Optimization Theory for Statistics and Machine Learning

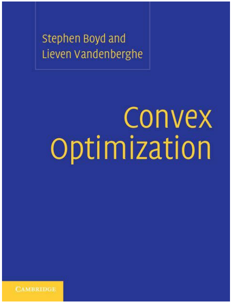
Dr. Hien Nguyen
hiendn.github.io

Lecturer, La Trobe University

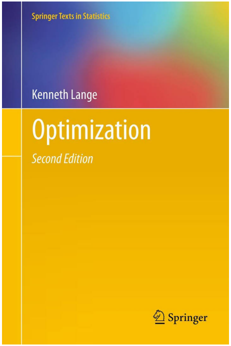
Contents of this course

- 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.

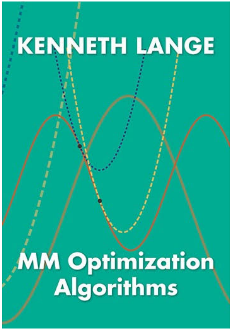
Key readings



(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 $f(\cdot) : \mathbb{T} \to \mathbb{R}$ 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 $\boldsymbol{\theta}$.

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

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

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

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

$$\boldsymbol{\beta}^\top = (\beta_1, \ldots, \beta_d).$$

We can, more concisely write the problem as:

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

An example of the regression problem

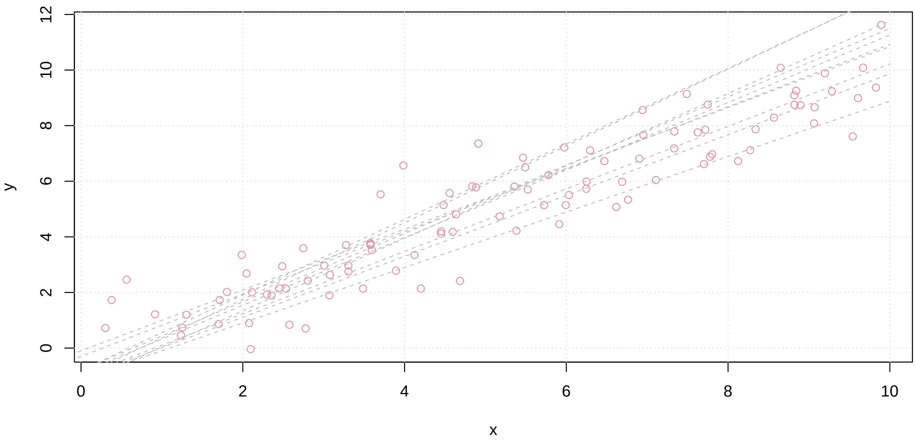


Figure 2: Example of 10 potential linear regression functions when d=1.

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 ℓ_1-LASSO of Wu and Lange (2008) ($p = 1$, $q = 1$, $\lambda > 0$).

Discrimination via optimal seperation hyperplanes

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

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

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

where $\lambda \in [0, \infty)$, and $l\left(y\alpha + \boldsymbol{\beta}^\top \boldsymbol{x}\right) = \left[y\left(\alpha + \boldsymbol{\beta}^\top \boldsymbol{x}\right) < 0\right]$ 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

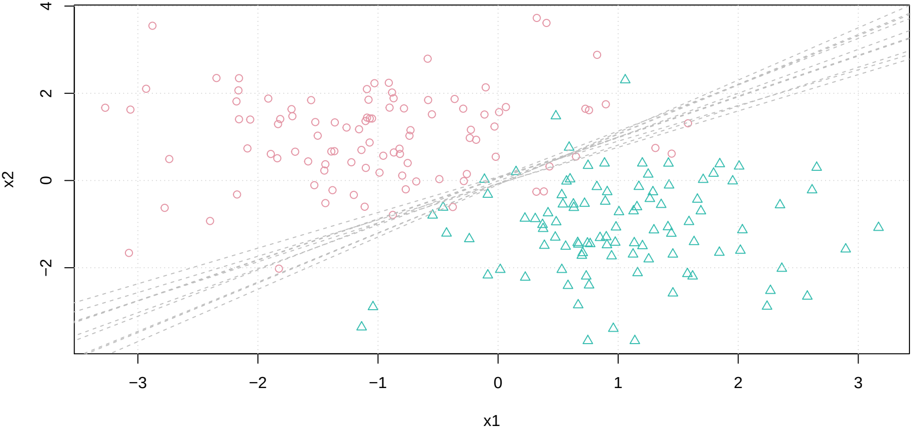


Figure 3: Example of 10 potential discriminant hyperplanes in 2 dimensions.

The support vector machine

The classification loss function

$$l\left(y, \alpha + \boldsymbol{\beta}^\top \boldsymbol{x}\right) = \left[y\left(\alpha + \boldsymbol{\beta}^\top \boldsymbol{x}\right) < 0\right]$$

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

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

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

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

The resulting optimization problem

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

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

General SVM problems

- **Logistic regression** is obtained by setting

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

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

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

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

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

A comparison of loss functions

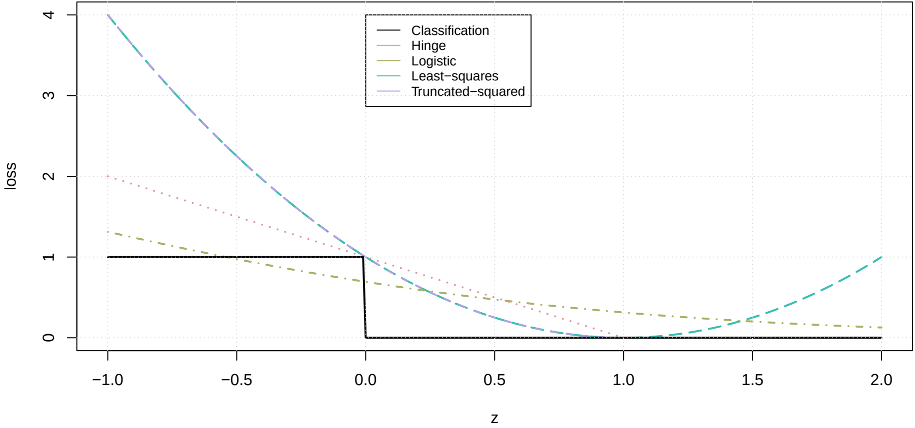


Figure 4: A comparison of SVM loss functions.

Maximum likelihood estimation

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

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

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

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

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

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

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

Latent variable problems

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

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

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

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

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

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

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 $\boldsymbol{X}$ and $\boldsymbol{Y}$.

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

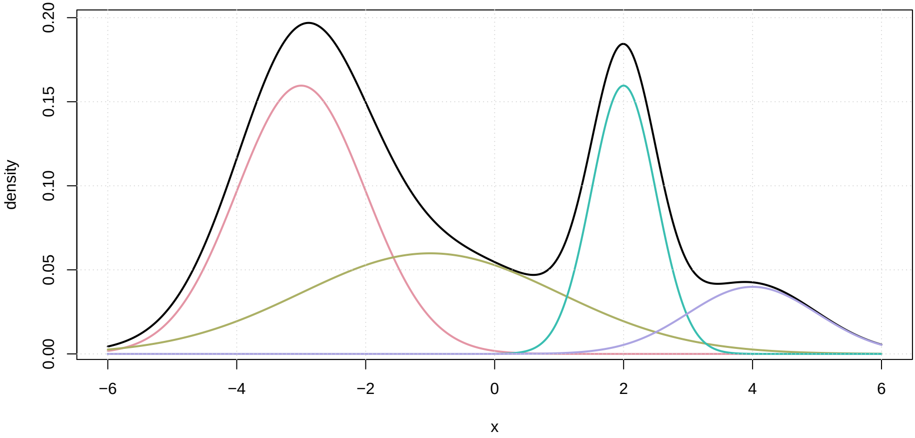


Figure 5: A 4-component mixture of normal PDFs.

Fundamental definitions and results

Global maxima and minima

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

$$f(\boldsymbol{\theta}^*) \geq f(\boldsymbol{\theta}),$$

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

If

$$f(\boldsymbol{\theta}^*) > f(\boldsymbol{\theta}),$$

for all $\boldsymbol{\theta} \neq \boldsymbol{\theta}^*$, then we say that $\boldsymbol{\theta}^*$ 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 ℓ_p vector norm by

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

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

Setting $p = 2$, we obtain the ℓ_2 norm $\|\cdot\|_2$, which is generally referred to as the **Euclidean norm**.

The Euclidean metric

We say that a function

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

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

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

It can ben shown that setting

$$\Delta\left(\boldsymbol{\theta},\boldsymbol{\upsilon}\right)=\left\|\boldsymbol{\theta}-\boldsymbol{\upsilon}\right\|_p$$

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

$$\Delta\left(\boldsymbol{\theta},\boldsymbol{\upsilon}\right)=\left\|\boldsymbol{\theta}-\boldsymbol{\upsilon}\right\|_2.$$

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 $\boldsymbol{\theta}^* \in \mathbb{T}$, such that there exists some $\epsilon > 0$ for which $f(\boldsymbol{\theta}^*) \geq f(\boldsymbol{\theta})$, for all

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

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

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

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

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* $\mathbb{R}^d \backslash \mathbb{T}$ is closed.

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

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

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 $f(\cdot) : [a, b] \to \mathbb{R}$ 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 optimimality theorem* generalizes the extreme value theorem, and states that if $\mathbb{T} \subset \mathbb{R}^d$ is compact and if $f(\cdot) : \mathbb{T} \to \mathbb{R}$ is continuous, then there exists $\boldsymbol{\psi}, \boldsymbol{\upsilon} \in \mathbb{T}$, such that

$$f(\boldsymbol{\psi}) \leq f(\boldsymbol{\theta}) \leq f(\boldsymbol{\upsilon}),$$

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**

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

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

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

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

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

If $\boldsymbol{\theta}^*$ 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 $\boldsymbol{\theta}^*$ is also a stationary point of f.

A second existence theorem

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

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

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

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

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

$$f(\boldsymbol{\psi}) \leq f(\boldsymbol{\theta}) \leq f(\boldsymbol{\upsilon}),$$

for all $\theta \in \mathbb{T}$. Furthermore, if $\boldsymbol{\psi}$ or $\boldsymbol{\upsilon}$ are interior points, then they must be stationary points of f. If $\boldsymbol{\psi}$ or $\boldsymbol{\upsilon}$ 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 $\boldsymbol{\psi}, \boldsymbol{\upsilon} \in \mathbb{T}$, and for any $\lambda \in [0, 1]$, we have

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

We say that $\boldsymbol{\theta}$ is a *convex combination* of the two points $\boldsymbol{\psi}$ and $\boldsymbol{\upsilon}$.

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

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

Convex functions

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

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

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

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

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

We say that a $d \times d$ matrix $\mathbf{A}$ is **positive definite** if for any $\boldsymbol{\theta} \in \mathbb{R}^d \backslash \{\mathbf{0}\}$, $\boldsymbol{\theta}^\top \mathbf{A} \boldsymbol{\theta} > 0$. 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 $f(\cdot) : \mathbb{T} \to \mathbb{R}$ is convex, over a convex domain $\mathbb{T}$, if for any $\boldsymbol{\psi}, \boldsymbol{\upsilon} \in \mathbb{T}$, such that $\boldsymbol{\psi} \neq \boldsymbol{\upsilon}$, we have

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

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

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

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

Furthermore, if $\boldsymbol{\theta}^* \in \mathbb{T}$ is a local minimizer of f, then $\boldsymbol{\theta}^*$ 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 $f(\cdot) : \mathbb{T} \to \mathbb{R}$ is convex. Denote the **subdifferential** of f at the point $\boldsymbol{\theta}^* \in \mathbb{T}$ by $\partial f(\boldsymbol{\theta}^*)$, where

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

When f is differentiable,

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

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

$$\mathbf{0} \in \partial f(\boldsymbol{\theta}^*).$$

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

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

Linear regression

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

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

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

in some sense.

The determination of the parameter $\boldsymbol{\theta}^\top = \left(\alpha, \boldsymbol{\beta}^\top\right) \in \mathbb{R}^{d+1}$ 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

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

Matrix notation

Write $\overline{\boldsymbol{x}}_i^\top = (1, \boldsymbol{x}_i)$ 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

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

If we further write $\mathbf{y}^\top = (y_1, \ldots, y_n)$ and let $\mathbf{X}$ be an $n \times d$ matrix with ith row $\overline{\boldsymbol{x}}_i^\top$, 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 $\boldsymbol{\theta}$. Using the rules of matrix differentiation from the *Matrix Cookbook* of Petersen and Pedersen (2012), we can write the gradient at any point $\boldsymbol{\theta}$ as

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

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

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

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

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

Rules of convexity

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

Checking convexity

Recall that

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

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

1. For each i, we know that $y_i-\alpha-\boldsymbol{\beta}^\top\boldsymbol{x}_i$ is affine, and thus convex.
2. Since $\left|\cdot\right|_2^2=\left(\cdot\right)^2$, it is convex.
3. The affine compositions $\left|y_i-\alpha-\boldsymbol{\beta}^\top\boldsymbol{x}_i\right|_2^2$ and $\left|\beta_j\right|_2^2$ 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 $\boldsymbol{\theta}^*$ 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 $\boldsymbol{x}_i$ that replaces the ℓ_2 loss by an ℓ_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

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

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 $f(\cdot) : \mathbb{T} \to \mathbb{R}$ 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

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

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 a approximate the behavior of f, evaluated at any $\psi \in \mathbb{T}$.

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

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

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

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

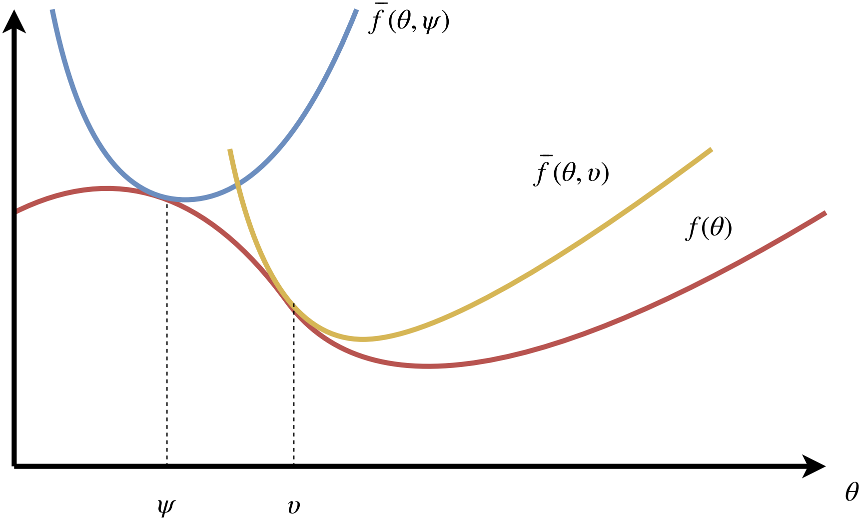


Figure 6: Example of majorizers of an arbitrary function.

The MM algorithm

Suppose that we wish to solve the minimization problem

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

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

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

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

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

via the scheme

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

Illustration of the MM algorithm

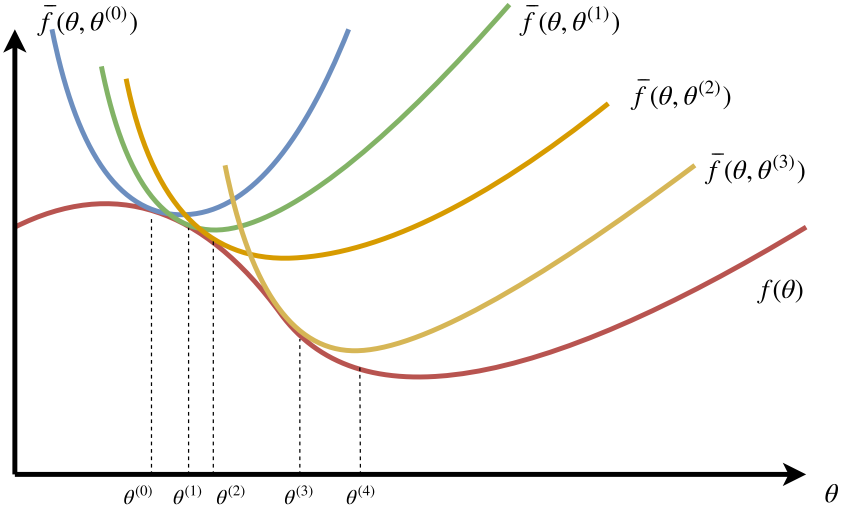


Figure 7: Four steps of an MM algorithm.

The descent property

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

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

By the first property, we have the equality

$$f\left(\boldsymbol{\theta}^{(r)}\right) = \overline{f}\left(\boldsymbol{\theta}^{(r)}, \boldsymbol{\theta}^{(r)}\right).$$

Since $\boldsymbol{\theta}^{(r+1)}$ minimizes $\overline{f}\left(\cdot, \boldsymbol{\theta}^{(r)}\right)$, we have

$$\overline{f}\left(\boldsymbol{\theta}^{(r)}, \boldsymbol{\theta}^{(r)}\right) \geq \overline{f}\left(\boldsymbol{\theta}^{(r+1)}, \boldsymbol{\theta}^{(r)}\right).$$

The second property then tells us that

$$\overline{f}\left(\boldsymbol{\theta}^{(r+1)}, \boldsymbol{\theta}^{(r)}\right) \geq f\left(\boldsymbol{\theta}^{(r+1)}\right),$$

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

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

The directional derivative

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

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

If f is differentiable, then

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

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

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

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

Some more technicalities

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

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

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

$$\boldsymbol{a}_r \in N_\epsilon\left(\boldsymbol{a}\right).$$

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

$$\boldsymbol{a}_r \in N_\epsilon\left(\boldsymbol{a}\right).$$

A first convergence result

Make the following assumptions:

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

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

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 $\overline{f}(\boldsymbol{\theta}, \boldsymbol{\psi})$ is continuous in $\left(\boldsymbol{\theta}^\top, \boldsymbol{\psi}^\top\right)$.

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 $\left\{\boldsymbol{\theta}^{(r)}\right\}$, then $\boldsymbol{\theta}^{(\infty)}$ is a *minimization stationary point* of f.

Proof of first result

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

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

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

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

Proof of first result (2)

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

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

which implies

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

By Assumption 3, we have

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

which completes the proof.

A second convergence result

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

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

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

$$\lim_{r\to\infty} \Delta\left(\boldsymbol{\theta}^{(r)}, \mathbb{T}^*\right) = 0,$$

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

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

Proof of second result

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

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

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

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

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

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

Catalog of majorizers

In Lange (2013), the following are listed as the must 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 $g : (0, \infty) \to \mathbb{R}$ be a convex function. Assume that $\boldsymbol{w}^\top = (w_1, \dots, w_d) \in \mathbb{T}$, $\boldsymbol{\theta}, \boldsymbol{\psi} \in \mathbb{T}$, where $\mathbb{T} = (0, \infty)^d$.

Then, we can *majorize* the function

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

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

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

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 $\boldsymbol{w}^\top = (w_1, \ldots, w_d) \in \mathbb{R}$ and $\boldsymbol{\theta}, \boldsymbol{\psi} \in \mathbb{R}$.

Then, we can *majorize* the function

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

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

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

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

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

An Example: Consider that the function $g(\theta) = \sqrt{\theta}$. Since $\mathrm{d}g/\mathrm{d}\theta = 1/\left(2\sqrt{\theta}\right)$, 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 $g : \mathbb{T} \to \mathbb{R}$ be *convex*, where $\mathbb{T} \subseteq \mathbb{R}^d$, and let $\boldsymbol{\theta}, \boldsymbol{\psi} \in \mathbb{T}$.

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

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

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

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

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

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

Closure properties

The following operations preserve the *majorization* property.

1. **Summation**. If $g_1(\boldsymbol{\theta}), \ldots, g_m(\boldsymbol{\theta})$ are respectively majorized by $\overline{g}_1(\boldsymbol{\theta}; \boldsymbol{\psi}), \ldots, \overline{g}_m(\boldsymbol{\theta}; \boldsymbol{\psi})$, then $f(\boldsymbol{\theta}) = \sum_{k=1}^{m} g_k(\boldsymbol{\theta})$ is majorized by
$$\bar{f}(\boldsymbol{\theta}; \boldsymbol{\psi}) = \sum_{k=1}^{m} \overline{g}_k(\boldsymbol{\theta}; \boldsymbol{\psi}).$$
2. **Non-negative product**. If $g_1(\boldsymbol{\theta}), \ldots, g_m(\boldsymbol{\theta}) \geq 0$ are respectively majorized by $\overline{g}_1(\boldsymbol{\theta}; \boldsymbol{\psi}), \ldots, \overline{g}_m(\boldsymbol{\theta}; \boldsymbol{\psi})$, then $f(\boldsymbol{\theta}) = \prod_{k=1}^{m} g_k(\boldsymbol{\theta})$ is majorized by
$$\bar{f}(\boldsymbol{\theta}; \boldsymbol{\psi}) = \prod_{k=1}^{m} \overline{g}_k(\boldsymbol{\theta}; \boldsymbol{\psi}).$$
3. **Increasing composition**. If $g(\boldsymbol{\theta})$ is majorized by $\overline{g}(\boldsymbol{\theta}, \boldsymbol{\psi})$, and if h is *increasing*, then $f(\boldsymbol{\theta}) = h(g(\boldsymbol{\theta}))$ is majorized by
$$\overline{f}(\boldsymbol{\theta}, \boldsymbol{\psi}) = h(\overline{g}(\boldsymbol{\theta}, \boldsymbol{\psi})).$$

A simple problem

Let $y_1, \ldots, y_n \in \mathbb{R}$ 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

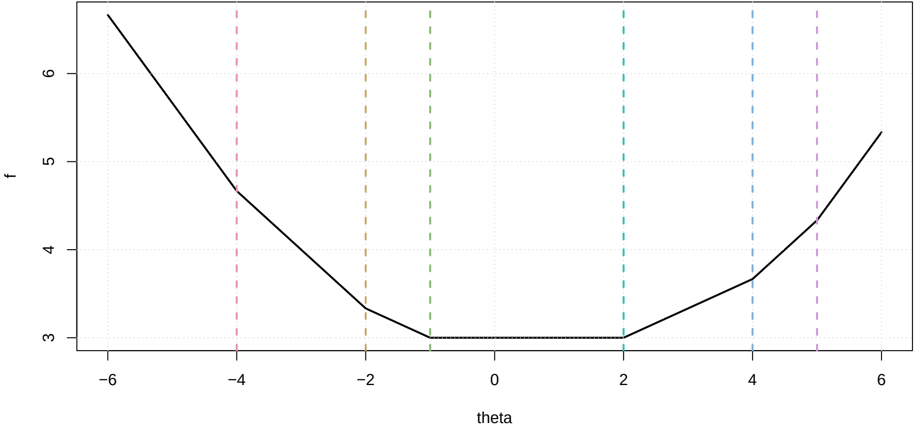


Figure 8: Example instance where the observations are -4, -2, -1, 2, 4, and 5.

Using subdifferentials

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

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

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

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

We observe that $\partial\left|y_i - \theta\right|$ 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 θ^* 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 $\boldsymbol{\theta}^*$ so that

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

Let $y_{(1)} < y_{(2)} < \cdots < y_{(n)}$. 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 θ^* is the **median** of the data.

A useful majorizer

Recall the majorizer

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

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

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

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

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

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

A majorizer for the median problem

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

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

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

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 rth iteration of the MM algorithm:

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

An MM algorithm for the median

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

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

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

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

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

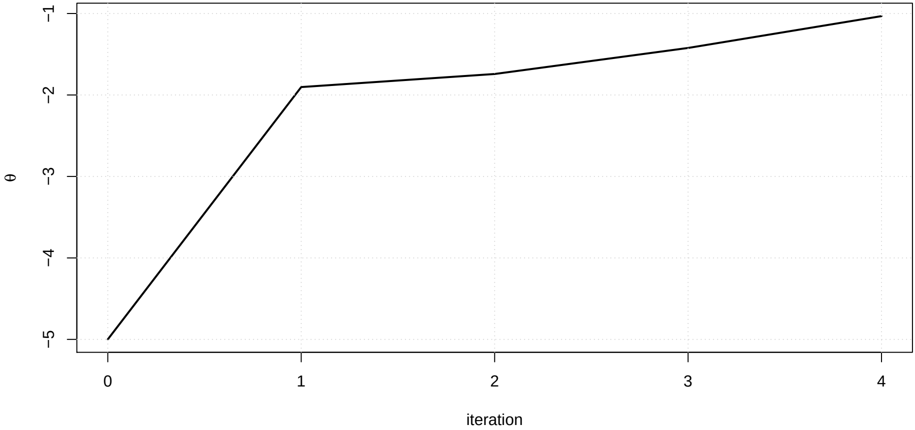


Figure 9: Sequence of MM algorithm iterations for the computation of the median from observations -4, -2, -1, 2, 4, and 5.

Visualization of the majorizers

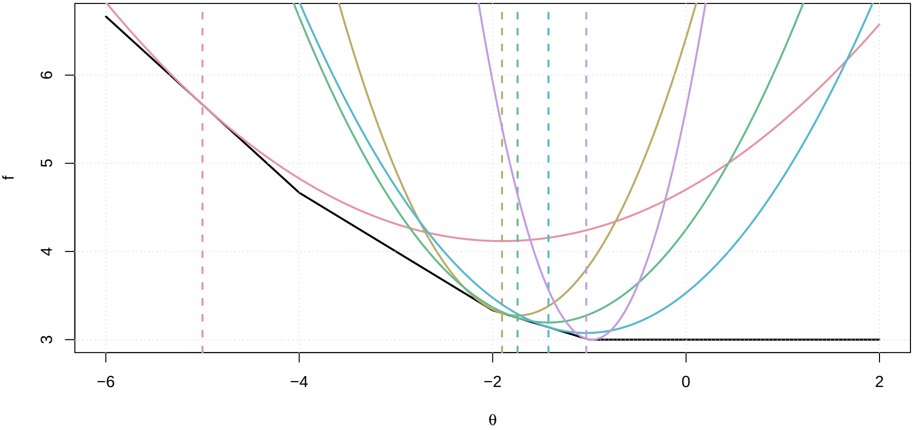


Figure 10: Visualization of the majorizers after 5 steps of the algorithm.

Regression problems

Least-absolute deviation regression

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

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

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

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

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

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

A solution

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

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

using a *linear upper bound*.

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

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

A solution (2)

As in the case of *ridge-regularized least squares*, we write $\mathbf{y}^\top = (y_1, \ldots, y_n)$ 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 ith diagonal element is equal to $1/\left|y_i - \boldsymbol{\theta}^{(r-1)\top}\bar{\mathbf{x}}_i\right|$. We can rewrite $\bar{f}\left(\boldsymbol{\theta}, \boldsymbol{\theta}^{(r-1)}\right)$ as:

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

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*

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

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

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

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 $f\left(\boldsymbol{\theta}, \boldsymbol{\theta}^{(r-1)}\right)$ by

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

which majorizes the *approximate objective function*

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

A problem in the solution (2)

Observe that we can similarly write

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

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

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

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

by solving the optimization problem

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

where $\lambda > 0$.

A first solution

Using the *linear upper bound inequality*, and letting $\boldsymbol{\theta}^{(r)}$ denote the rth 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

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

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

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

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

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

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

which is a convex quadratic function.

Solving the *first order condition*

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

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* $\boldsymbol{\beta}$,

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

is to serve as a **convex relaxation** of the so-called *"ℓ_0 norm regularizer"*

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

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

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

The LASSO regularizer is the only **sparsity inducing** regularizer of the form $\rho_q\left(\boldsymbol{\beta}\right)$.

Visualization of convex relaxation

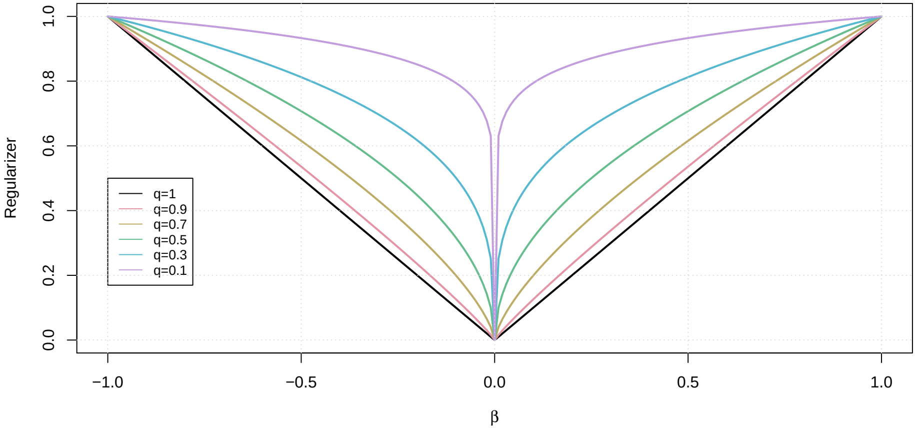


Figure 11: Visualization of various values of q.

Alternative interpretation of regularization

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

$$\rho\left(\boldsymbol{\beta}\right) \leq t,$$

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

That is, LASSO problem can be rewritten as

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

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

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

The Lagrange multiplier

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

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

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

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

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

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

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

Example of a LASSO regularization problem

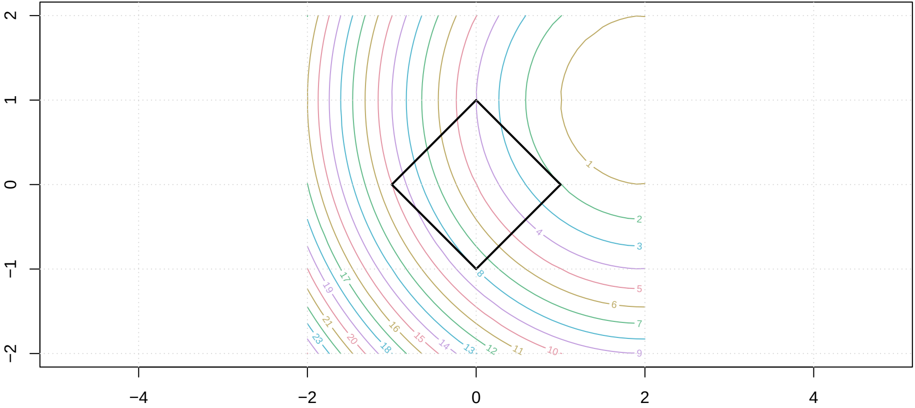


Figure 12: Visualization of the LASSO constraint and functional contours.

Example of an approximate LASSO regularization problem

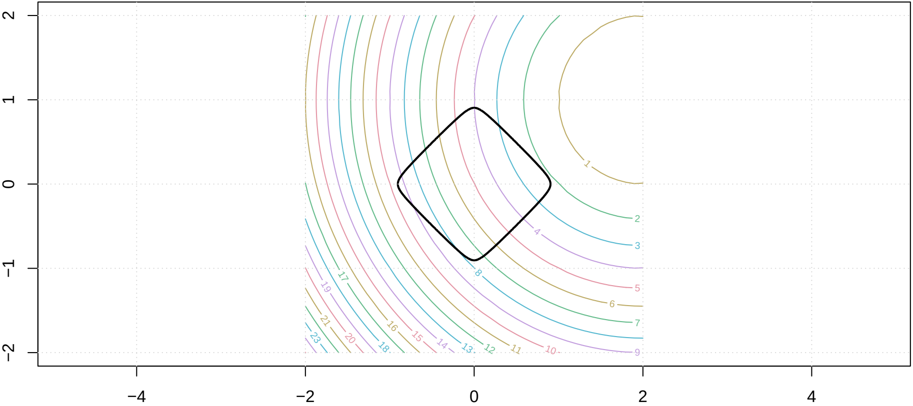


Figure 13: Visualization of the approximate LASSO constraint and functional contours.

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

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

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

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

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

of $f\left(\boldsymbol{\upsilon}\right) = g\left(\boldsymbol{\upsilon}^\top \boldsymbol{w}\right)$, for convex functions g.

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

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

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

A second solution to the LASSO problem (3)

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

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

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

Since $\bar{f}\left(\boldsymbol{\theta}, \boldsymbol{\theta}^{(r-1)}\right)$ 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*

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

A second solution to the LASSO problem (4)

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

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

This resolves to yield the 1st coordinate update:

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

A second solution to the LASSO problem (5)

For $j > 1$, we have

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

where

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

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 $\partial\left|\theta_j\right|$.

A second solution to the LASSO problem (6)

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

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

which yields the MM algorithm update

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

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

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

which yields

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

A second solution to the LASSO problem (7)

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

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

which implies

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

We thus obtain the result:

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

A second solution to the LASSO problem (8)

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

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

if

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

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

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

Ordinary least squares

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

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

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

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

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

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

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

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

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

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

An experimental setup

Let $\tau_1, \ldots, \tau_{100}$ be $n = 100$ equally spaced points between 0 to π. 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 $\boldsymbol{x}_i^\top = (x_{i1}, \ldots, x_{i4})$.

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

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

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

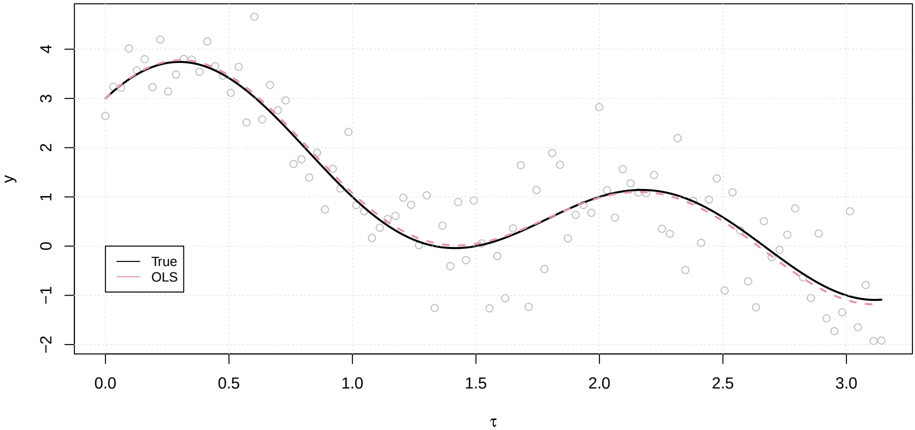


Figure 14: Experimental data, true model, and ordinary least squares fitted curve.

Example fit using the no inversion algorithm

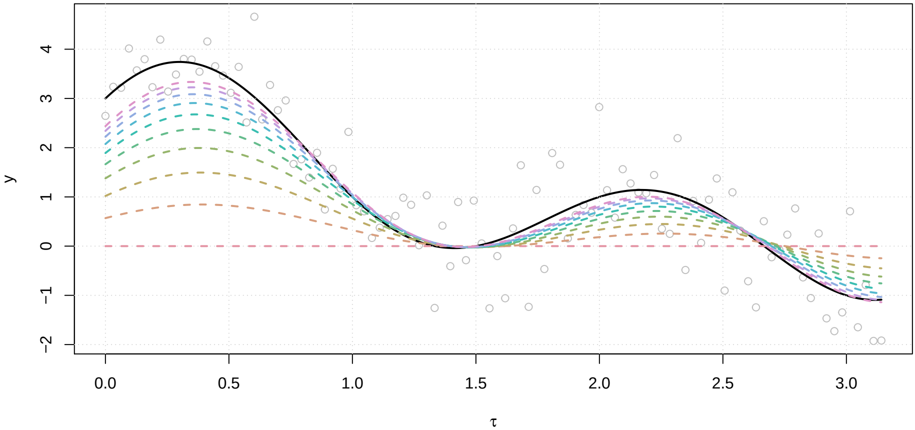


Figure 15: A visualization of 10 iterations of the MM algorithm linear regression.

Another experiment

Again, let $\tau_1, \ldots, \tau_{100}$ be $n = 100$ equally spaced points between 0 to π. 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 $\boldsymbol{x}_i^\top = (x_{i1}, \ldots, x_{i20})$.

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

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

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

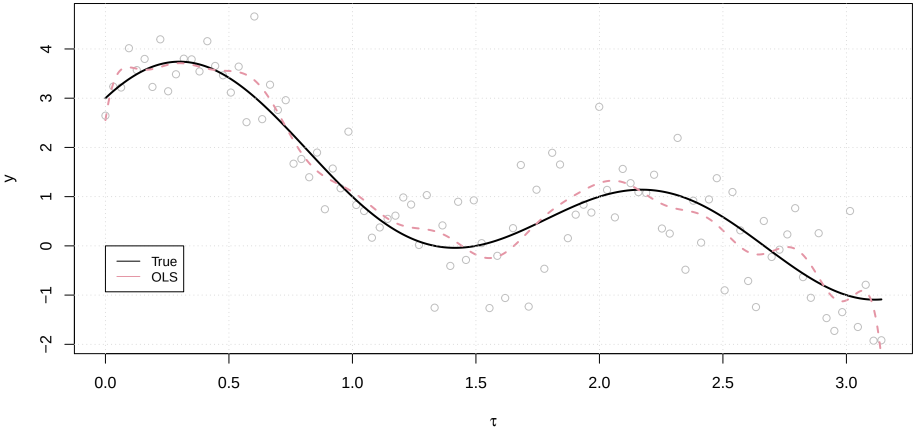


Figure 16: Data from the second experiment, true model, and ordinary least squares fitted curve.

Overfitted model

The data were generated with *regression coefficients*

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

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

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

We note that the generative vector β 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

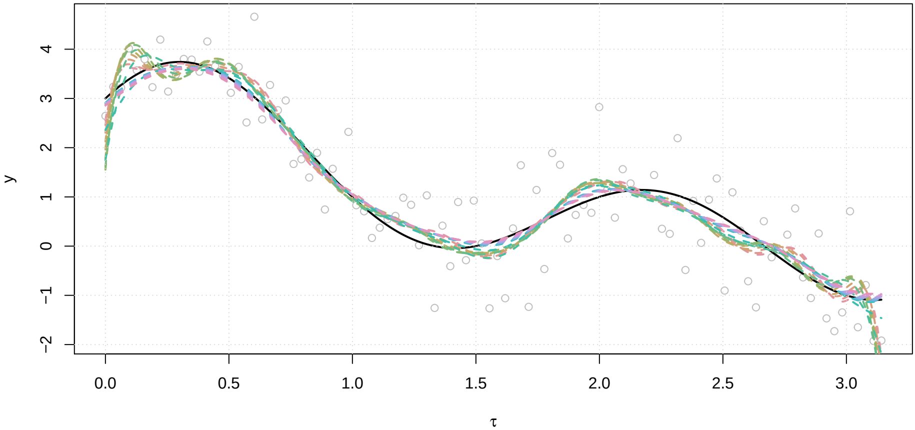


Figure 17: Fitted LASSO solutions for various levels of regularization.

Solution paths for the LASSO problem

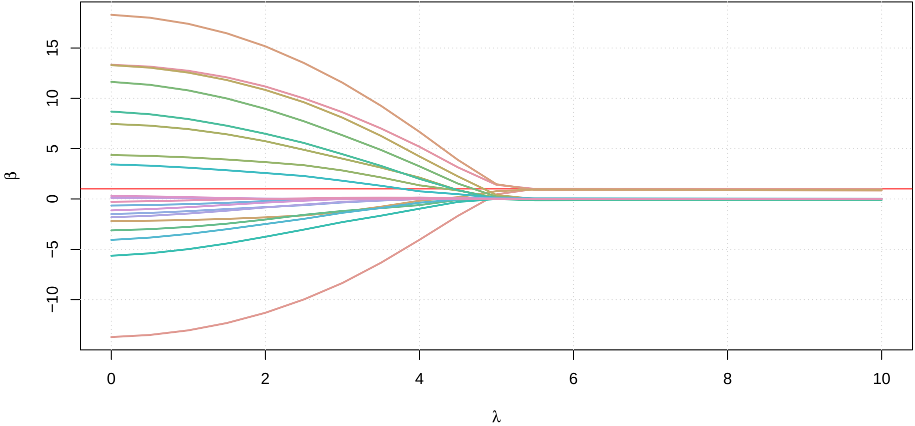


Figure 18: Visualization of the 20 LASSO solution paths of the regression coefficients.

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.

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.

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.

References II

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.

Rosset, S, J Zhu, and T Hastie. 2004. "Margin maximizing loss functions." In *Advances in Neural Information Processing Systems*.

References III

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.