dy = \mathbb{E}\left[\phi(Y) \right] && (64)
\end{aligned}
$$

Table of Contents

Maximum Likelihood Estimation (MLE)

Assume a set of iid samples $\{y_n\}_{n=1}^N$, sampled from a member of the exponential family with given $h(y)$, sufficient statistics $\phi(y)$, but unknown parameter $\boldsymbol{\eta}$.

Maximum Likelihood Estimation (MLE)

Assume a set of iid samples $\{y_n\}_{n=1}^{N}$, sampled from a member of the exponential family with given $h(y)$, sufficient statistics $\phi(y)$, but unknown parameter $\boldsymbol{\eta}$.

Goal: Estimate the natural parameter $\boldsymbol{\eta}$.

Maximum Likelihood Estimation (MLE)

Assume a set of iid samples $\{y_n\}_{n=1}^N$, sampled from a member of the exponential family with given $h(y)$, sufficient statistics $\phi(y)$, but unknown parameter $\boldsymbol{\eta}$.

Goal: Estimate the natural parameter $\boldsymbol{\eta}$.

How: MLE for $p(y|\boldsymbol{\eta}) = h(y) \exp\left[\boldsymbol{\eta}^\top \boldsymbol{\phi}(y) - A(\boldsymbol{\eta})\right]$.

Maximum Likelihood Estimation (MLE)

Assume a set of iid samples $\{y_n\}_{n=1}^N$, sampled from a member of the exponential family with given $h(y)$, sufficient statistics $\phi(y)$, but unknown parameter $\boldsymbol{\eta}$.

Goal: Estimate the natural parameter $\boldsymbol{\eta}$.

How: MLE for $p(y|\boldsymbol{\eta}) = h(y) \exp\left[\boldsymbol{\eta}^\top \boldsymbol{\phi}(y) - A(\boldsymbol{\eta})\right]$.

$$\mathcal{L}(\boldsymbol{\eta}) = -\frac{1}{N} \ln\left(p(y|\boldsymbol{\eta})\right) \tag{65}$$

$$= \frac{1}{N} \sum_{n=1}^{N} \left[-\ln\left(h(y_n)\right) - \boldsymbol{\eta}^\top \boldsymbol{\phi}(y_n) + A(\boldsymbol{\eta})\right] . \tag{66}$$

Maximum Likelihood Estimation (MLE)

Assume a set of iid samples $\{y_n\}_{n=1}^N$, sampled from a member of the exponential family with given $h(y)$, sufficient statistics $\phi(y)$, but unknown parameter $\boldsymbol{\eta}$.

Goal: Estimate the natural parameter $\boldsymbol{\eta}$.

How: MLE for $p(y|\boldsymbol{\eta}) = h(y) \exp\left[\boldsymbol{\eta}^\top \boldsymbol{\phi}(y) - A(\boldsymbol{\eta})\right]$.

$$\mathcal{L}(\boldsymbol{\eta}) = -\frac{1}{N} \ln\left(p(y|\boldsymbol{\eta})\right) \tag{65}$$

$$= \frac{1}{N} \sum_{n=1}^{N} \left[-\ln\left(h(y_n)\right) - \boldsymbol{\eta}^\top \boldsymbol{\phi}(y_n) + A(\boldsymbol{\eta})\right] . \tag{66}$$

$\implies$ The cost function $\mathcal{L}$ is a convex function in $\boldsymbol{\eta}$ since $A(\boldsymbol{\eta})$ is convex.

Given the definition

$$\mathcal{L}(\boldsymbol{\eta}) = \frac{1}{N}\sum_{n=1}^{N}\left[-\ln\left(h(y_n)\right) - \boldsymbol{\eta}^\top \boldsymbol{\phi}(y_n) + A(\boldsymbol{\eta})\right] \tag{67}$$

[2] It says that we should pick $\boldsymbol{\eta}$ s.t. the expected value of the sufficient statistics is equal to its empirical value!

Given the definition

$$\mathcal{L}(\boldsymbol{\eta}) = \frac{1}{N}\sum_{n=1}^{N}\left[-\ln\left(h(y_n)\right) - \boldsymbol{\eta}^\top \boldsymbol{\phi}(y_n) + A(\boldsymbol{\eta})\right] \tag{67}$$

Gradient:

$$\nabla_{\boldsymbol{\eta}}\mathcal{L}(\boldsymbol{\eta}) = -\frac{1}{N}\sum_{n=1}^{N}\boldsymbol{\phi}(y_n) + \mathbb{E}\left[\boldsymbol{\phi}(y)\right] , \tag{68}$$

[2] It says that we should pick $\boldsymbol{\eta}$ s.t. the expected value of the sufficient statistics is equal to its empirical value!

Given the definition

$$\mathcal{L}(\boldsymbol{\eta}) = \frac{1}{N}\sum_{n=1}^{N}\left[-\ln\left(h(y_n)\right) - \boldsymbol{\eta}^\top \boldsymbol{\phi}(y_n) + A(\boldsymbol{\eta})\right] \tag{67}$$

Gradient:

$$\nabla_{\boldsymbol{\eta}}\mathcal{L}(\boldsymbol{\eta}) = -\frac{1}{N}\sum_{n=1}^{N}\boldsymbol{\phi}(y_n) + \mathbb{E}\left[\boldsymbol{\phi}(y)\right]\,, \tag{68}$$

Stationary point:[2]

[2] It says that we should pick $\boldsymbol{\eta}$ s.t. the expected value of the sufficient statistics is equal to its empirical value!

Given the definition

$$\mathcal{L}(\boldsymbol{\eta}) = \frac{1}{N}\sum_{n=1}^{N}\left[-\ln\left(h(y_n)\right) - \boldsymbol{\eta}^\top \boldsymbol{\phi}(y_n) + A(\boldsymbol{\eta})\right] \tag{67}$$

Gradient:

$$\nabla_{\boldsymbol{\eta}}\mathcal{L}(\boldsymbol{\eta}) = -\frac{1}{N}\sum_{n=1}^{N}\boldsymbol{\phi}(y_n) + \mathbb{E}\left[\boldsymbol{\phi}(y)\right], \tag{68}$$

Stationary point:[2]

$$\boldsymbol{\mu} := \mathbb{E}\left[\boldsymbol{\phi}(y)\right] = \frac{1}{N}\sum_{n=1}^{N}\boldsymbol{\phi}(y_n), \tag{69}$$

[2] It says that we should pick $\boldsymbol{\eta}$ s.t. the expected value of the sufficient statistics is equal to its empirical value!

Given the definition

$$\mathcal{L}(\boldsymbol{\eta}) = \frac{1}{N}\sum_{n=1}^{N}\left[-\ln\left(h(y_n)\right) - \boldsymbol{\eta}^\top \boldsymbol{\phi}(y_n) + A(\boldsymbol{\eta})\right] \tag{67}$$

Gradient:

$$\nabla_{\boldsymbol{\eta}}\mathcal{L}(\boldsymbol{\eta}) = -\frac{1}{N}\sum_{n=1}^{N}\boldsymbol{\phi}(y_n) + \mathbb{E}\left[\boldsymbol{\phi}(y)\right]\,, \tag{68}$$

Stationary point:[2]

$$\boldsymbol{\mu} := \mathbb{E}\left[\boldsymbol{\phi}(y)\right] = \frac{1}{N}\sum_{n=1}^{N}\boldsymbol{\phi}(y_n)\,, \tag{69}$$

Closed-form: assume we have determined the link function $\mathbf{g}(\boldsymbol{\mu}) = \boldsymbol{\eta}$

$$\boldsymbol{\eta} = \mathbf{g}\left(\frac{1}{N}\sum_{n=1}^{N}\boldsymbol{\phi}(y_n)\right)\,, \tag{70}$$

and justify why we called $\phi(y)$ a sufficient statistics.

[2] It says that we should pick $\boldsymbol{\eta}$ s.t. the expected value of the sufficient statistics is equal to its empirical value!

Generalized Linear Models (GLM)

- Both linear and logistic regressions focus on the conditional relationship between X and Y

Generalized Linear Models (GLM)

- Both linear and logistic regressions focus on the conditional relationship between X and Y
  - LS: $Y \sim \mathcal{N}(\mathbf{x}^\top \mathbf{w}, \sigma^2)$

Generalized Linear Models (GLM)

- Both linear and logistic regressions focus on the conditional relationship between X and Y
  - LS: $Y \sim \mathcal{N}(\mathbf{x}^\top \mathbf{w}, \sigma^2)$
  - Logistic regression: $Y \sim \mathcal{B}(\sigma(\mathbf{x}^\top \mathbf{w}))$

Generalized Linear Models (GLM)

- Both linear and logistic regressions focus on the conditional relationship between X and Y
  - LS: $Y \sim \mathcal{N}(\mathbf{x}^\top \mathbf{w}, \sigma^2)$
  - Logistic regression: $Y \sim \mathcal{B}(\sigma(\mathbf{x}^\top \mathbf{w}))$
- Common feature of linear and logistic regression:

Generalized Linear Models (GLM)

- Both linear and logistic regressions focus on the conditional relationship between X and Y
  - LS: $Y \sim \mathcal{N}(\mathbf{x}^\top \mathbf{w}, \sigma^2)$
  - Logistic regression: $Y \sim \mathcal{B}(\sigma(\mathbf{x}^\top \mathbf{w}))$
- Common feature of linear and logistic regression:
  1. Model the conditional expectation as $\mu = f(\mathbf{x}^\top \mathbf{w})$

Generalized Linear Models (GLM)

- Both linear and logistic regressions focus on the conditional relationship between X and Y
  - LS: $Y \sim \mathcal{N}(\mathbf{x}^\top \mathbf{w}, \sigma^2)$
  - Logistic regression: $Y \sim \mathcal{B}(\sigma(\mathbf{x}^\top \mathbf{w}))$
- Common feature of linear and logistic regression:
  1. Model the conditional expectation as $\mu = f(\mathbf{x}^\top \mathbf{w})$
  2. Endow Y with a particular probability distribution having μ as parameter

Generalized Linear Models (GLM)

- Both linear and logistic regressions focus on the conditional relationship between X and Y
  - LS: $Y \sim \mathcal{N}(\mathbf{x}^\top \mathbf{w}, \sigma^2)$
  - Logistic regression: $Y \sim \mathcal{B}(\sigma(\mathbf{x}^\top \mathbf{w}))$
- Common feature of linear and logistic regression:
  1. Model the conditional expectation as $\mu = f(\mathbf{x}^\top \mathbf{w})$
  2. Endow Y with a particular probability distribution having μ as parameter
- The GLM framework extends these ideas to the general exponential family.

Generalized Linear Models (GLM): cont'd

Scenario:

Generalized Linear Models (GLM): cont'd

Scenario:

- We would like to build a model to estimate the number y of customers arriving in a store.

Generalized Linear Models (GLM): cont'd

Scenario:

- We would like to build a model to estimate the number y of customers arriving in a store.
- The Poisson distribution usually gives a good model for the number of visitors.

Generalized Linear Models (GLM): cont'd

Scenario:

- We would like to build a model to estimate the number y of customers arriving in a store.
- The Poisson distribution usually gives a good model for the number of visitors.
- How can we come up with a model for our problem?

Generalized Linear Models (GLM): cont'd

Scenario:

- We would like to build a model to estimate the number y of customers arriving in a store.
- The Poisson distribution usually gives a good model for the number of visitors.
- How can we come up with a model for our problem?
- Fortunately the Poisson is an exponential family distribution.

Generalized Linear Models (GLM): cont'd

Scenario:

- We would like to build a model to estimate the number y of customers arriving in a store.
- The Poisson distribution usually gives a good model for the number of visitors.
- How can we come up with a model for our problem?
- Fortunately the Poisson is an exponential family distribution.

We can apply a Generalized Linear Model (GLM)!

Constructing GLMs

To derive a GLM for a classification/regression problem (the conditional dist. of y given $\mathbf{x}$):

Constructing GLMs

To derive a GLM for a classification/regression problem (the conditional dist. of y given $\mathbf{x}$):

1. The natural parameter η and the observed inputs $\mathbf{x}$ are related linearly: $\eta = \mathbf{x}^\top \mathbf{w}$

Constructing GLMs

To derive a GLM for a classification/regression problem (the conditional dist. of y given $\mathbf{x}$):

1. The natural parameter η and the observed inputs $\mathbf{x}$ are related linearly: $\eta = \mathbf{x}^\top \mathbf{w}$
2. The conditional mean μ is represented as a function $f(\eta)$ of the linear combination $\boldsymbol{\eta}$

Constructing GLMs

To derive a GLM for a classification/regression problem (the conditional dist. of y given $\mathbf{x}$):

1. The natural parameter η and the observed inputs $\mathbf{x}$ are related linearly: $\eta = \mathbf{x}^\top \mathbf{w}$
2. The conditional mean μ is represented as a function $f(\eta)$ of the linear combination $\boldsymbol{\eta}$
3. The observed output y is assumed to be characterized by an exponential family distribution with conditional mean μ.

Constructing GLMs

To derive a GLM for a classification/regression problem (the conditional dist. of y given $\mathbf{x}$):

1. The natural parameter η and the observed inputs $\mathbf{x}$ are related linearly: $\eta = \mathbf{x}^\top \mathbf{w}$
2. The conditional mean μ is represented as a function $f(\eta)$ of the linear combination $\boldsymbol{\eta}$
3. The observed output y is assumed to be characterized by an exponential family distribution with conditional mean μ.

In summary:

- $g(\mu) = \eta$ (The link function g maps the mean μ to the linear predictor η)
- $\mu = f(\eta)$ (The response function f maps the linear predictor η back to the mean μ).

Constructing GLMs

To derive a GLM for a classification/regression problem (the conditional dist. of y given $\mathbf{x}$):

1. The natural parameter η and the observed inputs $\mathbf{x}$ are related linearly: $\eta = \mathbf{x}^\top \mathbf{w}$
2. The conditional mean μ is represented as a function $f(\eta)$ of the linear combination $\boldsymbol{\eta}$
3. The observed output y is assumed to be characterized by an exponential family distribution with conditional mean μ.

In summary:

- $g(\mu) = \eta$ (The link function g maps the mean μ to the linear predictor η)
- $\mu = f(\eta)$ (The response function f maps the linear predictor η back to the mean μ).

The condition probability is thus modeled as:

$$p(y|\mathbf{x};\mathbf{w}) = h(y_n)\exp\left(\eta\phi(y) - A(\eta)\right) \qquad \text{for} \quad \eta = g \circ f(\mathbf{x}^\top \mathbf{w}) \tag{71}$$

- Two choice points in the specification of a GLM:
  1. The choice of the exponential family distribution
     $\Longrightarrow$ Generally constrained by the nature of the data Y
  2. The choice of the response function f
     $\Longrightarrow$ Real degree of freedom!
     $\Longrightarrow$ Canonical response function: $f = g^{-1}$, uniquely associated with the given exponential family distribution.

- If we decide to use the canonical response function, the choice of the exponential family density completely determines the GLM:

$$p(y|\mathbf{x};\mathbf{w}) = h(y_n)\exp\left(\eta\phi(y) - A(\eta)\right) \qquad \text{for} \quad \eta = \mathbf{x}^\top\mathbf{w} \tag{72}$$

Negative log-likelihood estimation

Note that:

$$\mathcal{L}(\mathbf{w}) = -\frac{1}{N}\sum_{n=1}^{N} \ln p(y_n|\mathbf{x}_n^\top \mathbf{w}) = -\frac{1}{N}\sum_{n=1}^{N} \left(\ln(h(y_n)) + \eta_n \phi(y_n) - A(\eta_n)\right) \quad (73)$$

Negative log-likelihood estimation

Note that:

$$\mathcal{L}(\mathbf{w}) = -\frac{1}{N}\sum_{n=1}^{N} \ln p(y_n|\mathbf{x}_n^\top \mathbf{w}) = -\frac{1}{N}\sum_{n=1}^{N} \left(\ln(h(y_n)) + \eta_n \phi(y_n) - A(\eta_n)\right) \tag{73}$$

If we rewrite this sum by using the matrix notation, we get (exercise:)

$$\nabla_{\mathbf{w}} \mathcal{L}(\mathbf{w}) = -\frac{1}{N}\sum_{n=1}^{N} \left(\mathbf{x}_n \phi(y_n) - \nabla_{\mathbf{w}} A(\eta_n)\right)$$

(76)

Negative log-likelihood estimation

Note that:

$$\mathcal{L}(\mathbf{w}) = -\frac{1}{N}\sum_{n=1}^{N} \ln p(y_n|\mathbf{x}_n^\top \mathbf{w}) = -\frac{1}{N}\sum_{n=1}^{N} (\ln(h(y_n)) + \eta_n \phi(y_n) - A(\eta_n)) \tag{73}$$

If we rewrite this sum by using the matrix notation, we get (exercise:)

$$\nabla_{\mathbf{w}}\mathcal{L}(\mathbf{w}) = -\frac{1}{N}\sum_{n=1}^{N} (\mathbf{x}_n \phi(y_n) - \nabla_{\mathbf{w}} A(\eta_n)) = -\frac{1}{N}\sum_{n=1}^{N} \left(\mathbf{x}_n \phi(y_n) - A'(\mathbf{x}_n^\top \mathbf{w})\mathbf{x}_n\right) \tag{74}$$

(76)

Negative log-likelihood estimation

Note that:

$$\mathcal{L}(\mathbf{w}) = -\frac{1}{N}\sum_{n=1}^{N} \ln p(y_n|\mathbf{x}_n^\top \mathbf{w}) = -\frac{1}{N}\sum_{n=1}^{N} (\ln(h(y_n)) + \eta_n \phi(y_n) - A(\eta_n)) \tag{73}$$

If we rewrite this sum by using the matrix notation, we get (exercise:)

$$\nabla_{\mathbf{w}}\mathcal{L}(\mathbf{w}) = -\frac{1}{N}\sum_{n=1}^{N} (\mathbf{x}_n\phi(y_n) - \nabla_{\mathbf{w}} A(\eta_n)) = -\frac{1}{N}\sum_{n=1}^{N} \left(\mathbf{x}_n\phi(y_n) - A'(\mathbf{x}_n^\top \mathbf{w})\mathbf{x}_n\right) \tag{74}$$

$$= -\frac{1}{N}\sum_{n=1}^{N} (\mathbf{x}_n\phi(y_n) - \mathbb{E}\left[\phi(Y_n)\right]\mathbf{x}_n)$$

(76)

Negative log-likelihood estimation

Note that:

$$\mathcal{L}(\mathbf{w}) = -\frac{1}{N}\sum_{n=1}^{N} \ln p(y_n|\mathbf{x}_n^\top \mathbf{w}) = -\frac{1}{N}\sum_{n=1}^{N} \left(\ln(h(y_n)) + \eta_n \phi(y_n) - A(\eta_n)\right) \tag{73}$$

If we rewrite this sum by using the matrix notation, we get (exercise:)

$$\nabla_{\mathbf{w}}\mathcal{L}(\mathbf{w}) = -\frac{1}{N}\sum_{n=1}^{N} (\mathbf{x}_n \phi(y_n) - \nabla_{\mathbf{w}} A(\eta_n)) = -\frac{1}{N}\sum_{n=1}^{N} \left(\mathbf{x}_n \phi(y_n) - A'(\mathbf{x}_n^\top \mathbf{w})\mathbf{x}_n\right) \tag{74}$$

$$= -\frac{1}{N}\sum_{n=1}^{N} (\mathbf{x}_n \phi(y_n) - \mathbb{E}\left[\phi(Y_n)\right] \mathbf{x}_n) = -\frac{1}{N}\sum_{n=1}^{N} \left(\mathbf{x}_n \phi(y_n) - g^{-1}(\mathbf{x}_n^\top \mathbf{w})\mathbf{x}_n\right) \tag{75}$$

$$\tag{76}$$

Negative log-likelihood estimation

Note that:

$$\mathcal{L}(\mathbf{w}) = -\frac{1}{N}\sum_{n=1}^{N} \ln p(y_n|\mathbf{x}_n^\top \mathbf{w}) = -\frac{1}{N}\sum_{n=1}^{N} (\ln(h(y_n)) + \eta_n \phi(y_n) - A(\eta_n)) \tag{73}$$

If we rewrite this sum by using the matrix notation, we get (exercise:)

$$\nabla_{\mathbf{w}}\mathcal{L}(\mathbf{w}) = -\frac{1}{N}\sum_{n=1}^{N} (\mathbf{x}_n\phi(y_n) - \nabla_{\mathbf{w}}A(\eta_n)) = -\frac{1}{N}\sum_{n=1}^{N} \left(\mathbf{x}_n\phi(y_n) - A'(\mathbf{x}_n^\top \mathbf{w})\mathbf{x}_n\right) \tag{74}$$

$$= -\frac{1}{N}\sum_{n=1}^{N} (\mathbf{x}_n\phi(y_n) - \mathbb{E}\left[\phi(Y_n)\right]\mathbf{x}_n) = -\frac{1}{N}\sum_{n=1}^{N} \left(\mathbf{x}_n\phi(y_n) - g^{-1}(\mathbf{x}_n^\top \mathbf{w})\mathbf{x}_n\right) \tag{75}$$

$$= -\frac{1}{N}\mathbf{X}^\top \left[g^{-1}(\mathbf{X}\mathbf{w}) - \phi(\mathbf{y})\right] \tag{76}$$

Negative log-likelihood estimation

Note that:

$$\mathcal{L}(\mathbf{w}) = -\frac{1}{N}\sum_{n=1}^{N} \ln p(y_n|\mathbf{x}_n^\top \mathbf{w}) = -\frac{1}{N}\sum_{n=1}^{N} \left(\ln(h(y_n)) + \eta_n \phi(y_n) - A(\eta_n)\right) \tag{73}$$

If we rewrite this sum by using the matrix notation, we get (exercise:)

$$\nabla_{\mathbf{w}}\mathcal{L}(\mathbf{w}) = -\frac{1}{N}\sum_{n=1}^{N} \left(\mathbf{x}_n \phi(y_n) - \nabla_{\mathbf{w}} A(\eta_n)\right) = -\frac{1}{N}\sum_{n=1}^{N} \left(\mathbf{x}_n \phi(y_n) - A'(\mathbf{x}_n^\top \mathbf{w})\mathbf{x}_n\right) \tag{74}$$

$$= -\frac{1}{N}\sum_{n=1}^{N} \left(\mathbf{x}_n \phi(y_n) - \mathbb{E}\left[\phi(Y_n)\right]\mathbf{x}_n\right) = -\frac{1}{N}\sum_{n=1}^{N} \left(\mathbf{x}_n \phi(y_n) - g^{-1}(\mathbf{x}_n^\top \mathbf{w})\mathbf{x}_n\right) \tag{75}$$

$$= -\frac{1}{N}\mathbf{X}^\top \left[g^{-1}(\mathbf{X}\mathbf{w}) - \phi(\mathbf{y})\right] \tag{76}$$

In the case of Logistic Regression:

$$\nabla_{\mathbf{w}}\mathcal{L}(\mathbf{w}) = \frac{1}{N}\mathbf{X}^\top \left[\sigma(\mathbf{X}\mathbf{w}) - \mathbf{y}\right] \tag{77}$$

Some examples

- Gaussian distribution Least Squares

Some examples

- Gaussian distribution — Least Squares
- Bernoulli distribution — Logistic Regression

Some examples

• Gaussian distribution	Least Squares
• Bernoulli distribution	Logistic Regression
• Multi-nomial distribution	Softmax Regression

Last lecture:

- Logistic Regression

This lecture:

- Exponential Families
- Generalized Linear Models