

# Optimization Theory for Statistics and Machine Learning

Dr. Hien Nguyen  
[hiendn.github.io](https://github.com/hiendn)

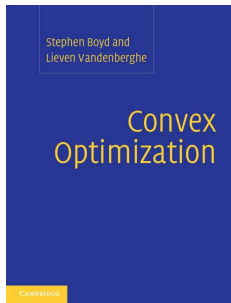
Lecturer, La Trobe University



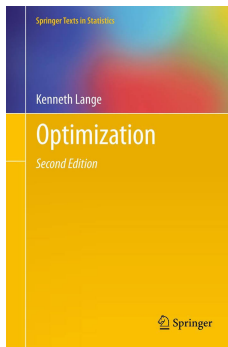
# Contents of this course

- Introduce interesting statistical and machine learning problems that can be solved via optimization.
- Present the core concepts of modern optimization theory that are required to solve these modern problems.
- Propose the *MM* algorithm framework as a unifying methodology for constructing optimization algorithms.
- Demonstrate how these algorithms can be implemented within the R programming language.
- All course contents can be found at <https://github.com/hiendn/CaenOptimization2018>.

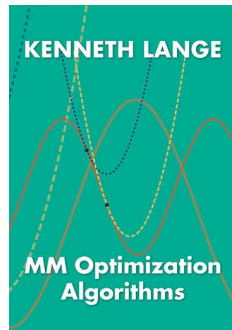
# Key readings



**(a)** Boyd and Vandenberghe, 2004



**(b)** Lange, 2013



**(c)** Lange, 2016

**Figure 1:** The contents of this course can mostly be found in the following books.

# What is an optimization problem?

Let  $f(\cdot) : \mathbb{T} \rightarrow \mathbb{R}$  be an **objective** function of interest, where  $\mathbb{T} \subseteq \mathbb{R}^d$ ,  $d \in \mathbb{N}$ , and  $\mathbb{N}$  and  $\mathbb{R}$  denote the **natural** and the **real** numbers, respectively.

We will generally denote a typical element of  $\mathbb{T}$  by  $\theta$ .

The general problem of mathematical **optimization** over real domains  $\mathbb{T} \subseteq \mathbb{R}^d$ , is find the either the maximum or the minimum values of  $f$  over  $\mathbb{T}$ .

# A fair warning

From the famous book of Nesterov (2004), the author gives the following two quotes in the first chapter.

1. Optimization is a very important and promising application theory. It covers almost *all* needs of operations research and numerical analysis.
2. In general, optimization problems are *unsolvable*.

## **Some examples of optimization problems**

# Regularized linear regression

Suppose that  $y_1, \dots, y_n \in \mathbb{R}$  are  $n \in \mathbb{N}$  observe **responses**, explained by their companion **covariates**  $\mathbf{x}_1, \dots, \mathbf{x}_n \in \mathbb{R}^d$ .

We wish to determine the coefficients  $\alpha \in \mathbb{R}$  and  $\beta \in \mathbb{R}^d$ , such that the quantity

$$\frac{1}{n} \sum_{i=1}^n |y_i - \alpha - \beta^\top \mathbf{x}_i|^p + \lambda \sum_{j=1}^d |\beta_j|^q,$$

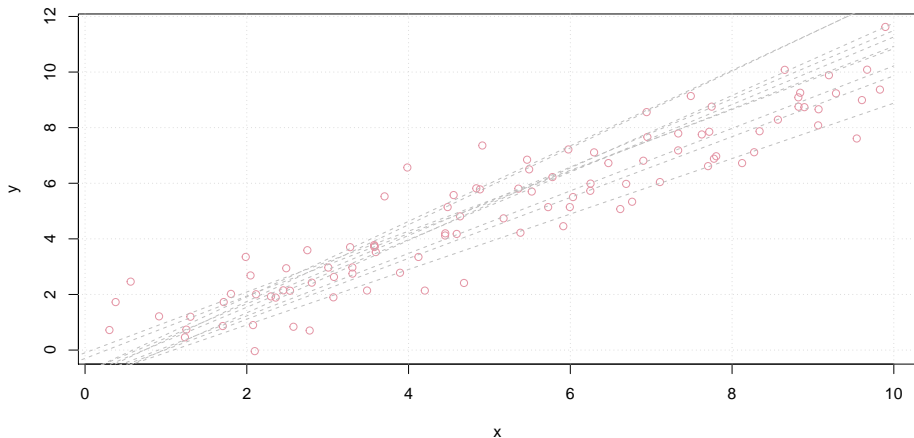
is **minimized**, where  $\lambda \in [0, \infty)$  is a **penalty**, for any  $p, q \in [1, \infty)$ . Here,  $(\cdot)^\top$  is the matrix transposition operator, and  $\theta^\top = (\alpha, \beta^\top) \in \mathbb{R}^{d+1}$ , where

$$\beta^\top = (\beta_1, \dots, \beta_d).$$

We can, more concisely write the problem as:

$$\min_{\theta \in \mathbb{R}^{d+1}} \frac{1}{n} \sum_{i=1}^n |y_i - \alpha - \beta^\top \mathbf{x}_i|^p + \lambda \sum_{j=1}^d |\beta_j|^q.$$

# An example of the regression problem



**Figure 2:** Example of 10 potential linear regression functions when  $d=1$ .



# Various regularized regression problems

- Ordinary least-squares regression ( $p = 2, \lambda = 0$ ).
- Least-absolute deviation regression ( $p = 2, \lambda = 0$ ).
- Ridge regression of Hoerl and Kennard (1970) ( $p = 2, q = 2, \lambda > 0$ ).
- LASSO of Tibshirani (1996) ( $p = 2, q = 1, \lambda > 0$ ).
- The  $\ell_1$ -LASSO of Wu and Lange (2008) ( $p = 1, q = 1, \lambda > 0$ ).

# Discrimination via optimal separation hyperplanes

Suppose that  $y_1, \dots, y_n \in \{-1, 1\}$  are  $n$  spin-binary variables, explained by their companion covariates  $\mathbf{x}_1, \dots, \mathbf{x}_n \in \mathbb{R}^d$ .

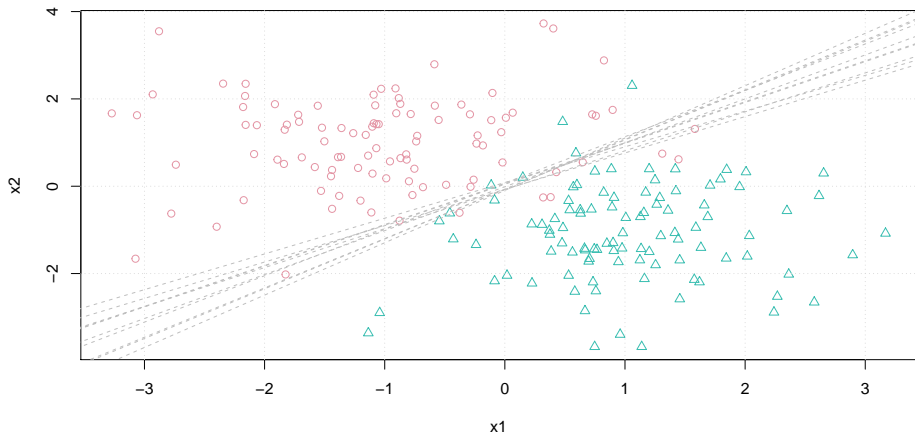
We wish to obtain an optimal hyperplane of the form  $\alpha + \beta^\top \mathbf{x}$ , where  $\alpha \in \mathbb{R}$ ,  $\beta \in \mathbb{R}^d$ ,  $\mathbf{x} \in \mathbb{R}^d$ , and  $\theta^\top = (\alpha, \beta^\top)$ , such that it minimizes the regularized average **loss**

$$\frac{1}{n} \sum_{i=1}^n l(y_i, \alpha + \beta^\top \mathbf{x}_i) + \lambda \sum_{j=1}^d |\theta_j|^2,$$

where  $\lambda \in [0, \infty)$ , and  $l(y, \alpha + \beta^\top \mathbf{x}) = [y(\alpha + \beta^\top \mathbf{x}) < 0]$  is the **classification** loss function.

Here,  $[\cdot]$  is the **Iverson bracket** notation which equals **1** if the content is true and **0**, otherwise.

# Example of hyperplane discrimination functions



**Figure 3:** Example of 10 potential discriminant hyperplanes in 2 dimensions.

# The support vector machine

The classification loss function

$$l(y, \alpha + \beta^\top \mathbf{x}) = \mathbb{I}[y(\alpha + \beta^\top \mathbf{x}) < 0]$$

is *irregular* due to its lack of **convexity** and lack of **differentiability** at the point where  $y(\alpha + \beta^\top \mathbf{x}) = 0$ , with respect to  $\theta$ .

In Cortes and Vapnik (1995), the authors proposed a convex approximation of the classification loss function, using the so-called **hinge** loss function

$$l(y, \alpha + \beta^\top \mathbf{x}) = [1 - y(\alpha + \beta^\top \mathbf{x})]_+,$$

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

The resulting optimization problem

$$\min_{\theta=(\alpha, \beta) \in \mathbb{R}^{d+1}} \frac{1}{n} \sum_{i=1}^n [1 - y_i(\alpha + \beta^\top \mathbf{x}_i)]_+ + \lambda \sum_{j=1}^d |\beta_j|^2,$$

is the original **support vector machine** (SVM) problem.

# General SVM problems

- **Logistic regression** is obtained by setting

$$l(y, \alpha + \beta^\top \mathbf{x}) = \log \left[ 1 + \exp \left( -y \left[ \alpha + \beta^\top \mathbf{x} \right] \right) \right].$$

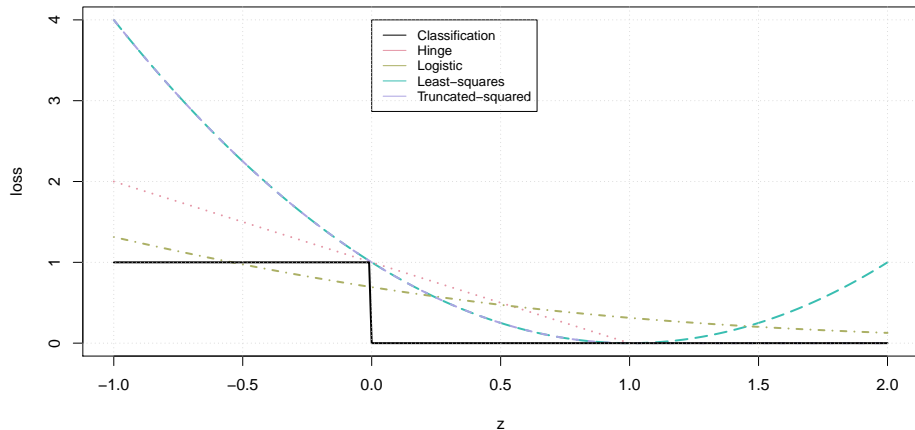
- The **least-squares** SVM of Suykens and Vandewalle (1999) is obtained by setting

$$l(y, \alpha + \beta^\top \mathbf{x}) = \left[ 1 - y \left( \alpha + \beta^\top \mathbf{x} \right) \right]^2.$$

- The **truncated-squared** loss SVM of Rosset, Zhu, and Hastie (2004) is obtained by setting

$$l(y, \alpha + \beta^\top \mathbf{x}) = \left[ 1 - y \left( \alpha + \beta^\top \mathbf{x} \right) \right]_+^2.$$

# A comparison of loss functions



**Figure 4:** A comparison of SVM loss functions.

# Maximum likelihood estimation

Let  $\mathbf{X} \in \mathbb{X}$  and  $\mathbf{Y} \in \mathbb{Y}$  be two random variables that share a joint *parametric probability density function* (PDF) of known form

$$f(\mathbf{x}, \mathbf{y}; \boldsymbol{\theta}),$$

where  $\boldsymbol{\theta} \in \mathbb{T}$  is a **parameter** vector that characterizes the relationship between  $\mathbf{X}$  and  $\mathbf{Y}$ .

If we observe both  $\mathbf{X}$  and  $\mathbf{Y}$  for a **data generating process** (DGP) that can be characterized by the PDF  $f(\mathbf{x}, \mathbf{y}; \boldsymbol{\theta}_0)$ , where  $\boldsymbol{\theta}_0$  is unknown, then we may estimate it via the method of **maximum likelihood estimation** (MLE), by solving the optimization problem

$$\max_{\boldsymbol{\theta} \in \mathbb{T}} \log f(\mathbf{X}, \mathbf{Y}; \boldsymbol{\theta}).$$

We say that the value of  $\boldsymbol{\theta}$  which solves the problem is the **maximum likelihood estimator** or **estimate** (MLE), and denote it by

$$\hat{\boldsymbol{\theta}} = \arg \max_{\boldsymbol{\theta} \in \mathbb{T}} \log f(\mathbf{X}, \mathbf{Y}; \boldsymbol{\theta}).$$

# Latent variable problems

Suppose that we only observe  $\mathbf{X}$  and not  $\mathbf{Y}$ , out of the pair. We say that  $\mathbf{X}$  is **observed** and  $\mathbf{Y}$  is **hidden** or **latent**.

In such a situation, we can characterize the DGP of what we observe via the *marginal* PDF

$$f(\mathbf{x}; \boldsymbol{\theta}) = \int_{\mathbb{Y}} f(\mathbf{x}, \mathbf{y}; \boldsymbol{\theta}) d\mathbf{y}.$$

We can still conduct MLE in order to estimate the value of  $\boldsymbol{\theta}_0$  by solving the problem

$$\max_{\boldsymbol{\theta} \in \mathbb{T}} \log f(\mathbf{X}; \boldsymbol{\theta}),$$

although the task is made more difficult due to the integration over  $\mathbf{Y}$ .

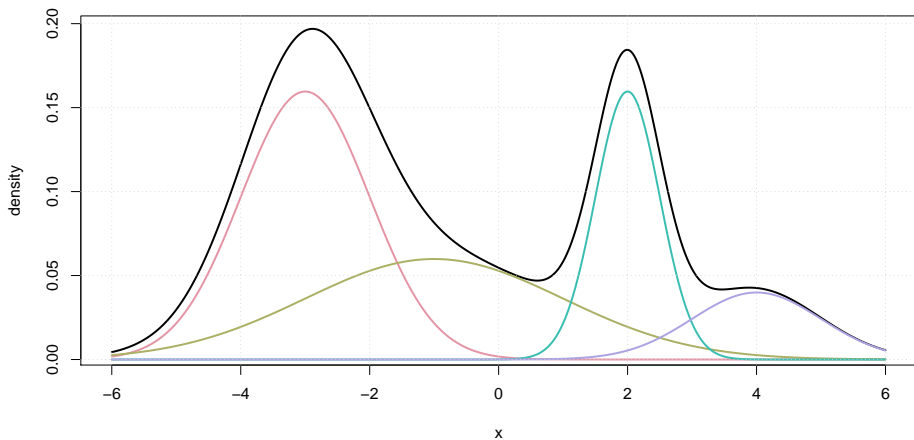
Such problems involving latent variables occur often in statistics, but may still be solvable via the famous *EM* algorithm of Dempster, Laird, and Rubin (1977) if enough structure is known regarding the relationship between  $\mathbf{X}$  and  $\mathbf{Y}$ .



# Examples of latent variable problems

- Elliptical density estimation.
- Factor analysis.
- Finite mixture models.
- Hidden Markov modeling.
- Linear mixed-effects modeling.
- Multiple missing data imputation.
- Non-negative matrix factorization.
- Probabilistic principal component analysis.
- Skew density estimation.

# Finite mixture models



**Figure 5:** A 4-component mixture of normal PDFs.

# Fundamental definitions and results

# Global maxima and minima

We say that a point  $\theta^*$  in the **domain** or **support** (i.e.  $\mathbb{T}$ ) of  $f(\cdot) : \mathbb{T} \rightarrow \mathbb{R}$  is a **global maximizer** if

$$f(\theta^*) \geq f(\theta),$$

for all  $\theta \in \mathbb{T}$ . We call the value  $f(\theta^*)$  the **global maximum**.

If

$$f(\theta^*) > f(\theta),$$

for all  $\theta \neq \theta^*$ , then we say that  $\theta^*$  is a **strict** global maximizer. Notice that by definition, a strict global maximizer must be *unique*, if it exists.

The definition of **global minimizer**, **global minimum**, and **strict** global minimizer can be obtained by reversing the inequalities.

# The Euclidean norm

For any  $p \in [1, \infty)$ , denote the  $\ell_p$  vector norm by

$$\|\boldsymbol{\theta}\|_p = \left( \sum_{j=1}^d |\theta_j|^p \right)^{1/p},$$

where  $\boldsymbol{\theta}^\top = (\theta_1, \dots, \theta_d) \in \mathbb{R}^d$ .

Setting  $p = 2$ , we obtain the  $\ell_2$  norm  $\|\cdot\|_2$ , which is generally referred to as the **Euclidean norm**.

# The Euclidean metric

We say that a function

$$\Delta(\cdot, \cdot) : \mathbb{R}^d \times \mathbb{R}^d \rightarrow [0, \infty)$$

is a **metric** if, for all  $\psi, \theta, v \in \mathbb{R}^d$ , it satisfies the conditions:

1.  $\Delta(\theta, v) \geq 0$ .
2.  $\Delta(\theta, v) = 0$  if and only if  $\theta = v$ .
3.  $\Delta(\theta, v) = \Delta(v, \theta)$ .
4.  $\Delta(\psi, v) \leq \Delta(\psi, \theta) + \Delta(\theta, v)$ .

It can be shown that setting

$$\Delta(\theta, v) = \|\theta - v\|_p$$

yields a metric for any  $p \in [0, \infty)$ . Again, in the case where  $p = 2$ , we obtain the **Euclidean metric**

$$\Delta(\theta, v) = \|\theta - v\|_2.$$

# Local maxima and minima

If we equip our real space  $\mathbb{T} \subseteq \mathbb{R}^d$  with the Euclidean norm, then we obtain the **Euclidean metric space**, which equips our space with *topological* properties that can be used to characterize functional behavior.

We now define a **local maximizer** as a point  $\theta^* \in \mathbb{T}$ , such that there exists some  $\epsilon > 0$  for which  $f(\theta^*) \geq f(\theta)$ , for all

$$\theta \in B_\epsilon(\theta^*) = \left\{ \theta \in \mathbb{R}^d : \|\theta - \theta^*\|_2 < \epsilon \right\}.$$

The value  $f(\theta^*)$  is then defined as a **local maximum**. Here, we say that the  $B_\epsilon(\theta^*)$  is the  $\epsilon$  (Euclidean) **ball** of  $\theta^*$ .

We can define a **strict** local maximizer by replacing the  $\geq$  symbol by a  $>$  symbol.

Furthermore, we can define **local minimizer**, **local minimum**, and **strict** local minimizer by reversing the inequalities.

## A bit of set theory

We say that a point  $\theta^* \in \mathbb{R}^d$  is a **limit point** of  $\mathbb{T}$  if for every ball  $N_\epsilon(\theta^*)$ , the following is true:

$$\mathbb{T} \cap N_\epsilon(\theta^*) \neq \{\}.$$

We can now define a **closed** set in a *real metric space* as a set that contains all of its limit points. Furthermore, we can say that a set  $\mathbb{T}$  is **open** if its *complement*  $\mathbb{R}^d \setminus \mathbb{T}$  is closed.

We say that a set  $\mathbb{T} \subset \mathbb{R}^d$  is **bounded** if there exists a finite  $\epsilon$  and some  $\theta \in \mathbb{R}^d$ , such that

$$\mathbb{T} \cap N_\epsilon(\theta) = \mathbb{T}.$$

By the famous *Heine-Borel theorem*, every closed and bounded set in the Euclidean metric space is **compact**.



# A first existence theorem

When  $\mathbb{T} \subset \mathbb{R}$ , the **extreme value theorem** in calculus states that if  $\mathbb{T} = [a, b]$ , where  $-\infty < a < b < \infty$ , and if  $f(\cdot) : [a, b] \rightarrow \mathbb{R}$  is *continuous*, then there exists  $c, d \in [a, b]$ , such that

$$f(c) \leq f(\theta) \leq f(d),$$

for all  $\theta \in [a, b]$ .

The famous *Weierstrass optimality theorem* generalizes the extreme value theorem, and states that if  $\mathbb{T} \subset \mathbb{R}^d$  is compact and if  $f(\cdot) : \mathbb{T} \rightarrow \mathbb{R}$  is continuous, then there exists  $\psi, \nu \in \mathbb{T}$ , such that

$$f(\psi) \leq f(\theta) \leq f(\nu),$$

for all  $\theta \in \mathbb{T}$ .

Thus, if  $\mathbb{T}$  is compact and  $f$  is continuous, then there exists at least one global minimizer and one global maximizer of  $f$ .

# Differentiable functions

Suppose now that  $f$  is **continuously differentiable** on any open subset of  $\mathbb{T}$ . That is, if  $\mathbb{S} \subseteq \mathbb{T}$  is open, then the **gradient**

$$\left[ \frac{\partial f}{\partial \theta}(\theta^*) \right]^\top = \left( \frac{\partial f}{\partial \theta_1}(\theta^*), \dots, \frac{\partial f}{\partial \theta_d}(\theta^*) \right)$$

exists for every  $\theta^* \in \mathbb{S}$ .

We say that  $\theta^* \in \mathbb{T}$  is a **stationary point** of  $f$ , if it satisfies the equation

$$\frac{\partial f}{\partial \theta}(\theta^*) = \mathbf{0},$$

where  $\mathbf{0}$  is a matrix or vector of zeros of appropriate dimensionality.

If  $\theta^*$  is a local maximum or local minimum of  $f$  in some open subset of  $\mathbb{T}$ , and if  $f$  is continuously differentiable, then it is *necessary* that  $\theta^*$  is also a stationary point of  $f$ .

## A second existence theorem

In a metric space, we say that  $\theta^*$  is an **interior point** of a set  $\mathbb{T}$  if there exists an  $\epsilon > 0$ , such that

$$\mathbb{T} \cap N_\epsilon(\theta^*) = N_\epsilon(\theta^*).$$

We then say that  $\theta^*$  is an **boundary point** of  $\mathbb{T}$  if for all  $\epsilon > 0$ ,

$$\mathbb{T} \cap N_\epsilon(\theta^*) \neq N_\epsilon(\theta^*).$$

We can extend the Weierstrass optimality theorem, as follows. If  $\mathbb{T} \subset \mathbb{R}^d$  is compact and if  $f(\cdot) : \mathbb{T} \rightarrow \mathbb{R}$  is continuously differentiable, then there exists  $\psi, v \in \mathbb{T}$ , such that

$$f(\psi) \leq f(\theta) \leq f(v),$$

for all  $\theta \in \mathbb{T}$ . Furthermore, if  $\psi$  or  $v$  are interior points, then they must be stationary points of  $f$ . If  $\psi$  or  $v$  are not stationary points, then they must be boundary points of  $f$ .

# Convex sets

A set  $\mathbb{T}$  is said to be **convex** if for all  $\psi, v \in \mathbb{T}$ , and for any  $\lambda \in [0, 1]$ , we have

$$\theta = \lambda\psi + (1 - \lambda)v \in \mathbb{T}.$$

We say that  $\theta$  is a *convex combination* of the two points  $\psi$  and  $v$ .

Some examples of convex sets in  $\mathbb{R}^d$  include:

- The real space,  $\mathbb{R}^d$ , itself.
- Any *half space*  $\{\theta \in \mathbb{R}^d : a^\top \theta < b\}$ , for  $a \in \mathbb{R}^d$  and  $b \in \mathbb{R}$ .
- Any *hyperplane*  $\{\theta \in \mathbb{R}^d : a^\top \theta = b\}$ , for  $a \in \mathbb{R}^d$  and  $b \in \mathbb{R}$ .
- Any ball  $\{\theta \in \mathbb{R}^d : \|\theta - \theta^*\|_p < \epsilon\}$ , for  $\theta^* \in \mathbb{R}^d$ ,  $\epsilon > 0$ , and  $p \geq 1$ .
- The intersection of any number of convex sets.

# Convex functions

We say that the function  $f(\cdot) : \mathbb{T} \rightarrow \mathbb{R}$  is **convex**, over a convex domain  $\mathbb{T}$ , if for all  $\psi, v \in \mathbb{T}$ , and for any  $\lambda \in [0, 1]$ , we have

$$f(\lambda\psi + (1 - \lambda)v) \leq \lambda f(\psi) + (1 - \lambda)f(v).$$

The function  $f$  is said to be **strictly convex** if we change the symbol  $\leq$  to the symbol  $<$ .

We then define a **concave** or **strictly concave** function by reversing the inequalities in the previous definitions.

It is not difficult to show that if  $f$  is a convex function, then  $-f$  is a concave function, and *vice versa*.

# The Hessian matrix and positive definiteness

Suppose that  $f(\cdot) : \mathbb{T} \rightarrow \mathbb{R}$  is now twice continuously differentiable over the convex domain  $\mathbb{T}$ .

Write the **Hessian** matrix of  $f$  at  $\theta^* \in \mathbb{T}$  as

$$\frac{\partial^2 f}{\partial \theta \partial \theta^\top}(\theta^*) = \begin{bmatrix} \frac{\partial^2 f}{\partial \theta_1^2}(\theta^*) & \frac{\partial^2 f}{\partial \theta_1 \partial \theta_2}(\theta^*) & \cdots & \frac{\partial^2 f}{\partial \theta_1 \partial \theta_d}(\theta^*) \\ \frac{\partial^2 f}{\partial \theta_1 \partial \theta_2}(\theta^*) & \frac{\partial^2 f}{\partial \theta_2^2}(\theta^*) & \cdots & \frac{\partial^2 f}{\partial \theta_2 \partial \theta_d}(\theta^*) \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial \theta_1 \partial \theta_d}(\theta^*) & \frac{\partial^2 f}{\partial \theta_2 \partial \theta_d}(\theta^*) & \cdots & \frac{\partial^2 f}{\partial \theta_d^2}(\theta^*) \end{bmatrix}.$$

We say that a  $d \times d$  matrix  $\mathbf{A}$  is **positive definite** if for any  $\theta \in \mathbb{R}^d \setminus \{0\}$ ,  $\theta^\top \mathbf{A} \theta > 0$ . A **positive semidefinite** matrix is defined by replacing the symbol  $>$  by  $\geq$ . The definition for **negative definite** and **negative semidefinite** matrices are obtained by reversing the inequalities.

# First and second order conditions

A continuously differentiable function  $f(\cdot) : \mathbb{T} \rightarrow \mathbb{R}$  is convex, over a convex domain  $\mathbb{T}$ , if for any  $\psi, v \in \mathbb{T}$ , such that  $\psi \neq v$ , we have

$$f(\psi) \geq f(v) + \left[ \frac{\partial f}{\partial \theta}(v) \right]^\top (\psi - v).$$

We obtain strict convexity by replacing the symbol  $\geq$  by  $>$ . First-order conditions for concavity and strict concavity are obtained by reversing the inequalities.

If  $f$  is twice continuously differentiable over the convex domain  $\mathbb{T}$ , then it is convex if its Hessian is positive semidefinite, for every  $\theta^* \in \mathbb{T}$ . It is strictly convex if the Hessian is positive definite.

The definitions for concavity of a twice continuously differentiable function can be obtained by replacing the word *positive* by the word *negative*.

## A third existence theorem

If  $f(\cdot) : \mathbb{T} \rightarrow \mathbb{R}$  is convex, over a convex domain  $\mathbb{T}$ , then a point  $\theta^* \in \mathbb{T}$  is a global minimizer if and only if

$$\left[ \frac{\partial f}{\partial \theta}(\theta^*) \right]^\top (\psi - \theta^*) \geq 0,$$

for every  $\psi \in \mathbb{T}$ .

Furthermore, if  $\theta^* \in \mathbb{T}$  is a local minimizer of  $f$ , then  $\theta^*$  is also a global minimizer of  $f$ . If  $f$  is strictly convex then it has at most one global minimizer.

Restatements of the results in terms of concave functions and maxima can be obtained by reversing the inequality.



# The subdifferential

We now only assume that  $f(\cdot) : \mathbb{T} \rightarrow \mathbb{R}$  is convex. Denote the **subdifferential** of  $f$  at the point  $\theta^* \in \mathbb{T}$  by  $\partial f(\theta^*)$ , where

$$\partial f(\theta^*) = \left\{ v \in \mathbb{R}^d : f(\theta) \geq f(\theta^*) + v^\top (\theta - \theta^*), \text{ for all } \theta \in \mathbb{T} \right\}.$$

When  $f$  is differentiable,

$$\partial f(\theta^*) = \{(\partial f / \partial \theta)(\theta^*)\}.$$

Using the notion of the subdifferential, we have the result that  $f$  has a global minimizer at  $\theta^*$  if and only if

$$0 \in \partial f(\theta^*).$$

Notice, in the case of continuously differentiable  $f$ , that this condition reduces to

$$\frac{\partial f}{\partial \theta}(\theta^*) = 0.$$

# Linear regression

Suppose that we observe responses  $y_1, \dots, y_n \in \mathbb{R}$  with companion covariates  $\mathbf{x}_1, \dots, \mathbf{x}_n \in \mathbb{R}^d$ .

We wish to explain the relationship between any arbitrary  $y \in \mathbb{R}$  and  $\mathbf{x} \in \mathbb{R}^d$  via a hyperplane  $\alpha + \beta^\top \mathbf{x}$ , such that

$$y \approx \alpha + \beta^\top \mathbf{x},$$

in some sense.

The determination of the parameter  $\boldsymbol{\theta}^\top = (\alpha, \beta^\top) \in \mathbb{R}^{d+1}$  is known as the **linear regression** problem and can be solved in a number of ways.

We will firstly consider the method of *ridge-regularized least squares*, as proposed by Hoerl and Kennard (1970), where the parameter  $\boldsymbol{\theta}$  is obtained by solving the problem

$$\min_{\boldsymbol{\theta}=(\alpha,\beta)\in\mathbb{R}^{d+1}} \frac{1}{n} \sum_{i=1}^n |y_i - \alpha - \beta^\top \mathbf{x}_i|_2^2 + \lambda \sum_{j=1}^d |\beta_j|_2^2.$$

# Matrix notation

Write  $\bar{\mathbf{x}}_i^\top = (1, \mathbf{x}_i)$  and

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

where  $\mathbf{I}$  is the identity matrix of appropriate dimensionality, in order to obtain the expression

$$\begin{aligned} f(\boldsymbol{\theta}) &= \frac{1}{n} \sum_{i=1}^n \left| y_i - \alpha - \boldsymbol{\beta}^\top \mathbf{x}_i \right|_2^2 + \lambda \sum_{j=1}^d |\beta_j|_2^2 \\ &= \frac{1}{n} \sum_{i=1}^n \left( y_i - \boldsymbol{\theta}^\top \bar{\mathbf{x}}_i \right)^2 + \lambda \boldsymbol{\theta}^\top \bar{\mathbf{I}} \boldsymbol{\theta}. \end{aligned}$$

If we further write  $\mathbf{y}^\top = (y_1, \dots, y_n)$  and let  $\mathbf{X}$  be an  $n \times d$  matrix with  $i$ th row  $\bar{\mathbf{x}}_i^\top$ , then we can further write

$$f(\boldsymbol{\theta}) = \frac{1}{n} (\mathbf{y} - \mathbf{X}\boldsymbol{\theta})^\top (\mathbf{y} - \mathbf{X}\boldsymbol{\theta}) + \lambda \boldsymbol{\theta}^\top \bar{\mathbf{I}} \boldsymbol{\theta}.$$

# Solving the first order condition

We note that  $f$  is continuously differentiable in  $\theta$ . Using the rules of matrix differentiation from the *Matrix Cookbook* of Petersen and Pedersen (2012), we can write the gradient at any point  $\theta$  as

$$\frac{\partial f}{\partial \theta}(\theta) = -\frac{2}{n} \mathbf{X}^\top (\mathbf{y} - \mathbf{X}\theta) + 2\lambda \bar{\mathbf{I}}\theta,$$

which we can use to solve for a stationary point  $\theta^*$  that satisfies

$$\frac{\partial f}{\partial \theta}(\theta^*) = \mathbf{0}.$$

By solving the first order condition, we obtain the stationary point

$$\theta^* = (\mathbf{X}^\top \mathbf{X} + n\lambda \bar{\mathbf{I}})^{-1} \mathbf{X}^\top \mathbf{y}.$$

# Rules of convexity

Assume that  $\theta \in \mathbb{T}$ , where  $\mathbb{T} \subseteq \mathbb{R}^d$  is convex. We can use the following rules for determining convexity (see Boyd and Vandenberghe (2004)):

- The (**affine**) function  $f(\theta) = \mathbf{a}^\top \theta + b$  for  $\mathbf{a} \in \mathbb{R}^d$  and  $b \in \mathbb{R}$ , is convex.
- The function  $f(\theta) = \theta^2$  is convex.
- If  $g(\cdot) : \mathbb{T} \rightarrow \mathbb{R}$  is affine and  $h(\cdot) : \mathbb{R} \rightarrow \mathbb{R}$  is convex, then  $f(\theta) = h(g(\theta))$  is convex.
- Positively weighted sums of convex functions is convex.

# Checking convexity

Recall that

$$f(\theta) = \frac{1}{n} \sum_{i=1}^n \left| y_i - \alpha - \beta^\top \mathbf{x}_i \right|_2^2 + \lambda \sum_{j=1}^d |\beta_j|_2^2.$$

By our third existence theorem, we can prove that  $\theta^*$  is a global minimizer if we can demonstrate that the objective function  $f$  is convex, in  $\theta$ .

1. For each  $i$ , we know that  $y_i - \alpha - \beta^\top \mathbf{x}_i$  is affine, and thus convex.
2. Since  $|\cdot|_2^2 = (\cdot)^2$ , it is convex.
3. The affine compositions  $\left| y_i - \alpha - \beta^\top \mathbf{x}_i \right|_2^2$  and  $|\beta_j|_2^2$  are convex, for each  $i$  and  $j$ .
4. Since,  $f$  is a positively weighted sum of convex functions, it is also convex.

We have therefore demonstrated that  $\theta^*$  is a global minimizer of  $f$ .

# Robust ridge regression

Suppose now that we wish to solve the linear regression problem using a measurement of loss between each  $y_i$  and  $\mathbf{x}_i$  that replaces the  $\ell_2$  loss by an  $\ell_p$  loss, where  $p \in [1, 2)$ . In particular, we are interested in the case where  $p = 1$  (*ridge regularized least-absolute deviation*).

Thus, we are interested in solving the problem

$$\min_{\theta=(\alpha,\beta)\in\mathbb{R}^{d+1}} f(\theta) = \frac{1}{n} \sum_{i=1}^n \left| y_i - \alpha - \beta^\top \mathbf{x}_i \right|_p^p + \lambda \sum_{j=1}^d |\beta_j|_2^2.$$

Unfortunately,  $f$  is no longer continuously differentiable, and thus we require an alternative approach to what we have used, previously.

# The MM algorithm



# Difficulties arising in optimization

Suppose that  $f(\cdot) : \mathbb{T} \rightarrow \mathbb{R}$  is a difficult function to manipulate. We are interested in two particular types of difficulties:

1. The function  $f$  is not differentiable.
2. The function  $f$  is differentiable, but the solution to the first order condition

$$\frac{\partial f}{\partial \theta}(\theta^*) = \mathbf{0},$$

does not exist in closed form.

In such cases, we can operate on *surrogates* of  $f$  instead of operating on  $f$ , directly.

# Majorization and minorization

Let  $\psi, \theta \in \mathbb{T}$  and suppose that we wish to approximate the behavior of  $f$ , evaluated at any  $\psi \in \mathbb{T}$ .

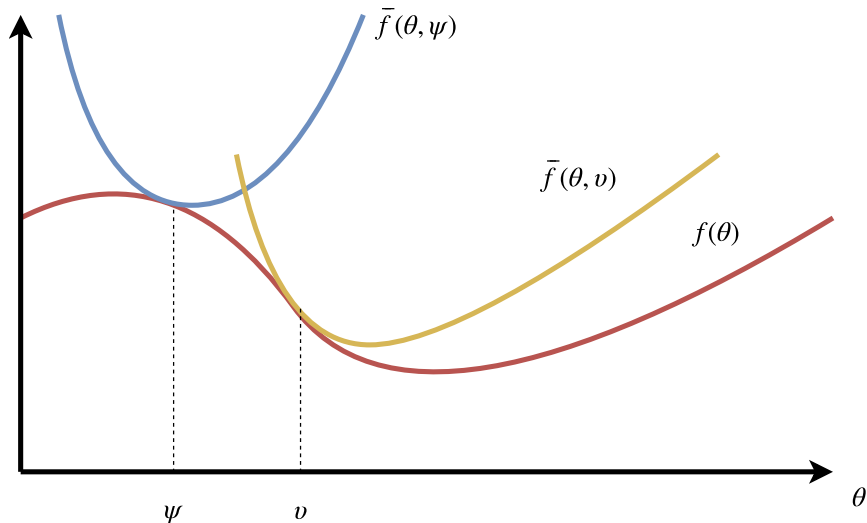
Introduce the function  $\bar{f}(\cdot, \cdot) : \mathbb{T} \times \mathbb{T} \rightarrow \mathbb{R}$ , and assume that  $\bar{f}$  satisfies the properties:

1. For any  $\theta \in \mathbb{T}$ ,  $\bar{f}(\theta, \theta) = f(\theta)$ .
2. For any  $\psi \neq \theta$ ,  $\bar{f}(\theta, \psi) \geq f(\theta)$ .

We call such a function a **majorizer** of  $f$ , and for any fixed  $\psi$ , we say that  $\bar{f}(\cdot, \psi) : \mathbb{T} \rightarrow \mathbb{R}$  **majorizes**  $f$ , at  $\psi$ .

The definition for a **minorizer** and the process of **minorization** can be obtained by reversing the inequality in the second condition.

# A visualization of the majorization process



**Figure 6:** Example of majorizers of an arbitrary function.

# The MM algorithm

Suppose that we wish to solve the minimization problem

$$\min_{\theta \in \mathbb{T}} f(\theta).$$

Let  $\theta^{(0)} \in \mathbb{T}$  be some **initialization** or *guess* of the solution to the problem. The **majorization-minimization (MM) algorithm** can be defined as follows. Let  $\theta^{(r)}$  be the  $r$ th iterate, obtained by the MM algorithm. We obtain this  $r$ th iterate by via the scheme

$$\theta^{(r)} \in \left\{ \theta^* \in \mathbb{T} : \bar{f}(\theta^*, \theta^{(r-1)}) = \min_{\theta \in \mathbb{T}} \bar{f}(\theta, \theta^{(r-1)}) \right\}.$$

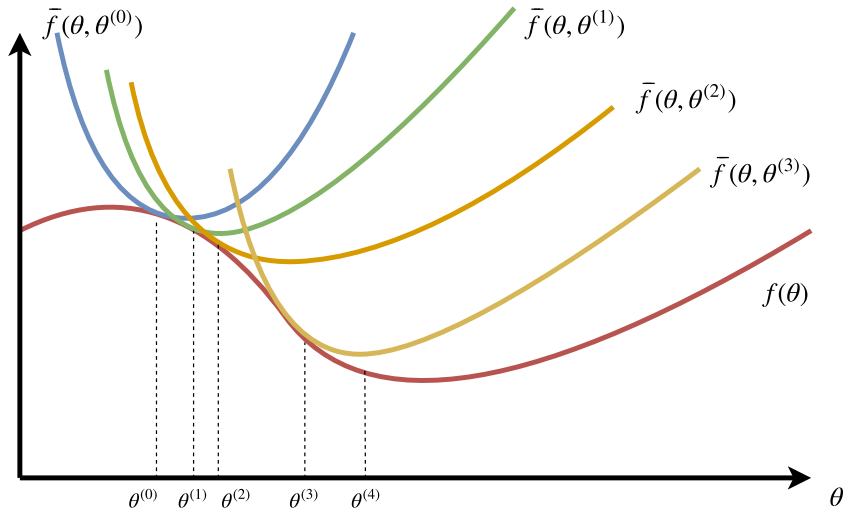
Alternatively, we can define the **minorization-maximization (MM) algorithm** for solving the problem

$$\max_{\theta \in \mathbb{T}} f(\theta),$$

via the scheme

$$\theta^{(r)} \in \left\{ \theta^* \in \mathbb{T} : \bar{f}(\theta^*, \theta^{(r-1)}) = \max_{\theta \in \mathbb{T}} \bar{f}(\theta, \theta^{(r-1)}) \right\}.$$

# Illustration of the MM algorithm



**Figure 7:** Four steps of an MM algorithm.

# The descent property

Let  $\theta^{(r)}$  and  $\theta^{(r+1)}$  be two consecutive iterates of the MM algorithm, and recall that a majorizer  $\bar{f}$  of  $f$  has the properties:

1. For any  $\theta \in \mathbb{T}$ ,  $\bar{f}(\theta, \theta) = f(\theta)$ .
2. For any  $\psi \neq \theta$ ,  $\bar{f}(\theta, \psi) \geq f(\theta)$ .

By the first property, we have the equality

$$f(\theta^{(r)}) = \bar{f}(\theta^{(r)}, \theta^{(r)}).$$

Since  $\theta^{(r+1)}$  minimizes  $\bar{f}(\cdot, \theta^{(r)})$ , we have

$$\bar{f}(\theta^{(r)}, \theta^{(r)}) \geq \bar{f}(\theta^{(r+1)}, \theta^{(r)}).$$

The second property then tells us that

$$\bar{f}(\theta^{(r+1)}, \theta^{(r)}) \geq f(\theta^{(r+1)}),$$

and hence, for any  $r \in \mathbb{N}$ ,

$$f(\theta^{(r)}) \geq f(\theta^{(r+1)}).$$

# The directional derivative

For convex  $\mathbb{T} \subseteq \mathbb{R}^d$ , and continuous function  $f$ , we say that  $f'(\cdot; \delta) : \mathbb{T} \rightarrow \mathbb{R}$  is the **directional derivative** of  $f$ , in the direction of  $\delta \in \mathbb{R}^d$ , and we write

$$f'(\theta; \delta) = \lim_{\lambda \downarrow 0} \frac{f(\theta + \lambda\delta) - f(\theta)}{\lambda}.$$

If  $f$  is differentiable, then

$$f'(\theta; \delta) = \delta^\top \frac{\partial f}{\partial \theta}(\theta).$$

For a *minimization problem*, we define a **stationary point**  $\theta^* \in \mathbb{T}$ , in an equivalent manner to the condition of the *third existence theorem*, by the condition

$$f'(\theta; \delta) \geq 0,$$

for all  $\theta + \delta \in \mathbb{T}$ . We define a stationary point for a *maximization problem* by reversing the inequality, above.

## Some more technicalities

Define the (Euclidean) distance from a point  $\theta^* \in \mathbb{T}$  to a set  $\mathbb{S} \subseteq \mathbb{T}$  by

$$\Delta(\theta^*, \mathbb{S}) = \inf_{\theta \in \mathbb{S}} \|\theta^* - \theta\|.$$

For a sequence  $\{\mathbf{a}_r\} = \mathbf{a}_1, \mathbf{a}_2, \dots \in \mathbb{R}^d$ , indexed by  $r \in \mathbb{N}$ , we say that  $\mathbf{a}$  is a **limit point** if for every  $\epsilon > 0$ , there are infinitely many  $r \in \mathbb{N}$ , such that

$$\mathbf{a}_r \in N_\epsilon(\mathbf{a}).$$

Thus the idea of limit points generalizes the idea of a **limit**, where we define  $\mathbf{a}$  to be a limit if for every  $\epsilon > 0$ , there exists a  $R_\epsilon > 0$  such that for all  $r > R_\epsilon$ ,

$$\mathbf{a}_r \in N_\epsilon(\mathbf{a}).$$



# A first convergence result

Make the following assumptions:

1.  $\bar{f}$  is a majorizer of the objective function  $f$ .
2. The majorizer  $\bar{f}(\boldsymbol{\theta}, \boldsymbol{\psi})$  is continuous in  $(\boldsymbol{\theta}^\top, \boldsymbol{\psi}^\top) \in \mathbb{T} \times \mathbb{T}$ .
3. For all  $\boldsymbol{\psi}$  and  $\boldsymbol{\delta}$ , such that  $\boldsymbol{\psi} + \boldsymbol{\delta} \in \mathbb{T}$ , we have

$$f'(\boldsymbol{\psi}; \boldsymbol{\delta}) = \bar{f}'(\boldsymbol{\theta}, \boldsymbol{\psi}; \boldsymbol{\delta})(\boldsymbol{\psi}).$$

Assumption 3 is satisfied if in addition to Assumptions 1 and 2, we also assume that  $f(\boldsymbol{\theta})$  is differentiable in  $\boldsymbol{\theta}$  and  $\bar{f}(\boldsymbol{\theta}, \boldsymbol{\psi})$  is continuous in  $(\boldsymbol{\theta}^\top, \boldsymbol{\psi}^\top)$ .

The first convergence theorem of Razaviyayn, Hong, and Luo (2013) states that: if Assumptions 1–3 are fulfilled, and if  $\boldsymbol{\theta}^{(\infty)}$  is a *limit point* of the *majorization-minimization* (MM) algorithm sequence  $\{\boldsymbol{\theta}^{(r)}\}$ , then  $\boldsymbol{\theta}^{(\infty)}$  is a *minimization stationary point* of  $f$ .

# Proof of first result

In a *metric space* a point  $\mathbf{a}$  is a *limit point* of  $\{\mathbf{a}_r\}$  if and only if it is a *limit* of some **subsequence** of  $\{\mathbf{a}_r\}$ .

We shall construct a proof using this idea of subsequences in mind. Let  $\{\boldsymbol{\theta}^{(r_s)}\}$  be a subsequence of  $\{\boldsymbol{\theta}^{(r)}\}$ , indexed by  $s \in \mathbb{N}$ , such that  $\lim_{s \rightarrow \infty} \boldsymbol{\theta}^{(r_s)} = \boldsymbol{\theta}^{(\infty)}$ , where  $\boldsymbol{\theta}^{(\infty)}$  is a limit point. Here  $r_s = r_1, r_2, \dots \in \mathbb{N}$  is an **increasing** sequence.

From the *descent property* and properties of the *majorizer*, for all  $\boldsymbol{\theta} \in \mathbb{T}$ , we have

$$\begin{aligned}\bar{f}(\boldsymbol{\theta}^{(r_{s+1})}, \boldsymbol{\theta}^{(r_{s+1})}) &= f(\boldsymbol{\theta}^{(r_{s+1})}) \leq f(\boldsymbol{\theta}^{(r_s+1)}) \\ &\leq \bar{f}(\boldsymbol{\theta}^{(r_s+1)}, \boldsymbol{\theta}^{(r_s)}) \leq \bar{f}(\boldsymbol{\theta}, \boldsymbol{\theta}^{(r_s)}).\end{aligned}$$

## Proof of first result (2)

By continuity, we can take the limit of the left and right hand side of the inequality, as  $s \rightarrow \infty$  to obtain, for all  $\theta \in \mathbb{T}$ ,

$$\bar{f}(\theta^{(\infty)}, \theta^{(\infty)}) \leq \bar{f}(\theta, \theta^{(\infty)}),$$

which implies

$$\begin{aligned}\bar{f}'(\theta, \theta^{(\infty)}; \delta)(\theta^{(\infty)}) &= \lim_{h \downarrow 0} \frac{\bar{f}(\theta^{(\infty)} + h\delta, \theta^{(\infty)}) - \bar{f}(\theta^{(\infty)}, \theta^{(\infty)})}{h} \\ &\geq 0\end{aligned}$$

By Assumption 3, we have

$$f'(\theta^{(\infty)}; \delta) = \bar{f}'(\theta, \theta^{(\infty)}; \delta)(\theta^{(\infty)}) \geq 0,$$

which completes the proof.

## A second convergence result

Define the **level set** of  $f(\cdot) : \mathbb{T} \rightarrow \mathbb{R}$  for given any point  $\theta^*$  as

$$\mathbb{T}(\theta^*) = \{\theta : f(\theta) \leq f(\theta^*)\}.$$

For any sequence of *majorization-minimization* (MM) algorithm sequence  $\{\theta^{(r)}\}$ , starting from some *initial guess*  $\theta^{(0)}$ , if  $\mathbb{T}(\theta^{(0)})$  is *compact* and if Assumptions 1–3 are fulfilled, then the sequence  $\{\theta^{(r)}\}$  satisfies the limit

$$\lim_{r \rightarrow \infty} \Delta(\theta^{(r)}, \mathbb{T}^*) = 0,$$

where  $\mathbb{T}^*$  is the set of stationary points

$$\{\theta^* \in \mathbb{T} : f'(\theta^*; \delta) \geq 0, \text{ for all } \theta^* + \delta \in \mathbb{T}\}.$$

## Proof of second result

By contradiction, suppose that there exists some subsequence  $\{\theta^{(r_s)}\}$ , indexed by  $s \in \mathbb{N}$ , such that

$$\Delta(\theta^{(r_s)}, \mathbb{T}^*) \geq c,$$

for some constant  $c > 0$ , for all indices  $s$ .

Since  $\mathbb{T}(\theta^{(0)})$  is *compact*,  $\{\theta^{(r_s)}\}$  must have its limit point  $\theta^{(\infty)} \in \{\theta^{(r_s)}\}$ , which implies that

$$\Delta(\theta^{(\infty)}, \mathbb{T}^*) \geq c.$$

But  $\theta^{(\infty)} \in \mathbb{T}^*$ , by the *first convergence result*.

# Catalog of majorizers

In Lange (2013), the following are listed as the most useful and fundamental majorizers.

1. The **Jensen's inequality** majorizer.
2. The **De Pierro** majorizer.
3. The **linear upper bound** majorizer.
4. The **quadratic upper bound** majorizer.

In the following descriptions, you can obtain *minorizers* by reversing inequalities, and changing the adjectives *positive* to *negative*, regarding the *definiteness* of matrices.

# Jensen's inequality

Let  $g : (0, \infty) \rightarrow \mathbb{R}$  be a convex function. Assume that  $\mathbf{w}^\top = (w_1, \dots, w_d) \in \mathbb{T}$ ,  $\boldsymbol{\theta}, \boldsymbol{\psi} \in \mathbb{T}$ , where  $\mathbb{T} = (0, \infty)^d$ .

Then, we can *majorize* the function

$$f(\boldsymbol{\theta}) = g(\mathbf{w}^\top \boldsymbol{\theta})$$

at  $\boldsymbol{\psi}$ , via the *majorizer*

$$\bar{f}(\boldsymbol{\theta}, \boldsymbol{\psi}) = \sum_{j=1}^d \frac{w_j \psi_j}{\mathbf{w}^\top \boldsymbol{\psi}} g\left(\frac{\mathbf{w}^\top \boldsymbol{\psi}}{\psi_j} \theta_j\right),$$

where  $\boldsymbol{\theta}^\top = (\theta_1, \dots, \theta_d)$  and  $\boldsymbol{\psi}^\top = (\psi_1, \dots, \psi_d)$ .

**An Example:** when  $f(\boldsymbol{\theta}) = \log\left(\sum_{j=1}^d \theta_j\right)$ , we can use

$$\bar{f}(\boldsymbol{\theta}, \boldsymbol{\psi}) = \sum_{j=1}^d \frac{\psi_j}{\sum_{k=1}^d \psi_k} \log\left(\frac{\sum_{k=1}^d \psi_k}{\psi_j} \theta_j\right).$$

# De Pierro

As the name suggests, this majorizer was studied by De Pierro (1993) in the context of *positron emissions tomography*. It is a generalization of the previous majorizer.

Let  $g : \mathbb{R} \rightarrow \mathbb{R}$  be a convex function. Assume that  $\mathbf{w}^\top = (w_1, \dots, w_d) \in \mathbb{R}$  and  $\boldsymbol{\theta}, \boldsymbol{\psi} \in \mathbb{R}$ .

Then, we can *majorize* the function

$$f(\boldsymbol{\theta}) = g(\mathbf{w}^\top \boldsymbol{\theta})$$

at  $\boldsymbol{\psi}$ , via the *majorizer*,

$$\bar{f}(\boldsymbol{\theta}, \boldsymbol{\psi}) = \sum_{j=1}^d v_j g\left(\frac{w_j}{v_j}(\theta_j - \psi_j) + \mathbf{w}^\top \boldsymbol{\psi}\right),$$

where  $v_j \geq 0$ ,  $\sum_{j=1}^d v_j = 1$ , and  $v_j > 0$  whenever  $w_j \neq 0$ .



# Linear upper bound

Let  $g : \mathbb{T} \rightarrow \mathbb{R}$  be concave, where  $\mathbb{T} \subseteq \mathbb{R}^d$ , and let  $\theta, \psi \in \mathbb{T}$ . Then, we can majorize the function  $f(\theta) = g(\theta)$ , at  $\psi$ , via the majorizer,

$$\bar{f}(\theta, \psi) = g(\psi) + \frac{\partial g}{\partial \theta}(\psi)(\theta - \psi).$$

**An Example:** Consider that the function  $g(\theta) = \sqrt{\theta}$ . Since  $dg/d\theta = 1/(2\sqrt{\theta})$ , we can majorize  $f = g$ , at  $\psi \in (0, \infty)$  by

$$\bar{f}(\theta, \psi) = \sqrt{\psi} + \frac{1}{2\sqrt{\psi}}(\theta - \psi).$$

# Quadratic upper bound

Let  $g : \mathbb{T} \rightarrow \mathbb{R}$  be *convex*, where  $\mathbb{T} \subseteq \mathbb{R}^d$ , and let  $\theta, \psi \in \mathbb{T}$ .

Recall that we can write the *Hessian* matrix of  $g$ , at any point  $\theta$  as

$$\frac{\partial^2 g}{\partial \theta \partial \theta^\top}(\theta).$$

Suppose that we can find a *postive definite* matrix  $\mathbf{H} \in \mathbb{R}^{d \times d}$ , such that for all  $\theta \in \mathbb{T}$ ,

$$\mathbf{H} - \frac{\partial^2 g}{\partial \theta \partial \theta^\top}(\theta)$$

is *positive semidefinite*. Then, we can *majorize* the function  $f(\theta) = g(\theta)$ , at  $\psi$ , via the *majorizer*,

$$\bar{f}(\theta, \psi) = g(\psi) + \frac{\partial g}{\partial \theta}(\psi)(\theta - \psi) + \frac{1}{2}(\theta - \psi)^\top \mathbf{H}(\theta - \psi).$$

# Closure properties

The following operations preserve the *majorization* property.

- 1. Summation.** If  $g_1(\theta), \dots, g_m(\theta)$  are respectively majorized by  $\bar{g}_1(\theta; \psi), \dots, \bar{g}_m(\theta; \psi)$ , then  $f(\theta) = \sum_{k=1}^m g_k(\theta)$  is majorized by

$$\bar{f}(\theta; \psi) = \sum_{k=1}^m \bar{g}_k(\theta; \psi).$$

- 2. Non-negative product.** If  $g_1(\theta), \dots, g_m(\theta) \geq 0$  are respectively majorized by  $\bar{g}_1(\theta; \psi), \dots, \bar{g}_m(\theta; \psi)$ , then  $f(\theta) = \prod_{k=1}^m g_k(\theta)$  is majorized by

$$\bar{f}(\theta; \psi) = \prod_{k=1}^m \bar{g}_k(\theta; \psi).$$

- 3. Increasing composition.** If  $g(\theta)$  is majorized by  $\bar{g}(\theta, \psi)$ , and if  $h$  is *increasing*, then  $f(\theta) = h(g(\theta))$  is majorized by

$$\bar{f}(\theta, \psi) = h(\bar{g}(\theta, \psi)).$$

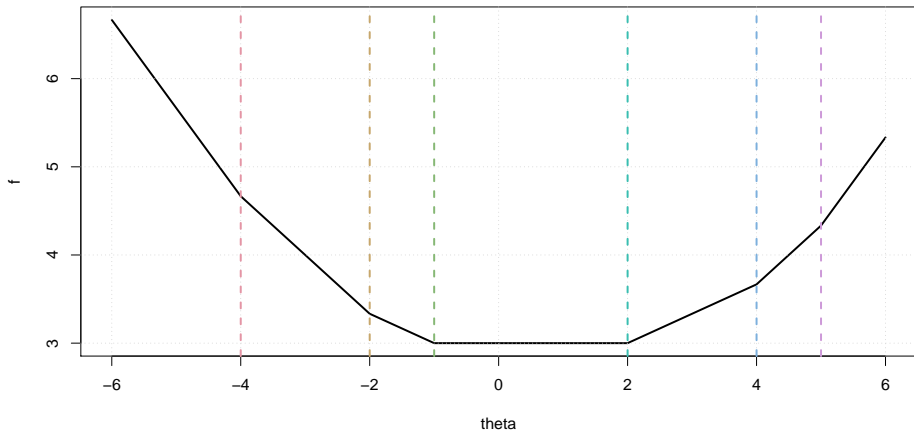
# A simple problem

Let  $y_1, \dots, y_n \in \mathbb{R}$  be a set of data and let  $\theta \in \mathbb{R}$  be a parameter of interest. Suppose that we wish to solve the **minimum absolute-deviation** problem

$$\min_{\theta \in \mathbb{R}} \left\{ f(\theta) = \frac{1}{n} \sum_{i=1}^n |y_i - \theta| \right\}.$$

Unfortunately  $f(\theta)$  is not differentiable, and thus we cannot solve for stationary points using the usual methods of calculus.

## Example instance



**Figure 8:** Example instance where the observations are -4, -2, -1, 2, 4, and 5.

# Using subdifferentials

Let  $f, g$  be convex. We have the following two facts about subdifferentials:

1. If  $a > 0$ , then  $\partial[af(\theta)] = a\partial f(\theta)$ .
2. We have  $\partial[f(\theta) + g(\theta)] = \partial f(\theta) + \partial g(\theta)$ .

Since the *absolute value* function is convex, we can apply the facts 1 and 2 to  $f(\theta) = \frac{1}{n} \sum_{i=1}^n |y_i - \theta|$ , to get:

$$\partial f(\theta) = \frac{1}{n} \sum_{i=1}^n \partial |y_i - \theta|.$$

We observe that  $\partial |y_i - \theta|$  equals  $\{1\}$  when  $y_i < \theta$ ,  $\{-1\}$  when  $y_i > \theta$ , and  $[-1, 1]$  when  $\theta = y_i$ .

Recall that a *stationary point* of the convex function can be obtained by finding a value of  $\theta^* \in \mathbb{R}$ , such that

$$0 \in \partial f(\theta^*).$$

## Solving for a stationary point

Unless all of the values of  $y_i$  are equal, we require a  $\theta^*$  such that there are equal numbers of  $\partial |y_i - \theta^*| = \{-1\}$  and  $\{1\}$  to make

$$0 \in \partial f(\theta^*) = \frac{1}{n} \sum_{i=1}^n \partial |y_i - \theta^*|.$$

We can achieve this by finding a  $\theta^*$  so that

$$\sum_{i=1}^n [y_i \leq \theta^*] = \sum_{i=1}^n [y_i \geq \theta^*].$$

Let  $y_{(1)} < y_{(2)} < \dots < y_{(n)}$ . For even  $n$ , we can set  $\theta^* \in [y_{(n/2)}, y_{(n/2+1)}]$  to get a suitable solution. When  $n$  is odd, we can set  $\theta^* = y_{(\lceil n/2 \rceil)}$ , where  $\lceil y \rceil = \min \{n \in \mathbb{Z} : n \geq y\}$  and  $\mathbb{Z}$  is the integers.

The solution  $\theta^*$  is the **median** of the data.

# A useful majorizer

Recall the majorizer

$$\bar{g}(v, \psi) = \sqrt{\psi} + \frac{1}{2\sqrt{\psi}} (v - \psi)$$

for  $g(v) = \sqrt{v}$ .

Set  $v = \theta^2$  and  $\psi = \theta^{(r-1)2}$  to obtain the majorizer

$$\bar{f}(\theta; \theta^{(r-1)}) = \sqrt{\theta^{(r-1)2}} + \frac{1}{2\sqrt{\theta^{(r-1)2}}} (\theta^2 - \theta^{(r-1)2}),$$

for the function  $f(\theta) = \sqrt{\theta^2} = |\theta|$ . This simplifies to

$$\bar{f}(\theta; \theta^{(r-1)}) = \frac{\theta^2}{2|\theta^{(r-1)}|} + \frac{1}{2} |\theta^{(r-1)}|.$$



# A majorizer for the median problem

We can substitute  $y_i - \theta$  for  $\theta$ , and  $y_i - \theta^{(r-1)}$  for  $\theta^{(r-1)}$  and use the *closure under summation* to obtain the majorizer

$$\bar{f}(\theta; \theta^{(r-1)}) = \frac{1}{2n} \sum_{i=1}^n \frac{(y_i - \theta)^2}{|y_i - \theta^{(r-1)}|} + \frac{1}{2n} \sum_{i=1}^n |y_i - \theta^{(r-1)}|,$$

for  $f(\theta) = (1/n) \sum_{i=1}^n |y_i - \theta|$ .

We notice that  $\bar{f}$  is differentiable and convex (since it is quadratic) and thus we only need to solve for a *stationary point* to obtain the  $r$ th iteration of the MM algorithm:

$$\theta^{(r)} \in \left\{ \theta^* \in \mathbb{T} : \bar{f}(\theta^*, \theta^{(r-1)}) = \min_{\theta \in \mathbb{T}} \bar{f}(\theta, \theta^{(r-1)}) \right\}.$$

# An MM algorithm for the median

Upon taking the derivative of  $\bar{f}$ , we obtain:

$$\frac{d\bar{f}(\cdot; \theta^{(r-1)})}{d\theta} = -\frac{1}{n} \sum_{i=1}^n \frac{y_i}{|y_i - \theta^{(r-1)}|} + \frac{\theta}{n} \sum_{i=1}^n \frac{1}{|y_i - \theta^{(r-1)}|}.$$

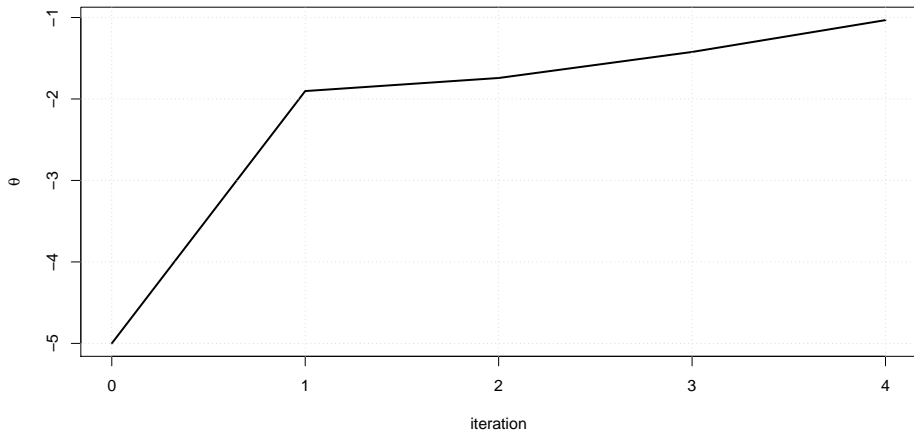
Solving for  $(d\bar{f}(\cdot; \theta^{(r-1)})/d\theta)(\theta^{(r)}) = 0$  then yields the MM algorithm iterations:

$$\theta^{(r)} = \left( \sum_{i=1}^n \frac{y_i}{|y_i - \theta^{(r-1)}|} \right) / \left( \sum_{i=1}^n \frac{1}{|y_i - \theta^{(r-1)}|} \right),$$

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

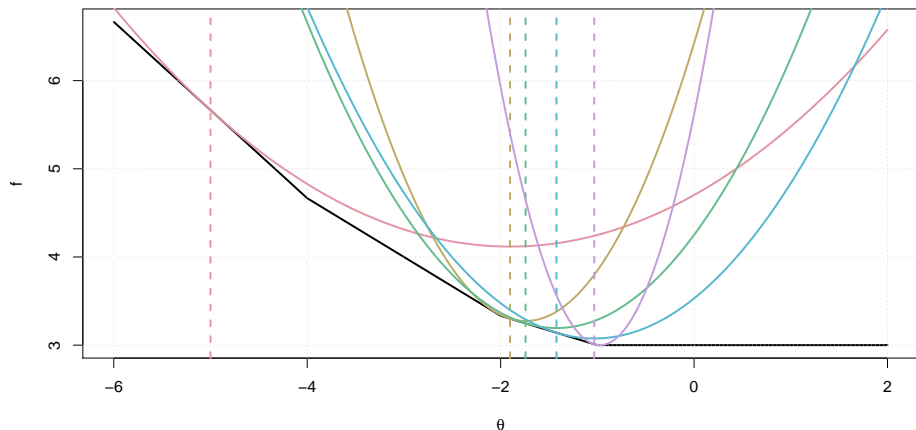
We have obtained an **iterative reweighting** algorithm for the computation of the *median*.

## Example output of the MM algorithm



**Figure 9:** Sequence of MM algorithm iterations for the computation of the median from observations -4, -2, -1, 2, 4, and 5.

# Visualization of the majorizers



**Figure 10:** Visualization of the majorizers after 5 steps of the algorithm.

# Regression problems

# Least-absolute deviation regression

As before,  $y_1, \dots, y_n \in \mathbb{R}$  are  $n \in \mathbb{N}$  observe responses, explained by their companion covariates  $\mathbf{x}_1, \dots, \mathbf{x}_n \in \mathbb{R}^d$ . We wish to explain any arbitrary  $y$  by its covariate  $\mathbf{x}$  via the relationship:

$$y \approx \alpha + \beta^\top \mathbf{x},$$

where  $\alpha \in \mathbb{R}$ ,  $\beta \in \mathbb{R}^d$ , and  $\boldsymbol{\theta}^\top = (\alpha, \beta^\top) \in \mathbb{R}^{d+1}$ .

In the case of *least-absolute deviation regression*, we obtain an estimate of the vector  $\boldsymbol{\theta}$  by solving the optimization problem

$$\min_{\boldsymbol{\theta} \in \mathbb{R}^{d+1}} \left\{ f(\boldsymbol{\theta}) = \sum_{i=1}^n |y_i - \alpha - \beta^\top \mathbf{x}_i| = \sum_{i=1}^n |y_i - \boldsymbol{\theta}^\top \bar{\mathbf{x}}_i| \right\},$$

where  $\bar{\mathbf{x}}_i^\top = (1, \mathbf{x}_i^\top) \in \mathbb{R}^{d+1}$ , for each  $i$ .

## A solution

Recall that we can majorize  $g(v) = \sqrt{v}$  by

$$\bar{g}(v, \psi) = \frac{\sqrt{\psi}}{2} + \frac{v}{2\sqrt{\psi}},$$

using a *linear upper bound*.

Let  $\theta^{(r)}$  denote the  $r$ th iteration of the MM algorithm, as usual. Upon substitutions  $v = (y_i - \theta^\top \bar{\mathbf{x}}_i)^2$  and  $\psi = (y_i - \theta^{(r-1)\top} \bar{\mathbf{x}}_i)^2$ , and upon using closure under summation, we obtain the majorizer  $f(\theta)$  by

$$\begin{aligned} f(\theta, \theta^{(r-1)}) &= \frac{1}{2} \sum_{i=1}^n \sqrt{(y_i - \theta^{(r-1)\top} \bar{\mathbf{x}}_i)^2} + \frac{1}{2} \sum_{i=1}^n \frac{(y_i - \theta^\top \bar{\mathbf{x}}_i)^2}{\sqrt{(y_i - \theta^{(r-1)\top} \bar{\mathbf{x}}_i)^2}} \\ &= \frac{1}{2} \sum_{i=1}^n |y_i - \theta^{(r-1)\top} \bar{\mathbf{x}}_i| + \frac{1}{2} \sum_{i=1}^n \frac{(y_i - \theta^\top \bar{\mathbf{x}}_i)^2}{|y_i - \theta^{(r-1)\top} \bar{\mathbf{x}}_i|}. \end{aligned}$$

## A solution (2)

As in the case of *ridge-regularized least squares*, we write  $\mathbf{y}^\top = (y_1, \dots, y_n)$  and we put the  $\bar{\mathbf{x}}_i$  into the rows of  $\mathbf{X}$ . In addition, we let  $\mathbf{W}^{(r-1)} \in \mathbb{R}^{n \times n}$  be a diagonal matrix, where the  $i$ th diagonal element is equal to  $1 / |y_i - \boldsymbol{\theta}^{(r-1)\top} \bar{\mathbf{x}}_i|$ . We can rewrite  $\bar{f}(\boldsymbol{\theta}, \boldsymbol{\theta}^{(r-1)})$  as:

$$\bar{f}(\boldsymbol{\theta}, \boldsymbol{\theta}^{(r-1)}) = C + \frac{1}{2} (\mathbf{y} - \mathbf{X}\boldsymbol{\theta})^\top \mathbf{W}^{(r-1)} (\mathbf{y} - \mathbf{X}\boldsymbol{\theta}),$$

where  $C$  is a constant that does not depend on  $\boldsymbol{\theta}$ .

We observe that the majorizer is a quadratic and thus convex. We therefore can obtain our MM algorithm update by solving the *first order condition*

$$\frac{\partial \bar{f}(\cdot, \boldsymbol{\theta}^{(r-1)})}{\partial \boldsymbol{\theta}} = -\mathbf{X}^\top \mathbf{W}^{(r-1)} (\mathbf{y} - \mathbf{X}\boldsymbol{\theta}) = \mathbf{0}.$$

We obtain the MM algorithm defined via the *iteratively reweighted least-squares* scheme:

$$\boldsymbol{\theta}^{(r)} = \left( \mathbf{X}^\top \mathbf{W}^{(r-1)} \mathbf{X} \right)^{-1} \mathbf{X}^\top \mathbf{W}^{(r-1)} \mathbf{y}.$$



# A problem in the solution

Upon inspection of the majorizer

$$\bar{f}(\boldsymbol{\theta}, \boldsymbol{\theta}^{(r-1)}) = \frac{1}{2} \sum_{i=1}^n \sqrt{(y_i - \boldsymbol{\theta}^{(r-1)\top} \bar{\mathbf{x}}_i)^2} + \frac{1}{2} \sum_{i=1}^n \frac{(y_i - \boldsymbol{\theta}^\top \bar{\mathbf{x}}_i)^2}{\sqrt{(y_i - \boldsymbol{\theta}^{(r-1)\top} \bar{\mathbf{x}}_i)^2}},$$

we note that it is not defined, when  $y_i - \boldsymbol{\theta}^{(r-1)\top} \bar{\mathbf{x}}_i = 0$  for any  $i \in [n]$ .

For  $\epsilon > 0$ , we propose to approximate  $f(\boldsymbol{\theta}, \boldsymbol{\theta}^{(r-1)})$  by

$$\bar{f}_\epsilon(\boldsymbol{\theta}, \boldsymbol{\theta}^{(r-1)}) = \frac{1}{2} \sum_{i=1}^n \sqrt{(y_i - \boldsymbol{\theta}^{(r-1)\top} \bar{\mathbf{x}}_i)^2 + \epsilon} + \frac{1}{2} \sum_{i=1}^n \frac{(y_i - \boldsymbol{\theta}^\top \bar{\mathbf{x}}_i)^2 + \epsilon}{\sqrt{(y_i - \boldsymbol{\theta}^{(r-1)\top} \bar{\mathbf{x}}_i)^2 + \epsilon}},$$

which majorizes the *approximate objective function*

$$f_\epsilon(\boldsymbol{\theta}) = \sum_{i=1}^n \sqrt{(y_i - \boldsymbol{\theta}^\top \bar{\mathbf{x}}_i)^2 + \epsilon}.$$

## A problem in the solution (2)

Observe that we can similarly write

$$\bar{f}_\epsilon(\boldsymbol{\theta}, \boldsymbol{\theta}^{(r-1)}) = C + \frac{1}{2} (\mathbf{y} - \mathbf{X}\boldsymbol{\theta})^\top \mathbf{W}_\epsilon^{(r-1)} (\mathbf{y} - \mathbf{X}\boldsymbol{\theta}),$$

where we  $\mathbf{W}_\epsilon^{(r-1)}$  is a diagonal matrix with  $i$ th element  $1/\sqrt{(y_i - \boldsymbol{\theta}^{(r-1)\top} \bar{\mathbf{x}}_i)^2 + \epsilon}$ .

This is again a convex quadratic function, and we derive the MM algorithm iteration

$$\boldsymbol{\theta}^{(r)} = \left( \mathbf{X}^\top \mathbf{W}_\epsilon^{(r-1)} \mathbf{X} \right)^{-1} \mathbf{X}^\top \mathbf{W}_\epsilon^{(r-1)} \mathbf{y},$$

by solving the *first order condition*.

This solution can be combined with our *ridge regression* solution in order to solve the *robust ridge regression* problem, that was proposed earlier.

# The LASSO

The *LASSO* stands for *least absolute shrinkage and selection operator* and aims to estimate the parameter  $\boldsymbol{\theta}^\top = (\alpha, \boldsymbol{\beta}^\top)$  in estimating equation

$$y \approx \alpha + \boldsymbol{\beta}^\top \mathbf{x},$$

by solving the optimization problem

$$\min_{\boldsymbol{\theta} \in \mathbb{R}^{d+1}} \left\{ f(\boldsymbol{\theta}) = \sum_{i=1}^n \left( y_i - \boldsymbol{\theta}^\top \bar{\mathbf{x}}_i \right)^2 + \lambda \sum_{j=1}^d |\beta_j| \right\},$$

where  $\lambda > 0$ .

# A first solution

Using the *linear upper bound inequality*, and letting  $\theta^{(r)}$  denote the  $r$ th iteration of the MM algorithm, we can approximately majorize each of the absolute values in the objective function. That is, for each  $j \in [d]$ , we majorize the approximation  $g_\epsilon(\beta_j) = \sqrt{\beta_j^2 + \epsilon}$  of  $|\beta_j|$  by

$$\bar{g}_\epsilon(\beta_j, \beta_j^{(r-1)}) = \frac{\sqrt{\beta_j^{(r-1)2} + \epsilon}}{2} + \frac{\beta_j^2 + \epsilon}{2\sqrt{\beta_j^{(r-1)2} + \epsilon}}.$$

We note that the approximation is perfect when  $\epsilon = 0$ .

By the *summation closure*, we obtain the approximate majorizer

$$\bar{f}(\theta, \theta^{(r-1)}) = \sum_{i=1}^n (y_i - \theta^\top \bar{\mathbf{x}}_i)^2 + \lambda \sum_{j=1}^d \left( \frac{\sqrt{\beta_j^{(r-1)2} + \epsilon}}{2} + \frac{\beta_j^2}{2\sqrt{\beta_j^{(r-1)2} + \epsilon}} \right).$$

## A first solution (2)

Let  $\overline{\mathbf{W}}_{\epsilon}^{(r-1)} \in \mathbb{R}^{d+1}$  be a diagonal matrix with 0 in its first entry, and  $1/\sqrt{\beta_j^{(r-1)^2} + \epsilon}$  in the  $(j+1)$ th entry, where  $j \in [d]$ .

We can now write

$$\bar{f}(\boldsymbol{\theta}, \boldsymbol{\theta}^{(r-1)}) = C + (\mathbf{y} - \mathbf{X}\boldsymbol{\theta})^{\top} (\mathbf{y} - \mathbf{X}\boldsymbol{\theta}) + \frac{\lambda}{2} \boldsymbol{\theta}^{\top} \overline{\mathbf{W}}_{\epsilon}^{(r-1)} \boldsymbol{\theta},$$

which is a convex quadratic function.

Solving the *first order condition*

$$\frac{\partial \bar{f}(\cdot, \boldsymbol{\theta}^{(r-1)})}{\partial \boldsymbol{\theta}} = -2\mathbf{X}^{\top} (\mathbf{y} - \mathbf{X}\boldsymbol{\theta}) + \lambda \overline{\mathbf{W}}_{\epsilon}^{(r-1)} \boldsymbol{\theta} = \mathbf{0},$$

and obtain the MM algorithm

$$\boldsymbol{\theta}^{(r)} = \left( \mathbf{X}^{\top} \mathbf{X} + \frac{\lambda}{2} \overline{\mathbf{W}}_{\epsilon}^{(r-1)} \right)^{-1} \mathbf{X}^{\top} \mathbf{y}.$$

# The LASSO regularizer

The purpose of the **LASSO regularizer**, for the *regression coefficient*  $\beta$ ,

$$\rho_1(\beta) = \sum_{j=1}^d |\beta_j|,$$

is to serve as a **convex relaxation** of the so-called “ $\ell_0$  norm regularizer”

$$\rho_0(\beta) = \sum_{j=1}^d [\beta_j \neq 0].$$

By letting  $\rho_q(\beta) = \sum_{j=1}^d |\beta_j|^q$ , we observe that

$$\lim_{q \downarrow 0} \rho_q(\beta) = \rho_0(\beta).$$

The LASSO regularizer is the only **sparsity inducing** regularizer of the form  $\rho_q(\beta)$ .

# Visualization of convex relaxation

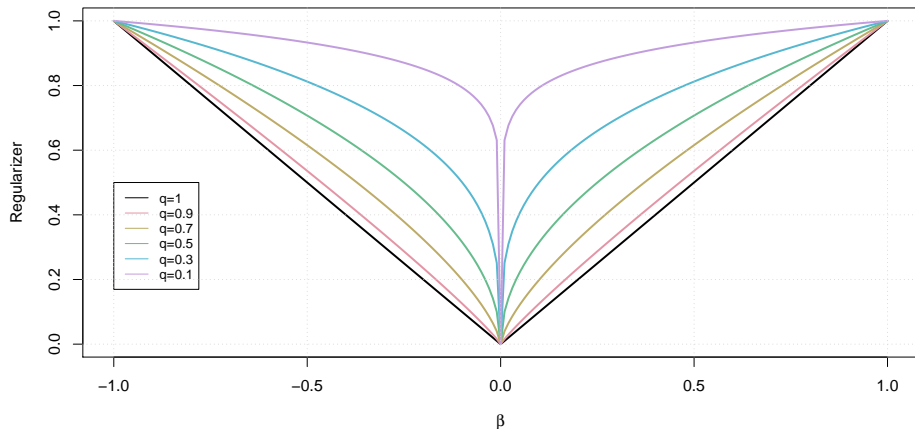


Figure 11: Visualization of various values of  $q$ .

# Alternative interpretation of regularization

Under sufficient *regularity conditions*, a minimization problem that is subject to the *regularizer*  $\lambda\rho(\beta)$ , for some  $\lambda > 0$  is equivalent to an unregularized problem under the **constraint**

$$\rho(\beta) \leq t,$$

for some  $t > 0$  (Bach et al. 2011).

That is, LASSO problem can be rewritten as

$$\min_{\theta=(\alpha,\beta)\in\mathbb{R}^{d+1}} \left\{ f(\theta) = \sum_{i=1}^n (y_i - \alpha - \beta^\top \mathbf{x}_i)^2 : \sum_{j=1}^d |\beta_j| \leq t \right\}.$$

This also implies that, for some  $t > 0$ , our approximate MM algorithm solves the *approximate LASSO* problem,

$$\min_{\theta=(\alpha,\beta)\in\mathbb{R}^{d+1}} \left\{ f(\theta) = \sum_{i=1}^n (y_i - \alpha - \beta^\top \mathbf{x}_i)^2 : \sum_{j=1}^d \sqrt{\beta_j^2 + \epsilon} \leq t \right\}.$$



# The Lagrange multiplier

When  $f(\cdot) : \mathbb{T} \rightarrow \mathbb{R}$  is differentiable and  $\rho(\cdot) : \mathbb{T} \rightarrow \mathbb{R}$  are both convex, under *regularity conditions*, **Lagrange multiplier theory** state that we can solve the problem

$$\min_{\theta \in \mathbb{T}} \{f(\theta) : \rho(\theta) \leq t\},$$

by solving the *first order condition* for the **Lagrangian**

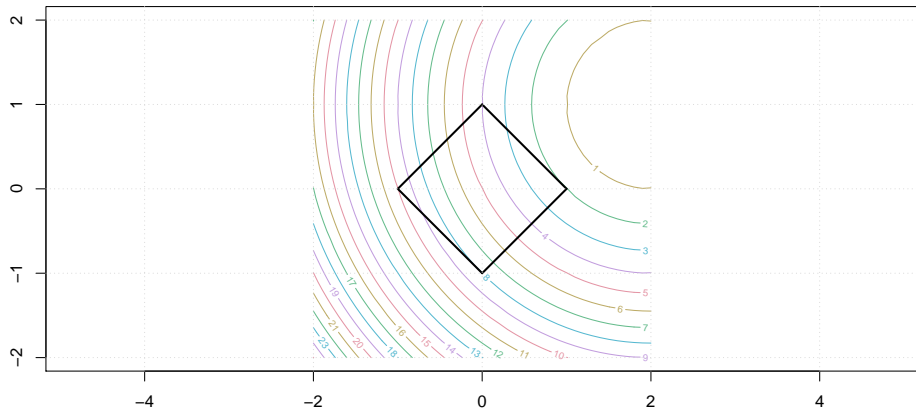
$$\mathcal{L}(\theta, \mu) = f(\theta) + \mu [\rho(\theta) - t],$$

where  $\mu \geq 0$ . That is, solving the simultaneous system:

$$\mathbf{0} \in \partial \mathcal{L}(\theta), \text{ and } \frac{\partial \mathcal{L}}{\partial \mu} = 0.$$

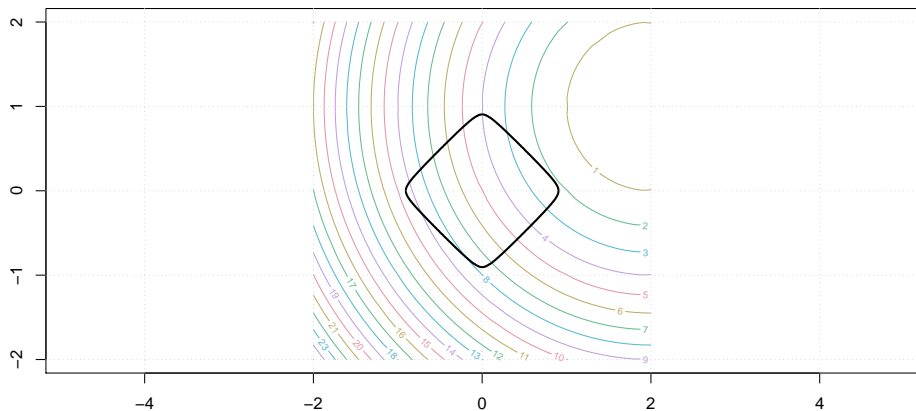
This implies that the solution to the problem  $\theta^* \in \mathbb{T}$  occurs along a contour of  $f$  that is *tangential* to the boundary of the constraint  $\rho(\theta) \leq t$ .

# Example of a LASSO regularization problem



**Figure 12:** Visualization of the LASSO constraint and functional contours.

## Example of an approximate LASSO regularization problem



**Figure 13:** Visualization of the approximate LASSO constraint and functional contours.

# A simplified problem

Consider the one-dimensional regularization problem

$$\min_{\theta \in \mathbb{R}} \left\{ f(\theta) = (z - \theta)^2 + \lambda |\theta| \right\},$$

where  $z \in \mathbb{R}$  and  $\lambda > 0$ .

Using the method of *subdifferentials*, we can solve the problem by finding a  $\theta^* \in \mathbb{R}$ , whereupon

$$0 \in \partial f(\theta^*) = -2(z - \theta^*) + \lambda \partial |\theta^*|,$$

where

$$\partial |\theta^*| = \begin{cases} -1 & \text{if } \theta^* < 0, \\ [-1, 1] & \text{if } \theta^* = 0, \\ 1 & \text{if } \theta^* > 0. \end{cases}$$

## A simplified problem (2)

We can solve for the root in the three cases, when  $\theta^* < 0$ ,  $\theta^* > 0$ , and  $\theta^* = 0$ .

In the  $\theta^* < 0$  case, we have

$$0 = -2z + 2\theta^* - \lambda \iff \theta^* = z + \frac{\lambda}{2}.$$

In the  $\theta^* > 0$  case, we have

$$0 = -2z + 2\theta^* + \lambda \iff \theta^* = z - \frac{\lambda}{2}.$$

In the  $\theta^* = 0$  case, we have

$$0 \in -2z + \lambda[-1, 1] \iff z \in \left[-\frac{\lambda}{2}, \frac{\lambda}{2}\right].$$

## A simplified problem (3)

Combining the three cases, we have the following *piecewise solution* to the problem

$$\theta^* = \begin{cases} z + \frac{\lambda}{2} & \text{if } z < -\lambda/2, \\ z - \frac{\lambda}{2} & \text{if } z > \lambda/2, \\ 0 & \text{if } z \in \left[-\frac{\lambda}{2}, \frac{\lambda}{2}\right], \end{cases}$$

to the problem

$$\min_{\theta \in \mathbb{R}} \left\{ f(\theta) = (z - \theta)^2 + \lambda |\theta| \right\}.$$

## A second solution to the LASSO problem

Recall that we seek a solution to the optimization problem

$$\min_{\boldsymbol{\theta} \in \mathbb{R}^{d+1}} \left\{ f(\boldsymbol{\theta}) = \sum_{i=1}^n \left( y_i - \boldsymbol{\theta}^\top \bar{\mathbf{x}}_i \right)^2 + \lambda \sum_{j=1}^d |\beta_j| \right\},$$

where  $\lambda > 0$ ,  $\boldsymbol{\theta}^\top = (\alpha, \boldsymbol{\beta}^\top)$ , and  $\bar{\mathbf{x}}_i^\top = (1, \mathbf{x}_i^\top)$ .

We seek a majorization scheme that will allow us to use the one dimensional solution to solve the problem.

## A second solution to the LASSO problem (2)

Setting  $v_j = 1/d$  for all  $j \in [d]$  in the *De Pierro* majorizer yields the majorizer

$$\bar{f}(\mathbf{v}, \boldsymbol{\psi}) = \sum_{j=1}^d \frac{1}{d} g\left(w_j d [v_j - \psi_j] + \boldsymbol{\psi}^\top \mathbf{w}\right),$$

of  $f(\mathbf{v}) = g(\mathbf{v}^\top \mathbf{w})$ , for convex functions  $g$ .

For each  $i$ , set  $g(\cdot) = (y_i - \cdot)^2$  and  $\mathbf{w} = \bar{\mathbf{x}}_i$ . Then, we can majorize  $f(\mathbf{v}) = g(\mathbf{v}^\top \bar{\mathbf{x}}_i) = (y_i - \mathbf{v}^\top \bar{\mathbf{x}}_i)^2$  by

$$\bar{f}(\mathbf{v}, \boldsymbol{\psi}) = \sum_{j=1}^{d+1} \frac{1}{d+1} \left( y_i - \bar{x}_{ij} (d+1) [v_j - \psi_j] - \boldsymbol{\psi}^\top \bar{\mathbf{x}}_i \right)^2,$$

where  $\bar{\mathbf{x}}_i^\top = (\bar{x}_{i1}, \dots, \bar{x}_{i,d+1}) = (1, x_{i1}, \dots, x_{id})$ .



## A second solution to the LASSO problem (3)

Now, make the substitutions  $\mathbf{v} = \boldsymbol{\theta}$  and  $\boldsymbol{\psi} = \boldsymbol{\theta}^{(r-1)}$ , and apply the summation rule to obtain the majorizer

$$\begin{aligned}\bar{f}(\boldsymbol{\theta}, \boldsymbol{\theta}^{(r-1)}) &= \sum_{j=1}^{d+1} \frac{1}{d+1} \sum_{i=1}^n \left( y_i - \bar{x}_{ij} (d+1) \left[ \theta_j - \theta_j^{(r-1)} \right] - \boldsymbol{\theta}^{(r-1)\top} \bar{\mathbf{x}}_i \right)^2 \\ &\quad + \lambda \sum_{j=2}^{d+1} |\theta_j|,\end{aligned}$$

for  $f(\boldsymbol{\theta})$ , where  $(\theta_1, \theta_2, \dots, \theta_{d+1}) = (\alpha, \beta_1, \dots, \beta_d)$ .

Since  $\bar{f}(\boldsymbol{\theta}, \boldsymbol{\theta}^{(r-1)})$  is additively separable in  $\boldsymbol{\theta}$ , we can now obtain an MM algorithm for the LASSO problem by *coordinate-wise* solving the *first order condition*

$$\mathbf{0} \in \partial f(\boldsymbol{\theta}, \boldsymbol{\theta}^{(r-1)}).$$

## A second solution to the LASSO problem (4)

Let  $\left[\partial \bar{f}(\boldsymbol{\theta}, \boldsymbol{\theta}^{(r-1)})\right]_j$  denote the subdifferential of the  $j$ th coordinate. For  $j = 1$ , there is no regularizing term, and thus the first order condition is simply:

$$\begin{aligned}\left[\partial \bar{f}(\boldsymbol{\theta}, \boldsymbol{\theta}^{(r-1)})\right]_1 &= -2 \sum_{i=1}^n \bar{x}_{ij} \left( y_i + \bar{x}_{ij} (d+1) \theta_1^{(r-1)} - \boldsymbol{\theta}^{(r-1)\top} \bar{\mathbf{x}}_i - \bar{x}_{ij} (d+1) \theta_1 \right) \\ &= 0.\end{aligned}$$

This resolves to yield the 1st coordinate update:

$$\theta_1^{(r)} = \theta_1^{(r-1)} + \frac{\sum_{i=1}^n \bar{x}_{i1} \left[ y_i - \boldsymbol{\theta}^{(r-1)\top} \bar{\mathbf{x}}_i \right]}{(d+1) \sum_{i=1}^n \bar{x}_{i1}^2}.$$

## A second solution to the LASSO problem (5)

For  $j > 1$ , we have

$$\begin{aligned} \left[ \partial \bar{f} \left( \boldsymbol{\theta}, \boldsymbol{\theta}^{(r-1)} \right) \right]_j &= -2 \sum_{i=1}^n \bar{x}_{ij} \left( y_i + \bar{x}_{ij} (d+1) \theta_j^{(r-1)} - \boldsymbol{\theta}^{(r-1)\top} \bar{\mathbf{x}}_i - \bar{x}_{ij} (d+1) \theta_j \right) \\ &\quad + \lambda \partial |\theta_j|, \end{aligned}$$

where

$$\partial |\theta_j| = \begin{cases} -1 & \text{if } \theta_j < 0, \\ [-1, 1] & \text{if } \theta_j = 0, \\ 1 & \text{if } \theta_j > 0. \end{cases}$$

We can explore the first order condition

$$0 \in \left[ \partial \bar{f} \left( \boldsymbol{\theta}, \boldsymbol{\theta}^{(r-1)} \right) \right]_j,$$

for the three possible cases of  $\partial |\theta_j|$ .

## A second solution to the LASSO problem (6)

In the case of  $\theta_j < 0$ , we have

$$-2 \sum_{i=1}^n \bar{x}_{ij} \left( y_i + \bar{x}_{ij} (d+1) \theta_j^{(r-1)} - \boldsymbol{\theta}^{(r-1)\top} \bar{\mathbf{x}}_i - \bar{x}_{ij} (d+1) \theta_j \right) - \lambda = 0,$$

which yields the MM algorithm update

$$\theta_j^{(r)} = \theta_j^{(r-1)} + \frac{\sum_{i=1}^n \bar{x}_{ij} \left[ y_i - \boldsymbol{\theta}^{(r-1)\top} \bar{\mathbf{x}}_i \right] + \lambda/2}{(d+1) \sum_{i=1}^n \bar{x}_{ij}^2}.$$

In the case of  $\theta_j > 0$ , we have

$$-2 \sum_{i=1}^n \bar{x}_{ij} \left( y_i + \bar{x}_{ij} (d+1) \theta_j^{(r-1)} - \boldsymbol{\theta}^{(r-1)\top} \bar{\mathbf{x}}_i - \bar{x}_{ij} (d+1) \theta_j \right) + \lambda = 0,$$

which yields

$$\theta_j^{(r)} = \theta_j^{(r-1)} + \frac{\sum_{i=1}^n \bar{x}_{ij} \left[ y_i - \boldsymbol{\theta}^{(r-1)\top} \bar{\mathbf{x}}_i \right] - \lambda/2}{(d+1) \sum_{i=1}^n \bar{x}_{ij}^2}.$$

## A second solution to the LASSO problem (7)

Lastly, for the  $\theta_j = 0$  case, we have

$$0 \in -2 \sum_{i=1}^n \bar{x}_{ij} \left( y_i + \bar{x}_{ij} (d+1) \theta_j^{(r-1)} - \boldsymbol{\theta}^{(r-1)\top} \bar{\mathbf{x}}_i \right) + \lambda [-1, 1],$$

which implies

$$(d+1) \theta_j^{(r-1)} \sum_{i=1}^n \bar{x}_{ij}^2 + \sum_{i=1}^n \bar{x}_{ij} \left[ y_i - \boldsymbol{\theta}^{(r-1)\top} \bar{\mathbf{x}}_i \right] \in [-\lambda/2, \lambda/2].$$

We thus obtain the result:

$$\theta_j^{(r-1)} + \frac{\sum_{i=1}^n \bar{x}_{ij} \left[ y_i - \boldsymbol{\theta}^{(r-1)\top} \bar{\mathbf{x}}_i \right]}{(d+1) \sum_{i=1}^n \bar{x}_{ij}^2} \in \left[ -\frac{\lambda/2}{(d+1) \sum_{i=1}^n \bar{x}_{ij}^2}, \frac{\lambda/2}{(d+1) \sum_{i=1}^n \bar{x}_{ij}^2} \right].$$

## A second solution to the LASSO problem (8)

At the  $r$ th iteration of the MM algorithm, for coordinate  $j > 1$ , we make the update

$$\theta_j^{(r)} = \theta_j^{(r-1)} + \frac{\sum_{i=1}^n \bar{x}_{ij} [y_i - \boldsymbol{\theta}^{(r-1)\top} \bar{\mathbf{x}}_i] \pm \lambda/2}{(d+1) \sum_{i=1}^n \bar{x}_{ij}^2},$$

if

$$\mp \left( \theta_j^{(r-1)} + \frac{\sum_{i=1}^n \bar{x}_{ij} [y_i - \boldsymbol{\theta}^{(r-1)\top} \bar{\mathbf{x}}_i]}{(d+1) \sum_{i=1}^n \bar{x}_{ij}^2} \right) > \frac{\lambda/2}{(d+1) \sum_{i=1}^n \bar{x}_{ij}^2},$$

respectively, and  $\theta_j^{(r)} = 0$  if

$$\left| \theta_j^{(r-1)} + \frac{\sum_{i=1}^n \bar{x}_{ij} [y_i - \boldsymbol{\theta}^{(r-1)\top} \bar{\mathbf{x}}_i]}{(d+1) \sum_{i=1}^n \bar{x}_{ij}^2} \right| \leq \frac{\lambda/2}{(d+1) \sum_{i=1}^n \bar{x}_{ij}^2}.$$

# Ordinary least squares

When  $\lambda = 0$ , the LASSO problem becomes the *ordinary least-squares* problem

$$\min_{\theta \in \mathbb{R}^{d+1}} \left\{ f(\theta) = \sum_{i=1}^n (y_i - \alpha - \beta^\top \mathbf{x}_i)^2 = \sum_{i=1}^n (y_i - \theta^\top \bar{\mathbf{x}}_i)^2 \right\}.$$

As before, letting  $\mathbf{y}^\top = (y_1, \dots, y_n)$ , putting  $\bar{\mathbf{x}}_i^\top = (1, \bar{\mathbf{x}}_i)$  into the  $i$ th row of  $\mathbf{X}$  yields the form:

$$f(\theta) = (\mathbf{y} - \mathbf{X}\theta)^\top (\mathbf{y} - \mathbf{X}\theta),$$

with first order condition  $\partial f / \partial \theta = -2\mathbf{X}^\top (\mathbf{y} - \mathbf{X}\theta) = \mathbf{0}$ , which yields the minimal solution

$$\theta^* = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}.$$

This solution requires the inversion of  $\mathbf{X}^\top \mathbf{X}$ , which can be *computationally intensive* for large matrices, and impossible when  $\mathbf{X}^\top \mathbf{X}$  is *singular*.

# Regression without matrix inversion

From our solution for the *LASSO* problem, we observe that setting  $\lambda = 0$  yields an MM algorithm for solving the *ordinary least-squares* problem, without matrix inversion.

Denote the  $r$ th iteration of the MM algorithm by  $\boldsymbol{\theta}^{(r)}$  and recall that

$$(\theta_1, \theta_2, \dots, \theta_{d+1}) = (\alpha, \beta_1, \dots, \beta_d).$$

The MM algorithm is defined via the following scheme: at the  $r$ th iteration, for each  $j \in [d+1]$ , make the update

$$\theta_j^{(r)} = \theta_j^{(r-1)} + \frac{\sum_{i=1}^n \bar{x}_{ij} [y_i - \boldsymbol{\theta}^{(r-1)\top} \bar{\mathbf{x}}_i]}{(d+1) \sum_{i=1}^n \bar{x}_{ij}^2}.$$



# An experimental setup

Let  $\tau_1, \dots, \tau_{100}$  be  $n = 100$  equally spaced points between 0 to  $\pi$ . Let  $k = 1, 2,$

$$x_{i,2k-1} = \sin\left(\frac{\pi\tau_i}{2}k\right), \text{ and } x_{i,2k} = \cos\left(\frac{\pi\tau_i}{2}k\right),$$

and  $j = 2k - 1, j = 2k$ , for the corresponding values of  $k$ . Set  $\mathbf{x}_i^\top = (x_{i1}, \dots, x_{i4})$ .

Let  $\alpha = 1$  and  $\beta^\top = (1, 1, 1, 1)$ . For each  $i \in [n]$ , we observe

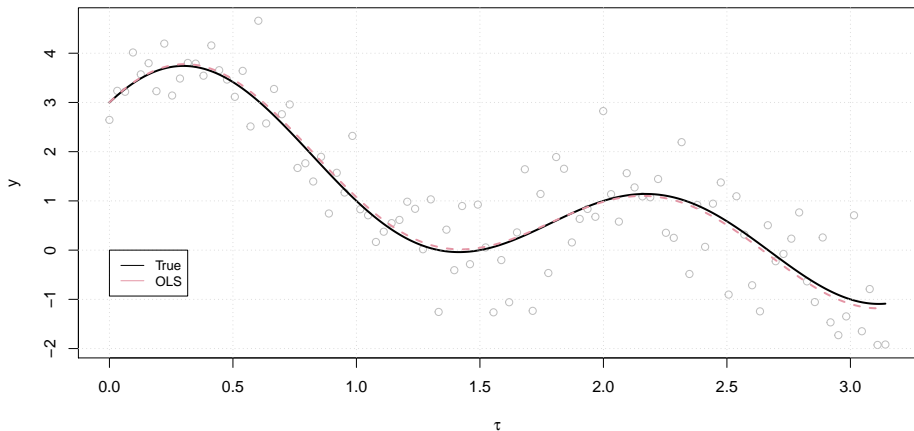
$$y_i = \alpha + \beta^\top \mathbf{x}_i + u_i,$$

where  $u_i$  is a *realization* of a normally distributed random variable with mean 0 and variance 1/2.

Using these data, we wish to estimate the model, which we know has form:

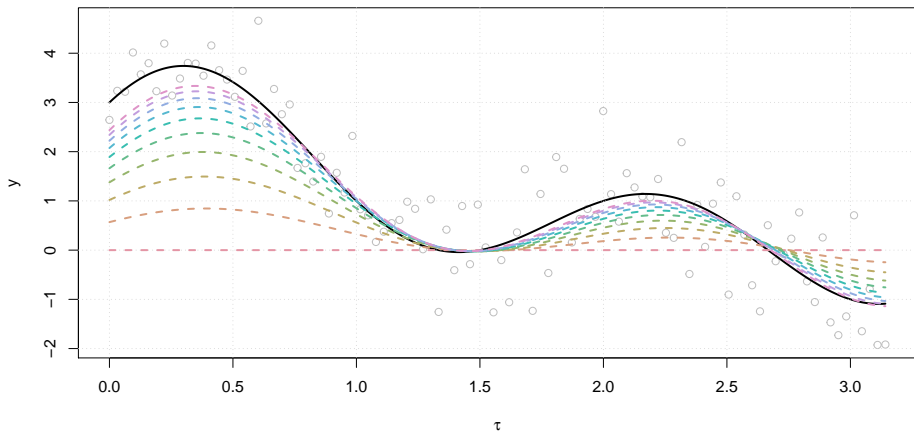
$$y(\tau) = \alpha + \sum_{k=1}^2 \beta_{2k-1} \sin\left(\frac{\pi\tau}{2}k\right) + \sum_{k=1}^2 \beta_{2k} \cos\left(\frac{\pi\tau}{2}k\right).$$

# Experimental data and fitted model



**Figure 14:** Experimental data, true model, and ordinary least squares fitted curve.

## Example fit using the no inversion algorithm



**Figure 15:** A visualization of 10 iterations of the MM algorithm linear regression.

## Another experiment

Again, let  $\tau_1, \dots, \tau_{100}$  be  $n = 100$  equally spaced points between 0 to  $\pi$ .

Let  $k \in [10]$ ,

$$x_{i,2k-1} = \sin\left(\frac{\pi\tau_i}{2}k\right), \text{ and } x_{i,2k} = \cos\left(\frac{\pi\tau_i}{2}k\right),$$

and  $j = 2k - 1, j = 2k$ , for the corresponding values of  $k$ . Set  $\mathbf{x}_i^\top = (x_{i1}, \dots, x_{i20})$ .

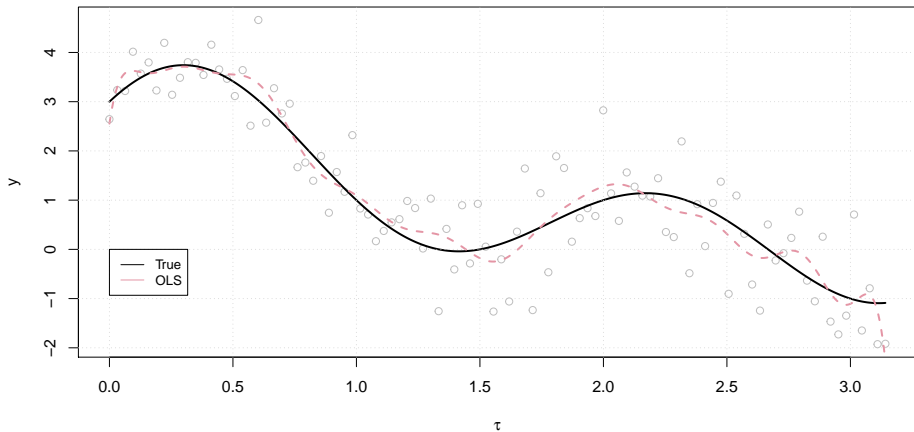
Let  $\alpha = 1$  and  $\beta^\top = (1, 1, 1, 1, \underbrace{0, \dots, 0}_{16})$ . For each  $i \in [n]$ , we observe

$$y_i = \alpha + \beta^\top \mathbf{x}_i + u_i,$$

where  $u_i$  is a *realization* of a normally distributed random variable with mean 0 and variance 1/2. Using these data, we wish to estimate the model, which we know has form:

$$y(\tau) = \alpha + \sum_{k=1}^{10} \beta_{2k-1} \sin\left(\frac{\pi\tau}{2}k\right) + \sum_{k=1}^{10} \beta_{2k} \cos\left(\frac{\pi\tau}{2}k\right).$$

## Second experimental data and fitted model



**Figure 16:** Data from the second experiment, true model, and ordinary least squares fitted curve.

# Overfitted model

The data were generated with *regression coefficients*

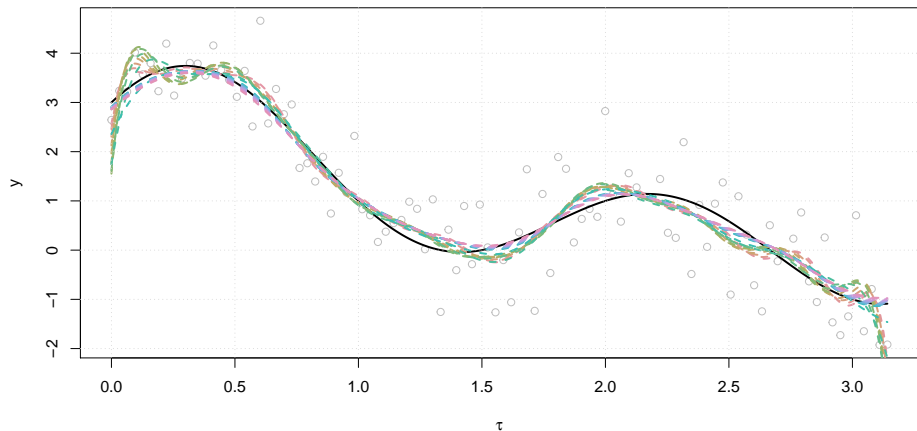
$$\beta^\top = (1, 1, 1, 1, \underbrace{0, \dots, 0}_{16}),$$

but the estimated *ordinary least-squares* estimates of the coefficients resolved to be

$$\hat{\beta}^\top = \begin{pmatrix} 13.3, & -13.7, & 18.3, & -2.2, & 13.3, \\ 7.5, & 4.4, & 11.6, & -3.1, & 8.7, \\ -5.6, & 3.4, & -4.1, & -0.7, & -1.5, \\ 1.8, & -0.1, & -1.1, & 0.3, & -0.3 \end{pmatrix}.$$

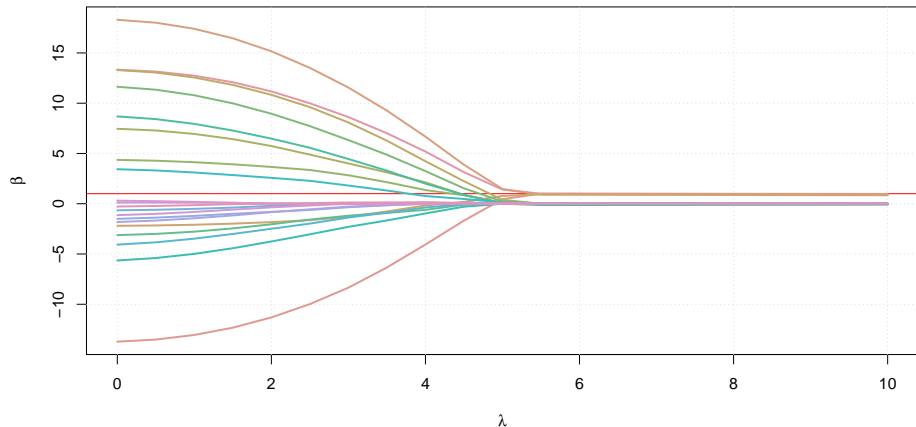
We note that the generative vector  $\beta$  is *sparse* in the sense that it has many elements that are exactly equal to 0. The *ordinary least squares* estimator  $\hat{\beta}$  does not generally yield a sparse solution, and is thus prone to overfitting the model.

# LASSO solutions



**Figure 17:** Fitted LASSO solutions for various levels of regularization.

# Solution paths for the LASSO problem



**Figure 18:** Visualization of the 20 LASSO solution paths of the regression coefficients.



# **Discrimination via optimal separation hyperplanes**

# Discriminant analysis setup

Suppose that  $y_1, \dots, y_n \in \{-1, 1\}$  are  $n$  *spin*-binary variables, explained by their companion covariates  $\mathbf{x}_1, \dots, \mathbf{x}_n \in \mathbb{R}^d$ .

For each  $i \in [n]$ , we wish to estimate  $y_i$  by a function of  $\mathbf{x}_i$ ,  $\hat{y}(\mathbf{x}_i) \in -1, 1$ , so that the **misclassification rate**

$$\frac{1}{n} \sum_{i=1}^n [y_i \times \hat{y}(\mathbf{x}_i) < 0]$$

is as small as possible.

We call the process of estimating the labels  $y_1, \dots, y_n$  via the function  $\hat{y}$  **discriminant analysis** or **classification**.

# Optimal separation hyperplanes

Let  $\alpha \in \mathbb{R}$  and  $\beta \in \mathbb{R}^d$ , and  $\theta^\top = (\alpha, \beta^\top)$ .

We wish to restrict the estimate  $\hat{y}$  to functions of the **separation hyperplane** form

$$\hat{y}(\mathbf{x}; \theta) = \text{sign}(\alpha + \beta^\top \mathbf{x}),$$

where  $\text{sign}(a) = a/|a|$ , for  $a \in \mathbb{R}$ .

The estimation of the labels  $y_1, \dots, y_n$  via a function of the form  $\hat{y}(\mathbf{x}; \theta) = \text{sign}(\alpha + \beta^\top \mathbf{x})$  is called *linear classification*.

When we estimate  $\theta$  by some  $\theta^*$  via some optimization process, we say that  $\hat{y}(\mathbf{x}; \theta^*)$  is an **optimal separation hyperplane**.

# Classification loss

Since our aim is to minimize the *misclassification rate*, it is natural to consider the process of obtaining an *optimal separation hyperplane* by solving the problem

$$\min_{\theta} \left\{ f(\theta) = \frac{1}{n} \sum_{i=1}^n l(y_i, \alpha + \beta^\top \mathbf{x}_i) + \lambda \sum_{j=1}^d \beta_j^2 \right\},$$

where

$$l(y, \alpha + \beta^\top \mathbf{x}) = \begin{cases} 0 & \text{if } y(\alpha + \beta^\top \mathbf{x}) \geq 0 \\ 1 & \text{if } y(\alpha + \beta^\top \mathbf{x}) < 0 \end{cases},$$

is the *classification loss* function, and  $\lambda \geq 0$ .

Here, we perform *ridge regularization* on the coefficients  $\beta$ , so that they do not become large and numerically unstable.

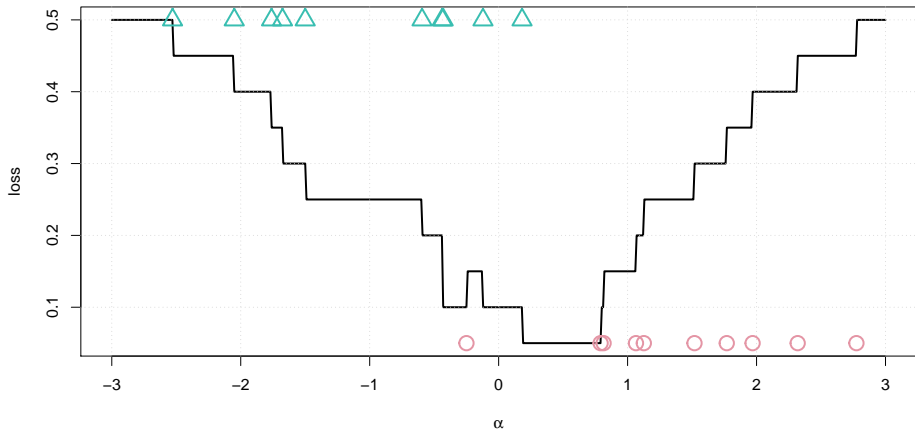
As noted before, the *classification loss* is not well-behaved as it is not convex in  $\theta$  and is not differentiable whenever  $\alpha + \beta^\top \mathbf{x} = 0$ .

# An experiment

Let  $y_1 = y_2 = \dots = y_{10} = -1$  and  $y_{11} = y_{12} = \dots = y_{20} = 1$ , and  $x_1, \dots, x_{10}$  and  $x_{11}, \dots, x_{20}$  be sets of covariates that are generated *uniformly* in the intervals  $[-3, 1]$  and  $[-1, 3]$ , respectively.

We set  $\beta = 1$  and assess the behavior of the average classification loss under linear classification, for different values of  $\alpha \in [-3, 3]$ .

# Visualization of the classification loss problem



**Figure 19:** Visualization of classification loss minimization problem.

# The hinge loss

Due to the difficulty of working with the *classification loss*, it is common to use a **convex relaxation**, whereupon the loss is replaced by a convex function that approximates it. In recent years, the most popular convex relaxation is the **hinge loss** function

$$l(y, \alpha + \beta^\top \mathbf{x}) = \left[1 - y(\alpha + \beta^\top \mathbf{x})\right]_+,$$

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

This yields the classic *soft-margin support vector machine (SVM)* problem of Cortes and Vapnik (1995):

$$\min_{\theta \in \mathbb{R}^{d+1}} \left\{ f(\theta) = \frac{1}{n} \sum_{i=1}^n \left[1 - y_i(\alpha + \beta^\top \mathbf{x}_i)\right]_+ + \lambda \sum_{j=1}^d \beta_j^2 \right\}.$$

# Convexity of the hinge loss

We can demonstrate that the *hinge loss* is a convex function of the parameter vector  $\theta$  by firstly noting the identity, for any  $a, b \in \mathbb{R}$ :

$$\max \{a, b\} = \frac{1}{2} |a - b| + \frac{1}{2}a + \frac{1}{2}b.$$

Using the identity, we can write

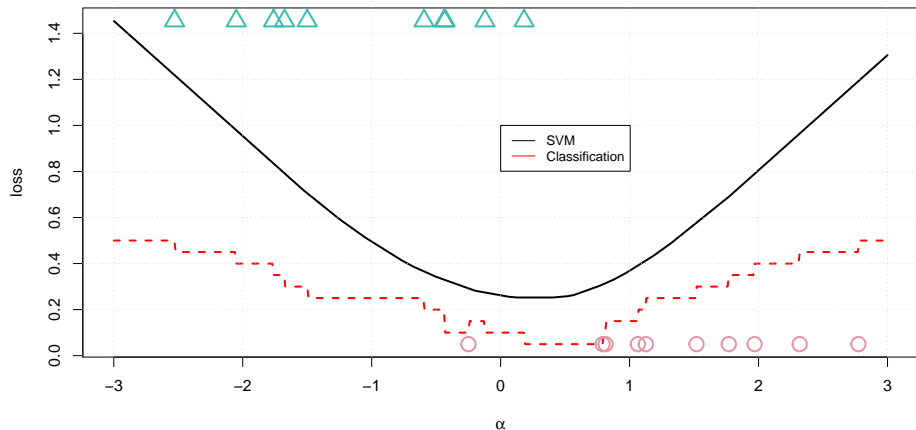
$$[a]_+ = \max \{a, 0\} = \frac{1}{2} |a| + \frac{1}{2}a.$$

We know that the absolute value function is convex, thus  $[\cdot]_+$  is convex. Since the *convex composition* of an *affine* function is convex, we obtain the convexity of the hinge loss.

We note that the function  $f$  is simply a sum of convex functions and therefore the *SVM* problem is a convex optimization problem.



# A visualization of the SVM problem



**Figure 20:** Visualization of the SVM loss in comparison to the classification loss.

# An MM algorithm for the SVM problem

Recall that for small  $\epsilon > 0$ , we can approximate  $g(v) = |v|$  by  $g_\epsilon(v) = \sqrt{v^2 + \epsilon}$ , which can be *majorized* at  $\psi$  by

$$\bar{g}_\epsilon(v, \psi) = \frac{\sqrt{\psi^2 + \epsilon}}{2} + \frac{v^2 + \epsilon}{2\sqrt{\psi^2 + \epsilon}}.$$

As before, write  $\bar{\mathbf{x}}_i^\top = (1, \mathbf{x}_i^\top)$ . For  $(r-1)$ th iteration vector  $\boldsymbol{\theta}^{(r-1)} \in \mathbb{R}^{d+1}$ , we can make the substitutions  $v = 1 - y_i \boldsymbol{\theta}^\top \bar{\mathbf{x}}_i$  and  $v = 1 - y_i \boldsymbol{\theta}^{(r)} \bar{\mathbf{x}}_i$  in order to obtain the majorizer

$$\bar{g}_\epsilon(\boldsymbol{\theta}, \boldsymbol{\theta}^{(r-1)}) = \frac{1}{4w_\epsilon^{(r-1)}} + \frac{1}{4}w_\epsilon^{(r-1)} \left(1 - y_i \boldsymbol{\theta}^\top \bar{\mathbf{x}}_i\right)^2 + \frac{1 - y_i \boldsymbol{\theta}^\top \bar{\mathbf{x}}_i}{2},$$

where  $w_\epsilon^{(r-1)} = 1/\sqrt{(1 - y_i \boldsymbol{\theta}^{(r-1)\top} \bar{\mathbf{x}}_i)^2 + \epsilon}$ , for the *approximation* of  $l(y_i, \alpha + \beta^\top \mathbf{x})$ :

$$g_\epsilon(\boldsymbol{\theta}) = \frac{1}{2}\sqrt{(1 - y_i \boldsymbol{\theta}^\top \bar{\mathbf{x}}_i)^2 + \epsilon} + \frac{1 - y_i \boldsymbol{\theta}^\top \bar{\mathbf{x}}_i}{2},$$

## An MM algorithm for the SVM problem (2)

Let  $f_\epsilon(\boldsymbol{\theta})$  be the approximation of  $f(\boldsymbol{\theta})$ , where  $l(y_i, \alpha + \beta^\top \mathbf{x})$  is replaced by  $g_\epsilon(\boldsymbol{\theta})$ .

We can majorize  $f$  at  $\boldsymbol{\theta}^{(r-1)}$  by

$$\begin{aligned}\bar{f}_\epsilon(\boldsymbol{\theta}; \boldsymbol{\theta}^{(r-1)}) = & C + \frac{1}{4n} \sum_{i=1}^n w_\epsilon^{(r-1)} \left(1 - y_i \boldsymbol{\theta}^\top \bar{\mathbf{x}}_i\right)^2 \\ & - \frac{1}{2n} \sum_{i=1}^n y_i \boldsymbol{\theta}^\top \bar{\mathbf{x}}_i + \lambda \sum_{j=1}^n \beta_j^2.\end{aligned}$$

Here, as usual  $C$  is a constant that does not depend on the active parameter of interest  $\boldsymbol{\theta}$ .

## An MM algorithm for the SVM problem (3)

For each  $i$ , write  $\bar{\mathbf{y}}_i = y_i \bar{\mathbf{x}}_i$ , and let  $\mathbf{Y}$  be the matrix with  $i$ th row  $\bar{\mathbf{y}}_i^\top$ . Also let  $\mathbf{1}$  be a vector of ones of appropriate dimensionality and recall that

$$\bar{\mathbf{I}}_d = \begin{bmatrix} 0 & 0 \\ 0 & \mathbf{I}_d \end{bmatrix}.$$

Further let  $\mathbf{W}_\epsilon^{(r-1)}$  be a diagonal matrix with  $i$ th term  $w_\epsilon^{(r-1)}$ .

In matrix notation, we can write:

$$\bar{f}_\epsilon(\boldsymbol{\theta}; \boldsymbol{\theta}^{(r-1)}) = C + \frac{1}{4n} (\mathbf{1} - \mathbf{Y}\boldsymbol{\theta})^\top \mathbf{W}_\epsilon^{(r-1)} (\mathbf{1} - \mathbf{Y}\boldsymbol{\theta}) - \frac{1}{2n} \mathbf{1}^\top \mathbf{Y}\boldsymbol{\theta} + \lambda \boldsymbol{\theta}^\top \bar{\mathbf{I}}_d \boldsymbol{\theta}.$$

## An MM algorithm for the SVM problem (4)

It is easy to see that  $\bar{f}_\epsilon$  is convex, since it is a positive definite quadratic. We only need to find a stationary point in order to obtain a *global optimizer*.

Thus, we can solve the *first order condition*:

$$\frac{\partial \bar{f}_\epsilon \left( \cdot; \boldsymbol{\theta}^{(r-1)} \right)}{\partial \boldsymbol{\theta}} = -\frac{1}{2n} \mathbf{Y}^\top \mathbf{W}_\epsilon^{(r-1)} (\mathbf{1} - \mathbf{Y}\boldsymbol{\theta}) - \frac{1}{2n} \mathbf{Y}^\top \mathbf{1} + 2\lambda \bar{\mathbf{I}}_d \boldsymbol{\theta} = \mathbf{0}.$$

in order to obtain the  $r$ th iteration of the MM algorithm.

Upon some rearranging, we obtain the MM algorithm of the *iterative least-squares* form

$$\boldsymbol{\theta}^{(r)} = \left( \mathbf{Y}^\top \mathbf{W}_\epsilon^{(r-1)} \mathbf{Y} + 4n\lambda \bar{\mathbf{I}}_d \right)^{-1} \mathbf{Y}^\top \left( \mathbf{1} + \mathbf{W}_\epsilon^{(r-1)} \mathbf{1} \right).$$

# Logistic regression

**Logistic regression** is a cornerstone of statistical inference and is still one of the most successful methods for classification, even though it dates back to the 1800s (Cramer 2002).

Within the framework of *optimal separation hyperplanes*, the logistic regression estimation problem can be phrased as

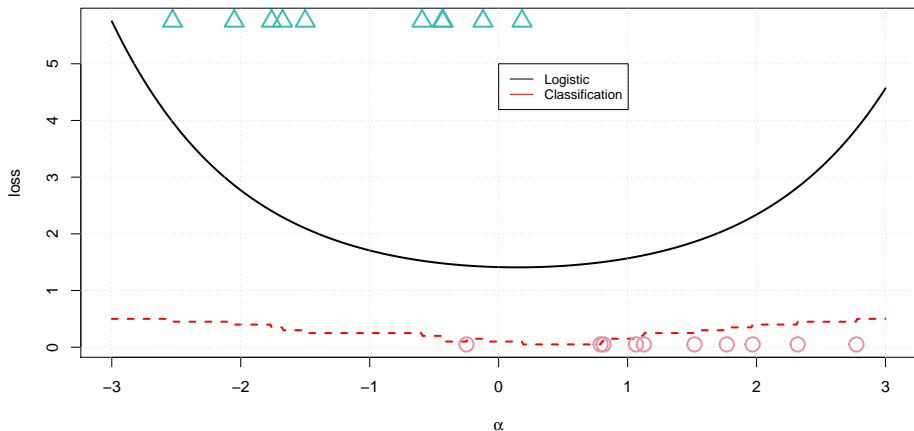
$$\min_{\theta} \left\{ f(\theta) = \frac{1}{n} \sum_{i=1}^n l(y_i, \alpha + \beta^\top \mathbf{x}_i) + \lambda \sum_{j=1}^d \beta_j^2 \right\},$$

where

$$l(y, \alpha + \beta^\top \mathbf{x}) = \log \left[ 1 + \exp \left( -y \left[ \alpha + \beta^\top \mathbf{x} \right] \right) \right].$$

We call this the **logistic loss**.

# A visualization of the logistic regression problem



**Figure 21:** Visualization of the logistic loss in comparison to the classification loss.

# Convexity of the logistic loss

We seek to show that

$$l(y, \alpha + \beta^\top \mathbf{x}) = \log \left[ 1 + \exp \left( -y \left[ \alpha + \beta^\top \mathbf{x} \right] \right) \right],$$

is convex in  $\theta$ . Since the convex composition of an affine function is convex, we are only required to show that

$$g(a) = \log[1 + \exp(a)],$$

is convex in  $a$ .

Since  $g$  is twice differentiable, we can do so by computing the derivative

$$\frac{\partial g}{\partial a} = \exp(a) / [1 + \exp(a)],$$

and verifying the positiveness of the second derivative

$$\frac{\partial^2 g}{\partial a^2} = \frac{1}{1 + \exp(a)} \times \frac{\exp(a)}{1 + \exp(a)} > 0.$$



# An MM algorithm for logistic regression

Recall the univariate form of the *quadratic upper bound* majorizer.

That is, if  $g(v)$  is twice differentiable with first derivative  $\partial g / \partial a$  and second derivative

$$\partial^2 g / \partial a^2 \leq H < \infty,$$

for some  $H < \infty$ , then it can be majorized at  $\psi$  by

$$g(v, \psi) = g(\psi) + \frac{\partial g}{\partial v}(\psi)(v - \psi) + \frac{H}{2}(v - \psi)^2.$$

We set  $g(v) = \log[1 + \exp(v)]$ , with the notation  $\bar{\mathbf{x}}_i^\top = (1, \mathbf{x}_i^\top)$  and  $\bar{\mathbf{y}}_i = y_i \bar{\mathbf{x}}_i$ , we can write

$$g(\boldsymbol{\theta}^{(r-1)}) = \log \left[ 1 + \exp \left( -\bar{\mathbf{y}}_i^\top \boldsymbol{\theta}^{(r-1)} \right) \right] = g_i^{(r-1)},$$

upon making the substitution  $\psi = -\bar{\mathbf{y}}_i^\top \boldsymbol{\theta}^{(r-1)}$ .

## An MM algorithm for logistic regression (2)

Using the expressions for the the first and second derivatives, we can write

$$\frac{\partial g}{\partial v} \left( \theta^{(r-1)} \right) = \frac{\exp \left( -\bar{\mathbf{y}}_i^\top \theta^{(r-1)} \right)}{1 + \exp \left( -\bar{\mathbf{y}}_i^\top \theta^{(r-1)} \right)} = P_i^{(r-1)},$$

and we note that

$$\partial^2 g / \partial v^2 = (1 - P) \times P < 1/4,$$

where  $P = \exp(v) / [1 + \exp(v)] \in (0, 1)$ .

Thus, for  $v = -\bar{\mathbf{y}}_i^\top \theta$ , we can set  $H = 1/4$  and majorize  $g(\theta)$  by

$$\begin{aligned} \bar{g} \left( \theta, \theta^{(r-1)} \right) &= g_i^{(r-1)} + P_i^{(r-1)} \bar{\mathbf{y}}_i^\top \left( \theta^{(r-1)} - \theta \right) \\ &\quad + \frac{1}{8} \left( \theta - \theta^{(r-1)} \right)^\top \bar{\mathbf{y}}_i \bar{\mathbf{y}}_i^\top \left( \theta - \theta^{(r-1)} \right) \end{aligned}$$

## An MM algorithm for logistic regression (3)

Summing over the  $n$  observations, we have the majorizer

$$\begin{aligned}\bar{f}(\boldsymbol{\theta}, \boldsymbol{\theta}^{(r-1)}) &= \frac{1}{n} \sum_{i=1}^n g_i^{(r-1)} + \frac{1}{n} \sum_{i=1}^n P_i^{(r-1)} \bar{\mathbf{y}}_i^\top (\boldsymbol{\theta}^{(r-1)} - \boldsymbol{\theta}) \\ &\quad + \frac{1}{8n} \sum_{i=1}^n (\boldsymbol{\theta} - \boldsymbol{\theta}^{(r-1)})^\top \bar{\mathbf{y}}_i \bar{\mathbf{y}}_i^\top (\boldsymbol{\theta} - \boldsymbol{\theta}^{(r-1)}) + \lambda \sum_{j=1}^n \beta_j^2,\end{aligned}$$

for  $f(\boldsymbol{\theta})$ .

Write  $\mathbf{H} = (1/4) \sum_{i=1}^n \bar{\mathbf{y}}_i \bar{\mathbf{y}}_i^\top$  and let  $\mathbf{p}^{(r-1)} \in \mathbb{R}^n$  have  $i$ th observation  $P_i^{(r-1)}$ . We can simplify the majorizer to the form

$$\begin{aligned}\bar{f}(\boldsymbol{\theta}, \boldsymbol{\theta}^{(r-1)}) &= C - \frac{1}{n} \mathbf{p}^{(r-1)\top} \mathbf{Y} \boldsymbol{\theta} + \frac{1}{2n} (\boldsymbol{\theta} - \boldsymbol{\theta}^{(r-1)})^\top \mathbf{H} (\boldsymbol{\theta} - \boldsymbol{\theta}^{(r-1)}) \\ &\quad + \lambda \boldsymbol{\theta}^\top \bar{\mathbf{I}}_d \boldsymbol{\theta},\end{aligned}$$

where  $C$  does not depend on  $\boldsymbol{\theta}$ .

## An MM algorithm for logistic regression (4)

Since  $\bar{f}$  is a positive definite quadratic function, we can obtain the  $r$ th iteration of the MM algorithm by obtaining a solution to the stationary point

$$\frac{\partial \bar{f}(\cdot, \boldsymbol{\theta}^{(r-1)})}{\partial \boldsymbol{\theta}} = -\frac{1}{n} \mathbf{Y}^\top \mathbf{p}^{(r-1)} + \frac{1}{n} \mathbf{H} (\boldsymbol{\theta} - \boldsymbol{\theta}^{(r-1)}) + 2\lambda \bar{\mathbf{I}}_d \boldsymbol{\theta} = \mathbf{0}.$$

Upon rearranging, we obtain the solution

$$\boldsymbol{\theta}^{(r)} = \left( \mathbf{H} + 2n\lambda \bar{\mathbf{I}}_d \right)^{-1} \left( \mathbf{H} \boldsymbol{\theta}^{(r-1)} + \mathbf{Y}^\top \mathbf{p}^{(r-1)} \right),$$

which is another *iterative least-squares* form.

## Another experiment

For  $n = 200$ , let us generate

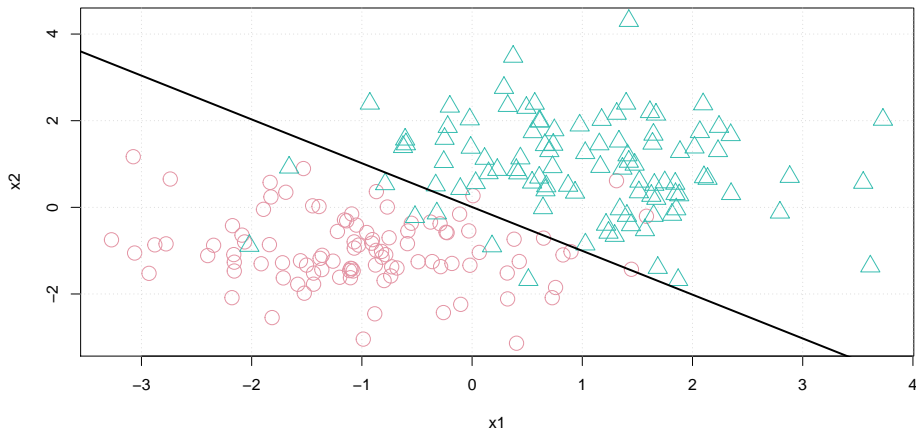
$$y_1, \dots, y_{n/2}, y_{n/2+1}, \dots, y_n = -1, \dots, -1, 1, \dots, 1.$$

For each  $i \in [n]$ , if  $y_i = -1$ , then generate  $\mathbf{x}_i$  from a multivariate normal distribution with mean  $\boldsymbol{\mu}_1^\top = (-1, -1)$  and covariance matrix  $\mathbf{I}$ . If  $y_i = 1$ , then generate  $\mathbf{x}_i$  from a multivariate normal distribution with mean  $\boldsymbol{\mu}_2^\top = (1, 1)$  and covariance matrix  $\mathbf{I}$ .

We wish to assess the performance of the *optimal separation hyperplane* classifiers, with respect to the *classification loss*.

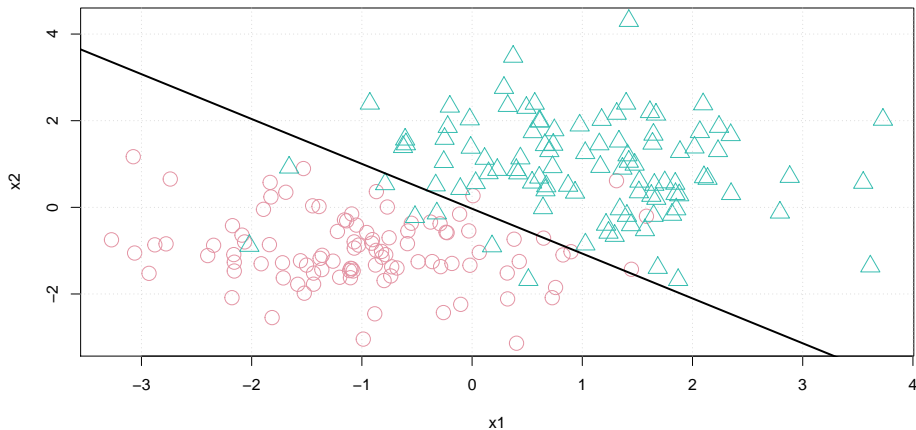
We set  $\epsilon = 0.01$  and  $\lambda = 1$ .

## SVM result



**Figure 22:** Visualization of the SVM separating hyperplane. The average classification loss is 0.055.

# Logistic regression result



**Figure 23:** Visualization of the logistic regression separating hyperplane. The average classification loss is 0.055.

# The Least squares loss

Suppose that we wish to solve the problem

$$\min_{\theta} \left\{ f(\theta) = \frac{1}{n} \sum_{i=1}^n l(y_i, \alpha + \beta^\top \mathbf{x}_i) + \lambda \sum_{j=1}^d \beta_j^2 \right\},$$

without requiring an iterative algorithm.

We can do so by using the **least-squares loss function**

$$l(y, \alpha + \beta^\top \mathbf{x}) = \left[ 1 - y (\alpha + \beta^\top \mathbf{x}) \right]^2.$$



## The Least squares loss (2)

Using our previous notation, we can write the function  $f(\boldsymbol{\theta})$  as:

$$f(\boldsymbol{\theta}) = \frac{1}{n} (\mathbf{1} - \mathbf{Y}\boldsymbol{\theta})^\top (\mathbf{1} - \mathbf{Y}\boldsymbol{\theta}) + \lambda \boldsymbol{\theta}^\top \bar{\mathbf{I}}_d \boldsymbol{\theta},$$

which can be minimized by solving the *first order condition*

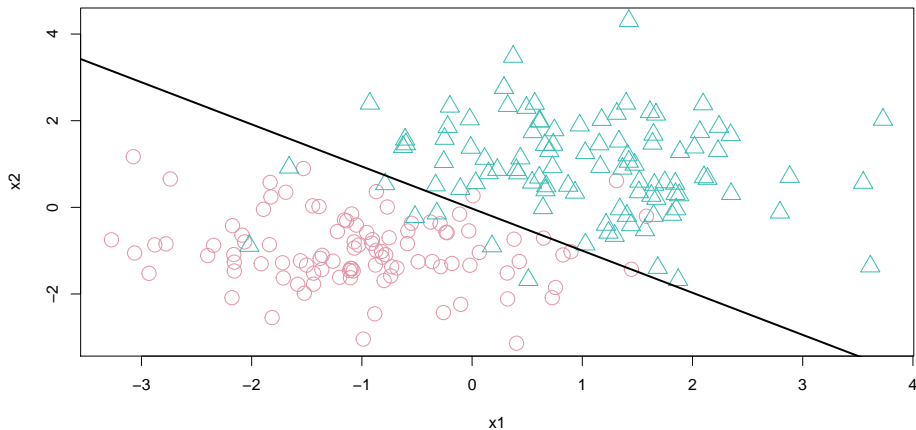
$$\frac{\partial f}{\partial \boldsymbol{\theta}} = -\frac{2}{n} \mathbf{Y}^\top (\mathbf{1} - \mathbf{Y}\boldsymbol{\theta}) + 2\lambda \bar{\mathbf{I}}_d \boldsymbol{\theta} = \mathbf{0}.$$

This resolves to give the solution:

$$\boldsymbol{\theta}^* = \left( \mathbf{Y}^\top \mathbf{Y} + n\lambda \bar{\mathbf{I}}_d \right)^{-1} \mathbf{Y}^\top \mathbf{1},$$

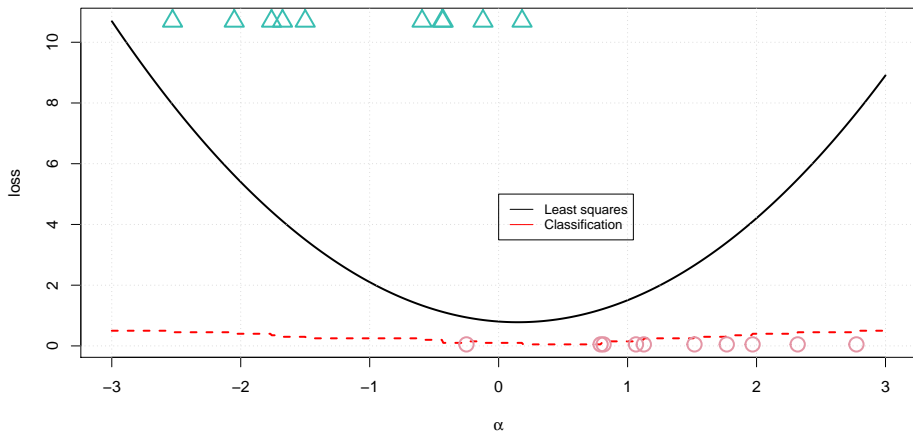
which is in the same form as a *ridge regularized regression*.

## Least squares result



**Figure 24:** Visualization of the least squares separating hyperplane. The average classification loss is 0.055.

# A visualization of least squares problem



**Figure 25:** Visualization of the least squares loss in comparison to the classification loss.

# References I

- Bach, F, R Jenatton, J Mairal, and G Obozinski. 2011. "Optimization with sparsity-inducing penalties." *Foundations and Trends in Machine Learning* 4:1–106.
- Boyd, S, and L Vandenberghe. 2004. *Convex Optimization*. Cambridge: Cambridge University Press.
- Cortes, C, and V Vapnik. 1995. "Support-vector networks." *Machine Learning* 20:273–97.
- Cramer, J S. 2002. "The origins of logistic regression." Tinbergen Institute, University of Amsterdam.
- Dempster, A P, N M Laird, and D B Rubin. 1977. "Maximum likelihood from incomplete data via the EM algorithm." *Journal of the Royal Statistical Society Series B* 39:1–38.

## References II

- De Pierro, A R. 1993. "On the relation between the ISRA and the EM algorithm for positron emission tomography." *IEEE Transactions on Medical Imaging* 12:328–33.
- Hoerl, A E, and R W Kennard. 1970. "Ridge regression: biased estimation for nonorthogonal problems." *Technometrics* 1:55–67.
- Lange, K. 2013. *Optimization*. New York: Springer.
- . 2016. *MM Optimization Algorithms*. Philadelphia: SIAM.
- Nesterov, Y. 2004. *Introductory Lectures on Convex Optimization: a Basic Course*. New York: Springer.
- Petersen, K B, and M S Pedersen. 2012. *Matrix Cookbook*.  
<https://www.math.uwaterloo.ca/~hwolkowi/matrixcookbook.pdf>.
- Razaviyayn, M, M Hong, and Z-Q Luo. 2013. "A unified convergence analysis of block successive minimization methods for nonsmooth optimization." *SIAM Journal of Optimization* 23:1126–53.
- <https://github.com/hiendn/CaenOptimization2018>

## References III

- Rosset, S, J Zhu, and T Hastie. 2004. "Margin maximizing loss functions." In *Advances in Neural Information Processing Systems*.
- Suykens, J A K, and J Vandewalle. 1999. "Least Squares Support Vector Machine." *Neural Processing Letters* 9:293–300.
- Tibshirani, R. 1996. "Regression shrinkage and selection via the Lasso." *Journal of the Royal Statistical Society Series B* 58:267–88.
- Wu, T T, and K Lange. 2008. "Coordinate descent algorithms for LASSO penalized regression." *Annals of Applied Statistics* 2:224–44.