# Majorization–Minimization (MM) Algorithms for Statistical Inference and Machine Learning Problems

**Hien D. Nguyen**[1]

[1]Department of Mathematics and Statistics, La Trobe University, Melbourne Australia

*Research Summer School in Statistics and Big Data Science (SBDS)*, University of Caen, 2017-06-08

LA TROBE UNIVERSITY

# Preface

This lecture is based on my recent review/tutorial paper:

- **H.D. Nguyen** (2017), An introduction to MM algorithms for machine learning and statistical estimation, *WIREs Data Mining and Knowledge Discovery*, vol. 7, e1198.

# Introduction

- Let $\mathbf{D}_n = \{\boldsymbol{D}_i\}_{i=1}^n$ be a sample of $n \in \mathbb{N}$ random observations from an unknown data generating process (DGP).
  - $i \in [n] = \{1, \ldots, n\}$.
  - Let $\mathbf{d}_n = \{\boldsymbol{d}_i\}_{i=1}^n$ be a realization of $\mathbf{D}_n$ (where $\boldsymbol{d}_i$ is a realization of $\boldsymbol{D}_i$).
- When conducting statistical inference or machine learning, we wish to obtain knowledge regarding some property/properties of the DGP of $\mathbf{D}_n$.
- Three typical problems arise out of this goal:
  - Density estimation.
  - Regression.
  - Classification.

# Introduction

- Let $\mathbf{D}_n = \{\boldsymbol{D}_i\}_{i=1}^n$ be a sample of $n \in \mathbb{N}$ random observations from an unknown data generating process (DGP).
  - $i \in [n] = \{1, \ldots, n\}$.
  - Let $\mathbf{d}_n = \{\boldsymbol{d}_i\}_{i=1}^n$ be a realization of $\mathbf{D}_n$ (where $\boldsymbol{d}_i$ is a realization of $\boldsymbol{D}_i$).
- When conducting statistical inference or machine learning, we wish to obtain knowledge regarding some property/properties of the DGP of $\mathbf{D}_n$.
- Three typical problems arise out of this goal:
  - Density estimation.
  - Regression.
  - Classification.

# Introduction

- Let $\mathbf{D}_n = \{\boldsymbol{D}_i\}_{i=1}^n$ be a sample of $n \in \mathbb{N}$ random observations from an unknown data generating process (DGP).
  - $i \in [n] = \{1, \ldots, n\}$.
  - Let $\mathbf{d}_n = \{\boldsymbol{d}_i\}_{i=1}^n$ be a realization of $\mathbf{D}_n$ (where $\boldsymbol{d}_i$ is a realization of $\boldsymbol{D}_i$).
- When conducting statistical inference or machine learning, we wish to obtain knowledge regarding some property/properties of the DGP of $\mathbf{D}_n$.
- Three typical problems arise out of this goal:
  - Density estimation.
  - Regression.
  - Classification.

# Density Estimation

- Let $\boldsymbol{D}_i = \boldsymbol{X}_i \in \mathbb{X} \subset \mathbb{R}^p$ $(p \in \mathbb{N})$.
- Suppose we know that $\boldsymbol{X}_i$ are each IID (independent and identically distributed) according to some distribution from a family that is characterized by a probability density in a function class $\mathscr{F}_{\boldsymbol{\theta}}$, that is parameterized by a vector $\boldsymbol{\theta} \in \mathbb{R}^d$ $(d \in \mathbb{N})$.
- In particular, let each $\boldsymbol{X}_i$ be generated from a distribution that is characterized by some density parameterized by an unknown vector $\boldsymbol{\theta}_0$ within $\mathscr{F}_{\boldsymbol{\theta}}$.
  - The task of (parametric) density estimation is to utilize $\mathbf{d}_n$ in order to estimate $\boldsymbol{\theta}_0$.

# Density Estimation

- Let $\boldsymbol{D}_i = \boldsymbol{X}_i \in \mathbb{X} \subset \mathbb{R}^p$ $(p \in \mathbb{N})$.
- Suppose we know that $\boldsymbol{X}_i$ are each IID (independent and identically distributed) according to some distribution from a family that is characterized by a probability density in a function class $\mathscr{F}_{\boldsymbol{\theta}}$, that is parameterized by a vector $\boldsymbol{\theta} \in \mathbb{R}^d$ $(d \in \mathbb{N})$.
- In particular, let each $\boldsymbol{X}_i$ be generated from a distribution that is characterized by some density parameterized by an unknown vector $\boldsymbol{\theta}_0$ within $\mathscr{F}_{\boldsymbol{\theta}}$.
  - The task of (parametric) density estimation is to utilize $\mathbf{d}_n$ in order to estimate $\boldsymbol{\theta}_0$.

# Density Estimation

- Let $\boldsymbol{D}_i = \boldsymbol{X}_i \in \mathbb{X} \subset \mathbb{R}^p$ $(p \in \mathbb{N})$.
- Suppose we know that $\boldsymbol{X}_i$ are each IID (independent and identically distributed) according to some distribution from a family that is characterized by a probability density in a function class $\mathscr{F}_{\boldsymbol{\theta}}$, that is parameterized by a vector $\boldsymbol{\theta} \in \mathbb{R}^d$ $(d \in \mathbb{N})$.
- In particular, let each $\boldsymbol{X}_i$ be generated from a distribution that is characterized by some density parameterized by an unknown vector $\boldsymbol{\theta}_0$ within $\mathscr{F}_{\boldsymbol{\theta}}$.
  - The task of (parametric) density estimation is to utilize $\mathbf{d}_n$ in order to estimate $\boldsymbol{\theta}_0$.

# Regression

- Let $\boldsymbol{D}_i^\top = \left(\boldsymbol{X}_i^\top, \boldsymbol{Y}_i^\top\right) \in \mathbb{X} \times \mathbb{Y}$, where $\mathbb{X} \subset \mathbb{R}^p$ and $\mathbb{Y} \subset \mathbb{R}^q$ $(p, q \in \mathbb{N})$.
  - $[\cdot]^\top$ is the matrix transposition operator.

- Suppose that $\mathbf{D}_n$ is an IID sample, and that it is known that the conditional probability of $\boldsymbol{Y}_i | \boldsymbol{X}_i = \boldsymbol{x}_i$ can be modeled by a density function $f_{\boldsymbol{\theta}}\left(\boldsymbol{y} | \boldsymbol{X} = \boldsymbol{x}\right)$ that is in some class of conditional density functions $\mathscr{F}_{\boldsymbol{\theta}}$ that is parameterized by a vector $\boldsymbol{\theta} \in \mathbb{R}^d$ $(d \in \mathbb{N})$.
  - We say that $\boldsymbol{Y}_i$ is the output (response) and $\boldsymbol{X}_i$ is the input (covariate), respectively.

- In particular, let each $\boldsymbol{D}_i$ be generated from a DGP that can be best characterized by a conditional density function by a density parameterized by some unknown vector $\boldsymbol{\theta}_0$ within $\mathscr{F}_{\boldsymbol{\theta}}$.
  - The task of (parametric) density estimation is to utilize $\mathbf{d}_n$ in order to estimate $\boldsymbol{\theta}_0$.

# Regression

- Let $\boldsymbol{D}_i^\top = \left(\boldsymbol{X}_i^\top, \boldsymbol{Y}_i^\top\right) \in \mathbb{X} \times \mathbb{Y}$, where $\mathbb{X} \subset \mathbb{R}^p$ and $\mathbb{Y} \subset \mathbb{R}^q$ $(p, q \in \mathbb{N})$.
  - $[\cdot]^\top$ is the matrix transposition operator.
- Suppose that $\mathbf{D}_n$ is an IID sample, and that it is known that the conditional probability of $\boldsymbol{Y}_i | \boldsymbol{X}_i = \boldsymbol{x}_i$ can be modeled by a density function $f_{\boldsymbol{\theta}}\left(\boldsymbol{y} | \boldsymbol{X} = \boldsymbol{x}\right)$ that is in some class of conditional density functions $\mathscr{F}_{\boldsymbol{\theta}}$ that is parameterized by a vector $\boldsymbol{\theta} \in \mathbb{R}^d$ $(d \in \mathbb{N})$.
  - We say that $\boldsymbol{Y}_i$ is the output (response) and $\boldsymbol{X}_i$ is the input (covariate), respectively.
- In particular, let each $\boldsymbol{D}_i$ be generated from a DGP that can be best characterized by a conditional density function by a density parameterized by some unknown vector $\boldsymbol{\theta}_0$ within $\mathscr{F}_{\boldsymbol{\theta}}$.
  - The task of (parametric) density estimation is to utilize $\mathbf{d}_n$ in order to estimate $\boldsymbol{\theta}_0$.

# Regression

- Let $\boldsymbol{D}_i^\top = \left(\boldsymbol{X}_i^\top, \boldsymbol{Y}_i^\top\right) \in \mathbb{X} \times \mathbb{Y}$, where $\mathbb{X} \subset \mathbb{R}^p$ and $\mathbb{Y} \subset \mathbb{R}^q$ $(p, q \in \mathbb{N})$.
  - $[\cdot]^\top$ is the matrix transposition operator.
- Suppose that $\mathbf{D}_n$ is an IID sample, and that it is known that the conditional probability of $\boldsymbol{Y}_i | \boldsymbol{X}_i = \boldsymbol{x}_i$ can be modeled by a density function $f_{\boldsymbol{\theta}}\left(\boldsymbol{y} | \boldsymbol{X} = \boldsymbol{x}\right)$ that is in some class of conditional density functions $\mathscr{F}_{\boldsymbol{\theta}}$ that is parameterized by a vector $\boldsymbol{\theta} \in \mathbb{R}^d$ $(d \in \mathbb{N})$.
  - We say that $\boldsymbol{Y}_i$ is the output (response) and $\boldsymbol{X}_i$ is the input (covariate), respectively.
- In particular, let each $\boldsymbol{D}_i$ be generated from a DGP that can be best characterized by a conditional density function by a density parameterized by some unknown vector $\boldsymbol{\theta}_0$ within $\mathscr{F}_{\boldsymbol{\theta}}$.
  - The task of (parametric) density estimation is to utilize $\mathbf{d}_n$ in order to estimate $\boldsymbol{\theta}_0$.

# Classification

- Let $\boldsymbol{D}_i^\top = \left(\boldsymbol{X}_i^\top, Y_i\right) \in \mathbb{X} \times \{-1, +1\}$, where $\mathbb{X} \subset \mathbb{R}^p$ $(p \in \mathbb{N})$.
- Suppose that $\mathbf{D}_n$ is an IID sample, and we know that the DGP of $\boldsymbol{X}_i | Y_i = -1$ is different to that of $\boldsymbol{X}_i | Y_i = +1$, for each $i \in [n]$.
- Let $f_{\boldsymbol{\theta}}(\boldsymbol{x})$ be a function in a class $\mathscr{F}_{\boldsymbol{\theta}}$ that is parameterized by the vector $\boldsymbol{\theta} \in \mathbb{R}^d$ $(d \in \mathbb{N})$, where the sign of $Y_i$ and $f_{\boldsymbol{\theta}}(\boldsymbol{X}_i)$ are strongly correlated some well chosen $\boldsymbol{\theta}$.
- Let $l(Y_i, f_{\boldsymbol{\theta}}(\boldsymbol{X}_i))$ be some loss function such that there exists an optimal $\boldsymbol{\theta}_0$, which minimizes the expected loss

$$\mathbb{E}\left[l(Y_i, f_{\boldsymbol{\theta}}(\boldsymbol{X}_i))\right],$$

under the DGP of $\boldsymbol{D}_i$.

  - The task of classification is to utilize $\mathbf{d}_n$ to estimate the value of $\boldsymbol{\theta}_0$.

# Classification

- Let $\boldsymbol{D}_i^\top = \left(\boldsymbol{X}_i^\top, Y_i\right) \in \mathbb{X} \times \{-1, +1\}$, where $\mathbb{X} \subset \mathbb{R}^p$ $(p \in \mathbb{N})$.
- Suppose that $\mathbf{D}_n$ is an IID sample, and we know that the DGP of $\boldsymbol{X}_i | Y_i = -1$ is different to that of $\boldsymbol{X}_i | Y_i = +1$, for each $i \in [n]$.
- Let $f_{\boldsymbol{\theta}}(\boldsymbol{x})$ be a function in a class $\mathscr{F}_{\boldsymbol{\theta}}$ that is parameterized by the vector $\boldsymbol{\theta} \in \mathbb{R}^d$ $(d \in \mathbb{N})$, where the sign of $Y_i$ and $f_{\boldsymbol{\theta}}(\boldsymbol{X}_i)$ are strongly correlated some well chosen $\boldsymbol{\theta}$.
- Let $l(Y_i, f_{\boldsymbol{\theta}}(\boldsymbol{X}_i))$ be some loss function such that there exists an optimal $\boldsymbol{\theta}_0$, which minimizes the expected loss

$$\mathbb{E}\left[l(Y_i, f_{\boldsymbol{\theta}}(\boldsymbol{X}_i))\right],$$

under the DGP of $\boldsymbol{D}_i$.

  - The task of classification is to utilize $\mathbf{d}_n$ to estimate the value of $\boldsymbol{\theta}_0$.

# Classification

- Let $\boldsymbol{D}_i^\top = \left(\boldsymbol{X}_i^\top, Y_i\right) \in \mathbb{X} \times \{-1, +1\}$, where $\mathbb{X} \subset \mathbb{R}^p$ $(p \in \mathbb{N})$.
- Suppose that $\mathbf{D}_n$ is an IID sample, and we know that the DGP of $\boldsymbol{X}_i | Y_i = -1$ is different to that of $\boldsymbol{X}_i | Y_i = +1$, for each $i \in [n]$.
- Let $f_{\boldsymbol{\theta}}(\boldsymbol{x})$ be a function in a class $\mathscr{F}_{\boldsymbol{\theta}}$ that is parameterized by the vector $\boldsymbol{\theta} \in \mathbb{R}^d$ $(d \in \mathbb{N})$, where the sign of $Y_i$ and $f_{\boldsymbol{\theta}}(\boldsymbol{X}_i)$ are strongly correlated some well chosen $\boldsymbol{\theta}$.
- Let $l(Y_i, f_{\boldsymbol{\theta}}(\boldsymbol{X}_i))$ be some loss function such that there exists an optimal $\boldsymbol{\theta}_0$, which minimizes the expected loss

$$\mathbb{E}\left[l(Y_i, f_{\boldsymbol{\theta}}(\boldsymbol{X}_i))\right],$$

under the DGP of $\boldsymbol{D}_i$.

  - The task of classification is to utilize $\mathbf{d}_n$ to estimate the value of $\boldsymbol{\theta}_0$.

# Classification

- Let $\boldsymbol{D}_i^\top = \left(\boldsymbol{X}_i^\top, Y_i\right) \in \mathbb{X} \times \{-1, +1\}$, where $\mathbb{X} \subset \mathbb{R}^p$ $(p \in \mathbb{N})$.
- Suppose that $\mathbf{D}_n$ is an IID sample, and we know that the DGP of $\boldsymbol{X}_i | Y_i = -1$ is different to that of $\boldsymbol{X}_i | Y_i = +1$, for each $i \in [n]$.
- Let $f_{\boldsymbol{\theta}}(\boldsymbol{x})$ be a function in a class $\mathscr{F}_{\boldsymbol{\theta}}$ that is parameterized by the vector $\boldsymbol{\theta} \in \mathbb{R}^d$ $(d \in \mathbb{N})$, where the sign of $Y_i$ and $f_{\boldsymbol{\theta}}(\boldsymbol{X}_i)$ are strongly correlated some well chosen $\boldsymbol{\theta}$.
- Let $l(Y_i, f_{\boldsymbol{\theta}}(\boldsymbol{X}_i))$ be some loss function such that there exists an optimal $\boldsymbol{\theta}_0$, which minimizes the expected loss

$$\mathbb{E}\left[l(Y_i, f_{\boldsymbol{\theta}}(\boldsymbol{X}_i))\right],$$

under the DGP of $\boldsymbol{D}_i$.

  - The task of classification is to utilize $\mathbf{d}_n$ to estimate the value of $\boldsymbol{\theta}_0$.

# Learning by Optimization

- Given a realization $\mathbf{d}_n$ of the data $\mathbf{D}_n$, many problems in statistical inference and machine learning can be phrased as

$$\arg\min_{\boldsymbol{\theta}\in\Theta} \frac{1}{n}\sum_{i=1}^{n} l(\boldsymbol{\theta}; \boldsymbol{d}_i) + \lambda R(\boldsymbol{\theta}), \tag{1}$$

  where $l(\cdot; \boldsymbol{\theta})$ is some loss function that is characterized by a vector $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^d$ $(d \in \mathbb{N})$.
  - $\lambda \geq 0$ and $R(\boldsymbol{\theta})$ is some regularization of $\boldsymbol{\theta}$.

- In machine learning, Problem (1) is often referred to as ERM (empirical risk minimization; see e.g. Vapnik, 1998).

- In statistics, Problem (1) goes by many names:
  - M-estimation (see e.g. Serfling, 1980).
  - Extremum estimation (Amemiya, 1985).
  - Minimum contrast estimation (Bickel and Doksum, 2001).

# Learning by Optimization

- Given a realization $\mathbf{d}_n$ of the data $\mathbf{D}_n$, many problems in statistical inference and machine learning can be phrased as

$$\arg\min_{\boldsymbol{\theta}\in\Theta} \frac{1}{n}\sum_{i=1}^{n} l(\boldsymbol{\theta}; \boldsymbol{d}_i) + \lambda R(\boldsymbol{\theta}), \tag{1}$$

  where $l(\cdot; \boldsymbol{\theta})$ is some loss function that is characterized by a vector $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^d$ $(d \in \mathbb{N})$.
  - $\lambda \geq 0$ and $R(\boldsymbol{\theta})$ is some regularization of $\boldsymbol{\theta}$.

- In machine learning, Problem (1) is often referred to as ERM (empirical risk minimization; see e.g. Vapnik, 1998).

- In statistics, Problem (1) goes by many names:
  - M-estimation (see e.g. Serfling, 1980).
  - Extremum estimation (Amemiya, 1985).
  - Minimum contrast estimation (Bickel and Doksum, 2001).

# Learning by Optimization

- Given a realization $\mathbf{d}_n$ of the data $\mathbf{D}_n$, many problems in statistical inference and machine learning can be phrased as

$$\arg\min_{\boldsymbol{\theta}\in\Theta} \frac{1}{n}\sum_{i=1}^{n} l(\boldsymbol{\theta}; \boldsymbol{d}_i) + \lambda R(\boldsymbol{\theta}), \tag{1}$$

  where $l(\cdot; \boldsymbol{\theta})$ is some loss function that is characterized by a vector $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^d$ $(d \in \mathbb{N})$.
  - $\lambda \geq 0$ and $R(\boldsymbol{\theta})$ is some regularization of $\boldsymbol{\theta}$.

- In machine learning, Problem (1) is often referred to as ERM (empirical risk minimization; see e.g. Vapnik, 1998).

- In statistics, Problem (1) goes by many names:
  - M-estimation (see e.g. Serfling, 1980).
  - Extremum estimation (Amemiya, 1985).
  - Minimum contrast estimation (Bickel and Doksum, 2001).

# Example 1: Maximum Likelihood Estimation for Gaussian Mixture Models

- Let $\mathbf{d}_n$ be a realization of the IID random sample $\mathbf{D}_n$, where $\boldsymbol{D}_i = \boldsymbol{X}_i \in \mathbb{R}^p$.
- For some known $g$, suppose that $\boldsymbol{X}_i$ arises from a distribution with densities from the family

$$\mathscr{F}_{\boldsymbol{\theta}} = \left\{ f_{\boldsymbol{\theta}} : f_{\boldsymbol{\theta}}(\boldsymbol{x}) = \sum_{z=1}^{g} \pi_z \phi_p(\boldsymbol{x}; \boldsymbol{\mu}_z, \boldsymbol{\Sigma}_z), \boldsymbol{\theta} \in \Theta \right\},$$

  that is parameterized by the vector $\boldsymbol{\theta}_0$.
  - $\phi_p(\cdot; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ is the $d$-dimensional Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$.
- $\Theta$ is the set of possible parameters that that meet the constraints:
  - $\sum_{z=1}^{g} \pi_z = 1$ and $\pi_z > 0$ for each $z \in [g]$.
  - $\boldsymbol{\mu}_z \in \mathbb{R}^p$ and $\boldsymbol{\Sigma}_z \in \mathbb{R}^{p \times p}$ is positive definite and has bounded Eigen values.

# Example 1: Maximum Likelihood Estimation for Gaussian Mixture Models

- Let $\mathbf{d}_n$ be a realization of the IID random sample $\mathbf{D}_n$, where $\boldsymbol{D}_i = \boldsymbol{X}_i \in \mathbb{R}^p$.
- For some known $g$, suppose that $\boldsymbol{X}_i$ arises from a distribution with densities from the family

$$\mathscr{F}_{\boldsymbol{\theta}} = \left\{ f_{\boldsymbol{\theta}} : f_{\boldsymbol{\theta}}(\boldsymbol{x}) = \sum_{z=1}^{g} \pi_z \phi_p(\boldsymbol{x}; \boldsymbol{\mu}_z, \boldsymbol{\Sigma}_z), \boldsymbol{\theta} \in \Theta \right\},$$

  that is parameterized by the vector $\boldsymbol{\theta}_0$.
  - $\phi_p(\cdot; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ is the $d$-dimensional Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$.
- $\Theta$ is the set of possible parameters that that meet the constraints:
  - $\sum_{z=1}^{g} \pi_z = 1$ and $\pi_z > 0$ for each $z \in [g]$.
  - $\boldsymbol{\mu}_z \in \mathbb{R}^p$ and $\boldsymbol{\Sigma}_z \in \mathbb{R}^{p \times p}$ is positive definite and has bounded Eigen values.

# Example 1: Maximum Likelihood Estimation for Gaussian Mixture Models

- Let $\mathbf{d}_n$ be a realization of the IID random sample $\mathbf{D}_n$, where $\boldsymbol{D}_i = \boldsymbol{X}_i \in \mathbb{R}^p$.
- For some known $g$, suppose that $\boldsymbol{X}_i$ arises from a distribution with densities from the family

$$\mathscr{F}_{\boldsymbol{\theta}} = \left\{ f_{\boldsymbol{\theta}} : f_{\boldsymbol{\theta}}(\boldsymbol{x}) = \sum_{z=1}^{g} \pi_z \phi_p(\boldsymbol{x}; \boldsymbol{\mu}_z, \boldsymbol{\Sigma}_z), \boldsymbol{\theta} \in \Theta \right\},$$

  that is parameterized by the vector $\boldsymbol{\theta}_0$.
  - $\phi_p(\cdot; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ is the $d$-dimensional Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$.
- $\Theta$ is the set of possible parameters that that meet the constraints:
  - $\sum_{z=1}^{g} \pi_z = 1$ and $\pi_z > 0$ for each $z \in [g]$.
  - $\boldsymbol{\mu}_z \in \mathbb{R}^p$ and $\boldsymbol{\Sigma}_z \in \mathbb{R}^{p \times p}$ is positive definite and has bounded Eigen values.

# Example 1: Maximum Likelihood Estimation for Gaussian Mixture Models

- We say that $\mathscr{F}_{\boldsymbol{\theta}}$ is the family $g$-component Gaussian mixture models (GMMs).
- In order to estimate $\boldsymbol{\theta}_0$ from $\mathbf{d}_n$, we can minimize the negative log-likelihood function to obtain the maximum likelihood estimate

$$\hat{\boldsymbol{\theta}}_n = \arg\min_{\boldsymbol{\theta}\in\Theta} \frac{1}{n}\sum_{i=1}^{n} l(\boldsymbol{\theta}; \boldsymbol{x}_i)$$

where

$$\begin{aligned} l(\boldsymbol{\theta}; \boldsymbol{x}_i) &= -\log f_{\boldsymbol{\theta}}(\boldsymbol{x}_i) \\ &= -\log \sum_{z=1}^{g} \pi_z \phi_p(\boldsymbol{x}_i; \boldsymbol{\mu}_z, \boldsymbol{\Sigma}_z). \end{aligned}$$

# Example 1: Maximum Likelihood Estimation for Gaussian Mixture Models

- We say that $\mathscr{F}_{\boldsymbol{\theta}}$ is the family $g$-component Gaussian mixture models (GMMs).
- In order to estimate $\boldsymbol{\theta}_0$ from $\mathbf{d}_n$, we can minimize the negative log-likelihood function to obtain the maximum likelihood estimate

$$\hat{\boldsymbol{\theta}}_n = \arg\min_{\boldsymbol{\theta}\in\Theta} \frac{1}{n}\sum_{i=1}^{n} l(\boldsymbol{\theta}; \boldsymbol{x}_i)$$

where

$$\begin{aligned} l(\boldsymbol{\theta}; \boldsymbol{x}_i) &= -\log f_{\boldsymbol{\theta}}(\boldsymbol{x}_i) \\ &= -\log \sum_{z=1}^{g} \pi_z \phi_p(\boldsymbol{x}_i; \boldsymbol{\mu}_z, \boldsymbol{\Sigma}_z). \end{aligned}$$

# Example 1: Maximum Likelihood Estimation for Gaussian Mixture Models

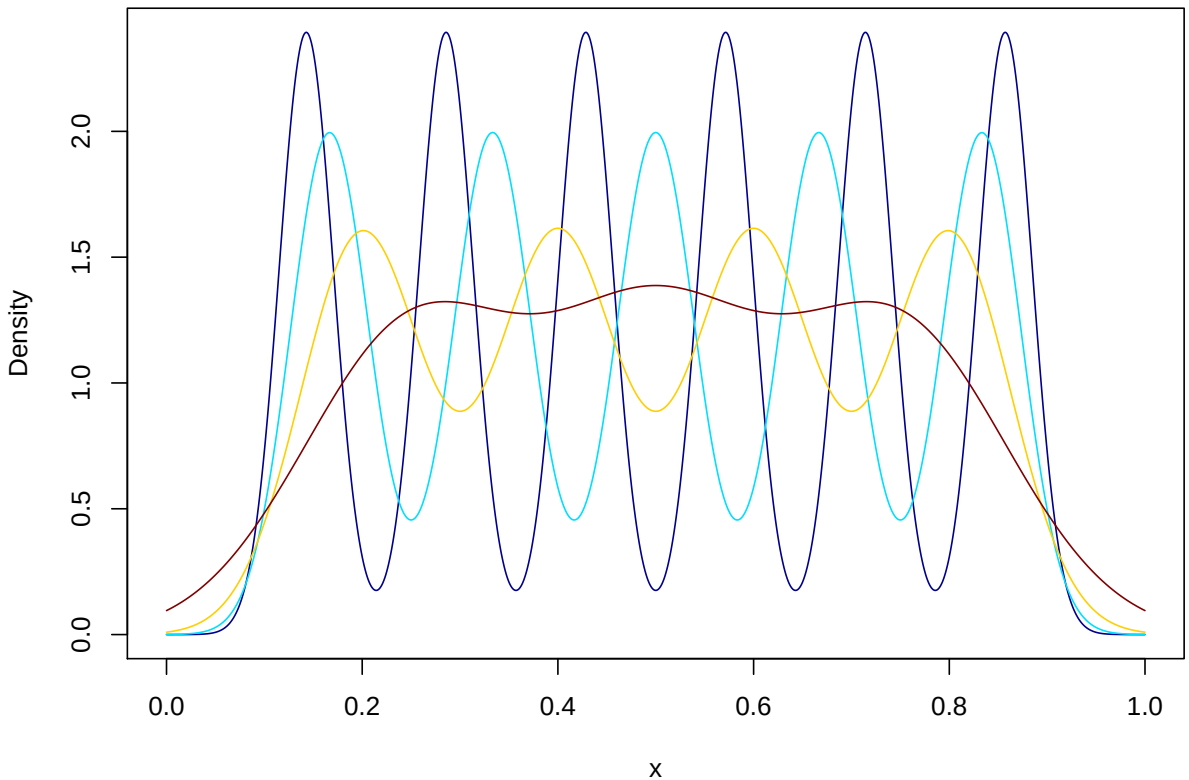


Figure 1: Examples of $g \in \{3, 4, 5, 6\}$ component Gaussian mixture models for $x \in \mathbb{R}$.

# Example 2: Maximum Quasi-Likelihood Estimation for Logistic Regressions

- Let $\mathbf{d}_n$ be a realization of the random sample $\mathbf{D}_n$, where $\boldsymbol{D}_i^\top = \left(\boldsymbol{X}_i^\top, Y_i\right)$, with $\boldsymbol{X}_i \in \mathbb{R}^p$ and $Y_i \in \{0,1\}$.
- Suppose that the DGP of $\boldsymbol{D}_i$ is unknown, but we can guess that the conditional distribution of $\boldsymbol{Y}_i|\boldsymbol{X}_i = \boldsymbol{x}_i$ can be best modeled by a density function $f_{\boldsymbol{\theta}}\left(y|\boldsymbol{X} = \boldsymbol{x}\right)$ in the family

$$\mathscr{F}_{\boldsymbol{\theta}} = \left\{ f_{\boldsymbol{\theta}} : f_{\boldsymbol{\theta}}\left(y|\boldsymbol{X} = \boldsymbol{x}\right) = \pi_{\boldsymbol{\theta}}^{y}\left(\boldsymbol{x}\right)\left[1 - \pi_{\boldsymbol{\theta}}\left(\boldsymbol{x}\right)\right]^{1-y}, \boldsymbol{\theta} \in \Theta \right\},$$

  for some $\boldsymbol{\theta}_0$.
- $\Theta = \mathbb{R}^p$ and

$$\pi_{\boldsymbol{\theta}}\left(\boldsymbol{x}\right) = \frac{\exp\left(\boldsymbol{\theta}^\top \boldsymbol{x}\right)}{1 + \exp\left(\boldsymbol{\theta}^\top \boldsymbol{x}\right)}$$

  is often known as the logistic link function.

# Example 2: Maximum Quasi-Likelihood Estimation for Logistic Regressions

- Let $\mathbf{d}_n$ be a realization of the random sample $\mathbf{D}_n$, where $\boldsymbol{D}_i^\top = \left(\boldsymbol{X}_i^\top, Y_i\right)$, with $\boldsymbol{X}_i \in \mathbb{R}^p$ and $Y_i \in \{0,1\}$.
- Suppose that the DGP of $\boldsymbol{D}_i$ is unknown, but we can guess that the conditional distribution of $\boldsymbol{Y}_i|\boldsymbol{X}_i = \boldsymbol{x}_i$ can be best modeled by a density function $f_{\boldsymbol{\theta}}\left(y|\boldsymbol{X} = \boldsymbol{x}\right)$ in the family

$$\mathscr{F}_{\boldsymbol{\theta}} = \left\{ f_{\boldsymbol{\theta}} : f_{\boldsymbol{\theta}}\left(y|\boldsymbol{X} = \boldsymbol{x}\right) = \pi_{\boldsymbol{\theta}}^{y}\left(\boldsymbol{x}\right)\left[1 - \pi_{\boldsymbol{\theta}}\left(\boldsymbol{x}\right)\right]^{1-y}, \boldsymbol{\theta} \in \Theta \right\},$$

  for some $\boldsymbol{\theta}_0$.

- $\Theta = \mathbb{R}^p$ and

$$\pi_{\boldsymbol{\theta}}\left(\boldsymbol{x}\right) = \frac{\exp\left(\boldsymbol{\theta}^\top \boldsymbol{x}\right)}{1 + \exp\left(\boldsymbol{\theta}^\top \boldsymbol{x}\right)}$$

  is often known as the logistic link function.

# Example 2: Maximum Quasi-Likelihood Estimation for Logistic Regressions

- Let $\mathbf{d}_n$ be a realization of the random sample $\mathbf{D}_n$, where $\boldsymbol{D}_i^\top = \left(\boldsymbol{X}_i^\top, Y_i\right)$, with $\boldsymbol{X}_i \in \mathbb{R}^p$ and $Y_i \in \{0,1\}$.

- Suppose that the DGP of $\boldsymbol{D}_i$ is unknown, but we can guess that the conditional distribution of $\boldsymbol{Y}_i|\boldsymbol{X}_i = \boldsymbol{x}_i$ can be best modeled by a density function $f_{\boldsymbol{\theta}}\left(y|\boldsymbol{X} = \boldsymbol{x}\right)$ in the family

$$\mathscr{F}_{\boldsymbol{\theta}} = \left\{ f_{\boldsymbol{\theta}} : f_{\boldsymbol{\theta}}\left(y|\boldsymbol{X} = \boldsymbol{x}\right) = \pi_{\boldsymbol{\theta}}^{y}\left(\boldsymbol{x}\right)\left[1 - \pi_{\boldsymbol{\theta}}\left(\boldsymbol{x}\right)\right]^{1-y}, \boldsymbol{\theta} \in \Theta \right\},$$

  for some $\boldsymbol{\theta}_0$.

- $\Theta = \mathbb{R}^p$ and

$$\pi_{\boldsymbol{\theta}}\left(\boldsymbol{x}\right) = \frac{\exp\left(\boldsymbol{\theta}^\top \boldsymbol{x}\right)}{1 + \exp\left(\boldsymbol{\theta}^\top \boldsymbol{x}\right)}$$

  is often known as the logistic link function.

# Example 2: Maximum Quasi-Likelihood Estimation for Logistic Regressions

- We say that $\mathscr{F}_{\boldsymbol{\theta}}$ is the family of logistic regression functions.
- Since we do not know the true DGP of $\boldsymbol{D}_i$, we can only say that the member $f_{\boldsymbol{\theta}_0}$ that best models the relationship between $\boldsymbol{X}_i$ and $Y_i$ is a misspecified model of the DGP.
- Under the misspecification framework of White (1982), we can estimate $\boldsymbol{\theta}_0$ under the potential misspecification by the minimum negative quasi-likelihood (maximum quasi-likelihood) estimator

$$\hat{\boldsymbol{\theta}}_n = \arg\min_{\boldsymbol{\theta}\in\Theta} \frac{1}{n}\sum_{i=1}^{n} l\left(\boldsymbol{\theta}; \boldsymbol{d}_i\right)$$

where

$$\begin{aligned} l\left(\boldsymbol{\theta}; \boldsymbol{d}_i\right) &= -\log f_{\boldsymbol{\theta}}\left(y_i | \boldsymbol{X}_i = \boldsymbol{x}_i\right) \\ &= -y_i \log \pi_{\boldsymbol{\theta}}\left(\boldsymbol{x}_i\right) - \left(1 - y_i\right) \log\left[1 - \pi_{\boldsymbol{\theta}}\left(\boldsymbol{x}_i\right)\right]. \end{aligned}$$

# Example 2: Maximum Quasi-Likelihood Estimation for Logistic Regressions

- We say that $\mathscr{F}_{\boldsymbol{\theta}}$ is the family of logistic regression functions.
- Since we do not know the true DGP of $\boldsymbol{D}_i$, we can only say that the member $f_{\boldsymbol{\theta}_0}$ that best models the relationship between $\boldsymbol{X}_i$ and $Y_i$ is a misspecified model of the DGP.
- Under the misspecification framework of White (1982), we can estimate $\boldsymbol{\theta}_0$ under the potential misspecification by the minimum negative quasi-likelihood (maximum quasi-likelihood) estimator

$$\hat{\boldsymbol{\theta}}_n = \arg\min_{\boldsymbol{\theta}\in\Theta} \frac{1}{n}\sum_{i=1}^{n} l\left(\boldsymbol{\theta}; \boldsymbol{d}_i\right)$$

where

$$\begin{aligned} l\left(\boldsymbol{\theta}; \boldsymbol{d}_i\right) &= -\log f_{\boldsymbol{\theta}}\left(y_i | \boldsymbol{X}_i = \boldsymbol{x}_i\right) \\ &= -y_i \log \pi_{\boldsymbol{\theta}}\left(\boldsymbol{x}_i\right) - \left(1 - y_i\right) \log\left[1 - \pi_{\boldsymbol{\theta}}\left(\boldsymbol{x}_i\right)\right]. \end{aligned}$$

# Example 2: Maximum Quasi-Likelihood Estimation for Logistic Regressions

- We say that $\mathscr{F}_{\boldsymbol{\theta}}$ is the family of logistic regression functions.
- Since we do not know the true DGP of $\boldsymbol{D}_i$, we can only say that the member $f_{\boldsymbol{\theta}_0}$ that best models the relationship between $\boldsymbol{X}_i$ and $Y_i$ is a misspecified model of the DGP.
- Under the misspecification framework of White (1982), we can estimate $\boldsymbol{\theta}_0$ under the potential misspecification by the minimum negative quasi-likelihood (maximum quasi-likelihood) estimator

$$\hat{\boldsymbol{\theta}}_n = \arg\min_{\boldsymbol{\theta}\in\Theta} \frac{1}{n}\sum_{i=1}^{n} l\left(\boldsymbol{\theta}; \boldsymbol{d}_i\right)$$

where

$$\begin{aligned} l\left(\boldsymbol{\theta}; \boldsymbol{d}_i\right) &= -\log f_{\boldsymbol{\theta}}\left(y_i | \boldsymbol{X}_i = \boldsymbol{x}_i\right) \\ &= -y_i \log \pi_{\boldsymbol{\theta}}\left(\boldsymbol{x}_i\right) - \left(1 - y_i\right) \log\left[1 - \pi_{\boldsymbol{\theta}}\left(\boldsymbol{x}_i\right)\right]. \end{aligned}$$

# Example 2: Maximum Quasi-Likelihood Estimation for Logistic Regressions

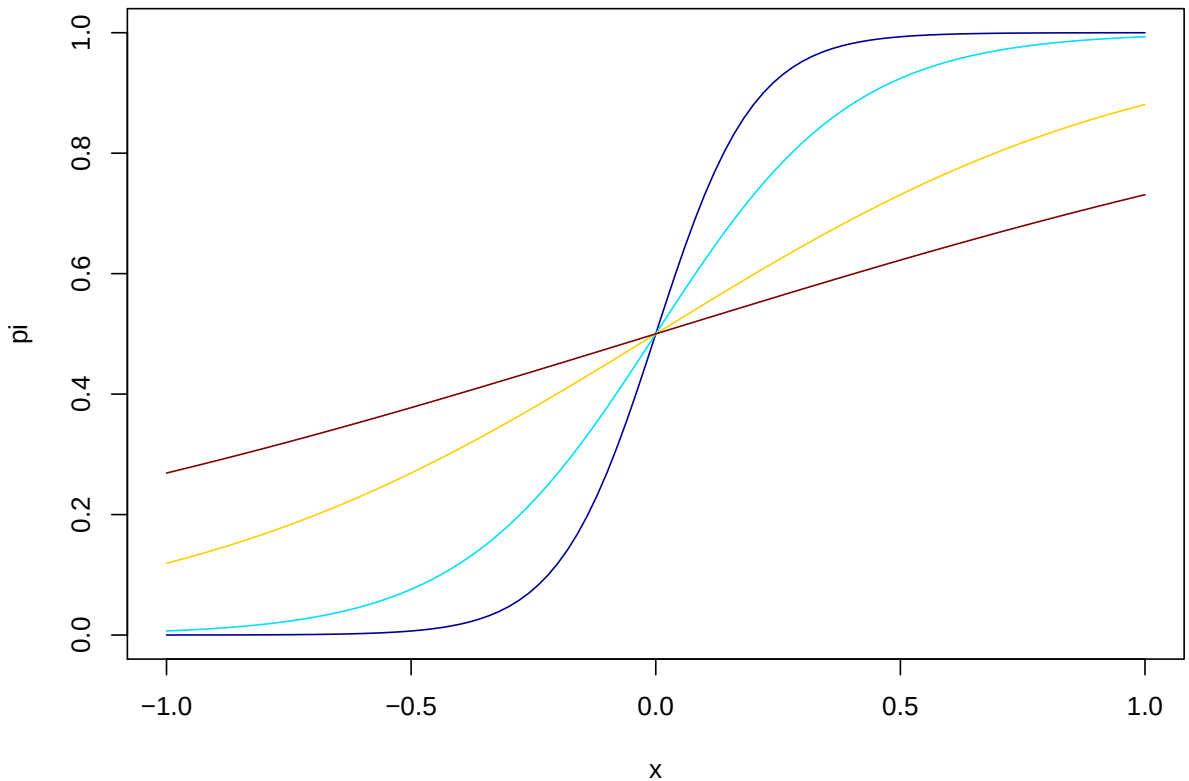


Figure 2: Examples of some logistic link functions for $x \in [-1, 1]$.

# Example 3: ERM for Linear Support Vector Machines

- Let $\mathbf{d}_n$ be a realization of the random sample $\mathbf{D}_n$, where $\boldsymbol{D}_i^\top = \left(\boldsymbol{X}_i^\top, Y_i\right)$, with $\boldsymbol{X}_i \in \mathbb{R}^p$ and $Y_i \in \{-1, 1\}$.
- Suppose that the DGP of $\boldsymbol{D}_i$ is unknown, but we know that the distribution of $\boldsymbol{X}_i | Y_i = -1$ is different to that of $\boldsymbol{X}_i | Y_i = +1$.
- Using $\mathbf{d}_n$, we wish to find a rule

$$\hat{y} = \begin{cases} +1 & \text{if } f_{\boldsymbol{\theta}}(\boldsymbol{x}) \geq 0, \\ -1 & \text{if } f_{\boldsymbol{\theta}}(\boldsymbol{x}) < 0, \end{cases}$$

  where $f_{\boldsymbol{\theta}}(\boldsymbol{x})$ arises is from the family of linear functions

$$\mathscr{F}_{\boldsymbol{\theta}} = \left\{ f_{\boldsymbol{\theta}} : f_{\boldsymbol{\theta}}(\boldsymbol{x}) = \alpha + \boldsymbol{\beta}^\top \boldsymbol{x}, \boldsymbol{\theta} \in \Theta \right\},$$

  such that for a new observation $Y_{n+1}$ from the same DGP, $\hat{Y}_{n+1}$ matches $Y_{n+1}$ with high probability.
- $\boldsymbol{\theta}^\top = \left(\alpha, \boldsymbol{\beta}^\top\right)$, where $\alpha \in \mathbb{R}$ and $\boldsymbol{\beta} \in \mathbb{R}^p$.

# Example 3: ERM for Linear Support Vector Machines

- Let $\mathbf{d}_n$ be a realization of the random sample $\mathbf{D}_n$, where $\boldsymbol{D}_i^\top = \left(\boldsymbol{X}_i^\top, Y_i\right)$, with $\boldsymbol{X}_i \in \mathbb{R}^p$ and $Y_i \in \{-1, 1\}$.
- Suppose that the DGP of $\boldsymbol{D}_i$ is unknown, but we know that the distribution of $\boldsymbol{X}_i | Y_i = -1$ is different to that of $\boldsymbol{X}_i | Y_i = +1$.
- Using $\mathbf{d}_n$, we wish to find a rule

$$\hat{y} = \begin{cases} +1 & \text{if } f_{\boldsymbol{\theta}}(\boldsymbol{x}) \geq 0, \\ -1 & \text{if } f_{\boldsymbol{\theta}}(\boldsymbol{x}) < 0, \end{cases}$$

  where $f_{\boldsymbol{\theta}}(\boldsymbol{x})$ arises is from the family of linear functions

$$\mathscr{F}_{\boldsymbol{\theta}} = \left\{ f_{\boldsymbol{\theta}} : f_{\boldsymbol{\theta}}(\boldsymbol{x}) = \alpha + \boldsymbol{\beta}^\top \boldsymbol{x}, \boldsymbol{\theta} \in \Theta \right\},$$

  such that for a new observation $Y_{n+1}$ from the same DGP, $\hat{Y}_{n+1}$ matches $Y_{n+1}$ with high probability.
- $\boldsymbol{\theta}^\top = \left(\alpha, \boldsymbol{\beta}^\top\right)$, where $\alpha \in \mathbb{R}$ and $\boldsymbol{\beta} \in \mathbb{R}^p$.

# Example 3: ERM for Linear Support Vector Machines

- Let $\mathbf{d}_n$ be a realization of the random sample $\mathbf{D}_n$, where $\boldsymbol{D}_i^\top = \left(\boldsymbol{X}_i^\top, Y_i\right)$, with $\boldsymbol{X}_i \in \mathbb{R}^p$ and $Y_i \in \{-1, 1\}$.
- Suppose that the DGP of $\boldsymbol{D}_i$ is unknown, but we know that the distribution of $\boldsymbol{X}_i | Y_i = -1$ is different to that of $\boldsymbol{X}_i | Y_i = +1$.
- Using $\mathbf{d}_n$, we wish to find a rule

$$\hat{y} = \begin{cases} +1 & \text{if } f_{\boldsymbol{\theta}}(\boldsymbol{x}) \geq 0, \\ -1 & \text{if } f_{\boldsymbol{\theta}}(\boldsymbol{x}) < 0, \end{cases}$$

  where $f_{\boldsymbol{\theta}}(\boldsymbol{x})$ arises is from the family of linear functions

$$\mathscr{F}_{\boldsymbol{\theta}} = \left\{ f_{\boldsymbol{\theta}} : f_{\boldsymbol{\theta}}(\boldsymbol{x}) = \alpha + \boldsymbol{\beta}^\top \boldsymbol{x}, \boldsymbol{\theta} \in \Theta \right\},$$

  such that for a new observation $Y_{n+1}$ from the same DGP, $\hat{Y}_{n+1}$ matches $Y_{n+1}$ with high probability.
- $\boldsymbol{\theta}^\top = \left(\alpha, \boldsymbol{\beta}^\top\right)$, where $\alpha \in \mathbb{R}$ and $\boldsymbol{\beta} \in \mathbb{R}^p$.

# Example 3: ERM for Linear Support Vector Machines

- Let $\mathbf{d}_n$ be a realization of the random sample $\mathbf{D}_n$, where $\boldsymbol{D}_i^\top = \left(\boldsymbol{X}_i^\top, Y_i\right)$, with $\boldsymbol{X}_i \in \mathbb{R}^p$ and $Y_i \in \{-1, 1\}$.
- Suppose that the DGP of $\boldsymbol{D}_i$ is unknown, but we know that the distribution of $\boldsymbol{X}_i | Y_i = -1$ is different to that of $\boldsymbol{X}_i | Y_i = +1$.
- Using $\mathbf{d}_n$, we wish to find a rule

$$\hat{y} = \begin{cases} +1 & \text{if } f_{\boldsymbol{\theta}}(\boldsymbol{x}) \geq 0, \\ -1 & \text{if } f_{\boldsymbol{\theta}}(\boldsymbol{x}) < 0, \end{cases}$$

  where $f_{\boldsymbol{\theta}}(\boldsymbol{x})$ arises is from the family of linear functions

$$\mathscr{F}_{\boldsymbol{\theta}} = \left\{ f_{\boldsymbol{\theta}} : f_{\boldsymbol{\theta}}(\boldsymbol{x}) = \alpha + \boldsymbol{\beta}^\top \boldsymbol{x}, \boldsymbol{\theta} \in \Theta \right\},$$

  such that for a new observation $Y_{n+1}$ from the same DGP, $\hat{Y}_{n+1}$ matches $Y_{n+1}$ with high probability.
- $\boldsymbol{\theta}^\top = \left(\alpha, \boldsymbol{\beta}^\top\right)$, where $\alpha \in \mathbb{R}$ and $\boldsymbol{\beta} \in \mathbb{R}^p$.

# Example 3: ERM for Linear Support Vector Machines

- In Cortes and Vapnik (1995), the hinge loss function

$$l(\boldsymbol{\theta}; \boldsymbol{d}_i) = [1 - y_i f_{\boldsymbol{\theta}}(\boldsymbol{x}_i)]_+$$

is suggested as a measurement of the loss incurred by a mismatched between $\hat{y}_i$ and $y_i$.

  - $[\cdot]_+ = \max\{0, \cdot\}$.

- Suppose that

$$\boldsymbol{\theta}_0 = \arg\max_{\boldsymbol{\theta}} \mathbb{E}[l(\boldsymbol{\theta}; \boldsymbol{D}_i)],$$

where $\mathbb{E}[\cdot]$ is the expectation operator under the DGP of $\boldsymbol{D}_i$.

- Cortes and Vapnik (1995) proposed the estimator

$$\hat{\boldsymbol{\theta}}_n = \arg\min_{\boldsymbol{\theta} \in \Theta} \left[ \frac{1}{n} \sum_{i=1}^{n} l(\boldsymbol{\theta}; \boldsymbol{d}_i) + \lambda \boldsymbol{\beta}^\top \boldsymbol{\beta} \right],$$

as an estimator for $\boldsymbol{\theta}_0$, where $\lambda \geq 0$ is a regulating constant.

# Example 3: ERM for Linear Support Vector Machines

- In Cortes and Vapnik (1995), the hinge loss function

$$l(\boldsymbol{\theta}; \boldsymbol{d}_i) = [1 - y_i f_{\boldsymbol{\theta}}(\boldsymbol{x}_i)]_+$$

is suggested as a measurement of the loss incurred by a mismatched between $\hat{y}_i$ and $y_i$.
  - $[\cdot]_+ = \max\{0, \cdot\}$.

- Suppose that

$$\boldsymbol{\theta}_0 = \arg\max_{\boldsymbol{\theta}} \mathbb{E}[l(\boldsymbol{\theta}; \boldsymbol{D}_i)],$$

where $\mathbb{E}[\cdot]$ is the expectation operator under the DGP of $\boldsymbol{D}_i$.

- Cortes and Vapnik (1995) proposed the estimator

$$\hat{\boldsymbol{\theta}}_n = \arg\min_{\boldsymbol{\theta} \in \Theta} \left[ \frac{1}{n} \sum_{i=1}^{n} l(\boldsymbol{\theta}; \boldsymbol{d}_i) + \lambda \boldsymbol{\beta}^\top \boldsymbol{\beta} \right],$$

as an estimator for $\boldsymbol{\theta}_0$, where $\lambda \geq 0$ is a regulating constant.

# Example 3: ERM for Linear Support Vector Machines

- In Cortes and Vapnik (1995), the hinge loss function

$$l(\boldsymbol{\theta}; \boldsymbol{d}_i) = [1 - y_i f_{\boldsymbol{\theta}}(\boldsymbol{x}_i)]_+$$

is suggested as a measurement of the loss incurred by a mismatched between $\hat{y}_i$ and $y_i$.
  - $[\cdot]_+ = \max\{0, \cdot\}$.

- Suppose that

$$\boldsymbol{\theta}_0 = \arg\max_{\boldsymbol{\theta}} \ \mathbb{E}[l(\boldsymbol{\theta}; \boldsymbol{D}_i)],$$

where $\mathbb{E}[\cdot]$ is the expectation operator under the DGP of $\boldsymbol{D}_i$.

- Cortes and Vapnik (1995) proposed the estimator

$$\hat{\boldsymbol{\theta}}_n = \arg\min_{\boldsymbol{\theta} \in \Theta} \ \left[ \frac{1}{n} \sum_{i=1}^{n} l(\boldsymbol{\theta}; \boldsymbol{d}_i) + \lambda \boldsymbol{\beta}^\top \boldsymbol{\beta} \right],$$

as an estimator for $\boldsymbol{\theta}_0$, where $\lambda \geq 0$ is a regulating constant.

# Example 3: ERM for Linear Support Vector Machines

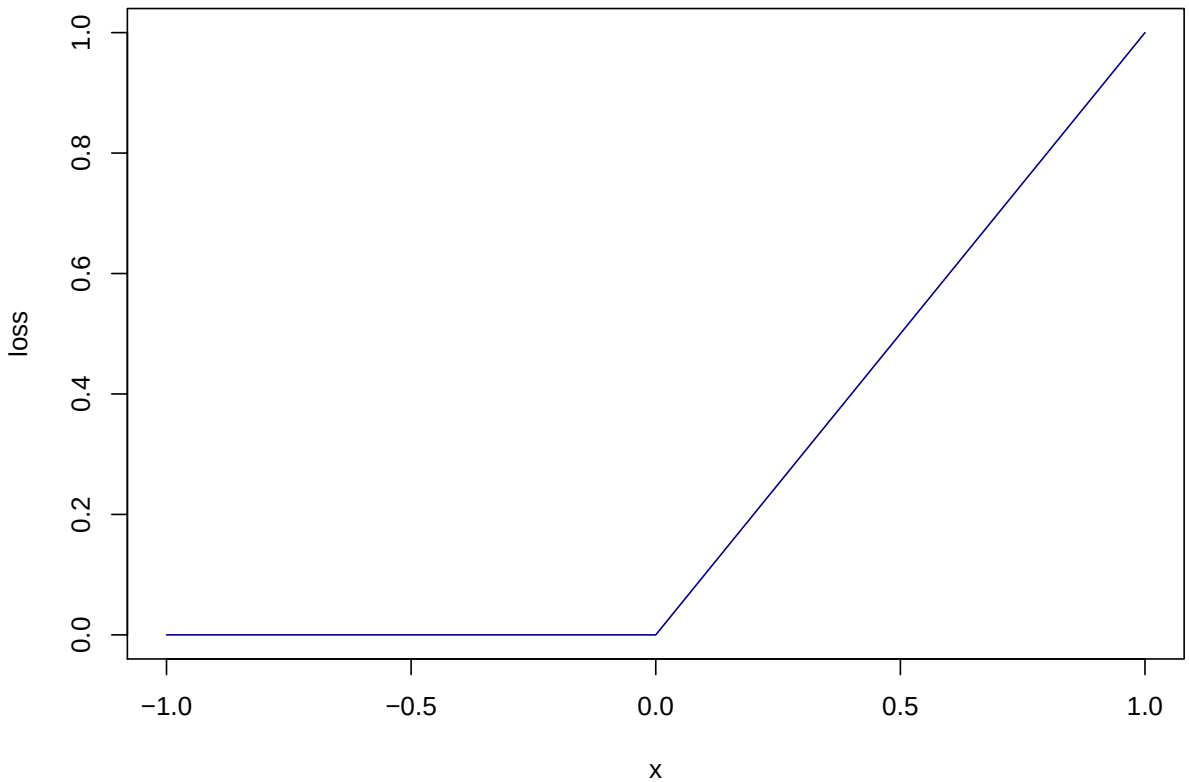


Figure 3: The hinge loss function.

# Solving Optimization Problems

- In general, to solve problems of form (1), that is

$$\arg\min_{\boldsymbol{\theta}\in\Theta} \frac{1}{n}\sum_{i=1}^{n} l(\boldsymbol{\theta};\boldsymbol{d}_i)+\lambda R(\boldsymbol{\theta}),$$

  we require a solution to the usual first order conditions (FOC) of calculus:

$$\nabla L(\boldsymbol{\theta};\mathbf{d}_n)=\mathbf{0}.$$

  - $L(\boldsymbol{\theta};\mathbf{d}_n)=\frac{1}{n}\sum_{i=1}^{n} l(\boldsymbol{\theta};\boldsymbol{d}_i)+\lambda R(\boldsymbol{\theta})$.
  - $\nabla[\cdot]$ is the gradient operator.

- Second order conditions may also be useful in cases where the objective $L(\cdot;\mathbf{d}_n)$ is not convex.

# Solving Optimization Problems

- In general, to solve problems of form (1), that is

$$\arg\min_{\boldsymbol{\theta}\in\Theta} \frac{1}{n}\sum_{i=1}^{n} l(\boldsymbol{\theta};\boldsymbol{d}_i)+\lambda R(\boldsymbol{\theta}),$$

  we require a solution to the usual first order conditions (FOC) of calculus:

$$\nabla L(\boldsymbol{\theta};\mathbf{d}_n)=\mathbf{0}.$$

  - $L(\boldsymbol{\theta};\mathbf{d}_n)=\frac{1}{n}\sum_{i=1}^{n} l(\boldsymbol{\theta};\boldsymbol{d}_i)+\lambda R(\boldsymbol{\theta})$.
  - $\nabla[\cdot]$ is the gradient operator.
- Second order conditions may also be useful in cases where the objective $L(\cdot;\mathbf{d}_n)$ is not convex.

# Difficulties (1)

- In Example 1 (GMM),

$$l(\boldsymbol{\theta}; \boldsymbol{x}_i) = -\log \sum_{z=1}^{g} \pi_z \phi_p(\boldsymbol{x}_i; \boldsymbol{\mu}_z, \boldsymbol{\Sigma}_z),$$

  where

$$\phi_p(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = |2\pi\boldsymbol{\Sigma}|^{-1/2} \exp\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right].$$

- $l(\cdot; \boldsymbol{x}_i)$ has the log-sum-exp form (cf. Boyd and Vandenberghe, 2004), which does not allow for a closed form solution to the FOC.

# Difficulties (1)

- In Example 1 (GMM),

$$l(\boldsymbol{\theta}; \boldsymbol{x}_i) = -\log \sum_{z=1}^{g} \pi_z \phi_p(\boldsymbol{x}_i; \boldsymbol{\mu}_z, \boldsymbol{\Sigma}_z),$$

  where

$$\phi_p(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = |2\pi\boldsymbol{\Sigma}|^{-1/2} \exp\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right].$$

- $l(\cdot; \boldsymbol{x}_i)$ has the log-sum-exp form (cf. Boyd and Vandenberghe, 2004), which does not allow for a closed form solution to the FOC.

# Difficulties (2)

- In Example 2 (Logistic),

$$\begin{aligned} l(\boldsymbol{\theta}; \boldsymbol{d}_i) &= -y_i \log \left[ \frac{\exp\left(\boldsymbol{\theta}^\top \boldsymbol{x}_i\right)}{1+\exp\left(\boldsymbol{\theta}^\top \boldsymbol{x}_i\right)} \right] \\ &\quad -(1-y_i) \log \left[ \frac{1}{1+\exp\left(\boldsymbol{\theta}^\top \boldsymbol{x}_i\right)} \right] \\ &= -y_i \boldsymbol{\theta}^\top \boldsymbol{x}_i + \log\left[1+\exp\left(\boldsymbol{\theta}^\top \boldsymbol{x}_i\right)\right]. \end{aligned}$$

- $l(\cdot; \boldsymbol{x}_i)$ again has the log-sum-exp form, which does not allow for a closed form solution to the FOC.

# Difficulties (2)

- In Example 2 (Logistic),

$$\begin{aligned} l(\boldsymbol{\theta}; \boldsymbol{d}_i) &= -y_i \log \left[ \frac{\exp\left(\boldsymbol{\theta}^\top \boldsymbol{x}_i\right)}{1+\exp\left(\boldsymbol{\theta}^\top \boldsymbol{x}_i\right)} \right] \\ &\quad -(1-y_i)\log \left[ \frac{1}{1+\exp\left(\boldsymbol{\theta}^\top \boldsymbol{x}_i\right)} \right] \\ &= -y_i \boldsymbol{\theta}^\top \boldsymbol{x}_i + \log\left[1+\exp\left(\boldsymbol{\theta}^\top \boldsymbol{x}_i\right)\right]. \end{aligned}$$

- $l(\cdot; \boldsymbol{x}_i)$ again has the log-sum-exp form, which does not allow for a closed form solution to the FOC.

# Difficulties (3)

- In Example 3 (SVM),

$$l(\boldsymbol{\theta}; \boldsymbol{d}_i) \quad = \quad [1 - y_i f_{\boldsymbol{\theta}}(\boldsymbol{x}_i)]_+ .$$

- $l(\cdot; \boldsymbol{x}_i)$ is not differentiable when $1 - y_i f_{\boldsymbol{\theta}}(\boldsymbol{x}_i) = 0$ and thus we cannot solve the usual FOC for a solution to (1).

# Difficulties (3)

- In Example 3 (SVM),

$$l(\boldsymbol{\theta}; \boldsymbol{d}_i) \quad = \quad [1 - y_i f_{\boldsymbol{\theta}}(\boldsymbol{x}_i)]_+ .$$

- $l(\cdot; \boldsymbol{x}_i)$ is not differentiable when $1 - y_i f_{\boldsymbol{\theta}}(\boldsymbol{x}_i) = 0$ and thus we cannot solve the usual FOC for a solution to (1).

# Majorization

- Suppose that we wish to minimize some objective function $o(\boldsymbol{\upsilon})$, where $\boldsymbol{\upsilon} \in \Upsilon \subset \mathbb{R}^d$, for some $d \in \mathbb{N}$, and where $o(\cdot)$ is difficult to operate on for some reason. For example:
  - FOC cannot be solved in closed form.
  - $o(\cdot)$ is not differentiable.
- Let $m(\boldsymbol{\upsilon}; \boldsymbol{\omega})$ be a majorizer to $o(\cdot)$, if $m(\boldsymbol{\upsilon}; \boldsymbol{\upsilon}) = o(\boldsymbol{\upsilon})$ and $m(\boldsymbol{\upsilon}; \boldsymbol{\omega}) \geq o(\boldsymbol{\upsilon})$ for all $\boldsymbol{\omega} \in \Upsilon$.
- We generally choose the majorizer to be a simpler function than $o(\cdot)$, in the active variable $\boldsymbol{\upsilon}$.

# Majorization

- Suppose that we wish to minimize some objective function $o(\boldsymbol{\upsilon})$, where $\boldsymbol{\upsilon} \in \Upsilon \subset \mathbb{R}^d$, for some $d \in \mathbb{N}$, and where $o(\cdot)$ is difficult to operate on for some reason. For example:
  - FOC cannot be solved in closed form.
  - $o(\cdot)$ is not differentiable.
- Let $m(\boldsymbol{\upsilon}; \boldsymbol{\omega})$ be a majorizer to $o(\cdot)$, if $m(\boldsymbol{\upsilon}; \boldsymbol{\upsilon}) = o(\boldsymbol{\upsilon})$ and $m(\boldsymbol{\upsilon}; \boldsymbol{\omega}) \geq o(\boldsymbol{\upsilon})$ for all $\boldsymbol{\omega} \in \Upsilon$.
- We generally choose the majorizer to be a simpler function than $o(\cdot)$, in the active variable $\boldsymbol{\upsilon}$.

# Majorization

- Suppose that we wish to minimize some objective function $o(\boldsymbol{\upsilon})$, where $\boldsymbol{\upsilon} \in \Upsilon \subset \mathbb{R}^d$, for some $d \in \mathbb{N}$, and where $o(\cdot)$ is difficult to operate on for some reason. For example:
  - FOC cannot be solved in closed form.
  - $o(\cdot)$ is not differentiable.
- Let $m(\boldsymbol{\upsilon}; \boldsymbol{\omega})$ be a majorizer to $o(\cdot)$, if $m(\boldsymbol{\upsilon}; \boldsymbol{\upsilon}) = o(\boldsymbol{\upsilon})$ and $m(\boldsymbol{\upsilon}; \boldsymbol{\omega}) \geq o(\boldsymbol{\upsilon})$ for all $\boldsymbol{\omega} \in \Upsilon$.
- We generally choose the majorizer to be a simpler function than $o(\cdot)$, in the active variable $\boldsymbol{\upsilon}$.

# Majorization

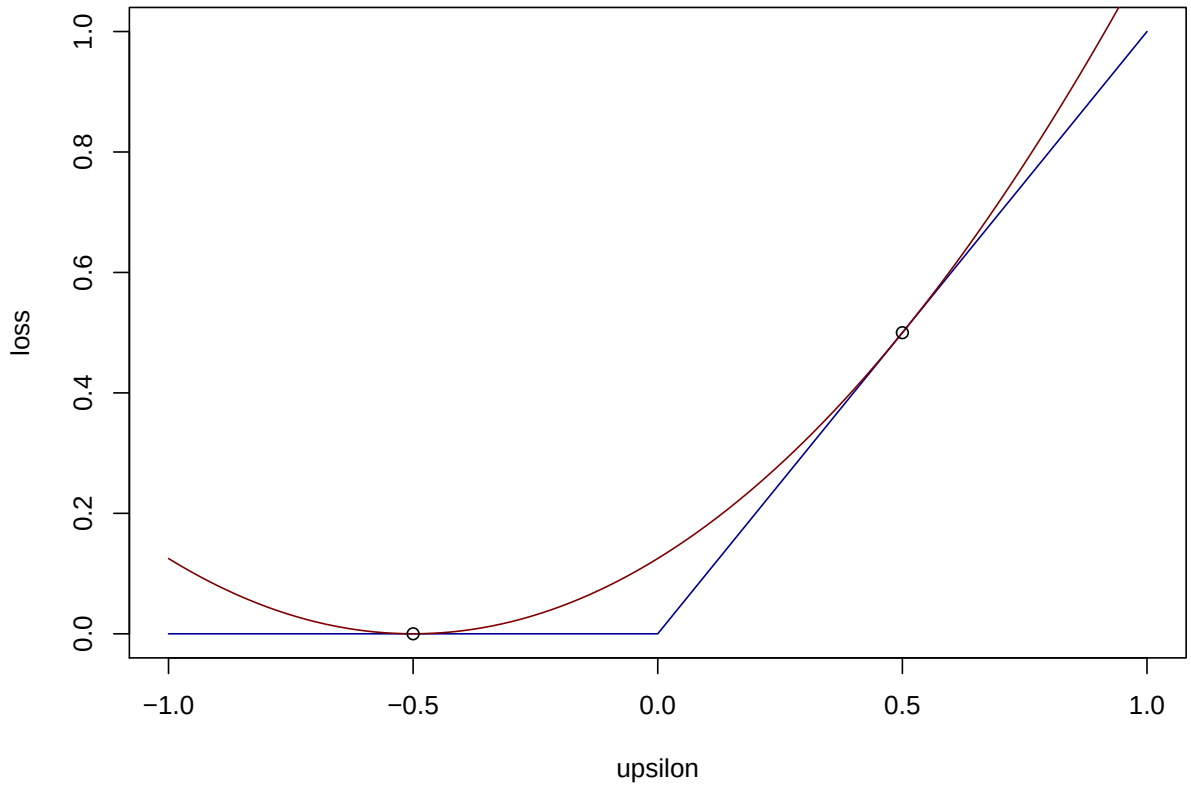


Figure 4: Hinge loss function ($o(\upsilon)$) and its majorizer $m(\upsilon;\omega)$ at $\omega \in \{-1/2, +1/2\}$.

# Majorization–Maximization Algorithm

- Since we choose $m(\cdot;\boldsymbol{\omega})$, it can be chosen to yield a simple solution to the problem

$$\min_{\boldsymbol{\upsilon}\in\Upsilon} \; m(\boldsymbol{\upsilon};\boldsymbol{\omega}).$$

- Starting from some initialization $\boldsymbol{\upsilon}^{(0)}$, let $\boldsymbol{\upsilon}^{(r)}$ be the $r$th iteration of our MM algorithm for minimizing $o(\cdot)$.

- The MM algorithm (using majorizer $m(\cdot,\boldsymbol{\omega})$) is then defined by the iterative scheme

$$\boldsymbol{\upsilon}^{(r+1)} = \arg\min_{\boldsymbol{\upsilon}\in\Upsilon} \; m\left(\boldsymbol{\upsilon};\boldsymbol{\upsilon}^{(r)}\right). \tag{2}$$

# Majorization–Maximization Algorithm

- Since we choose $m(\cdot;\boldsymbol{\omega})$, it can be chosen to yield a simple solution to the problem

$$\min_{\boldsymbol{\upsilon}\in\Upsilon} \; m(\boldsymbol{\upsilon};\boldsymbol{\omega}).$$

- Starting from some initialization $\boldsymbol{\upsilon}^{(0)}$, let $\boldsymbol{\upsilon}^{(r)}$ be the $r$th iteration of our MM algorithm for minimizing $o(\cdot)$.

- The MM algorithm (using majorizer $m(\cdot,\boldsymbol{\omega})$) is then defined by the iterative scheme

$$\boldsymbol{\upsilon}^{(r+1)} = \arg\min_{\boldsymbol{\upsilon}\in\Upsilon} \; m\left(\boldsymbol{\upsilon};\boldsymbol{\upsilon}^{(r)}\right). \tag{2}$$

# Majorization–Maximization Algorithm

- Since we choose $m(\cdot;\boldsymbol{\omega})$, it can be chosen to yield a simple solution to the problem

$$\min_{\boldsymbol{\upsilon}\in\Upsilon} \; m(\boldsymbol{\upsilon};\boldsymbol{\omega}).$$

- Starting from some initialization $\boldsymbol{\upsilon}^{(0)}$, let $\boldsymbol{\upsilon}^{(r)}$ be the $r$th iteration of our MM algorithm for minimizing $o(\cdot)$.

- The MM algorithm (using majorizer $m(\cdot,\boldsymbol{\omega})$) is then defined by the iterative scheme

$$\boldsymbol{\upsilon}^{(r+1)} = \arg\min_{\boldsymbol{\upsilon}\in\Upsilon} \; m\left(\boldsymbol{\upsilon};\boldsymbol{\upsilon}^{(r)}\right). \tag{2}$$

# Descent Property

- Recall that $m(\boldsymbol{v};\boldsymbol{\omega})$ majorizes $o(\cdot)$ if $m(\boldsymbol{v};\boldsymbol{v}) = o(\boldsymbol{v})$ and $m(\boldsymbol{v};\boldsymbol{\omega}) \geq o(\boldsymbol{v})$ for all $\boldsymbol{\omega} \in \Upsilon$.

- By (2), and the definition of a majorizer, we have the chain of inequalities

$$\begin{aligned} o\left(\boldsymbol{v}^{(r)}\right) &= m\left(\boldsymbol{v}^{(r)};\boldsymbol{v}^{(r)}\right) \\ &\geq m\left(\boldsymbol{v}^{(r+1)};\boldsymbol{v}^{(r)}\right) \\ &\geq o\left(\boldsymbol{v}^{(r+1)}\right), \end{aligned}$$

  for each $r \in \mathbb{N}$.

**Proposition 1**

*The MM algorithm sequence of objective evaluates $o\left(\boldsymbol{v}^{(r)}\right)$ is decreasing in $r$.*

# Descent Property

- Recall that $m(\boldsymbol{\upsilon};\boldsymbol{\omega})$ majorizes $o(\cdot)$ if $m(\boldsymbol{\upsilon};\boldsymbol{\upsilon}) = o(\boldsymbol{\upsilon})$ and $m(\boldsymbol{\upsilon};\boldsymbol{\omega}) \geq o(\boldsymbol{\upsilon})$ for all $\boldsymbol{\omega} \in \Upsilon$.

- By (2), and the definition of a majorizer, we have the chain of inequalities

$$\begin{aligned} o\left(\boldsymbol{\nu}^{(r)}\right) &= m\left(\boldsymbol{\upsilon}^{(r)};\boldsymbol{\upsilon}^{(r)}\right) \\ &\geq m\left(\boldsymbol{\upsilon}^{(r+1)};\boldsymbol{\upsilon}^{(r)}\right) \\ &\geq o\left(\boldsymbol{\upsilon}^{(r+1)}\right), \end{aligned}$$

  for each $r \in \mathbb{N}$.

**Proposition 1**

*The MM algorithm sequence of objective evaluates $o\left(\boldsymbol{\upsilon}^{(r)}\right)$ is decreasing in $r$.*

# Descent Property

- Recall that $m(\boldsymbol{v};\boldsymbol{\omega})$ majorizes $o(\cdot)$ if $m(\boldsymbol{v};\boldsymbol{v}) = o(\boldsymbol{v})$ and $m(\boldsymbol{v};\boldsymbol{\omega}) \geq o(\boldsymbol{v})$ for all $\boldsymbol{\omega} \in \Upsilon$.

- By (2), and the definition of a majorizer, we have the chain of inequalities

$$\begin{aligned} o\left(\boldsymbol{v}^{(r)}\right) &= m\left(\boldsymbol{v}^{(r)};\boldsymbol{v}^{(r)}\right) \\ &\geq m\left(\boldsymbol{v}^{(r+1)};\boldsymbol{v}^{(r)}\right) \\ &\geq o\left(\boldsymbol{v}^{(r+1)}\right), \end{aligned}$$

  for each $r \in \mathbb{N}$.

**Proposition 1**

*The MM algorithm sequence of objective evaluates $o\left(\boldsymbol{v}^{(r)}\right)$ is decreasing in $r$.*

# Global Convergence Property

- Define the directional derivative of $o(\boldsymbol{\upsilon})$ in the direction of $\boldsymbol{\delta}$ as

$$\dot{o}_{\boldsymbol{\delta}}(\boldsymbol{\upsilon}) = \lim_{\lambda \downarrow 0} \frac{o(\boldsymbol{\upsilon} + \lambda \boldsymbol{\delta}) - o(\boldsymbol{\upsilon})}{\lambda},$$

  and say that $\boldsymbol{\upsilon}^*$ is a stationary point of $o(\cdot)$ if $\dot{o}_{\boldsymbol{\delta}}(\boldsymbol{\upsilon}^*) \geq 0$ for all $\boldsymbol{\delta}$ such that $\boldsymbol{\upsilon}^* + \boldsymbol{\delta} \in \Upsilon$.

- Let $\boldsymbol{\upsilon}^{(\infty)} = \lim_{r \to \infty} \boldsymbol{\upsilon}^{(r)}$ be a limit point of MM algorithm (2).

**Theorem 1 (Razaviyayn, 2013)**

*If $o(\cdot)$ is differentiable, then every limit point $\boldsymbol{\upsilon}^{(\infty)}$ is a stationary point of $o(\cdot)$.*

# Global Convergence Property

- Define the directional derivative of $o(\boldsymbol{\upsilon})$ in the direction of $\boldsymbol{\delta}$ as

$$\dot{o}_{\boldsymbol{\delta}}(\boldsymbol{\upsilon}) = \lim_{\lambda \downarrow 0} \frac{o(\boldsymbol{\upsilon} + \lambda \boldsymbol{\delta}) - o(\boldsymbol{\upsilon})}{\lambda},$$

  and say that $\boldsymbol{\upsilon}^*$ is a stationary point of $o(\cdot)$ if $\dot{o}_{\boldsymbol{\delta}}(\boldsymbol{\upsilon}^*) \geq 0$ for all $\boldsymbol{\delta}$ such that $\boldsymbol{\upsilon}^* + \boldsymbol{\delta} \in \Upsilon$.

- Let $\boldsymbol{\upsilon}^{(\infty)} = \lim_{r \to \infty} \boldsymbol{\upsilon}^{(r)}$ be a limit point of MM algorithm (2).

**Theorem 1 (Razaviyayn, 2013)**

*If $o(\cdot)$ is differentiable, then every limit point $\boldsymbol{\upsilon}^{(\infty)}$ is a stationary point of $o(\cdot)$.*

# Global Convergence Property

- Define the directional derivative of $o(\boldsymbol{\upsilon})$ in the direction of $\boldsymbol{\delta}$ as

$$\dot{o}_{\boldsymbol{\delta}}(\boldsymbol{\upsilon}) = \lim_{\lambda \downarrow 0} \frac{o(\boldsymbol{\upsilon} + \lambda \boldsymbol{\delta}) - o(\boldsymbol{\upsilon})}{\lambda},$$

  and say that $\boldsymbol{\upsilon}^*$ is a stationary point of $o(\cdot)$ if $\dot{o}_{\boldsymbol{\delta}}(\boldsymbol{\upsilon}^*) \geq 0$ for all $\boldsymbol{\delta}$ such that $\boldsymbol{\upsilon}^* + \boldsymbol{\delta} \in \Upsilon$.

- Let $\boldsymbol{\upsilon}^{(\infty)} = \lim_{r \to \infty} \boldsymbol{\upsilon}^{(r)}$ be a limit point of MM algorithm (2).

**Theorem 1 (Razaviyayn, 2013)**

*If $o(\cdot)$ is differentiable, then every limit point $\boldsymbol{\upsilon}^{(\infty)}$ is a stationary point of $o(\cdot)$.*

# Global Convergence Property

(A1) $m(\boldsymbol{\upsilon};\boldsymbol{\omega})$ is continuous in both $\boldsymbol{\upsilon}$ and $\boldsymbol{\omega}$.

(A2) $\dot{o}_{\boldsymbol{\delta}}(\boldsymbol{\omega}) = \dot{m}_{\boldsymbol{\delta}}(\boldsymbol{\upsilon};\boldsymbol{\omega})|_{\boldsymbol{\upsilon}=\boldsymbol{\omega}}$ for all $\boldsymbol{\delta}$ with $\boldsymbol{\omega}+\boldsymbol{\delta} \in \Upsilon$.

**Theorem 2 (Razaviyayn, 2013)**

*If Assumptions (A1) and (A2) are satisfied, then every limit point $\boldsymbol{\upsilon}^{(\infty)}$ is a stationary point of $o(\cdot)$.*

# Global Convergence Property

(A1) $m(\boldsymbol{\upsilon};\boldsymbol{\omega})$ is continuous in both $\boldsymbol{\upsilon}$ and $\boldsymbol{\omega}$.

(A2) $\dot{o}_{\boldsymbol{\delta}}(\boldsymbol{\omega}) = \dot{m}_{\boldsymbol{\delta}}(\boldsymbol{\upsilon};\boldsymbol{\omega})|_{\boldsymbol{\upsilon}=\boldsymbol{\omega}}$ for all $\boldsymbol{\delta}$ with $\boldsymbol{\omega}+\boldsymbol{\delta} \in \Upsilon$.

**Theorem 2 (Razaviyayn, 2013)**

*If Assumptions (A1) and (A2) are satisfied, then every limit point $\boldsymbol{\upsilon}^{(\infty)}$ is a stationary point of $o(\cdot)$.*

# Global Convergence Property

(A1) $m(\boldsymbol{\upsilon};\boldsymbol{\omega})$ is continuous in both $\boldsymbol{\upsilon}$ and $\boldsymbol{\omega}$.

(A2) $\dot{o}_{\boldsymbol{\delta}}(\boldsymbol{\omega}) = \dot{m}_{\boldsymbol{\delta}}(\boldsymbol{\upsilon};\boldsymbol{\omega})|_{\boldsymbol{\upsilon}=\boldsymbol{\omega}}$ for all $\boldsymbol{\delta}$ with $\boldsymbol{\omega}+\boldsymbol{\delta} \in \Upsilon$.

**Theorem 2 (Razaviyayn, 2013)**

*If Assumptions (A1) and (A2) are satisfied, then every limit point* $\boldsymbol{\upsilon}^{(\infty)}$ *is a stationary point of* $o(\cdot)$.

# Useful Majorizers (1)

- Jensen's inequality:
  - Let $f(x)$ be a convex function, for $x \in \mathbb{R}$ and let $\boldsymbol{v}^\top = (v_1, \ldots, v_d)$, where $v_k > 0$, for $k \in [d]$.
  - The objective

$$o(\boldsymbol{v}) = f\left(\sum_{k=1}^{d} v_k\right)$$

can be majorized at $\boldsymbol{\omega}$ by

$$m(\boldsymbol{v}; \boldsymbol{\omega}) = \sum_{k=1}^{d} \frac{\omega_k}{\sum_{k'=1}^{d} \omega_{k'}} f\left(\frac{\sum_{k'=1}^{d} \omega_{k'}}{\omega_k} v_k\right),$$

where $\boldsymbol{\omega}^\top = (\omega_1, \ldots, \omega_d)$.

# Useful Majorizers (2)

- Quadratic upper bound:
  - Let $f(\boldsymbol{x})$ convex and twice differentiable in $\boldsymbol{x} \in \mathbb{R}^d$ and let $\mathbf{H}$ be some matrix such that

$$\mathbf{H} - \mathbb{H} f(\boldsymbol{x})$$

    is positive definite, where $\mathbb{H}[\cdot]$ is the Hessian operator.
  - The objective

$$o(\boldsymbol{v}) = f(\boldsymbol{v})$$

    can be majorized at $\boldsymbol{\omega}$ by

$$\begin{aligned} m(\boldsymbol{v}; \boldsymbol{\omega}) &= f(\boldsymbol{\omega}) + (\boldsymbol{v} - \boldsymbol{\omega})^\top \nabla f(\boldsymbol{\omega}) \\ &\quad + \frac{1}{2} (\boldsymbol{v} - \boldsymbol{\omega})^\top \mathbf{H} (\boldsymbol{v} - \boldsymbol{\omega}). \end{aligned}$$

# Useful Majorizers (3)

- Supporting hyperplane inequality:
  - Let $f(\boldsymbol{x})$ be a concave and differentiable in $\boldsymbol{x} \in \mathbb{R}^d$.
  - The objective

$$o(\boldsymbol{v}) = f(\boldsymbol{v})$$

    can be majorized at $\boldsymbol{\omega}$ by

$$m(\boldsymbol{v}; \boldsymbol{\omega}) \quad = \quad f(\boldsymbol{\omega}) + (\boldsymbol{v} - \boldsymbol{\omega})^\top \nabla f(\boldsymbol{\omega}).$$

# Relationship to the EM Algorithm

- Under the EM algorithm framework of Dempster et al. (1977), given some data $\mathbf{D}_n$, and some density function

$$f_{\boldsymbol{\upsilon}}(\mathbf{d}_n)$$

over $\mathbf{D}_n$ that characterizes the DGP of that is parameterized by $\boldsymbol{\upsilon} \in \Upsilon \subset \mathbb{R}^d$, we wish to compute the maximum likelihood estimator

$$\hat{\boldsymbol{\upsilon}}_n \quad = \quad \arg\min_{\boldsymbol{\upsilon} \in \Upsilon} \; o(\boldsymbol{\upsilon}),$$

where

$$o(\boldsymbol{\upsilon}) = -\log_{\boldsymbol{\upsilon}} f(\mathbf{d}_n).$$

# Relationship to the EM Algorithm

- Suppose that $o(\boldsymbol{\upsilon}) = -\log f_{\boldsymbol{\upsilon}}(\mathbf{d}_n)$ be a negative log-likelihood on $\mathbf{D}_n$ that is difficult to manipulate.
- Introduce some latent data $\mathbf{U}$, where we know that the joint probability density function between $\mathbf{D}_n$ and $\mathbf{U}$ has the form

$$f_{\boldsymbol{\upsilon}}(\mathbf{u}, \mathbf{d}_n). \tag{3}$$

- Using (3), we can majorize $o(\boldsymbol{\upsilon})$ at $\boldsymbol{\upsilon}^{(r)}$ by

$$\begin{aligned} m\left(\boldsymbol{\upsilon}; \boldsymbol{\upsilon}^{(r)}\right) &= -\int \log f_{\boldsymbol{\upsilon}}(\mathbf{u}, \mathbf{d}_n) f_{\boldsymbol{\upsilon}^{(r)}}(\mathbf{u}|\mathbf{d}_n) \, \mathrm{d}\mathbf{u} \\ &\quad + \int \log f_{\boldsymbol{\theta}^{(r)}}(\mathbf{u}|\mathbf{d}_n) f_{\boldsymbol{\theta}^{(r)}}(\mathbf{u}|\mathbf{d}_n) \, \mathrm{d}\mathbf{u}. \end{aligned}$$

- The update scheme defined by

$$\boldsymbol{\upsilon}^{(r+1)} = \arg\min_{\boldsymbol{\upsilon} \in \Upsilon} m\left(\boldsymbol{\upsilon}; \boldsymbol{\upsilon}^{(r)}\right)$$

is the EM algorithm.

# Relationship to the EM Algorithm

- Suppose that $o(\boldsymbol{\upsilon}) = -\log f_{\boldsymbol{\upsilon}}(\mathbf{d}_n)$ be a negative log-likelihood on $\mathbf{D}_n$ that is difficult to manipulate.
- Introduce some latent data $\mathbf{U}$, where we know that the joint probability density function between $\mathbf{D}_n$ and $\mathbf{U}$ has the form

$$f_{\boldsymbol{\upsilon}}(\mathbf{u}, \mathbf{d}_n). \tag{3}$$

- Using (3), we can majorize $o(\boldsymbol{\upsilon})$ at $\boldsymbol{\upsilon}^{(r)}$ by

$$\begin{aligned} m\left(\boldsymbol{\upsilon}; \boldsymbol{\upsilon}^{(r)}\right) &= -\int \log f_{\boldsymbol{\upsilon}}(\mathbf{u}, \mathbf{d}_n) f_{\boldsymbol{\upsilon}^{(r)}}(\mathbf{u}|\mathbf{d}_n) \, \mathrm{d}\mathbf{u} \\ &\quad + \int \log f_{\boldsymbol{\theta}^{(r)}}(\mathbf{u}|\mathbf{d}_n) f_{\boldsymbol{\theta}^{(r)}}(\mathbf{u}|\mathbf{d}_n) \, \mathrm{d}\mathbf{u}. \end{aligned}$$

- The update scheme defined by

$$\boldsymbol{\upsilon}^{(r+1)} = \arg\min_{\boldsymbol{\upsilon} \in \Upsilon} m\left(\boldsymbol{\upsilon}; \boldsymbol{\upsilon}^{(r)}\right)$$

is the EM algorithm.

# Relationship to the EM Algorithm

- Suppose that $o(\boldsymbol{\upsilon}) = -\log f_{\boldsymbol{\upsilon}}(\mathbf{d}_n)$ be a negative log-likelihood on $\mathbf{D}_n$ that is difficult to manipulate.
- Introduce some latent data $\mathbf{U}$, where we know that the joint probability density function between $\mathbf{D}_n$ and $\mathbf{U}$ has the form

$$f_{\boldsymbol{\upsilon}}(\mathbf{u}, \mathbf{d}_n). \tag{3}$$

- Using (3), we can majorize $o(\boldsymbol{\upsilon})$ at $\boldsymbol{\upsilon}^{(r)}$ by

$$\begin{aligned} m\left(\boldsymbol{\upsilon}; \boldsymbol{\upsilon}^{(r)}\right) &= -\int \log f_{\boldsymbol{\upsilon}}(\mathbf{u}, \mathbf{d}_n) f_{\boldsymbol{\upsilon}^{(r)}}(\mathbf{u}|\mathbf{d}_n) \, \mathrm{d}\mathbf{u} \\ &\quad + \int \log f_{\boldsymbol{\theta}^{(r)}}(\mathbf{u}|\mathbf{d}_n) f_{\boldsymbol{\theta}^{(r)}}(\mathbf{u}|\mathbf{d}_n) \, \mathrm{d}\mathbf{u}. \end{aligned}$$

- The update scheme defined by

$$\boldsymbol{\upsilon}^{(r+1)} = \arg\min_{\boldsymbol{\upsilon} \in \Upsilon} m\left(\boldsymbol{\upsilon}; \boldsymbol{\upsilon}^{(r)}\right)$$

is the EM algorithm.

# Relationship to the EM Algorithm

- Suppose that $o(\boldsymbol{\upsilon}) = -\log f_{\boldsymbol{\upsilon}}(\mathbf{d}_n)$ be a negative log-likelihood on $\mathbf{D}_n$ that is difficult to manipulate.
- Introduce some latent data $\mathbf{U}$, where we know that the joint probability density function between $\mathbf{D}_n$ and $\mathbf{U}$ has the form

$$f_{\boldsymbol{\upsilon}}(\mathbf{u}, \mathbf{d}_n). \tag{3}$$

- Using (3), we can majorize $o(\boldsymbol{\upsilon})$ at $\boldsymbol{\upsilon}^{(r)}$ by

$$\begin{aligned} m\left(\boldsymbol{\upsilon}; \boldsymbol{\upsilon}^{(r)}\right) &= -\int \log f_{\boldsymbol{\upsilon}}(\mathbf{u}, \mathbf{d}_n) f_{\boldsymbol{\upsilon}^{(r)}}(\mathbf{u}|\mathbf{d}_n) \, \mathrm{d}\mathbf{u} \\ &\quad + \int \log f_{\boldsymbol{\theta}^{(r)}}(\mathbf{u}|\mathbf{d}_n) f_{\boldsymbol{\theta}^{(r)}}(\mathbf{u}|\mathbf{d}_n) \, \mathrm{d}\mathbf{u}. \end{aligned}$$

- The update scheme defined by

$$\boldsymbol{\upsilon}^{(r+1)} = \arg\min_{\boldsymbol{\upsilon} \in \Upsilon} \; m\left(\boldsymbol{\upsilon}; \boldsymbol{\upsilon}^{(r)}\right)$$

is the EM algorithm.

# Relationship to the EM Algorithm

- We can prove that $m\left(\boldsymbol{\upsilon};\boldsymbol{\upsilon}^{(r)}\right)$ is a majorizer of $o\left(\boldsymbol{\upsilon}\right)$ by verifying the sequence of equalities and inequalities:

$$
\begin{aligned}
-\log_{\boldsymbol{\upsilon}} f\left(\mathbf{d}_n\right) &= -\log \int f_{\boldsymbol{\upsilon}}\left(\mathbf{d}_n|\mathbf{u}\right) f_{\boldsymbol{\upsilon}}\left(\mathbf{u}\right) \mathrm{d}\mathbf{u} \\
&= -\log \int \frac{f_{\boldsymbol{\upsilon}^{(r)}}\left(\mathbf{u}|\mathbf{d}_n\right) f_{\boldsymbol{\upsilon}}\left(\mathbf{d}_n|\mathbf{u}\right)}{f_{\boldsymbol{\upsilon}^{(r)}}\left(\mathbf{u}|\mathbf{d}_n\right)} f_{\boldsymbol{\upsilon}}\left(\mathbf{u}\right) \mathrm{d}\mathbf{u} \\
&= -\log \int \frac{f_{\boldsymbol{\upsilon}}\left(\mathbf{u}\right) f_{\boldsymbol{\upsilon}}\left(\mathbf{d}_n|\mathbf{u}\right)}{f_{\boldsymbol{\upsilon}^{(r)}}\left(\mathbf{u}|\mathbf{d}_n\right)} f_{\boldsymbol{\upsilon}^{(r)}}\left(\mathbf{u}|\mathbf{d}_n\right) \mathrm{d}\mathbf{u} \\
&\leq -\int \log \frac{f_{\boldsymbol{\upsilon}}\left(\mathbf{u}\right) f_{\boldsymbol{\upsilon}}\left(\mathbf{d}_n|\mathbf{u}\right)}{f_{\boldsymbol{\upsilon}^{(r)}}\left(\mathbf{u}|\mathbf{d}_n\right)} f_{\boldsymbol{\upsilon}^{(r)}}\left(\mathbf{u}|\mathbf{d}_n\right) \mathrm{d}\mathbf{u} \\
&= -\int \log f_{\boldsymbol{\upsilon}}\left(\mathbf{u},\mathbf{d}_n\right) f_{\boldsymbol{\upsilon}^{(r)}}\left(\mathbf{u}|\mathbf{d}_n\right) \mathrm{d}\mathbf{u} \\
&\quad + \int \log f_{\boldsymbol{\upsilon}^{(r)}}\left(\mathbf{u}|\mathbf{d}_n\right) f_{\boldsymbol{\upsilon}^{(r)}}\left(\mathbf{u}|\mathbf{d}_n\right) \mathrm{d}\mathbf{u} \\
&= m\left(\boldsymbol{\upsilon};\boldsymbol{\upsilon}^{(r)}\right).
\end{aligned}
$$

# Example Majorizers (1)

- Consider the loss from Example 1 of form:

$$l(\boldsymbol{\theta}; \boldsymbol{x}_i) = -\log \sum_{z=1}^{g} \pi_z \phi_p(\boldsymbol{x}_i; \boldsymbol{\mu}_z, \boldsymbol{\Sigma}_z),$$

where

$$\phi_p(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = |2\pi\boldsymbol{\Sigma}|^{-1/2} \exp\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right].$$

- Let $\boldsymbol{\theta}^{(r)}$ be the $r$th iterate of the MM algorithm and set

$$\upsilon_z = \pi_z \phi_p(\boldsymbol{x}_i; \boldsymbol{\mu}_z, \boldsymbol{\Sigma}_z),$$

$$\omega_z = \pi_z^{(r)} \phi_p\left(\boldsymbol{x}_i; \boldsymbol{\mu}_z^{(r)}, \boldsymbol{\Sigma}_z^{(r)}\right),$$

for each $z \in [g]$.

- Let $f(\cdot) = -\log(\cdot)$.

# Example Majorizers (1)

- Consider the loss from Example 1 of form:

$$l(\boldsymbol{\theta}; \boldsymbol{x}_i) = -\log \sum_{z=1}^{g} \pi_z \phi_p(\boldsymbol{x}_i; \boldsymbol{\mu}_z, \boldsymbol{\Sigma}_z),$$

where

$$\phi_p(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = |2\pi\boldsymbol{\Sigma}|^{-1/2} \exp\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right].$$

- Let $\boldsymbol{\theta}^{(r)}$ be the $r$th iterate of the MM algorithm and set

$$\upsilon_z = \pi_z \phi_p(\boldsymbol{x}_i; \boldsymbol{\mu}_z, \boldsymbol{\Sigma}_z),$$

$$\omega_z = \pi_z^{(r)} \phi_p\left(\boldsymbol{x}_i; \boldsymbol{\mu}_z^{(r)}, \boldsymbol{\Sigma}_z^{(r)}\right),$$

for each $z \in [g]$.

- Let $f(\cdot) = -\log(\cdot)$.

# Example Majorizers (1)

- Consider the loss from Example 1 of form:

$$l(\boldsymbol{\theta}; \boldsymbol{x}_i) = -\log \sum_{z=1}^{g} \pi_z \phi_p(\boldsymbol{x}_i; \boldsymbol{\mu}_z, \boldsymbol{\Sigma}_z),$$

  where

$$\phi_p(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = |2\pi\boldsymbol{\Sigma}|^{-1/2} \exp\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right].$$

- Let $\boldsymbol{\theta}^{(r)}$ be the $r$th iterate of the MM algorithm and set

$$\upsilon_z = \pi_z \phi_p(\boldsymbol{x}_i; \boldsymbol{\mu}_z, \boldsymbol{\Sigma}_z),$$

$$\omega_z = \pi_z^{(r)} \phi_p\left(\boldsymbol{x}_i; \boldsymbol{\mu}_z^{(r)}, \boldsymbol{\Sigma}_z^{(r)}\right),$$

  for each $z \in [g]$.

- Let $f(\cdot) = -\log(\cdot)$.

# Example Majorizers (1)

- Using the Jensen's inequality majorizer, we can majorize $l(\boldsymbol{\theta}; \boldsymbol{x}_i)$ in Example 1 at the $r$th iteration of the MM algorithm by

$$\begin{aligned} m_i\left(\boldsymbol{\theta}; \boldsymbol{\theta}^{(r)}\right) &= -\sum_{z=1}^{g} \tau_z^{(r)}(\boldsymbol{x}_i) \log \pi_z \\ &\quad -\sum_{z=1}^{g} \tau_z^{(r)}(\boldsymbol{x}_i) \log \phi_p(\boldsymbol{x}_i; \boldsymbol{\mu}_z, \boldsymbol{\Sigma}_z) + C_i^{(r)}, \end{aligned} \tag{4}$$

where

$$\tau_z^{(r)}(\boldsymbol{x}_i) = \frac{\pi_z^{(r)} \phi_p\left(\boldsymbol{x}_i; \boldsymbol{\mu}_z^{(r)}, \boldsymbol{\Sigma}_z^{(r)}\right)}{\sum_{z'=1}^{g} \pi_{z'}^{(r)} \phi_p\left(\boldsymbol{x}_i; \boldsymbol{\mu}_{z'}^{(r)}, \boldsymbol{\Sigma}_{z'}^{(r)}\right)},$$

and $C_i^{(r)}$ is a constant that does not depend on the active variable $\boldsymbol{\theta}$.

# Example Majorizers (2)

- Consider the loss from Example 2 (Logistic) of form:

$$l(\boldsymbol{\theta}; \boldsymbol{d}_i) = -y_i \boldsymbol{\theta}^\top \boldsymbol{x}_i + \log\left[1 + \exp\left(\boldsymbol{\theta}^\top \boldsymbol{x}_i\right)\right]. \tag{5}$$

- The gradient of (5) is

$$\nabla l(\boldsymbol{\theta}; \boldsymbol{d}_i) = -y_i \boldsymbol{x}_i + \pi_{\boldsymbol{\theta}}(\boldsymbol{x}_i)\, \boldsymbol{x}_i.$$

- The Hessian of (5) is

$$\mathbb{H} l(\boldsymbol{\theta}; \boldsymbol{d}_i) = \pi_{\boldsymbol{\theta}}(\boldsymbol{x}_i)\left[1 - \pi_{\boldsymbol{\theta}}(\boldsymbol{x}_i)\right] \boldsymbol{x}_i \boldsymbol{x}_i^\top.$$

# Example Majorizers (2)

- Consider the loss from Example 2 (Logistic) of form:

$$l(\boldsymbol{\theta}; \boldsymbol{d}_i) = -y_i \boldsymbol{\theta}^\top \boldsymbol{x}_i + \log\left[1 + \exp\left(\boldsymbol{\theta}^\top \boldsymbol{x}_i\right)\right]. \tag{5}$$

- The gradient of (5) is

$$\nabla l(\boldsymbol{\theta}; \boldsymbol{d}_i) = -y_i \boldsymbol{x}_i + \pi_{\boldsymbol{\theta}}(\boldsymbol{x}_i)\, \boldsymbol{x}_i.$$

- The Hessian of (5) is

$$\mathbb{H} l(\boldsymbol{\theta}; \boldsymbol{d}_i) = \pi_{\boldsymbol{\theta}}(\boldsymbol{x}_i)\left[1 - \pi_{\boldsymbol{\theta}}(\boldsymbol{x}_i)\right] \boldsymbol{x}_i \boldsymbol{x}_i^\top.$$

# Example Majorizers (2)

- Consider the loss from Example 2 (Logistic) of form:

$$l(\boldsymbol{\theta}; \boldsymbol{d}_i) = -y_i \boldsymbol{\theta}^\top \boldsymbol{x}_i + \log\left[1 + \exp\left(\boldsymbol{\theta}^\top \boldsymbol{x}_i\right)\right]. \tag{5}$$

- The gradient of (5) is

$$\nabla l(\boldsymbol{\theta}; \boldsymbol{d}_i) = -y_i \boldsymbol{x}_i + \pi_{\boldsymbol{\theta}}(\boldsymbol{x}_i)\, \boldsymbol{x}_i.$$

- The Hessian of (5) is

$$\mathbb{H} l(\boldsymbol{\theta}; \boldsymbol{d}_i) = \pi_{\boldsymbol{\theta}}(\boldsymbol{x}_i)\left[1 - \pi_{\boldsymbol{\theta}}(\boldsymbol{x}_i)\right] \boldsymbol{x}_i \boldsymbol{x}_i^\top.$$

# Example Majorizers (2)

- Let $\varphi = \pi_{\boldsymbol{\theta}}(\boldsymbol{x}_i)$ and consider that the function

$$g(\varphi) = \varphi(1-\varphi)$$

is maximized at $\varphi = 1/4$.

- Set $\mathbf{H}_i = \boldsymbol{x}_i \boldsymbol{x}_i^\top / 4$ and note that $\mathbf{H}_i - \mathbb{H} l(\boldsymbol{\theta}; \boldsymbol{d}_i)$ is positive definite for each $i \in [n]$.

- At the $r$th iteration of the MM algorithm, we have the following majorizer

$$\begin{aligned} m_i\left(\boldsymbol{\theta}; \boldsymbol{\theta}^{(r)}\right) &= l\left(\boldsymbol{\theta}^{(r)}; \boldsymbol{d}_i\right) + \left(\boldsymbol{\theta} - \boldsymbol{\theta}^{(r)}\right)^\top \nabla l\left(\boldsymbol{\theta}^{(r)}; \boldsymbol{d}_i\right) \\ &\quad + \frac{1}{2}\left(\boldsymbol{\theta} - \boldsymbol{\theta}^{(r)}\right)^\top \mathbf{H}_i \left(\boldsymbol{\theta} - \boldsymbol{\theta}^{(r)}\right), \end{aligned}$$

for (5) via a quadratic upper bound.

# Example Majorizers (2)

- Let $\varphi = \pi_{\boldsymbol{\theta}}(\boldsymbol{x}_i)$ and consider that the function

$$g(\varphi) = \varphi(1-\varphi)$$

  is maximized at $\varphi = 1/4$.

- Set $\mathbf{H}_i = \boldsymbol{x}_i \boldsymbol{x}_i^\top / 4$ and note that $\mathbf{H}_i - \mathbb{H} l(\boldsymbol{\theta}; \boldsymbol{d}_i)$ is positive definite for each $i \in [n]$.

- At the $r$th iteration of the MM algorithm, we have the following majorizer

$$\begin{aligned} m_i\left(\boldsymbol{\theta}; \boldsymbol{\theta}^{(r)}\right) &= l\left(\boldsymbol{\theta}^{(r)}; \boldsymbol{d}_i\right) + \left(\boldsymbol{\theta} - \boldsymbol{\theta}^{(r)}\right)^\top \nabla l\left(\boldsymbol{\theta}^{(r)}; \boldsymbol{d}_i\right) \\ &\quad + \frac{1}{2}\left(\boldsymbol{\theta} - \boldsymbol{\theta}^{(r)}\right)^\top \mathbf{H}_i \left(\boldsymbol{\theta} - \boldsymbol{\theta}^{(r)}\right), \end{aligned}$$

  for (5) via a quadratic upper bound.

# Example Majorizers (2)

- Let $\varphi = \pi_{\boldsymbol{\theta}}(\boldsymbol{x}_i)$ and consider that the function

$$g(\varphi) = \varphi(1-\varphi)$$

is maximized at $\varphi = 1/4$.

- Set $\mathbf{H}_i = \boldsymbol{x}_i \boldsymbol{x}_i^\top / 4$ and note that $\mathbf{H}_i - \mathbb{H} l(\boldsymbol{\theta}; \boldsymbol{d}_i)$ is positive definite for each $i \in [n]$.

- At the $r$th iteration of the MM algorithm, we have the following majorizer

$$\begin{aligned} m_i\left(\boldsymbol{\theta}; \boldsymbol{\theta}^{(r)}\right) &= l\left(\boldsymbol{\theta}^{(r)}; \boldsymbol{d}_i\right) + \left(\boldsymbol{\theta} - \boldsymbol{\theta}^{(r)}\right)^\top \nabla l\left(\boldsymbol{\theta}^{(r)}; \boldsymbol{d}_i\right) \\ &\quad + \frac{1}{2}\left(\boldsymbol{\theta} - \boldsymbol{\theta}^{(r)}\right)^\top \mathbf{H}_i \left(\boldsymbol{\theta} - \boldsymbol{\theta}^{(r)}\right), \end{aligned}$$

for (5) via a quadratic upper bound.

# Example Majorizers (3)

- Consider the loss from Example 3 (SVM) of form:

$$l(\boldsymbol{\theta}; \boldsymbol{d}_i) \quad = \quad [1 - y_i f_{\boldsymbol{\theta}}(\boldsymbol{x}_i)]_+ ,$$

where $[\cdot]_+ = \max\{0, \cdot\}$.

- For any $x, y \in \mathbb{R}$, note that

$$\max\{x, y\} = \frac{1}{2}|x - y| + \frac{1}{2}x + \frac{1}{2}y. \tag{6}$$

- Using (6), we can write

$$[x]_+ = \frac{1}{2}|x| + \frac{1}{2}x$$

for any $x \in \mathbb{R}$.

# Example Majorizers (3)

- Consider the loss from Example 3 (SVM) of form:

$$l(\boldsymbol{\theta}; \boldsymbol{d}_i) \quad = \quad [1 - y_i f_{\boldsymbol{\theta}}(\boldsymbol{x}_i)]_+ ,$$

where $[\cdot]_+ = \max\{0, \cdot\}$.

- For any $x, y \in \mathbb{R}$, note that

$$\max\{x, y\} = \frac{1}{2}|x - y| + \frac{1}{2}x + \frac{1}{2}y. \tag{6}$$

- Using (6), we can write

$$[x]_+ = \frac{1}{2}|x| + \frac{1}{2}x$$

for any $x \in \mathbb{R}$.

# Example Majorizers (3)

- Consider the loss from Example 3 (SVM) of form:

$$l(\boldsymbol{\theta}; \boldsymbol{d}_i) \quad = \quad [1 - y_i f_{\boldsymbol{\theta}}(\boldsymbol{x}_i)]_+ ,$$

where $[\cdot]_+ = \max\{0, \cdot\}$.

- For any $x, y \in \mathbb{R}$, note that

$$\max\{x, y\} = \frac{1}{2}|x - y| + \frac{1}{2}x + \frac{1}{2}y. \tag{6}$$

- Using (6), we can write

$$[x]_+ = \frac{1}{2}|x| + \frac{1}{2}x$$

for any $x \in \mathbb{R}$.

# Example Majorizers (3)

- Consider that we can write

$$|x| = \sqrt{x^2}$$

where $\sqrt{\cdot}$ is a concave function.

- Let $f(v) = \sqrt{v}$, with $\nabla f(v) = 1/\left(2\sqrt{v}\right)$.
- By the supporting hyperplane inequality, we can majorize

$$o(v) = f(v)$$

by

$$\begin{aligned} m(v;\omega) &= f(\omega) + (v-\omega)\nabla f(\omega) \\ &= \sqrt{\omega} + \frac{(v-\omega)}{2\sqrt{\omega}}. \end{aligned}$$

# Example Majorizers (3)

- Consider that we can write

$$|x| = \sqrt{x^2}$$

where $\sqrt{\cdot}$ is a concave function.

- Let $f(v) = \sqrt{v}$, with $\nabla f(v) = 1/\left(2\sqrt{v}\right)$.
- By the supporting hyperplane inequality, we can majorize

$$o(v) = f(v)$$

by

$$\begin{aligned} m(v;\omega) &= f(\omega) + (v-\omega)\nabla f(\omega) \\ &= \sqrt{\omega} + \frac{(v-\omega)}{2\sqrt{\omega}}. \end{aligned}$$

# Example Majorizers (3)

- Consider that we can write

$$|x| = \sqrt{x^2}$$

where $\sqrt{\cdot}$ is a concave function.

- Let $f(v) = \sqrt{v}$, with $\nabla f(v) = 1/\left(2\sqrt{v}\right)$.
- By the supporting hyperplane inequality, we can majorize

$$o(v) = f(v)$$

by

$$\begin{aligned} m(v;\omega) &= f(\omega) + (v-\omega)\nabla f(\omega) \\ &= \sqrt{\omega} + \frac{(v-\omega)}{2\sqrt{\omega}}. \end{aligned}$$

# Example Majorizers (3)

- We can therefore majorize

$$o(v) = [v]_+ = \frac{1}{2}\sqrt{v^2} + \frac{1}{2}v$$

by

$$\begin{aligned} m(v;\omega) &= \frac{\sqrt{\omega^2}}{2} + \frac{(v^2-\omega^2)}{4\sqrt{\omega^2}} + \frac{v^2}{2} \\ &= \frac{1}{4|\omega|}(v+|\omega|)^2. \end{aligned}$$

- Substituting in $v = 1 - y_i f_{\boldsymbol{\theta}}(\boldsymbol{x}_i)$ and $\omega = 1 - y_i f_{\boldsymbol{\theta}^{(r)}}(\boldsymbol{x}_i)$ yields the majorizer for $l(\boldsymbol{\theta}; \boldsymbol{d}_i)$:

$$m_i\left(\boldsymbol{\theta}; \boldsymbol{\theta}^{(r)}\right) = \frac{1}{4w_i^{(r)}}\left(1 - y_i f_{\boldsymbol{\theta}}(\boldsymbol{x}_i) + w_i^{(r)}\right)^2,$$

where $w_i^{(r)} = \left|1 - y_i f_{\boldsymbol{\theta}^{(r)}}(\boldsymbol{x}_i)\right|$.

# Example Majorizers (3)

- We can therefore majorize

$$o(v) = [v]_+ = \frac{1}{2}\sqrt{v^2} + \frac{1}{2}v$$

by

$$\begin{aligned} m(v;\omega) &= \frac{\sqrt{\omega^2}}{2} + \frac{\left(v^2 - \omega^2\right)}{4\sqrt{\omega^2}} + \frac{v^2}{2} \\ &= \frac{1}{4|\omega|}\left(v + |\omega|\right)^2. \end{aligned}$$

- Substituting in $v = 1 - y_i f_{\boldsymbol{\theta}}(\boldsymbol{x}_i)$ and $\omega = 1 - y_i f_{\boldsymbol{\theta}^{(r)}}(\boldsymbol{x}_i)$ yields the majorizer for $l(\boldsymbol{\theta}; \boldsymbol{d}_i)$:

$$m_i\left(\boldsymbol{\theta}; \boldsymbol{\theta}^{(r)}\right) = \frac{1}{4w_i^{(r)}}\left(1 - y_i f_{\boldsymbol{\theta}}(\boldsymbol{x}_i) + w_i^{(r)}\right)^2,$$

where $w_i^{(r)} = \left|1 - y_i f_{\boldsymbol{\theta}^{(r)}}(\boldsymbol{x}_i)\right|$.

# Example MM Algorithms (1)

- In Example 1 (GMM), we can now majorize $L(\boldsymbol{\theta}; \mathbf{d}_n)$ by

$$\begin{aligned} M\left(\boldsymbol{\theta}; \boldsymbol{\theta}^{(r)}\right) &= \frac{1}{n}\sum_{i=1}^{n} m_i\left(\boldsymbol{\theta}; \boldsymbol{\theta}^{(r)}\right) \\ &= -\frac{1}{n}\sum_{i=1}^{n}\sum_{z=1}^{g} \tau_z^{(r)}(\boldsymbol{x}_i)\log \pi_z \\ &\quad -\frac{1}{n}\sum_{i=1}^{n}\sum_{z=1}^{g} \tau_z^{(r)}(\boldsymbol{x}_i)\log \phi_p(\boldsymbol{x}_i; \boldsymbol{\mu}_z, \boldsymbol{\Sigma}_z) + C^{(r)}, \end{aligned}$$

  where $C^{(r)}$ does not depend on $\boldsymbol{\theta}$.

- We wish to obtain

$$\boldsymbol{\theta}^{(r+1)} = \arg\min_{\boldsymbol{\theta}\in\Theta} M\left(\boldsymbol{\theta}; \boldsymbol{\theta}^{(r)}\right),$$

  where $\Theta$ includes the restriction $\pi_z > 0$ and $\sum_{z=1}^{g} \pi_z = 1$.

# Example MM Algorithms (1)

- In Example 1 (GMM), we can now majorize $L(\boldsymbol{\theta}; \mathbf{d}_n)$ by

$$\begin{aligned} M\left(\boldsymbol{\theta}; \boldsymbol{\theta}^{(r)}\right) &= \frac{1}{n}\sum_{i=1}^{n} m_i\left(\boldsymbol{\theta}; \boldsymbol{\theta}^{(r)}\right) \\ &= -\frac{1}{n}\sum_{i=1}^{n}\sum_{z=1}^{g} \tau_z^{(r)}(\boldsymbol{x}_i)\log \pi_z \\ &\quad -\frac{1}{n}\sum_{i=1}^{n}\sum_{z=1}^{g} \tau_z^{(r)}(\boldsymbol{x}_i)\log \phi_p(\boldsymbol{x}_i; \boldsymbol{\mu}_z, \boldsymbol{\Sigma}_z) + C^{(r)}, \end{aligned}$$

  where $C^{(r)}$ does not depend on $\boldsymbol{\theta}$.

- We wish to obtain

$$\boldsymbol{\theta}^{(r+1)} = \arg\min_{\boldsymbol{\theta}\in\Theta} M\left(\boldsymbol{\theta}; \boldsymbol{\theta}^{(r)}\right),$$

  where $\Theta$ includes the restriction $\pi_z > 0$ and $\sum_{z=1}^{g} \pi_z = 1$.

# Example MM Algorithms (1)

- Expansion of $\phi_p$ yields the expression

$$\begin{aligned} M\left(\boldsymbol{\theta};\boldsymbol{\theta}^{(r)}\right) &= -\frac{1}{n}\sum_{i=1}^{n}\sum_{z=1}^{g}\tau_z^{(r)}(\boldsymbol{x}_i)\log\pi_z \\ &= +\frac{1}{2n}\sum_{i=1}^{n}\sum_{z=1}^{g}\tau_z^{(r)}(\boldsymbol{x}_i)\log|\boldsymbol{\Sigma}_z| \\ &\quad +\frac{1}{2n}\sum_{i=1}^{n}\sum_{z=1}^{g}\tau_z^{(r)}(\boldsymbol{x}_i)(\boldsymbol{x}_i-\boldsymbol{\mu}_z)^\top\boldsymbol{\Sigma}_z^{-1}(\boldsymbol{x}_i-\boldsymbol{\mu}_z). \end{aligned}$$

- To minimize $M\left(\cdot;\boldsymbol{\theta}^{(r)}\right)$ under the restriction $\sum_{z=1}^{g}\pi_z = 1$, we must consider the Lagrangian

$$\Lambda\left(\boldsymbol{\theta};\boldsymbol{\theta}^{(r)}\right) = M\left(\boldsymbol{\theta};\boldsymbol{\theta}^{(r)}\right) + \lambda\left(1-\sum_{z=1}^{g}\pi_z\right).$$

# Example MM Algorithms (1)

- Note that $\Lambda\left(\cdot;\boldsymbol{\theta}^{(r)}\right)$ is linearly separable with respect to the $g$ mixture components.
- Partial derivatives of $\Lambda\left(\cdot;\boldsymbol{\theta}^{(r)}\right)$ are given by

$$\nabla_{\pi_z}\Lambda\left(\boldsymbol{\theta};\boldsymbol{\theta}^{(r)}\right) = -\frac{1}{n}\frac{1}{\pi_z}\sum_{i=1}^{n}\tau_z^{(r)}(\boldsymbol{x}_i),$$

$$\nabla_{\boldsymbol{\mu}_z}\Lambda\left(\boldsymbol{\theta};\boldsymbol{\theta}^{(r)}\right) = -\frac{1}{n}\sum_{i=1}^{n}\tau_z^{(r)}(\boldsymbol{x}_i)\boldsymbol{\Sigma}_z^{-1}(\boldsymbol{x}_i-\boldsymbol{\mu}_z),$$

$$\begin{aligned}\nabla_{\boldsymbol{\Sigma}_z}\Lambda\left(\boldsymbol{\theta};\boldsymbol{\theta}^{(r)}\right) &= \frac{1}{n}\sum_{i=1}^{n}\tau_z^{(r)}(\boldsymbol{x}_i)\boldsymbol{\Sigma}_z^{-1} \\ &\quad -\frac{1}{n}\sum_{i=1}^{n}\tau_z^{(r)}(\boldsymbol{x}_i)\boldsymbol{\Sigma}_z^{-1}(\boldsymbol{x}_i-\boldsymbol{\mu}_z)(\boldsymbol{x}_i-\boldsymbol{\mu}_z)^{\top}\boldsymbol{\Sigma}_z^{-1},\end{aligned}$$

and

$$\nabla_{\lambda}\Lambda\left(\boldsymbol{\theta};\boldsymbol{\theta}^{(r)}\right) = \sum_{z=1}^{g}\pi_i - 1.$$

# Example MM Algorithms (1)

- Note that $\Lambda\left(\cdot;\boldsymbol{\theta}^{(r)}\right)$ is linearly separable with respect to the $g$ mixture components.
- Partial derivatives of $\Lambda\left(\cdot;\boldsymbol{\theta}^{(r)}\right)$ are given by

$$\nabla_{\pi_z}\Lambda\left(\boldsymbol{\theta};\boldsymbol{\theta}^{(r)}\right) = -\frac{1}{n}\frac{1}{\pi_z}\sum_{i=1}^{n}\tau_z^{(r)}(\boldsymbol{x}_i),$$

$$\nabla_{\boldsymbol{\mu}_z}\Lambda\left(\boldsymbol{\theta};\boldsymbol{\theta}^{(r)}\right) = -\frac{1}{n}\sum_{i=1}^{n}\tau_z^{(r)}(\boldsymbol{x}_i)\boldsymbol{\Sigma}_z^{-1}(\boldsymbol{x}_i-\boldsymbol{\mu}_z),$$

$$\begin{aligned}\nabla_{\boldsymbol{\Sigma}_z}\Lambda\left(\boldsymbol{\theta};\boldsymbol{\theta}^{(r)}\right) &= \frac{1}{n}\sum_{i=1}^{n}\tau_z^{(r)}(\boldsymbol{x}_i)\boldsymbol{\Sigma}_z^{-1} \\ &\quad -\frac{1}{n}\sum_{i=1}^{n}\tau_z^{(r)}(\boldsymbol{x}_i)\boldsymbol{\Sigma}_z^{-1}(\boldsymbol{x}_i-\boldsymbol{\mu}_z)(\boldsymbol{x}_i-\boldsymbol{\mu}_z)^{\top}\boldsymbol{\Sigma}_z^{-1},\end{aligned}$$

and

$$\nabla_{\lambda}\Lambda\left(\boldsymbol{\theta};\boldsymbol{\theta}^{(r)}\right) = \sum_{z=1}^{g}\pi_i - 1.$$

# Example MM Algorithms (1)

- Solving for the Lagrangian FOC

$$\nabla \Lambda \left( \boldsymbol{\theta}; \boldsymbol{\theta}^{(r)} \right) = \mathbf{0}$$

yields the MM algorithm update $\boldsymbol{\theta}^{(r+1)}$, which contains the solutions

$$\pi_z^{(r+1)} = n^{-1} \sum_{i=1}^{n} \tau_z^{(r)}(\boldsymbol{x}_i),$$

$$\boldsymbol{\mu}_z^{(r+1)} = \frac{\sum_{i=1}^{n} \tau_z^{(r)}(\boldsymbol{x}_i) \boldsymbol{x}_i}{\sum_{i=1}^{n} \tau_z^{(r)}(\boldsymbol{x}_i)},$$

and

$$\boldsymbol{\Sigma}_z^{(r+1)} = \frac{\sum_{i=1}^{n} \tau_z^{(r)}(\boldsymbol{x}_i) \left( \boldsymbol{x}_i - \boldsymbol{\mu}_z^{(r+1)} \right) \left( \boldsymbol{x}_i - \boldsymbol{\mu}_z^{(r+1)} \right)^\top}{\sum_{i=1}^{n} \tau_z^{(r)}(\boldsymbol{x}_i)}.$$

- These are exactly the usual EM (expectation–maximization) algorithm updates for Gaussian mixture models.

# Example MM Algorithms (1)

- Solving for the Lagrangian FOC

$$\nabla \Lambda \left( \boldsymbol{\theta}; \boldsymbol{\theta}^{(r)} \right) = \mathbf{0}$$

yields the MM algorithm update $\boldsymbol{\theta}^{(r+1)}$, which contains the solutions

$$\pi_z^{(r+1)} = n^{-1} \sum_{i=1}^{n} \tau_z^{(r)}(\boldsymbol{x}_i),$$

$$\boldsymbol{\mu}_z^{(r+1)} = \frac{\sum_{i=1}^{n} \tau_z^{(r)}(\boldsymbol{x}_i) \boldsymbol{x}_i}{\sum_{i=1}^{n} \tau_z^{(r)}(\boldsymbol{x}_i)},$$

and

$$\boldsymbol{\Sigma}_z^{(r+1)} = \frac{\sum_{i=1}^{n} \tau_z^{(r)}(\boldsymbol{x}_i) \left( \boldsymbol{x}_i - \boldsymbol{\mu}_z^{(r+1)} \right) \left( \boldsymbol{x}_i - \boldsymbol{\mu}_z^{(r+1)} \right)^{\top}}{\sum_{i=1}^{n} \tau_z^{(r)}(\boldsymbol{x}_i)}.$$

- These are exactly the usual EM (expectation–maximization) algorithm updates for Gaussian mixture models.

# Example MM Algorithms (2)

- In Example 2 (Logistic), we can now majorize $L(\boldsymbol{\theta}; \mathbf{d}_n)$ by

$$M\left(\boldsymbol{\theta}; \boldsymbol{\theta}^{(r)}\right) = \frac{1}{n}\sum_{i=1}^{n}\left(\boldsymbol{\theta} - \boldsymbol{\theta}^{(r)}\right)^{\top}\nabla l\left(\boldsymbol{\theta}^{(r)}; \boldsymbol{d}_i\right) + \frac{1}{2}\sum_{i=1}^{n}\left(\boldsymbol{\theta} - \boldsymbol{\theta}^{(r)}\right)^{\top}\mathbf{H}_i\left(\boldsymbol{\theta} - \boldsymbol{\theta}^{(r)}\right) + C^{(r)},$$

  where $C^{(r)}$ does not depend on $\boldsymbol{\theta}$.

- We wish to obtain

$$\boldsymbol{\theta}^{(r+1)} = \arg\min_{\boldsymbol{\theta}\in\Theta} M\left(\boldsymbol{\theta}; \boldsymbol{\theta}^{(r)}\right).$$

# Example MM Algorithms (2)

- In Example 2 (Logistic), we can now majorize $L(\boldsymbol{\theta}; \mathbf{d}_n)$ by

$$\begin{aligned} M\left(\boldsymbol{\theta}; \boldsymbol{\theta}^{(r)}\right) &= \frac{1}{n}\sum_{i=1}^{n}\left(\boldsymbol{\theta}-\boldsymbol{\theta}^{(r)}\right)^{\top}\nabla l\left(\boldsymbol{\theta}^{(r)}; \boldsymbol{d}_i\right) \\ &\quad + \frac{1}{2}\sum_{i=1}^{n}\left(\boldsymbol{\theta}-\boldsymbol{\theta}^{(r)}\right)^{\top}\mathbf{H}_i\left(\boldsymbol{\theta}-\boldsymbol{\theta}^{(r)}\right) + C^{(r)}, \end{aligned}$$

  where $C^{(r)}$ does not depend on $\boldsymbol{\theta}$.

- We wish to obtain

$$\boldsymbol{\theta}^{(r+1)} = \arg\min_{\boldsymbol{\theta}\in\Theta} M\left(\boldsymbol{\theta}; \boldsymbol{\theta}^{(r)}\right).$$

# Example MM Algorithms (2)

- The gradient of $M\left(\cdot;\boldsymbol{\theta}^{(r)}\right)$ can be computed as

$$\nabla M\left(\boldsymbol{\theta};\boldsymbol{\theta}^{(r)}\right)=\frac{1}{n}\sum_{i=1}^{n}\nabla l\left(\boldsymbol{\theta}^{(r)};\boldsymbol{d}_i\right)+\frac{1}{n}\sum_{i=1}^{n}\mathbf{H}_i\left(\boldsymbol{\theta}-\boldsymbol{\theta}^{(r)}\right).$$

- Solving the FOC, we obtain the MM update scheme:

$$\boldsymbol{\theta}^{(r+1)}=\boldsymbol{\theta}^{(r)}-\left[\sum_{i=1}^{n}\mathbf{H}_i\right]^{-1}\sum_{i=1}^{n}\left[\pi_{\boldsymbol{\theta}}\left(\boldsymbol{x}_i\right)-y_i\right]\boldsymbol{x}_i.$$

- Note that the MM update scheme is almost Newton-Raphson, except the Hessian is replaced by

$$\sum_{i=1}^{n}\mathbf{H}_i=\frac{1}{4}\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i^{\top}$$

which does not depend on $r$.

# Example MM Algorithms (2)

- The gradient of $M\left(\cdot;\boldsymbol{\theta}^{(r)}\right)$ can be computed as

$$\nabla M\left(\boldsymbol{\theta};\boldsymbol{\theta}^{(r)}\right)=\frac{1}{n}\sum_{i=1}^{n}\nabla l\left(\boldsymbol{\theta}^{(r)};\boldsymbol{d}_i\right)+\frac{1}{n}\sum_{i=1}^{n}\mathbf{H}_i\left(\boldsymbol{\theta}-\boldsymbol{\theta}^{(r)}\right).$$

- Solving the FOC, we obtain the MM update scheme:

$$\boldsymbol{\theta}^{(r+1)}=\boldsymbol{\theta}^{(r)}-\left[\sum_{i=1}^{n}\mathbf{H}_i\right]^{-1}\sum_{i=1}^{n}\left[\pi_{\boldsymbol{\theta}}\left(\boldsymbol{x}_i\right)-y_i\right]\boldsymbol{x}_i.$$

- Note that the MM update scheme is almost Newton-Raphson, except the Hessian is replaced by

$$\sum_{i=1}^{n}\mathbf{H}_i=\frac{1}{4}\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i^{\top}$$

which does not depend on $r$.

# Example MM Algorithms (2)

- The gradient of $M\left(\cdot;\boldsymbol{\theta}^{(r)}\right)$ can be computed as

$$\nabla M\left(\boldsymbol{\theta};\boldsymbol{\theta}^{(r)}\right)=\frac{1}{n}\sum_{i=1}^{n}\nabla l\left(\boldsymbol{\theta}^{(r)};\boldsymbol{d}_i\right)+\frac{1}{n}\sum_{i=1}^{n}\mathbf{H}_i\left(\boldsymbol{\theta}-\boldsymbol{\theta}^{(r)}\right).$$

- Solving the FOC, we obtain the MM update scheme:

$$\boldsymbol{\theta}^{(r+1)}=\boldsymbol{\theta}^{(r)}-\left[\sum_{i=1}^{n}\mathbf{H}_i\right]^{-1}\sum_{i=1}^{n}\left[\pi_{\boldsymbol{\theta}}\left(\boldsymbol{x}_i\right)-y_i\right]\boldsymbol{x}_i.$$

- Note that the MM update scheme is almost Newton-Raphson, except the Hessian is replaced by

$$\sum_{i=1}^{n}\mathbf{H}_i=\frac{1}{4}\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i^{\top}$$

which does not depend on $r$.

# Example MM Algorithms (3)

- In Example 3 (SVM), we can now majorize

$$\begin{aligned} O(\boldsymbol{\theta}) &= \frac{1}{n}\sum_{i=1}^{n}\left[1-y_i f_{\boldsymbol{\theta}}(\boldsymbol{x}_i)\right]_+ + \lambda\boldsymbol{\beta}^\top\boldsymbol{\beta} \\ &= \frac{1}{n}\sum_{i=1}^{n}\left[1-y_i\left(\alpha+\boldsymbol{\beta}^\top\boldsymbol{x}_i\right)\right]_+ + \lambda\boldsymbol{\beta}^\top\boldsymbol{\beta} \end{aligned}$$

by

$$\begin{aligned} M\left(\boldsymbol{\theta};\boldsymbol{\theta}^{(r)}\right) &= \sum_{i=1}^{n} m_i\left(\boldsymbol{\theta};\boldsymbol{\theta}^{(r)}\right) \\ &= \frac{1}{4n}\sum_{i=1}^{n}\frac{1}{w_i^{(r)}}\left[1-y_i\left(\alpha+\boldsymbol{\beta}^\top\boldsymbol{x}_i\right)+w_i^{(r)}\right]^2 + \lambda\boldsymbol{\beta}^\top\boldsymbol{\beta}. \end{aligned}$$

- We wish to obtain

$$\boldsymbol{\theta}^{(r+1)} = \arg\min_{\boldsymbol{\theta}\in\Theta} M\left(\boldsymbol{\theta};\boldsymbol{\theta}^{(r)}\right).$$

# Example MM Algorithms (3)

- In Example 3 (SVM), we can now majorize

$$\begin{aligned} O(\boldsymbol{\theta}) &= \frac{1}{n}\sum_{i=1}^{n}\left[1-y_i f_{\boldsymbol{\theta}}(\boldsymbol{x}_i)\right]_+ + \lambda\boldsymbol{\beta}^\top\boldsymbol{\beta} \\ &= \frac{1}{n}\sum_{i=1}^{n}\left[1-y_i\left(\alpha+\boldsymbol{\beta}^\top\boldsymbol{x}_i\right)\right]_+ + \lambda\boldsymbol{\beta}^\top\boldsymbol{\beta} \end{aligned}$$

by

$$\begin{aligned} M\left(\boldsymbol{\theta};\boldsymbol{\theta}^{(r)}\right) &= \sum_{i=1}^{n} m_i\left(\boldsymbol{\theta};\boldsymbol{\theta}^{(r)}\right) \\ &= \frac{1}{4n}\sum_{i=1}^{n}\frac{1}{w_i^{(r)}}\left[1-y_i\left(\alpha+\boldsymbol{\beta}^\top\boldsymbol{x}_i\right)+w_i^{(r)}\right]^2 + \lambda\boldsymbol{\beta}^\top\boldsymbol{\beta}. \end{aligned}$$

- We wish to obtain

$$\boldsymbol{\theta}^{(r+1)} = \arg\min_{\boldsymbol{\theta}\in\Theta} M\left(\boldsymbol{\theta};\boldsymbol{\theta}^{(r)}\right).$$

# Example MM Algorithms (3)

- Write

$$\boldsymbol{z}_i^\top = \left(y_i, y_i \boldsymbol{x}_i^\top\right),$$

$$v_i^{(r)} = w_i^{(r)} + 1,$$

$$\bar{\mathbf{I}} = \begin{bmatrix} 0 & 0 \\ 0 & \mathbf{I}_n \end{bmatrix},$$

and

$$\mathbf{W}^{(r)} = \frac{1}{4n} \begin{bmatrix} 1/w_1^{(r)} & 0 & \cdots & 0 \\ 0 & 1/w_2^{(r)} & & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & 1/w_n^{(r)} \end{bmatrix},$$

where $\mathbf{I}_n$ is the identity matrix in $\mathbb{R}^{n \times n}$.

# Example MM Algorithms (3)

- Let $\boldsymbol{v}^{(r)\top} = \left(v_1^{(r)}, \ldots, v_n^{(r)}\right)$.
- Put $\boldsymbol{z}_i$ into the $i$th row of a new matrix $\mathbf{Z} \in \mathbb{R}^{n \times (p+1)}$.
- Recalling that $\boldsymbol{\theta}^\top = \left(\alpha, \boldsymbol{\beta}^\top\right)$ and using the previous notation, we can now write

$$M\left(\boldsymbol{\theta}; \boldsymbol{\theta}^{(r)}\right) = \left(\boldsymbol{v}^{(r)} - \mathbf{Z}\boldsymbol{\theta}\right)^\top \mathbf{W}^{(r)} \left(\boldsymbol{v}^{(r)} - \mathbf{Z}\boldsymbol{\theta}\right) + \lambda \boldsymbol{\theta}^\top \bar{\mathbf{I}} \boldsymbol{\theta}.$$

# Example MM Algorithms (3)

- Let $\boldsymbol{v}^{(r)\top} = \left(v_1^{(r)}, \ldots, v_n^{(r)}\right)$.
- Put $\boldsymbol{z}_i$ into the $i$th row of a new matrix $\mathbf{Z} \in \mathbb{R}^{n \times (p+1)}$.
- Recalling that $\boldsymbol{\theta}^\top = \left(\alpha, \boldsymbol{\beta}^\top\right)$ and using the previous notation, we can now write

$$M\left(\boldsymbol{\theta}; \boldsymbol{\theta}^{(r)}\right) = \left(\boldsymbol{v}^{(r)} - \mathbf{Z}\boldsymbol{\theta}\right)^\top \mathbf{W}^{(r)} \left(\boldsymbol{v}^{(r)} - \mathbf{Z}\boldsymbol{\theta}\right) + \lambda \boldsymbol{\theta}^\top \bar{\mathbf{I}} \boldsymbol{\theta}.$$

# Example MM Algorithms (3)

- Let $\boldsymbol{v}^{(r)\top} = \left(v_1^{(r)}, \ldots, v_n^{(r)}\right)$.
- Put $\boldsymbol{z}_i$ into the $i$th row of a new matrix $\mathbf{Z} \in \mathbb{R}^{n \times (p+1)}$.
- Recalling that $\boldsymbol{\theta}^\top = \left(\alpha, \boldsymbol{\beta}^\top\right)$ and using the previous notation, we can now write

$$M\left(\boldsymbol{\theta}; \boldsymbol{\theta}^{(r)}\right) = \left(\boldsymbol{v}^{(r)} - \mathbf{Z}\boldsymbol{\theta}\right)^\top \mathbf{W}^{(r)} \left(\boldsymbol{v}^{(r)} - \mathbf{Z}\boldsymbol{\theta}\right) + \lambda \boldsymbol{\theta}^\top \bar{\mathbf{I}} \boldsymbol{\theta}.$$

# Example MM Algorithms (3)

- The gradient of $M\left(\cdot;\boldsymbol{\theta}^{(r)}\right)$ can be given as

$$\begin{aligned}\nabla M\left(\boldsymbol{\theta};\boldsymbol{\theta}^{(r)}\right) &= -2\mathbf{Z}^{\top}\mathbf{W}^{(r)}\left(\boldsymbol{v}^{(r)}-\mathbf{Z}\boldsymbol{\theta}\right)\\ &\quad +2\lambda\bar{\mathbf{I}}\boldsymbol{\theta}\end{aligned}$$

- The FOC solution to obtain MM algorithm update $\boldsymbol{\theta}^{(r+1)}$ is the usual least squares solution

$$\boldsymbol{\theta}^{(r+1)}=\left(\mathbf{Z}^{\top}\mathbf{W}^{(r)}\mathbf{Z}+\lambda\bar{\mathbf{I}}\right)^{-1}\mathbf{Z}^{\top}\mathbf{W}^{(r)}\boldsymbol{v}^{(r)}.$$

- The MM algorithm for linear SVMs is therefore an iteratively reweighted least squares algorithm.

# Example MM Algorithms (3)

- The gradient of $M\left(\cdot;\boldsymbol{\theta}^{(r)}\right)$ can be given as

$$\begin{aligned}\nabla M\left(\boldsymbol{\theta};\boldsymbol{\theta}^{(r)}\right) &= -2\mathbf{Z}^{\top}\mathbf{W}^{(r)}\left(\boldsymbol{v}^{(r)}-\mathbf{Z}\boldsymbol{\theta}\right)\\ &\quad +2\lambda\bar{\mathbf{I}}\boldsymbol{\theta}\end{aligned}$$

- The FOC solution to obtain MM algorithm update $\boldsymbol{\theta}^{(r+1)}$ is the usual least squares solution

$$\boldsymbol{\theta}^{(r+1)}=\left(\mathbf{Z}^{\top}\mathbf{W}^{(r)}\mathbf{Z}+\lambda\bar{\mathbf{I}}\right)^{-1}\mathbf{Z}^{\top}\mathbf{W}^{(r)}\boldsymbol{v}^{(r)}.$$

- The MM algorithm for linear SVMs is therefore an iteratively reweighted least squares algorithm.

# Example MM Algorithms (3)

- The gradient of $M\left(\cdot;\boldsymbol{\theta}^{(r)}\right)$ can be given as

$$\begin{aligned}\nabla M\left(\boldsymbol{\theta};\boldsymbol{\theta}^{(r)}\right) &= -2\mathbf{Z}^{\top}\mathbf{W}^{(r)}\left(\boldsymbol{v}^{(r)}-\mathbf{Z}\boldsymbol{\theta}\right)\\ &\quad +2\lambda\bar{\mathbf{I}}\boldsymbol{\theta}\end{aligned}$$

- The FOC solution to obtain MM algorithm update $\boldsymbol{\theta}^{(r+1)}$ is the usual least squares solution

$$\boldsymbol{\theta}^{(r+1)}=\left(\mathbf{Z}^{\top}\mathbf{W}^{(r)}\mathbf{Z}+\lambda\bar{\mathbf{I}}\right)^{-1}\mathbf{Z}^{\top}\mathbf{W}^{(r)}\boldsymbol{v}^{(r)}.$$

- The MM algorithm for linear SVMs is therefore an iteratively reweighted least squares algorithm.

# Example MM Algorithms (3)

- Notice that

$$\mathbf{W}^{(r)} = \frac{1}{4n} \begin{bmatrix} 1/w_1^{(r)} & 0 & \cdots & 0 \\ 0 & 1/w_2^{(r)} & & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & 1/w_n^{(r)} \end{bmatrix},$$

contains entries that may be infinite, since

$$w_i^{(r)} = \left|1 - y_i f_{\boldsymbol{\theta}^{(r)}}(\boldsymbol{x}_i)\right|,$$

may equal to zero.

- It is thus prudent to use the approximate weights

$$w_{i,\varepsilon}^{(r)} = w_i^{(r)} + \varepsilon$$

in $\mathbf{W}^{(r)}$ to avoid numerical instabilities, for small $\varepsilon > 0$ (e.g. $\varepsilon = 10^{-5}$).

# Example MM Algorithms (3)

- Notice that

$$\mathbf{W}^{(r)} = \frac{1}{4n} \begin{bmatrix} 1/w_1^{(r)} & 0 & \cdots & 0 \\ 0 & 1/w_2^{(r)} & & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & 1/w_n^{(r)} \end{bmatrix},$$

  contains entries that may be infinite, since

$$w_i^{(r)} = \left| 1 - y_i f_{\boldsymbol{\theta}^{(r)}}(\boldsymbol{x}_i) \right|,$$

  may equal to zero.

- It is thus prudent to use the approximate weights

$$w_{i,\varepsilon}^{(r)} = w_i^{(r)} + \varepsilon$$

  in $\mathbf{W}^{(r)}$ to avoid numerical instabilities, for small $\varepsilon > 0$ (e.g. $\varepsilon = 10^{-5}$).

# Global Convergence Analysis (1)

- For Example 1 (GMM), it can be shown that the updates that we derived satisfies

$$\boldsymbol{\theta}^{(r+1)} = \arg\min_{\boldsymbol{\theta}\in\Theta} M\left(\boldsymbol{\theta}; \boldsymbol{\theta}^{(r)}\right);$$

  see for example Nguyen and McLachlan (2015).

- Unfortunately, the negative log-likelihood $L(\cdot; \mathbf{d}_n)$ is known to be highly multimodal, due to the lack of identifiability of mixture models.

- By Theorem 1, we can only guarantee that if the sequence $\boldsymbol{\theta}^{(r)}$ converges to a stationary point of $L(\boldsymbol{\theta}; \mathbf{d}_n)$, then that stationary point is only a local maximum or a saddle point.

# Global Convergence Analysis (1)

- For Example 1 (GMM), it can be shown that the updates that we derived satisfies

$$\boldsymbol{\theta}^{(r+1)} = \arg\min_{\boldsymbol{\theta}\in\Theta} \; M\left(\boldsymbol{\theta}; \boldsymbol{\theta}^{(r)}\right);$$

  see for example Nguyen and McLachlan (2015).

- Unfortunately, the negative log-likelihood $L(\cdot; \mathbf{d}_n)$ is known to be highly multimodal, due to the lack of identifiability of mixture models.

- By Theorem 1, we can only guarantee that if the sequence $\boldsymbol{\theta}^{(r)}$ converges to a stationary point of $L(\boldsymbol{\theta}; \mathbf{d}_n)$, then that stationary point is only a local maximum or a saddle point.

# Global Convergence Analysis (1)

- For Example 1 (GMM), it can be shown that the updates that we derived satisfies

$$\boldsymbol{\theta}^{(r+1)} = \arg\min_{\boldsymbol{\theta}\in\Theta} \; M\left(\boldsymbol{\theta}; \boldsymbol{\theta}^{(r)}\right);$$

  see for example Nguyen and McLachlan (2015).

- Unfortunately, the negative log-likelihood $L(\cdot; \mathbf{d}_n)$ is known to be highly multimodal, due to the lack of identifiability of mixture models.

- By Theorem 1, we can only guarantee that if the sequence $\boldsymbol{\theta}^{(r)}$ converges to a stationary point of $L(\boldsymbol{\theta}; \mathbf{d}_n)$, then that stationary point is only a local maximum or a saddle point.

# Global Convergence Analysis (1)

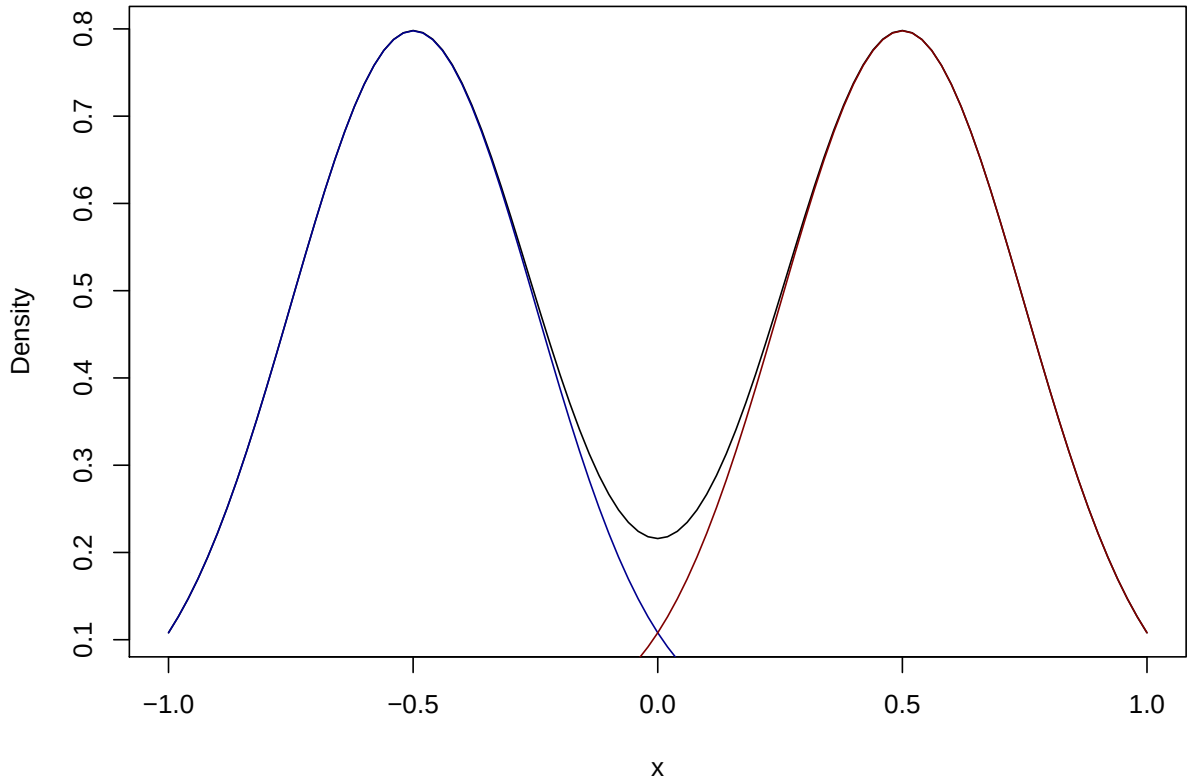


Figure 5: Note that $\frac{1}{2}\phi_1\left(x;\mu_1,1/4^2\right)+\frac{1}{2}\phi_1\left(x;\mu_1,1/4^2\right)$ is the same for $(\mu_1,\mu_2)=\left(\frac{1}{2},-\frac{1}{2}\right)$ or $(\mu_1,\mu_2)=\left(-\frac{1}{2},\frac{1}{2}\right)$.

# Global Convergence Analysis (2)

- For Example 2 (Logistic), it clear that

$$\boldsymbol{\theta}^{(r+1)} = \arg\min_{\boldsymbol{\theta}\in\Theta} M\left(\boldsymbol{\theta}; \boldsymbol{\theta}^{(r)}\right),$$

since $M\left(\boldsymbol{\theta}; \boldsymbol{\theta}^{(r)}\right)$ has a quadratic form.

- We can show that the negative log-quasi-likelihood function $L(\cdot; \mathbf{d}_n)$ is convex (cf. Albert and Anderson, 1984) and thus if it has a stationary point then the stationary point must be a global minimum.

- By Theorem 1, we have the sequence $\boldsymbol{\theta}^{(r)}$ converging to a stationary point of $L(\boldsymbol{\theta}; \mathbf{d}_n)$, which will be a global minimum by convexity.

# Global Convergence Analysis (2)

- For Example 2 (Logistic), it clear that

$$\boldsymbol{\theta}^{(r+1)} = \arg\min_{\boldsymbol{\theta}\in\Theta} M\left(\boldsymbol{\theta}; \boldsymbol{\theta}^{(r)}\right),$$

since $M\left(\boldsymbol{\theta}; \boldsymbol{\theta}^{(r)}\right)$ has a quadratic form.

- We can show that the negative log-quasi-likelihood function $L(\cdot; \mathbf{d}_n)$ is convex (cf. Albert and Anderson, 1984) and thus if it has a stationary point then the stationary point must be a global minimum.

- By Theorem 1, we have the sequence $\boldsymbol{\theta}^{(r)}$ converging to a stationary point of $L(\boldsymbol{\theta}; \mathbf{d}_n)$, which will be a global minimum by convexity.

# Global Convergence Analysis (2)

- For Example 2 (Logistic), it clear that

$$\boldsymbol{\theta}^{(r+1)} = \arg\min_{\boldsymbol{\theta}\in\Theta} M\left(\boldsymbol{\theta}; \boldsymbol{\theta}^{(r)}\right),$$

since $M\left(\boldsymbol{\theta}; \boldsymbol{\theta}^{(r)}\right)$ has a quadratic form.

- We can show that the negative log-quasi-likelihood function $L(\cdot; \mathbf{d}_n)$ is convex (cf. Albert and Anderson, 1984) and thus if it has a stationary point then the stationary point must be a global minimum.

- By Theorem 1, we have the sequence $\boldsymbol{\theta}^{(r)}$ converging to a stationary point of $L(\boldsymbol{\theta}; \mathbf{d}_n)$, which will be a global minimum by convexity.

# Global Convergence Analysis (3)

- For Example 3 (SVM), it clear that

$$\boldsymbol{\theta}^{(r+1)} = \arg\min_{\boldsymbol{\theta}\in\Theta} M\left(\boldsymbol{\theta}; \boldsymbol{\theta}^{(r)}\right),$$

since $M\left(\boldsymbol{\theta}; \boldsymbol{\theta}^{(r)}\right)$ is a least square problem.

- It is simple to check that

$$O(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n}\left[1 - y_i\left(\alpha + \boldsymbol{\beta}^\top \boldsymbol{x}_i\right)\right]_+ + \lambda\boldsymbol{\beta}^\top\boldsymbol{\beta}$$

is convex, since the first term is a composition of a convex and a linear function of $\boldsymbol{\theta}$ and the second term is a quadratic.

- By Theorem 2, and after checking Assumptions (A1) and (A2), we have the sequence $\boldsymbol{\theta}^{(r)}$ converging to a stationary point of $L(\boldsymbol{\theta}; \mathbf{d}_n)$, which will be a global minimum by convexity.

# Global Convergence Analysis (3)

- For Example 3 (SVM), it clear that

$$\boldsymbol{\theta}^{(r+1)} = \arg\min_{\boldsymbol{\theta}\in\Theta} M\left(\boldsymbol{\theta}; \boldsymbol{\theta}^{(r)}\right),$$

since $M\left(\boldsymbol{\theta}; \boldsymbol{\theta}^{(r)}\right)$ is a least square problem.

- It is simple to check that

$$O(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n}\left[1 - y_i\left(\alpha + \boldsymbol{\beta}^\top \boldsymbol{x}_i\right)\right]_+ + \lambda\boldsymbol{\beta}^\top\boldsymbol{\beta}$$

is convex, since the first term is a composition of a convex and a linear function of $\boldsymbol{\theta}$ and the second term is a quadratic.

- By Theorem 2, and after checking Assumptions (A1) and (A2), we have the sequence $\boldsymbol{\theta}^{(r)}$ converging to a stationary point of $L(\boldsymbol{\theta}; \mathbf{d}_n)$, which will be a global minimum by convexity.

# Global Convergence Analysis (3)

- For Example 3 (SVM), it clear that

$$\boldsymbol{\theta}^{(r+1)} = \arg\min_{\boldsymbol{\theta}\in\Theta} \; M\left(\boldsymbol{\theta}; \boldsymbol{\theta}^{(r)}\right),$$

since $M\left(\boldsymbol{\theta}; \boldsymbol{\theta}^{(r)}\right)$ is a least square problem.

- It is simple to check that

$$O(\boldsymbol{\theta}) \;=\; \frac{1}{n}\sum_{i=1}^{n}\left[1 - y_i\left(\alpha + \boldsymbol{\beta}^\top \boldsymbol{x}_i\right)\right]_+ + \lambda\boldsymbol{\beta}^\top\boldsymbol{\beta}$$

is convex, since the first term is a composition of a convex and a linear function of $\boldsymbol{\theta}$ and the second term is a quadratic.

- By Theorem 2, and after checking Assumptions (A1) and (A2), we have the sequence $\boldsymbol{\theta}^{(r)}$ converging to a stationary point of $L(\boldsymbol{\theta}; \mathbf{d}_n)$, which will be a global minimum by convexity.

# Simulation Study (1)

- Suppose that we have data $\mathbf{D}_n = \{X_i\}_{i=1}^n$ $(n = 10000)$, where $X_i \in \mathbb{R}$ is sampled from a 4-component Gaussian distribution with density function of the form

$$f_{\boldsymbol{\theta}_0}(x_i) = \sum_{z=1}^{4} \frac{1}{4} \phi_p \left( x_i; \frac{z+1}{6}, \frac{1}{16^2} \right).$$

- Using the MM algorithm for maximum likelihood estimation $\hat{\boldsymbol{\theta}}_n$, we wish to estimate the parameter vector $\boldsymbol{\theta}_0$ containing $\pi_z = 1/4$, $\Sigma_z = 1/16^2$, and $\mu_z = (z+1)/6$, for each $z \in [4]$.

- We can use the **normalmixEM** function from the **mixtools** package for **R** (Benaglia et al., 2009).

# Simulation Study (1)

- Suppose that we have data $\mathbf{D}_n = \{X_i\}_{i=1}^n$ $(n = 10000)$, where $X_i \in \mathbb{R}$ is sampled from a 4-component Gaussian distribution with density function of the form

$$f_{\boldsymbol{\theta}_0}(x_i) = \sum_{z=1}^{4} \frac{1}{4} \phi_p \left( x_i; \frac{z+1}{6}, \frac{1}{16^2} \right).$$

- Using the MM algorithm for maximum likelihood estimation $\hat{\boldsymbol{\theta}}_n$, we wish to estimate the parameter vector $\boldsymbol{\theta}_0$ containing $\pi_z = 1/4$, $\Sigma_z = 1/16^2$, and $\mu_z = (z+1)/6$, for each $z \in [4]$.

- We can use the **normalmixEM** function from the **mixtools** package for **R** (Benaglia et al., 2009).

# Simulation Study (1)

- Suppose that we have data $\mathbf{D}_n = \{X_i\}_{i=1}^n$ $(n = 10000)$, where $X_i \in \mathbb{R}$ is sampled from a 4-component Gaussian distribution with density function of the form

$$f_{\boldsymbol{\theta}_0}(x_i) = \sum_{z=1}^{4} \frac{1}{4} \phi_p \left( x_i; \frac{z+1}{6}, \frac{1}{16^2} \right).$$

- Using the MM algorithm for maximum likelihood estimation $\hat{\boldsymbol{\theta}}_n$, we wish to estimate the parameter vector $\boldsymbol{\theta}_0$ containing $\pi_z = 1/4$, $\Sigma_z = 1/16^2$, and $\mu_z = (z+1)/6$, for each $z \in [4]$.

- We can use the **normalmixEM** function from the **mixtools** package for **R** (Benaglia et al., 2009).

# Simulation Study (1)

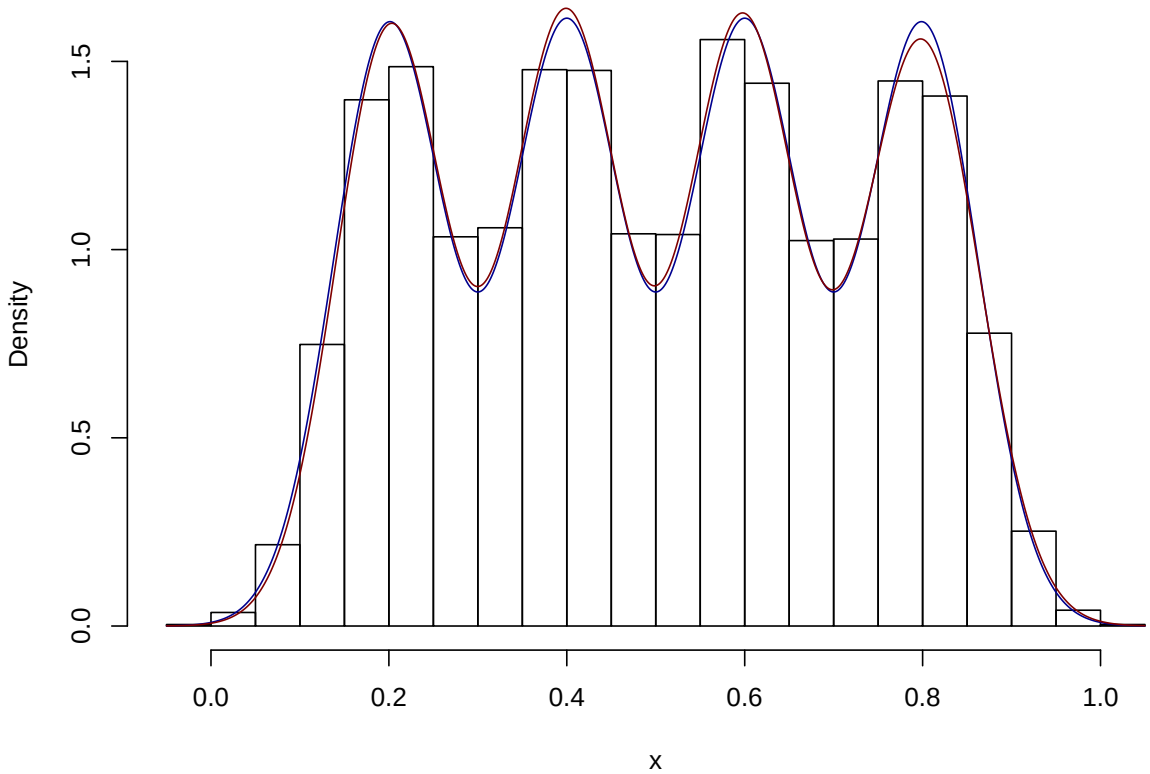


Figure 6: Histogram of simulated data $\mathbf{d}_n$ with the true density plotted in blue and density with estimated parameter vector plotted in red.

# Simulation Study (1)

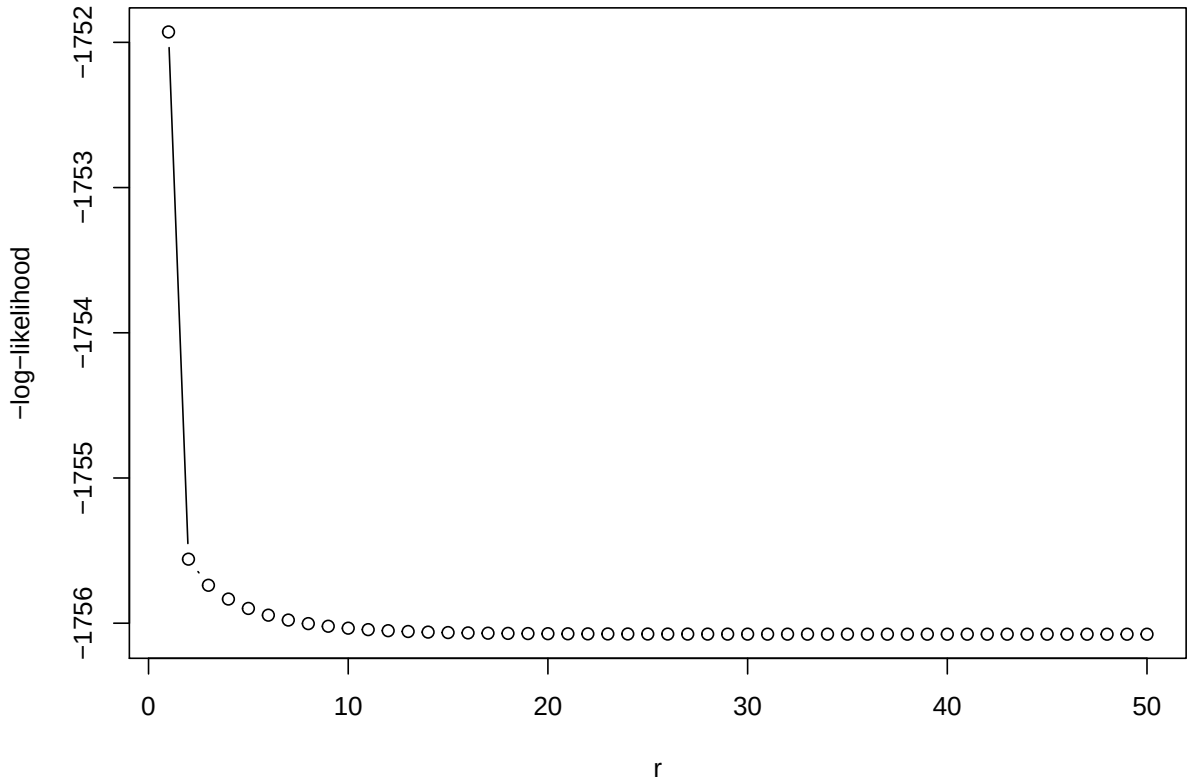


Figure 7: Sequence of negative log-likelihood evaluates $L\left(\boldsymbol{\theta}^{(r)}; \mathbf{d}_n\right)$ for $r \in [50]$.

# Simulation Study (2)

- Suppose that we have data $\mathbf{D}_n = \{(\mathbf{X}_i, Y_i)\}_{i=1}^n$ $(n = 100)$, where $\boldsymbol{X}_i^\top = (1, U_i)$ and $U_i$ is sampled from a uniform distribution on the interval $[-1, 1]$ and
- $$Y_i = \begin{cases} 0 & \text{if } X_i < 0, \\ 1 & \text{if } X_i \geq 0, \end{cases}$$

  for $i \in [n]$.
- Using the MM algorithm for maximum quasi-likelihood estimation $\hat{\boldsymbol{\theta}}_n$, we wish to estimate the parameter vector $\boldsymbol{\theta}_0$ that corresponds to the best logistic regression model.

# Simulation Study (2)

- Suppose that we have data $\mathbf{D}_n = \{(\mathbf{X}_i, Y_i)\}_{i=1}^n$ $(n = 100)$, where $\boldsymbol{X}_i^\top = (1, U_i)$ and $U_i$ is sampled from a uniform distribution on the interval $[-1, 1]$ and
- $$Y_i = \begin{cases} 0 & \text{if } X_i < 0, \\ 1 & \text{if } X_i \geq 0, \end{cases}$$

  for $i \in [n]$.
- Using the MM algorithm for maximum quasi-likelihood estimation $\hat{\boldsymbol{\theta}}_n$, we wish to estimate the parameter vector $\boldsymbol{\theta}_0$ that corresponds to the best logistic regression model.

# Simulation Study (2)

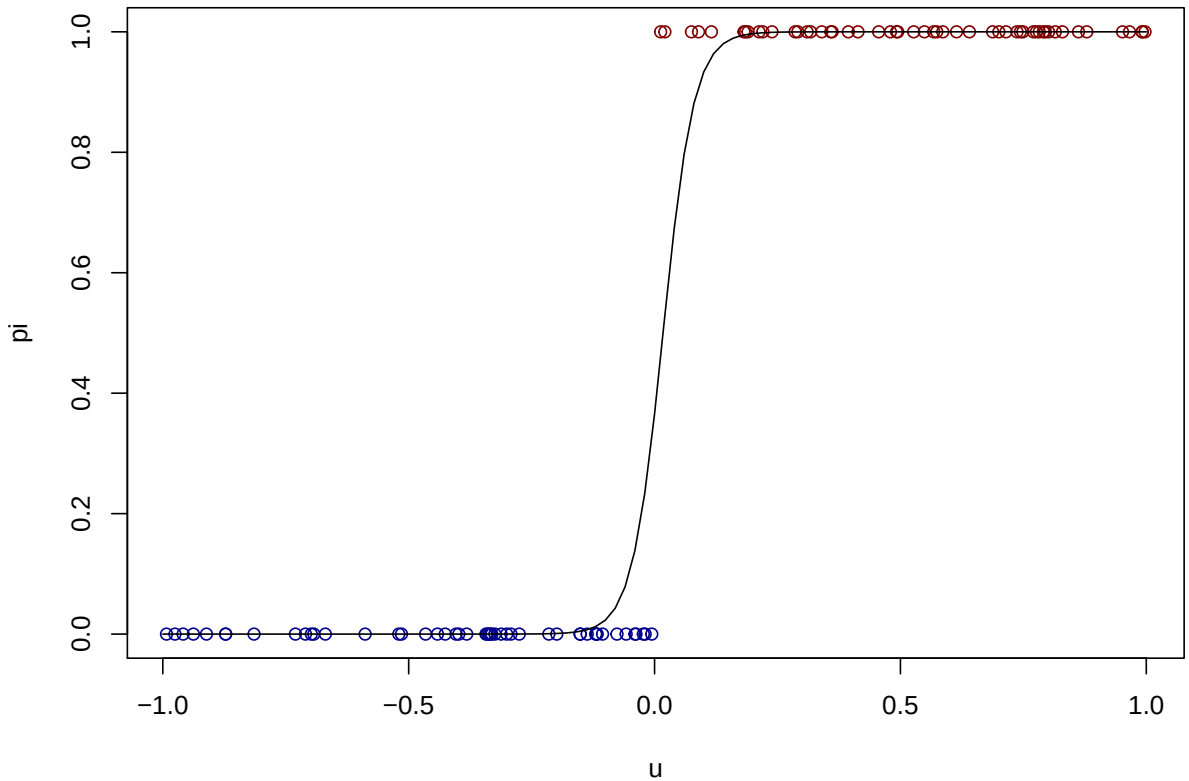


Figure 8: Data pairs $(u_i, y_i)$ for $i \in [100]$ with fitted logistic curve $\pi_{\hat{\boldsymbol{\theta}}_n}(\boldsymbol{x}_i)$, where $\hat{\boldsymbol{\theta}}_n^\top = (-0.552, 31.912)$.

# Simulation Study (2)

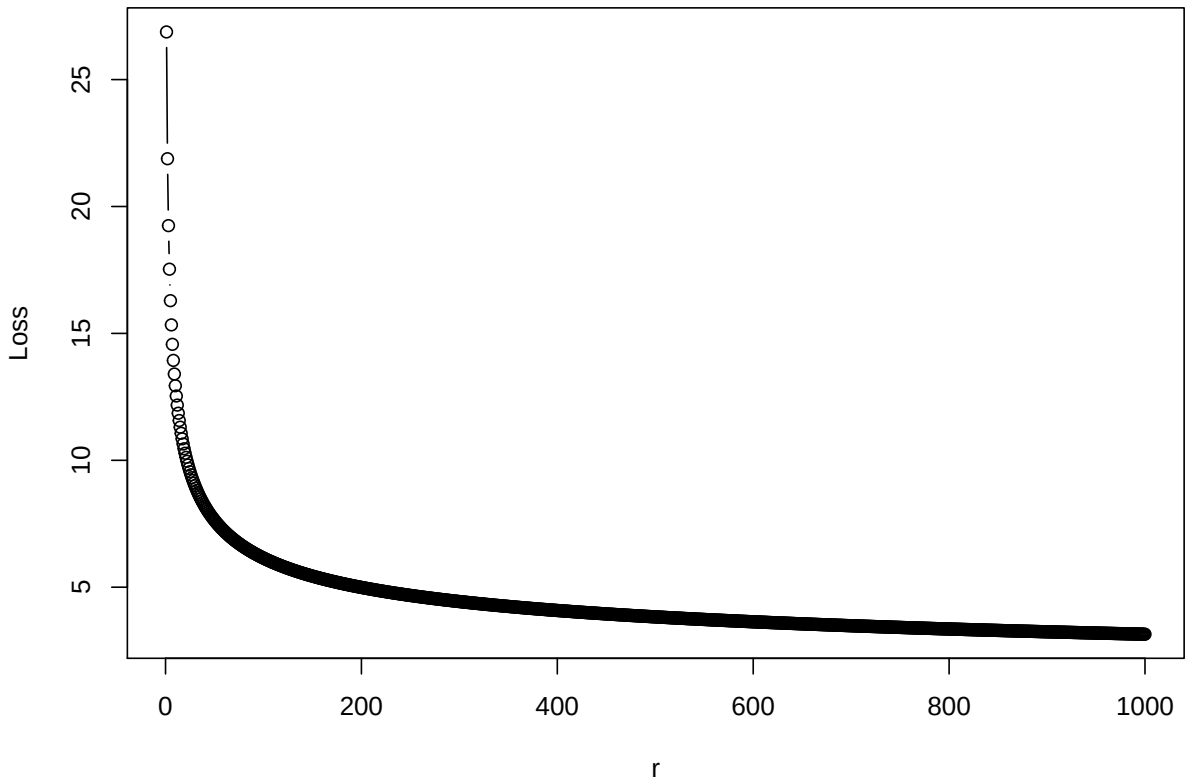


Figure 9: Sequence of negative log-quasi-likelihood evaluates $L\left(\boldsymbol{\theta}^{(r)}; \mathbf{d}_n\right)$ for $r \in [1000]$.

# Simulation Study (3)

- Suppose that we have data $\mathbf{D}_n = \{(\mathbf{X}_i, Y_i)\}_{i=1}^n$ $(n = 200)$, where $Y_i$ is generated with equal probability from the set $\{-1, 1\}$, and $\boldsymbol{X}_i | Y_i = -1$ and $\boldsymbol{X}_i | Y_i = +1$ are generated from distributions with densities

$$\phi_2\left(\boldsymbol{x}; \begin{bmatrix} -1 \\ -1 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\right)$$

and

$$\phi_2\left(\boldsymbol{x}; \begin{bmatrix} +1 \\ +1 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\right)$$

for $i \in [n]$.

- We utilize the MM algorithm for linear SVMs to obtain the optimal separation rule for discriminating between observations with different labels $Y_i$, with $\lambda = 0$.

# Simulation Study (3)

- Suppose that we have data $\mathbf{D}_n = \{(\mathbf{X}_i, Y_i)\}_{i=1}^n$ $(n = 200)$, where $Y_i$ is generated with equal probability from the set $\{-1, 1\}$, and $\boldsymbol{X}_i | Y_i = -1$ and $\boldsymbol{X}_i | Y_i = +1$ are generated from distributions with densities

$$\phi_2\left(\boldsymbol{x}; \begin{bmatrix} -1 \\ -1 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\right)$$

and

$$\phi_2\left(\boldsymbol{x}; \begin{bmatrix} +1 \\ +1 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\right)$$

for $i \in [n]$.

- We utilize the MM algorithm for linear SVMs to obtain the optimal separation rule for discriminating between observations with different labels $Y_i$, with $\lambda = 0$.

# Simulation Study (3)

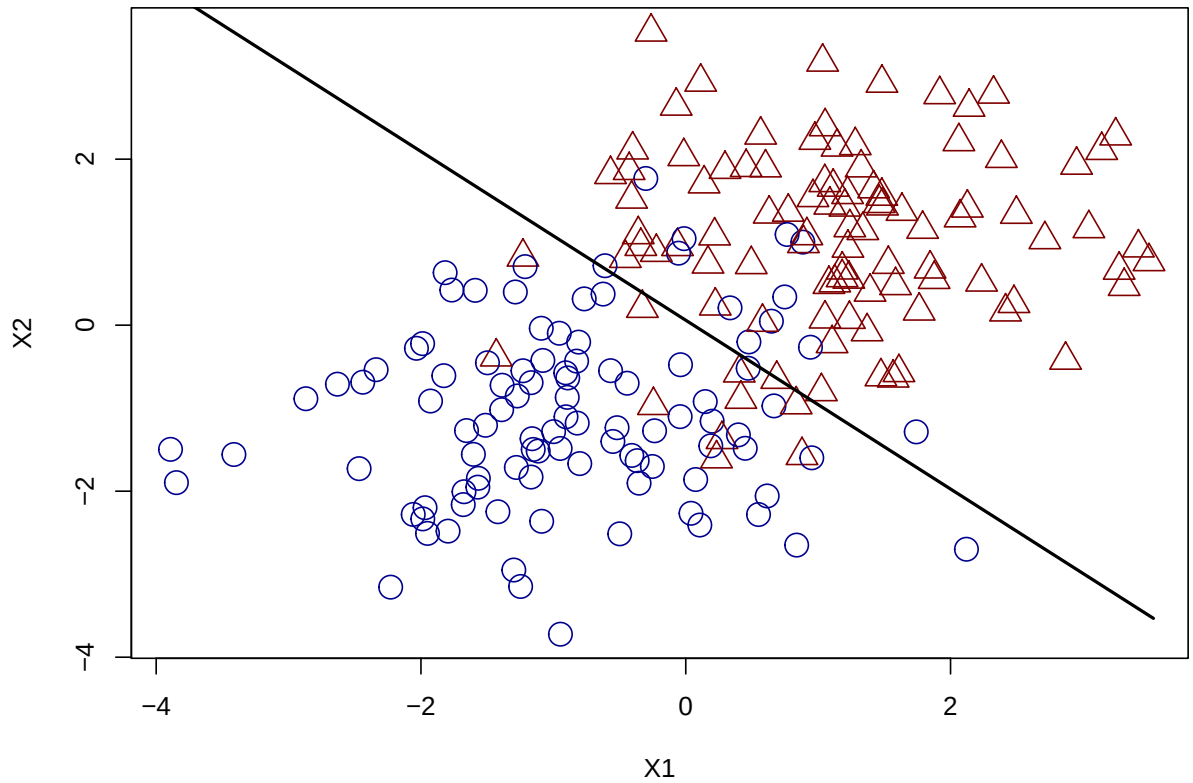


Figure 10: Observations $\boldsymbol{x}_i = (x_1, x_2)$ for $i \in [200]$ with fitted discriminate curve $f_{\hat{\boldsymbol{\theta}}_n}(\boldsymbol{x}_i)$, where $\hat{\boldsymbol{\theta}}_n^\top = (-0.055, 0.936, 0.921)$. Circles and triangles indicate observations with $y_i = -1$ and $y_i = +1$, respectively.

# Simulation Study (3)

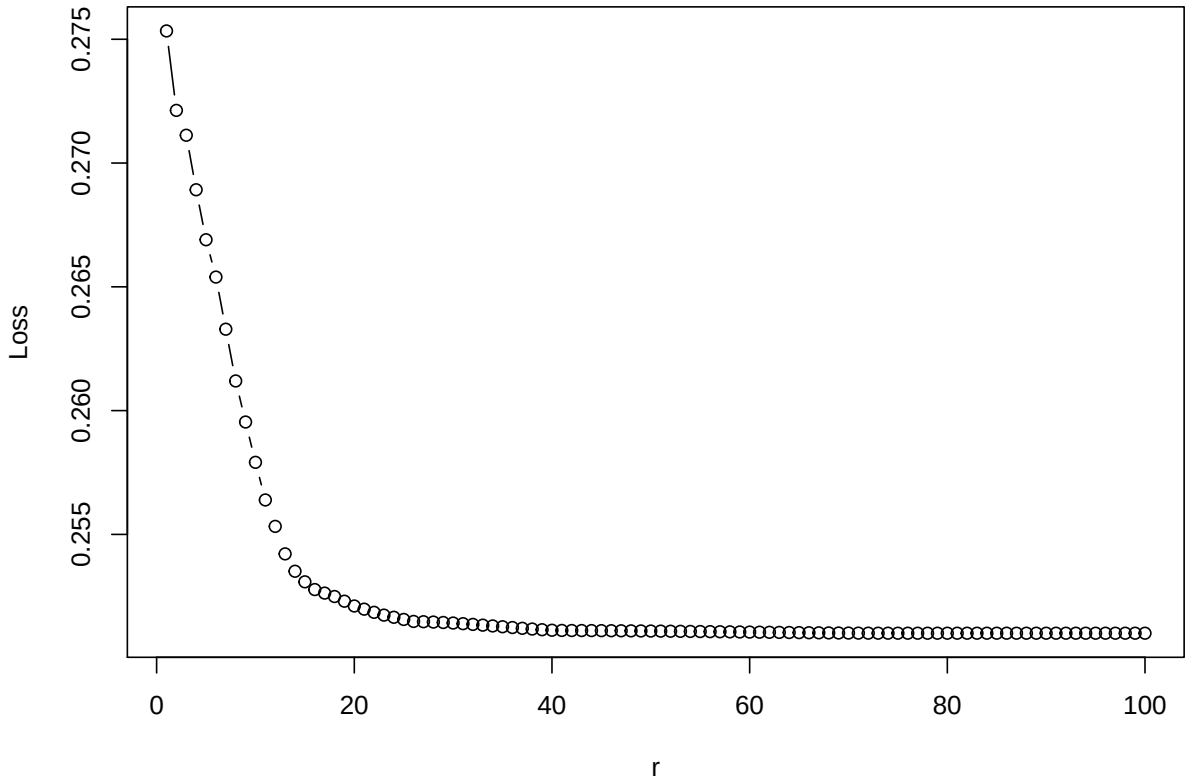


Figure 11: Sequence of SVM regularized loss evaluates $L\left(\boldsymbol{\theta}^{(r)}; \mathbf{d}_n\right)$ for $r \in [100]$.

# Further Readings

- Extensions of MM algorithms to include coordinate descent, block descent, and random block descent are considered in Razaviyayn et al. (2013) and Hong et al. (2016), along with the convergence analysis of said algorithms.
- Convergence rates for MM algorithms under Lipschitz, Lipschitz smooth, convexity, and strong convexity conditions are proved in Mairal (2015).
- Stochastic approximation versions of MM algorithms are proposed and analyzed under convexity conditions in Mairal (2013).

# Further Readings

- Extensions of MM algorithms to include coordinate descent, block descent, and random block descent are considered in Razaviyayn et al. (2013) and Hong et al. (2016), along with the convergence analysis of said algorithms.
- Convergence rates for MM algorithms under Lipschitz, Lipschitz smooth, convexity, and strong convexity conditions are proved in Mairal (2015).
- Stochastic approximation versions of MM algorithms are proposed and analyzed under convexity conditions in Mairal (2013).

# Further Readings

- Extensions of MM algorithms to include coordinate descent, block descent, and random block descent are considered in Razaviyayn et al. (2013) and Hong et al. (2016), along with the convergence analysis of said algorithms.
- Convergence rates for MM algorithms under Lipschitz, Lipschitz smooth, convexity, and strong convexity conditions are proved in Mairal (2015).
- Stochastic approximation versions of MM algorithms are proposed and analyzed under convexity conditions in Mairal (2013).

# Bibliographical Details

- A concise introduction to MM algorithms can be found in Hunter and Lange (2004).
- Convergence results for MM algorithms and their variants are comprehensively and cohesively studied in Razaviyayn et al. (2013) and Mairal (2015).
- Pedagogical references for the construction and application of MM algorithms are provided in Lange (2013) and Lange (2016).

# Bibliographical Details

- A concise introduction to MM algorithms can be found in Hunter and Lange (2004).
- Convergence results for MM algorithms and their variants are comprehensively and cohesively studied in Razaviyayn et al. (2013) and Mairal (2015).
- Pedagogical references for the construction and application of MM algorithms are provided in Lange (2013) and Lange (2016).

# Bibliographical Details

- A concise introduction to MM algorithms can be found in Hunter and Lange (2004).
- Convergence results for MM algorithms and their variants are comprehensively and cohesively studied in Razaviyayn et al. (2013) and Mairal (2015).
- Pedagogical references for the construction and application of MM algorithms are provided in Lange (2013) and Lange (2016).

# References I


Albert, A. and Anderson, J. A. (1984). On the existence of maximum likelihood estimates in logistic regression models. *Biometrika*, 71:1–10.

Amemiya, T. (1985). *Advanced Econometrics*. Harvard University Press, Cambridge.

Benaglia, T., Chauveau, D., Hunter, D. R., and Young, D. S. (2009). mixtools: An R package for analyzing finite mixture models. *Journal of Statistical Software*, 32:1–29.

Bickel, P. J. and Doksum, K. A. (2001). *Mathematical Statistics: Basic Ideas and Selected Topics*. Prentice Hall, Upper Saddle River.

Boyd, S. and Vandenberghe, L. (2004). *Convex Optimization*. Cambridge University Press, Cambridge.

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20:273–297.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society Series B*, 39:1–38.

# References II


Hong, M., Razaviyayn, M., Luo, Z.-Q., and Pang, J.-S. (2016). A unified algorithmic framework for block-structured optimization involving big data: with applications in machine learning and signal process. *IEEE Signal Processing Magazine*, 33:57–77.

Hunter, D. R. and Lange, K. (2004). A tutorial on MM algorithms. *The American Statistician*, 58:30–37.

Lange, K. (2013). *Optimization*. Springer, New York.

Lange, K. (2016). *MM Optimization Algorithms*. SIAM, Philadelphia.

Mairal, J. (2013). Stochastic majorization-minimization algorithms for large-scale optimization. In *Advances in Neural Information Processing*.

Mairal, J. (2015). Incremental majorization-minimization optimization with application to large-scale machine learning. *SIAM Journal of Optimization*, 25:829–855.

Nguyen, H. D. and McLachlan, G. J. (2015). Maximum likelihood estimation of Gaussian mixture models without matrix operations. *Advances in Data Analysis and Classification*, 9:371–394.

# References III


Razaviyayn, M., Hong, M., and Luo, Z.-Q. (2013). A unified convergence analysis of block successive minimization methods for nonsmooth optimization. *SIAM Journal of Optimization*, 23:1126–1153.

Serfling, R. J. (1980). *Approximation Theorems Of Mathematical Statistics*. Wiley, New York.

Vapnik, V. N. (1998). *Statistical Learning Theory*. Wiley, New York.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica*, 50:1–25.

# Thank You!

tinyurl.com/hiendnguyen