

# Optimization Theory for Statistics and Machine Learning

Dr. Hien Nguyen

[hiendn.github.io](https://hiendn.github.io)

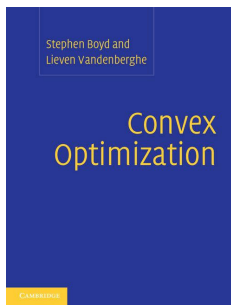
Lecturer, La Trobe University



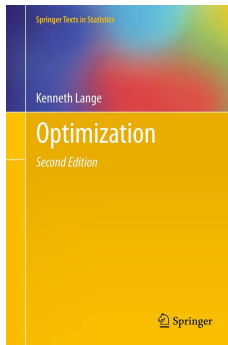
# Contents of this course

- 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>.

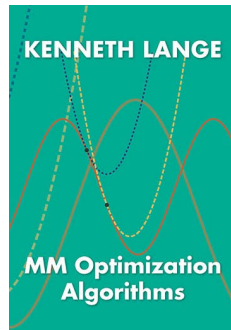
# Key readings



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

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

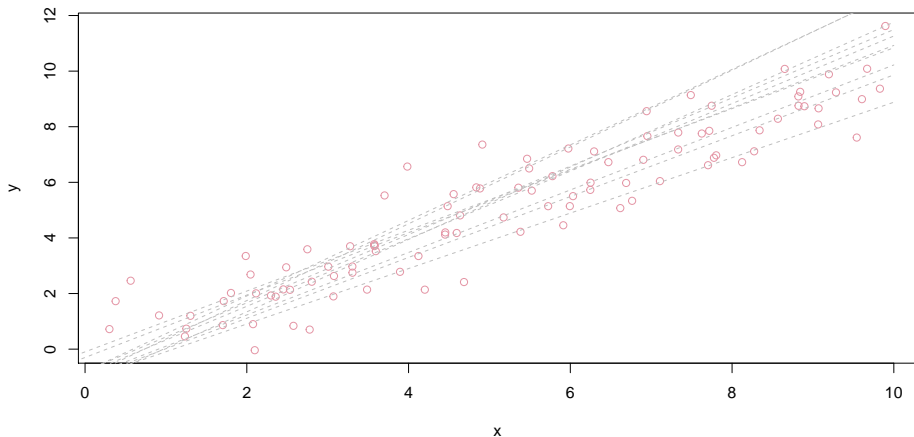
is **minimized**, where  $\lambda \in [0, \infty)$  is a **penalty**,  $|\theta|_p = |\theta|^p|^{1/p}$  for any  $\theta \in \mathbb{R}$  and  $p, q \in [1, \infty)$ . We call  $|\theta|_p$  the  $\ell_p$ -norm of the scalar  $\theta$ . Here,  $(\cdot)^\top$  is the matrix transposition operator, and  $\boldsymbol{\theta}^\top = (\alpha, \beta^\top) \in \mathbb{R}^{d+1}$ , where

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

We can, more concisely write the problem as:

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

# 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  $\mathbf{x}_1, \dots, \mathbf{x}_n \in \mathbb{R}^d$ .

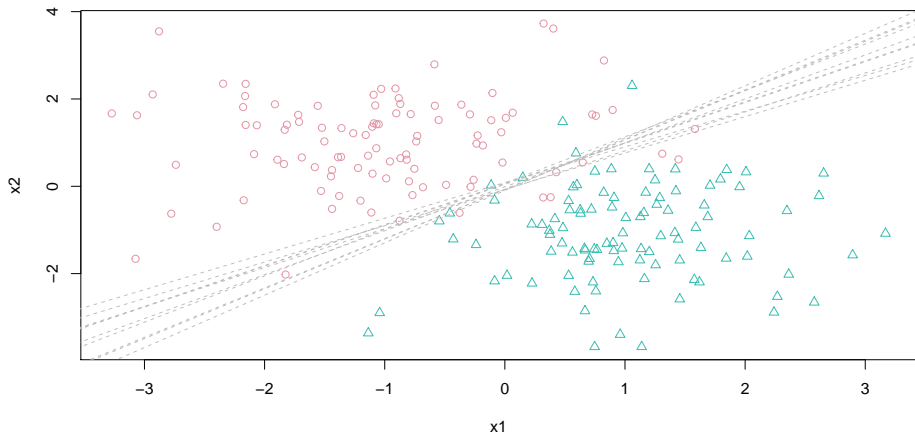
We wish to obtain an optimal hyperplane of the form  $\alpha + \beta^\top \mathbf{x}$ , where  $\alpha \in \mathbb{R}$ ,  $\beta \in \mathbb{R}^d$ ,  $\mathbf{x} \in \mathbb{R}^d$ , and  $\boldsymbol{\theta}^\top = (\alpha, \beta^\top)$ , such that it minimizes the regularized average **loss**

$$\frac{1}{n} l(y_i, \alpha + \beta^\top \mathbf{x}_i) + \lambda \sum_{j=1}^d |\theta_j|_2^2,$$

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

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

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

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

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

# General SVM problems

- **Logistic regression** is obtained by setting

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

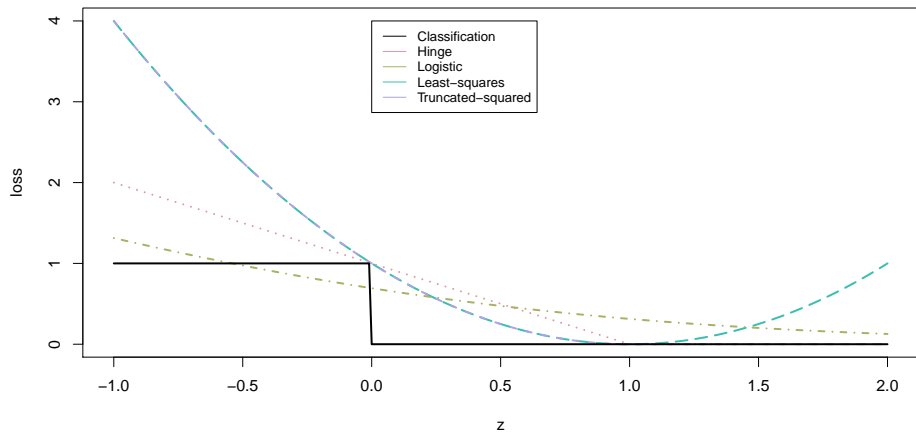
- The **least-squares** SVM of Suykens and Vandewalle (1999) is obtained by setting

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

- 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.

## 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

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

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  $N_\epsilon(\theta^*)$ , there exists a

$$\theta \in \mathbb{T} \cup N_\epsilon(\theta^*).$$

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

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

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 **partial derivatives**

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

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

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

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

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

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

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

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  $\theta \in \mathbb{R}^d$ ,  $\theta^\top \mathbf{A} \theta > 0$ . 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

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

When  $f$  is differentiable,  $\partial f(\theta^*) = \{\partial f(\theta^*) / \partial \theta\}$ .

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

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



# References I

- 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.
- 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.

## References II

- 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.