MATH4230-2022

该二维码 7 天内 (2月1日前) 有效，重新进入将更新

Optimization Theory

Math4230, 2022

The Chinese University of Hong Kong

zeng@math.cuhk.edu.hk

MATH4230-cuhk-2022

WhatsApp 群组

此群组二维码为您私有。如果您将它共享给他人，对方可以通过 WhatsApp 相机扫描二维码，来加入这个群组。

重置二维码

Math4230: Optimization Theory
Prerequisite Topics

This is meant to be a brief, informal refresher of some topics that will form building blocks in this course. The content of the first two sections of this document is mainly taken from Appendix A of B & V, with some supplemental information where needed. See the end for a list of potentially helpful resources you can consult for further information.

1 Real Analysis and Calculus

1.1 Properties of Functions

Limits You should be comfortable with the notion of limits, not necessarily because you will have to evaluate them, but because they are key to understanding other attributes of functions. Informally, $\lim_{x \to a} f(x)$ is the value that f approaches as x approaches the value a.

Continuity A function $f(x)$ is continuous at a particular point x' if, as a sequence $x_1, x_2, ...$ approaches x', the value $f(x_1), f(x_2), ...$ approaches $f(x')$. In limit notation: $\lim_{i \to \infty} f(x_i) = f(\lim_{i \to \infty} x_i)$. f is continuous if it is continuous at all points $x' \in \text{dom} f$.

Differentiability A function $f : \mathbb{R}^n \to \mathbb{R}$ is considered differentiable at $x \in \text{int dom} f$ if there exists a vector $\bigtriangledown f(x)$ that satisfies the following limit:

$$\lim_{z \in \text{dom} f, z \neq x, z \to x} \frac{||f(z) - f(x) - Df(x)(z - x)||_2}{||z - x||_2} = 0$$

We refer to $\bigtriangledown f(x)$ as the derivative of f, and it is the transpose of the gradient.

Smoothness f is smooth if the derivatives of f are continuous over all of $\text{dom} f$. We can describe smoothness of a certain order if the derivatives of f are continuous up to a certain derivative. It is also reasonable to talk about smoothness over a particular interval of the domain of f.

Lipschitz A function f is Lipschitz with Lipschitz constant L if $||f(x) - f(y)|| \leq L||x - y||$ $\forall x, y \in \text{dom} f$. If we refer to a function f as Lipschitz, we are making a stronger statement about the continuity of f. A Lipschitz function is not only continuous, but it does not change value very rapidly, either. This is obviously not unrelated to the smoothness of f, but a function can be Lipschitz but not smooth.

Taylor Expansion The first order Taylor expansion of a function gives us an easy way to form a linear approximation to that function:

$$f(y) \approx f(x) + \nabla f(x)(y - x)$$

And equivalent form that is often useful is the following:

$$f(y) = f(x) + \int_0^1 \nabla f(t(x - y) + y)(y - x)\, dt.$$

For a quadratic approximation, we add another term:

$$f(y) \approx f(x) + \nabla f(x)(y - x) + \frac{1}{2}(y - x)^T \nabla^2 f(x)(y - x)$$

Often when doing convergence analysis we will upper bound the Hessian and use the quadratic approximation to understand how well a technique does as a function of iterations.

1.2 Sets

Interior The interior int$C$ of the set C is the set of all points $x \in C$ for which $\exists \epsilon > 0$ s.t. $\{y || y - x||_2 \leq \epsilon\} \subseteq C$.

Closure The closure cl$C$ of a set C is the set of all x such that $\forall \epsilon > 0\ \exists y \in C$ s.t. $||x - y||_2 \leq \epsilon$. The closure only makes sense for closed sets (see below), and can be considered the union of the interior of C and the boundary of C.

Boundary The boundary is the set of points bd$C$ for which the following is true: $\forall \epsilon\ \exists y \in C, z \notin C$ s.t. $||y - x||_2 \leq \epsilon$ and $||z - x||_2 \leq \epsilon$.

Complement The complement of the set $C \subseteq \mathbb{R}^n$ is denoted by $\mathbb{R}^n \setminus C$. It is the set of all points not in C

Open vs Closed A set C is open if int$C = C$. A set is closed if its complement is open.

Equality You'll notice that above we used a notion of equality for sets. To show formally that sets A and B are equal, you must show $A \subseteq B$ *and* $B \subseteq A$.

1.3 Norms

See B & V for a much more detailed treatment of this topic. I am going to list the most common norms so that you are aware of the notation we will be using in this class:

ℓ_0 $||x||_0$ is the number of nonzero elements in x. We often want to minimize this, but it is non-convex (and actually, not a real norm), so we approximate it (you could say we relax it) to other norms (e.g. ℓ_1).

ℓ_p $||x||_p = (|x_1|^p + \cdots + |x_n|^p)^{1/p}$, where $p \geq 1$. Some common examples:

- $||x||_1 = \sum_{i=1}^n |x_i|$
- $||x||_2 = \sqrt{\sum_{i=1}^n x_i^2}$

- $||x||_\infty = \max_i |x_i|$

Spectral/Operator Norm $||X||_{op} = \sigma_1(X)$, the largest singular value of X.

Trace Norm $||X||_{tr} = \sum_{i=1}^{r} \sigma_r(X)$, the sum of all the singular values of X.

1.4 Linear/Affine Functions

In this course, a linear function will be a function $f(x) = a^T x$. Affine functions are linear functions with a nonzero intercept term: $g(x) = a^T x + b$.

1.5 Derivatives of Functions

See B & V for some nice examples. Consider the following for a function $f : \mathbb{R}^n \to \mathbb{R}$:

Gradient The i^th element of $\bigtriangledown f$ is the partial derivative of f w.r.t. the i^th dimension of the input x: $\bigtriangledown f(x)_i = \frac{\partial f(x)}{\partial x_i}$

Chain Rule Let $h(x) = g(f(x))$ for $g : \mathbb{R} \to \mathbb{R}$. We have: $\bigtriangledown h(x) = g'(f(x)) \bigtriangledown f(x)$

Hessian In the world of optimization, we denote the Hessian matrix as $\bigtriangledown^2 f(x) \in \mathbb{R}^{n \times n}$ (some of you have maybe seen this symbol used as the Laplace operator in other courses). The ij^th entry of the Hessian is given by: $\bigtriangledown f(x)_{ij} = \frac{\partial^2 f(x)}{\partial x_i \partial x_j}$

Matrix Differentials In general we will not be using these too much in class. The major differentials you need to know are:

- $\partial X^T X = X$
- $\frac{\partial}{\partial X} tr(XA) = A^T$

2 Linear Algebra

2.1 Matrix Subspaces

Row Space The row space of a matrix A is the subspace spanned of the rows of A.

Column Space The column space of a matrix A is the subspace spanned of the columns of A.

Null Space The null space of a matrix A is the set of all x such that $Ax = 0$.

Rank rank$A$ is the number of linearly independent columns in A (or, equivalently, the number of linearly independent rows). A matrix $A \in \mathbb{R}^{m \times n}$ is full rank if rank$A = \min\{m, n\}$. Recall that if A is square and full rank, it is invertible.

2.2 Orthogonal Subspaces

Two subspaces $S_1, S_2 \in \mathbb{R}^n$ are orthogonal if $s_1^T s_2 = 0 \; \forall \; s_1 \in S_1, s_2 \in S_2$.

2.3 Decomposition

Eigen Decomposition If $A \in S^n$, the set of real, symmetric, $n \times n$ matrices, then A can be factored:

$$A = Q\Lambda Q^T$$

Here Q is an orthogonal matrix, which means that $Q^TQ = I$. $\Lambda = \text{diag}(\lambda_1, \lambda_2, ..., \lambda_n)$, where the eigenvalues λ_i are ordered by decreasing value. Some useful facts about A that we can ascertain from the eigen decomposition:

- $|A| = \prod_{i=1}^n \lambda_i$
- $\text{tr}A = \sum_{i=1}^n \lambda_i$
- A is invertible iff (if and only if) all its eigenvalues are nonzero. Then $A^{-1} = Q\Lambda^{-1}Q^T$ (note that I have used the fact that for orthogonal Q, $Q^{-1} = Q^T$
- A is positive semidefinite if all its eigenvalues are nonnegative.

Singular Value Decomposition Any matrix $A \in \mathbb{R}^{m\times n}$ with rank r can be factored as:

$$A = U\Sigma V^T$$

Here $U \in \mathbb{R}^{m\times r}$ has the property that $U^TU = I$ and $V \in \mathbb{R}^{n\times r}$ likewise satisfies $V^TV = I$. $\Sigma = \text{diag}(\sigma_1, \sigma_2, ..., \sigma_r)$ where the singular values σ_i are ordered by decreasing value. Some useful facts that we can learn using this decomposition:

- The SVD of A has the following implication for the eigendecomposition of A^TA:

$$A^TA = [VW] \begin{bmatrix} \Sigma^2 & 0 \\ 0 & 0 \end{bmatrix} [VW]^T$$

  W is the matrix such that $[VW]$ is orthogonal.
- The *condition number* of A (an important concept for us in this course) is $\text{cond}A = \frac{\sigma_1}{\sigma_r}$

Pseudoinverse The SVD of a singular matrix A yields the pseudoinverse $A^\dagger = V\Sigma^{-1}U^T$.

3 Canonical ML Problems

3.1 Linear Regression

Linear regression is the problem of finding $f : X \to Y$, where $X \in \mathbb{R}^{n\times p}$, Y is an n-dimensional vector of real values and f is a linear function. Canonically, we find f by finding the vector $\hat{\beta} \in \mathbb{R}^p$ that minimizes the *least squares objective*:

$$\hat{\beta} = \underset{\beta}{\text{argmin}} \, ||X\beta - Y||_2^2$$

For $Y \in \mathbb{R}^{n\times q}$, the multiple linear regression problem, we find a matrix $\hat{B}$ that such that:

$$\hat{B} = \underset{B}{\text{argmin}} \, ||XB - Y||_F^2$$

Note that in its basic form, the linear regression problem can be solved in closed form.

3.2 Logistic Regression

Logistic regression is the problem of finding $f : X \to Y$, where Y is an n-dimensional vector binary values, and f has the form $f(x) = \text{logit}(\beta^T x)$. The logit function is defined as $\text{logit}(\alpha) = \frac{1}{1+\exp(-\alpha)}$. We typically solve for β by maximizing the likelihood of the observed data, which results in the following optimization problem:

$$\hat{\beta} = \underset{\beta}{\text{argmax}} \sum_{i=1}^{n} [y_i \beta^T x_i - \log(1 + \exp(-y_i \beta^T x_i)]$$

3.3 Support Vector Machines

Like logistic regression, SVMs attempt to find a function that linearly separates two classes. In this case, the elements of Y are either 1 or -1. SVMs frame the problem as the following constrained optimization problem (in primal form):

$$\hat{\beta} = \underset{\beta}{\text{argmin}} \frac{1}{2} ||\beta||_2^2$$

$$\text{s.t.} \qquad y_i(\beta^T x_i) \geq 1 \qquad \forall i = 1, ..., n$$

In its simplest form, the support vector machine seeks to find the hyperplane (parameterized by β) that separates the classes (encoded in the constraint) and does so in a way that creates the largest margin between the data points and the plane (encoded in the objective that is minimized).

3.4 Regularization/Penalization

Regularization (sometimes referred to as penalization) is a technique that can be applied to almost all machine learning problems. Most of the time, we regularize in an effort to simplify the learned function, often by forcing the parameters to be "small" (either in absolute size or in rank) and/or setting many of them to be zero. Regularization is also sometimes used to incorporate prior knowledge about the problem.

We incorporate regularization by adding either constraints or penalties to the existing optimization problem. This is easiest to see in the context of linear regression. Where previously we only had least squares loss, we can add penalties to create the following two variations:

Ridge Regression By adding an ℓ_2 penalty, our objective to minimize becomes:

$$\hat{\beta} = \underset{\beta}{\text{argmin}} \, ||X\beta - Y||_2^2 + \lambda ||\beta||_2$$

This will result in many elements of β being close to 0 (more so if λ is larger).

Lasso Regression By adding an ℓ_1 penalty, our objective to minimize becomes:

$$\hat{\beta} = \underset{\beta}{\text{argmin}} \, ||X\beta - Y||_2^2 + \lambda ||\beta||_1$$

This will result in many elements of β being 0 (more if λ is larger).

The first example is nice because it still can be solved in closed form. Notice however that the ℓ_1 penalty creates issues not only for a closed-form solution, but also for standard first-order methods, because it is not differentiable everywhere. We will study how to deal with this later in the course.

4 Further Resources

In addition to B & V, the following are good sources of information on these topics:

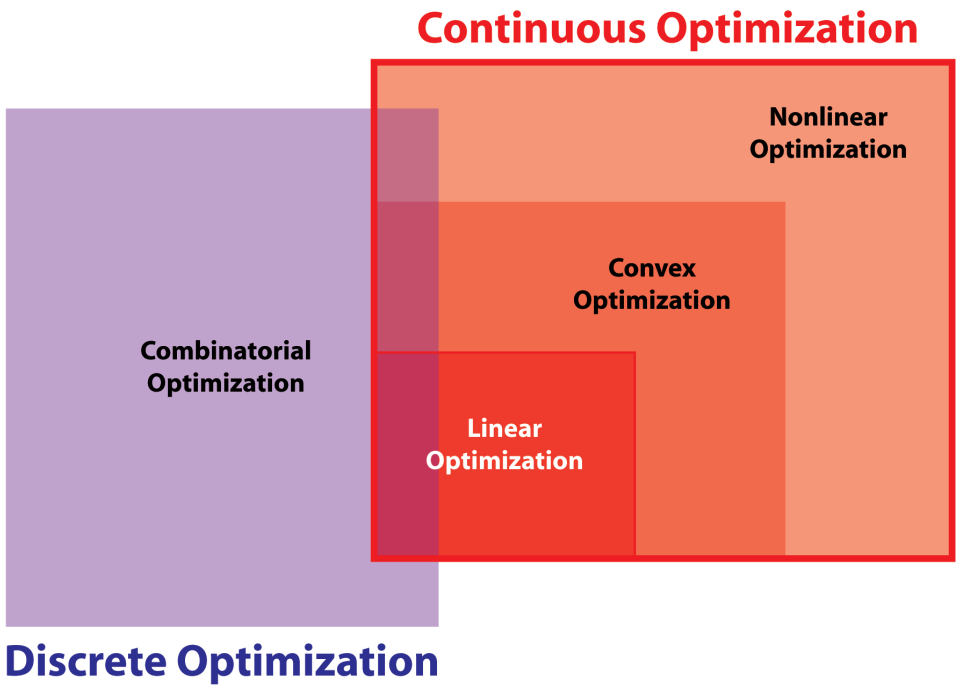
- Matrix Cookbook: https://www.math.uwaterloo.ca/~hwolkowi/matrixcookbook.pdf
- Linear Algebra Lectures by Zico Kolter: `http://www.cs.cmu.edu/~zkolter/course/linalg/index.html`
- Functional Analysis/Matrix Calculus Lectures by Aaditya Ramdas: `http://www.cs.cmu.edu/~aramdas/videos.html`

Outline

- What is the Course About
- Who Cares and Why
- Course Objective
- Convex Optimization History
- New Interest in the Topic
- Formal Introduction

What is the Course About?

- A special class of optimization (includes $Linear\ Programming$)



Who Cares and Why?

- Who?

  Anyone using or interested in computational aspects of optimization

- Why?
  - To understand the underlying basic terminology, principles, and methodology (to efficiently use the existing software tools)
  - To develop ability to modify tools when needed
  - To develop ability to design new algorithms or improve the efficiency of the existing ones

Course Objective

- The goal of this course is to provide you with working knowledge of convex optimization

- In particular, to provide you with skills and knowledge to
  - Recognize convex problems
  - Model problems as convex
  - Solve the problems

Convex Optimization History

- Convexity Theory and Analysis have being studied for a long time, mostly by mathematicians

- Until late 1980's:
  - Algorithmic development focused mainly on solving Linear Problems
    - Simplex Algorithm for linear programming (Dantzig, 1947)
    - Ellipsoid Method (Shor, 1970)
    - Interior-Point Methods for linear programming (Karmarkar, 1984)
  - Applications in operations research and few in engineering

- Since late 1980's: A new interest in Convex Optimization emerges

New Interest in the Topic

Recent developments stimulated a new interest in Convex Optimization

- The recognition that Interior-Point Methods can efficiently solve certain classes of convex problems, including semi-definite programs and second-order cone programs, almost *as easily as linear programs*
- The new technologies and their applications created a need for new models (convex models are often suitable)
- Convex Problems are now prevalent in practice
  - Automatic Control Systems
  - Estimation, Signal and Image Processing
  - Communication and Data Networks
  - Data Analysis and Modeling
  - Statistics and Finance

Formal Introduction

- Mathematical Formulation of Optimization
- Some Examples
- Solving Optimization Problems
  - Least-Squares
  - Linear Optimization
  - Convex Optimization
- Practical Example
- Ongoing Research in Convex Optimization

Mathematical Formulation of Optimization Problem

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & g_i(x) \leq 0, \quad i = 1, \ldots, m \\ & x \in X \end{array}$$

- Vector $x = (x_1, \ldots, x_n)$ represents optimization (decision) variables
- Function $f : \mathbb{R}^n \to \mathbb{R}$ is an objective function
- Functions $g_i : \mathbb{R}^n \to \mathbb{R}$, $i = 1, \ldots, m$ are constraint functions (represent inequality constraints)
- Set $X \subseteq \mathbb{R}^n$ is a constraint set

Optimal value: The smallest value of f among all vectors that satisfy the set and the inequality constraints

Optimal solution: A vector that achieves the optimal value of f and satisfies all the constraints

Some Examples

Communication Networks

- Variables: communication rates for users
- Constraints: link capacities
- Objective: user cost

Portfolio Optimization

- Variables: amounts invested in different assets
- Constraints: available budget, maximum/minimum investment per asset, minimum return, time constraints
- Objective: overall risk or return variance

Data Fitting

- Variables: model parameters
- Constraints: prior information, parameter limits
- Objective: measure of misfit or prediction error

Solving Optimization Problems

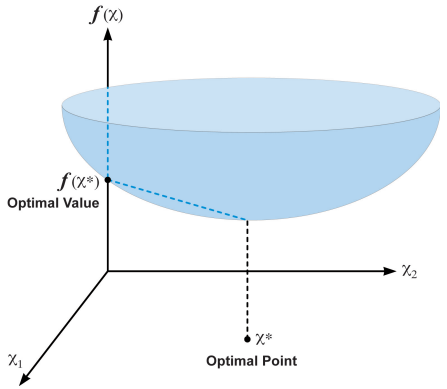
General Optimization Problem

- Very difficult to solve
- Existing methods involve trade offs between "time" and "accuracy", eg., very long computation time, or finding a sub-optimal solution

Exceptions: Certain problem classes can be solved efficiently and reliably

- Least-Squares Problems
- Linear Programming Problems
- Some classes of Convex Optimization Problems

Least-Squares



minimize $\|Ax - b\|^2$

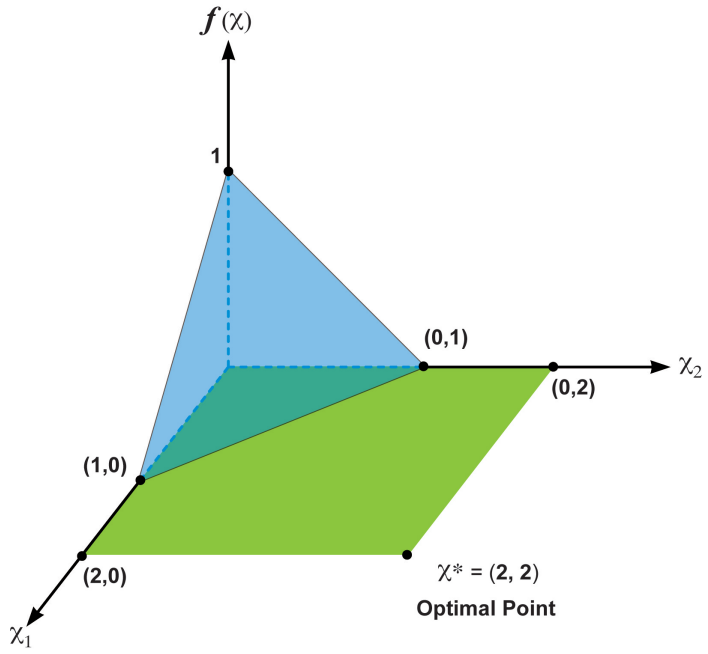
Solving Least-Squares Problems

- Analytical solution: $x^* = (A^T A)^{-1} A^T b$
- Reliable and efficient algorithms and software
- A mature technology
- Computation time proportional to $n^2 k$ ($A \in \mathbb{R}^{k \times n}$); less if structured

Using Least-Squares

- Least-squares problems are easy to recognize
- In regression analysis, optimal control, parameter estimation
- A few standard techniques increase its flexibility in applications (eg., including weights, regularization terms)

Linear Programming



minimize $c'x$

subject to $a_i'x \leq b_i, \quad 1 \leq i \leq m$

Solving Linear Programs

- No analytical solution
- Reliable and efficient algorithms and software
- A mature technology

Using Linear Programming

- Not as easy to recognize as least-squares problems (linear formulation possible but not always obvious)
- A few standard tricks used to convert problems into linear programs (eg., problems involving maximum norm, piecewise-linear functions)

Convex Optimization Problems

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & g_i(x) \leq 0, \quad i = 1, \ldots, m \\ & x \in X \end{array}$$

- Objective and constraint functions are convex
- Constraint set is convex
- Includes least-squares problems and linear programs as special cases

Solving Convex Optimization Problems

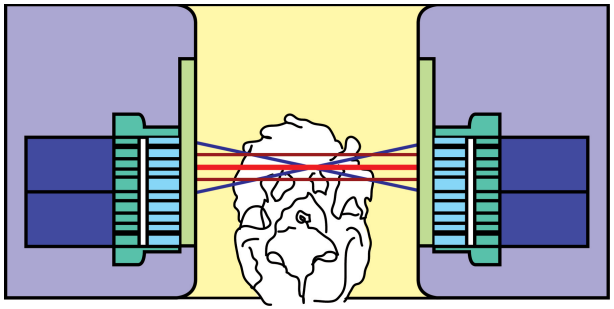
- No analytical solution
- Reliable and efficient algorithms for some classes
- Computation time (roughly) proportional to $\max\{n^3, n^2m, G\}$, where G is a cost of evaluating g_i's and their first and second derivatives
- Almost a technology

Using Convex Optimization

- Often difficult to recognize
- Many tricks for transforming problems into convex form
- Many practical problems can be modeled as convex optimization

Practical Example

Image Reconstruction in PET-scan [Ben-Tal, 2005]



- Maximum Likelihood Model results in convex optimization

$$\min_{x \geq 0,\ e'x \leq 1} \left\{ -\sum_{i=1}^{m} y_i \ln \left(\sum_{j=1}^{n} p_{ij} x_j \right) \right\}$$

  - x is a decision vector
  - y models measured data (by PET detectors)
  - p_{ij} probabilities modeling detections of emitted positrons

Ongoing Research in Convex Optimization and Beyond

- Distributed computations for large scale (nonsmooth) convex problems
- Approximation schemes with rate and error estimates
- Extending the methodology to non-convex problems

  Deep learning methods

Homework:

Download:

http://www.lix.polytechnique.fr/bigdata/mathbigdata/wp-content/uploads/2014/10/Lnotes_CvxAn_FullEn.pdf

Read:

1. Chapter 1
2. Chapter 2

Introduction: Why Optimization?

Convex Optimization

Prerequisites: no formal ones, but class will be fairly fast paced

Assume working knowledge of/proficiency with:

- Real analysis, calculus, linear algebra
- Core problems in Machine Learning and Statistics
- Programming (R, Python, Julia, your choice ...)
- Data structures, computational complexity
- Formal mathematical thinking

If you fall short on any one of these things, it's certainly possible to catch up; but don't hesitate to talk to us

Optimization in Machine Learning and Statistics

Optimization problems underlie nearly everything we do in Machine Learning and Statistics. In other courses, you learn how to:

translate  into $P : \min_{x \in D} f(x)$

Conceptual idea *Optimization problem*

Examples of this? Examples of the contrary?

This course: how to solve P, and why this is a good skill to have

Motivation: why do we bother?

Presumably, other people have already figured out how to solve

$$P \; : \; \min_{x \in D} \; f(x)$$

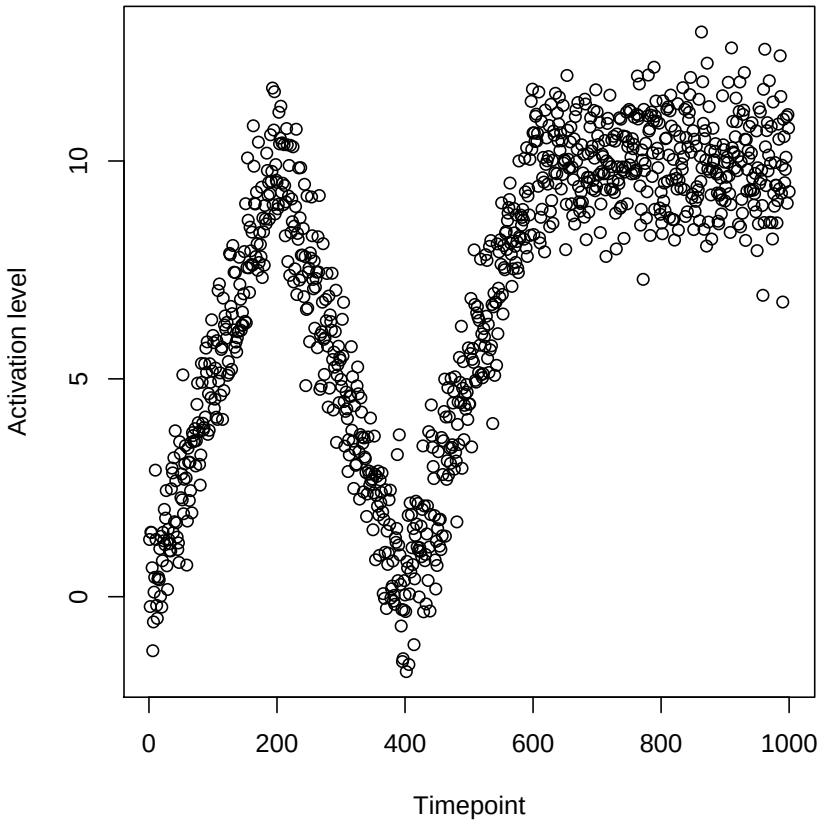
So why bother? Many reasons. Here's three:

1. Different algorithms can perform better or worse for different problems P (sometimes drastically so)
2. Studying P through an optimization lens can actually give you a deeper understanding of the task/procedure at hand
3. Knowledge of optimization can actually help you create a new problem P that is even more interesting/useful

Optimization moves quickly as a field. But there is still much room for progress, especially its intersection with ML and Stats

Example: algorithms for linear trend filtering

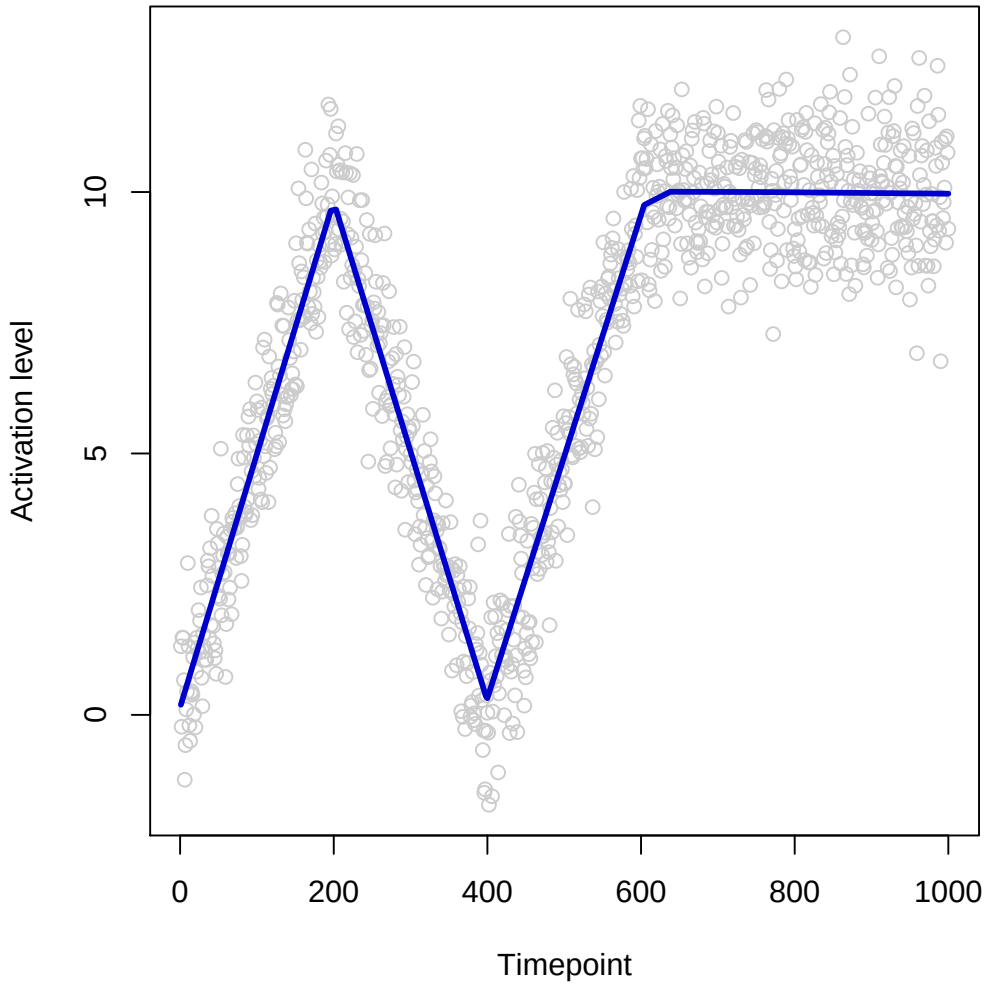
Given observations $y_i \in \mathbb{R}$, $i = 1, \ldots, n$ corresponding to underlying locations $x_i = i$, $i = 1, \ldots, n$



Linear trend filtering fits a piecewise linear function, with adaptively chosen knots (Steidl et al. 2006, Kim et al. 2009)

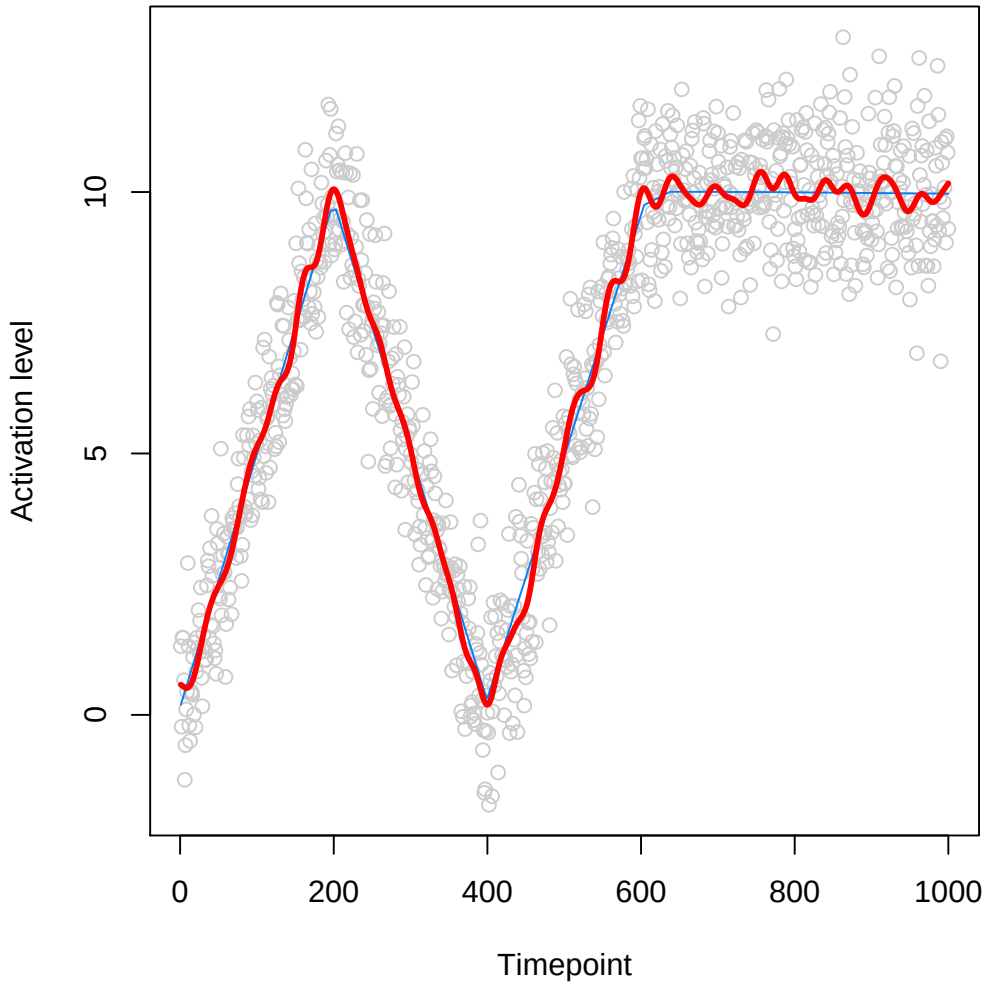
How? By solving $\min_\theta \frac{1}{2}\sum_{i=1}^{n}(y_i - \theta_i)^2 + \lambda \sum_{i=1}^{n-2} |\theta_i - 2\theta_{i+1} + \theta_{i+2}|$

Problem: $$\min_{\theta} \frac{1}{2}\sum_{i=1}^{n}(y_i-\theta_i)^2+\lambda\sum_{i=1}^{n-2}|\theta_i-2\theta_{i+1}+\theta_{i+2}|$$



Interior point method,
20 iterations

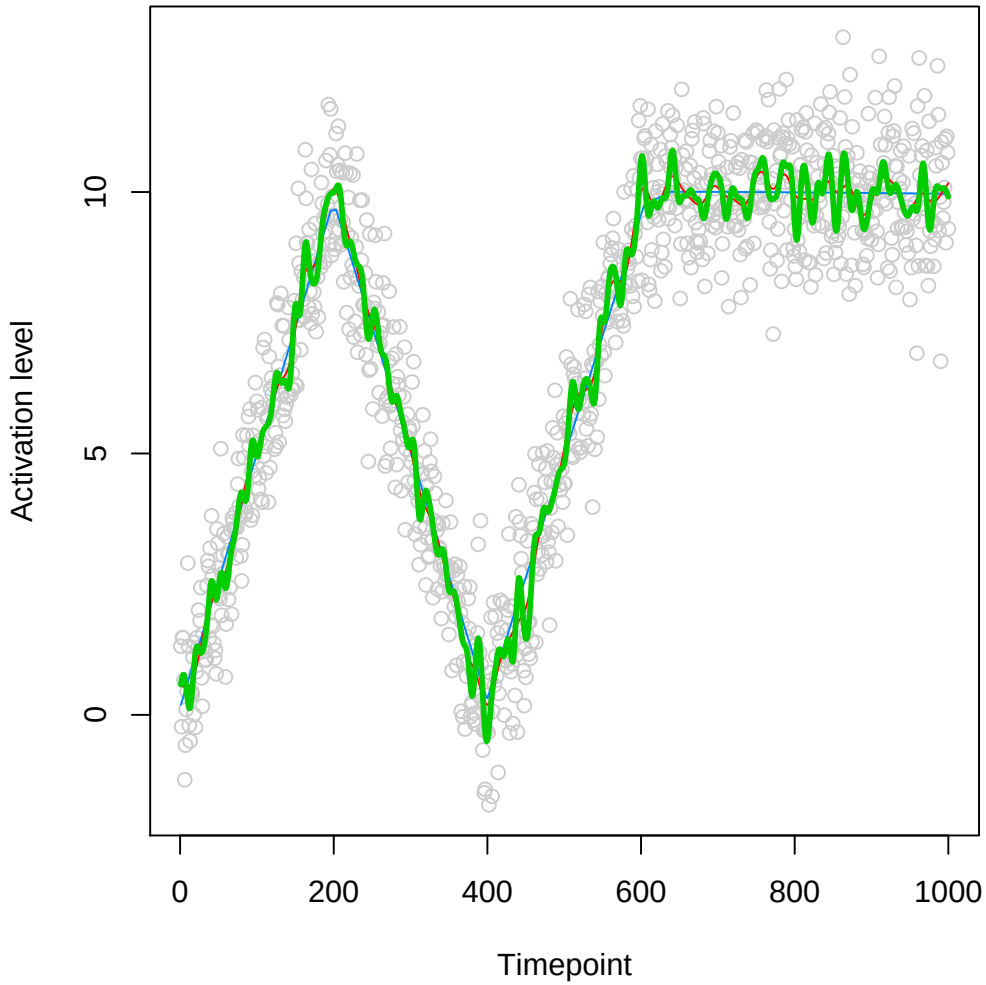
Problem: $$\min_{\theta} \frac{1}{2}\sum_{i=1}^{n}(y_i - \theta_i)^2 + \lambda \sum_{i=1}^{n-2} |\theta_i - 2\theta_{i+1} + \theta_{i+2}|$$



Interior point method, 20 iterations

Proximal gradient descent, 10K iterations

Problem: $$\min_{\theta} \frac{1}{2}\sum_{i=1}^{n}(y_i-\theta_i)^2+\lambda\sum_{i=1}^{n-2}|\theta_i-2\theta_{i+1}+\theta_{i+2}|$$

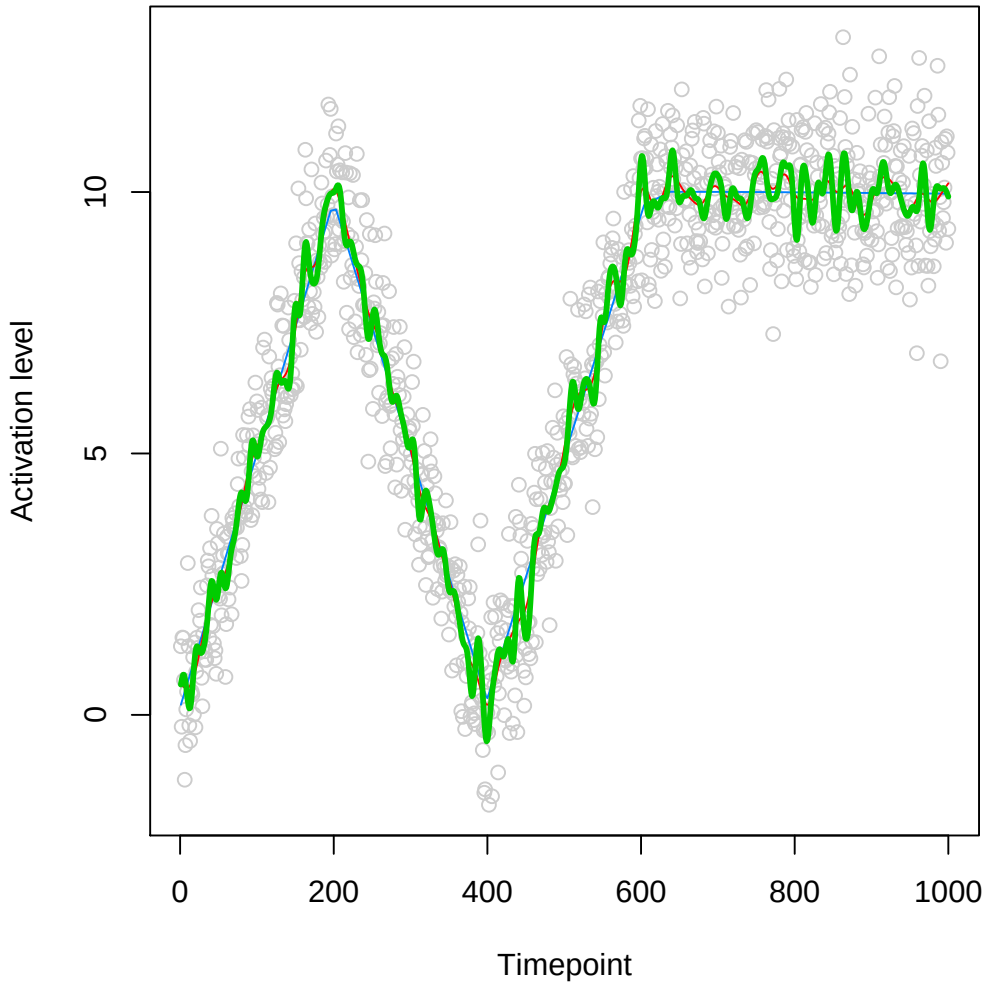


Interior point method, 20 iterations

Proximal gradient descent, 10K iterations

Coordinate descent, 1000 cycles

Problem: $$\min_{\theta} \frac{1}{2}\sum_{i=1}^{n}(y_i-\theta_i)^2+\lambda\sum_{i=1}^{n-2}|\theta_i-2\theta_{i+1}+\theta_{i+2}|$$



Interior point method, 20 iterations

Proximal gradient descent, 10K iterations

Coordinate descent, 1000 cycles

(all from the dual)

What's the message here?

So what's the right conclusion here?

Is primal-dual interior point method simply a better method than proximal gradient descent, coordinate descent? ... No

In fact, different algorithms will work better in different situations. We'll learn details throughout the course

In the linear trend filtering problem:

- Primal-dual: fast (structured linear systems)
- Proximal gradient: slow (conditioning)
- Coordinate descent: slow (large active set)

Central concept: convexity

Historically, linear programs were the focus in optimization

Initially, it was thought that the important distinction was between linear and nonlinear optimization problems. But some nonlinear problems turned out to be much harder than others ...

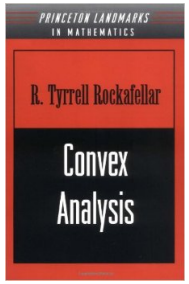
Now it is widely recognized that the right distinction is between convex and nonconvex problems

Your supplementary textbooks for the course:

Boyd and Vandenberghe (2004)

and

Rockafellar (1970)



Wisdom from Rockafellar (1993)

From Terry Rockafellar's 1993 SIAM Review survey paper:

> a convex set every locally optimal solution is global. Also, first-order necessary conditions for optimality turn out to be sufficient. A variety of other properties conducive to computation and interpretation of solutions ride on convexity as well. In fact the great watershed in optimization isn't between linearity and nonlinearity, but convexity and nonconvexity. Even for problems that aren't themselves of convex type, convexity may enter, for instance, in setting up subproblems as part of an iterative numerical scheme.

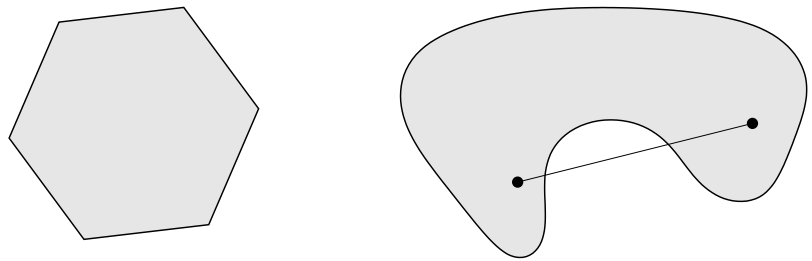
Credit to Nemirovski, Yudin, Nesterov, others for formalizing this

This view was dominant both within the optimization community and in many application domains for many decades (... currently being challenged by successes of neural networks?)

Convex sets and functions

Convex set: $C \subseteq \mathbb{R}^n$ such that

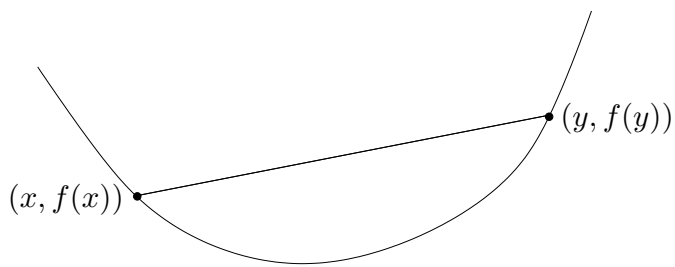
$$x, y \in C \implies tx + (1-t)y \in C \text{ for all } 0 \leq t \leq 1$$



Convex function: $f : \mathbb{R}^n \to \mathbb{R}$ such that $\mathrm{dom}(f) \subseteq \mathbb{R}^n$ convex, and

$$f(tx + (1-t)y) \leq tf(x) + (1-t)f(y) \quad \text{for all } 0 \leq t \leq 1$$

and all $x, y \in \mathrm{dom}(f)$



Convex optimization problems

Optimization problem:

$$\begin{array}{ll} \min\limits_{x \in D} & f(x) \\ \text{subject to} & g_i(x) \leq 0,\ i = 1, \ldots, m \\ & h_j(x) = 0,\ j = 1, \ldots, r \end{array}$$

Here $D = \mathrm{dom}(f) \cap \bigcap_{i=1}^m \mathrm{dom}(g_i) \cap \bigcap_{j=1}^p \mathrm{dom}(h_j)$, common domain of all the functions

This is a convex optimization problem provided the functions f and $g_i, i = 1, \ldots, m$ are convex, and $h_j, j = 1, \ldots, p$ are affine:

$$h_j(x) = a_j^T x + b_j, \quad j = 1, \ldots, p$$

Local minima are global minima

For convex optimization problems, local minima are global minima

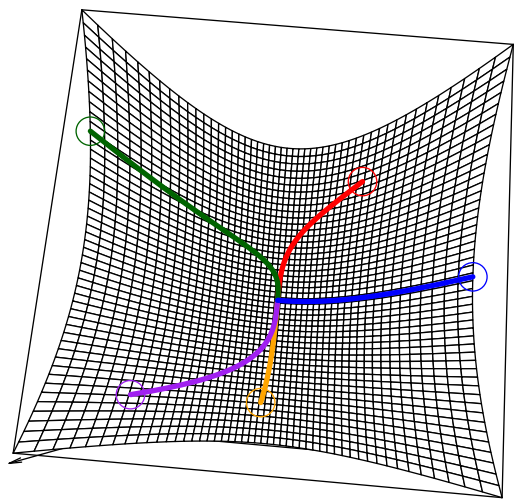
Formally, if x is feasible—$x \in D$, and satisfies all constraints—and minimizes f in a local neighborhood,

$$f(x) \leq f(y) \text{ for all feasible } y,\ \|x - y\|_2 \leq \rho,$$

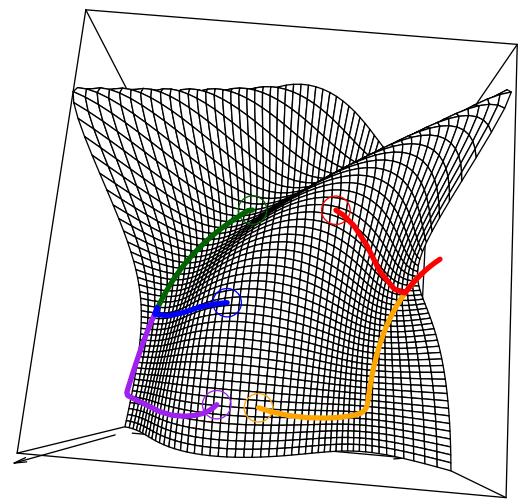
then

$$f(x) \leq f(y) \text{ for all feasible } y$$

This is a very useful fact and will save us a lot of trouble!



Convex



Nonconvex

Convexity I: Sets and Functions

Convex Optimization

See supplements for reviews of

- *basic real analysis*
- *basic multivariate calculus*
- *basic linear algebra*

Last time: why convexity?

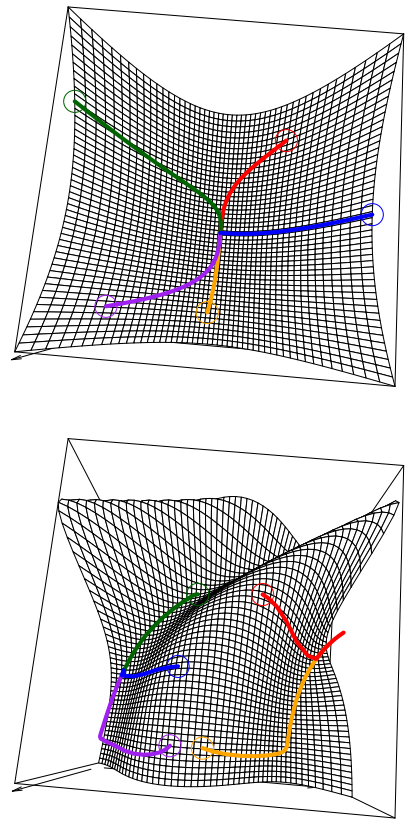
Why convexity? Simply put: because we can broadly understand and solve convex optimization problems

Nonconvex problems are mostly treated on a case by case basis

Reminder: a convex optimization problem is of the form

$$\begin{aligned} \min_{x \in D} \quad & f(x) \\ \text{subject to} \quad & g_i(x) \leq 0, \; i = 1, \ldots, m \\ & h_j(x) = 0, \; j = 1, \ldots, r \end{aligned}$$

where f and g_i, $i = 1, \ldots, m$ are all convex, and h_j, $j = 1, \ldots, r$ are affine. Special property: any local minimizer is a global minimizer



Outline

Today:

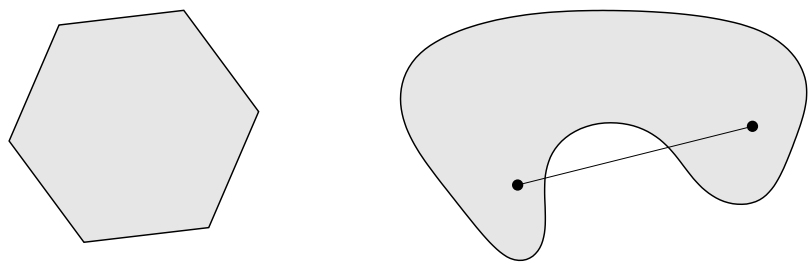
- Convex sets
- Examples
- Key properties
- Operations preserving convexity
- Same, for convex functions

Convex sets

Convex set: $C \subseteq \mathbb{R}^n$ such that

$$x, y \in C \implies tx + (1-t)y \in C \text{ for all } 0 \leq t \leq 1$$

In words, line segment joining any two elements lies entirely in set



Convex combination of $x_1, \ldots, x_k \in \mathbb{R}^n$: any linear combination

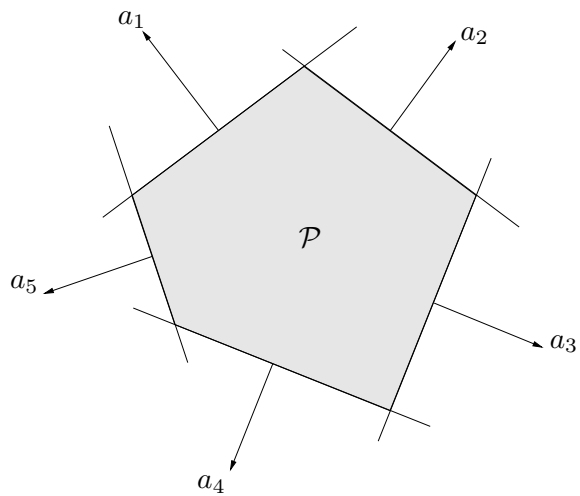
$$\theta_1 x_1 + \cdots + \theta_k x_k$$

with $\theta_i \geq 0$, $i = 1, \ldots, k$, and $\sum_{i=1}^k \theta_i = 1$. Convex hull of a set C, $\mathrm{conv}(C)$, is all convex combinations of elements. Always convex

Examples of convex sets

- Trivial ones: empty set, point, line
- Norm ball: $\{x : \|x\| \leq r\}$, for given norm $\|\cdot\|$, radius r
- Hyperplane: $\{x : a^T x = b\}$, for given a, b
- Halfspace: $\{x : a^T x \leq b\}$
- Affine space: $\{x : Ax = b\}$, for given A, b

- Polyhedron: $\{x : Ax \leq b\}$, where inequality $\leq$ is interpreted componentwise. Note: the set $\{x : Ax \leq b, Cx = d\}$ is also a polyhedron (why?)



- Simplex: special case of polyhedra, given by $\text{conv}\{x_0, \ldots, x_k\}$, where these points are affinely independent. The canonical example is the probability simplex,

$$\text{conv}\{e_1, \ldots, e_n\} = \{w : w \geq 0,\ 1^T w = 1\}$$

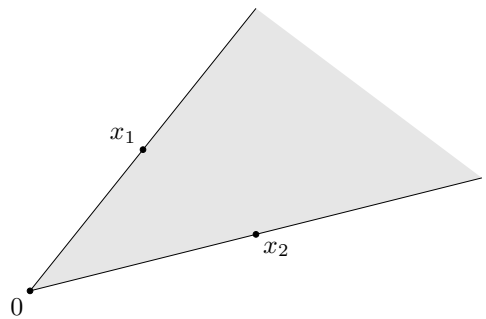
Cones

Cone: $C \subseteq \mathbb{R}^n$ such that

$$x \in C \implies tx \in C \text{ for all } t \geq 0$$

Convex cone: cone that is also convex, i.e.,

$$x_1, x_2 \in C \implies t_1 x_1 + t_2 x_2 \in C \text{ for all } t_1, t_2 \geq 0$$



Conic combination of $x_1, \ldots, x_k \in \mathbb{R}^n$: any linear combination

$$\theta_1 x_1 + \cdots + \theta_k x_k$$

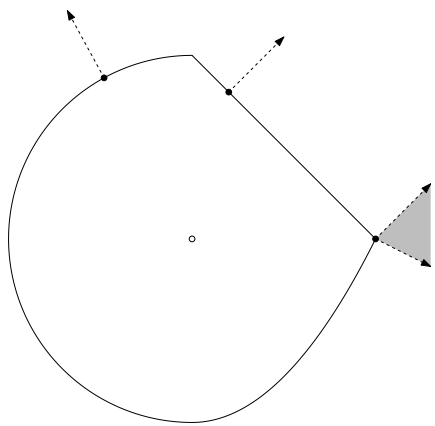
with $\theta_i \geq 0$, $i = 1, \ldots, k$. Conic hull collects all conic combinations

Examples of convex cones

- Norm cone: $\{(x,t) : \|x\| \leq t\}$, for a norm $\|\cdot\|$. Under the ℓ_2 norm $\|\cdot\|_2$, called second-order cone

- Normal cone: given any set C and point $x \in C$, we can define

$$\mathcal{N}_C(x) = \{g : g^T x \geq g^T y, \text{ for all } y \in C\}$$

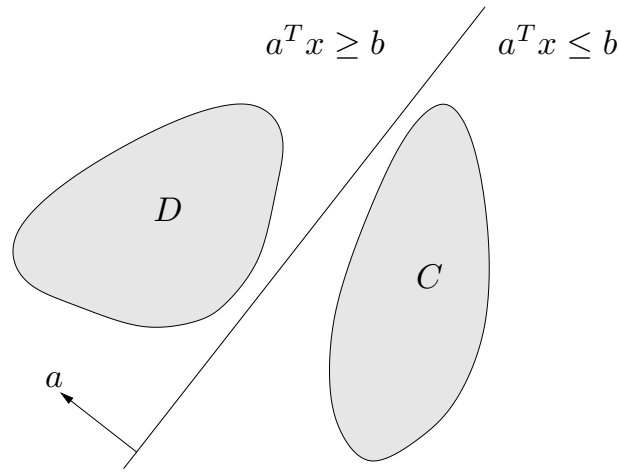


This is always a convex cone, regardless of C

- Positive semidefinite cone: $\mathbb{S}^n_+ = \{X \in \mathbb{S}^n : X \succeq 0\}$, where $X \succeq 0$ means that X is positive semidefinite (and $\mathbb{S}^n$ is the set of $n \times n$ symmetric matrices)

Key properties of convex sets

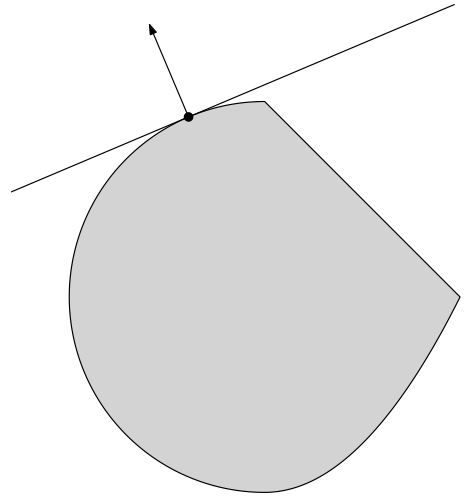
- Separating hyperplane theorem: two disjoint convex sets have a separating between hyperplane them



Formally: if C, D are nonempty convex sets with $C \cap D = \emptyset$, then there exists a, b such that

$$C \subseteq \{x : a^T x \leq b\}$$
$$D \subseteq \{x : a^T x \geq b\}$$

- Supporting hyperplane theorem: a boundary point of a convex set has a supporting hyperplane passing through it



Formally: if C is a nonempty convex set, and $x_0 \in \mathrm{bd}(C)$, then there exists a such that

$$C \subseteq \{x : a^T x \leq a^T x_0\}$$

Both of the above theorems (separating and supporting hyperplane theorems) have partial converses; see Section 2.5 of BV

Example: linear matrix inequality solution set

Given $A_1, \ldots, A_k, B \in \mathbb{S}^n$, a linear matrix inequality is of the form

$$x_1 A_1 + x_2 A_2 + \cdots + x_k A_k \preceq B$$

for a variable $x \in \mathbb{R}^k$. Let's prove the set C of points x that satisfy the above inequality is convex

Approach 1: directly verify that $x, y \in C \Rightarrow tx + (1-t)y \in C$. This follows by checking that, for any v,

$$v^T \left(B - \sum_{i=1}^{k} (tx_i + (1-t)y_i) A_i \right) v \geq 0$$

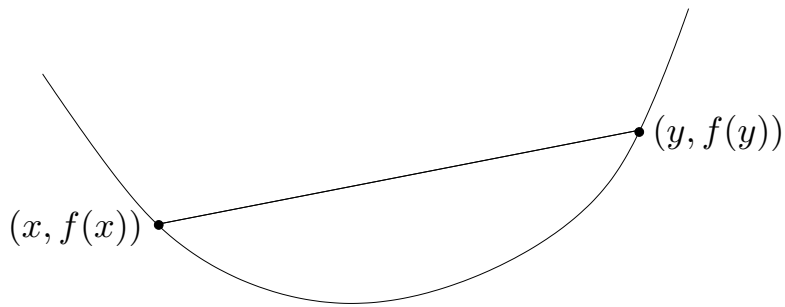
Approach 2: let $f : \mathbb{R}^k \to \mathbb{S}^n$, $f(x) = B - \sum_{i=1}^{k} x_i A_i$. Note that $C = f^{-1}(\mathbb{S}^n_+)$, affine preimage of convex set

Convex functions

Convex function: $f : \mathbb{R}^n \to \mathbb{R}$ such that $\mathrm{dom}(f) \subseteq \mathbb{R}^n$ convex, and

$$f(tx + (1-t)y) \leq tf(x) + (1-t)f(y) \quad \text{for } 0 \leq t \leq 1$$

and all $x, y \in \mathrm{dom}(f)$



In words, function lies below the line segment joining $f(x), f(y)$

Concave function: opposite inequality above, so that

$$f \text{ concave} \iff -f \text{ convex}$$

Important modifiers:

- Strictly convex: $f\big(tx + (1-t)y\big) < tf(x) + (1-t)f(y)$ for $x \neq y$ and $0 < t < 1$. In words, f is convex and has greater curvature than a linear function

- Strongly convex with parameter $m > 0$: $f - \frac{m}{2}\|x\|_2^2$ is convex. In words, f is at least as convex as a quadratic function

Note: strongly convex $\Rightarrow$ strictly convex $\Rightarrow$ convex

(Analogously for concave functions)

Examples of convex functions

- Univariate functions:
  - Exponential function: e^{ax} is convex for any a over $\mathbb{R}$
  - Power function: x^a is convex for $a \geq 1$ or $a \leq 0$ over $\mathbb{R}_+$ (nonnegative reals)
  - Power function: x^a is concave for $0 \leq a \leq 1$ over $\mathbb{R}_+$
  - Logarithmic function: $\log x$ is concave over $\mathbb{R}_{++}$
- Affine function: $a^T x + b$ is both convex and concave
- Quadratic function: $\frac{1}{2} x^T Q x + b^T x + c$ is convex provided that $Q \succeq 0$ (positive semidefinite)
- Least squares loss: $\|y - Ax\|_2^2$ is always convex (since $A^T A$ is always positive semidefinite)

- Norm: $\|x\|$ is convex for any norm; e.g., ℓ_p norms,

$$\|x\|_p = \left(\sum_{i=1}^{n} |x_i|^p \right)^{1/p} \quad \text{for } p \geq 1, \quad \|x\|_\infty = \max_{i=1,\ldots,n} |x_i|$$

and also operator (spectral) and trace (nuclear) norms,

$$\|X\|_{\text{op}} = \sigma_1(X), \quad \|X\|_{\text{tr}} = \sum_{i=1}^{r} \sigma_r(X)$$

where $\sigma_1(X) \geq \ldots \geq \sigma_r(X) \geq 0$ are the singular values of the matrix X

- Indicator function: if C is convex, then its indicator function

$$I_C(x) = \begin{cases} 0 & x \in C \\ \infty & x \notin C \end{cases}$$

  is convex

- Support function: for any set C (convex or not), its support function

$$I_C^*(x) = \max_{y \in C} \; x^T y$$

  is convex

- Max function: $f(x) = \max\{x_1, \ldots, x_n\}$ is convex

Key properties of convex functions

- A function is convex if and only if its restriction to any line is convex
- Epigraph characterization: a function f is convex if and only if its epigraph

$$\text{epi}(f) = \{(x, t) \in \text{dom}(f) \times \mathbb{R} : f(x) \leq t\}$$

  is a convex set
- Convex sublevel sets: if f is convex, then its sublevel sets

$$\{x \in \text{dom}(f) : f(x) \leq t\}$$

  are convex, for all $t \in \mathbb{R}$. The converse is not true

- First-order characterization: if f is differentiable, then f is convex if and only if $\mathrm{dom}(f)$ is convex, and

$$f(y) \geq f(x) + \nabla f(x)^T (y - x)$$

  for all $x, y \in \mathrm{dom}(f)$. Therefore for a differentiable convex function $\nabla f(x) = 0 \iff x$ minimizes f

- Second-order characterization: if f is twice differentiable, then f is convex if and only if $\mathrm{dom}(f)$ is convex, and $\nabla^2 f(x) \succeq 0$ for all $x \in \mathrm{dom}(f)$

- Jensen's inequality: if f is convex, and X is a random variable supported on $\mathrm{dom}(f)$, then $f(\mathbb{E}[X]) \leq \mathbb{E}[f(X)]$

Operations preserving convexity

- Nonnegative linear combination: $f_1, \ldots, f_m$ convex implies $a_1 f_1 + \cdots + a_m f_m$ convex for any $a_1, \ldots, a_m \geq 0$
- Pointwise maximization: if f_s is convex for any $s \in S$, then $f(x) = \max_{s \in S} f_s(x)$ is convex. Note that the set S here (number of functions f_s) can be infinite
- Partial minimization: if $g(x, y)$ is convex in x, y, and C is convex, then $f(x) = \min_{y \in C} g(x, y)$ is convex

Example: distances to a set

Let C be an arbitrary set, and consider the maximum distance to C under an arbitrary norm $\|\cdot\|$:

$$f(x) = \max_{y \in C} \|x - y\|$$

Let's check convexity: $f_y(x) = \|x - y\|$ is convex in x for any fixed y, so by pointwise maximization rule, f is convex

Now let C be convex, and consider the minimum distance to C:

$$f(x) = \min_{y \in C} \|x - y\|$$

Let's check convexity: $g(x, y) = \|x - y\|$ is convex in x, y jointly, and C is assumed convex, so apply partial minimization rule

More operations preserving convexity

- Affine composition: if f is convex, then $g(x) = f(Ax + b)$ is convex
- General composition: suppose $f = h \circ g$, where $g : \mathbb{R}^n \to \mathbb{R}$, $h : \mathbb{R} \to \mathbb{R}$, $f : \mathbb{R}^n \to \mathbb{R}$. Then:
  - f is convex if h is convex and nondecreasing, g is convex
  - f is convex if h is convex and nonincreasing, g is concave
  - f is concave if h is concave and nondecreasing, g concave
  - f is concave if h is concave and nonincreasing, g convex

  How to remember these? Think of the chain rule when $n = 1$:

$$f''(x) = h''(g(x))g'(x)^2 + h'(g(x))g''(x)$$

- Vector composition: suppose that

$$f(x) = h\big(g(x)\big) = h\big(g_1(x), \ldots, g_k(x)\big)$$

where $g : \mathbb{R}^n \to \mathbb{R}^k$, $h : \mathbb{R}^k \to \mathbb{R}$, $f : \mathbb{R}^n \to \mathbb{R}$. Then:

  - f is convex if h is convex and nondecreasing in each argument, g is convex
  - f is convex if h is convex and nonincreasing in each argument, g is concave
  - f is concave if h is concave and nondecreasing in each argument, g is concave
  - f is concave if h is concave and nonincreasing in each argument, g is convex

Example: log-sum-exp function

Log-sum-exp function: $g(x) = \log(\sum_{i=1}^k e^{a_i^T x + b_i})$, for fixed a_i, b_i, $i = 1, \ldots, k$. Often called "soft max", as it smoothly approximates $\max_{i=1,\ldots k} (a_i^T x + b_i)$

How to show convexity? First, note it suffices to prove convexity of $f(x) = \log(\sum_{i=1}^n e^{x_i})$ (affine composition rule)

Now use second-order characterization. Calculate

$$\nabla_i f(x) = \frac{e^{x_i}}{\sum_{\ell=1}^n e^{x_\ell}}$$

$$\nabla_{ij}^2 f(x) = \frac{e^{x_i}}{\sum_{\ell=1}^n e^{x_\ell}} 1\{i = j\} - \frac{e^{x_i} e^{x_j}}{(\sum_{\ell=1}^n e^{x_\ell})^2}$$

Write $\nabla^2 f(x) = \mathrm{diag}(z) - zz^T$, where $z_i = e^{x_i}/(\sum_{\ell=1}^n e^{x_\ell})$. This matrix is diagonally dominant, hence positive semidefinite

References and further reading

- S. Boyd and L. Vandenberghe (2004), "Convex optimization", Chapters 2 and 3
- J.P. Hiriart-Urruty and C. Lemarechal (1993), "Fundamentals of convex analysis", Chapters A and B
- R. T. Rockafellar (1970), "Convex analysis", Chapters 1–10,

Optimization Theory: Reading materials Spring 2022

Lecture: Convex Functions – Jan 18-19, 2022

Overview In these two lectures, we will introduce the concept of convex functions, and provide several ways to characterize convex functions, discuss some calculus that can be used to detect convexity of functions and compute the subgradients of convex function.

3.1 Definitions

Definition 3.1 (Convex function) *A function $f(x) : \mathbf{R}^n \to \mathbf{R}$ is convex if*

(i) $dom(f) \subseteq \mathbf{R}^n$ *s a convex set;*

(ii) $\forall x, y \in dom(f)$ *and* $\lambda \in [0, 1]$, $f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda)f(y)$.

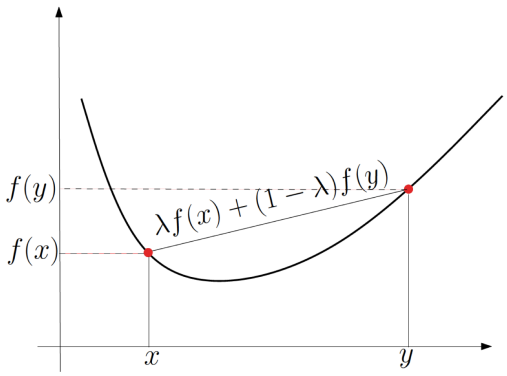


Figure 3.1: Example of convex function

A function is called strictly convex if (ii) holds with strict sign, i.e. $f(\lambda x + (1-\lambda)y) < \lambda f(x) + (1-\lambda)f(y)$.

A function is called α-strictly convex if $f(x) - \frac{\alpha}{2}\|x\|_2^2$ is convex.

A function is called concave if $-f(x)$ is convex.

For example, a linear function is both convex and concave. Any norm is convex.

Remark 1 (Extended value function). Conventionally, we can think of f as an extended value function from $\mathbf{R}^n$ to $\mathbf{R} \cup \{+\infty\}$ by setting $f(x) = +\infty$ if $x \notin dom(f)$, the condition (ii) is equivalent as

$$\forall x, y, \forall \lambda \in [0, 1], f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda)f(y).$$

Remark 2. (Slope inequality) What does convexity really mean? Let $z = \lambda x + (1-\lambda)y$, observe that $\|y-x\| : \|y-z\| : \|z-x\| = 1 : \lambda : (1-\lambda)$. Therefore

$$\begin{aligned} & f(z) \le \lambda f(x) + (1-\lambda) f(y) \\ \iff & \frac{f(z)-f(x)}{1-\lambda} \le f(y) - f(x) \le \frac{f(y)-f(z)}{\lambda} \\ \iff & \frac{f(z)-f(x)}{\|z-x\|} \le \frac{f(y)-f(x)}{\|y-x\|} \le \frac{f(y)-f(z)}{\|y-z\|} \end{aligned}$$

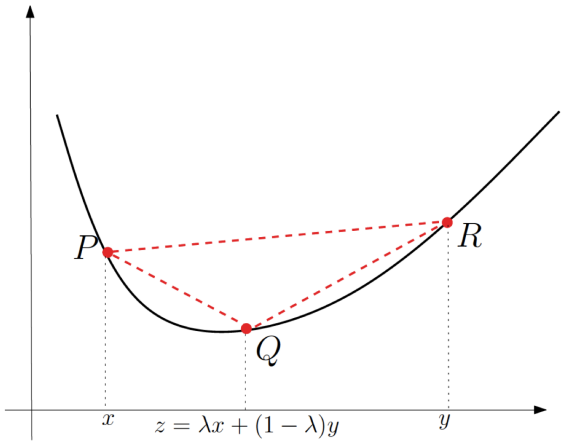


Figure 3.2: Slope PQ $\le$ Slope PR $\le$ Slope QR

3.2 Several Characterizations of Convex Functions

1. **Epigraph characterization**

   Proposition 3.2 *f is convex if and only if its epigraph*

   $$epi(f) := \{(x,t) \in \mathbf{R}^{n+1} : f(x) \le t\}$$

   is a convex set.

   Proof: This can be verified by using the definition of convex function and convex set.

   - ($\Longrightarrow$) Suppose $(x,t_1), (y,t_2) \in epi(f)$, then $f(x) \le t_1, f(y) \le t_2$. For any $\lambda \in [0,1]$, by convexity of f, $f(\lambda x + (1-\lambda)y) \le \lambda f(x) + (1-\lambda) f(y) \le \lambda t_1 + (1-\lambda) t_2$. Thi implies that $\lambda \cdot (x,t_1) + (1-\lambda) \cdot (y,t_2) \in epi(f)$. Hence, $epi(f)$ is a convex set.
   - ($\Longleftarrow$) Let $x, y \in \mathbf{R}^n$, since $(x, f(x))$ and $(y, f(y))$ lie in $epi(f)$, by convexity of epigraph set, we have for any $\lambda \in [0,1]$, $(\lambda x + (1-\lambda)y, \lambda(x) + (1-\lambda) f(y)) \in epi(f)$. By definition, $f(\lambda x + (1-\lambda)y) \le \lambda f(x) + (1-\lambda) f(y)$. Hence, function f is convex.

   ■

2. **Level set characterization**

   Proposition 3.3 *If f is convex, then the level set for any $t \in \mathbf{R}$*

   $$C_t(f) = \{x \in dom(f) : f(x) \leq t\}$$

   is a convex set.

   For example, the unit norm ball $\{x : \|x\| \leq 1\}$ is a convex set since $\|\cdot\|$ is convex.

   Remark. The reverse is not true. A function with convex level set is not always convex. In fact, it is known as a quasi-convex function.

3. **One-dimensional characterization**

   Proposition 3.4 *f is convex if and only if its restriction on any line, i.e. function*

   $$\phi(t) := f(x + th)$$

   is convex on the axis for any x and h.

   Remark. Convexity is a one-dimensional property. In order to detect the convexity of a function, it all boils down to check the convexity of a one-dimensional function on the axis. From basic calculus, we already know that

   $$\begin{aligned}
   & \phi(t) \text{ is convex on } (a, b) \\
   \iff \quad & \frac{\phi(s) - \phi(t_1)}{s - t_1} \leq \frac{\phi(t_2) - \phi(t_1)}{t_2 - t_1} \leq \frac{\phi(t_2) - \phi(s)}{t_2 - s}, \forall a < t_1 < s < t_2 < b && \text{(due to slope inequality)} \\
   \iff \quad & \phi'(t_1) \leq \phi'(t_2), \forall a < t_1 < t_2 < b && \text{(if } \phi \text{ is differentiable)} \\
   \iff \quad & \phi''(t) > 0, \forall a < t < b && \text{(if } \phi \text{ is twice-differentiable)}
   \end{aligned}$$

   Hence, if f is differentiable or twice-differentiable, we can characterize it by based on its first-order or second-order.

4. **First-order characterization for differentiable convex functions**

   Proposition 3.5 *Assume f is differentiable, then f is convex if and only if $dom(f)$ is convex and for any x, y,*

   $$f(x) \geq f(y) + \nabla f(y)^T (x - y). \qquad (\star)$$

   Proof:

   - ($\Longrightarrow$) If f is convex, letting $z = (1-\epsilon)y + \epsilon x = y + \epsilon(x - y)$ with $\epsilon \in (0, 1)$, from the slop inequality, we have
     $$\frac{f(x) - f(y)}{\|x - y\|} \geq \frac{f(z) - f(y)}{\|z - y\|} = \frac{f(y + \epsilon(x - y)) - f(y)}{\epsilon \|x - y\|}.$$
     Hence, letting $\epsilon \to 0+$, we have
     $$f(x) - f(y) \geq \lim_{\epsilon \to 0+} \frac{f(y + \epsilon(x - y)) - f(y)}{\epsilon} = \nabla f(y)^T (x - y).$$
     Therefore, $f(y) \geq f(x) + \nabla f(x)^T (y - x)$.

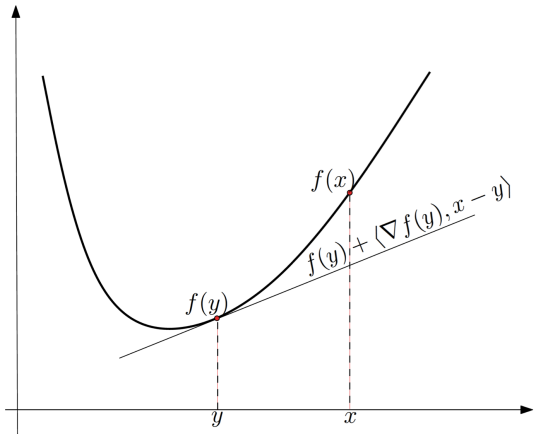


Figure 3.3: First-order condition

- ($\Longleftarrow$) If ($\star$) holds, letting $z = \lambda x + (1-\lambda)y$ for any $\lambda \in [0,1]$, we have
$$f(x) \geq f(z) + \nabla f(z)^T (x - z)$$
$$f(y) \geq f(z) + \nabla f(z)^T (y - z)$$
Adding the two inequalities with scalings λ and $(1-\lambda)$, it follows that
$$\lambda f(x) + (1-\lambda) f(y) \geq f(z) = f(\lambda x + (1-\lambda)y).$$
Hence, f is convex.

■

5. **Second-order characterization for twice-differentiable convex functions**

Proposition 3.6 *Assume f is twice-differentiable, then f is convex if and only if $dom(f)$ is convex and for any $x \in dom(f)$,*
$$\nabla^2 f(x) \succeq 0. \qquad (\star\star)$$

Proof:

- ($\Longrightarrow$) If f is convex, then for any x, h, $\phi(t) = f(x+th)$ is convex on the axis. Hence, $\phi''(t) \geq 0, \forall t$. Particularly,
$$\phi''(0) = h^T \nabla^2 f(x) h \geq 0.$$
This implies that $\nabla^2 f(x) \succeq 0$.
- ($\Longleftarrow$) It suffices to show that every one-dimensional function $\phi(t) := f(x + t(y-x))$ is convex for any $x, y \in dom(f)$. The latter is indeed true because $\phi''(t) = (y-x)^T \nabla^2 f(x+t(y-x))(y-x) \geq 0$ due to ($\star\star$).

■

6. **Subgradient characterization for non-differentiable convex functions**

Proposition 3.7 *f is convex if and only if $\forall x \in int(dom(f))$, there exists g, such that*
$$f(x) \geq f(y) + g^T (x - y)$$
i.e. the subdifferential set is non-empty.

To be discussed in Section 3.5.

3.3 Calculus of Convex Functions

The following operators preserve the convexity of functions, which can be easily verified based on the definition.

1. **Taking conic combination**: If $f_\alpha(x), \alpha \in \mathcal{A}$ are convex functions and $\{\lambda_\alpha\}_{\alpha\in\mathcal{A}} \geq 0$, then

$$\sum_{\alpha\in\mathcal{A}} \lambda_\alpha f_\alpha(x)$$

   is also a convex function.

2. **Taking affine composition** If $f(x)$ is convex on $\mathbf{R}^n$, and $\mathcal{A}(y) : y \mapsto Ay + b$ is an affine mapping from $\mathbf{R}^k$ to $\mathbf{R}^n$, then

$$g(y) := f(Ay + b)$$

   is convex on $\mathbf{R}^k$.

   The proofs are straightforward and hence omitted.

3. **Taking superposition**:

   - If f is a convex function on $\mathbf{R}^n$ and $F(\cdot)$ is a convex and non-decreasing function on $\mathbf{R}$, then $g(x) = F(f(x))$ is convex.
   - More generally, if $f_i(x), i = 1, \ldots, m$ are convex on $\mathbf{R}^n$ and $F(y_1, \ldots, y_m)$ is convex and non-decreasing (component-wise) on $\mathbf{R}^m$, then

$$g(x) = F(f_1(x), \ldots, f_m(x))$$

     is convex.

   Proof: By convexity of f_i, we have

$$f_i(\lambda x + (1-\lambda)y) \leq \lambda f_i(x) + (1-\lambda) f_i(y), \forall i, \forall \lambda \in [0,1].$$

   Hence, we have for any $\lambda \in [0,1]$,

$$\begin{aligned} g(\lambda x + (1-\lambda)y) &= F(f_1(\lambda x + (1-\lambda)y), \ldots, f_m(\lambda x + (1-\lambda)y)) \\ &\leq F(\lambda f_1(x) + (1-\lambda) f_1(y), \ldots, \lambda f_m(x) + (1-\lambda) f_m(y)) && (\text{ by monotonicity of } F) \\ &\leq \lambda F(f_1(x), \ldots, f_m(x)) + (1-\lambda) F(f_1(x), \ldots, f_m(x)) && (\text{ by convexity of } F) \\ &= \lambda g(x) + (1-\lambda) g(y) && (\text{ by definition of } g) \end{aligned}$$

   ■

4. **Taking supremum**: If $f_\alpha(x), \alpha \in \mathcal{A}$ are convex, then

$$\sup_{\alpha\in\mathcal{A}} f_\alpha(x)$$

   is convex.

   Note that when $\mathcal{A}$ is finite, this can be considered as a special superposition with $F(y_1, \ldots, y_m) = \max(y_1, \ldots, y_m)$, which can be easily shown to be monotonic and convex.

Proof: We show that

$$\begin{aligned} epi(\sup_{\alpha\in\mathcal{A}} f_\alpha) &= \{(x,t) : \sup_{\alpha\in\mathcal{A}} f_\alpha(x) \le t\} \\ &= \{(x,t) : f_\alpha(x) \le t, \forall\alpha \in \mathcal{A}\} \\ &= \cap_{\alpha\in\mathcal{A}}\{(x,t) : f_\alpha(x) \le t\} \\ &= \cap_{\alpha\in\mathcal{A}} epi(f_\alpha). \end{aligned}$$

Since f_α is convex, $epi(f_\alpha)$ is therefore a convex set for any $\alpha \in \mathcal{A}$. Their intersection remains convex, i.e. $epi(\sup_{\alpha\in\mathcal{A}} f_\alpha)$ is a convex set, i.e. $\sup_{\alpha\in\mathcal{A}} f_\alpha(x)$ is convex. ■

5. **Partial minimization**: If $f(x,y)$ is convex in $(x,y) \in \mathbf{R}^n$ and C is a convex set, then

$$g(x) = \inf_{y\in C} f(x,y)$$

is convex.

Proof: Given any x_1, x_2, by definition, for any $\epsilon > 0$,

$$\begin{aligned} \exists y_1 : f(x_1,y_1) \le g(x_1) + \epsilon/2 \\ \exists y_2 : f(x_2,y_2) \le g(x_2) + \epsilon/2 \end{aligned}$$

For any $\lambda \in [0,1]$, adding the two equations, we have

$$\lambda f(x_1,y_1) + (1-\lambda) f(x_2,y_2) \le \lambda g(x_1) + (1-\lambda) g(x_2) + \epsilon.$$

Invoking the convexity of $f(x,y)$, this implies

$$f(\lambda x_1 + (1-\lambda)x_2, \lambda y_1 + (1-\lambda)y_2) \le \lambda g(x_1) + (1-\lambda) g(x_2) + \epsilon.$$

Hence for any $\epsilon > 0$, $g(\lambda x_1 + (1-\lambda)x_2) \le \lambda g(x_1) + (1-\lambda) g(x_2) + \epsilon$. Letting $\epsilon \to 0$ leads to the convexity of g. ■

6. **Perspective function**: If $f(x)$ is convex , then the *perspective* of f

$$g(x,t) := tf(x/t)$$

is convex on its domain $dom(g) = \{(x,t) : x/t \in dom(f), t > 0\}$.

Proof: Observe that

$$(x,t,\tau) \in epi(g) \Longleftrightarrow tf(x/t) < \tau \Longleftrightarrow f(x/t) \le \tau/t \Longleftrightarrow (x/t, \tau/t) \in epi(f)$$

Define the perspective function $P : \mathbf{R}^n \times \mathbf{R}_{++} \times \mathbf{R} :\to \mathbf{R}^n \times \mathbf{R}, (x,t,\tau) \mapsto (x/t, \tau/t)$, then

$$epi(g) = P^{-1}(epi(f)).$$

Since f is convex, $epi(f)$ is a convex set. To show g is convex, it suffices to show that the inverse image of a convex set under the perspective function is convex.

Claim: If U is a convex set, then

$$P^{-1}(U) = \{(u,t) : u/t \in U, t > 0\}$$

is a convex set.
This is because if $(u,t) \in P^{-1}(U)$ and $(v,\mathrm{s}) \in P^{-1}(U)$, for any $\lambda \in [0,1]$,

$$\frac{\lambda u + (1-\lambda)v}{\lambda t + (1-\lambda)s} = \mu \cdot \frac{u}{t} + (1-\mu) \cdot \frac{v}{s} \in U$$

where $\mu = \frac{\lambda t}{\lambda t + (1-\lambda)s} \in [0,1]$. Hence, $\lambda \cdot (u,t) + (1-\lambda) \cdot (v,s) \in P^{-1}(U)$. ■

3.4 Examples of Convex Functions

Example 1. *Simple univariate functions:*

- $x^2, x^4, \ldots$
- e^{ax} for any a
- $-\log(x)$
- $x\log(x)$

Example 2. *Multi-variate functions*:

- $\|\cdot\|$
- $\frac{1}{2}x^TQx + b^Tx + c$, when $Q \succeq 0$
- $\|Ax - b\|_2^2$
- $\max(a_1^Tx + b_1, \ldots, a_k^Tx - b_k)$
- relative entropy function $g(x,t) : \mathbf{R}^2_{++} \to \mathbf{R}, (x,t) \mapsto t\log(t) - t\log(x)$
- $\log(\sum_{i=1}^k e^{a_i^Tx+b_i})$

  Proof: It suffices to show that $f(x) = \log(\sum_{i=1}^n e^{x_i})$ is convex. Observe that any h, we have

  $$h^T\nabla^2 f(x)h = \frac{\sum_i e^{x_i}h_i^2}{\sum_i e^{x_i}} - \frac{(\sum_i e^{x_i}h_i)^2}{(\sum_i e^{x_i})^2}.$$

  Let $p_i = \frac{e^{x_i}}{\sum_i e^{x_i}}$, we have

  $$h^T\nabla^2 f(x)h = \sum_i p_ih_i^2 - (\sum_i p_ih_i)^2 \geq \sum_i p_ih_i^2 - \sum_i(\sqrt{p_i})^2\sum_i(\sqrt{p_i}h_i)^2 = \sum_i p_ih_i^2 - 1\cdot\sum_i p_ih_i^2 = 0.$$

  The first inequality is due to Cauchy-Schwarz inequality. Hence, $\nabla^2 f(x) \succeq 0$. ■

- $-\log(det(X))$

  Proof: Let $f(X) = -\log(det(X))$, the domain $dom(f) = S^n_{++}$. Let $X, H \succ 0$, it is sufficient to show that $g(t) = f(X + tH)$ is convex on $dom(g) = \{t : X + tH \succ 0\}$. Since

  $$g(t) = -\log(det(X+tH)) = -\log(det(X^{1/2}(I+tX^{-1/2}HX^{-1/2})X^{1/2})) = -\sum_i \log(1+t\lambda_i) - \log(det(X))$$

  where $\lambda_1, \ldots, \lambda_n$ are the eigenvalues of $X^{-1/2}HX^{-1/2}$. Note that for each i, $-\log(1 + t\lambda_i)$ is convex in t, so $g(t)$ is also convex. ■

Example 3. Some distances:

- maximum distance to any set C: $d(x, C) := \max_{y\in C}\|x - y\|$
- minimum distance to a convex set C: $d(x, C) : \min_{y\in C}\|x - y\|$

Example 4. *Indicator and support functions*:

- indicator function of a convex set C: $I_C(x) := \begin{cases} 0, & x \in C \\ \infty, & x \notin C \end{cases}$
- support function of any set C (convex or not): $I_C^*(x) = \sup_{y \in C} x^T y$

3.5 Subgradients of Convex Functions

Definition 3.8 (Subgradient) *Let f be a convex function and $x \in dom(f)$, any vector g satisfying*

$$f(y) \geq f(x) + g^T(y - x)$$

is called a subgradient of f at x.

The set of all subgradients of f at x is called the subdifferential, denoted as $\partial f(x)$.

Example 1. If f is differentiable at $x \in dom(f)$, then $\nabla f(x)$ is the unique element of $\partial f(x)$.
Proof: Let $g \in \partial f(x)$, by definition, $\frac{f(x+td)-f(x)}{t} \geq g^T d, , \forall d$. Let $t \to 0$, we have $\nabla f(x)^T d \geq g^t d, \forall d$, which implies $\nabla f(x) = g$. ■

Example 2. Let $f(x) = |x|$, then $\partial f(0) = [-1, 1]$.
Proof: This is because $|x| \geq 0 + gx, \forall g \in [-1, 1]$. ■

Example 3. Let $f(x) = \|x\|_2$, then $\partial f(x) = \begin{cases} \frac{x}{\|x\|_2}, & x \neq 0 \\ \{g : \|g\|_2 \leq 1\}, & x = 0 \end{cases}$.
Proof: This is because $\|x\|_2 \geq 0 + g^T x, \forall \|g\|_2 \leq 1$. ■

Proposition 3.9 *If $\bar{x} \in int(dom(f))$, then $\partial f(\bar{x})$ is nonempty, closed, bounded, and convex.*

Proof:

- (Convexity and closedness): this is due to the fact that

  $$\partial f(\bar{x}) = \cap_x \{g : f(x) \geq f(\bar{x}) + g^T(x - \bar{x})\}$$

  is a infinite system of linear inequalities. The sub-differentiable set can be treated as the intersection of halfspaces, hence is closed and convex.

- (Non-emptyness): applying the separation theorem on $(\bar{x}, f(\bar{x}))$ and $epi(f) = \{(x, t) : f(x) \leq t\}$, we have

  $$\exists a, \beta, \text{s. t. } a^T \bar{x} + \beta f(\bar{x}) \leq a^T x + \beta t, \forall (x, t) \in epi(f).$$

  Claim: $\beta > 0$. We can first rule out $\beta \neq 0$ since $\bar{x} \in int(dom(f))$. We then rule out $\beta < 0$ by setting $x = \bar{x}$ and $t > f(\bar{x})$.
  Therefore, defining $g = \beta^{-1} a$, we have $f(x) \geq f(\bar{x}) + g^T(x - \bar{x})$, i.e. $g \in \partial(f)$.

- (Boundedness): if $\partial f(\bar{x})$ is unbounded, then there exist $s_k \in \partial f(\bar{x})$, such that $\|s_k\|_\infty \to \infty$ as $n \to \infty$. SInce $\bar{x} \in int(dom(f))$, there exits $\epsilon > 0$, such that $B(\bar{x}, \epsilon) = \{x : \|x - \bar{x}\| \leq \epsilon\} \subset dom(f)$. Hence, letting $y_k = \bar{x} + \epsilon \frac{s_k}{\|s_k\|}$, we have $y_k \in B(\bar{x}, \epsilon)$, and

  $$f(y_k) \geq f(\bar{x}) + s_k^T(y_k - \bar{x}) = f(\bar{x}) + \epsilon\|s_k\| \to \infty, \text{ as } k \to \infty.$$

  However, every convex function can be shown to be continuous on its interior; it is Lipschitz continuous on any convex compact subset on the domain. This implies that $f(x)$ is bounded on the compact ball $B(\bar{x}, \epsilon)$, which leads to a contradiction.

■

Remark. The reverse is also true. If $\forall x \in int(dom(f))$, $\partial f(x)$ is nonempty, then f is convex.

Proof: Let $x, y \in dom(f)$, $z = \lambda x + (1 - \lambda)y \in int(dom(f))$, we have

$$f(x) \geq f(z) + g^T(x - z)$$
$$f(y) \geq f(z) + g^T(y - z)$$

Hence, $\lambda f(x) + (1 - \lambda)f(y) \geq f(z) = f(\lambda x + (1 - \lambda)y)$. ■

3.6 Calculus of Sub-differential

Determining the subdifferentiable set of a convex function at a given point is in general very difficult. That's why calculus of subdifferentiable sets is particularly important in convex analyis.

1. **Taking conic combination**: If $h(x) = \lambda f(x) + \mu g(x)$, where $\lambda, \mu \geq 0$ and f, g are both convex, then

   $$\partial h(x) = \lambda \partial f(x) + \mu \partial g(x), \forall x \in int(dom(h)).$$

2. **Taking affine composition**: If $h(x) = f(Ax + b)$, where f is convex, then

   $$\partial h(x) = A^T \partial f(Ax + b).$$

3. **Taking supremum**: If $h(x) = \sup_{\alpha \in \mathcal{A}} f_\alpha(x)$ and each $f_\alpha(x)$ is convex, then

   $$\partial h(x) \supseteq conv\{\partial f_\alpha(x) | \alpha \in \alpha(x)\}$$

   where $\alpha(x) := \{\alpha : h(x) = f_\alpha(x)\}$.

4. **Taking superposition**: If $h(x) = F(f_1(x), \ldots, f_m(x))$, where $F(y_1, \ldots, y_m)$ is non-decreasing and convex, then

   $$\partial h(x) \supseteq \left\{ \sum_{i=1}^m d_i \partial f_i(x) : (d_1, \ldots, d_m) \in \partial F(y_1, \ldots, y_m) \right\}.$$

Example 1. Let $h(x) = \max_{1 \leq i \leq n}(a_i^T x + b_i)$, then $a_k \in \partial h(x)$ if k is some index such that $h(x) = a_k^T x + b_k$.

Example 2. Let $h(x) = \mathbb{E}[F(x, \xi)]$ be a convex function, then $g(x) = \int G(x, \xi) p(\xi) d\xi \in \partial h(x)$ if $G(x, \xi) \in \partial F(x, \xi)$ for each ξ.

Example 3. Let $h(x) = \max_{y \in C} f(x, y)$ where $f(x, y)$ is convex in x for any y and C is closed, then $\partial f(x, y_*(x)) \subset \partial h(x)$, where $y_*(x) = argmax_{y \in C} f(x, y)$.

This is because if $g \in \partial f(x, y_*(x))$, we have

$$h(z) \geq f(z, y_*(x)) \geq f(x, y_*(x)) + g^T(z - x) = h(z) + g^T(z - x).$$

3.7 Other Properties of Convex Functions

Jensen's inequality. Let f be a convex function, then

$$f(\sum_i \lambda_i x_i) \leq \sum_i \lambda_i f(x_i)$$

as long as $\lambda_i \geq 0, \forall i$ and $\sum_i \lambda_i = 1$.

Moreover, let f be a convex function and X be a random variable, then

$$f(\mathbb{E}[X]) \leq \mathbb{E}[f(X)].$$

Example . The Kullback-Liebler distance between two distributions is nonnegative: i.e.

$$KL(p||q) = \sum_i p_i \log\left(\frac{p_i}{q_i}\right) \geq 0$$

where $p_i \geq 0, q_i \geq 0, \sum_i p_i = \sum_i q_i = 1$.

Proof: Let $f(x) = -\log(x)$, f is convex, so

$$-\log(\sum_i p_i x_i) \leq -\sum_i p_i \log(x_i).$$

Plugging $x_i = q_i/p_i$, this leads to

$$0 = -\log(\sum_i q_i) \leq \sum_i p_i \log(p_i/q_i).$$

■

References

[BV04] Boyd, S., & Vandenberghe, L. (2004). *Convex optimization*. Cambridge university press.

[BN01] Ben-Tal, A., & Nemirovski, A. (2001). *Lectures on modern convex optimization: analysis, algorithms, and engineering applications*, (Vol. 2). SIAM.

Reading materials: Jan, 2022

1 Preliminaries

Definition We define $\mathbb{R}$ to be the set of real numbers, and we define $\mathbb{R}^n$ to be the set of ordered n-tuples, also called vectors. In other words it is the set of all elements $x = (x_1,\ x_2 \ldots,\ x_n)$ where $x_i \in \mathbb{R}$ for each $i = 1,\ 2,\ \ldots,\ n$ The zero vector is usually just denoted by 0.

For $x = (x_1,\ x_2,\ \ldots,\ x_n) \in \mathbb{R}^n$ and $y = (y_1,\ y_2,\ \ldots, y_n) \in \mathbb{R}^n$ we define the operation of *vector addition* by $x + y = (x_1 + y_1,\ x_2 + y_2,\ \ldots,\ x_n + y_n)$.

For $\lambda \in \mathbb{R}$, we define *scalar multiplication* by $\lambda x = (\lambda x_1,\ \lambda x_2,\ \ldots,\ \lambda x_n)$.

Given $x = (x_1,\ x_2,\ \ldots,\ x_n) \in \mathbb{R}^n$ and $y = (y_1,\ y_2,\ \ldots, y_n) \in \mathbb{R}^n$ we define the *inner product* by $\langle x,\ y\rangle = x_1y_1 + x_2y_2 + \ldots + x_ny_n$.

Proposition 1.1 Given $x,\ y,\ z \in \mathbb{R}^n$, the following hold:

(**i**) $\langle x,\ x\rangle \geq 0$, furthermore $\langle x,\ x\rangle = 0$ if and only if $x = 0$.
(**ii**) $\langle x,\ y\rangle = \langle y,\ x\rangle$.
(**iii**) $\langle \lambda x,\ y\rangle = \lambda\langle x,\ y\rangle$ for all $\lambda \in \mathbb{R}$.
(**iv**) $\langle x,\ y + z\rangle = \langle x,\ y\rangle + \langle x,\ z\rangle$.

Proof The proof is left as an exercise.

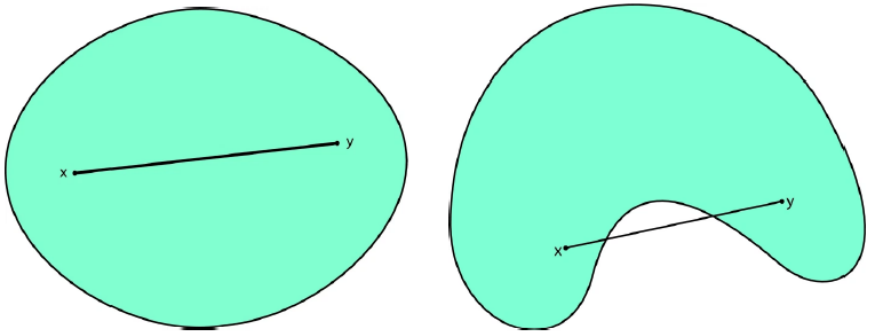
Definition For any $x \in \mathbb{R}^n$ we define the *norm* (https://en.wikipedia.org/wiki/Norm_(mathematics)) to be $\|x\| = \sqrt{\langle x, x \rangle} = \sqrt{x_1^2 + x_2^2 + ... + x_n^2}$.

Proposition 1.2 For any $x, y \in \mathbb{R}^n$ the following hold:

(**i**) $\|x\| \geq 0$ and $\|x\| = 0$ if and only if $x = 0$.
(**ii**) $\|\lambda x\| = |\lambda| \, \|x\|$ for all $\lambda \in \mathbb{R}$.
(**iii**) $\|x + y\| \leq \|x\| + \|y\|$ (Triangle Inequality).
(**iv**) $\langle x, y \rangle \leq \|x\| \, \|y\|$ (Cauchy-Schwartz Inequality).

Proof This proof is also left as a review exercise.

2 Convex Sets



Definition A subset Ω of $\mathbb{R}^n$ is called *convex (https://en.wikipedia.org/wiki/Convex_set)* if $\lambda x + (1-\lambda)y \in \Omega$ for all $x, y \in \Omega$ and $\lambda \in (0,1)$.

Given $a, b \in \mathbb{R}^n$, the line segment connecting a and b is the set

$$[a,b] = \{\lambda a + (1-\lambda)b \mid \lambda \in [0,1]\}.$$

It follows from the definition that Ω is convex if and only if $[a,b] \subset \Omega$ whenever $a, b \in \Omega$.

Proposition 2.1 If Ω_1 is a convex subset of $\mathbb{R}^n$ and Ω_2 is a convex subset of $\mathbb{R}^m$, then the Cartesian product $\Omega_1 \times \Omega_2$ is a convex subset of $\mathbb{R}^n \times \mathbb{R}^m$.

Proof Fix any $(a_1, a_2), (b_1, b_2) \in \Omega_1 \times \Omega_2$ and $\lambda \in (0, 1)$. Then $a_1, b_1 \in \Omega_1$ and $a_2, b_2 \in \Omega_2$. By the convexity of Ω_1 and the convexity of Ω_2, we have that $\lambda(a_1, a_2) + (1 - \lambda)(b_1, b_2) = (\lambda a_1 + (1 - \lambda)b_1, \lambda a_2 + (1 - \lambda)b_2) \in \Omega_1 \times \Omega_2$. Therefore, $\Omega_1 \times \Omega_2$ is a convex subset of $\mathbb{R}^n \times \mathbb{R}^m$. $\square$

Definition A mapping $B\colon \mathbb{R}^n \to \mathbb{R}^m$ is called an *affine mapping* (https://en.wikipedia.org/wiki/Affine_transformation) if there exist a linear mapping $A\colon \mathbb{R}^n \to \mathbb{R}^m$ and an element $b \in \mathbb{R}^m$ such that $B(x) = A(x) + b$ for all $x \in \mathbb{R}^n$.

The following proposition gives a characterization for affine mappings.

Proposition 2.2 A mapping $B\colon \mathbb{R}^n \to \mathbb{R}^m$ is an affine mapping if and only if

$$B(\lambda x_1 + (1-\lambda)x_2) = \lambda B(x_1) + (1-\lambda)B(x_2) \text{ for all } x_1, x_2 \in \mathbb{R}^n.$$

Proof Suppose that B is an affine mapping. Then there exist a linear mapping $A\colon \mathbb{R}^n \to \mathbb{R}^m$ and an element $b \in \mathbb{R}^m$ such that the equation above is satisfied. Given any $x_1, x_2 \in \mathbb{R}^n$ and $\lambda \in \mathbb{R}$, one has

$$\begin{aligned} B(\lambda x_1 + (1-\lambda)x_2) &= A(\lambda x_1 + (1-\lambda)x_2) + b \\ &= \lambda A(x_1) + (1-\lambda)A(x_2) + \lambda x_1 + (1-\lambda)x_2 \\ &= \lambda[A(x_1) + b] + (1-\lambda)[A(x_2) + b] \\ &= \lambda B(x_1) + (1-\lambda)B(x_2). \end{aligned}$$

Thus, the equation is satisfied. $\square$

In the proposition below, we show that the convexity of sets is preserved under affine mappings.

Proposition 2.3 Let $B\colon \mathbb{R}^n \to \mathbb{R}^m$ be an affine mapping. Then the following are true.
$(\mathbf{i})$ If Ω is a convex subset of $\mathbb{R}^n$, then $B(\Omega)$ is a convex subset of $\mathbb{R}^m$.
$(\mathbf{ii})$ If Θ is a convex subset of $\mathbb{R}^m$, then $B^{-1}(\Theta)$ is a convex subset of $\mathbb{R}^n$.
Proof We only prove $(\mathbf{i})$ and leave the proof of $(\mathbf{ii})$ as an exercise. Fix any $a, b \in B(\Omega)$ and $\lambda \in (0,1)$. Then $a = B(x)$ and $b = B(y)$ for $x, y \in \Omega$. By Proposition 2.2,
$\lambda a + (1-\lambda)b = \lambda B(x) + (1-\lambda)B(y) = B(\lambda x + (1-\lambda)y)$.
Since $\emptyset$ is convex, $\lambda x + (1-\lambda)y \in \Omega$, and hence $\lambda a + (1-\lambda)b \in B(\Omega)$, which shows that $B(\Omega)$ is convex. $\square$

For two subsets Ω_1 and Ω_2 of $\mathbb{R}^n$ and $\lambda \in \mathbb{R}$, define

$$\lambda\Omega_1 = \{\lambda x \mid x \in \Omega_2\}$$

$$\Omega_1 + \Omega_2 = \{x + y \mid x \in \Omega_1, y \in \Omega_2\}.$$

Corollary 2.4 Let Ω_1 and Ω_2 be convex subsets of $\mathbb{R}^n$ and let $\lambda \in \mathbb{R}$. Then $\lambda\Omega_1$ and $\Omega_1 + \Omega_2$ are convex subsets of $\mathbb{R}^n$.
Proof Define the mapping $B\colon \mathbb{R}^n \to \mathbb{R}^n$ by $B(x) = \lambda x$ for $x \in \mathbb{R}^n$. Then B is an affine mapping and $B(\Omega_1) = \lambda\Omega_1$. By Proposition 2.3, the set $B(\Omega_1)$ is convex.

Similarly, define $C\colon \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^n$ by $C(x, y) = x + y$. Then C is an affine mapping and $C(\Omega_1 \times \Omega_2) = \Omega_1 + \Omega_2$, justifying that $\Omega_1 + \Omega_2$ is a convex subset of $\mathbb{R}^n$. □

Next we proceed with intersections of convex sets.

Proposition 2.5 Let $\{\Omega_\alpha\}_{\alpha\in I}$ be a collection of convex subsets of $\mathbb{R}^n$. Then $\bigcap_{\alpha\in I} \Omega_\alpha$ is also a convex subset of $\mathbb{R}^n$.
Proof Taking any $a, b \in \bigcap_{\alpha\in I} \Omega_\alpha$, we get that $a, b \in \Omega_\alpha$ for all $\alpha \in I$. The convexity of each Ω_α ensures that $\lambda a + (1 - \lambda)b \in \Omega_\alpha$ for any $\lambda \in (0, 1)$. Thus $\lambda a + (1 - \lambda)b \in \bigcap_{\alpha\in I} \Omega_\alpha$ and the intersection $\bigcap_{\alpha\in I} \Omega_\alpha$ is convex. □

Definition (Convex Combination) A vector $x \in \mathbb{R}^n$ is called a *convex combination (https://en.wikipedia.org/wiki/Convex_combination)* of $x_1, \ldots, x_m \in \mathbb{R}^n$ if there exist $\lambda_1, \ldots, \lambda_m \geq 0$ such that

$$\sum_{i=1}^m \lambda_i = 1 \text{ and } x = \sum_{i=1}^m \lambda_i x_i.$$

Proposition 2.6 A subset Ω of $\mathbb{R}^n$ is convex if and only if it contains all convex combinations of its elements.

Proof If Ω contains all of its convex combinations then obviously it is convex. To prove the other implication, we show by induction that any convex combination $x = \sum_{i=1}^m \lambda_i \omega_i$ of elements in Ω is an element of Ω. This conclusion follows directly from the definition for $m = 1, 2$. Fix now a positive integer $m \geq 2$ and suppose that every convex combination of m elements from Ω belongs to Ω. Form the convex combination

$$y = \sum_{i=1}^{m+1} \lambda_i \omega_i, \ \sum_{i=1}^{m+1} \lambda_i = 1, \ \lambda_i \geq 0$$

and observe that if $\lambda_{m+1} = 1$, then $\lambda_1 = \lambda_2 = \ldots = \lambda_m = 0$, so $y = \omega_{m+1} \in \Omega$. In the case where $\lambda_{m+1} < 1$ we get the representations

$$\sum_{i=1}^{m} \lambda_i = 1 - \lambda_{m+1} \quad \text{and} \quad \sum_{i=1}^{m} \frac{\lambda_i}{1-\lambda_{m+1}} = 1,$$

which in turn implies the inclusion

$$z = \sum_{i=1}^{m} \frac{\lambda_i}{1-\lambda_{m+1}} \omega_i \in \Omega.$$

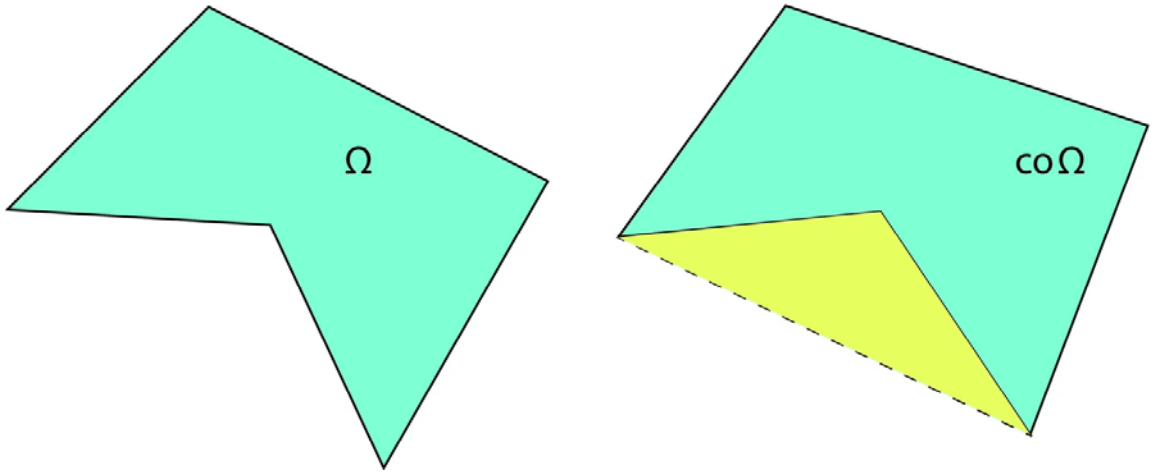
It yields therefore the relationships

$$y = (1-\lambda_{m+1}) \sum_{i=1}^{m} \frac{\lambda_i}{1-\lambda_{m+1}} \omega_i + \lambda_{m+1}\omega_{m+1} = (1-\lambda_{m+1})z + \lambda_{m+1}\omega_{m+1} \in \Omega$$

and thus completes the proof of the proposition. □

Definition (Convex Hull) Let Ω be a subset of $\mathbb{R}^n$. The *convex hull (https://en.wikipedia.org/wiki/Convex_hull)* of Ω is defined by

$$\mathrm{co}(\Omega) = \bigcap \Big\{ C \;\Big|\; C \text{ is convex and } \Omega \subset C \Big\}.$$



Convex Hull

Proposition 2.7 The convex hull $\mathrm{co}(\Omega)$ is the smallest convex set containing Ω.

Proof The convexity of the set $\operatorname{co}\Omega \supset \Omega$ follows from Proposition 2.5. On the other hand, for any convex set $C \subset \Omega$ we clearly have $\operatorname{co}(\Omega) \subset C$, which verifies the conclusion. $\square$

Proposition 2.8 For any subset Ω of $\mathbb{R}^n$, its convex hull admits the representation

$$\operatorname{co}(\Omega) = \Big\{ \sum_{i=1}^m \lambda_i a_i \;\Big|\; \sum_{i=1}^m \lambda_i = 1,\ \lambda_i \geq 0,\ a_i \in \Omega,\ m \in \mathbb{N} \Big\}.$$

Proof Denoting by C the right-hand side of the representation to prove, we obviously have $\Omega \subset C$. Let us check that the set C is convex. Take any $a, b \in C$ and get

$$a := \sum_{i=1}^p \alpha_i a_i, \quad b := \sum_{j=1}^q \beta_j b_j,$$

where $a_i, b_j \in \Omega$, $\alpha_i, \beta_j \geq 0$ with $\sum_{i=1}^p \alpha_i = \sum_{j=1}^q \beta_j = 1$, and $p, q \in \mathbb{N}$. It follows easily that for every number $\zeta \in (0,1)$, we have

$$\zeta a + (1-\zeta) b = \sum_{i=1}^p \zeta \alpha_i a_i + \sum_{j=1}^q (1-\zeta)\beta_j b_j.$$

Then the resulting equality

$$\sum_{i=1}^{p} \zeta\alpha_i + \sum_{j=1}^{q}(1-\zeta)\beta_j = \zeta\sum_{i=1}^{p}\alpha_i + (1-\zeta)\sum_{j=1}^{q}\beta_j = 1$$

ensures that $\zeta a + (1-\zeta)b \in C$, and thus $\operatorname{co}(\Omega) \subset C$ by the definition of $\operatorname{co}\Omega$. Fix now any $= \sum_{i=1}^{m} \lambda_i a_i \in C$ with $a_i \in \Omega \subset \operatorname{co}(\Omega)$ for $i = 1, \ldots, m$. Since the set $\operatorname{co}(\Omega)$ is convex, we conclude by Proposition 2.6 that $a \in \operatorname{co}(\Omega)$ and thus $\operatorname{co}(\Omega) = C$. $\square$

Definitions (Interior and Closure of a Set) Let $\Omega \subset \mathbb{R}^n$. We say that $\omega \in \text{int}(\Omega)$ (ω is in the interior of Ω) if there exists an $r > 0$ such that the open ball $B(\omega; 0) = \{x \in \mathbb{R}^n : \|x - \omega\| < r\}$ is contained in Ω.

The closure of Ω, denoted $\overline{\Omega}$, is the smallest closed set containing Ω. Alternatively, it is the intersection of all closed sets containing Ω.

Then $b \in \overline{\Omega}$ if and only if for all $r > 0$, $B(b; r) \cap \Omega \neq \emptyset$.

Proposition 2.9 The interior $\text{int}(\Omega)$ and closure $\overline{\Omega}$ of a convex set $\Omega \subset \mathbb{R}^n$ are also convex.

Proof Fix $a, b \in \text{int}\,\Omega$ and $\lambda \in (0, 1)$. Then find an open set V such that

$$a \in V \subset \Omega \text{ and so } \lambda a + (1 - \lambda)b \in \lambda V + (1 - \lambda)b \subset \Omega.$$

Since $\lambda V + (1 - \lambda)b$ is open, we get $\lambda a + (1 - \lambda)b \in \text{int}(\Omega)$, and thus the set $\text{int}(\Omega)$ is convex.

To verify the convexity of $\overline{\Omega}$, we fix $a, b \in \overline{\Omega}$ and $\lambda \in (0, 1)$ and then find sequences $\{a_k\}$ and $\{b_k\}$ in Ω converging to a and b, respectively. By the convexity of Ω, the sequence $\{\lambda a_k + (1 - \lambda)b_k\}$ lies entirely in Ω and converges

to $\lambda a + (1-\lambda)b$. This ensures the inclusion $\lambda a + (1-\lambda)b \in \overline{\Omega}$ and thus justifies the convexity of the closure $\overline{\Omega}$. $\square$

To proceed further, for any $a, b \in \mathbb{R}^n$, define the interval

$$[a, b) = \big\{\lambda a + (1-\lambda)b \;\big|\; \lambda \in (0, 1]\big\}.$$

We can also define the intervals $(a, b]$ and (a, b) in a similar way.

Lemma 2.10 For a convex set $\Omega \subset \mathbb{R}^n$ with nonempty interior, take any $a \in \operatorname{int}(\Omega)$ and $b \in \overline{\Omega}$. Then $[a, b) \subset \operatorname{int}(\Omega)$.

Proof Since $b \in \overline{\Omega}$, for any $\epsilon > 0$, we have $b \in \Omega + \epsilon B$. Take now $\lambda \in (0, 1]$ and let $x_\lambda := \lambda a + (1-\lambda)b$. Choosing $\epsilon > 0$ such that $a + \epsilon\dfrac{2-\lambda}{\lambda}B \subset \Omega$ gives us

$$\begin{aligned}
x_\lambda + \epsilon B &= \lambda a + (1-\lambda)b + \epsilon B \\
&\subset \lambda a + (1-\lambda)\big[\Omega + \epsilon B\big] + \epsilon B \\
&= \lambda a + (1-\lambda)\Omega + (1-\lambda)\epsilon B + \epsilon B \\
&\subset \lambda\Big[a + \epsilon\frac{2-\lambda}{\lambda}B\Big] + (1-\lambda)\Omega
\end{aligned}$$

$$\subset \lambda\Omega + (1-\lambda)\Omega \subset \Omega.$$

This shows that $x_\lambda \in \operatorname{int}(\Omega)$ and thus verifies the inclusion $[a,b) \subset \operatorname{int}(\Omega)$. $\square$

Now we establish relationships between the interior and closure of convex sets.

Proposition 2.11 Let $\Omega \subset \mathbb{R}^n$ be a convex set with nonempty interior. Then we have the following two properties:

$(\mathbf{i})$ $\overline{\operatorname{int}(\Omega)} = \overline{\Omega}$.
$(\mathbf{ii})$ $\operatorname{int}(\Omega) = \operatorname{int}(\overline{\Omega})$.
Proof $(\mathbf{i})$ Obviously, $\overline{\operatorname{int}(\Omega)} \subset \overline{\Omega}$. For any $b \in \bar{\Omega}$ and $a \in \operatorname{int}\Omega$, define the sequence $\{x_k\}$ by

$$x_k := \frac{1}{k}a + \Big(1 - \frac{1}{k}\Big)b, \ k \in \mathbb{N}.$$

Lemma 3.2 ensures that $x_k \in \operatorname{int}(\Omega)$. Since $x_k \to b$, we have $b \in \overline{\operatorname{int}(\Omega)}$ and thus we have that $(\mathbf{i})$ holds.

(**ii**) Since the inclusion $\mathrm{int}(\Omega) \subset \mathrm{int}(\overline{\Omega})$ is obvious, it remains to prove the opposite inclusion $\mathrm{int}(\overline{\Omega}) \subset \mathrm{int}(\Omega)$.

To proceed, fix any $b \in \mathrm{int}(\overline{\Omega})$ and $a \in \mathrm{int}(\Omega)$. If $\epsilon > 0$ is sufficiently small, then

$$c := b + \epsilon(b - a) \in \overline{\Omega} \text{ and } b = \frac{\epsilon}{1+\epsilon} a + \frac{1}{1+\epsilon} c \in (a, c) \subset \mathrm{int}(\Omega),$$

which verifies that $\mathrm{int}(\overline{\Omega}) \subset \mathrm{int}(\Omega)$ and thus completes the proof. $\square$

3 Convex Functions

Definitions We define the *extended real line* to be the set $\overline{\mathbb{R}} := (-\infty, \infty]$ where infinity is allowed as a value with the following conventions:

$\alpha + \infty = \infty$ where $\alpha \in \mathbb{R}$

$\alpha \cdot \infty = \infty$ where $\alpha > 0$

$0 \cdot \infty = \infty$

Let $\Omega \subset \mathbb{R}^n$ be a convex set and let $f : \Omega \to \overline{\mathbb{R}} := (-\infty, \infty]$ be an extended-real-valued function. Then the function f is said to be *convex (https://en.wikipedia.org/wiki/Convex_function)* on Ω if

$f\big(\lambda x + (1-\lambda)y\big) \leq \lambda f(x) + (1-\lambda) f(y)$ for all $x, y \in \Omega$ and $\lambda \in (0, 1)$.

If the inequality is strict for $x \neq y$, then f is said to be *strictly convex* on Ω.

The *domain* of a function is the set given by

$\operatorname{dom}(f) = \{x \in \mathbb{R}^n : f(x) < \infty\}$.

The *epigraph (https://en.wikipedia.org/wiki/Epigraph_(mathematics))* of a function is the set given by

$$\mathrm{epi}(f) = \{(x, \lambda) : x \in \mathbb{R}^n, \lambda \in \mathbb{R}, \lambda >= f(x)\}.$$

Proposition 3.1 Let $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ be a function.

(**i**) If f is a convex function, then the domain of f is a convex subset of $\mathbb{R}^n$.
(**ii**) f is a convex function if and only if

$$f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda) f(y) \text{ for all } x, y \in \mathrm{dom}(f), \lambda \in (0,1).$$

Proof (**i**) Let $x, y \in \mathrm{dom}(f)$ and let $0 \leq \lambda \leq 1$. Then we have that $f(x) < \infty$ and $f(y) < \infty$. Thus by the convexity of f we have that
$f\big(\lambda x + (1-\lambda)y\big) \leq \lambda f(x) + (1-\lambda) f(y)$ which is less than ∞.

(**ii**) is left as an exercise. □

The next theorem gives a geometric characterization of function convexity via the convexity of the associated epigraphical set.

Theorem 3.2 A function $f\colon \mathbb{R}^n \to \overline{\mathbb{R}}$ is convex if and only if its epigraph $\operatorname{epi}(f)$ is a convex subset of the product space $\mathbb{R}^n \times \mathbb{R}$.

Proof Assuming that f is convex, fix pairs $(x_1, t_1), (x_2, t_2) \in \operatorname{epi}(f)$ and a number $\lambda \in (0, 1)$. Then we have $f(x_i) \leq t_i$ for $i = 1, 2$. Thus the convexity of f ensures that

$$f\big(\lambda x_1 + (1-\lambda)x_2\big) \leq \lambda f(x_1) + (1-\lambda) f(x_2) \leq \lambda t_1 + (1-\lambda) t_2.$$

This implies that

$$\begin{aligned} &\lambda(x_1, t_1) + (1-\lambda)(x_2, t_2) \\ &= (\lambda x_1 + (1-\lambda)x_2, \lambda t_1 + (1-\lambda)t_2) \in \operatorname{epi}(f), \end{aligned}$$

and thus the epigraph $\operatorname{epi}(f)$ is a convex subset of $\mathbb{R}^n \times \mathbb{R}$.

Conversely, suppose that the set $\mathrm{epi}(f)$ is convex and fix $x_1, x_2 \in \mathrm{dom}\, f$ and a number $\lambda \in (0,1)$. Then $(x_1, f(x_1)), (x_2, f(x_2)) \in \mathrm{epi}(f)$. This tells us that

$$\big(\lambda x_1 + (1-\lambda)x_2, \lambda f(x_1) + (1-\lambda) f(x_2)\big)$$

$$= \lambda\big(x_1, f(x_1)\big) + (1-\lambda)\big(x_2, f(x_2)\big) \in \mathrm{epi}(f)$$

and thus we arrive at the inequality

$$f\big(\lambda x_1 + (1-\lambda)x_2\big) \leq \lambda f(x_1) + (1-\lambda) f(x_2).$$

This proves the convexity of f. $\square$

Theorem 3.3 (Jensen Inequality) A function $f\colon \mathbb{R}^n \to \overline{\mathbb{R}}$ is convex if and only if for any numbers $\lambda_i > 0$ as $i = 1, \ldots, m$ with $\sum_{i=1}^m \lambda_i = 1$ and for any elements $x_i \in \mathbb{R}^n$, with $i = 1, \ldots, m$, it holds that

$$f\Big(\sum_{i=1}^m \lambda_i x_i\Big) \leq \sum_{i=1}^m \lambda_i f(x_i).$$

Proof Since the inequality being satisfied immediately implies the convexity of f , we only need to prove that any convex function f satisfies the Jensen inequality. By Theorem 3.2, the set $\mathrm{epi}(f)$ is convex in $\mathbb{R}^n \times \mathbb{R}$. Fix $x_i \in \mathbb{R}^n$ and $\lambda_i > 0$ for $i = 1, \ldots, m$ with $\sum_{i=1}^m \lambda_i = 1$. It suffices to consider the case where $x_i \in \mathrm{dom}(f)$ for $i = 1, \ldots, m$. Then $(x_i, f(x_i)) \in \mathrm{epi}(f)$ for every $i = 1, \ldots, m$. From Proposition 2.6, one has

$$\textstyle\sum_{i=1}^m \lambda_i (x_i, f(x_i)) = (\sum_{i=1}^m \lambda_i x_i, \sum_{i=1}^m \lambda_i f(x_i)) \in \mathrm{epi}(f),$$

This implies the Jensen inequality completing the proof. □

Proposition 3.4 Let $f_i \colon \mathbb{R}^n \to \overline{\mathbb{R}}$ be convex functions for all $i = 1, \ldots, m$. Then the following functions are convex as well:

$(\mathbf{i})$ The multiplication by scalars λf for any $\lambda > 0$.
$(\mathbf{ii})$ The sum function $\sum_{i=1}^{m} f_i$.
$(\mathbf{iii})$ The maximum function $\max\limits_{1 \le i \le m} f_i$.

Proof The convexity of λf in $(\mathbf{i})$ follows directly from the definition. It is sufficient to prove $(\mathbf{ii})$ and $(\mathbf{iii})$ for $m = 2$, since the general cases immediately follow by induction.

$(\mathbf{ii})$ Fix any $x, y \in \mathbb{R}^n$ and $\lambda \in (0, 1)$. Then we have

$$
\begin{aligned}
&\big(f_1 + f_2\big)\big(\lambda x + (1-\lambda)y\big) \\
&= f_1\big(\lambda x + (1-\lambda)y\big) + f_2\big(\lambda x + (1-\lambda)y\big) \\
&\le \lambda f_1(x) + (1-\lambda) f_1(y) + \lambda f_2(x) + (1-\lambda) f_2(y) \\
&= \lambda (f_1 + f_2)(x) + (1-\lambda)(f_1 + f_2)(y),
\end{aligned}
$$

which verifies the convexity of the sum function $f_1 + f_2$.

(**iii**) Denote $g := \max\{f_1, f_2\}$ and get for any $x, y \in \mathbb{R}^n$ and $\lambda \in (0,1)$ that

$$\begin{aligned} f_i\big(\lambda x + (1-\lambda)y\big) &\le \lambda f_i(x) + (1-\lambda) f_i(y) \\ &\le \lambda g(x) + (1-\lambda) g(y) \end{aligned}$$

for $i = 1, 2$. This directly implies that

$$g\big(\lambda x + (1-\lambda)y\big) = \max\big\{f_1\big(\lambda x + (1-\lambda)y\big), f_2\big(\lambda x + (1-\lambda)y\big)\big\} \le \lambda g(x) + (1-\lambda) g(y),$$

which shows that the maximum function $g(x) = \max\{f_1(x), f_2(x)\}$ is convex. □

The next result concerns the preservation of convexity under function compositions.

Proposition 3.5 Suppose that $f : \mathbb{R}^n \to \mathbb{R}$ is convex and let $\phi : \mathbb{R} \to \overline{\mathbb{R}}$ is nondecreasing and convex on a convex set containing the range of the function f. Then the composition $\phi \circ f$ is convex.

Proof Take any $x_1, x_2 \in \mathbb{R}^n$ and $\lambda \in (0,1)$. Then we have by the convexity of f that

$$f\big(\lambda x_1 + (1-\lambda)x_2\big) \le \lambda f(x_1) + (1-\lambda) f(x_2).$$

By the nondecreasing convex properties of ϕ,

$$\begin{aligned}
&\big(\phi \circ f\big)\big(\lambda x_1 + (1-\lambda)x_2\big) \\
&= \phi\big(f\big(\lambda x_1 + (1-\lambda)x_2\big)\big) \\
&\le \phi\big(\lambda f(x_1) + (1-\lambda) f(x_2)\big) \\
&\le \lambda \phi\big(f(x_1)\big) + (1-\lambda)\phi\big(f(x_2)\big) \\
&= \lambda(\phi \circ f)(x_1) + (1-\lambda)(\phi \circ f)(x_2),
\end{aligned}$$

which verifies the convexity of the composition $\phi \circ f$. $\square$

Now we consider the composition of a convex function and an affine mapping.

Proposition 3.6 Let $B\colon \mathbb{R}^n \to \mathbb{R}^p$ be an affine mapping between linear spaces and let $f : \mathbb{R}^p \to \overline{\mathbb{R}}$ be a convex function. Then the composition $f \circ B$ is convex.

Proof Taking any $x, y \in \mathbb{R}^n$ and $\lambda \in (0, 1)$, we have

$$\begin{aligned} &\big(f \circ B\big)\big(\lambda x + (1-\lambda)y\big) \\ &= f(B(\lambda x + (1-\lambda)y)) = f\big(\lambda B(x) + (1-\lambda)B(y)\big) \\ &\leq \lambda f\big(B(x)\big) + (1-\lambda) f\big(B(y)\big) = \lambda (f \circ B)(x) + (1-\lambda)(f \circ B)(y) \end{aligned}$$

and thus we've shown the convexity of the composition $f \circ B$. $\square$

Lemma 3.7 Let $I \subset \mathbb{R}$ be an interval in $\mathbb{R}$ and let the function $f : I \to \mathbb{R}$ be convex. Then for any $a, b \in I$ such that $a < b$, and any $x \in \mathbb{R}$ such that $a < x < b$ we have the following inequality:

$$\frac{f(x) - f(a)}{x - a} \leq \frac{f(b) - f(a)}{b - a} \leq \frac{f(b) - f(x)}{b - x}.$$

Proof Fix $a, b \in I$ and let $a < x < b$. This implies that $0 < x - a < b - a$. If we let $t = \frac{x-a}{b-a}$ then you will notice that $t \in (0, 1)$. Then by the convexity of f we have that

$$\begin{aligned} f(x) = f(a + x - a) &= f\big(a + \tfrac{x-a}{b-a}(b - a)\big) \\ = f\big(a + t(b - a)\big) &= f\big(tb + (1 - t)a\big) \\ &\leq tf(b) + (1 - t)f(a). \end{aligned}$$

Thus we have that $f(x) \leq tf(b) + (1 - t)f(a)$. By subtracting $f(a)$ from both sides we get

$$f(x) - f(a) \leq t\big(f(b) - f(a)\big) = \frac{x-a}{b-a}\Big(f(b) - f(a)\Big).$$

Dividing both sides by $x - a$ we get the first inequality, namely

$$\frac{f(x) - f(a)}{x - a} \leq \frac{f(b) - f(a)}{b - a}.$$

We'll now prove the second inequality. From before we have that $f(x) \leq tf(b) + (1-t)f(a)$. Subtracting $f(b)$ from both sides gives us

$$f(x) - f(b) \leq (t-1)\big(f(b) - f(a)\big)$$

$$= \Big(\tfrac{x-a}{b-a} - 1\Big)(f(b) - f(a)) = \Big(\tfrac{x-b}{b-a}\Big)(f(b) - f(a)).$$

Notice that $x - b < 0$ thus when we divide both sides by $x - b$, we get

$$\frac{f(x) - f(b)}{x - b} \geq \frac{f(b) - f(a)}{b - a},$$

which gives us our second inequality completing the proof. $\square$

Theorem 3.8 Let $I \subset \mathbb{R}$ be an open interval and let $f : I \to \mathbb{R}$ be a differentiable function. Then f is convex if and only if f' is nondecreasing on I.

(f' is said to be nondecreasing if whenever $x \leq y$ for $x, y \in I$ we have that $f'(x) \leq f'(y)$).

Proof First we shall suppose f is convex on I and prove that f' is nondecreasing.

Fix $a, b \in I$ with $a < b$. Then by Lemma 3.7 we have that

$$\frac{f(x) - f(a)}{x - a} \leq \frac{f(b) - f(a)}{b - a}$$

for all $a < x < b$. Then

$$\lim_{x \to a^+} \frac{f(x) - f(a)}{x - a} \leq \frac{f(b) - f(a)}{b - a}.$$

But since f is differentiable by assumption, this is simply

$$f'(a) \leq \frac{f(b) - f(a)}{b - a}.$$

Similarly we have that

$$\frac{f(b)-f(a)}{b-a} \leq \frac{f(b)-f(x)}{b-x} = \frac{f(x)-f(b)}{x-b}$$

for all $a < x < b$. Then

$$\frac{f(b)-f(a)}{b-a} \leq \lim_{x \to b^-} \frac{f(b)-f(x)}{b-x}.$$

And again, by the differentiability of f, this gives us

$$\frac{f(b)-f(a)}{b-a} \leq f'(b).$$

Thus $f'(a) \leq f'(b)$ making f' nondecreasing.

Now we shall prove the other implication. We shall assume that f' is nondecreasing on I, and show that f is convex on I.

Fix $x, y \in I$ and $t \in (0,1)$. If $x = y$ then obviously
$f(tx + (1-t)y) \leq tf(x) + (1-t)f(y)$ so assume the case where $x < y$. If we let
$x_t = tx + (1-t)y$, then we have that $x < x_t < y$. By the Mean Value Theorem
there exists a c_1, c_2 such that $x < c_1 < x_t < c_2 < y$ and also we have that

$$f(x_t) - f(x) = f'(c_1)(x_t - x) = f'(c_1)(1-t)(y-x)$$

and

$$f(x_t) - f(y) = f'(c_2)(x_t - y) = f'(c_2)(t)(x-y).$$

Multiplying both sides of the first equation by t and both sides of the second
equation by $(1-t)$ and then adding the two equations together we get

$$f(x_t) - tf(x) - (t-t)f(y) = t(1-t)(y-x)\Big(f'(c_2) - f'(c_1)\Big).$$

But notice that because f' is nondecreasing we have that $f'(c_2) - f'(c_1) \leq 0$ and
also we have that $t(1-t)(y-x) > 0$. Thus we get the inequality

$$f(x_t) - tf(x) - (1-t)f(y) \leq 0.$$

Finally we get that $f(tx + (1-t)y) \leq tf(x) + (1-t)f(y)$. Therefore f is convex. $\square$

Corollary 3.9 Let $I \subset \mathbb{R}$ be an open interval and let the function $f : I \to \mathbb{R}$ be twice differentiable. Then f is convex on I if and only if $f''(x) \geq 0$ for all $x \in I$

Proof By Theorem 3.8 f is convex on I if and only if f' is nondecreasing on I. And we have that f' is nondecreasing on I if and only if $f''(x) \geq 0$ for all $x \in I$. $\square$

4 Distance Function

Definitions Let $\Omega \subset \mathbb{R}^n$ be a nonempty convex set. If $x \in \mathbb{R}^n$, then the *distance (https://en.wikipedia.org/wiki/Metric_(mathematics))* from x to Ω is given by the function

$$d(x;\Omega) = \inf\{\|x-y\| : y \in \Omega\}.$$

Let $A \subset \mathbb{R}$ be nonempty. Then $m \in \mathbb{R}$ is called a *lower bound* for A if $x \geq m$ for all $x \in A$.

A is *bounded below* if there exists a lower bound, furthermore $\inf(A)$ is the greatest lower bound.

Recall the following two properties.

Let $A \subset \mathbb{R}$ is nonempty and bounded below, and let $\alpha = \inf\{A\}$. Then for every $\epsilon > 0$ there exists an $a \in A$ such that $\alpha \leq a < a + \epsilon$.

Let $\Omega \subset \mathbb{R}^n$ be a nonempty set. Then $x \in \overline{\Omega}$ if and only if there exists a sequence $\{x_n\} \subset \Omega$ such that $\|x_n - x\| \to 0$ as $n \to \infty$. We shall use these facts to prove the next proposition.

Proposition 4.1 Let $\Omega \subset \mathbb{R}^n$ be nonempty. Then

(**i**) $d(x;\Omega) = 0$ if and only if $x \in \overline{\Omega}$.
(**ii**) $|d(x;\Omega) - d(u;\Omega)| \leq \|x - u\|$ for all $x, u \in \mathbb{R}$.

Proof
(**i**) First assume $d(x;\Omega) = \inf\{\|x - y\| : y \in \Omega\} = 0$. Then for each $k \in \mathbb{N}$ there exists a $y_k \subset \Omega$ such that.

$$0 \leq \|x - y_k\| < 0 + \tfrac{1}{k} = \tfrac{1}{k} \to 0 \text{ as } k \to \infty.$$

Thus $x \in \overline{\Omega}$.

Now let $x \in \overline{\Omega}$ and we will show that $d(x;\Omega) = 0$. Since $x \in \overline{\Omega}$ then there exists a $\{y_k\} \in \Omega$ such that y_k converges to x. Then

$$0 \leq d(x;\Omega) = \inf\{\|x - y\| : y \in \Omega\} \leq \|x - y_k\| \to 0 \text{ as } k \to \infty.$$

This implies that $d(x;\Omega) = 0$.

(**ii**) Fix any $x, u \in \mathbb{R}^n$. Then for any $y \in \Omega$ we have that

$$d(x;\Omega) \leq \|x-y\| = \|x-u+y-u\| \leq \|x-u\| + \|u-y\|.$$

Since this is true for all $y \in \Omega$ we have that

$$d(x;\Omega) \leq \|x-u\| + \inf\{\|u-y\| \colon y \in \Omega\} = \|x-u\| + d(u;\Omega).$$

Thus we have shown that $d(x;\Omega) - d(u;\Omega) \leq \|x-u\|$ and using the same argument we can show that $d(u;\Omega) - d(x;\Omega) \leq \|x-u\|$. This completes the proof of part (**ii**). $\square$

Definition Let $\Omega \subset \mathbb{R}^n$ be nonempty and let $x \in \mathbb{R}^n$. *The Euclidean Projection* from x to Ω is the set defined by

$$\Pi(x) = \{y \in \Omega \colon d(x;\Omega) = \|x-y\|\}.$$

It is the set of all elements in Ω on which the distance x from Ω is attained.

Proposition 4.2 Let $\Omega \subset \mathbb{R}^n$ be a nonempty closed set. Then $\Pi(x;\Omega) \neq \emptyset$ for all $x \in \mathbb{R}^n$.

Proof Notice that $d(x;\Omega) = \inf\{\|x-y\| : y \in \Omega\}$. Then for each $n \in \mathbb{N}$ there exists a $y_n \in \Omega$ such that

$$d(x;\Omega) \leq \|x-y_n\| < d(x;\Omega) + \tfrac{1}{n}.$$

Using the triangle inequality we have that

$$\|y_n\| - \|x\| < \|x-y_n\| < d(x;\Omega) + \tfrac{1}{n} \leq d(x;\Omega) + 1.$$

Then we have that $0 \leq \|y_n\| < d(x;\Omega) + \|x\| + 1$. Thus the sequence $\{y_n\}$ is bounded and there exists a convergent subsequence, we'll denote $\{y_{nl}\}$ and we'll let $y_{nl} \to y$.

Subtracting the inequality from before by $d(x;\Omega)$ and replacing y_n with y_{nl} we get

$$0 \leq \|x-y_{nl}\| - d(x;\Omega) < d(x;\Omega) - d(x;\Omega) + \tfrac{1}{n_l} = \tfrac{1}{n_l}.$$

Taking the $n_l \to \infty$ we get

$$0 \leq \|x-y\| - d(x;\Omega) \leq 0.$$

By the squeeze theorem we have that $\|x - y\| = d(x; \Omega)$, therefore $y \in \Pi(x; \Omega)$. $\square$

Corollary 4.3 Let $\Omega \subset \mathbb{R}^n$ be a closed convex set. Then $\Pi(x;\Omega)$ is a singleton for all $x \in \mathbb{R}^n$.

Proof In Proposition 4.2 we showed that for any $x \in \mathbb{R}^n$ we have $\Pi(x;\Omega)$ is nonempty when Ω is closed, so we simply need to prove there exists only one element.

Fix $x \in \mathbb{R}^n$. Suppose by contradiction that there exists $w_1, w_2 \in \Pi(x;\Omega)$ with $w_1 \neq w_2$. Then by definition $\|x - w_1\| = \|x - w_2\|$.

Recall that the parallelogram law gives us that $\frac{1}{2}\Big(\|a+b\|^2 + \|a-b\|^2\Big) = \|a\|^2 + \|b\|^2$. With this in mind, we have that

$$2\|x - w_1\|^2 = \|x - w_1\|^2 + \|x - w_1\|^2$$

$$= \|x - w_1\|^2 + \|x - w_2\|^2 = \tfrac{1}{2}\Big(\|2x - (w_1 + w_2)\|^2 + \|w_1 - w_2\|^2\Big)$$

$$= 2\Big(\|x - \tfrac{w_1+w_2}{2}\|^2 + \tfrac{1}{4}\|w_1 - w_2\|^2\Big).$$

Then we have that

$$\|x - w_1\|^2 = \|x - \tfrac{w_1+w_2}{2}\|^2 + \tfrac{1}{2}\|w_1 - w_2\|^2.$$

This implies that

$$\begin{aligned}\|x - \tfrac{w_1+w_2}{2}\|^2 &= \|x - w_1\|^2 - \tfrac{1}{2}\|w_1 - w_2\|^2 \\ &< \|x - w_1\|^2 = d(x;\Omega)^2,\end{aligned}$$

and by the convexity of Ω we have that $\frac{w_1+w_2}{2} \in \Omega$, but we just showed that $\|x - \frac{w_1+w_2}{2}\| < d(x;\Omega)$ which gives us our contradiction. Therefore $\Pi(x;\Omega)$ must be a singleton. $\square$

Proposition 4.4 Let $\Omega \subset \mathbb{R}^n$ be a nonempty convex set. Then the function given by $f(x) = d(x;\Omega)$ is a convex function.
Proof Fix $x_1, x_2 \in \mathbb{R}^n$ and let $0 \leq t \leq 1$.

Let $\epsilon > 0$. By definition, there exists a $y_1 \in \Omega$ such that $\|x_1 - y_1\| < d(x_1;\Omega) + \epsilon$.

Similarly, there exists a $y_2 \in \Omega$ such that $\|x_2 - y_2\| < d(x_2;\Omega) + \epsilon$. With this in mind we have

$$\begin{aligned}
&\Big\| \big[tx_1 + (1-t)x_2\big] - \big[ty_1 + (1-t)y_2\big] \Big\| \\
&= \Big\| t\big[x_1 - y_1\big] + (1-t)\big[x_2 - y_2\big] \Big\| \\
&\leq t\|x_1 - y_1\| + (1-t)\|x_2 - y_2\| < t\Big(d(x_1;\Omega) + \epsilon\Big) + (1-t)\Big(d(x_2;\Omega) + \epsilon\Big) \\
&= td(x_1;\Omega) + (1-t)d(x_2;\Omega) + \epsilon = f(x_1) + f(x_2) + \epsilon.
\end{aligned}$$

Since Ω is convex we have that $ty_1 + (1-t)y_2 \in \Omega$. Thus

$$\begin{aligned}
f(tx_1 + (1-t)x_2) &= d(tx_1 + (1-t)x_2;\Omega) \\
&\leq \Big\| \big[tx_1 + (1-t)x_2\big] - \big[ty_1 + (1-t)y_2\big] \Big\| \\
&< tf(x_1) + (1-t)f(x_2) + \epsilon.
\end{aligned}$$

Since this is true for any $\epsilon > 0$ we then have that
$f(tx_1 + (1-t)x_2) \leq tf(x_1) + (1-t)f(x_2)$. Thus f is a convex function. $\square$

Proposition 4.5 Let Ω be a nonempty convex subset of $\mathbb{R}^n$, and let $u \in \Omega$. Then $u \in \Pi(x;\Omega)$ for some $x \in \mathbb{R}^n$ if and only if $\langle x-u, v-u\rangle \leq 0$ for all $v \in \Omega$.
Proof First assume that $u \in \Pi(x;\Omega)$. For any $\lambda \in (0,1)$ we have that
$\lambda v + (1-\lambda)u = u + \lambda(v-u) \in \Omega$ by the convexity of Ω. Then

$$\|x-u\|^2 = d(x;\Omega)^2 \leq \big\|x - \big(u + \lambda(v-u)\big)\|^2$$

$$= \|x-u\|^2 - 2\lambda\langle x-u, v-u\rangle + \lambda^2\|v-u\|^2.$$

This implies that $2\langle x-u, v-u\rangle \leq \lambda\|v-u\|^2$. And since this is true for all $\lambda \in (0,1)$ we have that $\langle x-u, v-u\rangle \leq 0$.

The other implication we will leave as an exercise. $\square$

Proposition 4.6 Let $\Omega \subset \mathbb{R}^n$ be a nonempty closed and convex set, and let $x_1, x_2 \in \mathbb{R}^n$. Then we have that $\|\Pi(x_1;\Omega) - \Pi(x_2;\Omega)\| \leq \|x_1 - x_2\|$.
Proof Let $\Pi(x_1;\Omega) = w_1$ and $\Pi(x_2;\Omega) = w_2$. Then by Proposition 4.5 we have that $\langle x_1 - w_1, w_2 - w_1\rangle \leq 0$ and also that $\langle x_2 - w_2, w_1 - w_2\rangle \leq 0$. Adding these two inequalities together we get

$$\langle w_1 - x_1 + x_2 - w_2, w_1 - w_2\rangle \leq 0.$$

This implies that

$$-\langle x_1 - x_2, w_1 - w_2\rangle + \langle w_1 - w_2, w_1 - w_2\rangle \leq 0.$$

Then using the Cauchy-Schwartz Inequality we get that

$$\|w_1 - w_2\|^2 \leq \langle x_1 - x_2, w_1 - w_2\rangle \leq \|x_1 - x_2\|\|w_1 - w_2\|.$$

Thus we get that $\Pi(x_1;\Omega) - \Pi(x_2;\Omega) \leq \|x_1 - x_2\|^2$. $\square$

Mathematical Programming 24 (1982) 225–228.
North-Holland Publishing Company

SHORT COMMUNICATION

A NOTE ON THE EXISTENCE OF SUBGRADIENTS*

J.M. BORWEIN

Carnegie-Mellon University, Pittsburgh, U.S.A.

and

Dalhousie University, Halifax, Canada

Received 19 March 1982
Revised manuscript received 19 April 1982

We describe an apparently novel way of constructing the subgradient of a convex function defined on a finite dimensional vector space.

Key words: Convex Function, Subgradient, Max-formula.

The existence of subgradients for continuous convex functions plays a central role in optimization theory [1, 3, 5–8]. Typically the existence of subgradients is shown by some separation argument which itself relies on a substratum of topology and calculus. In this note we describe an apparently novel construction of the "max-formula" for subgradients which relies only on the definition of the directional derivative and on linear algebra. This also makes the subgradient immediately accessible for pedagogical purposes. As is well known, the existence of subgradients is itself equivalent to all the other standard separation or duality principles [3, 4]. Thus our result can be satisfactorily used to base most further analysis.

A few preliminary definitions and notations (essentially as in [6]) need to be reviewed. We let X be a finite dimensional real vector space and let X^* denote the linear functionals on X. Let $f : X \to]-\infty, \infty]$ be a *proper convex* function. This is to say that the *effective domain* of f, $\operatorname{dom} f := \{x \in X \mid f(x) < \infty\}$, is non-empty and that

$$t_1 f(x^1) + t_2 f(x^2) \geq f(t_1 x^1 + t_2 x^2) \tag{1}$$

for all x_1, x_2 in $\operatorname{dom} f$ and $t_1, t_2 \geq 0$ with $t_1 + t_2 = 1$. If (1) holds for all non-negative t_1 and t_2, then f is said to be *sublinear*. Recall that the *core* of a convex set C, denoted by $\operatorname{core} C$, is the set of points x in C with the property that, for each y

* Research partially funded on NSERC grant A5116.

in X, one can find $\delta > 0$ with $x + ty \in C$ for $0 < t < \delta$. Also, the *directional derivative* (or *minorant*) of f at x_0 is defined by

$$\nabla f(x_0; h) := \inf_{t>0} \frac{f(x_0 + th) - f(x_0)}{t} \tag{2}$$

for each h in X. Now $\nabla f(x_0; \cdot)$ is always positively homogeneous and well defined with values in $[-\infty, \infty]$. For completeness we include a self-contained proof of the following proposition [6].

Proposition A. *Let $f : X \to]\infty, \infty]$ be convex and proper. Let x_0 lie in* core(dom f). *Then $\nabla f(x_0; \cdot)$ is an everywhere finite sublinear function.*

Proof. Let $r_t(h) := t^{-1}[f(x_0 + th) - f(x_0)]$ for t non-zero in $\mathbb{R}$. Then r_t is a convex function which (because f is convex) satisfies

$$r_t(h) \geq r_s(h) \geq r_{-s}(h) \geq r_{-t}(h) \tag{3}$$

for $0 < s < t$ and h in X. Since x_0 lies in core(dom f) one can find t with both $r_{-t}(h)$ and $r_t(h)$ finite. It follows from (3) that $\nabla f(x_0; h)$ is always finite. Moreover, if h, k lie in X one has (again because f is convex) that

$$\nabla f(x_0; h + k) \leq r_s(h + k) \leq r_{2s}(h) + r_{2s}(k) \leq r_{2s}(h) + r_{2t}(k)$$

for $0 < s < t$. Taking infima first with respect to s and then with respect to t shows that $\nabla f(x_0; h + k) \leq \nabla f(x_0; h) + \nabla f(x_0; k)$. Since $\nabla f(x_0; \cdot)$ is positively homogeneous, this shows that $\nabla f(x_0; \cdot)$ is sublinear.

Now let us recall that the *subgradient set* of f at x_0 is defined by

$$\partial f(x_0) := \{x^* \in X^* \mid x^*(x - x_0) \leq f(x) - f(x_0), \text{ for all } x \in \text{dom } f\}. \tag{4}$$

We may now state and establish our result.

Theorem B. *Let $f : X \to]-\infty, \infty]$ be proper and convex. Let x_0 lie in* core(dom f). *Then, for each h in X,*

$$\nabla f(x_0; h) = \max\{x^*(h) \mid x^* \in \partial f(x_0)\}. \qquad \text{MAX FORMULA} \tag{5}$$

In particular $\partial f(x_0)$ is non-empty.

Proof. Let us fix h in X. It is easily verified that $\nabla f(x_0; h)$ is an upper bound for the right-hand side of (5). Thus it suffices to establish the existence of a subgradient x^* with $\nabla f(x_0; h) = x^*(h)$. Let us consider a basis $B := \{e_k \mid 1 \leq k \leq n\}$ for X with $e_1 := h$. Recursively define

$$\begin{aligned} &\text{(i)} && p_0 := \nabla f(x_0; \cdot), \\ &\text{(ii)} && p_k := \nabla p_{k-1}(e_k; \cdot) \end{aligned} \tag{6}$$

for $1 \leq k \leq n$. It follows from Proposition A that each p_k is sublinear and finite. Moreover, for x in X and $1 \leq k \leq n$,

$$p_n(x) \leq p_k(x) \leq p_{k-1}(e_k + x) - p_{k-1}(e_k) \leq p_{k-1}(x) \leq p_0(x). \tag{7}$$

Now the definition of p_k and (7) shows that, for $1 \leq k \leq m \leq n$,

$$0 \leq p_m(e_k) + p_m(-e_k) \leq p_k(e_k) + p_k(-e_k) = p_{k-1}(e_k) + (-p_{k-1}(e_k)) = 0, \tag{8}$$

since each p_m is sublinear. Then (8) shows that $p_m(-e_k) = -p_m(e_k)$ for $1 \leq k \leq m \leq n$. This implies that p_m being sublinear is actually linear on the span of $\{e_k \mid 1 \leq k \leq m\}$. In particular p_n must be linear. Set $x^* := p_n$. Now (7) shows that

$$x^*(x - x_0) \leq p_0(x - x_0) \leq f(x) - f(x_0) \tag{9}$$

for x in X; and so $x^* \in \partial f(x_0)$. Finally (7) and (8) show that

$$-x^*(h) = x^*(-e_1) \leq p_1(-e_1) = -p_0(e_1) = -p_0(h). \tag{10}$$

This implies that $x^*(h) = \nabla f(x_0; h)$ as required.

The same argument in combination with the appropriate maximality principle can be used to establish Theorem B in arbitrary vector spaces or for convex operators [2]. The basic iteration remains unchanged. One indexes a basis for X by the ordinals preceding some cardinal δ and defines a "sequence" of sublinear operators (p_α) by (i) using (6) for successor ordinals, and (ii) defining $p_\alpha := \inf\{p_\beta \mid \beta < \alpha\}$ when α is a limit ordinal. The proof is essentially unchanged.

Geometrically the proof is very simple. Each directional derivative minorizes the previous one and is guaranteed to be linear in at least one more direction. After n steps we must produce an appropriate linear minorant. In general, many fewer than n steps will be needed. After all f is differentiable almost everywhere in core(dom f) [6]. At such points $\nabla f(x_0; \cdot)$ is itself the appropriate function. The following simple example shows that n iterations may be needed.

Example C. Let $X := \mathbb{R}^n$. Let $\{\delta_j \mid 1 \leq j \leq n\}$ be the usual basis and let f be defined by $f(x) := \max\{x_j \mid 1 \leq j \leq n\}$. Let $x_0 := 0$ and $h := \sum_{j=1}^n \delta_j$. Consider the iteration in (6) with $e_k := \sum_{j=1}^{n+1-k} \delta_j$. Then $p_0 = f$ and $p_k := \max\{x_j \mid 1 \leq j \leq n - k + 1\}$ for $1 \leq k \leq n$. This is easily established by induction. It follows that p_n is linear, as promised, but no previous p_k is. Notice also that $p_n \in \partial f(0)$ and

$$1 = p_n(h) = p_0(h) = \nabla f(x_0; h),$$

as claimed by (5).

An immediate consequence of (5) is that $\partial f(x_0)$ is singleton exactly when $\nabla f(x_0; \cdot)$ is linear (and f is Gateaux differentiable at x_0).

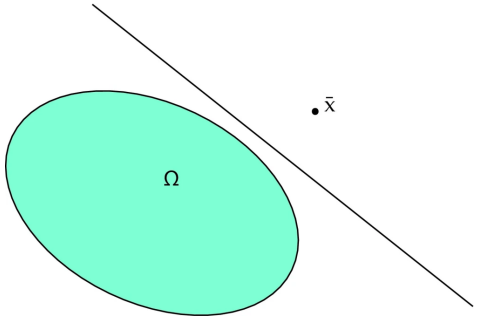
We leave an open question as to whether (6) has any possible utility in making a numerical estimate of a subgradient of a convex function?

References

[1] M.S. Bazaraa and C.M. Shetty, *Nonlinear programming. Theory and algorithms* (Wiley, New York, 1979).
[2] J.M. Borwein, "Subgradients of convex operators", Research Report 82-3, Department of Mathematics, Carnegie-Mellon University (Pittsburgh, PA, 1982).
[3] R.B. Holmes, *Geometric functional analysis and its applications* (Springer, New York, 1975).
[4] D.G. Luenberger, *Optimization by vector space methods* (Wiley, New York, 1969).
[5] A.W. Roberts and D.E. Varberg, *Convex functions* (Academic Press, New York, 1973).
[6] R.T. Rockafellar, *Convex analysis* (Princeton University Press, Princeton, NJ, 1970).
[7] J.F. Shapiro, *Mathematical programming: Structures and algorithms* (Wiley, New York, 1979).
[8] J. Stoer and C. Witzgall, *Convexity and optimization in finite dimensions I* (Springer, Berlin, 1970).

5 Convex Separation

In the next proposition we will prove that any point outside of a closed convex set can be separated from that set with a hyperplane (https://en.wikipedia.org/wiki/Hyperplane).



Proposition 5.1 Let $\Omega \subset \mathbb{R}^n$ be a nonempty closed convex set and let $\bar{x} \notin \Omega$. Then there exists a nonzero vector $v \in \mathbb{R}^n$ such that

$$\sup\{\langle v, x\rangle \mid x \in \Omega\} < \langle v, \bar{x}\rangle.$$

Proof Denote $\bar{w} = \pi(\bar{x}; \Omega)$, let $v = \bar{x} - \bar{w}$ and fix any $x \in \Omega$. By Proposition 4.5,

$$\langle v, x - \bar{w}\rangle = \langle \bar{x} - \bar{w}, x - \bar{w}\rangle \leq 0.$$

It follows that

$$\langle v, x - \bar{w}\rangle = \langle v, x - \bar{x} + \bar{x} - \bar{w}\rangle = \langle v, x - \bar{x} + v\rangle \leq 0.$$

The last inequality implies

$$\langle v, x\rangle \le \langle v, \bar{x}\rangle - \|v\|^2.$$

Therefore, $\sup\{\langle v, x\rangle \mid x \in \Omega\} < \langle v, \bar{x}\rangle$, which completes the proof. □

The next theorem extends the result of Proposition 5.1 to the case of two nonempty, closed, convex sets at least one of which is bounded.

Theorem 5.2 Let Ω_1 and Ω_2 be nonempty, closed, convex subsets of $\mathbb{R}^n$ with $\Omega_1 \cap \Omega_2 = \emptyset$. If Ω_1 is bounded or Ω_2 is bounded, then there is a nonzero element $v \in \mathbb{R}^n$ such that

$$\sup\big\{\langle v, x\rangle \;\big|\; x \in \Omega_1\big\} < \inf\big\{\langle v, y\rangle \;\big|\; y \in \Omega_2\big\}.$$

Proof Denote $\Omega := \Omega_1 - \Omega_2$. Then Ω is a nonempty, closed, convex set and $0 \notin \Omega$. Applying Proposition 5.1 to Ω and $\bar{x} = 0$, we have

$$\gamma := \sup\big\{\langle v, x\rangle \;\big|\; x \in \Omega\big\} < 0 = \langle v, \bar{x}\rangle \;\text{ with some }\; 0 \ne v \in \mathbb{R}^n.$$

For any $x \in \Omega_1$ and $y \in \Omega_2$, we have $\langle v, x - y\rangle \le \gamma$, and so $\langle v, x\rangle \le \gamma + \langle v, y\rangle$. Therefore,

$$\sup\big\{\langle v, x\rangle \;\big|\; x \in \Omega_1\big\} \le \gamma + \inf\big\{\langle v, y\rangle \;\big|\; y \in \Omega_2\big\} < \inf\big\{\langle v, y\rangle \;\big|\; y \in \Omega_2\big\},$$

which completes the proof of the theorem.

Convex Seperations - Pt 2

Remark 5.3 If Ω is a nonempty convex set in $\mathbb{R}^n$ and $\bar{x} \notin \bar{\Omega}$, applying Proposition 5.1 for the convex set $\bar{\Omega}$ gives a nonzero vector $v \in \mathbb{R}^n$ such that

$$\sup\{\langle v, x\rangle \mid x \in \Omega\} \leq \sup\{\langle v, x\rangle \mid x \in \bar{\Omega}\} < \langle v, \bar{x}\rangle.$$

The next property presents a separation property in a subspace of $\mathbb{R}^n$ instead of in $\mathbb{R}^n$.

Proposition 5.4 Let L be a subspace of $\mathbb{R}^n$ and let $\Omega \subset L$ be a nonempty convex set with $\bar{x} \in L$ and $\bar{x} \notin \bar{\Omega}$. Then there exists $v \in L$, $v \neq 0$, such that

$$\sup\{\langle v, x\rangle \mid x \in \Omega\} < \langle v, \bar{x}\rangle.$$

Proof By Proposition 5.1, there exists $w \in \mathbb{R}^n$ such that

$$\sup\{\langle w, x\rangle \mid x \in \Omega\} < \langle w, \bar{x}\rangle.$$

It is well-known that $\mathbb{R}^n$ can be represented as $\mathbb{R}^n = L \oplus L^{\perp}$, where

$$L^{\perp} = \{u \in \mathbb{R}^n \mid \langle u, x\rangle = 0 \text{ for all } x \in L\}.$$

Thus, we have the representation $w = u + v$, where $u \in L^{\perp}$ and $v \in L$. For any $x \in \Omega \subset L$, one has $\langle u, x\rangle = 0$ and

$$\langle y, x\rangle = \langle u, x\rangle + \langle v, x\rangle = \langle u + v, x\rangle = \langle w, x\rangle$$

$$\leq \sup\{\langle w, x\rangle \mid x \in \Omega\} < \langle w, \bar{x}\rangle = \langle u + v, \bar{x}\rangle$$

$$= \langle u, \bar{x}\rangle + \langle v, \bar{x}\rangle = \langle v, \bar{x}\rangle.$$

It follows that $\sup\{\langle v, x\rangle \mid x \in \Omega\} < \langle v, \bar{x}\rangle$, which also implies $v \neq 0$. □

We continue with another important separation property called proper separation.

Definition We say that two nonempty convex sets Ω_1 and Ω_2 can be *properly separated* (https://en.wikipedia.org/wiki/Hyperplane_separation_theorem)if there exists a nonzero vector $v \in \mathbb{R}^n$ such that

$$\sup\{\langle v, x\rangle \mid x \in \Omega_1\} \leq \inf\{\langle v, y\rangle \mid y \in \Omega_2\}$$

and

$$\inf\{\langle v, x\rangle \mid x \in \Omega_1\} < \sup\{\langle v, y\rangle \mid y \in \Omega_2\}.$$

Definition (Affine Sets) A subset Ω of X is called *affine (https://en.wikipedia.org/wiki/Affine_space)* if for any $a, b \in \Omega$ we have

$$\big\{\lambda a + (1-\lambda)b \;\big|\; \lambda \in \mathbb{R}\big\}. \subset \Omega.$$

The *affine hull (https://en.wikipedia.org/wiki/Affine_hull)* of an arbitrary set $\Omega \subset \mathbb{R}^n$ is

$$\text{aff}(\Omega) := \bigcap \big\{C \;\big|\; C \text{ is affine and } \;\Omega \subset C\big\}.$$

Proposition 5.5 The following assertions hold:

(**i**) A set $\Omega \subset \mathbb{R}^n$ is affine if and only if Ω contains all affine combinations of its elements.

(**ii**) If Ω_1 is an affine subset of $\mathbb{R}^n$ and Ω_2 is an affine subsets of $\mathbb{R}^m$, then $\Omega_1 \times \Omega_2$ is an affine subset of $\mathbb{R}^n \times \mathbb{R}^m$.

(**iii**) Let $B\colon \mathbb{R}^n \to \mathbb{R}^m$ be an affine mapping. If Ω is an affine subset of $\mathbb{R}^n$ and Θ is an affine subset of $\mathbb{R}^m$, then the image $B(\Omega)$ is an affine subset of $\mathbb{R}^m$ and the inverse image $B^{-1}(\Theta)$ is an affine subset of $\mathbb{R}^n$.

(**iv**) Let Ω, Ω_1, and Ω_2 be affine subsets of $\mathbb{R}^n$. Then the sum $\Omega_1 + \Omega_2$ and the scalar product $\lambda\Omega$ for any $\lambda \in \mathbb{R}$ are also affine subsets of $\mathbb{R}^n$.

(**v**) Given $\Omega \subset \mathbb{R}^n$, its affine hull is the smallest affine set containing Ω. We have

$$\operatorname{aff}(\Omega) = \Big\{ \textstyle\sum_{i=1}^m \lambda_i w_i \;\Big|\; \sum_{i=1}^m \lambda_i = 1,\ w_i \in \Omega,\ m \in \mathbb{N} \Big\}.$$

$(\mathbf{vi})$ A set Ω is a linear subspace of $\mathbb{R}^n$ if and only if Ω is an affine set containing the origin.

Proof The proof is similar to the proof for properties of convex hulls and is left as an exercise. $\square$

Definition (Relative Interior) Let $\Omega \subset \mathbb{R}^n$ be a convex set. We say that an element $v \in \Omega$ belongs to the *relative interior (https://en.wikipedia.org/wiki/Relative_interior)*, $\mathrm{ri}(\Omega)$ of Ω, if there exists $\epsilon > 0$ such that $\mathbb{B}(v;\epsilon) \cap \mathrm{aff}(\Omega) \subset \Omega$.

Definition (Affine Independence) The elements $v_0, \ldots, v_m$ in $\mathbb{R}^n$, $m \geq 1$, are *affinely independent* if

$$\Big[\sum_{i=0}^m \lambda_i v_i = 0,\ \sum_{i=0}^m \lambda_i = 0\Big] \Longrightarrow \big[\lambda_i = 0 \ \text{ for all } \ i = 0, \ldots, m\big].$$

Definition (Simplex) Let $v_0, \ldots, v_m$ be affinely independent in $\mathbb{R}^n$. Then the set

$$\Delta_m := \operatorname{co}\{v_i \mid i = 0, \dots, m\}$$

is called an *m–simplex* (https://en.wikipedia.org/wiki/Simplex) in $\mathbb{R}^n$ with the vertices v_i, $i = 0, \dots, m$.}

Proposition 5.6 Let Δ_m be an m-simplex in $\mathbb{R}^n$ with some $m \geq 1$. Then $\operatorname{ri}(\Delta_m) \neq \emptyset$.

Proof Consider the vertices $v_0, \dots, v_m$ of the simplex Δ_m and denote

$$v := \frac{1}{m+1} \sum_{i=0}^{m} v_i.$$

We prove the proposition by showing that $v \in \operatorname{ri} \Delta_m$. Define

$$L := \operatorname{span}\{v_i - v_0 \mid i = 1, \dots, m\}$$

and observe that L is the m-dimensional subspace of $\mathbb{R}^n$ parallel to $\operatorname{aff} \Delta_m = \operatorname{aff}\{v_0, \dots, v_m\}$. It is easy to see that for every $x \in L$, there is a unique collection $(\lambda_0, \dots, \lambda_m) \in \mathbb{R}^{m+1}$ with

$$x=\sum_{i=0}^m \lambda_i v_i, \quad \sum_{i=0}^m \lambda_i=0.$$

Consider the mapping $A:L\to\mathbb{R}^{m+1}$, which maps x to the corresponding coefficients $(\lambda_0,\ldots,\lambda_m)\in\mathbb{R}^{m+1}$ as above. Then A is linear, and hence it is continuous. Since $A(0)=0$, we can choose $\delta>0$ such that

$$\|A(u)\|<\frac{1}{m+1} \quad \text{whenever} \quad \|u\|\le\delta.$$

Let us now show that $(v+\delta\mathbb{B})\cap\operatorname{aff}\Delta_m\subset\Delta_m$, which means that $v\in\operatorname{ri}\Delta_m$. To proceed, fix any $x\in(v+\delta\mathbb{B})\cap\operatorname{aff}\Delta_m$ and get that $x=v+u$ for some $u\in\delta\mathbb{B}$. Since $v,x\in\operatorname{aff}\Delta_m$ and $u=x-v$, we have $u\in L$. Denoting $A(u):=(\alpha_0,\ldots,\alpha_m)$ gives us the representation $u=\sum_{i=0}^m\alpha_i v_i$ with $\sum_{i=0}^m\alpha_i=0$ and the estimate

$$|\alpha_i|\le\|A(u)\|<\frac{1}{m+1} \quad \text{for all} \quad i=0,\ldots,m.$$

Then implies in turn that

$$v+u=\sum_{i=0}^{m}(\frac{1}{m+1}+\alpha_i)v_i=\sum_{i=0}^{m}\mu_i v_i,$$

where $\mu_i:=\dfrac{1}{m+1}+\alpha_i\geq 0$ for $i=0,\ldots,m$. Since $\sum_{i=0}^{m}\mu_i=1$, this ensures that $x\in\Delta_m$. Thus $(v+\delta\mathbb{B})\cap\operatorname{aff}\Delta_m\subset\Delta_m$ and therefore $v\in\operatorname{ri}\Delta_m$. $\square$

Lemma 5.7 Let Ω be a nonempty, convex set in $\mathbb{R}^n$ of dimension $m\geq 1$. Then there exist $m+1$ affinely independent elements $v_0,\ldots,v_m$ in Ω.

Proof Let $\Delta_k:=\{v_0,\ldots,v_k\}$ be a $k-$simplex of maximal dimension contained in Ω. Then $v_0,\ldots,v_k$ are affinely independent. To verify now that $k=m$, form $K:=\operatorname{aff}\{v_0,\ldots,v_k\}$ and observe that $K\subset\operatorname{aff}\Omega$ since $\{v_0,\ldots,v_k\}\subset\Omega$. The opposite inclusion also holds since we have $\Omega\subset K$. Justifying it, we argue by contradiction and suppose that there exists $w\in\Omega$ such that $w\notin K$. Then a direct application of the definition of affine independence shows that $v_0,\ldots,v_k,w$ are affinely independent being a subset of Ω, which is a contradiction. Thus $K=\operatorname{aff}\Omega$, and hence we get $k=\dim K=\dim\operatorname{aff}\ \Omega=\dim\Omega=m$. $\square$

Theorem 5.8 Let $\Omega \subset \mathbb{R}^n$ be a nonempty, convex set. The following assertions hold:

($\mathbf{i}$) We always have $\operatorname{ri}\Omega \neq \emptyset$.
($\mathbf{ii}$) We have $[a,b) \subset \operatorname{ri}\Omega$ for any $a \in \operatorname{ri}\Omega$ and $b \in \bar{\Omega}$.

Proof ($\mathbf{i}$) Let m be the dimension of Ω. Observe first that the case where $m = 0$ is trivial since in this case Ω is a singleton and $\operatorname{ri}\Omega = \Omega$. Suppose that $m \geq 1$ and find $m+1$ affinely independent elements $v_0, \ldots, v_m$ in Ω. Consider further the m-simplex

$$\Delta_m := \operatorname{co}\{v_0, \ldots, v_m\}.$$

We can show that $\operatorname{aff}\Delta_m = \operatorname{aff}\Omega$. To complete the proof, take $v \in \operatorname{ri}\Delta_m$ and get for any small $\epsilon > 0$ that

$$\mathbb{B}(v,\epsilon) \cap \operatorname{aff}\Omega = \mathbb{B}(v,\epsilon) \cap \operatorname{aff}\Delta_m \subset \Delta_m \subset \Omega.$$

This verifies that $v \in \operatorname{ri}\Omega$ by the definition of relative interior.

(**ii**) Let L be the subspace of $\mathbb{R}^n$ parallel to $\operatorname{aff}\Omega$ and let $m := \dim L$. Then there is a bijective linear mapping $A : L \to \mathbb{R}^m$ such that both A and A^{-1} are continuous. Fix $x_0 \in \operatorname{aff}\Omega$ and define the mapping $f : \operatorname{aff}\Omega \to \mathbb{R}^m$ by $f(x) := A(x - x_0)$. It is easy to check that f is a bijective affine mapping and that both f and f^{-1} are continuous. We also see that $a \in \operatorname{ri}\Omega$ if and only if $f(a) \in \operatorname{int} f(\Omega)$, and that $b \in \bar{\Omega}$ if and only if $f(b) \in \overline{f(\Omega)}$. Then $[f(a), f(b)) \subset \operatorname{int} f(\Omega)$. This shows that $[a, b) \subset \operatorname{ri}\Omega$. □

Theorem 5.9 Let Ω_1, Ω_2 be nonempty convex subsets of $\mathbb{R}^n$. Then we have that

$$\operatorname{ri}(\Omega_1 - \Omega_2) = \operatorname{ri}\Omega_1 - \operatorname{ri}\Omega_2.$$

Proof We leave this proof as an exercise. □

Lemma 5.10 Let Ω be a nonempty convex subset of $\mathbb{R}^n$. Suppose that $0 \in \overline{\Omega} \setminus \operatorname{ri}(\Omega)$. Then $\operatorname{aff}(\Omega)$ is a subspace of $\mathbb{R}^n$, and there exists a sequence $\{x_k\} \subset \operatorname{aff}(\Omega)$ such that $x_k \notin \overline{\Omega}$ for all $k \in \mathbb{N}$ and $x_k \to 0$ as $k \to \infty$.

Proof Suppose $0 \in \overline{\Omega} \setminus \mathrm{ri}(\Omega)$. By Theorem 5.7 $(\mathbf{i})$, the relative interior of Ω is nonempty, so there exists $x_0 \in \mathrm{ri}(\Omega)$. Then $-tx_0 \notin \overline{\Omega}$ for all $t > 0$. Indeed, by contradiction suppose that $-tx_0 \in \overline{\Omega}$ for some $t > 0$. It follows from Theorem 5.7 $(\mathbf{ii})$ that

$$0 = \tfrac{t}{1+t}x_0 + \tfrac{1}{1+t}(-tx_0) \in \mathrm{ri}(\Omega).$$

This is a contradiction because $0 \notin \mathrm{ri}(\Omega)$ by the assumption. Thus $-tx_0 \notin \overline{\Omega}$ for all $t > 0$. Let $x_k = -\frac{x_0}{k}$. Then $x_k \notin \overline{\Omega}$ for every k and $x_k \to 0$ as $k \to \infty$.

Since $\Omega \subset \mathrm{aff}(\Omega)$ and $\mathrm{aff}(\Omega)$ is closed, one has

$$0 \in \overline{\Omega} \subset \overline{\mathrm{aff}(\Omega)} = \mathrm{aff}(\Omega).$$

Thus, $\mathrm{aff}(\Omega)$ is a subspace of $\mathbb{R}^n$. This also implies that $x_k \in \mathrm{aff}(\Omega)$ for all $k \in \mathbb{N}$. $\square$

Reading materials: Proof of Theorem 5.9

1. Let $B : \mathbb{R}^n \to \mathbb{R}^p$ be an affine mapping and let Ω be a convex subset of $\mathbb{R}^n$. Prove the equality

$$B(\text{ri } \Omega) = \text{ri } B(\Omega).$$

Proof. Let $y \in B(\text{ri } \Omega)$, then there exits $x \in \text{ri } \Omega$ such that $y = Bx$. By the prolongation lemma, for any $\bar{x} \in \Omega$, there exists $\gamma > 0$ such that $x + \gamma(x - \bar{x}) \in \Omega$. Hence $y + \gamma(y - \bar{y}) = B(x + \gamma(x - \bar{x})) \in B(\Omega)$, where $\bar{y} = B\bar{x}$. Since $\bar{x}$ is arbitrary, by the prolongation lemma again, $y \in \text{ri } B(\Omega)$. Hence $B(\text{ri } \Omega) \subseteq \text{ri } B(\Omega)$.
To show the other direction, we first show that $\overline{B(\Omega)} = \overline{B(\text{ri } \Omega)}$. Note that $\overline{\Omega} = \overline{\text{ri } \Omega}$, hence we have

$$B(\Omega) \subseteq B(\overline{\Omega}) = B(\overline{\text{ri } \Omega}) \subseteq \overline{B(\text{ri } \Omega)},$$

where the last inclusion follows from the continuity of B. This shows that $\overline{B(\Omega)} \subseteq \overline{B(\text{ri } \Omega)}$. Since $\overline{B(\text{ri } \Omega)} \subseteq \overline{B(\Omega)}$, we have $\overline{B(\Omega)} = \overline{B(\text{ri } \Omega)}$.
Now since $\overline{B(\Omega)} = \overline{B(\text{ri } \Omega)}$, $\text{ri } B(\Omega) = \text{ri } B(\text{ri } \Omega)$ (see tutorial notes). Hence

$$\text{ri } B(\Omega) = \text{ri } B(\text{ri } \Omega) \subseteq B(\text{ri } \Omega).$$

□

2. Let Ω_1, Ω_2 be convex subsets of $\mathbb{R}^n$. Show that $\text{ri}(\Omega_1 - \Omega_2) = \text{ri } \Omega_1 - \text{ri } \Omega_2$.

Proof. Consider $B : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^n$ given by $B(x, y) = x - y$. Then $\Omega_1 - \Omega_2 = B(\Omega_1 \times \Omega_2)$. The equality can then be obtained by applying the previous result to B and $\Omega = \Omega_1 \times \Omega_2$. □

Exercise 1.27 **(i)** Let $B : \mathbb{R}^n \to \mathbb{R}^p$ be an affine mapping and let Ω be a convex subset of $\mathbb{R}^n$. Prove the equality

$$B(\text{ri } \Omega) = \text{ri } B(\Omega).$$

(ii) Let Ω_1 and Ω_2 be convex subsets of $\mathbb{R}^n$. Show that $\text{ri } (\Omega_1 - \Omega_2) = \text{ri } \Omega_1 - \text{ri } \Omega_2$.

Lemma 5.11 A nonempty subset Ω of $\mathbb{R}^n$ is affine if and only if $\Omega - w$ is a subspace of $\mathbb{R}^n$ for any $w \in \Omega$.

Proof Suppose that a nonempty set $\Omega \subset \mathbb{R}^n$ is affine. It follows from Proposition 5.5 (v) that $\Omega - w$ is a subspace for any $w \in \Omega$. Conversely, fix $w \in \Omega$ and suppose that $\Omega - w$ is a subspace denoted by L. Then the set $\Omega = w + L$ is obviously affine. $\square$

Proposition 5.12 Let Ω be a nonempty convex set. Suppose that $\bar{x} \in \mathrm{ri}(\Omega)$ and $\bar{y} \in \Omega$. Then there exists $t > 0$ such that

$$\bar{x} + t(\bar{x} - \bar{y}) \in \Omega.$$

Proof Choose a number $\gamma > 0$ such that

$$\mathbb{B}(\bar{x}; \gamma) \cap \mathrm{aff}(\Omega) \subset \Omega$$

and note that $\bar{x} + t(\bar{x} - \bar{y}) = (1 + t)\bar{x} + (-t)\bar{y} \in \mathrm{aff}(\Omega)$ for all $t \in \mathbb{R}$ as it is an affine combination of $\bar{x}$ and $\bar{y}$. Select $t > 0$ small enough that $\bar{x} + t(\bar{x} - \bar{y}) \in \mathbb{B}(\bar{x}; \gamma)$. Then we have $\bar{x} + t(\bar{x} - \bar{y}) \in \mathbb{B}(x; \gamma) \cap \mathrm{aff}(\Omega) \subset \Omega$. $\square$

Proposition 5.13 Let Ω be a nonempty convex set in $\mathbb{R}^n$. Then $0 \notin \mathrm{ri}(\Omega)$ if and only if the sets Ω and $\{0\}$ can be properly separated, i.e., there exists $v \in \mathbb{R}^n$, $v \neq 0$, such that

$$\sup\{\langle v, x\rangle \mid x \in \Omega\} \leq 0$$

and

$$\inf\{\langle v, x\rangle \mid x \in \Omega\} < 0.$$

Proof We consider two cases.

Case 1: First suppose$0 \notin \overline{\Omega}$. By Remark 5.3 with $\bar{x} = 0$, there exists $v \neq 0$ such that

$$\sup\{\langle v, x\rangle \mid x \in \Omega\} < \langle v, \bar{x}\rangle = 0.$$

It follows that Ω and $\{0\}$ can be properly separated.

Case 2: Now suppose $0 \in \overline{\Omega} \setminus \mathrm{ri}(\Omega)$. Let $L = \mathrm{aff}(\Omega)$. By Lemma 5.10, L is a subspace of $\mathbb{R}^n$, and there exists a sequence $\{x_k\} \subset L$ with $x_k \notin \overline{\Omega}$ for every k and $x_k \to 0$ as $k \to \infty$. By Proposition 5.4, there exists a sequence $\{v_k\} \subset L$ with $v_k \neq 0$ for all k and

$$\sup\{\langle v_k, x\rangle \mid x \in \Omega\} < \langle v_k, x_k\rangle.$$

Let $w_k = \frac{v_k}{\|v_k\|}$ and observe that $\|w_k\| = 1$ for all $k \in \mathbb{N}$. Then

$$\langle w_k, x\rangle < \langle w_k, x_k\rangle \text{ for all } x \in \Omega.$$

We can assume without loss of generality that $w_k \to v \in L$ with $\|v\| = 1$ as $k \to \infty$. Letting $k \to \infty$ in the inequality above with the observation that $|\langle w_k, x_k\rangle| \leq \|w_k\|\|x_k\| = \|x_k\| \to 0$, one has

$$\sup\{\langle v, x\rangle \mid x \in \Omega\} \leq 0.$$

To show that the condition $\inf\{\langle v, x\rangle \mid x \in \Omega\} < 0$ is satisfied, it suffices to show that there exists $x \in \Omega$ with $\langle v, x\rangle < 0$. Suppose by contradiction that $\langle v, x\rangle \geq 0$ for all $x \in \Omega$. Since $\sup\{\langle v, x\rangle \mid x \in \Omega\} \leq 0$, it follows that $\langle v, x\rangle = 0$ for all $x \in \Omega$. Since $v \in L = \operatorname{aff}(\Omega)$, we can write $v = \sum_{i=1}^{m} \lambda_i w_i$, where $\sum_{i=1}^{m} \lambda_i = 1$ and $w_i \in \Omega$ for each $i = 1, \ldots, m$. Then

$$\|v\|^2 = \langle v, v\rangle = \textstyle\sum_{i=1}^{m} \lambda_i \langle v, w_i\rangle = 0.$$

This is a contradiction because $\|v\| = 1$. Thus, Ω and $\{0\}$ can be properly separated.

Now suppose that Ω and $\{0\}$ can be properly separated. Then there exists $0 \neq v \in \mathbb{R}^n$ such that

$$\sup\{\langle v, x\rangle \mid x \in \Omega\} \leq 0,$$

and there exists $\bar{x} \in \Omega$ with $\langle v, \bar{x}\rangle < 0$. Suppose by contradiction that $0 \in \operatorname{ri}(\Omega)$. By Proposition 5.11, $0 + t(0 - \bar{x}) = -t\bar{x} \in \Omega$ for some $t > 0$.

This implies

$$\langle v, -t\bar{x}\rangle \leq \sup\{\langle v, x\rangle \mid x \in \Omega\} \leq 0.$$

Then $\langle v, \bar{x}\rangle \geq 0$, which is a contradiction. Therefore, $0 \notin \operatorname{ri}(\Omega)$. $\square$

Theorem 5.14 Let Ω_1 and Ω_2 be two nonempty convex subsets of $\mathbb{R}^n$. Then Ω_1 and Ω_2 can be properly separated if and only if $\mathrm{ri}(\Omega_1) \cap \mathrm{ri}(\Omega_2) = \emptyset$.

Proof Define $\Omega = \Omega_1 - \Omega_2$ and note that $\mathrm{ri}(\Omega_1) \cap \mathrm{ri}(\Omega_2) = \emptyset$ if and only if

$$0 \notin \mathrm{ri}(\Omega_1 - \Omega_2) = \mathrm{ri}(\Omega_1) - \mathrm{ri}(\Omega_2).$$

First, suppose that $\mathrm{ri}(\Omega_1) \cap \mathrm{ri}(\Omega_2) = \emptyset$. Then $0 \notin \mathrm{ri}(\Omega_1 - \Omega_2) = \mathrm{ri}(\Omega)$. By Proposition 5.13, the sets Ω and $\{0\}$ can be properly separated, so there exists $v \in \mathbb{R}^n$ such that $\langle v, x\rangle \leq 0$ for all $x \in \Omega$, and there exists $y \in \Omega$ such that $\langle v, y\rangle < 0$. For any $w_1 \in \Omega_1$ and $w_2 \in \Omega_2$, one has $x = w_1 - w_2 \in \Omega$, and hence

$$\langle v, w_1 - w_2\rangle = \langle v, x\rangle \leq 0.$$

This implies $\langle v, w_1\rangle \leq \langle v, w_2\rangle$. Choose $\bar{w}_1 \in \Omega_1$ and $\bar{w}_2 \in \Omega_2$ such that $y = \bar{w}_1 - \bar{w}_2$. Then

$$\langle v, \bar{w}_1 - \bar{w}_2\rangle = \langle v, y\rangle < 0,$$

which implies that $\langle v, \bar{w}_1\rangle < \langle v, \bar{w}_2\rangle$. Therefore, Ω_1 and Ω_2 can be properly separated.

Next, suppose that Ω_1 and Ω_2 can be properly separated. It follows that $\Omega = \Omega_1 - \Omega_2$ and $\{0\}$ can be properly separated. Applying Proposition 5.13 again yields

$$0 \notin \operatorname{ri}(\Omega) = \operatorname{ri}(\Omega_1 - \Omega_2) = \operatorname{ri}(\Omega_1) - \operatorname{ri}(\Omega_2).$$

Therefore, $\operatorname{ri}(\Omega_1) \cap \operatorname{ri}(\Omega_2) = \emptyset$.

The proof is now complete. $\square$

Math4230: Optimization Theory
(for reading)
Handout: Separating hyperplane theorem

Strict separation

For $x, y \in \mathbb{R}^n$, we write $d(x, y) = \|x - y\|$. For subsets $A, B \subseteq \mathbb{R}^n$, we define $d(A, x) = d(x, A) = \inf_{a \in A} d(x, a)$ and $d(A, B) = \inf_{a \in A, b \in B} d(a, b)$. Let $\operatorname{diam} A = \sup_{x,y \in A} d(x, y)$. For $A \subseteq \mathbb{R}^n$ and $x \in \mathbb{R}^n$, we define $\langle A, x \rangle = \{\langle a, x \rangle : a \in A\}$. In Euclidean space $\mathbb{R}^n$, the term *compact set* refers to any set that is closed and bounded.

Theorem 1 (Separating hyperplane theorem, strict case)**.** *Let $C, K \subseteq \mathbb{R}^n$ be nonempty convex sets with $C \cap K = \varnothing$. If C is closed and K compact, then there exists $\psi \in \mathbb{R}^n$ with*

$$\inf \langle C, \psi \rangle > \sup \langle K, \psi \rangle.$$

Proof. The strategy of the proof is illustrated in Figure 1. We start by proving the existence of a pair of closest points x^* and y^*, where $x \in C$ and $y \in K$. We then show that the hyperplane with normal vector $\psi = x^* - y^*$ separates the two convex sets. Details follow.

Claim 2. *There exist $x^* \in C$ and $y^* \in K$ such that $d(x^*, y^*) = d(C, K)$.*

Proof. For this, pick an arbitrary point $x_0 \in C$ and define $r = 2d(x_0, K) + \operatorname{diam} K$. By the triangle inequality, $d(x, x_0) \leqslant d(x, K) + \operatorname{diam} K + d(x_0, K)$. It follows that any point x with $d(x, x_0) > r$ satisfies

$$\begin{aligned} d(x, K) &\geqslant d(x, x_0) - \operatorname{diam} K - d(x_0, K) \\ &> d(x_0, K). \end{aligned}$$

As a result, the compact set $C' = C \cap \{x : d(x, x_0) \leqslant r\}$ obeys $d(C, K) = d(C', K)$. Since $d(\cdot, \cdot)$ is a continuous function on the compact $C' \times K$, it must attain its infimum on $C' \times K$, i.e., there must exist $(x^*, y^*) \in C' \times K$ with $d(x^*, y^*) = d(C', K) = d(C, K)$. □

In the remainder of the proof, fix x^* and y^* as in Claim 2, and define $\psi = x^* - y^*$.

Claim 3. $\inf \langle C, \psi \rangle \geqslant \langle x^*, \psi \rangle$.

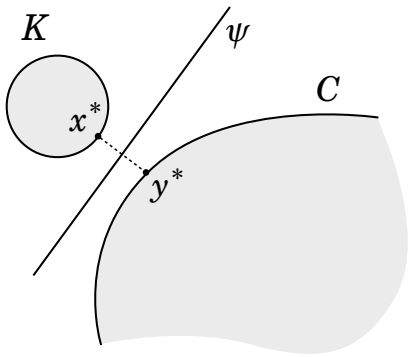


Figure 1: Separating two convex sets by a hyperplane.

Proof. For the sake of contradiction, suppose that $\langle x, \psi \rangle < \langle x^*, \psi \rangle$ for some $x \in C$. This is equivalent to

$$\langle x - x^*, x^* - y^* \rangle < 0. \tag{1}$$

For $\varepsilon \in (0, 1)$, the point $x_\varepsilon = (1 - \varepsilon)x^* + \varepsilon x$ is contained in C by convexity. However,

$$\begin{aligned} \|x_\varepsilon - y^*\|^2 &= \langle x^* - y^* + \varepsilon(x - x^*), x^* - y^* + \varepsilon(x - x^*) \rangle \\ &= \|x^* - y^*\|^2 + 2\varepsilon \underbrace{\langle x - x^*, x^* - y^* \rangle}_{<0 \text{ by } (1)} + \varepsilon^2 \|x - x^*\|^2. \end{aligned}$$

Hence $d(x_\varepsilon, y^*) < d(x^*, y^*)$ for $\varepsilon > 0$ small enough, contradicting $d(x^*, y^*) = d(C, K)$. □

Claim 4. $\langle x^*, \psi \rangle > \langle y^*, \psi \rangle$.

Proof. We have $\langle x^* - y^*, \psi \rangle = \|x^* - y^*\|^2 > 0$, where the last step uses the fact that $x^* \neq y^*$ by the disjointness of C and K. □

Claim 5. $\langle y^*, \psi \rangle \geqslant \sup \langle K, \psi \rangle$.

Proof. The proof is analogous to Claim 3. Specifically, suppose for the sake of contradiction that $\langle y, \psi \rangle > \langle y^*, \psi \rangle$ for some $y \in K$. This is equivalent to

$$\langle y^* - y, x^* - y^* \rangle < 0. \tag{2}$$

For $\varepsilon \in (0, 1)$, the point $y_\varepsilon = (1 - \varepsilon)y^* + \varepsilon y$ is contained in K by convexity. However,

$$\begin{aligned} \|x^* - y_\varepsilon\|^2 &= \langle x^* - y^* + \varepsilon(y^* - y), x^* - y^* + \varepsilon(y^* - y) \rangle \\ &= \|x^* - y^*\|^2 + 2\varepsilon \underbrace{\langle y^* - y, x^* - y^* \rangle}_{<0 \text{ by } (2)} + \varepsilon^2 \|y^* - y\|^2. \end{aligned}$$

Hence $d(x^*, y_\varepsilon) < d(x^*, y^*)$ for $\varepsilon > 0$ small enough, contradicting $d(x^*, y^*) = d(C, K)$. □

By Claims 3–5, the proof is complete. □

Nonstrict separation

The proofs below use the following property of compact sets $K \subset \mathbb{R}^n$: given any sequence $x_1, x_2, \ldots, x_n, \ldots \in K$, there is a subsequence $x_{i_1}, x_{i_2}, \ldots, x_{i_n}, \ldots$ and some $x^* \in K$ such that $x_{i_n} \to x^*$ as $n \to \infty$. In other words, every sequence in a compact set has a convergent subsequence. The *closure* of a set $A \subseteq \mathbb{R}^n$ is a superset of A defined by $\operatorname{cl} A = \{x \in \mathbb{R}^n : d(x, A) = 0\}$. Put differently, $\operatorname{cl} A$ is the smallest closed set that contains A. A point x is called an *interior* point of A if there exists $\varepsilon > 0$ such that $\{y \in \mathbb{R}^n : d(x, y) < \varepsilon\} \subseteq A$. The set of all interior points of A is denoted $\operatorname{int} A$.

Lemma 6. *Let $M \in \mathbb{R}^{n \times (n+1)}$ be given by*

$$M = \begin{bmatrix} 1 & 0 & 0 & & 0 & -1 \\ 0 & 1 & 0 & & 0 & -1 \\ 0 & 0 & 1 & & 0 & -1 \\ \vdots & \vdots & \vdots & \ddots & 0 & -1 \\ 0 & 0 & 0 & & 1 & -1 \end{bmatrix}.$$

Let $\{M_k\}$ be a sequence with $M_k \to M$. Then for some k, there exists a vector $\lambda \in (0, \infty)^{n+1}$ with $M_k \lambda = 0$.

Proof. Since the nullspace of every M_k is nonempty, we can fix a sequence $\{\lambda_k\}$ of unit vectors with $M_k \lambda_k = 0$. By passing to a subsequence if necessary, we may assume that $\lambda_k \to \lambda^*$. But then λ^* is a unit vector with $M\lambda^* = 0$, which forces

$$\lambda^* = \frac{\pm 1}{\sqrt{n+1}} \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}.$$

In particular, for all k large enough, the components of λ_k are either all positive or all negative, so that either λ_k or $-\lambda_k$ is the desired vector. □

Theorem 7 (Separating hyperplane theorem, nonstrict case)**.** *Let $X, Y \subseteq \mathbb{R}^n$ be nonempty convex subsets. If $X \cap Y = \varnothing$, then there exists a nonzero $\psi \in \mathbb{R}^n$ with*

$$\inf \langle X, \psi \rangle \geqslant \sup \langle Y, \psi \rangle.$$

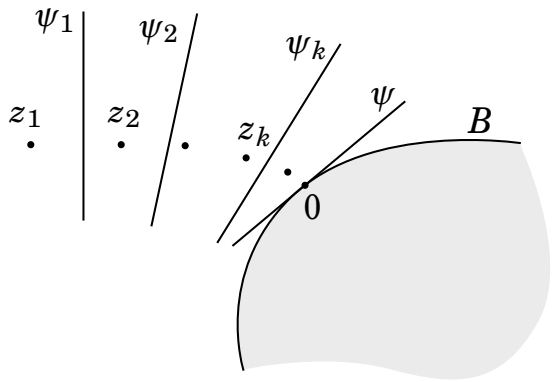


Figure 2: Separating 0 from B by a hyperplane.

Proof. Consider the convex set $A = X - Y = \{x - y : x \in X, y \in Y\}$. Then $0 \notin A$, and our objective is to find a nonzero $\psi \in \mathbb{R}^n$ with $\inf \langle A, \psi \rangle \geqslant 0$. Let $B = \operatorname{cl} A$ be the closure of A.

First of all, we claim that $0 \notin \operatorname{int} B$. For the sake of contradiction, suppose otherwise. Then for $\varepsilon > 0$ small enough, B contains the ball $\{v : \|v\|_\infty \leqslant \varepsilon\}$. In particular, B contains $\varepsilon v_1, \varepsilon v_2, \ldots, \varepsilon v_{n+1}$, where v_i is the ith column of the matrix M in Lemma 6. Recall that each v_i is the limit of a sequence in A. By Lemma 6, it follows that some $\tilde{v}_1, \tilde{v}_2, \ldots, \tilde{v}_{n+1} \in A$ obey $\sum \lambda_i v_i = 0$ for some positive coefficients $\lambda_1, \lambda_2, \ldots, \lambda_{n+1}$. Since A is convex, we conclude that $0 \in A$, a contradiction. Hence $0 \notin \operatorname{int} B$, as claimed.

The remainder of the proof is illustrated in Figure 2. By the claim just settled, we can fix a sequence of points $\{z_k\}$ outside of B with $z_k \to 0$. By the strict version of the separating hyperplane theorem, for each k there exists a unit vector ψ_k with

$$\inf \langle B, \psi_k \rangle > \langle z_k, \psi_k \rangle. \tag{3}$$

Passing to a subsequence if necessary, we may assume that $\psi_k \to \psi$ for some unit vector ψ. We now claim that $\inf \langle B, \psi \rangle \geqslant 0$. Indeed, for every $v \in B$,

$$\begin{aligned}
\langle v, \psi \rangle &= \lim_{k \to \infty} \langle v, \psi_k \rangle && \text{since } \psi_k \to \psi \\
&\geqslant \lim_{k \to \infty} \langle z_k, \psi_k \rangle && \text{by (3)} \\
&= 0 && \text{since } \|\psi_k\| = 1 \text{ and } z_k \to 0. \qquad \square
\end{aligned}$$

6 Normal Cones to Convex Sets

Definition (Normal Cone) Let $\Omega \subset \mathbb{R}^n$ be a convex set with $\bar{x} \in \Omega$. The *normal cone (https://en.wikipedia.org/wiki/Normal_(geometry))* to Ω at $\bar{x}$ is

$$N(\bar{x};\Omega) := \big\{x^* \in \mathbb{R}^n \big| \ \langle x^*, x - \bar{x}\rangle \leq 0 \ \text{ for all } \ x \in \Omega\big\}.$$

By convention, we let $N(\bar{x};\Omega) := \emptyset$ for $\bar{x} \notin \Omega$.

Definition A set $\Omega \subset \mathbb{R}^n$ is called a *cone* if $\lambda x \in \Omega$ whenever $x \in \Omega$ and $\lambda \geq 0$. If Ω is convex, then it is called a *convex cone (https://en.wikipedia.org/wiki/Convex_cone)*.

Proposition 6.1 Let $\Omega \subset \mathbb{R}^n$ be a convex set and let $\bar{x} \in \Omega$. Then $N(\bar{x};\Omega)$ is a closed convex cone.
Proof First we'll prove that $N(\bar{x};\Omega)$ is a cone. Fix $v \in N(\bar{x};\Omega)$ and let $\lambda \geq 0$. By the definition of normal cone, we have that $\langle v, x - \bar{x}\rangle \leq 0$ for any $x \in \Omega$. Then we have that $\lambda\langle v, x - \bar{x}\rangle = \langle \lambda v, x - \bar{x}\rangle \leq 0$ for any $x \in \Omega$. Thus we have that $\lambda v \in N(\bar{x};\Omega)$ whenever $v \in N(\bar{x};\Omega)$ and $\lambda \geq 0$.

Now we will show that the normal cone is convex. Let $v_1, v_2 \in N(\bar{x}; \Omega)$ and $0 \leq \lambda \leq 1$. Then we have that $\langle v_1, x - \bar{x} \rangle \leq 0$ and $\langle v_2, x - \bar{x} \rangle \leq 0$ for any $x \in \Omega$. Then if we take

$$\langle \lambda v_1 + (1 - \lambda) v_2, x - \bar{x} \rangle = \lambda \langle v_1, x - \bar{x} \rangle + (1 - \lambda) \langle v_2, x - \bar{x} \rangle \leq 0$$

for any $x \in \Omega$. Thus it follows that $\lambda v_1 + (1 - \lambda) v_2 \in N(\bar{x}; \Omega)$ and therefore the normal cone is convex.

We leave closure as an exercise. $\square$

Proposition 6.2 Let $\Omega \subset \mathbb{R}^n$ be convex and let $\bar{x} \in \operatorname{int} \Omega$. Then we have that $N(\bar{x}; \Omega) = \{0\}$.
Proof Because $\bar{x} \in \operatorname{int} \Omega$ then there exists a $\delta > 0$ such that the ball centered at $\bar{x}$ with radius δ is contained in Ω, so we have $\mathbb{B}(\bar{x}, \delta) \subset \Omega$.

Let $v \in N(\bar{x}; \Omega)$. Then we have that $\langle v, x - \bar{x} \rangle \leq 0$ for all $x \in \Omega$. Let $x \in \Omega$ and let $t > 0$ be small enough such that $\bar{x} + tx \in \mathbb{B}(\bar{x}, \delta)$. Then by the definition of normal cone we have that

$$\langle v, \bar{x} + tx - \bar{x} \rangle = t\langle v, x \rangle \leq 0 \text{for all} x \in \Omega.$$

Thus $\langle v, v \rangle = \|v\|^2 \leq 0$ which implies that $v = 0$. $\square$

Lemma Let C_1 and C_2 be nonempty convex sets. We have

$$\mathrm{ri}(C_1) \cap \mathrm{ri}(C_2) \subseteq \mathrm{ri}(C_1 \cap C_2),\ \overline{C_1 \cap C_2} \subseteq \overline{C_1} \cap \overline{C_2}.$$

Furthermore, if $\mathrm{ri}(C_1) \cap \mathrm{ri}(C_2) \neq \emptyset$, then

$$\mathrm{ri}(C_1) \cap \mathrm{ri}(C_2) = \mathrm{ri}(C_1 \cap C_2),\ \overline{C_1 \cap C_2} = \overline{C_1} \cap \overline{C_2}.$$

Proof. Let $x \in \mathrm{ri}(C_1) \cap \mathrm{ri}(C_2)$, $y \in C_1 \cap C_2$. By the prolongation lemma, the line segment connecting x and y can be prolonged beyond x without leaving C_1 and C_2. Hence, by the prolongation lemma again, $x \in \mathrm{ri}(C_1 \cap C_2)$. Since $C_1 \cap C_2 \subseteq \overline{C_1} \cap \overline{C_2}$, which is closed, we have $\overline{C_1 \cap C_2} \subseteq \overline{C_1} \cap \overline{C_2}$. Now suppose $\mathrm{ri}(C_1) \cap \mathrm{ri}(C_2) \neq \emptyset$ and let $x \in \mathrm{ri}(C_1) \cap \mathrm{ri}(C_2)$ and $y \in \overline{C_1} \cap \overline{C_2}$. Consider $\alpha_k \to 0$ and $y_k = \alpha_k x + (1 - \alpha_k)y$, then $y_k \to y$. By the line segment property, $y_k \in \mathrm{ri}(C_1) \cap \mathrm{ri}(C_2)$. Hence $y \in \overline{\mathrm{ri}(C_1) \cap \mathrm{ri}(C_2)}$. Then

$$\overline{C_1} \cap \overline{C_2} \subseteq \overline{\mathrm{ri}(C_1) \cap \mathrm{ri}(C_2)} \subseteq \overline{C_1 \cap C_2}.$$

Hence $\overline{C_1 \cap C_2} = \overline{C_1} \cap \overline{C_2}$. Moreover, the closure of $\mathrm{ri}(C_1) \cap \mathrm{ri}(C_2)$ and $C_1 \cap C_2$ are the same. Hence, they have the same relative interior. Then

$$\mathrm{ri}(C_1 \cap C_2) = \mathrm{ri}(\mathrm{ri}(C_1) \cap \mathrm{ri}(C_2)) \subseteq \mathrm{ri}(C_1) \cap \mathrm{ri}(C_2).$$

□

Lemma If $f : \mathbb{R}^n \to (-\infty, \infty]$ is a convex function then we have that

$$\mathrm{ri}(\mathrm{epi}\ f) = \{(x, \lambda) \in \mathbb{R}^n \times \mathbb{R} :\ x \in \mathrm{ri}(\mathrm{dom}\ f),\ \lambda > f(x)\}.$$

Proof. Let P be the projection on the x component, i.e. $P(x, \lambda) = x$. Then $P(\mathrm{epi}\ f) = \mathrm{dom}\ f$. Since P is linear, by previous proposition, $P(\mathrm{ri}(\mathrm{epi}\ f)) = \mathrm{ri}\ P(\mathrm{epi}\ f) = \mathrm{ri}(\mathrm{dom}\ f)$. Let $F_x := \{(x, \lambda) : \lambda \in \mathbb{R}\}$. Then

$$\mathrm{ri}(\mathrm{epi}\ f) = \bigcup_{x \in \mathrm{ri}(\mathrm{dom}\ f)} \big(F_x \cap \mathrm{ri}(\mathrm{epi}\ f)\big).$$

Note that $\mathrm{ri}\ F_x = F_x$ and $F_x \cap \mathrm{ri}(\mathrm{epi}\ f) \neq \emptyset$ for $x \in \mathrm{ri}(\mathrm{dom}\ f)$, by the above lemma, we have

$$F_x \cap \mathrm{ri}(\mathrm{epi}\ f) = \mathrm{ri}\ F_x \cap \mathrm{ri}(\mathrm{epi}\ f) = \mathrm{ri}(F_x \cap \mathrm{epi}\ f) = \mathrm{ri}\ (\mathrm{epi}\ f)_x,$$

where $(\mathrm{epi}\ f)_x := \{\lambda : (x, \lambda) \in \mathrm{epi}\ f\} = \{\lambda : \lambda \geq f(x)\}$. Hence

$$\mathrm{ri}(\mathrm{epi}\ f) = \bigcup_{x \in \mathrm{ri}(\mathrm{dom}\ f)} \{(x, \lambda) : \lambda \in \mathrm{ri}\ (\mathrm{epi}\ f)_x\}.$$

One can easily observe that the relative interior of the set $(\mathrm{epi}\ f)_x$ is $\{\lambda : \lambda > f(x)\}$. Hence

$$\mathrm{ri}(\mathrm{epi}\ f) = \{(x, \lambda) \in \mathbb{R}^n \times \mathbb{R} :\ x \in \mathrm{ri}(\mathrm{dom}\ f),\ \lambda > f(x)\}.$$

□

Theorem 6.3 Let $\Omega_1, \Omega_2 \subset \mathbb{R}^n$ be convex sets satisfying the relative interior condition

$$\mathrm{ri}(\Omega_1) \cap \mathrm{ri}(\Omega_2) \neq \emptyset,$$

If $\bar{x} \in \Omega_1 \cap \Omega_2$, then we have the intersection rule

$$N\Big(\bar{x}; \Omega_1 \cap \Omega_2)\Big) = N(\bar{x}; \Omega_1) + N(\bar{x}; \Omega_2).$$

Proof First we will prove that $N(\bar{x}; \Omega_1) + N(\bar{x}; \Omega_2) \subset N(\bar{x}; \Omega_1 \cap \Omega_2)$. Fix $v \in N(\bar{x}; \Omega_1) + N(\bar{x}; \Omega_2)$. Then there exists $v_1 \in N(\bar{x}; \Omega_1)$ and $v_2 \in N(\bar{x}; \Omega_2)$ such that $v = v_1 + v_2$. Let $x \in \Omega_1 \cap \Omega_2$ be arbitrary. Then we have that

$$\langle v, x - \bar{x} \rangle = \langle v_1 + v_2, x - \bar{x} \rangle = \langle v_1, x - \bar{x} \rangle + \langle v_2, x - \bar{x} \rangle.$$

However, since $v_1 \in N(\bar{x}; \Omega_1)$ and $x \in \Omega_1 \cap \Omega_2 \subset \Omega_1$, we have that $\langle v_1, x - \bar{x} \rangle \leq 0$ and similarly we have that $\langle v_2, x - \bar{x} \rangle \leq 0$. Thus we have that $\langle v, x - \bar{x} \rangle \leq 0$ and since this is true for any $x \in \Omega_1 \cap \Omega_2$ then we have $v \in N(\bar{x}; \Omega_1 \cap \Omega_2)$.

Now we will show that $N(\bar{x}; \Omega_1 \cap \Omega_2) \subset N(\bar{x}; \Omega_1) + N(\bar{x}; \Omega_2)$. To do this recall the following two facts.

If $f : \mathbb{R}^n \to (-\infty, \infty]$ is a convex function then we have that

$$\operatorname{ri}(\operatorname{epi} f) = \{(x, \lambda) \in \mathbb{R}^n \times \mathbb{R} \colon x \in \operatorname{ri}(\operatorname{dom} f), \lambda > 0\}.$$

Also recall that two convex sets can be properly separated if and only if $\operatorname{ri}(\Omega_1) \cap \operatorname{ri}(\Omega_2) = \emptyset$. We will now begin the proof of the second inclusion.

Fixing $\bar{x} \in \Omega_1 \cap \Omega_2$ and $v \in N(\bar{x}; \Omega_1 \cap \Omega_2)$, we get by the normal cone definition that

$$\langle v, x - \bar{x} \rangle \leq 0 \ \text{ for all } \ x \in \Omega_1 \cap \Omega_2.$$

Denote the sets

$$\Theta_1 := \Omega_1 \times [0, \infty)$$

and

$$\Theta_2 := \{(x, \lambda) \mid x \in \Omega_2, \ \lambda \leq \langle v, x - \bar{x} \rangle\}.$$

It follows that $\mathrm{ri}(\Theta_1) = \mathrm{ri}(\Omega_1) \times (0, \infty)$ and

$$\mathrm{ri}(\Theta_2) = \big\{(x, \lambda) \;\big|\; x \in \mathrm{ri}(\Omega_2), \ \lambda < \langle v, x - \bar{x} \rangle\big\}.$$

We will now show by contradiction, that $\mathrm{ri}(\Theta_1) \cap \mathrm{ri}(\Theta_2) = \emptyset$. Suppose there exists an $(x, \lambda) \in \mathrm{ri}(\Theta_1) \cap \mathrm{ri}(\Theta_2)$. Then we have that $x \in \mathrm{ri}(\Omega_1)$ and $0 < \lambda$, and we also have that $x \in \mathrm{ri}(\Omega_2)$ and $\lambda < \langle v, x - \bar{x} \rangle$. But then we have

$x \in \Omega_1 \cap \Omega_2$ and $0 < \lambda < \langle v, x - \bar{x} \rangle \leq 0$ which is a contradiction. Therefore $\text{ri}(\Theta_1) \cap \text{ri}(\Theta_2) = \emptyset$.

Then applying the Proper Separation Theorem gives us $(w, \gamma) \in \mathbb{R}^n \times \mathbb{R}$ such that

$$\langle w, x \rangle + \lambda_1 \gamma \leq \langle w, y \rangle + \lambda_2 \gamma \ \text{ for all } \ (x, \lambda_1) \in \Theta_1, \ (y, \lambda_2) \in \Theta_2.$$

Moreover, there are $(\tilde{x}, \tilde{\lambda}_1) \in \Theta_1$ and $(\tilde{y}, \tilde{\lambda}_2) \in \Theta_2$ satisfying

$$\langle w, \tilde{x} \rangle + \tilde{\lambda}_1 \gamma < \langle w, \tilde{y} \rangle + \tilde{\lambda}_2 \gamma.$$

Observe that $\gamma \leq 0$. Notice that $(\bar{x}, 1) \in \Theta_1$ and $(\bar{x}, 0) \in \Theta_2$ and using the inequality above we have that

$$\langle w, \bar{x} \rangle + \gamma(1) \leq \langle w, \bar{x} \rangle + \gamma(0).$$

Solving for γ we get that $\gamma \leq 0$.

Let us now show that $\gamma < 0$. Again arguing by contradiction, suppose that $\gamma = 0$. Then we get $\langle w, x\rangle \leq \langle w, y\rangle$ for all $x \in \Omega_1,\ y \in \Omega_2$ and $\langle w, \tilde{x}\rangle < \langle w, \tilde{y}\rangle$ with $\tilde{x} \in \Omega_1,\ \tilde{y} \in \Omega_2$. This means that the sets Ω_1 and Ω_2 can be properly separated, which in turn implies that $\mathrm{ri}(\Omega_1) \cap \mathrm{ri}(\Omega_2) = \emptyset$. This gives us a contradiction, thus $\gamma < 0$.

To proceed further, notice that $(x, 0) \in \Theta_1$ for any $x \in \Omega_1$ and and that $(\bar{x}, 0) \in \Theta_2$. Thus we have the inequality

$$\langle w, x\rangle \leq \langle w, \bar{x}\rangle \ \text{ for all } \ x \in \Omega_1.$$

This gives us that $\langle w, x - \bar{x}\rangle$ for all $x \in \Omega_1$ and therefore $w \in N(\bar{x}; \Omega_1)$. Because $N(\bar{x}; \Omega_1)$ is a cone and $-\gamma > 0$ we have that $\dfrac{w}{-\gamma} \in N(\bar{x}; \Omega_1)$.

We also get that since $(\bar{x}, 0) \in \Theta_1$ and $(y, \lambda) \in \Theta_2$ for all $y \in \Omega_2$ and $\lambda = \langle v, y - \bar{x}\rangle$, we have

$$\langle w, \bar{x} \rangle \leq \langle w, y \rangle + \gamma \langle v, y - \bar{x} \rangle \ \text{ whenever } \ y \in \Omega_2.$$

Dividing both sides by γ, we get the inequality

$$\left\langle \frac{w}{\gamma} + v, y - \bar{x} \right\rangle \leq 0 \ \text{ for all } \ y \in \Omega_2,$$

and thus $\dfrac{w}{\gamma} + v \in N(\bar{x}; \Omega_2)$. This gives us

$$v \in \frac{w}{-\gamma} + N(\bar{x}; \Omega_2).$$

However as we have shown $\dfrac{w}{-\gamma} \in N(\bar{x}; \Omega_1)$ and $v + \dfrac{w}{\gamma} \in N(\bar{x}; \Omega_2)$, thus we have that $v = \left(v + \dfrac{w}{\gamma}\right) + \dfrac{w}{-\gamma} \in N(\bar{x}; \Omega_1) + N(\bar{x}; \Omega_2)$.

completing the proof. $\square$

7 Subgradients of Convex Functions

Definition Let $f\colon \mathbb{R}^n \to \overline{\mathbb{R}}$ be a convex function with $\bar{x} \in \operatorname{dom}(f)$. An element $x^* \in \mathbb{R}^n$ is called a *subgradient* of f at $\bar{x}$ if

$$\langle x^*, x - \bar{x} \rangle \leq f(x) - f(\bar{x}) \ \text{ for all } \ x \in \mathbb{R}^n.$$

The collection of all the subgradients of f at $\bar{x}$ is called the *subdifferential (https://en.wikipedia.org/wiki/Subderivative#The_subgradient)* of the function at $\bar{x}$ and is denoted by $\partial f(\bar{x})$.

Definition A function $f : \mathbb{R}^n \to \mathbb{R}$ is said to be *differentiable (https://en.wikipedia.org/wiki/Differentiable_function)* at $\bar{x}$ if there exists a $v \in \mathbb{R}^n$ such that

$$\lim_{x \to \bar{x}} \frac{f(x) - f(\bar{x}) - \langle v, x - \bar{x} \rangle}{\|x - \bar{x}\|} = 0.$$

Any such v is called the *gradient* of f and is denoted by ∇f.

Proposition 7.1 Let $f : \mathbb{R}^n \to \mathbb{R}$ be convex, with f differentiable at $\bar{x}$. Then we have that

$$\langle \nabla f(\bar{x}), x - \bar{x} \rangle \leq f(x) - f(\bar{x}) \text{ for all } x \in \mathbb{R}^n.$$

Proof Let $0 < t < 1$. Using the fact that f is convex, we have that

$$\frac{f\big(\bar{x}+t(x-\bar{x})\big)-f(\bar{x})-\langle\nabla f(\bar{x}),t(x-\bar{x})\rangle}{\|t(x-\bar{x})\|}$$

$$=\frac{f\big((1-t)\bar{x}+tx)\big)-f(\bar{x})-\langle\nabla f(\bar{x}),t(x-\bar{x})\rangle}{\|t(x-\bar{x})\|}$$

$$\leq\frac{(1-t)f(\bar{x})+tf(x)-f(\bar{x})-\langle\nabla f(\bar{x}),t(x-\bar{x})\rangle}{\|t(x-\bar{x})\|}$$

$$=\frac{t\Big(f(x)-f(\bar{x})-\langle\nabla f(\bar{x}),(x-\bar{x})\rangle\Big)}{t\|(x-\bar{x})\|}$$

$$=\frac{\Big(f(x)-f(\bar{x})-\langle\nabla f(\bar{x}),(x-\bar{x})\rangle\Big)}{\|(x-\bar{x})\|}.$$

Since this is true for all $0 < t < 1$, taking the limit as t approaches 0 we get, by the definition of differentiability that

$$\lim_{t \to 0} \frac{f\big(\bar{x} + t(x - \bar{x})\big) - f(\bar{x}) - \langle \nabla f(\bar{x}), t(x - \bar{x}) \rangle}{\|t(x - \bar{x})\|}$$

$$= 0 \leq \frac{\Big(f(x) - f(\bar{x}) - \langle \nabla f(\bar{x}), (x - \bar{x}) \rangle\Big)}{\|(x - \bar{x})\|}.$$

Thus we have that $\langle \nabla f(\bar{x}), x - \bar{x} \rangle \leq f(x) - f(\bar{x})$ as desired. $\square$

Lemma 7.2 Let $\delta > 0$ and $\epsilon > 0$. Suppose that $\langle v, h \rangle \leq \epsilon \|h\|$ whenever $\|h\| < \delta$. Then $\|v\| \leq \epsilon$.

Proof Let $h = \dfrac{\delta}{2} \dfrac{v}{\|v\|}$. Then we have that

$$\left\langle v, \frac{\delta}{2}\frac{v}{\|v\|}\right\rangle \leq \epsilon\frac{\delta}{2}.$$

This implies that

$$\frac{\langle v, v\rangle}{\|v\|} \leq \epsilon.$$

Thus we have that $\|v\| \leq \epsilon$. $\square$

Proposition 7.3 Let $f : \mathbb{R}^n \to \mathbb{R}$ be a convex function. If f is differentiable at $\bar{x} \in \mathbb{R}^n$ then we have that $\partial f(\bar{x}) = \{\nabla f(\bar{x})\}$.
Proof By Proposition 7.1 we have that $\nabla f(\bar{x}) \in \partial f(\bar{x})$, so we only need to show the opposite conclusion.

Let $v \in \partial f(\bar{x})$, then $\langle v, x - \bar{x}\rangle \leq f(x) - f(\bar{x})$. Since f is differentiable at $\bar{x}$, we have that

$$\lim_{x \to \bar{x}} \frac{f(x) - f(\bar{x}) - \langle \nabla f(\bar{x}), x - \bar{x} \rangle}{\|x - \bar{x}\|} = 0.$$

Thus for any $\epsilon > 0$ there exists a $\delta > 0$ such that

$$\left| \frac{f(x) - f(\bar{x}) - \langle \nabla f(\bar{x}), x - \bar{x} \rangle}{\|x - \bar{x}\|} \right| < \epsilon \text{ whenever } \|x - \bar{x}\| < \delta.$$

Multiplying both sides by $\|x - \bar{x}\|$ and then adding both sides by $\langle \nabla f(\bar{x}), x - \bar{x} \rangle$ we get

$$f(x) - f(\bar{x}) < \langle \nabla f(\bar{x}), x - \bar{x} \rangle - \epsilon \|x - \bar{x}\| \text{ whenever } \|x - \bar{x}\| < \delta.$$

But then using the inequality we obtained above, we have that

$$\langle v, x - \bar{x} \rangle \leq f(x) - f(\bar{x}) < \langle \nabla f(\bar{x}), x - \bar{x} \rangle - \epsilon \|x - \bar{x}\| \text{ whenever } \|x - \bar{x}\| < \delta.$$

Rearranging we get

$$\langle v - \nabla f(\bar{x}), x - \bar{x} \rangle < \epsilon \|x - \bar{x}\| \text{ whenever } \|x - \bar{x}\| < \delta.$$

By Lemma 7.2 we get that $\|v - \nabla f(\bar{x})\| < \epsilon$. Since this is true for any $\epsilon > 0$ we have that $\|v - \nabla f(\bar{x})\| = 0$. Thus we arrive at the conclusion that $v = \nabla f(\bar{x})$, and since this is true for all $v \in \partial f(\bar{x})$ then the conclusion holds.
□

8 The Subdifferential Sum Rule

Proposition 8.1 Let $f : \mathbb{R}^n \to (-\infty, \infty]$ be a convex function, and let $\bar{x} \in \operatorname{dom}(f)$. Then we have

$$\partial f(\bar{x}) = \{v \in \mathbb{R}^n \mid (v, -1) \in N((\bar{x}, f(\bar{x})); \operatorname{epi}(f))\}.$$

Proof Fix any subgradient $v \in \partial f(\bar{x})$ and then get from the definition of the subdifferential that

$$\langle v, x - \bar{x} \rangle \leq f(x) - f(\bar{x}) \text{ for all } \ x \in \mathbb{R}^n.$$

To show that $(v, -1) \in N((\bar{x}, f(\bar{x})); \operatorname{epi}(f))$, fix any $(x, \lambda) \in \operatorname{epi}(f)$ and observe that due to $\lambda \geq f(x)$ we have the relationships

$$\Big\langle (v, -1), (x, \lambda) - \big(\bar{x}, f(\bar{x})\big) \Big\rangle = \langle v, x - \bar{x} \rangle + (-1)(\lambda - f(\bar{x}))$$

$$= \langle v, x - \bar{x} \rangle - (\lambda - f(\bar{x})) \leq \langle v, x - \bar{x} \rangle - (f(x) - f(\bar{x})) \leq 0,$$

To verify the opposite inclusion, take $(v, -1) \in N((\bar{x}, f(\bar{x})); \mathrm{epi}(f))$ and fix any $x \in \mathrm{dom}(f)$. Then $(x, f(x)) \in \mathrm{epi}(f)$ and hence

$$\Big\langle (v, -1), (x, f(x)) - (\bar{x}, f(\bar{x})) \Big\rangle \leq 0,$$

which in turn implies the inequality

$$\langle v, x - \bar{x} \rangle - (f(x) - f(\bar{x})) \leq 0.$$

Thus $v \in \partial f(\bar{x})$, which completes the proof of the proposition. $\square$

Proposition 2.11 **(i)** *Let Ω_1 and Ω_2 be nonempty, convex subsets of $\mathbb{R}^n$ and $\mathbb{R}^p$, respectively. For $(\bar{x}_1, \bar{x}_2) \in \Omega_1 \times \Omega_2$, we have*

$$N\big((\bar{x}_1, \bar{x}_2); \Omega_1 \times \Omega_2\big) = N(\bar{x}_1; \Omega_1) \times N(\bar{x}_2; \Omega_2).$$

(ii) *Let Ω_1 and Ω_2 be convex subsets of $\mathbb{R}^n$ with $\bar{x}_i \in \Omega_i$ for $i = 1, 2$. Then*

$$N(\bar{x}_1 + \bar{x}_2; \Omega_1 + \Omega_2) = N(\bar{x}_1; \Omega_1) \cap N(\bar{x}_2; \Omega_2).$$

Proof. To verify **(i)**, fix $(v_1, v_2) \in N((\bar{x}_1, \bar{x}_2); \Omega_1 \times \Omega_2)$ and get by the definition that

$$\langle (v_1, v_2), (x_1, x_2) - (\bar{x}_1, \bar{x}_2) \rangle = \langle v_1, x_1 - \bar{x}_1 \rangle + \langle v_2, x_2 - \bar{x}_2 \rangle \le 0 \tag{2.5}$$

whenever $(x_1, x_2) \in \Omega_1 \times \Omega_2$. Putting $x_2 := \bar{x}_2$ in (2.5) gives us

$$\langle v_1, x_1 - \bar{x}_1 \rangle \le 0 \ \text{ for all } \ x_1 \in \Omega_1,$$

which means that $v_1 \in N(\bar{x}_1; \Omega_1)$. Similarly, we obtain $v_2 \in N(\bar{x}_2; \Omega_2)$ and thus justify the inclusion "$\subset$" in (2.5). The opposite inclusion is obvious.

Let us now verify **(ii)**. Fix $v \in N(\bar{x}_1 + \bar{x}_2; \Omega_1 + \Omega_2)$ and get by the definition that

$$\langle v, x_1 + x_2 - (\bar{x}_1 + \bar{x}_2) \rangle \le 0 \ \text{ whenever } \ x_1 \in \Omega_1, \ x_2 \in \Omega_2.$$

Putting there $x_1 := \bar{x}_1$ and $x_2 := \bar{x}_2$ gives us $v \in N(\bar{x}_1; \Omega_1) \cap N(\bar{x}_2; \Omega_2)$. The opposite inclusion in **(ii)** is also straightforward. □

Now we are ready to deduce the following subdifferential sum rule for convex functions from the intersection rule of Theorem 6.3 for normal cones.

Theorem 8.2 Let $f_1, f_2 \colon \mathbb{R}^n \to (-\infty, \infty]$, be extended-real-valued convex functions satisfying the relative interior qualification condition

$$\operatorname{ri}(\operatorname{dom} f) \cap \operatorname{ri}(\operatorname{dom} f) \neq \emptyset.$$

Then for all $\bar{x} \in \operatorname{dom}(f_1) \cap \operatorname{dom}(f_2)$ we have the sum rule

$$\partial\Big(f_1 + f_2\Big)(\bar{x}) = \partial f_1(\bar{x}) + \partial f_2(\bar{x}).$$

Proof Observing that the inclusion "$\supset$" above directly follows from the subdifferential definition, we proceed with the proof of the opposite inclusion. Pick any $v \in \partial(f_1 + f_2)(\bar{x})$. Then we have

$$\langle v, x - \bar{x} \rangle \leq (f_1 + f_2)(x) - (f_1 + f_2)(\bar{x}) \text{ for all } \ x \in \mathbb{R}^n.$$

Define the following convex subsets of $\mathbb{R}^{n+2}$ by

$$\Omega_1 := \{(x, \lambda_1, \lambda_2) \in \mathbb{R}^n \times \mathbb{R} \times \mathbb{R} \mid \lambda_1 \geq f_1(x)\} = \text{epi}(f_1) \times \mathbb{R},$$

$$\Omega_2 := \{(x, \lambda_1, \lambda_2) \in \mathbb{R}^n \times \mathbb{R} \times \mathbb{R} \mid \lambda_2 \geq f_2(x)\}.$$

We can easily verify by the normal cone definition that

$$(v, -1, -1) \in N((\bar{x}, f_1(\bar{x}), f_2(\bar{x})); \Omega_1 \cap \Omega_2).$$

To apply Theorem 6.3 (Normal Cone Intersection Rule) to these sets, let us check that $\text{ri}(\Omega_1) \cap \text{ri}(\Omega_2) \neq \emptyset$. Indeed, we get

$$\begin{aligned}
\text{ri}(\Omega_1) &= \{(x, \lambda_1, \lambda_2) \in \mathbb{R}^n \times \mathbb{R} \times \mathbb{R} \mid x \in \text{ri}(\text{dom}(f_1)),\ \lambda_1 > f_1(x)\} \\
&= \text{ri}(\text{epi}(f_1) \times \mathbb{R}, \\
\text{ri}(\Omega_2) &= \{(x, \lambda_1, \lambda_2) \in \mathbb{R}^n \times \mathbb{R} \times \mathbb{R} \mid x \in \text{ri}(\text{dom}(f_2)),\ \lambda_2 > f_2(x)\}
\end{aligned}$$

by Proposition 8.1. Then choosing $z \in \operatorname{ri}(\operatorname{dom}(f_1)) \cap \operatorname{ri}(\operatorname{dom}(f_2))$, it is not hard to see that

$$(z, f_1(z) + 1, f_2(z) + 1) \in \operatorname{ri}(\Omega_1) \cap \operatorname{ri}(\Omega_2) \neq \emptyset.$$

Applying now Theorem 6.3 to the above set intersection gives us

$$N((\bar{x}, f_1(\bar{x}), f_2(\bar{x})); \Omega_1 \cap \Omega_2) = N((\bar{x}, f_1(\bar{x}), f_2(\bar{x})); \Omega_1) + N((\bar{x}, f_1(\bar{x}), f_2(\bar{x})); \Omega_2).$$

It follows from how we defined the sets Ω_1 and Ω_2 that

$$(v, -1, -1) = (v_1, -\gamma_1, 0) + (v_2, 0, -\gamma_2)$$

with $(v_1, -\gamma_1) \in N((\bar{x}, f_1(\bar{x})); \operatorname{epi}(f_1))$ and $(v_2, -\gamma_2) \in N((\bar{x}, f_2(\bar{x})); \operatorname{epi}(f_2))$. Thus

$$v = v_1 + v_2, \quad \gamma_1 = \gamma_2 = 1,$$

and we have by Proposition 8.1 that $v_1 \in \partial f_1(\bar{x})$ and $v_2 \in \partial f_2(\bar{x})$. This ensures the inclusion $\partial(f_1 + f_2)(\bar{x}) \subset \partial f_1(\bar{x}) + \partial f_2(\bar{x})$ and hence completes the proof. $\square$

9 The Subdifferential Chain Rule

Lemma 9.1 Let $B : \mathbb{R}^n \to \mathbb{R}^p$ be an affine mapping given by $B(x) := Ax + b$, where A is a $p \times n$ matrix and $b \in \mathbb{R}^p$. Then for any $(\bar{x}, \bar{y}) \in \operatorname{gph}(B)$ we have

$$N\big((\bar{x}, \bar{y}); \operatorname{gph}(B)\big) = \big\{(u, v) \in \mathbb{R}^n \times \mathbb{R}^p \;\big|\; u = -A^\top v\big\}.$$

Proof It is clear that $\operatorname{gph}(B)$ is convex and $(u, v) \in N((\bar{x}, \bar{y}) \operatorname{gph}(B))$ if and only if

$$\langle u, x - \bar{x}\rangle + \langle v, B(x) - B(\bar{x})\rangle \le 0 \;\text{ for all } x \in \mathbb{R}^n.$$

It follows directly from the definitions that

$$\begin{aligned}\langle u, x - \bar{x}\rangle + \langle v, B(x) - B(\bar{x})\rangle &= \langle u, x - \bar{x}\rangle + \langle v, A(x) - A(\bar{x})\rangle \\ &= \langle u, x - \bar{x}\rangle + \langle A^\top v, x - \bar{x}\rangle = \langle u + A^\top v, x - \bar{x}\rangle.\end{aligned}$$

This implies that $(u, v) \in N\big((\bar{x}, \bar{y}); \mathrm{gph}(B)\big)$ is equivalent to $\langle u + A^\top v, x - \bar{x}\rangle \leq 0$ for all $x \in \mathbb{R}^n$, and so we have $u = -A^\top v$. $\square$

Proposition 9.2 Let $f : \mathbb{R}^p \to (-\infty, \infty]$ be a convex function, and let $B : \mathbb{R}^n \to \mathbb{R}^p$ be given by $B(x) := Ax + b$, where A is a $p \times n$ matrix and $b \in \mathbb{R}^p$. Fix $\bar{x} \in \mathbb{R}^n$ such that $B(\bar{x}) \in \mathrm{dom}(f)$. Denote $\bar{y} := B(\bar{x})$. Then we have that

$$\partial(f \circ B)(\bar{x}) \supset A^\top\big(\partial f(\bar{y})\big) = \big\{A^\top v \;\big|\; v \in \partial f(\bar{y})\big\}.$$

Proof Fix $u \in A^\top \partial f(\bar{y})$. By definition we have that there exists a $v \in \partial f(\bar{y})$ such that $u = A^\top v$. Then we have that

$$\langle u, x - \bar{x}\rangle = \langle A^\top v, x - \bar{x}\rangle$$

$$= \langle v, Ax - A\bar{x}\rangle = \langle v, B(x) - B(\bar{x})\rangle,$$

and because $v \in \partial f(\bar{y})$ and $\bar{y} = B(\bar{x})$ then we have that

$$\langle u, x - \bar{x} \rangle \leq f\big(B(x)\big) - f\big(B(\bar{x})\big) = (f \circ B)(x) - (f \circ B)(\bar{x}).$$

Thus we have that $u \in \partial(f \circ B)(\bar{x})$. $\square$

We will now prove the main result of this section.

Theorem 9.3 Let $f : \mathbb{R}^p \to (-\infty, \infty]$ be a convex function, and let $B : \mathbb{R}^n \to \mathbb{R}^p$ be given by $B(x) := Ax + b$, where A is a $p \times n$ matrix and $b \in \mathbb{R}^p$. Fix $\bar{x} \in \mathbb{R}^n$ such that $B(\bar{x}) \in \mathrm{dom}(f)$. Denote $\bar{y} := B(\bar{x})$ and assume that the range of B contains a point of $\mathrm{ri}(\mathrm{dom}(f))$. Then we have that

$$\partial(f \circ B)(\bar{x}) = A^\top \big(\partial f(\bar{y})\big) = \big\{A^\top v \;\big|\; v \in \partial f(\bar{y})\big\}.$$

Proof The first inclusion was proved in Proposition 9.2, so we only need to prove that

$$\partial(f \circ B)(\bar{x}) \subset A^\top \big(\partial f(\bar{y})\big).$$

Fix $v \in \partial(f \circ B)(\bar{x})$ and form the subsets of $\mathbb{R}^n \times \mathbb{R}^p \times \mathbb{R}$ by

$$\Omega_1 := \operatorname{gph}(B) \times \mathbb{R} \;\text{ and }\; \Omega_2 := \mathbb{R}^n \times \operatorname{epi}(f).$$

Then we clearly get the relationships

$$\operatorname{ri}(\Omega_1) = \Omega_1 = \operatorname{gph}(B) \times \mathbb{R},$$

$$\operatorname{ri}(\Omega_2) = \{(x, y, \lambda) \mid x \in \mathbb{R}^n,\ y \in \operatorname{ri}(\operatorname{dom}(f)),\ \lambda > f(y)\},$$

and thus the assumption of the theorem tells us that $\operatorname{ri}(\Omega_1) \cap \operatorname{ri}(\Omega_2) \neq \emptyset$.

Further, it follows from the definitions of the subdifferential and of the normal cone that $(v, 0, -1) \in N((\bar{x}, \bar{y}, \bar{z}); \Omega_1 \cap \Omega_2)$, where $\bar{z} := f(\bar{y})$. Indeed, for any $(x, y, \lambda) \in \Omega_1 \cap \Omega_2$ we have $y = B(x)$ and $\lambda \geq f(y)$, and so $\lambda \geq f(B(x))$. Thus

$$\langle v, x - \bar{x}\rangle + 0(y - \bar{y}) + (-1)(\lambda - \bar{z}) \leq \langle v, x - \bar{x}\rangle - [f(B(x)) - f(B(\bar{x}))] \leq 0.$$

Employing the intersection rule of Theorem 6.3 to the above sets gives us

$$(v, 0, -1) \in N\big((\bar{x}, \bar{y}, \bar{z}); \Omega_1\big) + N\big((\bar{x}, \bar{y}, \bar{z}); \Omega_2\big),$$

which tells us that $(v, 0, -1) = (v, -w, 0) + (0, w, -1)$ with $(v, -w) \in N((\bar{x}, \bar{y}); \operatorname{gph}(B))$ and $(w, -1) \in N((\bar{y}, \bar{z}); \operatorname{epi}(f))$. Then we get

$$v = A^\top w \;\text{ and }\; w \in \partial f(\bar{y}),$$

which implies that $v \in A^\top(\partial f(\bar{y}))$ and hence verifies the inclusion "$\subset$". $\square$

10 The Subdifferential Maximum Rule

Let $f_i \colon \mathbb{R}^n \to (-\infty, \infty]$ with $i = 1, ..., m$ be a collection of convex functions. Define $f(x) = \max\{f_i(x); i = 1, ..., m\}$.

Given $\bar{x} \in \mathbb{R}^n$, we define the *active index set* of $x \in \mathbb{R}^n$ by $I(x) = \{i = 1, ..., m; f_i(x) = f(x)\}$.

In this section we will ultimately prove that if f_i is continuous at $\bar{x}$ for each $i = 1, ..., m$, then $\partial f(\bar{x}) = \text{co}\{\bigcup_{i \in I(\bar{x})} \partial f_i(\bar{x})\}$. In the next proposition we will prove the first inclusion.

Proposition 10.1 For any $\bar{x} \in \text{dom}(f)$ we have that

$$\partial f(\bar{x}) \supset \text{co}\{\bigcup_{i \in I(\bar{x})} \partial f_i(\bar{x})\}.$$

Proof First, notice that $\partial f(\bar{x})$ is a convex set. Fix any $i \in I(\bar{x})$ and $v \in \partial f_i(\bar{x})$. Then

$$\langle v, x - \bar{x} \rangle \leq f_i(x) - f_i(\bar{x}) \text{ for all } x \in \mathbb{R}^n.$$

Since $i \in I(\bar{x})$, then $f_i(\bar{x}) = f(\bar{x})$ and furthermore $f_i(x) \leq f(x)$ for all $x \in \mathbb{R}^n$. It follows that

$$\langle v, x - \bar{x} \rangle \leq f_i(x) - f_i(\bar{x}) \leq f(x) - f(\bar{x}) \text{ for all } x \in \mathbb{R}^n.$$

Thus $\bigcup_{i \in I(\bar{x})} \partial f_i(\bar{x}) \subset \partial f(\bar{x})$ and by the convexity of $\partial f(\bar{x})$ we have that

$$\text{co}\Big\{ \bigcup_{i \in I(\bar{x})} \partial f_i(\bar{x}) \Big\} \subset \partial f(\bar{x}). \ \square$$

Lemma 10.2
$(\mathbf{i})$ Let Ω be a convex set in $\mathbb{R}^n$. Then $\text{int}\,(\Omega) = \text{ri}(\Omega)$ provided that $\text{int}\,(\Omega) \neq \emptyset$. Moreover, we have that $N(\bar{x}; \Omega) = \{0\}$ whenever $\bar{x} \in \text{int}(\Omega)$.

(**ii**) Let $f\colon \mathbb{R}^n \to (-\infty,\infty]$ be a convex function, which is continuous at $\bar{x} \in \operatorname{dom}(f)$. Then we have $\bar{x} \in \operatorname{int}(\operatorname{dom}(f))$ with the implication

$$(v,-\lambda) \in N((\bar{x},f(\bar{x}));\operatorname{epi}(f)) \Longrightarrow [\lambda \geq 0 \ \text{ and } \ v \in \lambda \partial f(\bar{x})].$$

Proof (**i**) Suppose that $\operatorname{int}(\Omega) \neq \emptyset$ and check that $\operatorname{aff}(\Omega) = \mathbb{R}^n$. Indeed, picking $\bar{x} \in \operatorname{int}(\Omega)$ and fixing $x \in \mathbb{R}^n$, find $t > 0$ with $tx + (1-t)\bar{x} = \bar{x} + t(x - \bar{x}) \in \operatorname{int}(\Omega) \subset \operatorname{aff}(\Omega)$. It yields

$$x = \tfrac{1}{t}(tx + (1-t)\bar{x}) + (1 - \tfrac{1}{t})\bar{x} \in \operatorname{aff}(\Omega),$$

which justifies the claimed statement due to the definition of relative interior.

To verify the second statement in (**i**), take $v \in N(\bar{x};\Omega)$ with $\bar{x} \in \operatorname{int}(\Omega)$ and get

$$\langle v, x - \bar{x} \rangle \leq 0 \text{ for all } \ x \in \Omega.$$

Choosing $\delta > 0$ such that $\bar{x} + tv \in \mathbb{B}(\bar{x}; \delta) \subset \Omega$ for $t > 0$ sufficiently small, gives us

$$\langle v, \bar{x} + tv - \bar{x} \rangle = t\|v\|^2 \leq 0,$$

which implies $v = 0$ and thus completes the proof of the second assertion in part $(\mathbf{i})$.

$(\mathbf{ii})$ The continuity of f allows us to find $\delta > 0$ such that

$$|f(x) - f(\bar{x})| < 1 \text{ whenever } \ x \in \mathbb{B}(\bar{x}; \delta).$$

This yields $\mathbb{B}(\bar{x}; \delta) \subset \text{dom}(f)$ and shows therefore that $\bar{x} \in \text{int}(\text{dom}(f))$. Now suppose that $(v, -\lambda) \in N((\bar{x}, f(\bar{x})); \text{epi}(f))$. Then

$$\langle v, x - \bar{x} \rangle - \lambda(t - f(\bar{x})) \leq 0 \text{ whenever } \ (x, t) \in \text{epi}(f).$$

Employing this inequality with $x = \bar{x}$ and $t = f(\bar{x}) + 1$ yields $\lambda \geq 0$.

If $\lambda > 0$, we readily get $(v/\lambda, -1) \in N((\bar{x}, f(\bar{x})); \mathrm{epi}(f))$. It follows from Proposition 8.1 that $v/\lambda \in \partial f(\bar{x})$, and hence $v \in \lambda \partial f(\bar{x})$. In the case where $\lambda = 0$, we deduce that $v \in N(\bar{x}; (\mathrm{dom}(f)) = \{0\}$, and so the inclusion $v \in \lambda \partial f(\bar{x})$ is also valid. $\square$

Theorem 10.3 Let $f_i\colon \mathbb{R}^n \to (-\infty,\infty]$, $i=1,\ldots,m$, be convex functions, and let $\bar{x} \in \bigcap_{i=1}^m \operatorname{dom} f_i$ be such that each f_i is continuous at $\bar{x}$. Then we have the maximum rule:

$$\partial\big(\max f_i\big)(\bar{x}) = \operatorname{co}\Big(\bigcup_{i\in I(\bar{x})} \partial f_i(\bar{x})\Big).$$

Proof Let f be the maximum function defined by $f(x)=\max\big(f_i(x)\big)$ for which we obviously have

$$\operatorname{epi}(f) = \bigcap_{i=1}^m \operatorname{epi}(f_i).$$

Employing Lemma 10.2 $(\mathbf{i})$ give us the equalities

$$\begin{aligned}\operatorname{ri}(\operatorname{epi}(\mathrm{f_i})) &= \{(x,\lambda) \mid x \in \operatorname{ri}(\operatorname{dom}(f_i)),\ \lambda > f_i(x)\}\\ &= \{(x,\lambda) \mid x \in \operatorname{int}(\operatorname{dom}(f_i)),\ \lambda > f_i(x)\},\end{aligned}$$

which implies that

$$(\bar{x}, f(\bar{x})+1) \in \bigcap_{i=1}^m \operatorname{int}(\operatorname{epi}(f_i)) = \bigcap_{i=1}^m \operatorname{ri}(\operatorname{epi}(f_i)).$$

Furthermore, since $f_i(\bar{x}) < f(\bar{x}) = \bar{\alpha}$ for any $i \notin I(\bar{x})$, there exists a neighborhood U of $\bar{x}$ and $\gamma > 0$ such that $f_i(x) < \alpha$ whenever $(x, \alpha) \in U \times (\bar{\alpha} - \gamma, \bar{\alpha} + \gamma)$. It follows that $(\bar{x}, \bar{\alpha}) \in \operatorname{int}(\operatorname{epi}(f_i))$, and so $N((\bar{x}, \bar{\alpha}); \operatorname{epi}(f_i)) = \{(0,0)\}$ for such indices i. Thus the normal cone intersection rule tells us that

$$N\big((\bar{x}, f(\bar{x})); \operatorname{epi}(f)\big) = \sum_{i=1}^m N\big((\bar{x}, \bar{\alpha}); \operatorname{epi}(f_i)\big) = \sum_{i \in I(\bar{x})} N\big((\bar{x}, f_i(\bar{x})); \operatorname{epi}(f_i)\big).$$

Picking $v \in \partial f(\bar{x})$, we have that $(v, -1) \in N((\bar{x}, f(\bar{x})); \operatorname{epi} f)$, which allows us to find $(v_i, -\lambda_i) \in N((\bar{x}, f_i(\bar{x})); \operatorname{epi} f_i)$ for $i \in I(\bar{x})$ such that

$$(v, -1) = \sum_{i \in I(\bar{x})} (v_i, -\lambda_i).$$

This yields $\sum_{i \in I(\bar{x})} \lambda_i = 1$, $\lambda_i \geq 0$, $v = \sum_{i \in I(\bar{x})} v_i$, and $v_i \in \lambda_i \partial f_i(\bar{x})$ by Lemma 10.2 (**ii**). Thus $v = \sum_{i \in I(\bar{x})} \lambda_i u_i$, where $u_i \in \partial f_i(\bar{x})$ and $\sum_{i \in I(\bar{x})} \lambda_i = 1$. This verifies that

$$v \in \operatorname{co}\left(\bigcup_{i \in I(\bar{x})} \partial f_i(\bar{x})\right).$$

The opposite inclusion in the maximum rule follows from

$$\partial f_i(\bar{x}) \subset \partial f(\bar{x}) \text{ for all } \ i \in I(\bar{x}),$$

which in turn follows directly from the definitions. □

Lagrangian Duality

March 2021

Outline

Lagrangian Duality

- Given a nonlinear programming problem, known as the *primal problem*, there exists another nonlinear programming problem, closely related to it, that receives the name of the *Lagrangian dual problem*.

- Under certain convexity assumptions and suitable constraint qualifications, the primal and dual problems have equal optimal objective values.

The Primal Problem

Consider the following nonlinear programming problem:

Primal Problem P

$$
\begin{aligned}
&\text{minimise } f(x), && (1)\\
&\text{subject to:}\\
&g_i(x) \leq 0 \quad \text{for } i = 1, \ldots, m,\\
&h_i(x) = 0 \quad \text{for } i = 1, \ldots, \ell,\\
&x \in X.
\end{aligned}
$$

The Dual Problem

Then the *Lagrangian dual problem* is defined as the following nonlinear programming problem.

Lagrangian Dual Problem D

$$\begin{aligned} &\text{maximise } \theta(u, v), \\ &\text{subject to:} \\ &u \geq 0, \end{aligned} \tag{2}$$

where,

$$\theta(u, v) = \inf\{f(x) + \sum_{i=1}^{m} u_i g_i(x) + \sum_{i=1}^{\ell} v_i h_i(x) : x \in X\}, \tag{3}$$

is the *Lagrangian dual function*.

The Dual Problem

- In the dual problem (2)–(3), the vectors u and v have as their components the Lagrange multipliers u_i for $i = 1, \ldots, m$, and v_i for $i = 1, \ldots, \ell$.

- Note that the Lagrange multipliers u_i, corresponding to the inequality constraints $g_i(x) \leq 0$, are restricted to be nonnegative, whereas the Lagrange multipliers v_i, corresponding to the equality constraints $h_i(x) = 0$, are unrestricted in sign.

- Given the primal problem P (1), several Lagrangian dual problems D of the form of (2)–(3) can be devised, depending on which constraints are handled as $g_i(x) \leq 0$ and $h_i(x) = 0$, and which constraints are handled by the set X. (An appropriate selection of the set X must be made, depending on the nature of the problem.)

Geometric Interpretation

Consider the following primal problem P:

Primal Problem P

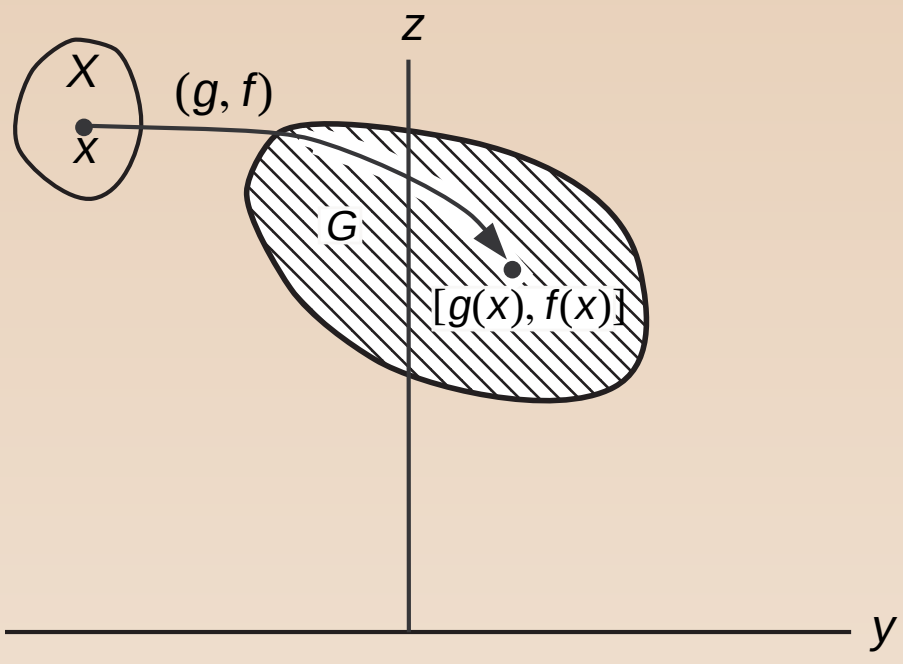
minimise $f(x)$,

subject to:

$g(x) \leq 0$,

$x \in X$,

where $f : \mathbb{R}^n \to \mathbb{R}$ and $g : \mathbb{R}^n \to \mathbb{R}$.



Define the following set in $\mathbb{R}^2$:

$$G = \{(y, z) : y = g(x), z = f(x) \text{ for some } x \in X\},$$

that is, G is the image of X under the (g, f) map.

Geometric Interpretation

$$G = \{(y, z) : y = g(x), z = f(x) \text{ for some } x \in X\},$$

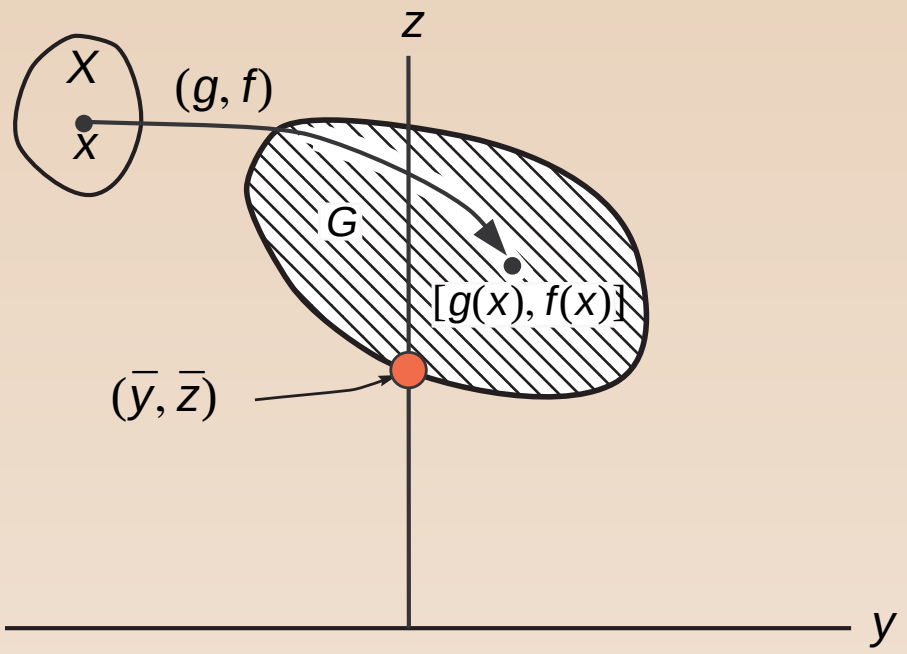
Primal Problem P

minimise $f(x)$,

subject to:

$g(x) \leq 0$,

$x \in X$.



Then, the primal problem consists in finding a point in G with $y \leq 0$ that has minimum ordinate z.

Obviously this point is $(\bar{y}, \bar{z})$.

Geometric Interpretation

Lagrangian Dual Problem D

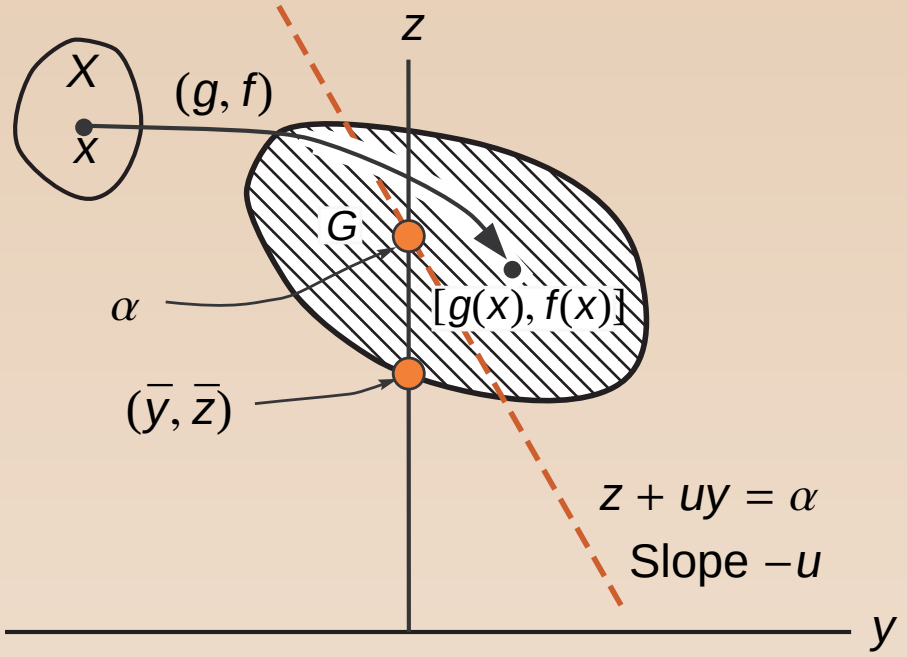
maximise $\theta(u)$,

subject to:

$u \geq 0$,

where (*Lagrangian dual subproblem*):

$\theta(u) = \inf\{f(x)+ug(x) : x \in X\}$.



Given $u \geq 0$, the Lagrangian dual subproblem is equivalent to minimise $z + uy$ over points (y, z) in G. Note that $z + uy = \alpha$ is the equation of a straight line with slope $-u$ that intercepts the z-axis at α.

Geometric Interpretation

Lagrangian Dual Problem D

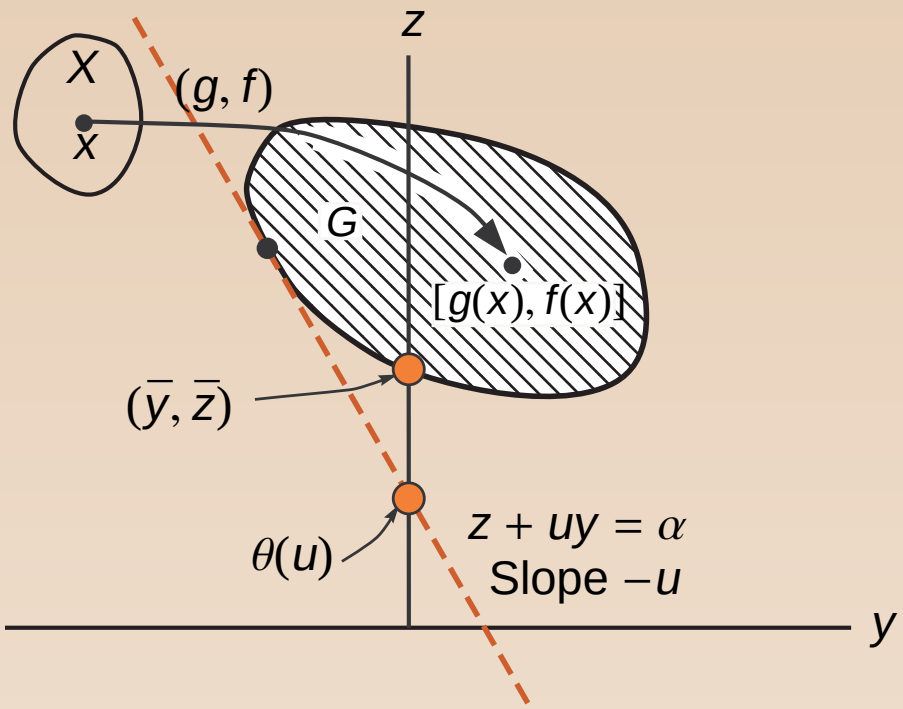
maximise $\theta(u)$,

subject to:

$u \geq 0$,

where (*Lagrangian dual subproblem*):

$\theta(u) = \inf\{f(x)+ug(x) : x \in X\}$.



In order to minimise $z + uy$ over G we need to move the line $z + uy = \alpha$ parallel to itself as far down as possible, whilst it remains in contact with G. The last intercept on the z-axis thus obtained is the value of $\theta(u)$ corresponding to the given $u \geq 0$.

Geometric Interpretation

Lagrangian Dual Problem D

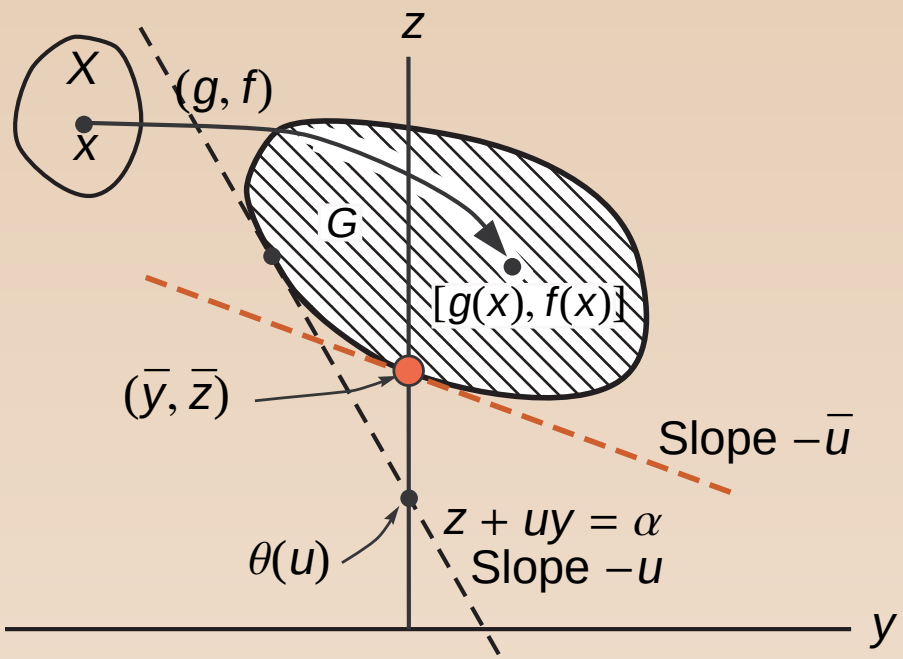
maximise $\theta(u)$,

subject to:

$u \geq 0$,

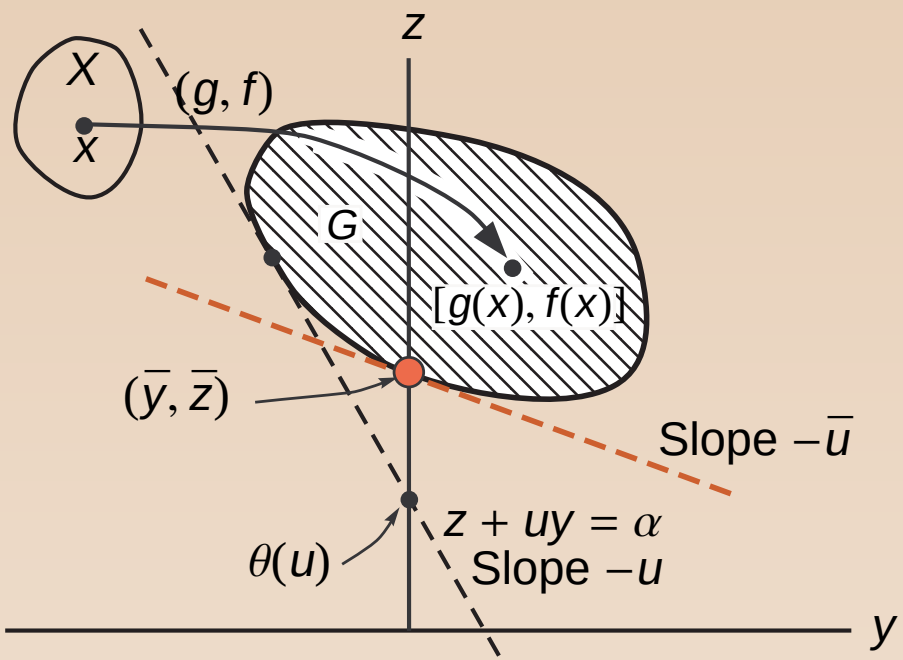
where (*Lagrangian dual subproblem*):

$\theta(u) = \inf\{f(x)+ug(x) : x \in X\}$.



Finally, to solve the dual problem, we have to find the line with slope $-u$ ($u \geq 0$) such that the last intercept on the z-axis, $\theta(u)$, is maximal. Such a line has slope $-\bar{u}$ and supports the set G at the point $(\bar{y}, \bar{z})$. Thus, the solution to the dual problem is $\bar{u}$, and the optimal dual objective value is $\bar{z}$.

Geometric Interpretation



- The solution of the Primal problem is $\bar{z}$, and the solution of the Dual problem is also $\bar{z}$.
- It can be seen that, in the example illustrated, the optimal primal and dual objective values are equal. In such cases, it is said that there is no *duality gap* (strong duality).

Weak Duality

The following result shows that the objective value of any feasible solution to the dual problem constitutes a lower bound for the objective value of any feasible solution to the primal problem.

Theorem (Weak Duality Theorem)

Consider the primal problem P *given by* (1) *and its Lagrangian dual problem* D *given by* (2)*. Let* x *be a feasible solution to* P*; that is,* $x \in X$*,* $g(x) \leq 0$*, and* $h(x) = 0$*. Also, let* (u, v) *be a feasible solution to* D*; that is,* $u \geq 0$*. Then:*

$$f(x) \geq \theta(u, v).$$

$\circ$

Weak Duality

Proof.

We use the definition of θ given in (3), and the facts that $x \in X$, $u \geq 0$, $g(x) \leq 0$ and $h(x) = 0$. We then have

$$\begin{aligned}\theta(u, v) &= \inf\{f(\tilde{x}) + u^{\mathsf{T}} g(\tilde{x}) + v^{\mathsf{T}} h(\tilde{x}) : \tilde{x} \in X\} \\ &\leq f(x) + u^{\mathsf{T}} g(x) + v^{\mathsf{T}} h(x) \leq f(x),\end{aligned}$$

and the result follows. □

Weak Duality

We then have, as a corollary of the previous theorem, the following result.

Corollary

$$\inf\{f(x) : x \in X, g(x) \leq 0, h(x) = 0\} \geq \sup\{\theta(u, v) : u \geq 0\}.$$

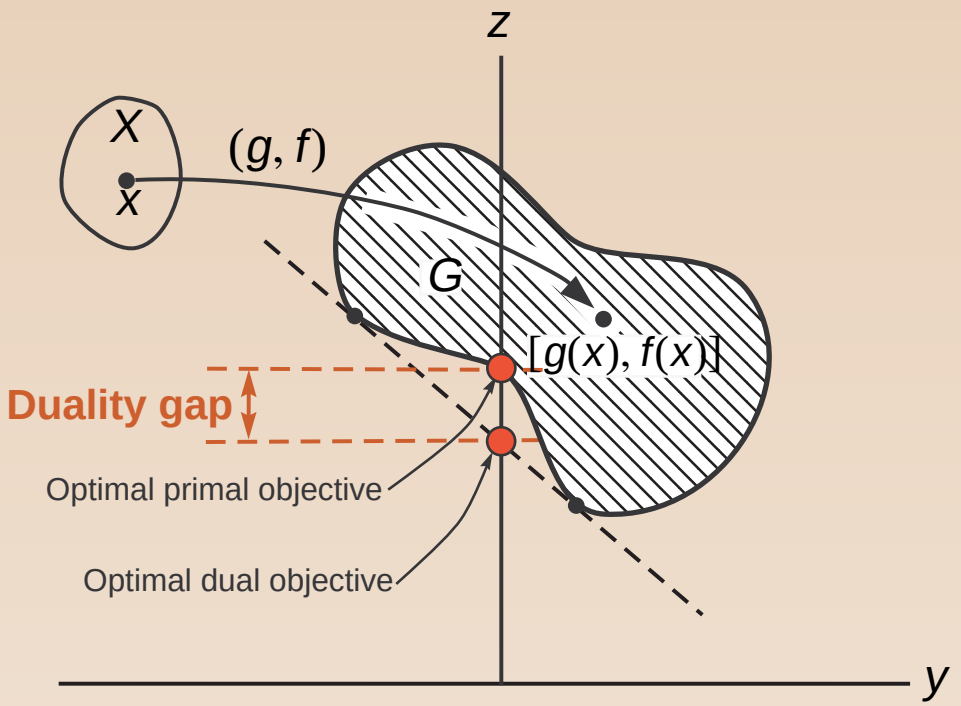
$\circ$

Note from the corollary that the optimal objective value of the primal problem is greater than or equal to the optimal objective value of the dual problem.

If the inequality holds as a *strict* inequality, then it is said that there exists a *duality gap*.

Weak Duality

The figure shows an example of the geometric interpretation of the primal and dual problems.



Notice that, in the case shown in the figure, there exists a duality gap due to the nonconvexity of the set *G*.

We will see, in the **Strong Duality Theorem**, that if some suitable convexity conditions are satisfied, then there is no duality gap between the primal and dual optimisation problems.

Strong Duality

Before stating the conditions that guarantee the absence of a duality gap, we need the following result.

Lemma

Let X be a nonempty convex set in $\mathbb{R}^n$. Let $\alpha : \mathbb{R}^n \to \mathbb{R}$ and $g : \mathbb{R}^n \to \mathbb{R}^m$ be (componentwise) convex, and $h : \mathbb{R}^n \to \mathbb{R}^\ell$ be affine (that is, assume h is of the form $h(x) = Ax - b$). Also, let u_0 be a scalar, $u \in \mathbb{R}^m$ and $v \in \mathbb{R}^\ell$. Consider the following two systems:

System 1: $\alpha(x) < 0, \quad g(x) \leq 0, \quad h(x) = 0 \quad$ *for some* $x \in X$.

System 2: $u_0\alpha(x) + u^\intercal g(x) + v^\intercal h(x) \geq 0$ *for some* $(u_0, u, v) \neq (0, 0, 0)$, $(u_0, u) \geq (0, 0)$ *and for all* $x \in X$.

If System 1 has no solution x, then System 2 has a solution (u_0, u, v). Conversely, if System 2 has a solution (u_0, u, v) with $u_0 > 0$, then System 1 has no solution.

Proof of the Lemma

Outline of the proof:

Assume first that

System 1: $\alpha(x) < 0, \quad g(x) \leq 0, \quad h(x) = 0 \quad$ for some $x \in X$,

has no solution.

Define the set:

$$S = \{(p, q, r) : p > \alpha(x), q \geq g(x), r = h(x) \quad \text{for some } x \in X\}.$$

The set S is convex, since X, α and g are convex and h is affine. Since System 1 has no solution, we have that $(0, 0, 0) \notin S$.

Proof of the Lemma (Ctd.)

Example

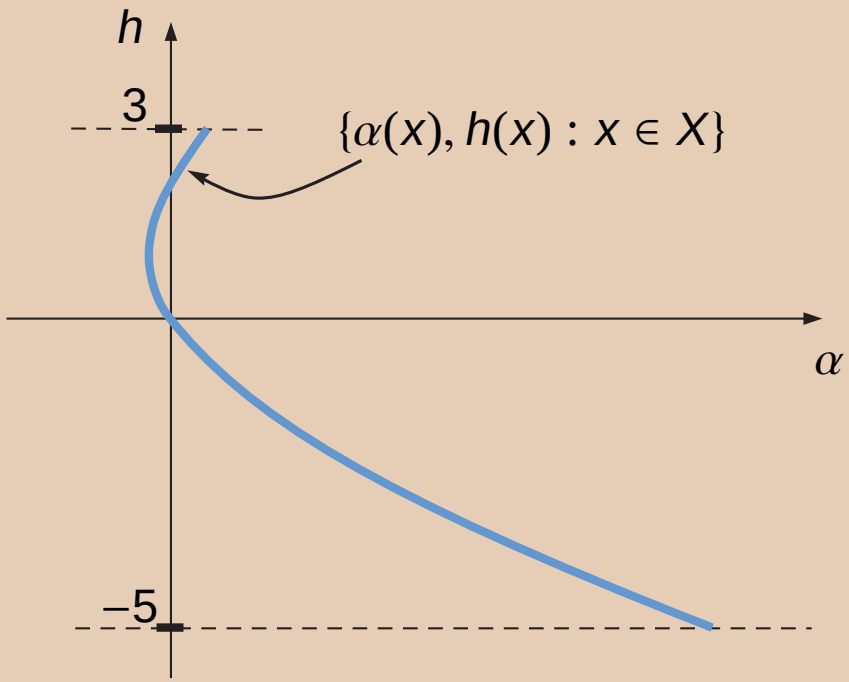
Consider the functions:

$\alpha(x) = (x-1)^2 - \frac{1}{4}$,

$h(x) = 2x - 1$,

and the set

$X = \{x \in \mathbb{R} : |x| \leq 2\}$.



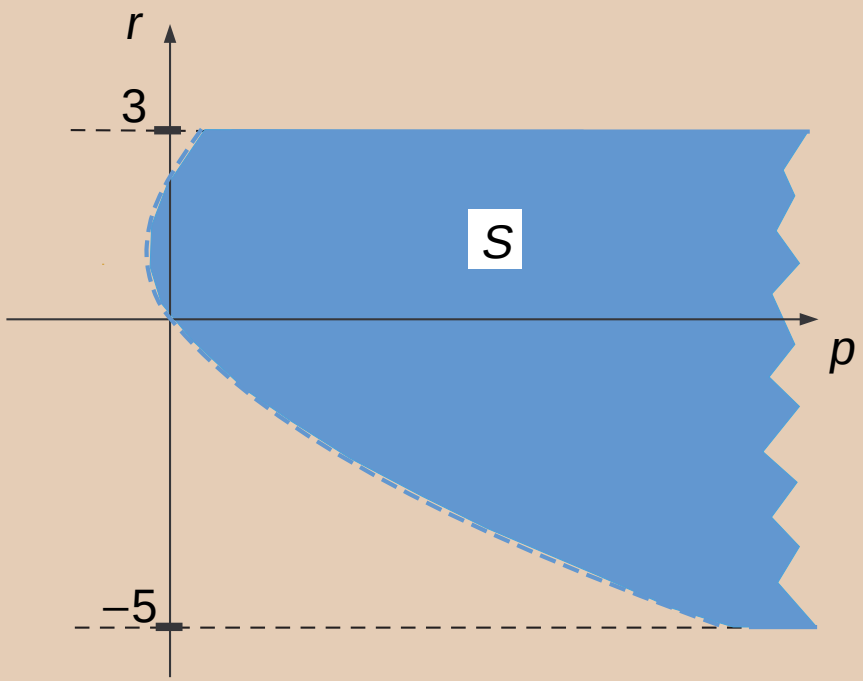
Proof of the Lemma (Ctd.)

Example (Ctd.)

$\alpha(x) = (x-1)^2 - \frac{1}{4}$,

$h(x) = 2x - 1$,

$X = \{x \in \mathbb{R} : |x| \leq 2\}$.



$$S = \{(p, r) : p > \alpha(x), r = h(x) \quad \textit{for some } x \in X\}$$

Proof of the Lemma (Ctd.)

Continuing with the proof of the Lemma, we have the convex set:

$$S = \{(p, q, r) : p > \alpha(x), q \geq g(x), r = h(x) \quad \text{for some } x \in X\},$$

and that $(0, 0, 0) \notin S$.

Recall the following corollary of the **Supporting Hyperplane Theorem**:

Corollary

Let S be a nonempty convex set in $\mathbb{R}^n$ and $\bar{x} \notin \operatorname{int} S$. Then there is a nonzero vector p such that $p^{\mathsf{T}}(x - \bar{x}) \leq 0$ for each $x \in \operatorname{cl} S$.

$\circ$

Proof of the Lemma (Ctd.)

We then have, from the above corollary, that there exists a nonzero vector (u_0, u, v) such that

$$(u_0, u, v)^{\mathsf{T}}[(p, q, r) - (0, 0, 0)] = u_0 p + u^{\mathsf{T}} q + v^{\mathsf{T}} r \geq 0, \qquad (4)$$

for each $(p, q, r) \in \mathrm{cl}\, S$.

Now, fix an $x \in X$. Noticing, from the definition of S, that p and q can be made arbitrarily large, we have that in order to satisfy (4), we must have $u_0 \geq 0$ and $u \geq 0$.

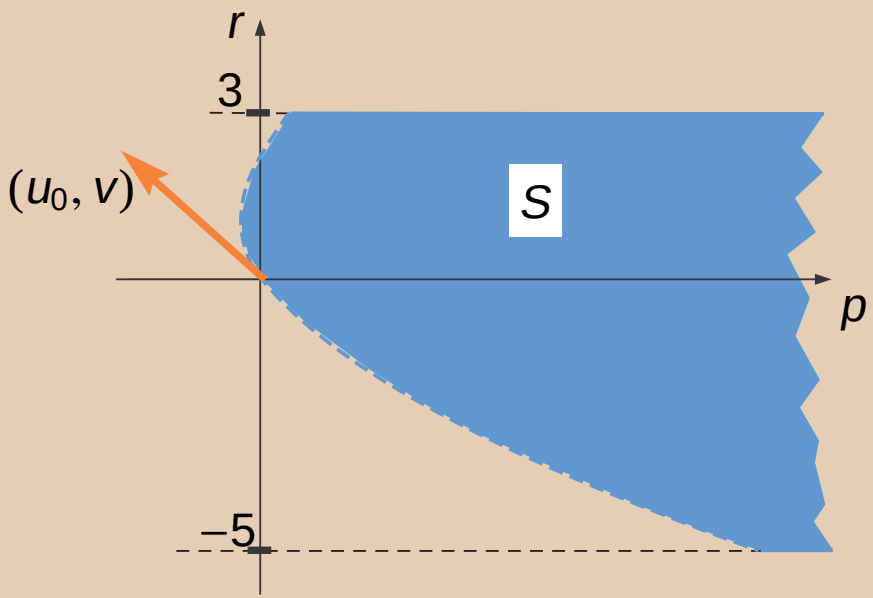
Proof of the Lemma (Ctd.)

Example (Ctd.)

$\alpha(x) = (x-1)^2 - \frac{1}{4}$,

$h(x) = 2x - 1$,

$X = \{x \in \mathbb{R} : |x| \leq 2\}$.



We can see that u_0 *cannot be* $u_0 < 0$ *and satisfy:*

$$(u_0, v)^\intercal[(p, r) - (0, 0)] = (u_0, v)^\intercal(p, r) = u_0 p + v^\intercal r \geq 0,$$

for each $(p, q, r) \in \operatorname{cl} S$.

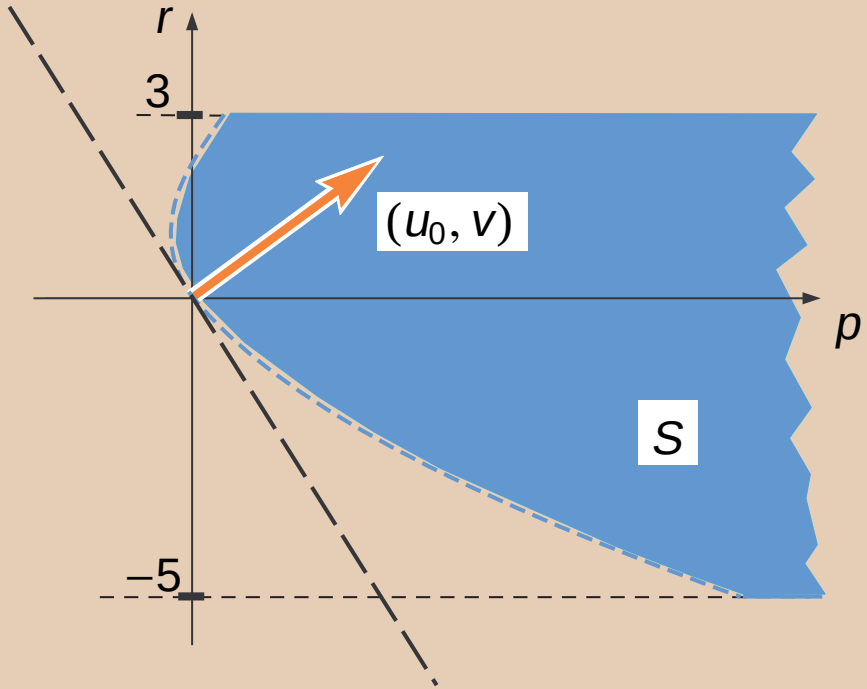
Proof of the Lemma (Ctd.)

Example (Ctd.)

$\alpha(x) = (x - 1)^2 - \frac{1}{4}$,

$h(x) = 2x - 1$,

$X = \{x \in \mathbb{R} : |x| \leq 2\}$.



We conclude that $u_0 \geq 0$ *and*

$$(u_0, v)^\intercal[(p, r) - (0, 0)] = (u_0, v)^\intercal(p, r) = u_0 p + v^\intercal r \geq 0,$$

for each $(p, q, r) \in \operatorname{cl} S$.

Proof of the Lemma (Ctd.)

We have that there exists a nonzero vector (u_0, u, v) with $(u_0, u) \geq (0, 0)$ such that

$$(u_0, u, v)^{\mathsf{T}}[(p, q, r) - (0, 0, 0)] = u_0 p + u^{\mathsf{T}} q + v^{\mathsf{T}} r \geq 0,$$

for each $(p, q, r) \in \operatorname{cl} S$.

Also, note that $[\alpha(x), g(x), h(x)] \in \operatorname{cl} S$ and we have from the above inequality that

$$u_0 \alpha(x) + u^{\mathsf{T}} g(x) + v^{\mathsf{T}} h(x) \geq 0.$$

Since the above inequality is true for each $x \in X$, we conclude that

System 2: $u_0 \alpha(x) + u^{\mathsf{T}} g(x) + v^{\mathsf{T}} h(x) \geq 0$ for some $(u_0, u, v) \neq (0, 0, 0)$, $(u_0, u) \geq (0, 0)$ and for all $x \in X$.

has a solution.

Proof of the Lemma (Ctd.)

To prove the converse, assume that

System 2: $u_0\alpha(x) + u^{\mathsf{T}}g(x) + v^{\mathsf{T}}h(x) \geq 0$ for some $(u_0, u, v) \neq (0, 0, 0)$, $(u_0, u) \geq (0, 0)$ and for all $x \in X$,

has a solution (u_0, u, v) such that $u_0 > 0$.

Suppose that $x \in X$ is such that $g(x) \leq 0$ and $h(x) = 0$.

From the previous inequality we conclude that $u_0\alpha(x) \geq -u^{\mathsf{T}}g(x) \geq 0$, since $u \geq 0$ and $g(x) \leq 0$. But, since $u_0 > 0$, we must then have that $\alpha(x) \geq 0$.

Hence,

System 1: $\alpha(x) < 0, \quad g(x) \leq 0, \quad h(x) = 0$ for some $x \in X$.

has no solution and this completes the proof. □

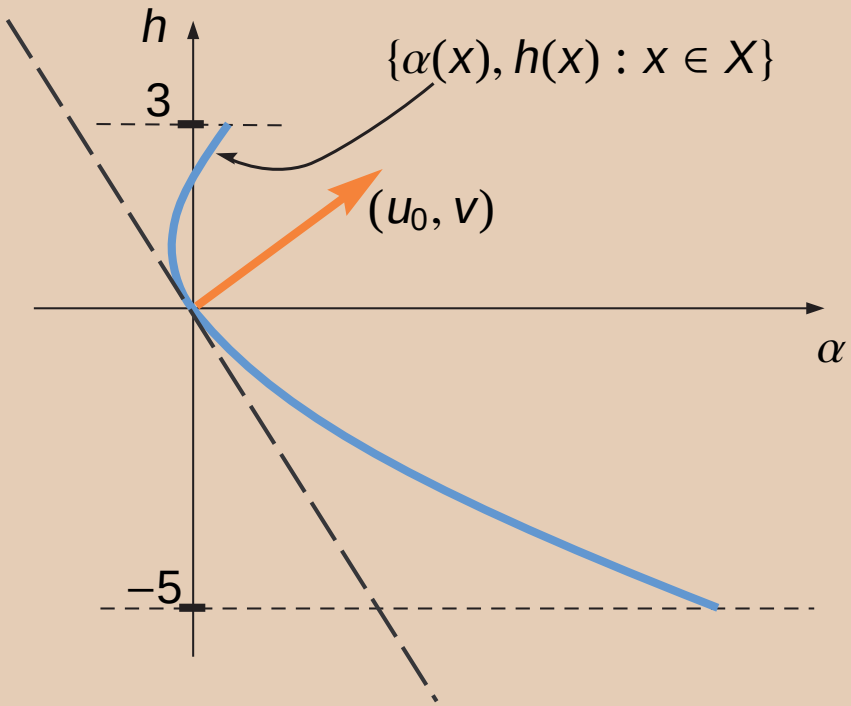
Proof of the Lemma (Ctd.)

Example (Ctd.)

$\alpha(x) = (x-1)^2 - \frac{1}{4}$,

$h(x) = 2x - 1$,

$X = \{x \in \mathbb{R} : |x| \leq 2\}$.



If $\Big\{$ ***System 2:*** $u_0\alpha(x) + v^{\mathsf{T}}h(x) \geq 0$ *for some* $(u_0, v) \neq (0,0)$, $u_0 \geq 0$ *and for all* $x \in X$ $\Big\}$, *has a solution such that* $u_0 > 0$, *and* $x \in X$ *is such that* $h(x) = 0$, *we can see that* $\alpha(x)$ *must be* $\alpha(x) \geq 0$, *and hence* ***System 1*** *has no solution.*

Strong Duality

The following result, known as the *strong duality theorem*, shows that, under suitable convexity assumptions and under a constraint qualification, there is no *duality gap* between the primal and dual optimal objective function values.

Theorem (Strong Duality Theorem)

Let X be a nonempty convex set in $\mathbb{R}^n$. Let $f : \mathbb{R}^n \to \mathbb{R}$ and $g : \mathbb{R}^n \to \mathbb{R}^m$ be convex, and $h : \mathbb{R}^n \to \mathbb{R}^\ell$ be affine. Suppose that the following constraint qualification is satisfied. There exists an $\hat{x} \in X$ such that $g(\hat{x}) < 0$ and $h(\hat{x}) = 0$, and $0 \in \operatorname{int} h(X)$, where $h(X) = \{h(x) : x \in X\}$. Then,

$$\inf\{f(x) : x \in X, g(x) \leq 0, h(x) = 0\} = \sup\{\theta(u, v) : u \geq 0\}, \qquad (5)$$

where $\theta(u, v) = \inf\{f(x) + u^\intercal g(x) + v^\intercal h(x) : x \in X\}$. Furthermore, if the inf *is finite, then* $\sup\{\theta(u, v) : u \geq 0\}$ *is achieved at $(\overline{u}, \overline{v})$ with $\overline{u} \geq 0$. If the* inf *is achieved at $\overline{x}$, then $\overline{u}^\intercal g(\overline{x}) = 0$.*

Proof of the Strong Duality Theorem

Let $\gamma = \inf\{f(x) : x \in X, g(x) \leq 0, h(x) = 0\}$.

By assumption there exists a feasible solution $\hat{x}$ for the primal problem and hence $\gamma < \infty$.

If $\gamma = -\infty$, we then conclude from the corollary of the **Weak Duality Theorem** that $\sup\{\theta(u, v) : u \geq 0\} = -\infty$ and, hence, (5) is satisfied.

Thus, suppose that γ is finite, and consider the following system:

$$f(x) - \gamma < 0, \qquad g(x) \leq 0 \qquad h(x) = 0, \quad \text{for some } x \in X.$$

By the definition of γ, this system has no solution. Hence, from the previous lemma, there exists a nonzero vector (u_0, u, v) with $(u_0, u) \geq (0, 0)$ such that

$$u_0[f(x) - \gamma] + u^{\mathsf{T}} g(x) + v^{\mathsf{T}} h(x) \geq 0 \qquad \text{for all } x \in X. \qquad (6)$$

Proof of the Strong Duality Theorem (Ctd.)

We will next show that $u_0 > 0$. Suppose, by contradiction that $u_0 = 0$.

By assumption, there exists an $\hat{x} \in X$ such that $g(\hat{x}) < 0$ and $h(\hat{x}) = 0$. Substituting in (6) we obtain $u^\intercal g(\hat{x}) \geq 0$. But, since $g(\hat{x}) < 0$ and $u \geq 0$, $u^\intercal g(\hat{x}) \geq 0$ is only possible if $u = 0$.

From (6), $u_0 = 0$ and $u = 0$ imply that $v^\intercal h(x) \geq 0$ for all $x \in X$. But, since $0 \in \operatorname{int} h(X)$, we can choose an $x \in X$ such that $h(x) = -\lambda v$, where $\lambda > 0$. Therefore, $0 \leq v^\intercal h(x) = -\lambda \|v\|^2$, which implies that $v = 0$.

Thus, it has been shown that $u_0 = 0$ implies that $(u_0, u, v) = (0, 0, 0)$, which is a contradiction. We conclude, then, that $u_0 > 0$.

Proof of the Strong Duality Theorem (Ctd.)

Dividing (6) by u_0 and denoting $\bar{u} = u/u_0$ and $\bar{v} = v/u_0$, we obtain

$$f(x) + \bar{u}^T g(x) + \bar{v}^T h(x) \geq \gamma \qquad \text{for all } x \in X. \tag{7}$$

This implies that $\theta(\bar{u}, \bar{v}) = \inf\{f(x) + \bar{u}^T g(x) + \bar{v}^T h(x) : x \in X\} \geq \gamma$.

We then conclude, from the **Weak Duality Theorem**, that $\theta(\bar{u}, \bar{v}) = \gamma$. And, from the corollary of the **Weak Duality Theorem**, we conclude that $(\bar{u}, \bar{v})$ solves the dual problem.

Finally, to complete the proof, assume that $\bar{x}$ is an optimal solution to the primal problem; that is, $\bar{x} \in X$, $g(\bar{x}) \leq 0$, $h(\bar{x}) = 0$ and $f(\bar{x}) = \gamma$.

From (7), letting $x = \bar{x}$, we get $\bar{u}^T g(\bar{x}) \geq 0$. Since $\bar{u} \geq 0$ and $g(\bar{x}) \leq 0$, we get $\bar{u}^T g(\bar{x}) = 0$.

This completes the proof. □

Example of Strong Duality

Example

Consider the following optimisation problem:

Primal Problem P

$$\begin{aligned} &\textit{minimise } (x-1)^2, \\ &\textit{subject to:} \\ &2x - 1 = 0, \\ &x \in X = \{x \in \mathbb{R} : |x| \leq 2\}. \end{aligned}$$

It is clear that the optimal value of the objective function is equal to $\left(\frac{1}{2} - 1\right)^2 = \frac{1}{4}$, *since the feasible set is the singleton* $\left\{\frac{1}{2}\right\}$.

Example of Strong Duality

Example (Ctd.)

Lagrangian Dual Problem D

maximise $\theta(v)$,

where the Lagrangian dual function is,

$\theta(v) = \inf\{(x-1)^2 + v(2x-1) : |x| \leq 2\}$.

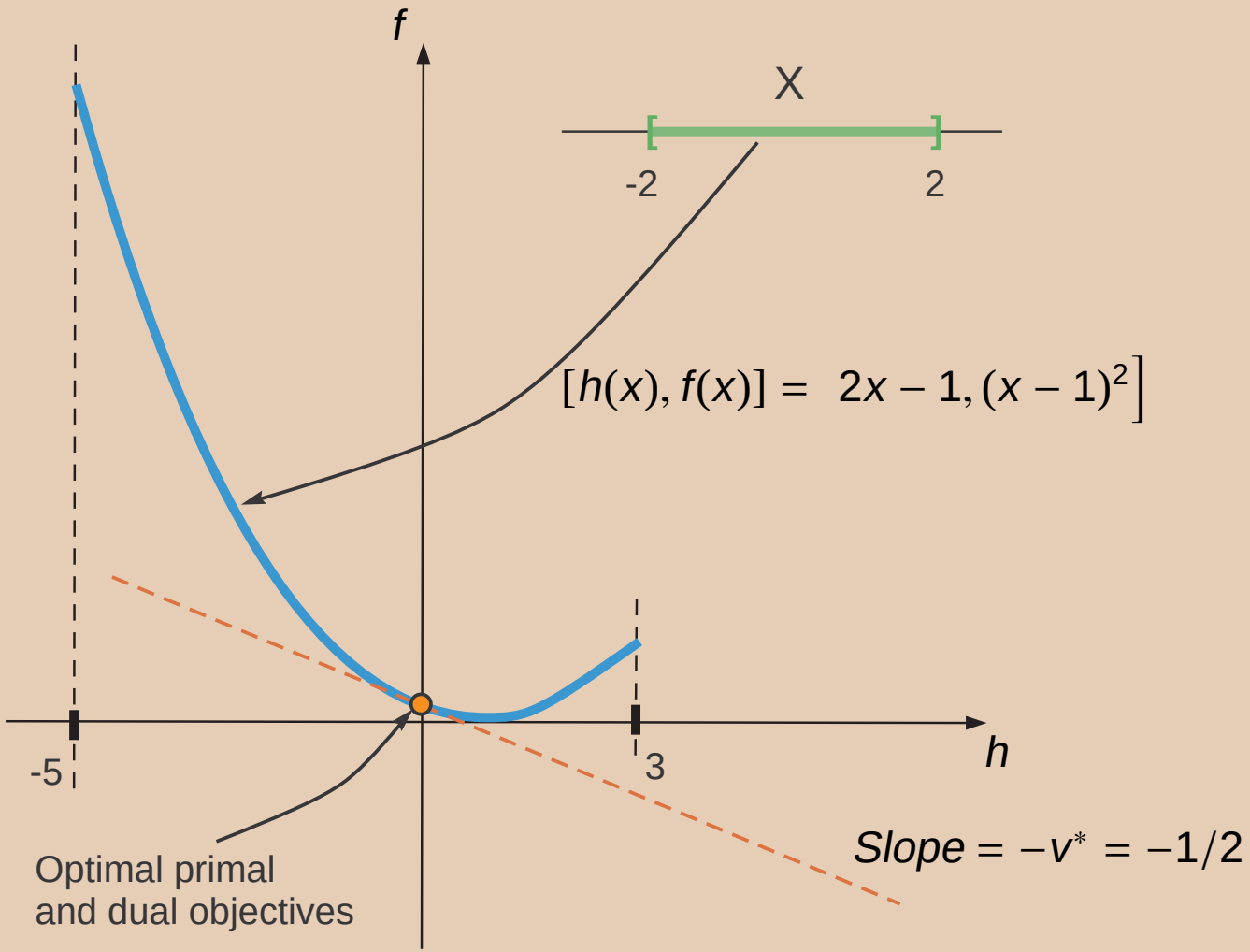
Differentiating w.r.t. x and equating to zero, we get that the optimiser of the dual Lagrangian subproblem is $x^* = -v + 1$ *(if* $-1 \leq v \leq 3$*).*

Hence $\theta(v) = (-v+1-1)^2 + v(-2v+2-1) = -v^2 + v$.

Differentiating w.r.t. v and equating to zero, we get that the optimiser of the dual problem is $v^* = \frac{1}{2}$ *and the optimal value of the dual problem is* $-v^{*2} + v^* = \frac{1}{4}$. *Thus, there is no duality gap.*

Example of Strong Duality

Example (Ctd.)



11 Continuity of Convex Functions

A function $f\colon \mathbb{R}^n \to (-\infty,\infty]$ be an extended-real-valued function. We say that f is *continuous* at $x_0 \in \mathrm{dom}(f) := \{x \in \mathbb{R}^n \mid f(x) < \infty\}$ if for any $\epsilon > 0$, there exists $\delta > 0$ such that

$$|f(x) - f(x_0)| < \epsilon \text{ whenever } \|x - x_0\| < \delta.$$

It follows directly from the definition that if f is continuous at $x_0 \in \mathrm{dom}(f)$, then $x_0 \in \mathrm{int}(\mathrm{dom}(f))$.

Lemma 11.1 If a convex function $f\colon \mathbb{R}^n \to (-\infty,\infty]$ is bounded above on $\mathbb{B}(\bar{x};\delta)$ for some element $\bar{x} \in \mathrm{dom}(f)$ and number $\delta > 0$, then f is bounded on $\mathbb{B}(\bar{x};\delta)$.

Proof Denote $m := f(\bar{x})$ and take $M > 0$ with $f(x) \le M$ for all $x \in \mathbb{B}(\bar{x};\delta)$. Picking any $u \in \mathbb{B}(\bar{x};\delta)$, consider the element $x := 2\bar{x} - u$. Then $x \in \mathbb{B}(\bar{x};\delta)$ and

$$m = f(\bar{x}) = f\big(\frac{x+u}{2}\big) \le \frac{1}{2}f(x) + \frac{1}{2}f(u),$$

which shows that $f(u) \geq 2m - f(x) \geq 2m - M$ and thus f is bounded on $\mathbb{B}(\bar{x}; \delta)$. $\square$

Definition We say that $f : \mathbb{R}^n \to (-\infty, \infty]$ is *Lipschitz continuous (https://en.wikipedia.org/wiki/Lipschitz_continuity)* on $\Omega \subset \operatorname{dom}(f)$ if there exists $\delta > 0$ and $\ell \geq 0$ such that

$$|f(x) - f(u)| \leq \ell \|x - u\| \text{ whenever } x, u \in \Omega.$$

We say that f is locally Lipschitz continuous around $x_0 \in \operatorname{dom}(f)$ if there exists $\delta > 0$ and $\ell \geq 0$ such that

$$|f(x) - f(u)| \leq \ell \|x - u\| \text{ whenever } x, u \in \mathbb{B}(x_0; \delta).$$

Theorem 11.2 Let $f : \mathbb{R}^n \to (-\infty, \infty]$ be convex with $\bar{x} \in \operatorname{dom}(f)$. If f is bounded above on $\mathbb{B}(\bar{x}; \delta)$ for some $\delta > 0$, then f is Lipschitz continuous on $\mathbb{B}(\bar{x}; \delta/2)$.

Proof Fix $x, y \in \mathbb{B}(\bar{x}; \delta/2)$ with $x \neq y$ and we define the element u by

$$u := x + \frac{\delta}{2\|x - y\|}\Big(x - y\Big).$$

Since $e := \dfrac{x-y}{\|x-y\|} \in \mathbb{B}$, we have

$$u = x + \frac{\delta}{2}e \in \bar{x} + \frac{\delta}{2}\mathbb{B} + \frac{\delta}{2}\mathbb{B} \subset \bar{x} + \delta\mathbb{B}.$$

If we denote $\alpha := \dfrac{\delta}{2\|x-y\|}$, this gives us $u = x + \alpha(x - y)$, and thus

$$x = \frac{1}{\alpha+1}u + \frac{\alpha}{\alpha+1}y.$$

It follows from the convexity of f that

$$f(x) \le \frac{1}{\alpha+1} f(u) + \frac{\alpha}{\alpha+1} f(y),$$

which implies in turn the inequalities

$$f(x) - f(y) \le \frac{1}{\alpha+1}(f(u) - f(y)) \le 2M \frac{1}{\alpha+1}$$

$$= 2M \frac{2\|x-y\|}{\delta + 2\|x-y\|} \le \frac{4M}{\delta}\|x-y\|$$

with $M := \sup\{|f(x)| \mid x \in \mathbb{B}(\bar{x};\delta)\} < \infty$ by Lemma 11.2. Interchanging the role of x and y above, we arrive at the estimate

$$|f(x) - f(y)| \le \frac{4M}{\delta}\|x-y\|$$

and thus verify the Lipschitz continuity of f on $\mathbb{B}(\bar{x};\delta/2)$. $\square$

Lemma 11.3 Let $\{e_i \mid i = 1, \ldots, n\}$ be the standard orthonormal basis of $\mathbb{R}^n$. Denote

$$A := \big\{\bar{x} \pm \epsilon e_i \;\big|\; i = 1, \ldots, n\big\}, \quad \epsilon > 0.$$

Then the following properties hold:

(**i**) $\bar{x} + \gamma e_i \in \operatorname{co} A$ for $|\gamma| \leq \epsilon$ and $i = 1, \ldots, n$.
(**ii**) $\mathbb{B}(\bar{x}; \epsilon/n) \subset \operatorname{co} A$.

Proof (**i**) For $|\gamma| \leq \epsilon$, find $t \in [0, 1]$ with $\gamma = t(-\epsilon) + (1 - t)\epsilon$. Then $\bar{x} \pm \epsilon e_i \in A$ gives us

$$\bar{x} + \gamma e_i = t(\bar{x} - \epsilon e_i) + (1 - t)(\bar{x} + \epsilon e_i) \in \operatorname{co} A.$$

(**ii**) For $x \in \mathbb{B}(\bar{x}; \epsilon/n)$, we have $x = \bar{x} + (\epsilon/n)u$ with $\|u\| \leq 1$. Represent u via $\{e_i\}$ by

$$u = \sum_{i=1}^{n} \lambda_i e_i,$$

where $|\lambda_i| \le \sqrt{\sum_{i=1}^{n} \lambda_i^2} = \|u\| \le 1$ for every i. This gives us

$$x = \bar{x} + \frac{\epsilon}{n} u = \bar{x} + \sum_{i=1}^{n} \frac{\epsilon \lambda_i}{n} e_i = \sum_{i=1}^{n} \frac{1}{n} \Big(\bar{x} + \epsilon \lambda_i e_i \Big).$$

Denoting $\gamma_i := \epsilon \lambda_i$ yields $|\gamma_i| \le \epsilon$. It follows from $(\mathbf{i})$ that $\bar{x} + \epsilon \lambda_i e_i = \bar{x} + \gamma_i e_i \in \operatorname{co} A$, and thus $x \in \operatorname{co} A$ since it is a convex combination of elements in $\operatorname{co} A$. $\square$

Theorem 11.4 Let $f : \mathbb{R}^n \to (-\infty, \infty]$ be a convex function such that $\operatorname{int}(\operatorname{dom} f) \neq \emptyset$. Then f is locally Lipschitz continuous on in the interior of its domain $\operatorname{int}(\operatorname{dom} f)$.

Proof Pick $\bar{x} \in \operatorname{int}(\operatorname{dom} f)$ and choose $\epsilon > 0$ such that $\bar{x} \pm \epsilon e_i \in \operatorname{dom} f$ for every i. Considering the set A from Lemma 11.3 $(\mathbf{ii})$, we get $\mathbb{B}(\bar{x}; \epsilon/n) \subset \operatorname{co} A$. Denote $M := \max\{f(a) \mid a \in A\} < \infty$ over the finite set A. Using the representation

$$x = \sum_{i=1}^{m} \lambda_i a_i \ \text{ with } \ \lambda_i \geq 0 \ \sum_{i=1}^{m} \lambda_i = 1, \ a_i \in A$$

for any $x \in \mathbb{B}(\bar{x}; \epsilon/n)$ which shows that

$$f(x) \leq \sum_{i=1}^{m} \lambda_i f(a_i) \leq \sum_{i=1}^{m} \lambda_i M = M,$$

and so f is bounded above on $\mathbb{B}(\bar{x}; \epsilon/n)$. Then Theorem 11.2 tells us that f is Lipschitz continuous on $\mathbb{B}(\bar{x}; \epsilon/2n)$ and thus it is locally Lipschitz continuous on $\text{int}\,(\text{dom} f)$. □

It immediately follows that any finite convex function on $\mathbb{R}^n$ is always locally Lipschitz continuous on the whole space.

Corollary 11.5 If $f : \mathbb{R}^n \to \mathbb{R}$ is convex, it is locally Lipschitz continuous on $\mathbb{R}^n$.
The next corollary is also a direct consequence of Theorem 11.4.

Corollary 11.6 Let $f : \mathbb{R}^n \to (-\infty, \infty]$ be a convex function and let $\bar{x} \in \text{dom}(f)$. Then the following properties are equivalent.

(**i**) f is continuous at $\bar{x}$.
(**ii**) $\bar{x} \in \text{int}(\text{dom}(f))$.
(**iii**) f is locally Lipschitz continuous around $\bar{x}$.

12 Fenchel Conjugates

Definition Given a function $f\colon \mathbb{R}^n \to (-\infty, \infty]$ (not necessarily convex), its *Fenchel conjugate (https://en.wikipedia.org/wiki/Convex_conjugate)*, $f^*\colon \mathbb{R}^n \to [-\infty, \infty]$ is given by

$$f^*(v) := \sup\big\{\langle v, x\rangle - f(x) \;\big|\; x \in \mathbb{R}^n\big\}.$$

The proposition below shows that the Fenchel conjugate of a proper function f (f is proper if it has nonempty domain), is a convex function even if f is nonconvex.

Proposition 12.1 Let $f\colon \mathbb{R}^n \to (-\infty, \infty]$ be a function with $\operatorname{dom}(f) \neq \emptyset$, or equivalently there exists an $\bar{x} \in \mathbb{R}^n$ such that $f(\bar{x}) < \infty$. Then $f^*\colon \mathbb{R}^n \to (-\infty, \infty]$ is a convex function.

Proof Fix $\bar{x} \in \operatorname{dom}(f)$ so that $f(\bar{x}) \in \mathbb{R}$. For any $v \in \mathbb{R}^n$ we have that,

$$f^*(v) = \sup\big\{\langle v, x\rangle - f(x) \;:\; x \in \mathbb{R}^n\big\} \geq \langle v, \bar{x}\rangle - f(\bar{x}) > -\infty.$$

Observe that $\langle v, x\rangle - f(x) = -\infty$ if $x \notin \operatorname{dom}(f)$, and so

$$f^*(v) = \sup\big\{\langle v, x\rangle - f(x) \ : \ x \in \mathbb{R}^n\big\} = \sup\big\{\langle v, x\rangle - f(x) \ : \ x \in \operatorname{dom}(f)\big\}$$

$$= \sup\big\{\phi_x(v) \ : \ x \in \operatorname{dom}(f)\big\},$$

where $\phi_x(v) := \langle v, x\rangle - f(x)$. Then f^* is convex on $\mathbb{R}^n$ as it is the supremum of a family of convex functions defined on $\mathbb{R}^n$. $\square$

Proposition 12.2 Let $f, g\colon \mathbb{R}^n \to (-\infty, \infty]$ be such that $f(x) \le g(x)$ for all $x \in \mathbb{R}^n$. Then we have $f^*(v) \ge g^*(v)$ for all $v \in \mathbb{R}^n$.

Proof For any fixed $v \in \mathbb{R}^n$, it follows from the assumption that $f(x) \le g(x)$ for all $x \in \mathbb{R}^n$ that

$$\langle v, x\rangle - f(x) \ge \langle v, x\rangle - g(x), \quad x \in \mathbb{R}^n.$$

This readily implies the relationships

$$f^*(v) = \sup\big\{\langle v, x\rangle - f(x) \ \big| \ x \in \mathbb{R}^n\big\} \ge \sup\big\{\langle v, x\rangle - g(x) \ \big| \ x \in \mathbb{R}^n\big\} = g^*(v).$$

Since this is true for all $v \in \mathbb{R}^n$, then $f^* \geq g^*$ on $\mathbb{R}^n$. $\square$

Proposition 12.3 Let $f, g \colon \mathbb{R}^n \to (-\infty, \infty]$ with $\mathrm{dom}(f) \neq 0$. Then
$\langle v, x \rangle \leq f(x) + f^*(v) \;\; \text{for all} \;\; x \in \mathbb{R}^n$ and $v \in \mathbb{R}^n$.
Proof Observe first that if $f(x) = \infty$ the conclusion follows. If $x \in \mathrm{dom} f$, we get from the definition of Fenchel conjugates that

$$f^*(v) \geq \langle v, x \rangle - f(x).$$

The conclusion follows when we add $f(x)$ to both sides. $\square$

Definition Let $f : \mathbb{R}^n \to (-\infty, \infty]$ be a function with $\mathrm{dom}(f) \neq 0$. Then for $x \in \mathbb{R}^n$ we have the function $f^{**}(x) = (f^*)^*(x)$ given by

$$f^{**}(x) = \sup\{\langle x, v\rangle - f^*(v) \mid v \in \mathbb{R}^n\}.$$

Then $f^{**} : \mathbb{R}^n \to [-\infty, \infty]$ is called the second order Fenchel conjugate of f.

Proposition 12.4 Let $f : \mathbb{R}^n \to (-\infty, \infty]$. Then for all $x \in \mathbb{R}^n$ we have that $f^{**}(x) \leq f(x)$.
Proof Fix any $v, x \in \mathbb{R}^n$. Then from Proposition 12.3 we have

$$\langle v, x\rangle \leq f(x) + f^*(v).$$

When we subtract both sides by $f^*(v)$ we get the inequality

$$\sup\{\langle v, x\rangle - f^*(v) \mid v \in \mathbb{R}^n\} \leq f(x),$$

and since this is true for any $v, x \in \mathbb{R}^n$ then the proof is complete. $\square$

The following important result reveals a close relationship between subgradients and Fenchel conjugates of convex functions.

Proposition 12.5 For any convex function $f\colon \mathbb{R}^n \to (-\infty, \infty]$ and any $\bar{x} \in \operatorname{dom}(f)$, we have that $v \in \partial f(\bar{x})$ if and only if

$$f(\bar{x}) + f^*(v) = \langle v, \bar{x} \rangle.$$

Proof Taking any $v \in \partial f(\bar{x})$ and using the definition of the subdifferential we get

$$f(\bar{x}) + \langle v, x \rangle - f(x) \le \langle v, \bar{x} \rangle \ \text{ for all } \ x \in \mathbb{R}^n.$$

This readily implies the inequality

$$f(\bar{x}) + f^*(v) = f(\bar{x}) + f^*(v) = \sup\big\{\langle v, x \rangle - f(x) \ : \ x \in \mathbb{R}^n\big\} \le \langle v, x \rangle.$$

Since the opposite inequality holds by Proposition 12.4, we arrive at the equation,

$$f(\bar{x}) + f^*(v) = \langle v, \bar{x} \rangle.$$

Conversely, suppose that $f(\bar{x}) + f^*(v) = \langle v, \bar{x} \rangle$. Applying Proposition 12.3, we get the estimate $f^*(v) \geq \langle v, x \rangle - f(x)$ for every $x \in \mathbb{R}^n$. Then

$$f(\bar{x}) + f^*(v) = \langle v, \bar{x} \rangle \geq f(\bar{x}) + \langle v, x \rangle - f(x) \text{ for all } x \in \mathbb{R}^n.$$

This shows that $v \in \partial f(\bar{x})$. $\square$

Proposition 12. 6 Let $f : \mathbb{R}^n \to (-\infty, \infty]$ and let $\bar{x} \in \operatorname{dom}(f)$. Suppose that $\partial f(\bar{x}) \neq \emptyset$. Then we have the equality $f^{**}(\bar{x}) = f(\bar{x})$.
Proof Proposition 12.4 gives us the first inequality thus it suffices to verify the opposite inequality. Fix $v \in \partial f(\bar{x})$ and get $\langle v, \bar{x} \rangle = f(\bar{x}) + f^*(v)$ by the preceding theorem. This shows that

$$f(\bar{x}) = \langle v, \bar{x} \rangle - f^*(v) \leq \sup\big\{ \langle \bar{x}, v \rangle - f^*(v) \;\big|\; v \in \mathbb{R}^n \big\} = f^{**}(\bar{x}),$$

which completes the proof of this proposition. $\square$

13 Directional Derivatives of Convex Functions

Definition Let $f : \mathbb{R}^n \to (-\infty, \infty]$ with $\bar{x} \in \operatorname{dom}(f)$. The *directional derivative (https://en.wikipedia.org/wiki/Directional_derivative)* of the function f at the point $\bar{x}$ in the direction $d \in \mathbb{R}^n$ is given by

$$f'(\bar{x}; d) := \lim_{t \to 0^+} \frac{f(\bar{x} + td) - f(\bar{x})}{t}.$$

If this limit exists as a real number or $\pm\infty$, then it is called the right directional derivative. We can similarly define the *left directional derivative* as

$$f'_-(\bar{x}; d) := \lim_{t \to 0^-} \frac{f(\bar{x} + td) - f(\bar{x})}{t}.$$

It is easy to see from the definitions that

$$f'_-(\bar{x}; d) = -f'(\bar{x}; -d) \quad \text{for all} \quad d \in \mathbb{R}^n,$$

Lemma 13.1 Given a convex function $f : \mathbb{R}^n \to (-\infty, \infty]$ with $\bar{x} \in \operatorname{dom}(f)$ and given $d \in \mathbb{R}^n$, define

$$\phi(t) := \frac{f(\bar{x}+td)-f(\bar{x})}{t}, \quad t>0.$$

Then the function ϕ is nondecreasing on $(0,\infty)$.

Proof Fix any numbers $0<t_1<t_2$ and get the representation

$$\bar{x}+t_1d=\frac{t_1}{t_2}\Big(\bar{x}+t_2d\Big)+\Big(1-\frac{t_1}{t_2}\Big)\bar{x}.$$

It follows from the convexity of f that

$$f(\bar{x}+t_1d)\le\frac{t_1}{t_2}f(\bar{x}+t_2d)+\Big(1-\frac{t_1}{t_2}\Big)f(\bar{x}),$$

which implies in turn the inequality

$$\phi(t_1) = \frac{f(\bar{x} + t_1 d) - f(\bar{x})}{t_1} \leq \frac{f(\bar{x} + t_2 d) - f(\bar{x})}{t_2} = \phi(t_2).$$

This shows us that ϕ is nondecreasing on $(0, \infty)$. $\square$

Proposition 13.2 For any convex function $f : \mathbb{R}^n \to (-\infty, \infty]$ and any $\bar{x} \in \mathrm{dom}(f)$, the directional derivative $f'(\bar{x}; d)$ exists for every direction $d \in \mathbb{R}^n$. Furthermore, if $\bar{x} \in \mathrm{int}(\mathrm{dom} f)$ then $f'(\bar{x}; d)$ is a real number for every $d \in \mathbb{R}^n$.

Proof From what we showed in Lemma 13.1, we have that the function ϕ is nondecreasing. Thus we have

$$f'(\bar{x}; d) = \lim_{t \to 0+} \frac{f(\bar{x} + td) - f(\bar{x})}{t} = \lim_{t \to 0^+} \phi(t) = \inf_{t > 0} \phi(t),$$

thus $f'(\bar{x}; d)$ is either a real number or $\pm\infty$.

Now let $\bar{x} \in \operatorname{int}(\operatorname{dom} f)$. This implies that f is locally Lipschitz around $\bar{x}$.
Thus there exists an $\ell \geq 0$ and a $\delta > 0$ such that

$$\big|f(x) - f(y)\big| \leq \ell \|x - y\|$$

whenever $x, y \in \mathbb{B}(\bar{x}; \delta)$. Then for t sufficiently small we have

$$\big|f(\bar{x} + td) - f(\bar{x})\big| \leq \ell t \|d\|.$$

Dividing both sides by t gives us that for sufficiently small t,

$$\Big|\frac{f(\bar{x} + td) - f(\bar{x})}{t}\Big| \leq \ell \|d\|.$$

Thus when we take the limit as $t \to 0^+$, we get

$$f'(\bar{x}; d) \leq \ell \|d\|.$$

Thus the directional derivative is a real number as desired. $\square$

Lemma 13.3 Let $f\colon \mathbb{R}^n \to (-\infty,\infty]$ be a convex function with $\bar{x} \in \operatorname{dom}(f)$. Then we have

$$f'(\bar{x};d) \le f(\bar{x}+d) - f(\bar{x}) \ \text{ whenever } \ d \in \mathbb{R}^n.$$

Proof Using Lemma 13.1, we have for the function ϕ defined in the lemma that

$$\phi(t) \le \phi(1) = f(\bar{x}+d) - f(\bar{x}) \ \text{ for all } \ t \in (0,1),$$

which justifies the claimed property due to the fact that $f'(\bar{x};d) = \inf\limits_{t>0}\phi(t) \le \phi(1)$. $\square$

Theorem 13.4 Let $f\colon \mathbb{R}^n \to (-\infty,\infty]$ be convex with $\bar{x} \in \operatorname{dom}(f)$. The following are equivalent:

$(\mathbf{i})\, v \in \partial f(\bar{x})$;
$(\mathbf{ii})\, \langle v, d\rangle \le f'(\bar{x}; d)$ for all $d \in \mathbb{R}^n$;
$(\mathbf{iii})\, f'_-(\bar{x}; d) \le \langle v, d\rangle \le f'(\bar{x}; d)$ for all $d \in \mathbb{R}^n$.

Proof

$(\mathbf{i}) \Longrightarrow (\mathbf{ii})$ Picking any $v \in \partial f(\bar{x})$ and $t > 0$, we get

$$\langle v, td\rangle \le f(\bar{x} + td) - f(\bar{x}) \text{ whenever } d \in \mathbb{R}^n.$$

By taking the limit as $t \to 0^+$ we arrive at our desired result.

$(\mathbf{ii}) \Longrightarrow (\mathbf{i})$ Assuming now that assertion $(\mathbf{ii})$ holds, we get by the previous lemma that

$$\langle v, d\rangle \le f'(\bar{x}; d) \le f(\bar{x} + d) - f(\bar{x}) \text{ for all } d \in \mathbb{R}^n.$$

It ensures by the definition of the subdifferential that $v \in \partial f(\bar{x})$, and thus assertions $(\mathbf{i})$ and $(\mathbf{ii})$ are equivalent. It is obvious that $(\mathbf{iii})$ implies $(\mathbf{ii})$, so we will only prove the reverse implication.

$(\mathbf{ii}) \Longrightarrow (\mathbf{iii})$ Assume that $(\mathbf{ii})$ is satisfied, then for $d \in \mathbb{R}^n$ we have

$$\langle v, -d\rangle \le f'(\bar{x}; -d),$$

and thus

$$f'_-(\bar{x}; d) = -f'(\bar{x}; -d) \le \langle v, d\rangle \ \text{ for any } \ d \in \mathbb{R}^n.$$

This justifies the validity of $(\mathbf{iii})$ and completes the proof of the theorem. $\square$

We will now prove some properties of the directional derivative as a function of the direction.

Proposition 13.5 Let $f\colon \mathbb{R}^n \to (-\infty, \infty]$ be a convex function with $\bar{x} \in \operatorname{dom}(f)$, we define the function $\psi\colon \mathbb{R}^n \to (-\infty, \infty]$ by $\psi(d) := f'(\bar{x}; d)$, which satisfies the following properties:

$(\mathbf{i})\, \psi(d_1 + d_2) \leq \psi(d_1) + \psi(d_2)$ for all $d_1, d_2 \in \mathbb{R}^n$.
$(\mathbf{ii})\, \psi(\alpha d) = \alpha\psi(d)$ whenever $d \in \mathbb{R}^n$ and $\alpha > 0$.
$(\mathbf{iii})\, \partial f(\bar{x}) = \partial\psi(0)$.

Proof $(\mathbf{i})$ Fix $d_1, d_2 \in \mathbb{R}^n$. Then we have that

$$\psi(d_1 + d_2) = \lim_{t \to 0^+} \frac{f(\bar{x} + t(d_1 + d_2)) - f(\bar{x})}{t}$$

$$= \lim_{t\to 0^+} \frac{f(\frac{\bar{x}+2td_1+\bar{x}+2td_2}{2}) - f(\bar{x})}{t}$$

$$\leq \lim_{t\to 0^+} \frac{f(\bar{x}+2td_1)-f(\bar{x})}{2t} + \lim_{t\to 0^+} \frac{f(\bar{x}+2td_2)-f(\bar{x})}{2t} = \psi(d_1)+\psi(d_2).$$

(**ii**) Now fix $\alpha > 0$ and $d \in \mathbb{R}^n$. Then using the definition of directional derivative we get

$$\psi(\alpha d) = \lim_{t\to 0^+} \frac{f(\bar{x}+t(\alpha d)) - f(\bar{x})}{t}$$

$$= \lim_{t\to 0^+} \alpha \frac{f\big(\bar{x}+(\alpha t)d\big) - f(\bar{x})}{(\alpha t)} = \alpha\,\psi(d).$$

(**iii**) Recall from Theorem 13.4 that $v \in \partial f(\bar{x})$ if and only if

$$\langle v, d\rangle \leq f'(\bar{x}, d) = \psi(d) \text{ for all } d \in \mathbb{R}^n.$$

It can be easily shown that $\psi(0) = 0$, thus we have that the inequality above is equivalent to

$$\langle v, d - 0 \rangle \leq \psi(d) - \psi(0) \text{ for all } d \in \mathbb{R}^n.$$

Applying the definition of the subdifferential to the above inequality tells us that $v \in \partial\psi(0)$. Thus we have shown that $v \in \partial f(\bar{x})$ if and only if $v \in \partial\psi(0)$, which completes the proof. $\square$

Lemma 13.6 Let the function $\psi \colon \mathbb{R}^n \to \mathbb{R}$ be subadditive and positively homogeneous. Then

$$\psi^*(v) = \delta_\Omega(v) \text{ for all } v \in \mathbb{R}^n, \text{ where } \Omega := \partial\psi(0).$$

Proof First we will show that $\psi^*(v) = 0$ for all $v \in \Omega = \partial\psi(0)$. Fixing $v \in \Omega$, we have

$$\psi^*(v) = \sup\big\{\langle v, d\rangle - \psi(d) \;\big|\; d \in \mathbb{R}^n\big\} \geq \langle v, 0\rangle - \psi(0) = 0.$$

We also have

$$\langle v, d\rangle = \langle v, d - 0\rangle \leq \psi(d) - \psi(0) = \psi(d), \text{ for all } d \in \mathbb{R}^n$$

which gives us that

$$\psi^*(v) = \sup\big\{\langle v, d\rangle - \psi(d) \;\big|\; d \in \mathbb{R}^n\big\} \leq 0$$

and so ensures the validity of $\psi^*(v) = 0$ for any $v \in \partial\psi(0)$.

It remains to verify that $\psi^*(v) = \infty$ if $v \notin \partial\psi(0)$. For such an element v, find $d_0 \in \mathbb{R}^n$ with $\langle v, d_0\rangle > \psi(d_0)$. Since ψ is positively homogeneous by assumption, it follows that

$$\psi^*(v) = \sup\big\{\langle v, d\rangle - \psi(d) \;\big|\; d \in \mathbb{R}^n\big\} \geq \sup_{t>0}(\langle v, td_0\rangle - \psi(td_0))$$

$$= \sup_{t>0} t(\langle v, d_0\rangle - \psi(d_0)) = \infty.$$

Thus we have shown that $\psi^*(v) = \delta_\Omega(v)$ for any $v \in \mathbb{R}^n$. $\square$

The next theorem gives us a characterization for the directional derivative in terms of the subdifferential.

Theorem 13.5 Let $f\colon \mathbb{R}^n \to (-\infty, \infty]$ be a convex function and let $\bar{x} \in \operatorname{int}(\operatorname{dom} f$. Then

$$f'(\bar{x}; d) = \sup\{\langle v, d\rangle \colon v \in \partial f(\bar{x})\}$$

$$= \max\{\langle v, d\rangle \colon v \in \partial f(\bar{x})\}.$$

Proof Consider $\psi\colon \mathbb{R}^n \to \mathbb{R}$ given by $\psi(d) = f'(\bar{x}; d)$. Then ψ is subadditive and positively homogeneous. Furthermore we have shown that $\partial f(\bar{x}) = \partial\psi(0)$ and from the previous lemma we have that $\psi^* = \delta_\Omega$ where $\Omega = \partial f(\bar{x})$. Then given any $d \in \mathbb{R}^n$, we have

$$\psi(d) = \psi^{**}(d) = \delta_\Omega{}^*(d)$$

$$= \sup\big\{\langle v, d\rangle - \delta_\Omega(v) \colon v \in \mathbb{R}^n\big\}$$

$$= \sup\big\{\langle v, d\rangle \colon v \in \Omega\big\}$$

$$= \sup\big\{\langle v, d\rangle \colon v \in \partial f(\bar{x})\big\}$$

$$= \max\big\{\langle v, d\rangle \colon v \in \partial f(\bar{x})\big\}.$$

Thus we have proved that $\psi(d) = f'(\bar{x}; d) = \max\big\{\langle v, d\rangle \colon v \in \partial f(\bar{x})\big\}$ as desired. $\square$

14 Subdifferential Characterization for Differentiability

Definition Let $f\colon \mathbb{R}^n \to (-\infty, \infty]$ be a function and let $\bar{x} \in \operatorname{int}(\operatorname{dom} f)$. We say that f is *Gâteaux differentiable (https://en.wikipedia.org/wiki/G%C3%A2teaux_derivative)* at $\bar{x}$ if there exists a $v \in \mathbb{R}^n$ such that

$$\lim_{t \to 0} \frac{f(\bar{x} + td) - f(\bar{x}) - t\langle v, d\rangle}{t} = 0 \text{ for all } d \in \mathbb{R}^n.$$

If f is Gâteaux differentiable at $\bar{x}$, then the element v is unique and is called the Gâteaux derivative, which we denote by $f'_G(\bar{x})$.

Proposition 14.1 Let $f\colon \mathbb{R}^n \to (-\infty, \infty]$ be a function and let $\bar{x} \in \operatorname{int}(\operatorname{dom} f)$. Then f is Gâteaux differentiable at $\bar{x}$ if and only if the directional derivative $f'(\bar{x}; d)$ exists and

$$f'(\bar{x}; d) = \langle v, d\rangle \text{ for all } d \in \mathbb{R}^n$$

where $v = f'_G(\bar{x})$.

Proof First assume that $v = f'_G(\bar{x})$ exists. Then from the definition we have that

$$\langle v, d \rangle = \lim_{t \to 0} \frac{f(\bar{x} + td) - f(\bar{x})}{t} \text{ for all } d \in \mathbb{R}^n.$$

If the limit exists, then it is equivalent to taking the limit from the right, so we have

$$\langle v, d \rangle = \lim_{t \to 0^+} \frac{f(\bar{x} + td) - f(\bar{x})}{t} \text{ for all } d \in \mathbb{R}^n.$$

From the definition of directional derivatives, this shows us that $f'(\bar{x}; d)$ exists for all $d \in \mathbb{R}^n$ and that $\langle v, d \rangle = f'(\bar{x}; d)$.

To prove the other implication, suppose that $f'(\bar{x};d)$ exists and that $f'(\bar{x};d) = \langle v, d\rangle$. Notice that $f'(\bar{x};-d) = \langle v, -d\rangle$ for all $d \in \mathbb{R}^n$, which implies that $-f'(\bar{x};-d) = f_-{}'(\bar{x};d) = \langle v, d\rangle$ for all $d \in \mathbb{R}^n$. Thus we have

$$\langle v, d\rangle = \lim_{t\to 0^+} \frac{f(\bar{x}+td) - f(\bar{x})}{t} \text{ for all } d \in \mathbb{R}^n,$$

and we also get

$$\langle v, d\rangle = \lim_{t\to 0^-} \frac{f(\bar{x}+td) - f(\bar{x})}{t} \text{ for all } d \in \mathbb{R}^n.$$

Therefore we have shown that the limit from both the right and the left are equivalent and thus the Gâteaux derivative exists, completing the proof. □

Definition Let $f\colon \mathbb{R}^n \to (-\infty, \infty]$ be a function and let $\bar{x} \in \operatorname{int}(\operatorname{dom} f)$. f is said to be *Fréchet differentiable* (https://en.wikipedia.org/wiki/Fr%C3%A9chet_derivative) at $\bar{x}$ if there exists

a $v \in \mathbb{R}^n$ such that

$$\lim_{h \to 0} \frac{f(\bar{x} + h) - f(\bar{x}) - \langle v, h \rangle}{\|h\|} = 0.$$

The element v is called the Fréchet derivative of f at $\bar{x}$, and it is denoted by $f'(\bar{x})$.

In general Gâteaux differentiability is not in general equivalent to Fréchet differentiability, however under certain conditions it is as we shall see in the next theorem.

Theorem 14.2 Let $f : \mathbb{R}^n \to (-\infty, \infty]$ and let $\bar{x} \in \operatorname{int}(\operatorname{dom} f)$. If f is Fréchet differentiable at $\bar{x}$ then it is also Gâteaux differentiable at $\bar{x}$. Furthermore if f is convex then the converse is also true.

Proof Suppose that f is Fréchet differentiable at $\bar{x}$ and let $v = f'(\bar{x})$. Then

$$\lim_{h\to 0}\frac{f(\bar{x}+h)-f(\bar{x})-\langle v,h\rangle}{\|h\|}=0.$$

Fix any $0\neq d\in\mathbb{R}^n$ and let $h=td$. Then we have that

$$\lim_{t\to 0}\frac{f(\bar{x}+td)-f(\bar{x})-t\langle v,d\rangle}{t}$$

$$=\lim_{t\to 0}\Big(\frac{f(\bar{x}+td)-f(\bar{x})-\langle v,td\rangle}{t\|d\|}\Big)\|d\|=0.$$

Thus we have that f is Gâteaux differentiable.

Now suppose that $f\colon\mathbb{R}^n\to(-\infty,\infty]$ is convex and Gâteaux differentiable with $v=f_G{}'$ and we shall prove that f is Fréchet differentiable. Suppose by contradiction that f is not Fréchet differentiable. Then there exists a $\epsilon_0>0$ and a sequence $\{h_k\}$ with $h_k\to 0$ such that

$$\epsilon_0 \leq \frac{\left|f(\bar{x}+h_k) - f(\bar{x}) - \langle v, h_k\rangle\right|}{\|h_k\|}.$$

Define $t_k = \|h_k\|$ and let $d_k = \dfrac{h_k}{\|h_k\|}$. Then you will notice that $t_k \to 0$ as $k \to \infty$ and also that $\|d_k\| = 1$ for all $k \in \mathbb{N}$. Thus $\{d_k\}$ is a bounded sequence which implies it has a convergent subsequence. Without loss of generality suppose that $d_k \to d$ as $k \to \infty$. Rewriting the inequality above we get

$$\epsilon_0 \leq \frac{\left|f(\bar{x}+h_k) - f(\bar{x}) - \langle v, h_k\rangle\right|}{\|h_k\|}$$

$$= \frac{\left|f(\bar{x}+\|h_k\|\frac{h_k}{\|h_k\|}) - f(\bar{x}) - \langle v, \|h_k\|\frac{h_k}{\|h_k\|}\rangle\right|}{\|h_k\|}$$

$$= \frac{\big|f(\bar{x}+t_k d_k) - f(\bar{x}) - \langle v, t_k d_k\rangle\big|}{t_k}$$

$$= \frac{\big|f(\bar{x}+t_k d_k) - f(\bar{x}+t_k d) + f(\bar{x}+t_k d) - f(\bar{x}) - \langle v, t_k d_k\rangle + \langle v, t_k d\rangle - \langle v, t_k d\rangle\big|}{t_k}$$

$$\leq \frac{\big|f(\bar{x}+t_k d_k) - f(\bar{x}+t_k d)\big|}{t_k} + \frac{\big|f(\bar{x}+t_k d) - f(\bar{x}) - \langle v, t_k d\big|}{t_k} + \frac{\big|\langle v, t_k d_k\rangle - \langle v, t_k d\rangle\big|}{t_k}. \quad (1)$$

Now, because $\bar{x} \in \operatorname{int}(\operatorname{dom} f)$, then f is locally Lipchitz continuous around $\bar{x}$. Thus there exists a $\ell \geq 0$ and a $\delta > 0$ such that

$$\big|f(x) - f(u)\big| \leq \ell \|x - u\| \text{ for all } x, u \in \mathbb{B}(\bar{x}; u).$$

Applying this fact and the Cauchy-Schwarz inequality to (1) gives us

$$\epsilon_0 \leq \frac{\ell\|(\bar{x}+t_k d_k)-(\bar{x}+t_k d)\|}{t_k}+\frac{\left|f(\bar{x}+t_k d)-f(\bar{x})-\langle v, t_k d\right|}{t_k}+\frac{\|v\|\|t_k d_k - t_k d\|}{t_k}$$

$$= \ell\|(d_k - d)\| + \frac{\left|f(\bar{x}+t_k d)-f(\bar{x})-\langle v, t_k d\right|}{t_k} + \|v\|\|d_k - d\|. \quad (2)$$

Using the Gâteaux differentiability of f at $\bar{x}$ on the middle term we can see that (2) goes to 0 as $k \to \infty$. But this contradicts that fact that (2) was greater than ϵ_0 for all $k \in \mathbb{N}$. Thus we have completed the proof. $\square$

Theorem 14.4 Let $f : \mathbb{R}^n \to (-\infty, \infty]$ be a convex function and let $\bar{x} \in \operatorname{int}(\operatorname{dom} f)$. Then the following are equivalent:

(**i**) f is Fréchet differentiable at $\bar{x}$.
(**ii**) f is Gâteaux differentiable at $\bar{x}$.
(**iii**) $\partial f(\bar{x})$ is a singleton.

Proof (**i**) $\iff$ (**ii**) was shown in the last theorem so we only need to show that (**ii**) $\iff$ (**iii**).

$\big($(**ii**) $\implies$ (**iii**)$\big)$ Suppose that f is Gâteaux differentiable at $\bar{x}$ and let $v = {f'}_G(\bar{x})$. Then

$$f'(\bar{x}; d) = \langle v, d \rangle = {f'}_-(\bar{x}; d) \text{ for all } d \in \mathbb{R}^n.$$

Fix any $w \in \partial f(\bar{x})$. Then

$$f'_-(\bar{x};d) \leq \langle w, d\rangle \leq f'(\bar{x};d), \text{ for all } d \in \mathbb{R}^n.$$

Thus for any $d \in \mathbb{R}^n$ we have that $\langle v, d\rangle = \langle w, d\rangle$. This implies that

$$\langle v - w, d\rangle = 0, \text{ for all } d \in \mathbb{R}^n.$$

Choosing $d = v - w$ gives us $\|w - v\|^2 = 0$ which implies that $v = w$. Therefore $\partial f(\bar{x}) = \{v\}$ and thus is a singleton.

$\big((\textbf{iii}) \implies (\textbf{ii})\big)$ Suppose that $\partial f(\bar{x}) = \{v\}$. Then we have that

$$f'(\bar{x};d) = \max\big\{\langle w, d\rangle \colon w \in \partial f(\bar{x})\big\} = \langle v, d\rangle, \text{ for all } d \in \mathbb{R}^n.$$

By the relationship between Gâteaux differentiability and directional differentiability established in Lemma 1.1 we can see that f is Gâteaux differentiable and that $f'_G(\bar{x}) = v$. $\square$

15 Lower Semicontinuity and the Existence of Minimizers

Definition Let $f\colon \mathbb{R}^n \to (-\infty, \infty]$ be a function. We say that f is *lower semicontinuous (https://en.wikipedia.org/wiki/Semi-continuity)*, or l.s.c., at $\bar{x}$ if for any $\lambda < f(\bar{x})$, there exists a $\delta > 0$ such that

$$\lambda < f(x) \text{ for all } x \in \mathbb{B}(\bar{x}; \delta).$$

If f is l.s.c. at all $\bar{x} \in \mathbb{R}^n$, then we say that f is l.s.c.

Proposition 15.1 Let $f\colon \mathbb{R}^n \to (-\infty, \infty]$ be a function and let $\bar{x} \in \operatorname{dom}(f)$. Then f is l.s.c. at $\bar{x}$ if and only if for any $\epsilon > 0$, there exists a $\delta > 0$ such that

$$f(\bar{x}) - \epsilon < f(x) \text{ for all } x \in \mathbb{B}(\bar{x}; \delta).$$

Proof First assume that f is l.s.c. at $\bar{x}$. Then, letting $\epsilon > 0$, we have that

$$\lambda := f(\bar{x}) - \epsilon < f(\bar{x}).$$

By the definition of lower semicontinuity there exists a $\delta > 0$ such that

$$\lambda = f(\bar{x}) - \epsilon < f(x) \text{ for all } x \in \mathbb{B}(\bar{x}; \delta),$$

which proves the first implication. Now suppose that for any $\epsilon > 0$ there exists a $\delta > 0$ such that

$$f(\bar{x}) - \epsilon < f(x) \text{ for all } x \in \mathbb{B}(\bar{x}; \delta).$$

Take $\lambda < f(\bar{x})$. We can choose a $\epsilon > 0$ such that $\lambda < f(\bar{x}) - \epsilon$. Then there exists a $\delta > 0$ such that

$$\lambda < f(\bar{x}) - \epsilon < f(x) \text{ for all } x \in \mathbb{B}(\bar{x}; \delta).$$

Thus we have that f is l.s.c. $\square$

Proposition 15.2 Let $f : \mathbb{R}^n \to (-\infty, \infty]$ be a function. Then f is l.s.c. if and only if for any $\lambda > 0$, the set $\mathcal{L}_\lambda$ is closed where

$$\mathcal{L}_\lambda := \big\{x \in \mathbb{R}^n \colon f(x) \leq \lambda\big\}.$$

Proof First suppose that f is l.s.c. and fix $\lambda \in \mathbb{R}^n$. We will show that $\mathcal{L}_\lambda$ is closed by proving that the compliment, $(\mathcal{L}_\lambda)^c$, is open. Take any $\bar{x} \in (\mathcal{L}_\lambda)^c$, then by definition we have that $\lambda < f(\bar{x})$. Because f is l.s.c at $\bar{x}$, there exists a $\delta > 0$ such that

$$\lambda < f(x) \text{ for all } x \in \mathbb{B}(\bar{x}; \delta).$$

Thus $\mathbb{B}(\bar{x}; \delta) \subset (\mathcal{L}_\lambda)^c$, and since this is true for any $\bar{x} \in (\mathcal{L}_\lambda)^c$ we have verified that $(\mathcal{L}_\lambda)^c$ is open which in turn implies that $\mathcal{L}_\lambda$ is closed.

To prove the other implication, suppose that $\mathcal{L}_\lambda$ is closed for any $\lambda \in \mathbb{R}^n$, and we will show that f is lower semicontinuous. Let $\bar{x} \in \mathbb{R}^n$ and let $\lambda < f(\bar{x})$. Then we have that $\bar{x} \in (\mathcal{L}_\lambda)^c$. But by assumption $(\mathcal{L}_\lambda)^c$ is open so there must exists a $\delta > 0$ such that $\mathbb{B}(\bar{x}; \delta) \subset (\mathcal{L}_\lambda)^c$. Thus we have

$$\lambda < f(x) \text{ for all } x \in \mathbb{B}(\bar{x}; \delta).$$

This shows us that is l.s.c. as desired.

Proposition 15.3 Let $f\colon \mathbb{R}^n \to (-\infty, \infty]$be a function. Then $f$is l.s.c. if and only if $\operatorname{epi}(f)$is closed.

Proof First assume that $f$is l.s.c. and we will show that $\operatorname{epi}(f)$is closed by showing that $\big(\operatorname{epi}(f)\big)^c$is open. Take any $(\bar{x}, \lambda) \in \big(\operatorname{epi}(f)\big)^c$. Then from the definition of the epigraph, we get that $\lambda < f(\bar{x})$ Choose $\epsilon > 0$small enough such that $\lambda + \epsilon < f(\bar{x})$ By the lower semicontinuity of f, there exists a $\delta > 0$ such that

$$\lambda + \epsilon < f(x) \text{ for all } x \in \mathbb{B}(\bar{x}; \delta).$$

Then it can easily be seen that

$$\mathbb{B}(\bar{x}; \delta) \times (\lambda - \epsilon, \lambda + \epsilon) \subset \big(\operatorname{epi}(f)\big)^c.$$

Since such an open ball exists for any $(\bar{x}, \lambda) \in \big(\operatorname{epi}(f)\big)^c$, then we have that $\big(\operatorname{epi}(f)\big)^c$is open, which proves that $\operatorname{epi}(f)$is closed.

To prove the converse we assume that $\text{epi}(f)$ is closed. Taking any $\lambda \in \mathbb{R}$ we will show that

$$\mathcal{L}_\lambda := \{x \in \mathbb{R}^n : f(x) \leq \lambda\}$$

is closed, which by Proposition 15.2 will give us the lower semicontinuity of f. Fix any sequence $\{x_k\}$ in $\mathcal{L}_\lambda$ that converges to some point $\bar{x} \in \mathbb{R}^n$. Then $f(x_k) \leq \lambda$ for all k which shows us that $(x_k, \lambda) \in \text{epi}(f)$ and we also have that $(x_k, \lambda) \to (\bar{x}, \lambda)$ as $k \to \infty$. By the closure of $\text{epi}(f)$ we have that $(\bar{x}, \lambda) \in \text{epi}(f)$. Thus $f(\bar{x}) \leq \lambda$ which gives us that $\bar{x} \in \mathcal{L}_\lambda$. Thus we have completed the proof. □

Proposition 15.4 Let $f : \mathbb{R}^n \to (-\infty, \infty]$ be a function. Then f is l.s.c. at $\bar{x}$ if and only if for any sequence in $\mathbb{R}^n$, $\{x_k\}$ with $x_k \to \bar{x}$, we have that

$$\liminf_{k \to \infty} f(x_k) \geq f(\bar{x}).$$

Proof Suppose that f is l.s.c. at $\bar{x}$ and take any sequence such that $x_k \to \bar{x}$ as $k \to \infty$. Fix any $\lambda < f(\bar{x})$. Then there exists $\delta > 0$ such that

$$\lambda < f(x) \text{ for all } x \in \mathbb{B}(\bar{x}; \delta).$$

Since $x_k \to \bar{x}$ then there exists a $k_0 \in \mathbb{N}$ such that $x_k \in \mathbb{B}(\bar{x}; \delta)$ for all $k \geq k_0$. Then $\lambda < f(x_k)$ for all $k \geq k_0$, which in turn gives us that

$$\lambda \leq \liminf_{k \to \infty} f(x_k),$$

and since this is true for all $\lambda < f(\bar{x})$, then we have that

$$f(\bar{x}) \leq \liminf_{k \to \infty} f(x_k).$$

To prove the converse let $\liminf_{k \to \infty} f(x_k) \geq f(\bar{x})$ for any sequence $\{x_k\}$ which converges to $\bar{x}$. Suppose by contradiction that l.s.c. at $\bar{x}$. Then there exists a $\lambda < f(\bar{x})$ such that for any $\delta > 0$, there exists a $x_\delta \in \mathbb{B}(\bar{x}; \delta)$ with $\lambda > f(x_\delta)$.

Therefore letting $\delta = \frac{1}{k}$, we have that there exists a sequence $x_k \in \mathbb{B}(\bar{x}; \frac{1}{k})$ such that $\lambda \geq f(x_k)$. Thus $x_k \to \bar{x}$ and

$$\liminf_{k \to \infty} f(x_k) \leq \lambda \leq f(\bar{x}).$$

This contradicts our previous assumption completing the proof. $\square$

Definition Let $f : \mathbb{R}^n \to (-\infty, \infty]$ be a function with nonempty domain. An element $\overline{x} \in \mathrm{dom}(f)$ is called an *absolute minimizer* of f if

$$f(\overline{x}) \leq f(x) \text{ for all } x \in \mathbb{R}^n.$$

Theorem 15.5 Let $f : \mathbb{R}^n \to (-\infty, \infty]$ be lower semicontinuous with $\mathrm{dom}(f) \neq \emptyset$. Suppose that for any $\lambda \in \mathbb{R}$ the sublevel set

$$\mathcal{L}_\lambda = \{x \in \mathbb{R}^n : f(x) \leq \lambda\}$$

is bounded. Then f has an absolute minimizer.

Proof Let $\alpha = \inf_{x \in \mathbb{R}^n} f(x)$. Because the domain of f is nonempty, we have that $\alpha < \infty$. Then there exists a sequence $\{x_n\} \subset \mathbb{R}^n$ such that

$$\lim_{n \to \infty} f(x_n) = \alpha.$$

So there exists a $\lambda \in \mathbb{R}$ and $k_0 \in \mathbb{N}$ such that

$$f(x_k) \leq \lambda \text{ for all } k \geq k_0.$$

Then $x_k \in \mathcal{L}_\lambda$ for all $k \geq k_0$. Since $\mathcal{L}_\lambda$ is bounded then the sequence $\{x_k\}$ is a bounded sequence and thus has a convergent subsequence, $\{x_{kl}\}$, which converges to some point, $\overline{x} \in \mathbb{R}^n$. Thus we have

$$\begin{aligned} f(\overline{x}) &\leq \liminf_{l\to\infty} f(x_{kl}) = \lim_{l\to\infty} f(x_{kl}) \\ &= \alpha = \inf_{x\in\mathbb{R}^n} f(x) \\ &\leq f(x) \text{ for all } x \in \mathbb{R}^n. \end{aligned}$$

Thus $\overline{x}$ is an absolute minimizer. $\square$

Next we will examine the Constrained Optimization Problem. Let $f : \mathbb{R}^n \to \mathbb{R}$ and let Ω be a nonempty subset of $\mathbb{R}^n$. Consider the problem:

Minimize $f(x)$ $\qquad\qquad\qquad\qquad\qquad\qquad \Big(\mathcal{P}\Big)$

Subject to $x \in \Omega$

Any $\overline{x} \in \Omega$, is called a *feasible solution* of $\Big(\mathcal{P}\Big)$. Any $\overline{x} \in \Omega$ such that $f(\overline{x}) \leq f(x)$ for all $x \in \Omega$ is called a *global optimal solution* of $\Big(\mathcal{P}\Big)$.

Theorem 15.6 Consider the problem $\Big(\mathcal{P}\Big)$. Suppose Ω is a compact set and f is l.s.c., then $\Big(\mathcal{P}\Big)$ has a global optimal solution.

Proof Let $g : \mathbb{R}^n \to (-\infty, \infty]$ defined by

$$g(x) = f(x) + \delta(x; \Omega)$$

where δ is the indicator function. Then g is lower semicontinuous.
Moreover, for any $\lambda \in \mathbb{R}$, the set

$$\begin{aligned}\mathcal{L}^g{}_\lambda &:= \{x \in \mathbb{R}^n : g(x) \leq \lambda\} \\ &= \Omega \bigcap \{x \in \mathbb{R}^n : f(x) \leq \lambda\} \\ &= \Omega \bigcap \mathcal{L}_\lambda.\end{aligned}$$

Since Ω is compact then $\mathcal{L}^g{}_\lambda$ is bounded. Then g has an absolute minimizer denote by $\overline{x}$. Then we have

$$\begin{aligned}g(\overline{x}) &= f(\overline{x}) + \delta(\overline{x}; \Omega) \\ &\leq f(x) + \delta(x; \Omega) \text{ for all } x \in \mathbb{R}^n\end{aligned}$$

Let $u \in \Omega$ then

$$g(\overline{x}) = f(\overline{x}) + \delta(\overline{x}; \Omega)$$

$$f(u) + \delta(u; \Omega) = f(u) \leq \infty.$$

Thus $\overline{x} \in \Omega$ and we have that $f(\overline{x}) \leq f(u)$ for all $u \in \Omega$ which proves that $\overline{x}$ is a global optimal solution. $\square$

Remark It follows from the previous theorem, that $\overline{x} \in \Omega$ is a global optimal solution if and only if $\overline{x}$ is an absolute minimizer for the function $g : \mathbb{R}^n \to (-\infty, \infty]$ given by

$$g(x) = f(x) + \delta(x; \Omega).$$

Corollary 15.7 Let $f : \mathbb{R}^n \to \mathbb{R}$ be a l.s.c. function. Suppose that $\lim_{\|x\| \to \infty} f(x) = \infty$. Then f has an absolute minimizer.

Proof To prove this we only need to show that the sublevel set is bounded for any $\lambda \in \mathbb{R}$. Fix $\lambda \in \mathbb{R}$. Since $\lim_{\|x\| \to \infty} f(x) = \infty$, there exists an $r > 0$ such that

$$\Big[x \in \mathbb{R}^n\,,\ \|x\| > r\Big] \implies \Big[f(x) > \lambda\Big].$$

Then

$$\mathcal{L}_\lambda \subset \{x \in \mathbb{R}^n : \|x\| \leq r\} = \mathbb{B}(0; r).$$

This proves that $\mathcal{L}_\lambda$ is bounded which completes the proof. $\square$.

16 Optimality Conditions and the Methods of Lagrange Multipliers

Theorem 16.1 Let $f : \mathbb{R}^n \to (-\infty, \infty]$ be a convex function and let $\overline{x} \in \mathrm{dom}(f)$. Then f has an absolute minimum at $\overline{x}$ if and only if f has a local minimum at $\overline{x}$.

Proof Notice that if $\overline{x}$ is an absolute minimum then it is a local minimum so we only need to prove the opposite implication. Let f have a local minimum at $\overline{x}$. Then by the definition of local minimum there exists a $\delta > 0$ such that

$$f(\overline{x}) \leq f(x) \text{ for all } x \in B(\overline{x}; \delta).$$

Fix any $u \in \mathbb{R}^n$. Then we can find a $0 < t < 1$ such that

$$x := \overline{x} + t(u - \overline{x}) = tu + (1 - t)\overline{x} \in B(\overline{x}; \delta).$$

Then we have

$$f(\overline{x}) \leq f(x) = f\big(tu + (1 - t)\overline{x}\big)$$

$$\leq tf(u) + (1 - t)f(\overline{x}) = f(\overline{x}) + t\big(f(u) - f(\overline{x})\big).$$

Subtracting both sides of the inequality by $f(\overline{x})$ gives us

$$0 \leq t\big(f(u) - f(\overline{x})\big).$$

Dividing both sides by t we get,

$$0 \leq f(u) - f(\overline{x}).$$

Thus we get that $f(\overline{x}) \leq f(u)$ and since this is true for any $u \in \mathbb{R}^n$ then $\overline{x}$ is an absolute minimum for f. $\square$

Recall the Constrained Optimization Problem from the last section. Namely, letting $f : \mathbb{R}^n \to \mathbb{R}$ and Ω be a nonempty subset of $\mathbb{R}^n$

Minimize $f(x)$ $\qquad\qquad\qquad\qquad (\mathcal{P})$

Subject to $x \in \Omega$.

Theorem 16.2 Consider the problem $\Big(\mathcal{P}\Big)$ where f is a convex function and Ω is a nonempty convex set. Then $\overline{x}$ is an optimal solution of $\Big(\mathcal{P}\Big)$ if and only if

$$0 \in \partial f(\overline{x}) + N(\overline{x}; \Omega).$$

Proof We proved in the last section that $\overline{x}$ is an optimal solution of $\Big(\mathcal{P}\Big)$ if and only if $\overline{x}$ is an absolute minimizer of $g(x)$ where

$$g(x) := f(x) + \delta(x; \Omega).$$

This is true if and only if $0 \in \partial g(\overline{x})$. Then we have that

$$0 \in \partial g(\overline{x})$$

$$= \partial f(\overline{x}) + \partial\big(\delta(\overline{x}; \Omega)\big) = \partial f(\overline{x}) + N(\overline{x}; \Omega).$$

Thus $\overline{x}$ is an optimal solution of $\Big(\mathcal{P}\Big)$. $\square$

Corollary 16.3 Consider the problem $\Big(\mathcal{P}\Big)$ in which f is a convex function and

$$\Omega = \{x \in \mathbb{R}^n : Ax = b\},$$

where A is an $m \times n$ matrix and $b \in \mathbb{R}^n$. Let $\overline{x} \in \Omega$. Then $\overline{x}$ is an optimal solution of $\Big(\mathcal{P}\Big)$ if and only if

$$0 \in \partial f(\overline{x}) + A^\top(\mathbb{R}^m).$$

Proof It is easy to show that

$$N(\overline{x}; \Omega) = A^\top(\mathbb{R}^m) = \{A^\top \lambda : \lambda \in \mathbb{R}^m\}. \ \square$$

Consider the problem,

Minimize $f(x)$ $\quad\quad (\mathcal{P})$

Subject to $g_i(x) \leq 0$ for all $i = 1, ..., m$ and $x \in \Omega$.

where $f, g_1, ..., g_m : \mathbb{R}^n \to \mathbb{R}$ are convex functions and Ω is a nonempty convex subset of $\mathbb{R}^n$.

We say that $\overline{x} \in \mathbb{R}^n$ is a feasible solution of $(\mathcal{P})$ if $\overline{x} \in \Omega$ and $g_i(\overline{x}) \leq 0$ for all $i = 1, ..., m$. For $\overline{x} \in \mathbb{R}^n$ define

$$A(\overline{x}) = \{i = 1, ..., m : g_i(\overline{x}) = 0\}.$$

Theorem 16.4 Consider the problem $(\mathcal{P})$. Suppose that $\overline{x} \in \mathbb{R}^n$ is an optimal solution of $(\mathcal{P})$. Then there exists multipliers $\lambda_0, \lambda_1, ..., \lambda_m$ with at least one $i \in \{1, ..., m\}$ that has $\lambda_i \neq 0$, such that

$$0 \in \lambda_0 \partial f(\overline{x}) + \left(\sum_{i=1}^{m} \lambda_i \partial g_i(\overline{x}) \right) + N(\overline{x}; \Omega),$$

$\lambda_i \geq 0$ for all $i = 0, ..., m$, and furthermore $\lambda_i g_i(\overline{x}) = 0$ for all $i = 1, ..., m$.

Proof Let $\overline{x}$ be an optimal solution of $\left(\mathcal{P}\right)$. Define the function $\phi : \mathbb{R}^n \to \mathbb{R}$ by

$$\phi(x) = \max_{1 \leq i \leq m} \{f(x) - f(\overline{x}), g_i(x)\}.$$

Observe that $\phi(\overline{x}) = 0$, and $\overline{x}$ is an optimal solution of the problem,

Minimize $\phi(x)$

Subject to $x \in \Omega$.

Equivalently, $\phi(\overline{x}) \leq \phi(x)$ for all $x \in \Omega$. By the optimality condition,

$$0 \in \partial\phi(\overline{x}) + N(\overline{x}; \Omega)$$

Define $g_0(x) = f(x) - f(\overline{x})$. Then for $x \in \mathbb{R}^n$ we have $\phi(x) = \max\limits_{0 \leq i \leq 1}\{g_i(x)\}$. Recall from the section on the subdifferential maximum rule the definition of the index set,

$$I(\overline{x}) := \{i = 0, ..., m : g_i(x) = \phi(x) = 0\} = \{0\} \cup A(\overline{x}).$$

By the subdifferential formula for the maximum function, we have

$$0 \in \text{co}\big(\partial g_0(\overline{x}) \cup \Big[\bigcup_{i \in A(\overline{x})} \partial g_i(\overline{x})\Big]\big) + N(\overline{x}; \Omega)$$

$$= \text{co}\big(\partial f(\overline{x}) \cup \Big[\bigcup_{i \in A(\overline{x})} \partial g_i(\overline{x})\Big]\big) + N(\overline{x}; \Omega).$$

Then there exists a $\lambda_0 \geq 0$ and $\lambda_i \geq 0$ for $i \in A(\overline{x})$ with $\lambda_0 + \sum_{i \in A(\overline{x})} \lambda_i = 1$ and

$$0 \in \lambda_0 \partial f(\overline{x}) + \Big[\sum_{i \in A(\overline{x})} \lambda_i \partial g_i(\overline{x}) \Big] + N(\overline{x}; \Omega).$$

Set $\lambda_i = 0$ if $i \notin A(\overline{x})$ for $i \in \{1, ..., m\}$. Then $\lambda_i g_i(\overline{x}) = 0$ for all $i = 1, ..., m$. Then

$$0 \in \lambda_0 \partial f(\overline{x}) + \Big[\sum_{i=1}^{m} \lambda_i \partial g_i(\overline{x}) \Big] + N(\overline{x}; \Omega),$$

where $\lambda_i g_i(\overline{x}) = 0$ for all $i = 1, ..., m$, thus completing the proof. $\square$

Definition We say that the Slater condition is satisfied for the problem $\Big(\mathcal{P}\Big)$ if there exists a $u \in \Omega$ such that

$$g_i(u) < 0 \text{ for all } i = 1, ..., m.$$

Theorem 16.5 Consider problem $\Big(\mathcal{P}\Big)$ and let $\overline{x}$ be a feasible solution of $\Big(\mathcal{P}\Big)$. Then $\overline{x}$ is an optimal solution if and only if there exists multipliers $\lambda_1, ..., \lambda_m \geq 0$ such that

$$0 \in \partial f(\overline{x}) + \Big[\sum_{i=1}^{m} \lambda_i \partial g_i(\overline{x})\Big] + N(\overline{x}; \Omega),$$

$\lambda_i \geq 0$ for all $i = 0, ..., m$, and furthermore $\lambda_i g_i(\overline{x}) = 0$ for all $i = 1, ..., m$.

Proof of Theorem 16.5 First assume that $\overline{x}$ is an optimal solution. It suffices to prove that $\lambda_0 \neq 0$. By contradiction assume that $\lambda_0 = 0$. Thus there exists $\lambda_i g_i(\overline{x})$ for $i = 1, ..., m$ and $v \in N(\overline{x}; \Omega)$ such that

$$0 = \sum_{i=1}^{m} \lambda_i v_i + v,$$

with $\lambda_i \geq 0$ and $\lambda_i g_i(\overline{x})$ for all $i = 1, ..., m$. Then

$$\langle 0, u - \overline{x} \rangle = \langle \sum_{i=1}^{m} \lambda_i v_i + v, u - \overline{x} \rangle = 0$$

where $u \in \Omega$ satisfies the Slaters condition, i.e. $g_i(u) < 0$ for all $i = 1, ..., m$. Thus

$$0 = \sum_{i=1}^{m} \lambda_i \langle v_i, u - \overline{x} \rangle + \langle v, u - \overline{x} \rangle$$

$$\leq \sum_{i=1}^{m} \lambda_i \big(g_i(u) - g_i(\overline{x}) \big) + \langle v, u - \overline{x} \rangle$$

$$= \sum_{i=1}^{m} \lambda_i g_i(u) + \langle v, u - \overline{x} \rangle < 0$$

which gives us our contradiction.

To prove the other implication assume that

$$0 \in \partial f(\overline{x}) + \sum_{i=1}^{m} \lambda_i \partial g_i(\overline{x}) + N(\overline{x}; \Omega),$$

where $\lambda_i \geq 0$ and $\lambda_i g_i(\overline{x}) = 0$ for all $i = 1, ..., m$, and furthermore assume that $\overline{x}$ is a feasible solution. We will show that $\overline{x}$ is an optimal solution of $\Big(\mathcal{P}\Big)$. Fix any feasible solution $x \in \mathbb{R}^n$. Then there exists $v_0 \in \partial f(\overline{x})$, $v_i \in \partial g_i(\overline{x})$ for $i = 1, ..., m$ and $v \in N(\overline{x}; \Omega)$ such that

$$0 = v_0 + \sum_{i=1}^{m} \lambda_i v_i + v.$$

Thus we have that

$$0 = \langle 0, x - \overline{x} \rangle$$

$$
\begin{aligned}
&= \langle v_0 + \textstyle\sum_{i=1}^{m} \lambda_i v_i + v, x - \overline{x} \rangle \\
&= \langle v_0, x - \overline{x} \rangle + \textstyle\sum_{i=1}^{m} \lambda_i \langle v_i, x - \overline{x} \rangle + \langle v, x - \overline{x} \\
&\leq \Big(f(x) - f(\overline{x}) \Big) + \textstyle\sum_{i=1}^{m} \lambda_i \Big(g_i(x) - g_i(\overline{x}) \Big) + \langle v, x - \overline{x} \rangle \\
&= f(x) - f(\overline{x}) + \textstyle\sum_{i=1}^{m} \lambda_i g_i(x) + \langle v, x - \overline{x} \rangle \\
&\leq f(x) - f(\overline{x}).
\end{aligned}
$$

Thus we have shown that $f(\overline{x}) \leq f(x)$. $\square$

Consider the problem

Minimize $f(x)$ $\qquad\qquad \Big(\mathcal{Q} \Big)$

Subject to $g_i(x) \leq 0$ for all $i = 1, ..., m$ and $x \in \Omega$

where $f, g_1, ..., g_2 : \mathbb{R}^n \to \mathbb{R}$ are differentiable convex functions.

Theorem 16.6 Consider the problem $\left(\mathcal{Q}\right)$ and let $\overline{x}$ be a feasible solution of $\left(\mathcal{Q}\right)$. Suppose that $\{g_i(\overline{x}) : i \in A(\overline{x})\}$ is a linearly independent set. Then $\overline{x}$ is an optimal solution of $\left(\mathcal{Q}\right)$ if and only if there exists $\lambda_1, ..., \lambda_m \geq 0$ such that

$$\nabla f(\overline{x}) + \sum_{i=1}^{m} \lambda_i \nabla g_i(\overline{x}) = 0$$

with $\lambda_i g_i(\overline{x}) = 0$ for all $i = 1, ..., m$.

Proof From the previous proof we have that

$$0 = \lambda_0 \nabla f(\overline{x}) + \sum_{i=1}^{m} \lambda_i \nabla g_i(\overline{x})$$

$$\lambda_0 \nabla f(\overline{x}) + \sum_{i \in A(\overline{x})} \lambda_i \nabla g_i(\overline{x}).$$

Assume by contradiction that $\lambda_0 = 0$. Then

$$\sum_{i \in A(\overline{x})} \lambda_i \nabla g_i(\overline{x}) = 0$$

and by the linear independence of $\{g_i(\overline{x}) : i \in A(\overline{x})\}$ we have that $\lambda_i = 0$ for all $i \in A(\overline{x})$. This gives us a contradiction because by assumption there exists a $\lambda_i > 0$ for some $i \in \{1, ..., m\}$. $\square$

KKT Example

Consider

$$\begin{aligned} \min\, & x_1^2 + x_2^2 \\ \text{s.t. } & x_1 + x_2 = 1, \\ & x_2 \le \alpha, \end{aligned}$$

where $(x_1, x_2) \in \mathbb{R}^2$, $\alpha \in \mathbb{R}$.

The Lagrangian function is

$$L(x_1, x_2, \lambda, \mu) = x_1^2 + x_2^2 + \lambda(1 - x_1 - x_2) + \mu(x_2 - \alpha).$$

KKT conditions are

$$\begin{aligned} \frac{\partial L}{\partial x_i} = 0, \ i = 1, 2, \\ x_1 + x_2 = 1, \\ x_2 - \alpha \le 0, \\ \mu \ge 0, \\ \mu(x_2 - \alpha) = 0. \end{aligned}$$

KKT Example

Setting the partial derivatives zero, we get

$$\frac{\partial L}{\partial x_1} = 2x_1 - \lambda = 0, \ \frac{\partial L}{\partial x_2} = 2x_2 - \lambda + \mu = 0.$$

Therefore, $x_1 = \frac{\lambda}{2}$, $x_2 = \frac{\lambda - \mu}{2}$. Substituting into the equality constraint:

$$x_1 + x_2 = \lambda - \frac{\mu}{2} = 1.$$

So $\lambda = \frac{\mu}{2} + 1$. We get

$$x_1 = \frac{\mu}{4} + \frac{1}{2}, \ x_2 = -\frac{\mu}{4} + \frac{1}{2}.$$

Combining with the inequality constraint, we get $-\frac{\mu}{4} + \frac{1}{2} \leq \alpha$, that is $\mu \geq 2 - 4\alpha$. We consider 3 cases.

KKT Example

- $\alpha > \frac{1}{2}$: We can check that $\mu = 0 > 2 - 4\alpha$ satisfies all the KKT conditions. So $x_1^* = x_2^* = \frac{1}{2}$ is a strictly feasible solution and the minimum value is $\frac{1}{2}$.
- $\alpha = \frac{1}{2}$: Similar to case 1, $\mu = 0 = 2 - 4\alpha$. $x_1^* = x_2^* = \frac{1}{2}$ is a boundary solution and the minimum value is $\frac{1}{2}$.
- $\alpha < \frac{1}{2}$: In this case $\mu = 2 - 4\alpha > 0$. Then $x_1^* = 1 - \alpha,\ x_2^* = \alpha$. The minimum value is $(1 - \alpha)^2 + \alpha^2$.

Computation of KKT Points

There seems to be confusion on how one computes KKT points. In general this is a hard problem. The problems I give you to do by hand are not necessarily easy, but they are doable. The basic is idea is to make some reasonable guesses and then to use elimination techniques. I will illustrate this with the following homework problem.

Problem: Locate all of the KKT points for the following problem. Are these points local solutions? Are they global solutions?

$$\begin{array}{ll} \text{minimize} & x_1^2 + x_2^2 - 4x_1 - 4x_2 \\ \text{subject to} & x_1^2 \leq x_2 \\ & x_1 + x_2 \leq 2 \ . \end{array}$$

Solution: First write the problem in the standard form required for the application of the KKT theory:

$$\begin{array}{lll} \text{minimize} & f_0(x) & \\ \text{subject to} & f_i(x) \leq 0, & i = 1, 2, \ldots, s \\ & f_i(x) = 0, & i = s+1, s+2, \ldots, m. \end{array}$$

In our example there are no equality constraints, so $s = m = 2$ and we have

$$\begin{aligned} f_0(x_1, x_2) &= x_1^2 + x_2^2 - 4x_1 - 4x_2 = (x_1 - 2)^2 + (x_2 - 2)^2 - 8 \\ f_1(x_1, x_2) &= x_1^2 - x_2 \\ f_2(x_1, x_2) &= x_1 + x_2 - 2 \ . \end{aligned}$$

Note that we can ignore the constant term in the objective function since it does not effect the optimal solution, so henceforth $f_0(x_1, x_2) = x_1 - 2)^2 + (x_2 - 2)^2$. At this point it is often helpful to graph the solution set if possible, as it is in this case. It is a slice of a parabola.

Since all of these functions are convex, this is an example of a convex programming problem and so the KKT conditions are both necessary and sufficient for global optimality. Hence, if we locate a KKT point we know that it is necessarily a globally optimal solution.

The Lagrangian for this problem is

$$L((x_1, x_2), (u_1, u_2)) = (x_1 - 2)^2 + (x_2 - 2)^2 + u_1(x_1^2 - x_2) + u_2(x_1 + x_2 - 2) \ .$$

Let us now write the KKT conditions for this problem.

1. (Primal Feasibility) $x_1^2 \leq x_2$ and $x_1 + x_2 \leq 2$
2. (Dual Feasibility) $0 \leq u_1$ and $0 \leq u_2$
3. (Complementarity) $u_1(x_1^2 - x_2) = 0$ and $u_2(x_1 + x_2 - 2) = 0$
4. (Stationarity of the Lagrangian)

$$0 = \nabla_x L((x_1, x_2), (u_1, u_2)) = \begin{pmatrix} 2(x_1 - 2) + 2u_1 x_1 + u_2 \\ 2(x_2 - 2) - u_1 + u_2 \end{pmatrix},$$

or equivalently

$$\begin{aligned} 4 &= 2x_1 + 2u_1 x_1 + u_2 \\ 4 &= 2x_2 - u_1 + u_2. \end{aligned}$$

Next observe that the global minimizer for the objective function is $(x_1, x_2) = (2, 2)$. Thus, if this point are feasible, it would be the global solution and the multipliers would both be zero. But it is not feasible. Indeed, both constraints are violated by this point. Hence, we conjecture that both constraints are active at the solution. In this case, the KKT pair $((x_1, x_2), (u_1, u_2))$ must satisfy the following 4 key equations

$$\begin{aligned} x_2 &= x_2^2 \\ 2 &= x_1 + x_2 \\ 4 &= 2x_1 + 2u_1x_1 + u_2 \\ 4 &= 2x_2 - u_1 + u_2. \end{aligned}$$

This is 4 equations in 4 unknowns that we can try to solve by elimination. Using the first equation to eliminate x_2 from the second equation, we see that x_1 must satisfy

$$0 = x_1^2 + x_1 - 2 = (x_1 + 2)(x_1 - 1),$$

so $x_1 = -2$ or $x_1 = 1$. Thus, either $(x_1, x_2) = (-2, 4)$ or $(x_1, x_2) = (1, 1)$. Since $(1, 1)$ is closer the global minimizer of the objective f_0, let us first investigate $(x_1, x_2) = (1, 1)$ to see if it is a KKT point. For this we must find the KKT multipliers (u_1, u_2).

By plugging $(x_1, x_2) = (1, 1)$ into the second of the key equations given above, we get

$$2 = 2u_1 + u_2 \quad \text{and} \quad 2 = -u_1 + u_2 \ .$$

By subtracting these two equations, we get $0 = 3u_1$ so $u_1 = 0$ and $u_2 = 2$. Since both of these values are non-negative, we have found a KKT pair for the original problem. Hence, by convexity we know that $(x_1, x_2) = (1, 1)$ is the global solution to the problem.

Algorithm

April 2022

Content

- Convergence of gradient descent
- Accelerated first order methods
- Newton’s method
- Subgradients and the subgradient method
- Proximal algorithms
- Proximal gradient methods

Convergence of gradient descent

Here we will prove convergence guarantees for **gradient descent**, where we find a minimizer[1] of

$$\underset{\boldsymbol{x}\in\mathbb{R}^N}{\text{minimize}}\, f(\boldsymbol{x})$$

using our generic iterative algorithm choosing the direction to move as

$$\boldsymbol{d}_k = -\nabla f\left(\boldsymbol{x}_k\right),$$

resulting in the update rule

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \alpha_k \nabla f\left(\boldsymbol{x}_k\right).$$

Our goal is to establish the convergence rate of gradient descent. This can be measured in many different ways. One way is to establish

$$\begin{aligned} f(\boldsymbol{x}_k) - f(\boldsymbol{x}^\star) &\leq \text{some function that decreases to 0 as } k \to \infty \\ &:= g(k) \end{aligned}$$

This established convergence of the *function values* to the minimum. With a result like this in hand, you can ask

How many iterations do we need to be within ϵ of a solution?

and the answer is

$$k \geq g^{-1}(\epsilon) \text{ iterations will suffice.}$$

[1] In this section, we will always assume that a minimizer exists.

For example, if we establish

$$f(\boldsymbol{x}_k) - f(\boldsymbol{x}^\star) \leq 5/k^2$$

then we know that

$$k \geq \sqrt{\frac{5}{\epsilon}} \quad \Rightarrow \quad f(\boldsymbol{x}_k) - f(\boldsymbol{x}^\star) \leq \epsilon.$$

Note that the $g(k)$ we derive will in general be monotonically decreasing and hence invertible.

If we know that there is a unique solution $\boldsymbol{x}^\star$, we might also bound

$$\|\boldsymbol{x}_k - \boldsymbol{x}^\star\|_2 \leq \text{some function that decreases to 0 as } k \to \infty.$$

The bounds we develop will depend on the structural properties of the function f. In the mathematical optimization literature, there are results for all different kinds of structure on f. In this set of notes, we will consider two cases: convex differentiable f that

1. have an L-Lipschitz gradient map, i.e.

$$\|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})\|_2 \leq L\|\boldsymbol{x} - \boldsymbol{y}\|_2, \quad \text{for all } \boldsymbol{x}, \boldsymbol{y};$$

2. have an L-Lipschitz gradient and in addition are μ-strongly convex, i.e.

$$f(\boldsymbol{y}) \;\geq\; f(\boldsymbol{x}) + \langle \boldsymbol{y} - \boldsymbol{x}, \nabla f(\boldsymbol{x}) \rangle + \frac{\mu}{2}\|\boldsymbol{y} - \boldsymbol{x}\|_2^2 \quad \text{for all } \boldsymbol{x}, \boldsymbol{y}.$$

We will see that the additional structure added in the second case makes a dramatic difference in convergence rate.

Convergence of gradient descent: f smooth

As we have discussed before, having an L-Lipschitz gradient is akin to the function being smooth: if the derivative changes in a controlled manner as we move from point to point, the function itself will be very well-behaved.

On the homework, you showed that

$$\|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})\|_2 \leq L\|\boldsymbol{x} - \boldsymbol{y}\|_2, \tag{1}$$

means that we have the pointwise quadratic upper bound

$$f(\boldsymbol{y}) \leq f(\boldsymbol{x}) + \langle \boldsymbol{y} - \boldsymbol{x}, \nabla f(\boldsymbol{x})\rangle + \frac{L}{2}\|\boldsymbol{y} - \boldsymbol{x}\|_2^2 \tag{2}$$

This provides some intuition for what kind of structure the Lipschitz gradient condition imposes on f. Recall that for any convex function, we have

$$f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \langle \boldsymbol{y} - \boldsymbol{x}, \nabla f(\boldsymbol{x})\rangle.$$

So if f is convex, then at any point $\boldsymbol{x}$ we can bound f from *below* by a linear approximation. If in addition, if f has a Lipschitz gradient, (2) we can also bound it from *above* using a quadratic approximation. We will often refer to functions that obey (1) as *L-smooth*.

Now, let's consider running gradient descent on such a function with a **fixed step size**[2] $\alpha_k = 1/L$. Recall that the central gradient descent iteration is just

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \frac{1}{L}\nabla f(\boldsymbol{x}_k).$$

[2]This requires that you know L, which may not be possible in practice. In fact, if $\alpha < 1/L$ you will still get convergence, it will simply be slower. Moreover, it is not too hard to extend this approach to get a similar guarantee when using a backtracking line search.

From our assumption that f is L-smooth, we know that f satisfies (2), and thus plugging in $\boldsymbol{y} = \boldsymbol{x}_{k+1}$, we obtain

$$\begin{aligned} f(\boldsymbol{x}_{k+1}) &\leq f(\boldsymbol{x}_k) + \left\langle -\frac{1}{L}\nabla f(\boldsymbol{x}_k), \nabla f(\boldsymbol{x}_k) \right\rangle + \frac{L}{2}\left\| \frac{1}{L}\nabla f(\boldsymbol{x}_k) \right\|_2^2 \\ &= f(\boldsymbol{x}_k) - \frac{1}{L}\|\nabla f(\boldsymbol{x}_k)\|_2^2 + \frac{1}{2L}\|\nabla f(\boldsymbol{x}_k)\|_2^2 \\ &= f(\boldsymbol{x}_k) - \frac{1}{2L}\|\nabla f(\boldsymbol{x}_k)\|_2^2. \end{aligned} \tag{3}$$

Note that (3) shows that $f(\boldsymbol{x}_{k+1}) < f(\boldsymbol{x}_k)$ as long as we are not already at the solution, so we are at least guaranteed to make some progress at each iteration. In fact, it says a bit more, giving us a guarantee regarding *how much* progress we are making, namely that

$$f(\boldsymbol{x}_k) - f(\boldsymbol{x}_{k+1}) \geq \frac{1}{2L}\|\nabla f(\boldsymbol{x}_k)\|_2^2,$$

so that if the gradient is large we are guaranteed to make a large amount of progress.

In the Technical Details section at the end of these notes, we show that by combining this result with the definition of convexity and doing some clever manipulations, we can get a guarantee of the form

$$f(\boldsymbol{x}_k) - f(\boldsymbol{x}^\star) \leq \frac{L}{2k}\|\boldsymbol{x}_0 - \boldsymbol{x}^\star\|_2^2.$$

Thus, for L-smooth functions, we can guarantee that the error is $O(1/k)$ after k iterations. Another way to put this is to say that we can guarantee accuracy

$$f(\boldsymbol{x}_k) - f(\boldsymbol{x}^\star) \leq \epsilon$$

as long as

$$k \geq \frac{L}{2\epsilon}\|\boldsymbol{x}_0 - \boldsymbol{x}^\star\|_2^2.$$

Note that if ϵ is very small, this says we can expect to need a very large number of iterations.

Convergence of gradient descent: smooth and strongly convex

We will now show that the convergence rate is much faster if f is strongly convex in addition to being smooth. Recall that for a μ-strongly convex function, we have

$$f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \langle \boldsymbol{y} - \boldsymbol{x}, \nabla f(\boldsymbol{x}) \rangle + \frac{\mu}{2}\|\boldsymbol{y} - \boldsymbol{x}\|_2^2. \tag{4}$$

for all $\boldsymbol{x}, \boldsymbol{y}$.

We will use the same fixed step size $\alpha_k = 1/L$, and begin our analysis in the same way as before, in which we derived the intermediate result (3) that the L-smoothness of f implies

$$f(\boldsymbol{x}_{k+1}) \leq f(\boldsymbol{x}_k) - \frac{1}{2L}\|\nabla f(\boldsymbol{x}_k)\|_2^2.$$

We can now use strong convexity to obtain a lower bound on $\|\nabla f(\boldsymbol{x})\|_2^2$.

We can obtain a simpler lower bound for $f(\boldsymbol{y})$ by determining the smallest value that the right-hand side of (4) could ever take over all possible choices of $\boldsymbol{y}$. To do this, we simply minimize this lower bound by taking the gradient with respect to $\boldsymbol{y}$ and setting it equal to zero:

$$\nabla f(\boldsymbol{x}) + \mu(\boldsymbol{y} - \boldsymbol{x}) = 0,$$

From this we obtain that the lower bound in (4) will be minimized by

$$\boldsymbol{y} - \boldsymbol{x} = -\frac{1}{\mu}\nabla f(\boldsymbol{x}).$$

Plugging this into (4) yields

$$\begin{aligned} f(\boldsymbol{y}) &\geq f(\boldsymbol{x}) - \frac{1}{\mu}\|\nabla f(\boldsymbol{x})\|_2^2 + \frac{1}{2\mu}\|\nabla f(\boldsymbol{x})\|_2^2 \\ &= f(\boldsymbol{x}) - \frac{1}{2\mu}\|\nabla f(\boldsymbol{x})\|_2^2. \end{aligned}$$

In particular, this applies when $\boldsymbol{y} = \boldsymbol{x}^\star$, which after some rearranging yields

$$\|\nabla f(\boldsymbol{x})\|_2^2 \geq 2\mu\left(f(\boldsymbol{x}) - f(\boldsymbol{x}^\star)\right). \tag{PL}$$

This is a famous and useful result, often referred to as the **Polyak-Łojasiewicz inequality**.

Combining the PL inequality with (3) we obtain

$$\begin{aligned} f(\boldsymbol{x}_{k+1}) - f(\boldsymbol{x}^\star) &\leq f(\boldsymbol{x}_k) - f(\boldsymbol{x}^\star) - \frac{\mu}{L}\left(f(\boldsymbol{x}_k) - f(\boldsymbol{x}^\star)\right) \\ &= \left(1 - \frac{\mu}{L}\right)\left(f(\boldsymbol{x}_k) - f(\boldsymbol{x}^\star)\right). \end{aligned}$$

That is, the gap between the current value of the objective function and the optimal value is cut down by a factor of $1 - \mu/L < 1$ at each iteration. (Note that (2) and (4) imply that $L \geq \mu$.)

This is an example of *linear convergence*; it is easy to apply the above iteratively to show that

$$f(\boldsymbol{x}_k) - f(\boldsymbol{x}^\star) \;\leq\; \left(1 - \frac{\mu}{L}\right)^k \left(f(\boldsymbol{x}_0) - f(\boldsymbol{x}^\star)\right). \tag{5}$$

If we use $\epsilon_0 = f(\boldsymbol{x}_0) - f(\boldsymbol{x}^\star)$ to denote the initial error, this means that we can guarantee that

$$f(\boldsymbol{x}_k) - f(\boldsymbol{x}^\star) \leq \epsilon$$

for

$$\begin{aligned} k &\geq \frac{\log(\epsilon/\epsilon_0)}{\log(1-\mu/L)} \\ &\geq \frac{L}{\mu} \log\left(\frac{\epsilon_0}{\epsilon}\right), \end{aligned}$$

where the second inequality uses the fast that $-\log(1-\alpha) \geq \alpha$ for all $0 \leq \alpha < 1$.

Let's step back for a moment, and compare

$$\frac{1}{\epsilon} \quad \text{versus} \quad \log\left(\frac{1}{\epsilon}\right).$$

What are these quantities when $\epsilon = 10^{-2}$? What about 10^{-6}? This is all to say that the performance guarantees for gradient descent are dramatically better when f is strictly convex than when it is not.

We can also use (5) to characterize the convergence of the iterates $\boldsymbol{x}_k$ to the unique solution $\boldsymbol{x}^\star$. Applying (4) with $\boldsymbol{x} = \boldsymbol{x}^\star$ and $\boldsymbol{y} = \boldsymbol{x}_k$ yields (after noting $\nabla f(\boldsymbol{x}^\star) = \boldsymbol{0}$)

$$f(\boldsymbol{x}_k) - f(\boldsymbol{x}^\star) \geq \frac{\mu}{2}\|\boldsymbol{x}_k - \boldsymbol{x}^\star\|_2^2,$$

while applying (2) with $\boldsymbol{x} = \boldsymbol{x}^\star$ and $\boldsymbol{y} = \boldsymbol{x}_0$ yields

$$f(\boldsymbol{x}_0) - f(\boldsymbol{x}^\star) \leq \frac{L}{2}\|\boldsymbol{x}_0 - \boldsymbol{x}^\star\|_2^2.$$

Combining these with (5) yields

$$\|\boldsymbol{x}_k - \boldsymbol{x}^\star\|_2^2 \;\leq\; \frac{L}{\mu}\left(1 - \frac{\mu}{L}\right)^k \|\boldsymbol{x}_0 - \boldsymbol{x}^\star\|_2^2,$$

so $\boldsymbol{x}_k \to \boldsymbol{x}^\star$ at a linear rate as well. I will note that a more careful analysis (which we won't go into here) can also remove the factor of L/μ in front, yielding

$$\|\boldsymbol{x}_k - \boldsymbol{x}^\star\|_2^2 \;\leq\; \left(1 - \frac{\mu}{L}\right)^k \|\boldsymbol{x}_0 - \boldsymbol{x}^\star\|_2^2.$$

Finally, we also note that the PL inequality above also provides some guidance in terms of setting a stopping criterion. Specifically, if we declare convergence when $\|\nabla f(\boldsymbol{x}_k)\|_2 \leq \epsilon$ then the PL inequality allows us to conclude that

$$f(\boldsymbol{x}_k) - f(\boldsymbol{x}^\star) \leq \frac{1}{2\mu}\|\nabla f(\boldsymbol{x}_k)\|_2^2 \leq \frac{\epsilon^2}{2\mu}.$$

This provides a principled way of declaring convergence.

Technical Details: L-smooth convergence

Here we complete the convergence analysis for gradient descent on L-smooth functions that is summarized above. Specifically, recall that above in (3) we showed that if f is L-smooth then

$$f(\boldsymbol{x}_{k+1}) \leq f(\boldsymbol{x}_k) - \frac{1}{2L}\|\nabla f(\boldsymbol{x}_k)\|_2^2.$$

Moreover, by the convexity of f,

$$f(\boldsymbol{x}_k) \leq f(\boldsymbol{x}^\star) + \langle \boldsymbol{x}_k - \boldsymbol{x}^\star, \nabla f(\boldsymbol{x}_k)\rangle,$$

where $\boldsymbol{x}^\star$ is a minimizer of f, and so we have

$$f(\boldsymbol{x}_{k+1}) \leq f(\boldsymbol{x}^\star) + \langle \boldsymbol{x}_k - \boldsymbol{x}^\star, \nabla f(\boldsymbol{x})\rangle - \frac{1}{2L}\|\nabla f(\boldsymbol{x}_k)\|_2^2.$$

Substituting $\nabla f(\boldsymbol{x}_k) = L(\boldsymbol{x}_k - \boldsymbol{x}_{k+1})$ then yields

$$f(\boldsymbol{x}_{k+1}) - f(\boldsymbol{x}^\star) \leq L\langle \boldsymbol{x}_k - \boldsymbol{x}^\star, \boldsymbol{x}_k - \boldsymbol{x}_{k+1}\rangle - \frac{L}{2}\|\boldsymbol{x}_k - \boldsymbol{x}_{k+1}\|_2^2. \quad (6)$$

We can re-write this in a slightly more convenient way using the fact that

$$\|\boldsymbol{a} - \boldsymbol{b}\|_2^2 = \|\boldsymbol{a}\|_2^2 - 2\langle \boldsymbol{a}, \boldsymbol{b}\rangle + \|\boldsymbol{b}\|_2^2$$

and thus

$$2\langle \boldsymbol{a}, \boldsymbol{b}\rangle - \|\boldsymbol{b}\|_2^2 = \|\boldsymbol{a}\|_2^2 - \|\boldsymbol{a} - \boldsymbol{b}\|_2^2.$$

Setting $\boldsymbol{a} = \boldsymbol{x}_k - \boldsymbol{x}^\star$ and $\boldsymbol{b} = \boldsymbol{x}_k - \boldsymbol{x}_{k+1}$ and applying this to (6), we obtain the bound

$$f(\boldsymbol{x}_{k+1}) - f(\boldsymbol{x}^\star) \leq \frac{L}{2}\left(\|\boldsymbol{x}_k - \boldsymbol{x}^\star\|_2^2 - \|\boldsymbol{x}_{k+1} - \boldsymbol{x}^\star\|_2^2\right).$$

This result bounds how far away $f(\boldsymbol{x}_{k+1})$ is from the optimal $f(\boldsymbol{x}^\star)$ in terms (primarily) of the error in the previous iteration: $\|\boldsymbol{x}_k - \boldsymbol{x}^\star\|_2^2$. We can use this result to bound $f(\boldsymbol{x}_{k+1}) - f(\boldsymbol{x}^\star)$ in terms of the initial error $\|\boldsymbol{x}_0 - \boldsymbol{x}^\star\|_2^2$ by a clever argument.

Specifically, this bound holds not only for iteration k, but for all iterations $i = 1, \ldots, k$, so we can write down k inequalities and then sum them up to obtain

$$\sum_{i=1}^{k} f(\boldsymbol{x}_i) - f(\boldsymbol{x}^\star) \leq \frac{L}{2} \left(\sum_{i=1}^{k} \|\boldsymbol{x}_{i-1} - \boldsymbol{x}^\star\|_2^2 - \|\boldsymbol{x}_i - \boldsymbol{x}^\star\|_2^2 \right).$$

The right-hand side of this inequality is what is called a *telescopic sum*: each successive term in the sum cancels out part of the previous term. Once you write this out, all the terms cancel except for two (one component from the $i = 1$ term and one from the $i = k$ term) giving us:

$$\begin{aligned} \sum_{i=1}^{k} f(\boldsymbol{x}_i) - f(\boldsymbol{x}^\star) &\leq \frac{L}{2} \left(\|\boldsymbol{x}_0 - \boldsymbol{x}^\star\|_2^2 - \|\boldsymbol{x}_k - \boldsymbol{x}^\star\|_2^2 \right) \\ &\leq \frac{L}{2} \|\boldsymbol{x}_0 - \boldsymbol{x}^\star\|_2^2. \end{aligned}$$

Since, as noted above, $f(\boldsymbol{x}_i)$ is monotonically decreasing in i, we also have that

$$k \left(f(\boldsymbol{x}_k) - f(\boldsymbol{x}^\star) \right) \leq \sum_{i=1}^{k} f(\boldsymbol{x}_i) - f(\boldsymbol{x}^\star),$$

and thus

$$f(\boldsymbol{x}_k) - f(\boldsymbol{x}^\star) \leq \frac{L}{2k} \|\boldsymbol{x}_0 - \boldsymbol{x}^\star\|_2^2,$$

which is exactly what we wanted to show.

Accelerated first-order methods

In the last lecture we provided convergence guarantees for gradient descent under two different assumptions. Under the stronger assumption that f was both L-smooth *and* strongly convex with parameter μ, we showed that convergence to a tolerance of ϵ was possible in $O(\frac{L}{\mu}\log(1/\epsilon))$ iterations. Under the weaker assumption where we only assume that f is L-smooth, we were able to show that $O(L/\epsilon)$ iterations would be sufficient.

In this lecture we show that there are small changes we can make to gradient descent that can dramatically improve its performance, both in theory (resulting in improvements on the bounds above) and in practice. We will talk about two of these here: the heavy ball method and Nesterov's "optimal algorithm." Both of these strategies incorporate the idea of *momentum*, although in subtly different ways.

Momentum

One way to interpret gradient descent is as a discretization to the *gradient flow* differential equation

$$\begin{aligned} \boldsymbol{x}'(t) &= -\nabla f(\boldsymbol{x}(t)), \\ \boldsymbol{x}(0) &= \boldsymbol{x}_0. \end{aligned} \tag{1}$$

The solution to (1) is a curve that tracks the direction of steepest descent directly to the minimizer, where it arrives at a fixed point (where $\nabla f(\boldsymbol{x}) = \boldsymbol{0}$). To see how gradient descent arises as a discretization of (1), suppose we approximate the derivative with a forward difference

$$\boldsymbol{x}'(t) \approx \frac{\boldsymbol{x}(t+h) - \boldsymbol{x}(t)}{h},$$

for some small h. So if we think of $\boldsymbol{x}_{k+1}$ and $\boldsymbol{x}_k$ as closely spaced time points, we can interpret

$$\frac{1}{\alpha}\left(\boldsymbol{x}_{k+1} - \boldsymbol{x}_k\right) = -\nabla f(\boldsymbol{x}_k),$$

as a discrete approximation to gradient flow. Re-arranging the equation above yields the gradient descent iteration $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \alpha \nabla f(\boldsymbol{x}_k)$.

The problem is once we perform this discretization, the path tends to oscillate. One way to get a more regular path is to consider an alternative differential equation that also has a fixed point where $\nabla f(\boldsymbol{x}) = 0$ but also incorporates a second-order term:

$$m\boldsymbol{x}''(t) + \boldsymbol{x}'(t) = -\nabla f(\boldsymbol{x}(t)). \tag{2}$$

From a physical perspective, this is a model for a particle with mass m moving in a potential field with friction. This results in trajectories that develop momentum (a heavy ball will move down a hill faster than a light one in the presence of friction). In the case where $m = 0$ we recover (1), but in general the inclusion of the mass term above will result in a more accelerated trajectory towards the solution.

We can discretize the dynamics as before by setting

$$\boldsymbol{x}''(t) \approx \frac{\boldsymbol{x}_{k+1} - 2\boldsymbol{x}_k + \boldsymbol{x}_{k-1}}{h_1}, \quad \boldsymbol{x}'(t) \approx \frac{\boldsymbol{x}_k - \boldsymbol{x}_{k-1}}{h_2}.$$

If we plug these into (2) and rearrange we obtain an update rule of the form

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \beta_k(\boldsymbol{x}_k - \boldsymbol{x}_{k-1}) - \alpha_k \nabla f(\boldsymbol{x}_k), \tag{3}$$

where $\beta = h_1/h_2 m$ and $\alpha = h_1/m$. This is the core iteration for the **heavy ball method**, introduced by Polyak in 1964 [Pol64]. The $\boldsymbol{x}_k - \boldsymbol{x}_{k-1}$ term above adds a little bit of the last step $\boldsymbol{x}_k - \boldsymbol{x}_{k-1}$ direction into the new step direction $\boldsymbol{x}_{k+1} - \boldsymbol{x}_k$ – this method is also referred to as *gradient descent with momentum*.

Convergence of the heavy ball method

In the previous lecture we showed that if $f(\boldsymbol{x})$ is L-smooth and strongly convex, then we can obtain a bound of the form

$$f(\boldsymbol{x}_{k+1}) - f(\boldsymbol{x}^\star) \leq \left(1 - \frac{1}{\kappa}\right)^k \left(f(\boldsymbol{x}_0) - f(\boldsymbol{x}^\star)\right),$$

where $\kappa = L/\mu$ is the "condition number." From this we showed that we can guarantee

$$\frac{f(\boldsymbol{x}_k) - f(\boldsymbol{x}^\star)}{f(\boldsymbol{x}_0) - f(\boldsymbol{x}^\star)} \leq \epsilon$$

provided that

$$k \geq \kappa \log\left(1/\epsilon\right).$$

In the Technical Details at the end of these notes we also provide an alternative argument for the convergence of gradient descent that begins by showing that

$$\|\boldsymbol{x}_k - \boldsymbol{x}^\star\|_2 \leq \left(\frac{\kappa - 1}{\kappa + 1}\right)^k \|\boldsymbol{x}_0 - \boldsymbol{x}^\star\|_2.$$

Using a similar argument as before, we can use this to show that

$$\frac{\|\boldsymbol{x}_k - \boldsymbol{x}^\star\|_2}{\|\boldsymbol{x}_0 - \boldsymbol{x}^\star\|_2} \leq \epsilon$$

provided that

$$k \geq \kappa \log(1/\epsilon).$$

(Note that

$$\frac{\kappa - 1}{\kappa + 1} = 1 - \frac{2}{\kappa + 1} \leq 1 - 1/\kappa = 1 - \mu/L,$$

where the inequality comes from the fact that $\kappa \geq 1$.)

The heavy ball method significantly improves on this result in terms of its dependence on κ.

Specifically, under the same assumptions as before (L-smoothness and strong convexity), in the Technical Details section we show (for the quadratic case) that for the heavy ball method with

$$\alpha_k = \frac{4}{(\sqrt{L} + \sqrt{\mu})^2} \quad \text{and} \quad \beta_k = \left(\frac{\sqrt{L} - \sqrt{\mu}}{\sqrt{L} + \sqrt{\mu}} \right)^2$$

we can achieve

$$\frac{\|\boldsymbol{x}_k - \boldsymbol{x}^\star\|_2}{\|\boldsymbol{x}_0 - \boldsymbol{x}^\star\|_2} \leq \epsilon \quad \text{when} \quad k \gtrsim \sqrt{\kappa}\, \log(1/\epsilon).$$

The difference with gradient descent can be significant. When $\kappa = 10^2$, we are asking for $\approx 100 \log(1/\epsilon)$ iterations for gradient descent, as compared with $\approx 10 \log(1/\epsilon)$ from the heavy ball method.

Conjugate gradients

If you are familiar with the *method of conjugate gradients* (CG), some of this may feel vaguely familiar. If you have never heard of CG, I highly recommend reading through the tutorial "An introduction to the conjugate gradient method without the agonizing pain" [She94].

The CG method was developed for minimizing quadratic functions of the form $f(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^{\mathrm{T}}\boldsymbol{Q}\boldsymbol{x} - \boldsymbol{x}^{\mathrm{T}}\boldsymbol{b}$. While it is normally presented in quite a different fashion, it ultimately boils down to being a variant of

the heavy ball method that is particularly well-suited to minimizing quadratic functions. To see this connection, note that the core CG iteration can be expressed[1] as

$$\begin{aligned} \boldsymbol{d}_k &= -\nabla f(\boldsymbol{x}_k) + \beta_k \boldsymbol{d}_{k-1} \\ \boldsymbol{x}_{k+1} &= \boldsymbol{x}_k + \alpha_k \boldsymbol{d}_k, \end{aligned}$$

where we start with $\boldsymbol{d}_0 = -\nabla f(\boldsymbol{x}_0)$. In CG, the β_k are set as

$$\beta_k = \frac{\|\nabla f(\boldsymbol{x}_k)\|_2^2}{\|\nabla f(\boldsymbol{x}_{k-1})\|_2^2}.$$

If $f(\boldsymbol{x})$ is a quadratic function this choice ensures that at each iteration $\boldsymbol{d}_k$ is *conjugate* to $\boldsymbol{d}_0, \ldots, \boldsymbol{d}_{k-1}$. We won't worry about saying more about this beyond the fact that this is a good idea *if* $f(\boldsymbol{x})$ *is quadratic*. Once β_k is fixed, α_k can then be chosen using a line search. Again, if $f(\boldsymbol{x})$ is quadratic, there is a simple closed form solution for this (which we have previously derived).

While CG is parameterized differently than the heavy ball method as described in (3), they are fundamentally the same. To see this note that we can also write

$$\begin{aligned} \boldsymbol{x}_{k+1} &= \boldsymbol{x}_k + \alpha_k \left(-\nabla f(\boldsymbol{x}_k) + \beta_k \boldsymbol{d}_{k-1}\right) \\ &= \boldsymbol{x}_k - \alpha_k \nabla f(\boldsymbol{x}_k) + \alpha_k \beta_k \frac{\boldsymbol{x}_k - \boldsymbol{x}_{k-1}}{\alpha_{k-1}}. \end{aligned}$$

This is precisely the same iteration as (3), but with a slightly different way of parameterizing the weight being applied to the momentum term.

[1] You will typically see this algorithm described specifically for the quadratic case, in which case $\nabla f(\boldsymbol{x}) = Q\boldsymbol{x} - \boldsymbol{b}$ and these calculations are carefully broken up to re-use as many calculations as possible and avoid any unnecessary matrix-vector multiplies, so it may initially look quite different.

If you are trying to minimize a quadratic function, CG is the way to go. The convergence guarantees you get for CG when minimizing a quadratic function are just as good (but not actually better than) what you have for the heavy ball method, but you don't need to know anything like Lipschitz or strong convexity parameters (which would correspond to the maximum and minimum eigenvalues of $\boldsymbol{Q}$) in order to choose the α_k and β_k.

However, if you are trying to minimize *anything else* CG is not necessarily a good choice. The choices for α_k and β_k are highly tuned to the quadratic setting and can yield unstable results in general.

Nesterov's "optimal" method

In the case where f is strictly convex, you can come up with examples that show that the convergence rate of the heavy ball method can't be improved in general. For non-strictly convex f, the story is more complicated.

Recall that we also have a convergence result for gradient descent in the case where we only assume L-smoothness. In particular, last time we showed that for a fixed step size $\alpha = 1/L$,

$$f(\boldsymbol{x}_k) - f(\boldsymbol{x}^\star) \leq \frac{L}{2k} \|\boldsymbol{x}_0 - \boldsymbol{x}^\star\|_2^2.$$

Thus, to reduce the error by a factor of ϵ requires

$$k \geq \frac{L}{2\epsilon}$$

iterations.

In 1983, Yuri Nesterov proposed a slight variation on the heavy ball method that can improve on this theory, and often works better in

practice [Nes83].[2] Specifically, recall the heavy ball method, which can be represented via the iteration:

$$\begin{aligned} \boldsymbol{p}_k &= \beta_k \left(\boldsymbol{x}_k - \boldsymbol{x}_{k-1} \right) \\ \boldsymbol{x}_{k+1} &= \boldsymbol{x}_k + \boldsymbol{p}_k - \alpha_k \nabla f(\boldsymbol{x}_k), \end{aligned}$$

where we start with $\boldsymbol{p}_0 = \mathbf{0}$. Nesterov's method makes a subtle, but significant, change to this iteration:

$$\begin{aligned} \boldsymbol{p}_k &= \beta_k \left(\boldsymbol{x}_k - \boldsymbol{x}_{k-1} \right) \\ \boldsymbol{x}_{k+1} &= \boldsymbol{x}_k + \boldsymbol{p}_k - \alpha_k \nabla f(\boldsymbol{x}_k + \boldsymbol{p}_k). \end{aligned} \tag{4}$$

Notice that this is the same as heavy ball *except* that there is also a momentum term *inside* the gradient expression. With this iteration, we will show that (for a suitable choice of α_k and β_k)

$$f(\boldsymbol{x}_k) - f(\boldsymbol{x}^\star) \lesssim \frac{L}{k^2} \left\| \boldsymbol{x}_0 - \boldsymbol{x}^\star \right\|_2^2,$$

meaning that we can reduce the error by a factor of ϵ in

$$k \gtrsim \frac{1}{\sqrt{\epsilon}},$$

iterations. When $\epsilon \sim 10^{-4}$, this is much, much better than $1/\epsilon$.

Nesterov's method is called "optimal" because it is impossible to beat the $1/k^2$ rate using only function and gradient evaluations. There are careful demonstrations of this in the literature (e.g., in [Nes04]).

Note that in practice, α_k can be chosen using a standard line search, and a good choice of β_k (both in practice, and as we will show below,

[2] Note that this method remained to a large extent unknown in the wider community until his 2004 publication (in English) of [Nes04].

in theory) turns out to be

$$\beta_k = \frac{k-1}{k+2}. \tag{5}$$

This tells us that we should initially not provide much weight to the momentum term, which makes intuitive sense as the initial gradients may not be pushing us in the right direction, but as we proceed we should have increased confidence that we are headed in the right direction and increase how much weight we place on the momentum term.

Significantly, note that in setting β_k we do *not* need to know anything about the function we are minimizing (such as strong convexity parameters). This represents an important advantage compared to the heavy ball method described above.

Convergence analysis of Nesterov's method

Analyzing the convergence of Nesterov's method under the assumption of L-smoothness is a little more involved than for gradient descent, but the overall approach is the same and contains many of the same elements, so we will start by recalling the main building blocks that we used in analyzing gradient descent.

Consequences of convexity and L-smoothness

First, we recall some basic facts that hold for any $\boldsymbol{x}, \boldsymbol{y} \in \operatorname{dom} f$. Since f is convex we have

$$f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \langle \boldsymbol{y} - \boldsymbol{x}, \nabla f(\boldsymbol{x}) \rangle. \tag{6}$$

Since f is L-smooth we have

$$f(\boldsymbol{y}) \leq f(\boldsymbol{x}) + \langle \boldsymbol{y} - \boldsymbol{x}, \nabla f(\boldsymbol{x}) \rangle + \frac{L}{2}\|\boldsymbol{y} - \boldsymbol{x}\|_2^2. \tag{7}$$

As a consequence of (7) (by setting $\boldsymbol{y} = \boldsymbol{x} - \frac{1}{L}\nabla f(\boldsymbol{x})$), we have that for any $\boldsymbol{x}$,

$$f\left(\boldsymbol{x} - \frac{\nabla f(\boldsymbol{x})}{L}\right) \leq f(\boldsymbol{x}) - \frac{\|\nabla f(\boldsymbol{x})\|_2^2}{2L}. \tag{8}$$

Combining this with the upper bound on $f(\boldsymbol{x})$ that you can obtain by rearranging (6), we obtain

$$f\left(\boldsymbol{x} - \frac{\nabla f(\boldsymbol{x})}{L}\right) \leq f(\boldsymbol{y}) + \langle \boldsymbol{x} - \boldsymbol{y}, \nabla f(\boldsymbol{x})\rangle - \frac{\|\nabla f(\boldsymbol{x})\|_2^2}{2L}. \tag{9}$$

As we will see below, this inequality is the foundation of our analysis of both gradient descent and Nesterov's method. By plugging in different choices for $\boldsymbol{y}$ (such as $\boldsymbol{x}_k$ or $\boldsymbol{x}^\star$) we can obtain both *lower* bounds on how much progress we make when we take a gradient step as well as *upper* bounds on how far away we are from a global optimum.

Convergence of gradient descent

Recall that in our analysis for gradient we assume a fixed step size $\alpha = 1/L$, resulting in an update rule of

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \frac{\nabla f(\boldsymbol{x}_k)}{L}.$$

Thus, setting $\boldsymbol{x} = \boldsymbol{x}_k$ and $\boldsymbol{y} = \boldsymbol{x}^\star$ in (9) implies that

$$f(\boldsymbol{x}_{k+1}) \leq f(\boldsymbol{x}^\star) + L\langle \boldsymbol{x}_k - \boldsymbol{x}^\star, \boldsymbol{x}_k - \boldsymbol{x}_{k+1}\rangle - \frac{L}{2}\|\boldsymbol{x}_k - \boldsymbol{x}_{k+1}\|_2^2.$$

From this, if we define $\delta_k = f(\boldsymbol{x}_k) - f(\boldsymbol{x}^\star)$ and do some algebraic manipulation (see the previous notes) we get a bound of the form

$$\delta_{k+1} \leq \frac{L}{2}\left(\|\boldsymbol{x}_k - \boldsymbol{x}^\star\|_2^2 - \|\boldsymbol{x}_{k+1} - \boldsymbol{x}^\star\|_2^2\right).$$

This yields the *telescopic sum*

$$\begin{aligned}\sum_{i=0}^{k-1} \delta_{i+1} &\leq \frac{L}{2}\left(\sum_{i=0}^{k-1} \|\boldsymbol{x}_i - \boldsymbol{x}^\star\|_2^2 - \|\boldsymbol{x}_{i+1} - \boldsymbol{x}^\star\|_2^2\right) \\ &= \frac{L}{2}\left(\|\boldsymbol{x}_0 - \boldsymbol{x}^\star\|_2^2 - \|\boldsymbol{x}_k - \boldsymbol{x}^\star\|_2^2\right) \\ &\leq \frac{L}{2}\|\boldsymbol{x}_0 - \boldsymbol{x}^\star\|_2^2.\end{aligned}$$

The proof for gradient descent concludes by noting that

$$\delta_k \leq \frac{1}{k}\sum_{i=0}^{k-1} \delta_{i+1} \leq \frac{L}{2k}\|\boldsymbol{x}_0 - \boldsymbol{x}^\star\|_2^2.$$

Convergence of Nesterov's method

We will follow a similar argument to analyze Nesterov's method. We will again take $\alpha_k = 1/L$, but we will see that the analysis suggests a natural choice for β_k. With this choice of α_k, the main iteration from (4) is

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \boldsymbol{p}_k - \frac{1}{L}\nabla f(\boldsymbol{x}_k + \boldsymbol{p}_k).$$

It will be convenient to define

$$\boldsymbol{g}_k = -\frac{1}{L}\nabla f(\boldsymbol{x}_k + \boldsymbol{p}_k),$$

so that the main iteration becomes simply $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \boldsymbol{p}_k + \boldsymbol{g}_k$. With this notation, by setting $\boldsymbol{x} = \boldsymbol{x}_k + \boldsymbol{p}_k$ in (9) we obtain the bound

$$f(\boldsymbol{x}_{k+1}) \leq f(\boldsymbol{y}) - L\langle \boldsymbol{x}_k - \boldsymbol{p}_k - \boldsymbol{y}, \boldsymbol{g}_k\rangle - \frac{L}{2}\|\boldsymbol{g}_k\|_2^2. \tag{10}$$

If we set $\boldsymbol{y} = \boldsymbol{x}^\star$ in (10) and again let δ_k denote $f(\boldsymbol{x}_k) - f(\boldsymbol{x}^\star)$ we obtain

$$\delta_{k+1} \leq \frac{L}{2}\left(2\langle \boldsymbol{x}^\star - \boldsymbol{x}_k - \boldsymbol{p}_k, \boldsymbol{g}_k\rangle - \|\boldsymbol{g}_k\|_2^2\right). \tag{11}$$

In our analysis of gradient descent, we then tried to rearrange an analogous bound to obtain a telescopic sum, but that doesn't quite work here. Instead we will need to combine (11) with another bound. Noting that $\delta_k - \delta_{k+1} = f(\boldsymbol{x}_k) - f(\boldsymbol{x}_{k+1})$, we observe that setting $\boldsymbol{y} = \boldsymbol{x}_k$ in (10) yields

$$\delta_k - \delta_{k+1} \geq \frac{L}{2}\left(2\langle \boldsymbol{p}_k, \boldsymbol{g}_k\rangle + \|\boldsymbol{g}_k\|_2^2\right). \tag{12}$$

We now consider the inequality formed by adding together (11) and $1 - \lambda_k$ times (12) (where λ_k is something we will choose later, but satisfies $\lambda_k \geq 1$, so that this multiplication switches the direction of the inequality). The left-hand side of the sum will be

$$\delta_{k+1} + (1-\lambda_k)(\delta_k - \delta_{k+1}) = \lambda_k \delta_{k+1} - (\lambda_k - 1)\delta_k.$$

The right-hand side of the sum will be

$$\begin{aligned}
&\frac{L}{2}\left(2\langle \boldsymbol{x}^\star - \boldsymbol{x}_k - \boldsymbol{p}_k + (1-\lambda_k)\boldsymbol{p}_k, \boldsymbol{g}_k\rangle - \|\boldsymbol{g}_k\|_2^2 + (1-\lambda_k)\|\boldsymbol{g}_k\|_2^2\right) \\
&\quad = \frac{L}{2}\left(2\langle \boldsymbol{x}^\star - \boldsymbol{x}_k - \lambda_k\boldsymbol{p}_k, \boldsymbol{g}_k\rangle - \lambda_k\|\boldsymbol{g}_k\|_2^2\right) \\
&\quad = \frac{L}{2\lambda_k}\left(2\langle \boldsymbol{x}^\star - \boldsymbol{x}_k - \lambda_k\boldsymbol{p}_k, \lambda_k\boldsymbol{g}_k\rangle - \|\lambda_k\boldsymbol{g}_k\|_2^2\right) \\
&\quad = \frac{L}{2\lambda_k}\left(\|\boldsymbol{x}^\star - \boldsymbol{x}_k - \lambda_k\boldsymbol{p}_k\|_2^2 - \|\boldsymbol{x}^\star - \boldsymbol{x}_k - \lambda_k\boldsymbol{p}_k - \lambda_k\boldsymbol{g}_k\|_2^2\right),
\end{aligned}$$

where the last equality follows from the easily verified fact that $2\langle \boldsymbol{a}, \boldsymbol{b}\rangle - \|\boldsymbol{b}\|_2^2 = \|\boldsymbol{a}\|_2^2 - \|\boldsymbol{a} - \boldsymbol{b}\|_2^2$. If we make the substitution

$\boldsymbol{u}_k = \boldsymbol{x}_k + \lambda_k \boldsymbol{p}_k$, then combining these yields the inequality

$$\lambda_k^2 \delta_{k+1} - (\lambda_k^2 - \lambda_k)\delta_k \leq \frac{L}{2}\left(\|\boldsymbol{x}^\star - \boldsymbol{u}_k\|_2^2 - \|\boldsymbol{x}^\star - \boldsymbol{u}_k - \lambda_k \boldsymbol{g}_k\|_2^2\right). \tag{13}$$

We will now show that if we choose λ_k and β_k appropriately, (13) will yield a telescopic sum on both sides. This will occur on right-hand side of (13) if

$$\boldsymbol{u}_k + \lambda_k \boldsymbol{g}_k = \boldsymbol{u}_{k+1}.$$

Noting that $\boldsymbol{p}_{k+1} = \beta_{k+1}(\boldsymbol{x}_{k+1} - \boldsymbol{x}_k) = \beta_{k+1}(\boldsymbol{p}_k + \boldsymbol{g}_k)$, we can write

$$\begin{aligned}
\boldsymbol{u}_{k+1} &= \boldsymbol{x}_{k+1} + \lambda_{k+1}\boldsymbol{p}_{k+1} \\
&= \boldsymbol{x}_k + \boldsymbol{p}_k + \boldsymbol{g}_k + \lambda_{k+1}\beta_{k+1}(\boldsymbol{p}_k + \boldsymbol{g}_k) \\
&= \boldsymbol{x}_k + (1 + \lambda_{k+1}\beta_{k+1})(\boldsymbol{p}_k + \boldsymbol{g}_k).
\end{aligned}$$

Thus, to make $\boldsymbol{u}_{k+1}$ equal to $\boldsymbol{u}_k + \lambda_k \boldsymbol{g}_k = \boldsymbol{x}_k + \lambda_k(\boldsymbol{p}_k + \boldsymbol{g}_k)$ we simply need to have

$$\lambda_k = 1 + \lambda_{k+1}\beta_{k+1} \Rightarrow \beta_{k+1} = \frac{\lambda_k - 1}{\lambda_{k+1}}. \tag{14}$$

For β_k satisfying (14), if we sum (13) from $i = 0$ to $k - 1$ we thus have

$$\begin{aligned}
\sum_{i=0}^{k-1} \lambda_i^2 \delta_{i+1} - (\lambda_i^2 - \lambda_i)\delta_i &\leq \frac{L}{2}\left(\|\boldsymbol{x}^\star - \boldsymbol{u}_0\|_2^2 - \|\boldsymbol{x}^\star - \boldsymbol{u}_k\|_2^2\right) \\
&\leq \frac{L}{2}\|\boldsymbol{x}^\star - \boldsymbol{u}_0\|_2^2 \\
&= \frac{L}{2}\|\boldsymbol{x}^\star - \boldsymbol{x}_0\|_2^2.
\end{aligned} \tag{15}$$

Next, one possible approach is to choose the λ_k so as to obtain a telescopic sum on the left-hand side of the inequality as well. This is the approach you will see most often in analyzing the convergence of Nesterov's method, but it is a little involved and leads to a recursive formula for λ_k (and hence β_k) instead of a simple closed form expression. Instead we will choose a simpler λ_k that yields essentially the same bound.

Specifically, suppose that we set $\lambda_k = (k+2)/2$. First, note that from (14) this yields

$$\beta_{k+1} = \frac{\frac{k+2}{2} - 1}{\frac{k+1}{2}} = \frac{k}{k+3},$$

which coincides with the rule for setting β_k given in (5). Next, note that we can write

$$\sum_{i=0}^{k-1} \lambda_i^2 \delta_{i+1} - (\lambda_i^2 - \lambda_i)\delta_i = (\lambda_0 - \lambda_0^2)\delta_0 + \lambda_{k-1}^2 \delta_k + \sum_{i=1}^{k-1} (\lambda_{i-1}^2 - \lambda_i^2 + \lambda_i)\delta_i.$$

Plugging in $\lambda_i = (i+2)/2$ yields

$$\begin{aligned}\sum_{i=0}^{k-1} \lambda_i^2 \delta_{i+1} - (\lambda_i^2 - \lambda_i)\delta_i &= \left(\frac{k+1}{2}\right)^2 \delta_k + \frac{1}{4}\sum_{i=0}^{k-1} \delta_i \\ &\geq \left(\frac{k+1}{2}\right)^2 \delta_k,\end{aligned}$$

where the inequality follows since $\delta_i = f(\boldsymbol{x}_i) - f(\boldsymbol{x}^\star) \geq 0$. Combining this lower bound with (15) yields

$$\left(\frac{k+1}{2}\right)^2 \delta_k \leq \frac{L}{2}\|\boldsymbol{x}^\star - \boldsymbol{x}_0\|_2^2$$

or equivalently

$$f(\boldsymbol{x}_k) - f(\boldsymbol{x}^\star) \leq \frac{2L}{(k+1)^2}\|\boldsymbol{x}^\star - \boldsymbol{x}_0\|_2^2,$$

which is exactly the $O(1/k^2)$ convergence rate we wanted.

Technical Details: Analysis of the heavy ball method

We will analyze the heavy ball method for the special case of a quadratic function:

$$\underset{\boldsymbol{x}\in\mathbb{R}^N}{\text{minimize}} \;\; f(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^{\mathrm{T}}\boldsymbol{Q}\boldsymbol{x} + \boldsymbol{b}^{\mathrm{T}}\boldsymbol{x}$$

We will assume that the eigenvalues of $\boldsymbol{Q}$ are in $[\mu, L]$, and so $f(\boldsymbol{x})$ is both L-smooth and μ-strongly convex.

Gradient descent revisited

We will warm up for our analysis on the heavy ball method by quickly revisiting standard gradient descent. In the quadratic case, there is an easy argument that

$$\begin{aligned}\|\boldsymbol{x}_{k+1} - \boldsymbol{x}^\star\|_2 &\leq \frac{L-\mu}{L+\mu}\|\boldsymbol{x}_k - \boldsymbol{x}^\star\|_2 \\ &= \frac{\kappa - 1}{\kappa + 1}\|\boldsymbol{x}_k - \boldsymbol{x}^\star\|_2,\end{aligned}$$

where $\kappa = L/\mu$ is the condition number of $\boldsymbol{Q}$.

Since $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \alpha_k \nabla f(\boldsymbol{x}_k)$ and $\nabla f(\boldsymbol{x}^\star) = \boldsymbol{0}$, we have

$$\begin{aligned}\|\boldsymbol{x}_{k+1} - \boldsymbol{x}^\star\|_2 &= \|\boldsymbol{x}_k - \alpha_k \nabla f(\boldsymbol{x}_k) - \boldsymbol{x}^\star\|_2 \\ &= \|\boldsymbol{x}_k - \boldsymbol{x}^\star - \alpha_k \left(\nabla f(\boldsymbol{x}_k) - \nabla f(\boldsymbol{x}^\star)\right)\|_2 \\ &= \|(\mathbf{I} - \alpha_k \boldsymbol{Q})(\boldsymbol{x}_k - \boldsymbol{x}^\star)\|_2 \\ &\leq \|\mathbf{I} - \alpha_k \boldsymbol{Q}\| \cdot \|\boldsymbol{x}_k - \boldsymbol{x}^\star\|_2\end{aligned}$$

Since we have a bound on the eigenvalues of $\boldsymbol{Q}$, we know that the maximum eigenvalue of the symmetric matrix $\mathbf{I} - \alpha_k \boldsymbol{Q}$ is no more than

$$\|\mathbf{I} - \alpha_k \boldsymbol{Q}\| \;\leq\; \max\left(|1 - \alpha_k \mu|, |1 - \alpha_k L|\right).$$

If we take $\alpha_k = 2/(L + \mu)$, we obtain

$$\|\mathbf{I} - \alpha_k \boldsymbol{Q}\| \;\leq\; \frac{L - \mu}{L + \mu} = \frac{\kappa - 1}{\kappa + 1},$$

and so

$$\|\boldsymbol{x}_{k+1} - \boldsymbol{x}^\star\|_2 \leq \left(\frac{\kappa - 1}{\kappa + 1}\right) \|\boldsymbol{x}_k - \boldsymbol{x}^\star\|_2,$$

and by induction on k

$$\|\boldsymbol{x}_k - \boldsymbol{x}^\star\|_2 \leq \left(\frac{\kappa - 1}{\kappa + 1}\right)^k \|\boldsymbol{x}_0 - \boldsymbol{x}^\star\|_2.$$

Heavy ball

For the heavy ball method, we have a similar analysis[3] that ends in a better result. Recall the heavy ball iteration

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \beta_k(\boldsymbol{x}_k - \boldsymbol{x}_{k-1}) - \alpha_k \nabla f(\boldsymbol{x}_k),$$

We will derive a bound on how quickly $\|\boldsymbol{x}_{k+1} - \boldsymbol{x}^\star\|^2 + \|\boldsymbol{x}_k - \boldsymbol{x}^\star\|_2^2$ goes to zero for fixed values of $\alpha_k = \alpha,\ \beta_k = \beta$ which we will choose

[3] These notes are derived from [Wri18].

later. Rewriting the iteration above, we have

$$\underbrace{\begin{bmatrix} \boldsymbol{x}_{k+1} - \boldsymbol{x}^\star \\ \boldsymbol{x}_k - \boldsymbol{x}^\star \end{bmatrix}}_{\boldsymbol{z}_{k+1}} = \begin{bmatrix} \boldsymbol{x}_k + \beta(\boldsymbol{x}_k - \boldsymbol{x}_{k-1}) - \boldsymbol{x}^\star \\ \boldsymbol{x}_k - \boldsymbol{x}^\star \end{bmatrix} - \alpha \begin{bmatrix} \nabla f(\boldsymbol{x}_k) - \nabla f(\boldsymbol{x}^\star) \\ \boldsymbol{0} \end{bmatrix}$$

$$= \begin{bmatrix} \boldsymbol{x}_k + \beta(\boldsymbol{x}_k - \boldsymbol{x}_{k-1}) - \boldsymbol{x}^\star \\ \boldsymbol{x}_k - \boldsymbol{x}^\star \end{bmatrix} - \alpha \begin{bmatrix} \boldsymbol{Q}\left(\boldsymbol{x}_k - \boldsymbol{x}^\star\right) \\ \boldsymbol{0} \end{bmatrix}$$

$$= \underbrace{\begin{bmatrix} (1+\beta)\mathbf{I} - \alpha\boldsymbol{Q} & -\beta\mathbf{I} \\ \mathbf{I} & \boldsymbol{0} \end{bmatrix}}_{\boldsymbol{T}} \underbrace{\begin{bmatrix} \boldsymbol{x}_k - \boldsymbol{x}^\star \\ \boldsymbol{x}_{k-1} - \boldsymbol{x}^\star \end{bmatrix}}_{\boldsymbol{z}_k},$$

We have $\boldsymbol{z}_k = \boldsymbol{T}^k \boldsymbol{z}_0$, and so

$$\|\boldsymbol{z}_k\|_2 \ \leq \ \|\boldsymbol{T}^k\| \cdot \|\boldsymbol{z}_0\|_2,$$

so we want to bound the spectral norm (largest singular value) of $\boldsymbol{T}^k_k$.

We are now analyzing the rate of convergence (to zero) of a linear dynamical system. We know that the eigenvalues of $\boldsymbol{T}^k$ are the eigenvalues of $\boldsymbol{T}$ raised to the kth power. The only complicating factor is that $\boldsymbol{T}$ is not symmetric, and so the eigenvalues and singular values are not the same thing. We reconcile this using the *spectral radius*

$$\rho(\boldsymbol{T}) = \text{maximum magnitude of eigenvalues of } \boldsymbol{T}.$$

Two key results from linear algebra and dynamical systems are that $\rho(\boldsymbol{T}) \leq \|\boldsymbol{T}\|$ and

$$\rho(\boldsymbol{T}) = \lim_{k\to\infty} \|\boldsymbol{T}^k\|^{1/k}.$$

That is, for any given $\delta > 0$, there exists an n such that

$$\|\boldsymbol{T}^k\|^{1/k} \leq \rho(\boldsymbol{T}) + \delta,$$

for all $k \geq n$. Thus if we define the constant

$$C = \max_{0 \leq k \leq n} \frac{\|\boldsymbol{T}^k\|}{(\rho(\boldsymbol{T}) + \delta)^k},$$

we will have

$$\|\boldsymbol{T}^k\| \leq C \left(\rho(\boldsymbol{T}) + \delta\right)^k. \tag{16}$$

We are left with the task of bounding $\rho(\boldsymbol{T}) < 1$ and choosing an appropriate δ. (Note that if $\boldsymbol{T}$ were symmetric, we would simply have $\rho(\boldsymbol{T}) = \|\boldsymbol{T}\|$ and $\|\boldsymbol{T}^k\| = \|\boldsymbol{T}\|^k = \rho(\boldsymbol{T})^k$.)

We can get a start on this by taking an eigenvalue decomposition of the symmetric positive definite matrix $\boldsymbol{Q} = \boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^{\mathrm{T}}$. Since $\boldsymbol{V}\boldsymbol{V}^{\mathrm{T}} = \mathbf{I}$, we can write

$$\begin{bmatrix} (1+\beta)\mathbf{I} - \alpha\boldsymbol{Q} & -\beta\mathbf{I} \\ \mathbf{I} & \mathbf{0} \end{bmatrix} = \begin{bmatrix} \boldsymbol{V} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{V} \end{bmatrix} \begin{bmatrix} (1+\beta)\mathbf{I} - \alpha\boldsymbol{\Lambda} & -\beta\mathbf{I} \\ \mathbf{I} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \boldsymbol{V}^{\mathrm{T}} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{V}^{\mathrm{T}} \end{bmatrix}.$$

Since $\begin{bmatrix} \boldsymbol{V} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{V} \end{bmatrix}$ is orthonormal, its application on the right of a matrix and its transpose (inverse) on the left does not change the eigenvalues, and so we can study the spectral radius of

$$\boldsymbol{T}' = \begin{bmatrix} (1+\beta)\mathbf{I} - \alpha\boldsymbol{\Lambda} & -\beta\mathbf{I} \\ \mathbf{I} & \mathbf{0} \end{bmatrix}.$$

Notice that this a $2N \times 2N$ matrix divided into 4 blocks, each of which is an $N \times N$ diagonal matrix. As such, there is also a permutation matrix $\boldsymbol{P}$ that we can apply on both the rows and columns to make this a block diagonal matrix (with 2×2 blocks along the diagonal):

$$\boldsymbol{P}\boldsymbol{T}'\boldsymbol{P}^{\mathrm{T}} = \begin{bmatrix} \boldsymbol{T}'_1 & & \\ & \cdots & \\ & & \boldsymbol{T}'_N \end{bmatrix}, \quad \boldsymbol{T}'_n = \begin{bmatrix} 1+\beta-\alpha\lambda_n & -\beta \\ 1 & 0 \end{bmatrix}.$$

Since again the application of a matrix and its inverse on either side does not change the eigenvalues, we can compute the spectral radius of the matrix on the right. Since it is block diagonal, we know the spectral radius is the maximum of the individual spectral radii of the blocks. That is, we now have

$$\rho(\boldsymbol{T}) = \max_{1 \leq n \leq N} \rho(\boldsymbol{T}'_n).$$

Since it is a 2×2 matrix, we can compute the eigenvalues of $\boldsymbol{T}'_n$ exactly. We know that γ is an eigenvalue of $\boldsymbol{T}'_n$ if $\det(\boldsymbol{T}'_n - \gamma \mathbf{I}) = 0$, i.e. if

$$\gamma^2 - (1 + \beta - \alpha\lambda_n)\gamma + \beta = 0,$$

which means the eigenvalues are

$$(\gamma_1, \gamma_2) = \frac{1}{2}\left(1 + \beta - \alpha\lambda_n \pm \sqrt{(1 + \beta - \alpha\lambda_n)^2 - 4\beta}\right).$$

If we choose β so that the eigenvalues are complex,

$$4\beta > (1 + \beta - \alpha\lambda_n)^2 \tag{17}$$

then we have

$$(\gamma_1, \gamma_2) = \frac{1}{2}\left(1 + \beta - \alpha\lambda_n \pm j\sqrt{4\beta - (1 + \beta - \alpha\lambda_n)^2}\right),$$

and $|\gamma_1| = |\gamma_2| = \beta$, and hence $\rho(\boldsymbol{T}'_n) = \beta$. Using that fact that $\mu \leq \lambda_n \leq L$, we can ensure (17) holds when

$$\beta = \min(|1 - \sqrt{\alpha\mu}|^2, |1 - \sqrt{\alpha L}|^2).$$

We can now choose α so that these two terms are equal,

$$\alpha = \frac{4}{(\sqrt{L} + \sqrt{\mu})^2} \quad \Rightarrow \quad 1 - \sqrt{\alpha\mu} = -(1 - \sqrt{\alpha L}) = \frac{\sqrt{L} - \sqrt{\mu}}{\sqrt{L} + \sqrt{\mu}},$$

and so

$$\beta = \left(\frac{\sqrt{L}-\sqrt{\mu}}{\sqrt{L}+\sqrt{\mu}}\right)^2 = \left(1 - \frac{2}{\sqrt{\kappa}+1}\right)^2.$$

Taking $\delta = 1/(\sqrt{\kappa}+1)$ in (16) above and using $\beta^2 \leq \beta$, we have

$$\|\boldsymbol{z}_k\|_2 \;\leq\; C\left(1 - \frac{1}{\sqrt{\kappa}+1}\right)^k \|\boldsymbol{z}_0\|_2.$$

This means we are guaranteed that $\|\boldsymbol{z}_k\|_2 \leq \epsilon$ when

$$\begin{aligned} k &\geq (\sqrt{\kappa}+1)\log(C\epsilon_0/\epsilon), \quad \epsilon_0 = \|\boldsymbol{z}_0\|_2, \\ &\gtrsim \sqrt{\kappa}\log(\epsilon_0/\epsilon). \end{aligned}$$

References

[Nes83] Y. Nesterov. A method of solving a convex programming problem with convergence rate $O(1/k^2)$. *Proc. USSR Acad. Sci.*, 269:543–547, 1983.

[Nes04] Y. Nesterov. *Introductory Lectures on Convex Optimization*. Springer Science+Business Media, 2004.

[Pol64] B. Polyak. Some methods of speeding up the convergence of iteration methods. *USSR Computational Mathematics and Mathematical Physics*, 4(5):1–17, 1964.

[She94] J. Shewchuk. An introduction to the conjugate gradient method without the agonizing pain. 1994.

[Wri18] S. Wright. Gradient methods for optimization. Slides, Madison Summer School, July 2018.

Newton's Method

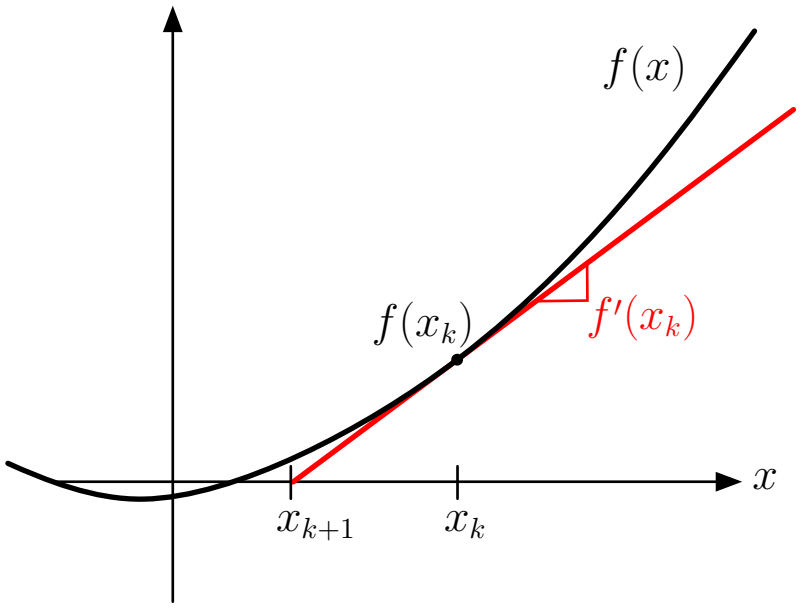
Newton's method is a classical technique for finding the root of a general differentiable function $f(x) : \mathbb{R} \to \mathbb{R}$. That is, we want to find an $x \in \mathbb{R}$ such that

$$f(x) = 0.$$

As you probably learned in high school, one technique for doing this is to start at some guess x_0, and then follow the iteration

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)}.$$

This update results from taking a simple linear approximation at each step:



Of course, there can be many roots, and which one we converge to will depend on what we choose for x_0. It is also very much possible that the iterations do not converge for certain (or even almost all) initial values x_0.

However there is a classical convergence theory that says that once we are close enough to a particular root x_0, we will have

$$\underbrace{|x_0 - x_{k+1}|}_{\epsilon_{k+1}} \leq C \cdot \underbrace{(x_0 - x_k)^2}_{\epsilon_k^2},$$

where the constant C depends on the ratio between the first and second derivatives in the interval[1] around the root x_0:

$$C = \sup_{x \in \mathcal{I}} \frac{|f''(x)|}{2|f'(x)|}.$$

The take-away here is that close to the solution, Newton's methods exhibits *quadratic convergence*: the error at the next iteration is proportional to the square of the error at the last iteration. Since we are concerned with ϵ_k small, $\epsilon_k \ll 1$, this means that under the right conditions, the error goes down in dramatic fashion from iteration to iteration.

Notice that applying the technique requires that f is differentiable, but the convergence guarantee depends on f be twice (continuously) differentiable.

When $f(x)$ is convex, twice differentiable, and has a minimizer, we can find a minimizer by applying Newton's method to the derivative. We start at some initial guess x_0, and then take

$$x_{k+1} = x_k - \frac{f'(x_k)}{f''(x_k)}. \tag{1}$$

[1] There are various technical conditions that f must obey on $\mathcal{I}$ for this result to hold, including the second derivative being continuous and the first derivative not being equal to zero. Also, the condition "close enough" is characterized by looking at ratios of derivatives at the root and on $\mathcal{I}$. The Wikipedia article on this is not bad: https://en.wikipedia.org/wiki/Newton's_method.

Again, if f is three-times continuously differentiable, we converge to the global minimizer quadratically with a constant that depends on

$$C = \sup_{x \in \mathcal{I}} \frac{1}{2} \frac{|f'''(x)|}{|f''(x)|},$$

for an appropriate interval $\mathcal{I}$ around the solution. Again, applying the method relies on us being able to compute first and second derivatives of f, and the analysis relies on f being three-times differentiable.

We can interpret the iteration (1) above in the following way:

1. At x_k, approximate $f(x)$ using the Taylor expansion

$$f(x) \approx f(x_k) + f'(x_k)(x - x_k) + \frac{1}{2} f''(x_k)(x - x_k)^2.$$

2. Find the exact minimizer of this quadratic approximation. Taking the derivative of the expansion above and setting it equal to zero yields the following optimality condition for $\widehat{x}$ to be a minimizer:

$$f''(x_k) \cdot (\widehat{x} - x_k) = -f'(x_k).$$

   This is just a re-arrangement of the iteration (1).

3. Take $x_{k+1} = \widehat{x}$.

This last interpretation extends naturally to the case where $f(\boldsymbol{x})$ is a function of many variables, $f : \mathbb{R}^N \to \mathbb{R}$. We know that if f is convex and twice differentiable, we have a minimizer $\boldsymbol{x}^\star$ when $\nabla f(\boldsymbol{x}^\star) = \mathbf{0}$. Newton's method to find such a minimizer proceeds as above. We start with an initial guess $\boldsymbol{x}_0$, and use the following iteration:

1. Take a Taylor approximation around $f(\boldsymbol{x}_k)$:

$$f(\boldsymbol{x}) \approx f(\boldsymbol{x}_k) + \langle \boldsymbol{x} - \boldsymbol{x}_k, \boldsymbol{g} \rangle + \frac{1}{2}(\boldsymbol{x} - \boldsymbol{x}_k)^{\mathrm{T}} \boldsymbol{H} (\boldsymbol{x} - \boldsymbol{x}_k)$$

where

$$\boldsymbol{g} = \nabla f(\boldsymbol{x}_k) = N \times 1 \text{ gradient vector at } \boldsymbol{x}_k$$
$$\boldsymbol{H} = \nabla^2 f(\boldsymbol{x}_k) = N \times N \text{ Hessian matrix at } \boldsymbol{x}_k.$$

2. Find the exact minimizer $\widehat{\boldsymbol{x}}$ to this approximation. This gives us the problem

$$\underset{\boldsymbol{x} \in \mathbb{R}^N}{\text{minimize}} \ \boldsymbol{g}^{\mathrm{T}}(\boldsymbol{x} - \boldsymbol{x}_k) + \frac{1}{2}(\boldsymbol{x} - \boldsymbol{x}_k)^{\mathrm{T}} \boldsymbol{H} (\boldsymbol{x} - \boldsymbol{x}_k).$$

Since $\boldsymbol{H} \in \mathcal{S}_+^N$ (since we are assuming f is convex), we know that the conditions for $\widehat{\boldsymbol{x}}$ being a minimizer[2] are

$$\boldsymbol{H}(\boldsymbol{x} - \boldsymbol{x}_k) = -\boldsymbol{g}.$$

If $\boldsymbol{H}$ is invertible (i.e., $\boldsymbol{H} \in \mathcal{S}_{++}^N$), then we have a unique minimizer and

$$\widehat{\boldsymbol{x}} = \boldsymbol{x}_k - \boldsymbol{H}^{-1}\boldsymbol{g}.$$

3. Take $\boldsymbol{x}_{k+1} = \widehat{\boldsymbol{x}}$.

This procedure is often referred to as a *pure Newton step*, as it does not involve the selection of a step size. In practice, however, it is often beneficial to choose the step direction as

$$\boldsymbol{d}_k = -\left(\nabla^2 f(\boldsymbol{x}_k)\right)^{-1} \nabla f(\boldsymbol{x}_k),$$

and then choose a step size α_k using a backtracking line search, and then take

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \alpha_k \boldsymbol{d}_k$$

as before.

[2] Take the gradient of this new expression and set it equal to $\mathbf{0}$.

Convergence of Newton's Method

Suppose that $f(\boldsymbol{x})$ is strongly convex,

$$\mu \mathbf{I} \preceq \nabla^2 f(\boldsymbol{x}) \preceq L\mathbf{I}, \quad \forall \boldsymbol{x} \in \mathbb{R}^N,$$

and that its Hessian is Lipschitz,

$$\|\nabla^2 f(\boldsymbol{x}) - \nabla^2 f(\boldsymbol{y})\| \ \leq \ Q\|\boldsymbol{x} - \boldsymbol{y}\|_2.$$

(The norm on the left-hand side above is the standard operator norm.) We will show that the Newton algorithm coupled with an exact line search[3] provides a solution with precision ϵ:

$$f(\boldsymbol{x}_k) - p^\star \leq \epsilon,$$

provided that the number of iterations satisfies

$$k \ \geq \ C_1 \left(f(\boldsymbol{x}_0) - p^\star\right) + \log_2 \log_2(\epsilon_0/\epsilon),$$

where we can take the constants above to be $C_1 = 2L^2Q^2/\mu^5$ and $\epsilon_0 = 2\mu^3/Q^2$. Qualitatively, this says that Newton's method takes a constant number of iterations to converge to any reasonable precision — we can bound $\log_2 \log_2(\epsilon_0/\epsilon) \leq 6$ (say) for ridiculously small values of ϵ.

To establish this result, we break the analysis into two stages. In the first, the *damped Newton stage*, we are far from the solution (as measured by $\|\nabla f(\boldsymbol{x}_k)\|_2$), but we make constant progress towards the answer. Specifically, we will show that in this stage,

$$f(\boldsymbol{x}_k) - f(\boldsymbol{x}_{k+1}) \ \geq \ 1/C_1.$$

[3] These results are easily extended to backtracking line searches; we are just using an exact line search to make the exposition easier. See [BV04, Sec. 9.5.3] for the analysis with backtracking.

This implies that when we are far from the solution, we reduce the gap $f(\boldsymbol{x}_k) - p^\star$ by at least $1/C_1$ at each iteration. It should be clear, then, that the number of damped Newton steps is no greater than $C_1 \left(f(\boldsymbol{x}_0) - p^\star\right)$.

We will then show that when $\|\nabla f(\boldsymbol{x}_k)\|_2$ is small enough, the gap closes dramatically at every iteration. We call this the *quadratic convergence stage*, as we will be able to show that once the algorithm enters this stage at iteration ℓ, for all $k > \ell$,

$$\|\nabla f(\boldsymbol{x}_k)\|_2 \ \leq \ C_2 \cdot 2^{-2^{k-\ell}},$$

where $C_2 = Q/(2\mu^2)$ is another constant.

Damped phase

We are in this stage when

$$\|\nabla f(\boldsymbol{x}_k)\|_2 \geq \mu^2/Q.$$

We take $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \alpha_{\text{exact}} \boldsymbol{d}_k$, where

$$\boldsymbol{d}_k = -\nabla^2 f(\boldsymbol{x}_k)^{-1} \nabla f(\boldsymbol{x}_k),$$

and α_{exact} is the result of an exact line search[4]:

$$\alpha_{\text{exact}} = \operatorname*{arg\,min}_{0 \leq \alpha \leq 1} f(\boldsymbol{x}_k + \alpha \boldsymbol{d}_k).$$

We define the current *Newton decrement* as

$$\lambda_k = \sqrt{\nabla f(\boldsymbol{x}_k)^{\mathrm{T}} (\nabla^2 f(\boldsymbol{x}_k))^{-1} \nabla f(\boldsymbol{x}_k)},$$

[4]For convenience, we are not letting α be larger than 1, just as in a backtracking method.

and note that $\lambda_k^2 = -\nabla f(\boldsymbol{x}_k)^{\mathrm{T}} \boldsymbol{d}_k$. Moreover, strong convexity implies that the eigenvalues of $(\nabla^2 f(\boldsymbol{x}_k))^{-1}$ are at least $1/L$ and at most $1/\mu$, yielding the bounds

$$\|\boldsymbol{d}_k\|_2^2 \leq \frac{1}{\mu}\lambda_k^2 \qquad \text{and} \qquad \frac{1}{L}\|\nabla f(\boldsymbol{x}_k)\|_2^2 \leq \lambda_k^2,$$

which we will use below. From the L-smoothness of the gradient of f, we know that for any t we have

$$\begin{aligned} f(\boldsymbol{x}_k + t\boldsymbol{d}_k) &\leq f(\boldsymbol{x}_k) + \langle t\boldsymbol{d}_k, \nabla f(\boldsymbol{x}_k \rangle + \frac{L}{2}\|t\boldsymbol{d}_k\|_2^2 \\ &= f(\boldsymbol{x}_k) - t\lambda_k^2 + \frac{Lt^2}{2}\|\boldsymbol{d}_k\|_2^2 \\ &\leq f(\boldsymbol{x}_k) - t\lambda_k^2 + \frac{Lt^2}{2\mu}\lambda_k^2. \end{aligned}$$

Plugging in $t = \mu/L$ above yields

$$\begin{aligned} f(\boldsymbol{x}_k + \alpha_{\text{exact}}\boldsymbol{d}_k) - f(\boldsymbol{x}_k) &\leq f\left(\boldsymbol{x}_k + \frac{\mu}{L}\boldsymbol{d}_k\right) - f(\boldsymbol{x}_k) \\ &\leq -\frac{\mu}{2L}\lambda_k^2 \\ &\leq -\frac{\mu}{2L^2}\|\nabla f(\boldsymbol{x}_k)\|_2^2 \\ &\leq -\frac{\mu^5}{2Q^2L^2}. \end{aligned}$$

Quadratic convergence

When

$$\|\nabla f(\boldsymbol{x}_k)\|_2 < \mu^2/Q,$$

we start to settle things very quickly. We will assume that in this stage, we choose the step size to be $\alpha_k = 1$. In fact, you can show that under very mild assumptions on the backtracking parameter ($c < 1/3$, to be specific), backtracking will indeed not backtrack at all and return $\alpha_k = 1$ (see [BV04, p. 490]).

We start by pointing out that by construction,

$$\nabla^2 f(\boldsymbol{x}_k)\boldsymbol{d}_k = -\nabla f(\boldsymbol{x}_k),$$

and so by the fundamental theorem of calculus,

$$\begin{aligned}
\nabla f(\boldsymbol{x}_{k+1}) &= \nabla f(\boldsymbol{x}_k + \boldsymbol{d}_k) - \nabla f(\boldsymbol{x}_k) - \nabla^2 f(\boldsymbol{x}_k)\boldsymbol{d}_k \\
&= \int_0^1 \nabla^2 f(\boldsymbol{x}_k + t\boldsymbol{d}_k)\boldsymbol{d}_k dt - \nabla^2 f(\boldsymbol{x}_k)\boldsymbol{d}_k \\
&= \int_0^1 \left[\nabla^2 f(\boldsymbol{x}_k + t\boldsymbol{d}_k) - \nabla^2 f(\boldsymbol{x}_k)\right] \boldsymbol{d}_k \, dt.
\end{aligned}$$

Thus, we obtain

$$\begin{aligned}
\|\nabla f(\boldsymbol{x}_{k+1})\|_2 &\leq \int_0^1 \|\nabla^2 f(\boldsymbol{x}_k + t\boldsymbol{d}_k) - \nabla^2 f(\boldsymbol{x}_k)\| \cdot \|\boldsymbol{d}_k\|_2 \, dt \\
&\leq \int_0^1 tQ\|\boldsymbol{d}_k\|_2^2 \, dt \\
&= \frac{Q}{2}\|[\nabla^2 f(\boldsymbol{x}_k)]^{-1}\nabla f(\boldsymbol{x}_k)\|_2^2 \\
&\leq \frac{Q}{2\mu^2}\|\nabla f(\boldsymbol{x}_k)\|_2^2,
\end{aligned}$$

where the second inequality follows from the Lipschitz assumption on the Hessian and the last inequality follows from the fact that the maximum eigenvalue of $(\nabla^2 f(\boldsymbol{x}_k))^{-2}$ is less than $1/\mu^2$. Thus we have

$$\frac{Q}{2\mu^2}\|\nabla f(\boldsymbol{x}_{k+1})\|_2 \;\leq\; \left(\frac{Q}{2\mu^2}\|\nabla f(\boldsymbol{x}_k)\|_2\right)^2 \;\leq\; \left(\frac{1}{2}\right)^2,$$

where the last inequality follows since $\|\nabla f(\boldsymbol{x}_k)\|_2 \leq \mu^2/Q$. That is, at every iteration, we are **squaring** the error (which is less than $1/2$). If we entered this stage at iteration ℓ, this means

$$\frac{Q}{2\mu^2}\|\nabla f(\boldsymbol{x}_k)\|_2 \;\leq\; \left(\frac{Q}{2\mu^2}\|\nabla f(\boldsymbol{x}^{(\ell)})\|_2\right)^{2^{k-\ell}} \;\leq\; \left(\frac{1}{2}\right)^{2^{k-\ell}}.$$

Then using the strong convexity of f,

$$f(\boldsymbol{x}_k) - p^\star \;\leq\; \frac{1}{2\mu}\|\nabla f(\boldsymbol{x}_k)\|_2^2 \;\leq\; \frac{2\mu^3}{Q^2}\left(\frac{1}{2}\right)^{2^{k-\ell+1}}.$$

The right hand side above is less than ϵ when

$$k - \ell + 1 \geq \log_2 \log_2(\epsilon_0/\epsilon), \quad \epsilon_0 = 2m^3/L^2,$$

so we spend no more than $\log_2 \log_2(\epsilon_0/\epsilon)$ iterations in this phase.

Note that

$$\epsilon = 10^{-20}\epsilon_0 \quad \Rightarrow \quad \log_2 \log_2(\epsilon_0/\epsilon) = 6.0539.$$

Convergence criteria: the Newton decrement

We know that at the minimia of a smooth convex functional we will have $\nabla f(\boldsymbol{x}) = \mathbf{0}$. So a natural test for convergence is to measure how far away $\nabla f(\boldsymbol{x})$ is from $\mathbf{0}$; that is, we say we are converged when the norm of $\nabla f(\boldsymbol{x})$ is below some threshold (call it ϵ):

$$\text{stop when } \|\nabla f(\boldsymbol{x}_k)\| \leq \epsilon.$$

Which norm should we use?

The natural instinct here is to go with the standard Euclidean (ℓ_2) norm, stopping when

$$\|\nabla f(\boldsymbol{x}_k)\|_2 \leq \epsilon,$$

and in fact, this quantity played a key role in our analysis above. But there is something that is unsatisfying about using the Euclidean norm, and this problem also extends to the way we approached the analysis in the previous section. An interesting feature of Newton's method is that it is *affine invariant*; if we simply change the coordinates, the iterates change accordingly. For example, let $\boldsymbol{T}$ be a $N \times N$ invertible matrix, and set $\tilde{f}(\boldsymbol{x}) = f(\boldsymbol{T}\boldsymbol{x})$. Suppose we run Newton's method to try to find a minima of f starting at $\boldsymbol{x}_0$ and computing iterates $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots$. Then we run Newton's method on $\tilde{f}$ starting at $\boldsymbol{T}^{-1}\boldsymbol{x}_0$ and compute iterates $\widetilde{\boldsymbol{x}}_1, \widetilde{\boldsymbol{x}}_2, \ldots$. This second set of iterates will follow the same progression as the first under transformation by $\boldsymbol{T}$:

$$\tilde{\boldsymbol{x}}_k = \boldsymbol{T}^{-1}\boldsymbol{x}_k, \quad k = 1, 2, \ldots$$

The problem, then, with the the Euclidean norm of the gradient is that it is not affinely invariant:

$$\|\nabla \tilde{f}(\boldsymbol{x})\|_2 \neq \|\nabla f(\boldsymbol{T}\boldsymbol{x})\|_2 \quad \text{for general } \boldsymbol{T}.$$

(Apply the chain rule.)

A criteria that is affinely invariant is the Newton decrement:

$$\lambda(\boldsymbol{x}) = \sqrt{\boldsymbol{g}^{\mathrm{T}} \boldsymbol{H}^{-1} \boldsymbol{g}}, \quad \boldsymbol{g} = \nabla f(\boldsymbol{x}), \quad \boldsymbol{H} = \nabla^2 f(\boldsymbol{x}).$$

(Again, you can work this out with a little effort by applying the chain rule.) These are various ways you can interpret this: one is as size of the gradient in the norm induced by $\boldsymbol{H}^{-1}$:

$$\lambda(\boldsymbol{x}) = \|\nabla f(x)\|_{\boldsymbol{H}^{-1}}.$$

Of course, the norm itself depends on the point $\boldsymbol{x}$. You can also think of it as the directional derivative in the direction we are taking a Newton step; if $\boldsymbol{d} = -(\nabla^2 f(\boldsymbol{x}))^{-1} \nabla f(\boldsymbol{x})$, then

$$\langle \boldsymbol{d}, \nabla f(\boldsymbol{x}) \rangle = -\lambda(\boldsymbol{x})^2.$$

At any rate, the convergence criteria for Newton's method is usually whether $\lambda(\boldsymbol{x}_k)$ is below some threshold.

Self-concordant functions

There is an alternative analysis of Newton's method that is more satisfying in that it gives an affinely invariant bound, and it does not depend on the constants μ, L, Q that are usually unknown. The analysis holds for functions that are self-concordant, a term that we define below.

Definition. We say that a convex function of one variable $f : \mathbb{R} \to \mathbb{R}$ is *self-concordant* if

$$|f'''(x)| \leq 2f''(x)^{3/2}, \quad \text{for all} \quad x \in \operatorname{dom} f.$$

We say that a convex function of multiple variables $f : \mathbb{R}^N \to \mathbb{R}$ is self-concordant if

$$g(t) = f(\boldsymbol{x} + t\boldsymbol{v}) \text{ is self-concordant for all } \boldsymbol{x} \in \operatorname{dom} f, \ \boldsymbol{v} \in \mathbb{R}^N.$$

We should note that the constant 2 that appears in front of the $f''(x)$ above is somewhat arbitrary — if there is any uniform bound on the ratio of $|f'''(x)|$ to $f''(x)^{3/2}$, then f can be made self-concordant simply by re-scaling.

We mention a few important examples (see [BV04, Chapter 9.6] for many more).

- Since the third derivative of all linear and quadratic functionals is zero, they are self-concordant.
- $f(x) = -\log(x)$ is self-concordant
- $f(\boldsymbol{X}) = -\log\det \boldsymbol{X}$ for $\boldsymbol{X} \in S^N_{++}$ is self-concordant
- Self-concordance is preserved under composition with an affine

transformation, so for example

$$f(\boldsymbol{x}) = -\sum_{m=1}^{M} \log(b_m - \boldsymbol{a}_m^{\mathrm{T}}\boldsymbol{x}) \quad \text{on } \{\boldsymbol{x} : \boldsymbol{a}_m^{\mathrm{T}} \leq b_m,\ m = 1, \ldots, M\}$$

is self-concordant. Functions of the above form will play a major role when we talk about log-barrier methods for contrained optimization.

Using a line of argumentation not too different than in the classical analysis in the last section, we have the following result for the convergence of Newton's method (again, see [BV04, Chapter 9] for the details):

If $f(\boldsymbol{x}) : \mathbb{R}^N \to \mathbb{R}$ is self-concordant, then Newton iterations starting from $\boldsymbol{x}_0$ coupled with standard backtracking line search will have

$$f(\boldsymbol{x}_k) - p^\star \leq \epsilon$$

when

$$k \geq C\epsilon_0 + \log_2 \log_2(1/\epsilon), \quad \epsilon_0 = f(\boldsymbol{x}_0) - p^\star.$$

The constant C above depends only on the backtracking parameters.

You may more fully appreciate this result when we talk about log barrier techniques a little later.

References

[BV04] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

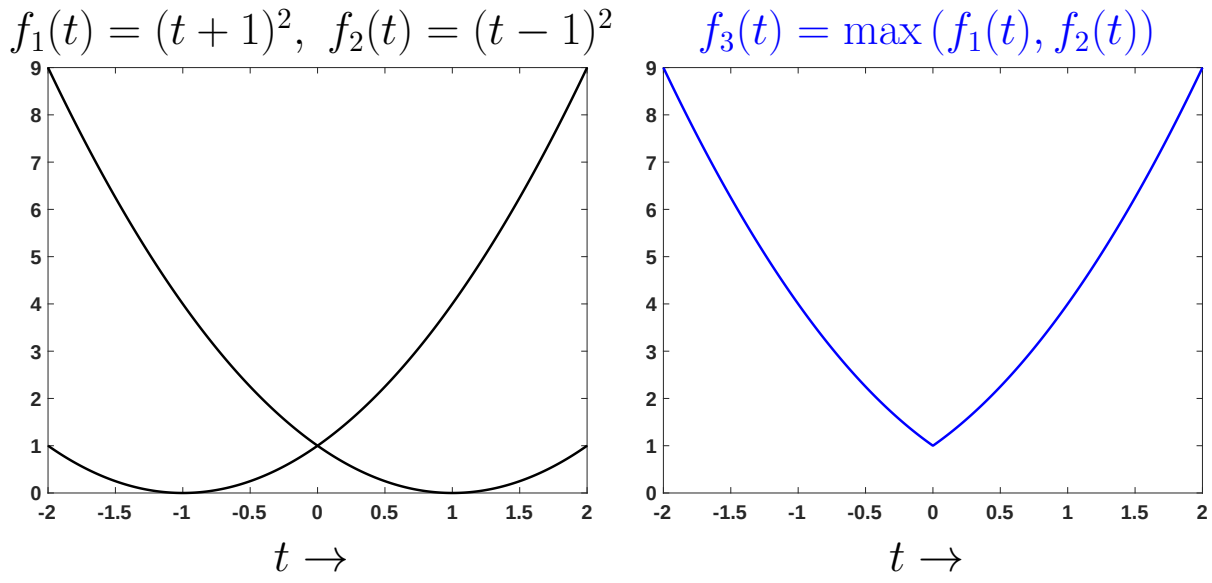
Nonsmooth optimization

Most of the theory and algorithms that we have explored for convex optimization have assumed that the functions involved are differentiable – that is, smooth.

This is not always the case in interesting applications. In fact, nonsmooth functions can arise quite naturally in applications. We already have looked at optimization programs involving the hinge loss $\max(\boldsymbol{a}^{\mathrm{T}}\boldsymbol{x} + b, 0)$, the ℓ_1 norm, the ℓ_∞ norm, and the nuclear norm — none of these is differentiable. As another example, suppose $f_1, \ldots, f_Q$ are all perfectly smooth convex functions. Then the pointwise maximum

$$f(\boldsymbol{x}) = \max_{1 \le q \le Q} \; f_q(\boldsymbol{x})$$

is in general not smooth.

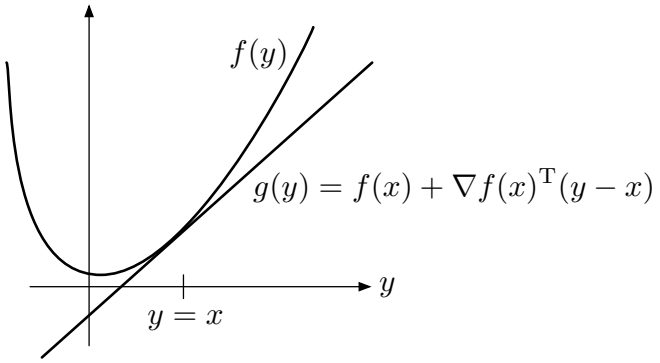


Fortunately, the theory for nonsmooth optimization is not too different than for smooth optimization. We really just need one new concept: that of a subgradient.

Subgradients

If you look back through the notes so far, you will see that the vast majority of the time we use the gradient of a convex function, it is in the context of the inequality

$$f(\boldsymbol{y}) \;\geq\; f(\boldsymbol{x}) + \langle \boldsymbol{y} - \boldsymbol{x}, \nabla f(\boldsymbol{x}) \rangle, \quad \text{for all } \boldsymbol{x}, \boldsymbol{y} \in \operatorname{dom} f.$$



This is a very special property of convex functions, and it led to all kinds of beautiful results.

When a convex f is not differentiable at a point $\boldsymbol{x}$, we can more or less reproduce the entire theory using subgradients. A **subgradient** of f at $\boldsymbol{x}$ is a vector $\boldsymbol{g}$ such that

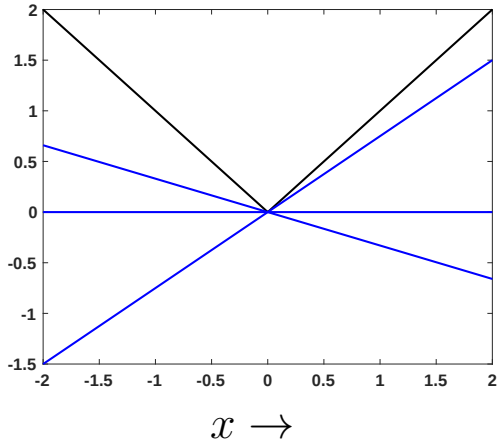
$$f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \langle \boldsymbol{y} - \boldsymbol{x}, \boldsymbol{g} \rangle, \quad \text{for all } \boldsymbol{y} \in \operatorname{dom} f.$$

Unlike gradients for smooth functions, there can be more than one subgradient of a nonsmooth function at a point. We call the collection of subgradients the **subdifferential** at $\boldsymbol{x}$:

$$\partial f(\boldsymbol{x}) = \{\boldsymbol{g} \;:\; f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \langle \boldsymbol{y} - \boldsymbol{x}, \boldsymbol{g} \rangle, \quad \text{for all } \boldsymbol{y} \in \operatorname{dom} f\}.$$

Example:

$$f(x) = |x|, \quad \partial f(x) = \begin{cases} -1, & x < 0 \\ [-1, 1], & x = 0 \\ 1, & x > 0. \end{cases}$$



black: $f(x) = |x|$
blue: $f(0) + g(x - 0)$ for a few $g \in \partial f(0)$

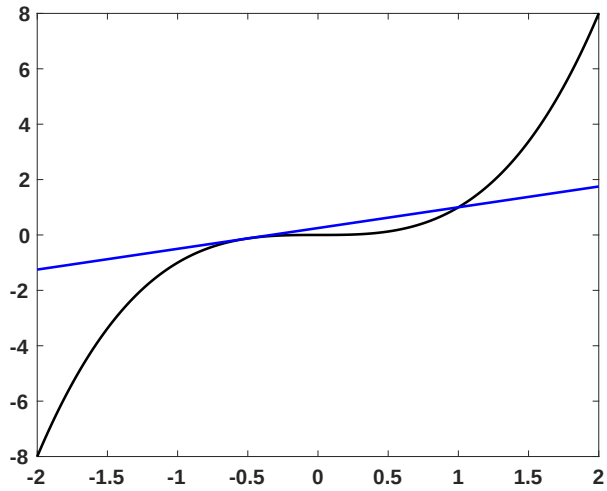
Facts for subdifferentials of convex functions:

1. If f is convex and differentiable at $\boldsymbol{x}$, then the subdifferential contains exactly one vector: the gradient,

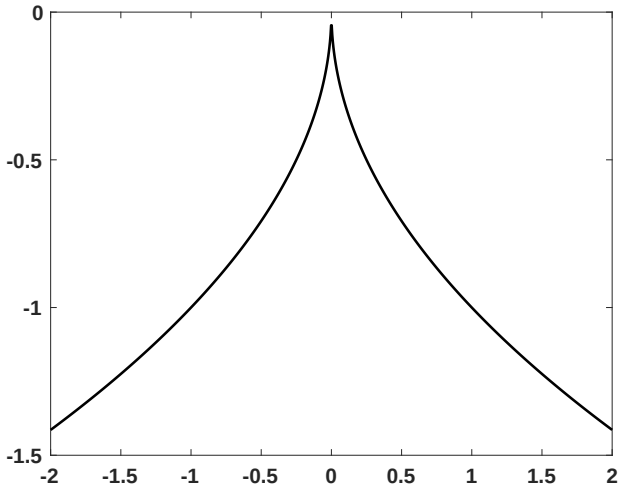
$$\partial f(\boldsymbol{x}) = \{\nabla f(\boldsymbol{x})\}.$$

2. If f is convex on $\operatorname{dom} f$, then the subdifferential is non-empty and bounded at all $\boldsymbol{x}$ in the interior of $\operatorname{dom} f$.

For non-convex f, these two points do not hold in general. The gradient at a point is not necessarily a subgradient:



and there can also be points where neither the gradient nor subgradient exist, e.g. $f(x) = -\sqrt{|x|}$ for $x \in \mathbb{R}$



Example: The ℓ_1 norm

Consider the function

$$f(\boldsymbol{x}) = \|\boldsymbol{x}\|_1.$$

The ℓ_1 norm is not differentiable at any $\boldsymbol{x}$ that has at least one coordinate equal to zero. We will see that optimization problems involving the ℓ_1 norm very often have solutions that are sparse, meaning that they have many zeros. This is a big problem – the nonsmoothness is kicking in at exactly the points we are interested in.

What does the subdifferential $\partial\|\boldsymbol{x}\|_1$ look like in such a case? To see, recall that by definition, if a vector $\boldsymbol{u} \in \partial\|\boldsymbol{x}\|_1$, at the point $\boldsymbol{x}$, then we must have

$$\|\boldsymbol{y}\|_1 \geq \|\boldsymbol{x}\|_1 + \langle \boldsymbol{y} - \boldsymbol{x}, \boldsymbol{u} \rangle \tag{1}$$

for all $\boldsymbol{y} \in \mathbb{R}^N$. To understand what this means in terms of $\boldsymbol{x}$, it is useful to introduce the notation $\Gamma(\boldsymbol{x})$ to denote the set of indexes where $\boldsymbol{x}$ is non-zero:

$$\Gamma(\boldsymbol{x}) = \{n \ : \ x_n \neq 0\}.$$

Using this, we can re-write the right-hand side of (1) as

$$\begin{aligned}
\|\boldsymbol{x}\|_1 + \langle \boldsymbol{y} - \boldsymbol{x}, \boldsymbol{u} \rangle &= \sum_{n=1}^{N} |x_n| + \sum_{n=1}^{N} u_n(y_n - x_n) \\
&= \sum_{n \in \Gamma} |x_n| - u_n x_n + \sum_{n=1}^{N} u_n y_n.
\end{aligned}$$

Note that if

$$u_n = \operatorname{sign}(x_n) = \begin{cases} 1 & \text{if } x_n \geq 0, \\ -1 & \text{if } x_n < 0, \end{cases}$$

then $u_n x_n = |x_n|$. Thus, if $u_n = \text{sign}(x_n)$ for all $n \in \Gamma$, we have

$$\sum_{n \in \Gamma} |x_n| - u_n x_n = \sum_{n \in \Gamma} |x_n| - |x_n| = 0.$$

Thus, if we set $u_n = \text{sign}(x_n)$ for all $n \in \Gamma$, then (1) reduces to

$$\|\boldsymbol{y}\|_1 \geq \langle \boldsymbol{y}, \boldsymbol{u} \rangle.$$

As long as $|u_n| \leq 1$ for all n, then this will hold. Hence, if a vector $\boldsymbol{u}$ satisfies

$$\begin{aligned} u_n &= \text{sign}(x_n) && \text{if } n \in \Gamma, \\ |u_n| &\leq 1 && \text{if } n \notin \Gamma, \end{aligned}$$

then $\boldsymbol{u} \in \partial \|\boldsymbol{x}\|_1$. It is not hard to show that for any $\boldsymbol{u}$ that violates these conditions, we can construct a $\boldsymbol{y}$ such that (1) is violated, and thus this is a complete description of all vectors in $\boldsymbol{u} \in \partial \|\boldsymbol{x}\|_1$.

Example: The ℓ_2 norm

While the function $\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x} = \|\boldsymbol{x}\|_2^2$ is the prototypical differentiable ($\nabla f(\boldsymbol{x}) = \boldsymbol{x}$), smooth, and strongly convex function ($\nabla^2 f(\boldsymbol{x}) = \mathbf{I}$), the function $f(\boldsymbol{x}) = \|\boldsymbol{x}\|_2$ is not as nice; it is not strongly convex, and it is not differentiable at $\boldsymbol{x} = \mathbf{0}$ (to appreciate this latter point, consider that a 1D slice of the function $s(t) = \|t\boldsymbol{v}\|_2 = |t|\|\boldsymbol{v}\|_2$ looks like the absolute value function as function of t).

For $\boldsymbol{x} \neq \mathbf{0}$, an easy calculation[1] shows that

$$\nabla \|\boldsymbol{x}\|_2 = \frac{\boldsymbol{x}}{\|\boldsymbol{x}\|_2}.$$

At $\boldsymbol{x} = \mathbf{0}$, we know that $\boldsymbol{u} \in \partial\|\boldsymbol{x}\|_2$ if

$$\|\boldsymbol{y}\|_2 \geq \|\mathbf{0}\|_2 + \langle \boldsymbol{y} - \mathbf{0}, \boldsymbol{u} \rangle = \langle \boldsymbol{y}, \boldsymbol{u} \rangle \quad \text{for all} \quad \boldsymbol{y} \in \mathbb{R}^N. \tag{2}$$

We can find $\boldsymbol{u}$ that meet these conditions using the Cauchy-Schwarz inequality. Note that

$$\langle \boldsymbol{y}, \boldsymbol{u} \rangle \leq \|\boldsymbol{y}\|_2 \|\boldsymbol{u}\|_2,$$

so (2) will hold when $\|\boldsymbol{u}\|_2 \leq 1$. On the other hand, if $\|\boldsymbol{u}\|_2 > 1$, then for $\boldsymbol{y} = \boldsymbol{u}$, we have

$$\langle \boldsymbol{y}, \boldsymbol{u} \rangle = \|\boldsymbol{y}\|_2^2 > \|\boldsymbol{y}\|_2,$$

and (2) does not hold. Therefore

$$\partial\|\boldsymbol{x}\|_2 = \begin{cases} \{\boldsymbol{u} \ : \ \|\boldsymbol{u}\|_2 \leq 1\}, & \boldsymbol{x} = \mathbf{0} \\ \frac{\boldsymbol{x}}{\|\boldsymbol{x}\|_2}, & \boldsymbol{x} \neq \mathbf{0}. \end{cases}$$

[1] Use the fact that $\frac{d}{dx}\sqrt{x^2 + a} = x/\sqrt{x^2 + a}$.

General norms at $\boldsymbol{x} = \mathbf{0}$

Norms in general are not differentiable at $\boldsymbol{x} = \mathbf{0}$, again because they look like an absolute value function along a line: $s(t) = \|t\boldsymbol{v}\| = |t| \cdot \|\boldsymbol{v}\|$ for any valid norm $\|\cdot\|$. We can generalize the result for the ℓ_2 norm at $\boldsymbol{x} = \mathbf{0}$ using the concept of a **dual norm**.

The dual norm $\|\cdot\|_*$ of a norm $\|\cdot\|$ is

$$\|\boldsymbol{y}\|_* = \max_{\|\boldsymbol{x}\| \leq 1} \langle \boldsymbol{x}, \boldsymbol{y} \rangle.$$

Since sublevel sets of norms in $\mathbb{R}^N$ are compact, we know that this maximum is achieved, and it is an easy exercise to show that $\|\cdot\|_*$ is a valid norm. You can also verify the following easy facts at home

- the dual of $\|\cdot\|_2$ is again $\|\cdot\|_2$
- the dual of $\|\cdot\|_1$ is $\|\cdot\|_\infty$
- the dual of $\|\cdot\|_\infty$ is $\|\cdot\|_1$

It is also a fact (for norms on $\mathbb{R}^N$) that the dual of $\|\cdot\|_*$ is the original norm $\|\cdot\|$, i.e. $\|\boldsymbol{x}\|_{**} = \|\boldsymbol{x}\|$. We also have the generalized Cauchy-Schwarz inequality

$$|\langle \boldsymbol{x}, \boldsymbol{y} \rangle| \ \leq \ \|\boldsymbol{x}\| \cdot \|\boldsymbol{y}\|_*.$$

We can use these facts with an argument similar to the ℓ_2 case above to compute the subdifferential of any norm at $\mathbf{0}$ as

$$\partial \|\mathbf{0}\| = \{\boldsymbol{u} \ : \ \|\boldsymbol{u}\|_* \leq 1\}.$$

Properties of subdifferentials

Here are some properties of the subdifferential that we will state without proof (but are easy to prove). Below, we assume that all functions are well-defined on all of $\mathbb{R}^N$.

Summation: If $f(\boldsymbol{x}) = f_1(\boldsymbol{x}) + f_2(\boldsymbol{x})$, then

$$\partial f(\boldsymbol{x}) = \partial f_1(\boldsymbol{x}) + \partial f_2(\boldsymbol{x}).$$

That is, the set of all subgradients (at $\boldsymbol{x}$) of f is the set of vectors that can be written as a sum of a vector from $\partial f_1(\boldsymbol{x})$ plus a vector from $\partial f_2(\boldsymbol{x})$.

Chain rule for affine transformations: If $h(\boldsymbol{x}) = f(\boldsymbol{A}\boldsymbol{x}+\boldsymbol{b})$, then

$$\partial h(\boldsymbol{x}) = \boldsymbol{A}^{\mathrm{T}} \partial f(\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}).$$

That is, we compute the subgradients of f at the point $\boldsymbol{A}\boldsymbol{x}+\boldsymbol{b}$, then map them through $\boldsymbol{A}^{\mathrm{T}}$.

Max of functions: If $f(\boldsymbol{x}) = \max\{f_1(\boldsymbol{x}), \ldots, f_M(\boldsymbol{x})\}$, then

$$\partial f(\boldsymbol{x}) = \operatorname{conv}\left(\bigcup_{m \in \Gamma(\boldsymbol{x})} \partial f_i(\boldsymbol{x}) \right),$$

where $\Gamma(\boldsymbol{x}) = \{m \ : \ f_m(\boldsymbol{x}) = f(\boldsymbol{x})\}$, and conv takes the convex hull:

$$\operatorname{conv}(\mathcal{X}) = \left\{ \sum_{p=1}^{P} \lambda_p \boldsymbol{x}_p, \ \boldsymbol{x}_p \in \mathcal{X}, \lambda_p \geq 0, \sum_{p=1}^{P} \lambda_p = 1, \ \forall P \right\}$$

Exercise: Compute $\partial f(\boldsymbol{x})$ for $f(\boldsymbol{x}) = \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_1$.

Answer: Set $\Gamma(\boldsymbol{x}) = \{m \ : \ \boldsymbol{a}_m^{\mathrm{T}}\boldsymbol{x} = y_m\}$, where $\boldsymbol{a}_m^{\mathrm{T}}$ is the mth row of $\boldsymbol{A}$. Then $\partial f(\boldsymbol{x})$ is the set of vectors that can be written as

$$\boldsymbol{u} = \sum_{m \notin \Gamma(\boldsymbol{x})} \mathrm{sgn}(\boldsymbol{a}_m^{\mathrm{T}}\boldsymbol{x} - y_m)\boldsymbol{a}_m + \sum_{m \in \Gamma(\boldsymbol{x})} \beta_m \boldsymbol{a}_m$$

for any β_m with $|\beta_m| \leq 1$.

Exercise: Compute $\partial f(x)$ for $f(x) = \max(x, 0)$.

Answer:

$$\partial f(x) = \begin{cases} 0, & x < 0, \\ [0, 1], & x = 0, \\ 1, & x > 0. \end{cases}$$

Exercise: Compute $\partial f(x)$ for $f(x) = \max((x+1)^2, (x-1)^2)$.

Answer:

$$\partial f(x) = \begin{cases} 2(x-1) & x < 0, \\ [-2, 2], & x = 0, \\ 2(x+1), & x > 0. \end{cases}$$

Exercise: Compute $\partial f(\boldsymbol{x})$ for $f(\boldsymbol{x}) = \|\boldsymbol{x}\|_\infty$.

Optimality conditions for unconstrained optimization

With the right definition in place, it is very easy to re-derive the central mathematical results in this course for general[2] convex functions.

Let $f(\boldsymbol{x})$ be a general convex function. Then $\boldsymbol{x}^\star$ is a solution to the unconstrained problem

$$\underset{\boldsymbol{x}\in\mathbb{R}^N}{\text{minimize}} \;\; f(\boldsymbol{x})$$

if and only if

$$\mathbf{0} \in \partial f(\boldsymbol{x}^\star).$$

The proof of this statement is so easy you could do it in your sleep. Suppose $\mathbf{0} \in \partial f(\boldsymbol{x}^\star)$. Then

$$\begin{aligned} f(\boldsymbol{y}) &\geq f(\boldsymbol{x}^\star) + \langle \boldsymbol{y} - \boldsymbol{x}^\star, \mathbf{0} \rangle \\ &= f(\boldsymbol{x}^\star) \end{aligned}$$

for all $\boldsymbol{y} \in \operatorname{dom} f$. Thus $\boldsymbol{x}^\star$ is optimal. Likewise, if $f(\boldsymbol{y}) \geq f(\boldsymbol{x}^\star)$ for all $\boldsymbol{y} \in \operatorname{dom} f$, then of course it must also be true that $f(\boldsymbol{y}) \geq f(\boldsymbol{x}^\star) + \langle \boldsymbol{y} - \boldsymbol{x}, \mathbf{0} \rangle$ for all $\boldsymbol{y}$, and so $\mathbf{0} \in \partial f(\boldsymbol{x}^\star)$.

Example: The LASSO

Consider the ℓ_1 regularized least-squares problem

$$\underset{\boldsymbol{x}\in\mathbb{R}^N}{\text{minimize}} \;\; \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_2^2 + \tau\|\boldsymbol{x}\|_1.$$

[2] Meaning not necessarily differentiable.

We can quickly translate the general result $\mathbf{0} \in \partial f(\boldsymbol{x}^\star)$ into a useful set of optimality conditions. We need to compute the subdifferential of $f(\boldsymbol{x}) = \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_2^2 + \tau\|\boldsymbol{x}\|_1$. The first term is smooth, so the subdifferential just contains the gradient:

$$\partial f(\boldsymbol{x}) = \boldsymbol{A}^{\mathrm{T}}(\boldsymbol{A}\boldsymbol{x} - \boldsymbol{y}) + \tau\partial\|\boldsymbol{x}\|_1.$$

As shown above $\partial\|\boldsymbol{x}\|_1$ is the set of all vectors $\boldsymbol{u}$ such that

$$\begin{aligned} u_n &= \operatorname{sign}(x_n) && \text{if } x_n \neq 0, \\ |u_n| &\leq 1 && \text{if } x_n = 0. \end{aligned}$$

Thus the optimality condition

$$\mathbf{0} \in \boldsymbol{A}^{\mathrm{T}}(\boldsymbol{A}\boldsymbol{x}^\star - \boldsymbol{y}) + \tau\partial\|\boldsymbol{x}^\star\|_1,$$

means that $\boldsymbol{x}^\star$ is optimal if and only if

$$\begin{aligned} \boldsymbol{a}_n^{\mathrm{T}}(\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}^\star) &= \tau \operatorname{sign} x_n^\star && \text{if } x_n^\star \neq 0, \\ |\boldsymbol{a}_n^{\mathrm{T}}(\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}^\star)| &\leq \tau && \text{if } x_n^\star = 0. \end{aligned}$$

where here $\boldsymbol{a}_n$ is the n^{th} column of $\boldsymbol{A}$.

Note that this doesn't quite give us a closed form expression for $\boldsymbol{x}^\star$ (except when $\boldsymbol{A}$ is an orthonormal matrix), but it is useful both algorithmically (for checking if a candidate $\boldsymbol{x}$ is a solution) and theoretically (for understanding and analyzing the properties of the solution to this optimization problem.)

The subgradient method

The subgradient method is the non-smooth version of gradient descent. The basic algorithm is straightforward, consisting of the iteration

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \alpha_k \boldsymbol{d}_k, \tag{3}$$

where $\boldsymbol{d}_k$ is *any* subgradient at $\boldsymbol{x}_k$, i.e., $\boldsymbol{d}_k \in \partial f(\boldsymbol{x}_k)$. Of course, there could be many choices for $\boldsymbol{d}_k$ at every step, and the progress you make at that iteration could very dramatically with this choice. Making this determination, though, is often very difficult, and whether or not it can even be done it very problem dependent. Thus the analytical results for the subgradient method just assume we have any subgradient at a particular step.

With the right choice of step sizes $\{\alpha_k\}$, some simple analysis (which we will get to in a minute) shows that the subgradient method converges. The convergence rate, though, is very slow. This is also evidenced in most practical applications of this method: it can take many iterations on even a medium-sized problem to arrive at a solution that is even close to optimal.

Here is what we know about this algorithm for solving the general unconstrained program

$$\underset{\boldsymbol{x}\in\mathbb{R}^N}{\text{minimize}}\, f(\boldsymbol{x}). \tag{4}$$

We will look at one particular case here; for more detailed results see [Nes04, Chapter 3]. Along with f being convex, we will assume that it has at least one minimizer. The results also assume that f is Lipschitz:

$$|f(\boldsymbol{x}) - f(\boldsymbol{y})| \;\leq\; L\|\boldsymbol{x} - \boldsymbol{y}\|_2.$$

Note that here we are assuming that f is Lipschitz, not that f has Lipschitz gradients (since the gradient does not even necessarily exist). A direct consequence of f being Lipschitz is that the norms of the subgradients are bounded:

$$\|\boldsymbol{d}\|_2 \leq L, \quad \text{for all } \boldsymbol{d} \in \partial f(\boldsymbol{x}), \quad \text{for all } \boldsymbol{x} \in \mathbb{R}^N. \tag{5}$$

The results below used pre-determined step sizes. Thus the iteration (3) does not necessarily decrease the functional $f(\boldsymbol{x})$ at every step. We will keep track of the best value we have up to the current iteration with

$$f_k^{\text{best}} = \min \left\{ f(\boldsymbol{x}_i), \quad 0 \leq i < k \right\}.$$

We will let $\boldsymbol{x}^\star$ be any solution to (4) and set $f^\star = f(\boldsymbol{x}^\star)$.

Our analytical results stem from a careful look at what happens during a single iteration. Note that

$$\begin{aligned} \|\boldsymbol{x}_{i+1} - \boldsymbol{x}^\star\|_2^2 &= \|\boldsymbol{x}_i - \alpha_i \boldsymbol{d}_i - \boldsymbol{x}^\star\|_2^2 \\ &= \|\boldsymbol{x}_i - \boldsymbol{x}^\star\|_2^2 - 2\alpha_i \langle \boldsymbol{x}_i - \boldsymbol{x}^\star, \boldsymbol{d}_i \rangle + \alpha_i^2 \|\boldsymbol{d}_i\|_2^2 \\ &\leq \|\boldsymbol{x}_i - \boldsymbol{x}^\star\|_2^2 - 2\alpha_i (f(\boldsymbol{x}_i) - f^\star) + \alpha_i^2 \|\boldsymbol{d}_i\|_2^2, \end{aligned}$$

where the inequality follows from the definition of a subgradient:

$$f^\star \geq f(\boldsymbol{x}_i) + \langle \boldsymbol{x}^\star - \boldsymbol{x}_i, \boldsymbol{d}_i \rangle.$$

Rearranging the bound above we have

$$2\alpha_i \left(f(\boldsymbol{x}_i) - f^\star \right) \leq \|\boldsymbol{x}_i - \boldsymbol{x}^\star\|_2^2 - \|\boldsymbol{x}_{i+1} - \boldsymbol{x}^\star\|_2^2 + \alpha_i^2 \|\boldsymbol{d}_i\|_2^2,$$

and so of course

$$2\alpha_i \left(f_i^{\text{best}} - f^\star \right) \leq \|\boldsymbol{x}_i - \boldsymbol{x}^\star\|_2^2 - \|\boldsymbol{x}_i - \boldsymbol{x}^\star\|_2^2 + \alpha_i^2 \|\boldsymbol{d}_i\|_2^2.$$

Since f_i^{best} is monotonically decreasing, at iteration k we have

$$2\alpha_i \left(f_k^{\text{best}} - f^\star\right) \leq \|\boldsymbol{x}_i - \boldsymbol{x}^\star\|_2^2 - \|\boldsymbol{x}_{i+1} - \boldsymbol{x}^\star\|_2^2 + \alpha_i^2 \|\boldsymbol{d}_i\|_2^2,$$

for all $i \leq k$. To understand what has happened after k iterations, we sum both sides of the expression above from $i = 0$ to $i = k - 1$. Notice that the two error terms on the right hand side give us the telescoping sum:

$$\begin{aligned}\sum_{i=0}^{k-1} \left(\|\boldsymbol{x}_i - \boldsymbol{x}^\star\|_2^2 - \|\boldsymbol{x}_{i+1} - \boldsymbol{x}^\star\|_2^2\right) &= \|\boldsymbol{x}_0 - \boldsymbol{x}^\star\|_2^2 - \|\boldsymbol{x}_k - \boldsymbol{x}^\star\|_2^2 \\ &\leq \|\boldsymbol{x}_0 - \boldsymbol{x}^\star\|_2^2\end{aligned}$$

and so

$$f_k^{\text{best}} - f^\star \leq \frac{\|\boldsymbol{x}_0 - \boldsymbol{x}^\star\|_2^2 + \sum_{i=0}^{k-1} \alpha_i^2 \|\boldsymbol{d}_i\|_2^2}{2\sum_{i=0}^{k-1} \alpha_i} \tag{6}$$

We can now specialize this result to general step-size strategies.

Fixed step size. Suppose that $\alpha_k = \alpha > 0$ for all k. Then (6) becomes

$$f_k^{\text{best}} - f^\star \leq \frac{\|\boldsymbol{x}_0 - \boldsymbol{x}^\star\|_2^2}{2k\alpha} + \frac{L^2\alpha}{2},$$

where we have also used the Lipschitz property (5). Note that in this case, no matter how small we choose α, **the subgradient algorithm is not guaranteed to converge**. This is, in fact, standard in practice as well. The problem is that, unlike gradients for smooth functions, the subgradients do not have to vanish as we approach the solution. Even at the solution, there can be subgradients that are large.

Fixed step length. A similar result holds if we always move the same amount, taking

$$\alpha_k = s/\|\boldsymbol{d}_k\|_2.$$

This means that

$$\|\boldsymbol{x}_{k+1} - \boldsymbol{x}_k\|_2 = s.$$

Of course, with this strategy it is self-evident that it will never converge, since we move some fixed amount at every step. We can bound the suboptimality at step k as

$$f_k^{\text{best}} - f^\star \leq \frac{L\|\boldsymbol{x}_0 - \boldsymbol{x}^\star\|_2^2}{2ks} + \frac{Ls}{2},$$

which is not necessarily worse than the fixed step size result. In fact, notice that even though you are moving some fixed amount, you will never move too far from an optimal point.

Decreasing step size. The results above suggest that we might want to decrease the step size as k increases, so we can get rid of this constant offset term. To make the terms in (6) work out, we let $\alpha_k \to 0$, but not too fast. Specifically, we choose a sequence $\{\alpha_k\}$ such that

$$\sum_{k=1}^{\infty} \alpha_k = \infty, \quad \text{and} \quad \frac{\sum_{i=0}^{k-1} \alpha_i^2}{\sum_{i=0}^{k-1} \alpha_i} \to 0.$$

Looking at (6) above, we can see that under these conditions $f_k^{\text{best}} \to f^\star$. It is an exercise (but a nontrivial one) to show that it is enough to choose $\{\alpha_k\}$ such that

$$\alpha_k \to 0 \text{ as } k \to \infty, \quad \text{and} \quad \sum_{k=1}^{\infty} \alpha_k = \infty. \tag{7}$$

To get an idea of the tradeoffs involved here, suppose that $\alpha_k = \alpha/(k+1)$. Then for large k, we have the approximations

$$\sum_{i=0}^{k-1} \alpha_i \sim \alpha \log k, \quad \text{and} \quad \sum_{i=0}^{k-1} \alpha_i^2 \sim \text{Const} = \alpha^2 \pi^2/6$$

that are good as upper and lower bounds to within constants. In this case, the convergence result (6) becomes

$$f_k^{\text{best}} - f^\star \lesssim \frac{\|\boldsymbol{x}_0 - \boldsymbol{x}^\star\|_2^2}{\alpha \log k} + \text{Const} \cdot \frac{\alpha L^2}{\log k}.$$

So the convergence is extraordinarily slow – *logarithmic* in k.

You can get much better rates than this (but still not great) by decreasing the stepsize more slowly. Consider now $\alpha_k = \alpha/\sqrt{k+1}$. Then for large k

$$\sum_{i=0}^{k-1} \alpha_i \sim (\alpha+1)\sqrt{k}, \quad \text{and} \quad \sum_{i=0}^{k-1} \alpha_i^2 \sim \alpha^2 \log k,$$

and so

$$f_k^{\text{best}} - f^\star \lesssim \frac{\|\boldsymbol{x}_0 - \boldsymbol{x}^\star\|_2^2}{(\alpha+1)\sqrt{k}} + \text{Const} \cdot \frac{\alpha L^2 \log k}{\sqrt{k}}.$$

This is something like $O(1/\sqrt{k})$ convergence. This means that if we want to guarantee $f_k^{\text{best}} - f^\star \leq \epsilon$, we need $k = O(1/\epsilon^2)$ iterations.

In [Nes04, Chapter 3], it is shown that there is no better rate of convergence than $O(1/\sqrt{k})$ that holds uniformly across all problems.

Example. Consider the "ℓ_1 approximation problem"

$$\underset{\boldsymbol{x} \in \mathbb{R}^N}{\text{minimize}} \ \|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\|_1.$$

We have already looked at the subdifferential of $\|\boldsymbol{x}\|_1$. Specifically, we showed that $\boldsymbol{u}$ is a subgradient of $\|\boldsymbol{x}\|_1$ at $\boldsymbol{x}$ if it satisfies

$$\begin{aligned} u_n &= \text{sign}(x_n) \quad \text{if } x_n \neq 0, \\ |u_n| &\leq 1 \qquad\quad \text{if } x_n = 0. \end{aligned}$$

In the exercise above, we also derived the subdifferential for $\|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\|_1$. We quickly re-derive it here using "guess and check". First consider a vector $\boldsymbol{z}$ that satisfies

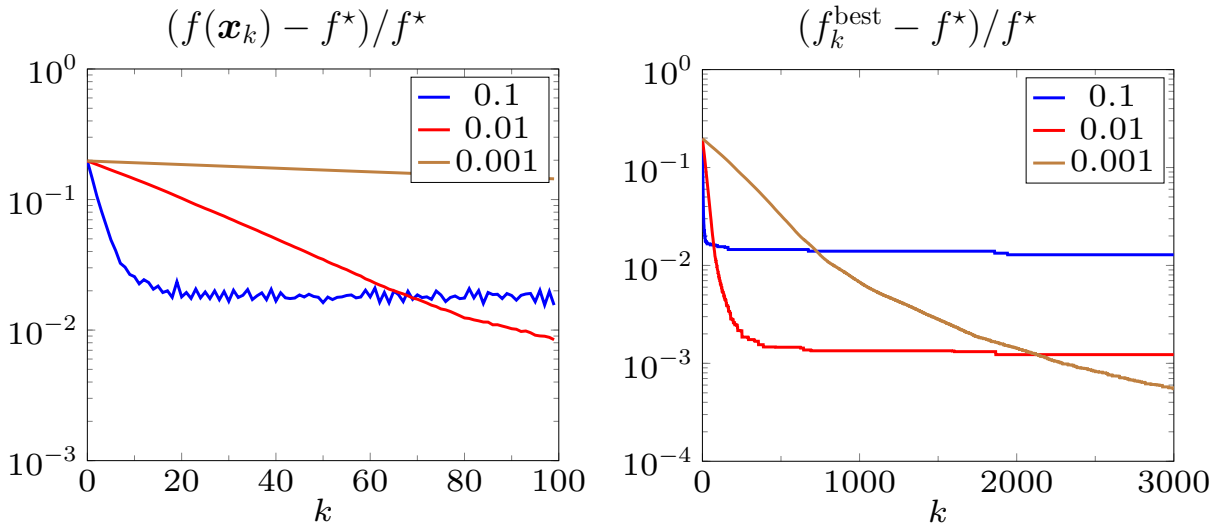
$$\begin{aligned} z_m &= \text{sign}(\boldsymbol{a}_m^{\mathrm{T}}\boldsymbol{x} - b_m) \quad \text{if } \boldsymbol{a}_m^{\mathrm{T}}\boldsymbol{x} - b_m \neq 0, \\ |z_m| &\leq 1 \qquad\qquad\qquad\quad \text{if } \boldsymbol{a}_m^{\mathrm{T}}\boldsymbol{x} - b_m = 0. \end{aligned}$$

Now consider the vector $\boldsymbol{u} = \boldsymbol{A}^{\mathrm{T}}\boldsymbol{z}$. Note that

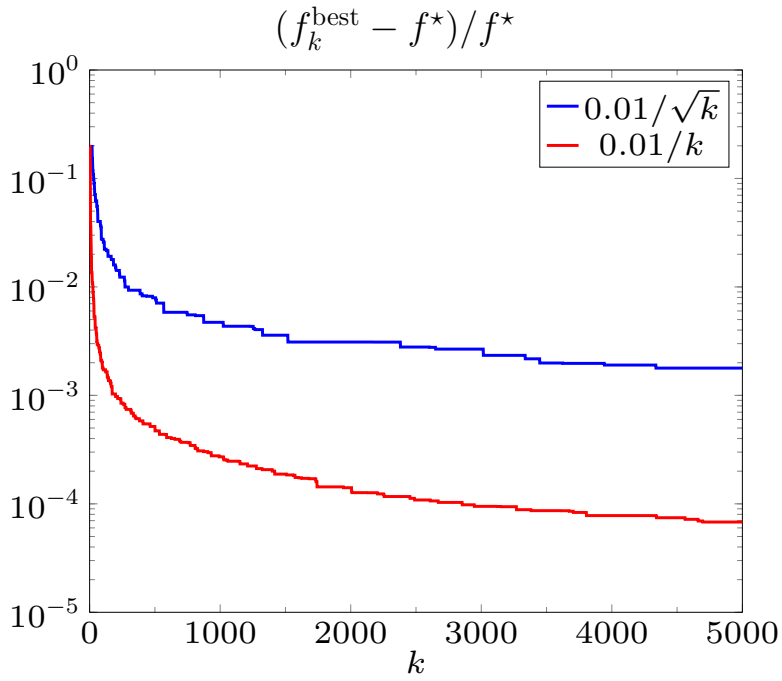
$$\begin{aligned} \boldsymbol{u}^{\mathrm{T}}(\boldsymbol{y} - \boldsymbol{x}) &= \boldsymbol{z}^{\mathrm{T}}\boldsymbol{A}(\boldsymbol{y} - \boldsymbol{x}) \\ &= \boldsymbol{z}^{\mathrm{T}}(\boldsymbol{A}\boldsymbol{y} - \boldsymbol{b} + \boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}) \\ &= \boldsymbol{z}^{\mathrm{T}}(\boldsymbol{A}\boldsymbol{y} - \boldsymbol{b}) - \boldsymbol{z}^{\mathrm{T}}(\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}) \\ &= \boldsymbol{z}^{\mathrm{T}}(\boldsymbol{A}\boldsymbol{y} - \boldsymbol{b}) - \|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\|_1 \\ &\leq \|\boldsymbol{A}\boldsymbol{y} - \boldsymbol{b}\|_1 - \|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\|_1. \end{aligned}$$

Rearranging this shows that $\boldsymbol{u}$ is a subgradient of $\|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\|_1$. Using this we can construct a subgradient at each step $\boldsymbol{x}_k$.

Below we illustrate the performance of this approach for a randomly generated example with $\boldsymbol{A} \in \mathbb{R}^{500\times 100}$ and $\boldsymbol{b} \in \mathbb{R}^{1000}$. For three different sizes of fixed step length, $s = 0.1, 0.01, 0.001$, we make quick progress at the beginning, but then saturate, just as the theory predicts:



Here is a run using two different decreasing step size strategies: $\alpha_k = .01/\sqrt{k}$ and $\alpha_k = .01/k$.



As you can see, even though the theoretical worst case bound makes a stepsize of $\sim 1/\sqrt{k}$ look better, in this particular case, a stepsize $\sim 1/k$ actually performs better.

Qualitatively, the takeaways for the subgradient method are:

1. It is a natural extension of the gradient descent formulation
2. In general, it does not converge for fixed stepsizes.
3. If the stepsizes decrease, you can guarantee convergence.
4. Theoretical convergence rates are slow.
5. Convergence rates in practice are also very slow, but depend a lot on the particular example.

References

[Nes04] Y. Nesterov. *Introductory Lectures on Convex Optimization*. Springer Science+Business Media, 2004.

Proximal algorithms

The subgradient algorithm is one generalization of gradient descent. It is simple, but the convergence is typically very slow (and it does not even converge in general for a fixed step size).

One way to deal with this is to add a smooth *regularization* term. Specifically, it is easy to show that if $\boldsymbol{x}^\star$ is a minimizer of $f(\boldsymbol{x})$, then it is also the minimizer of

$$\underset{\boldsymbol{x}\in\mathbb{R}^N}{\text{minimize}}\, f(\boldsymbol{x}) + \delta\|\boldsymbol{x}-\boldsymbol{x}^\star\|_2^2,$$

where $\delta > 0$. While the resulting optimization problem is still nonsmooth, it is now strongly convex, and we know that strongly convex functions are generally much easier to minimize. The "only" challenge is that it requires us to already know the solution $\boldsymbol{x}^\star$, which would seem to limit the practical applicability of this idea.

We can turn this into an actual algorithm by adopting an iterative approach. The **proximal algorithm** or **proximal point method** uses the following iteration:

$$\boldsymbol{x}_{k+1} = \underset{\boldsymbol{x}\in\mathbb{R}^N}{\arg\min}\left(f(\boldsymbol{x}) + \frac{1}{2\alpha_k}\|\boldsymbol{x}-\boldsymbol{x}_k\|_2^2\right). \tag{1}$$

As noted above, when f is convex, $f(\boldsymbol{x}) + \delta\|\boldsymbol{x}-\boldsymbol{z}\|_2^2$ is strictly convex for all $\delta > 0$ and $\boldsymbol{z}\in\mathbb{R}^N$, so the mapping from $\boldsymbol{x}_k$ to $\boldsymbol{x}_{k+1}$ is well-defined. We will sometimes use the "prox operator" to denote this mapping:

$$\text{prox}_{\alpha_k f}(\boldsymbol{z}) = \underset{\boldsymbol{x}\in\mathbb{R}^N}{\arg\min}\left(f(\boldsymbol{x}) + \frac{1}{2\alpha_k}\|\boldsymbol{x}-\boldsymbol{z}\|_2^2\right).$$

It can be shown (and in fact we give a proof later in these notes) that the iterations above do find a minimizer of convex f for an appropriate choice of "step sizes" α_k.

At this point, you would be forgiven for having doubts about what we are really doing here. We have taken an optimization problem and turned it into... a sequence of *many* optimization problems. However, these problems can sometimes be far easier to solve that the original problem. One way to think about the additional $\frac{1}{2\alpha_k}\|\boldsymbol{x}-\boldsymbol{x}_k\|$ term is as a *regularizer* that makes each subproblem computationally easier to solve, and whose influence naturally disappears as we approach the solution, even for a fixed "step size" $\alpha_k = \alpha$.

A very nice a detailed review of proximal algorithms can be found in [PB14].

Implicit gradient descent ("backward Euler")

The proximal point method can also be interpreted as a variation on gradient descent. To see this, let us return for a moment to the differential equations for the "gradient flow" of f:

$$\boldsymbol{x}'(t) = -\nabla f(\boldsymbol{x}(t)), \quad \boldsymbol{x}(0) = \boldsymbol{x}_0. \tag{2}$$

The equilibrium points for this system are the $\boldsymbol{x}$ such that $\nabla f(\boldsymbol{x}) = \boldsymbol{0}$, which are precisely the minimizers for $f(\boldsymbol{x})$.

As we first discussed in the context of momentum-based methods, we can interpret gradient descent as a first-order numerical method for tracing the path from $\boldsymbol{x}_0$ to a solution $\boldsymbol{x}^\star$. This comes from discretizing the derivative on the right using a forward finite difference:

$$\frac{\boldsymbol{x}(t+h) - \boldsymbol{x}(t)}{h} \approx -\nabla f(\boldsymbol{x}(t)) \quad \text{for small } h.$$

Thus the gradient descent iterations

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - h\nabla f(\boldsymbol{x}_k)$$

approximate the solution at equispaced times spaced h seconds apart — the step size in gradient descent can be interpreted as the time scale to which we are approximating the derivative. This is known as the *forward Euler method* for discretizing (2).

But now suppose we used a *backward difference* to approximate the derivative:

$$\frac{\boldsymbol{x}(t) - \boldsymbol{x}(t-h)}{h} \approx -\nabla f(\boldsymbol{x}(t)) \quad \text{for small } h.$$

Now the iterates must obey

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - h\nabla f(\boldsymbol{x}_{k+1}).$$

This is an equally valid technique for discretizing (2) known as the *backward Euler method*. However, computing the iterates is not as straightforward – we can't just compute the gradient at the current point, we have to find the next point by finding an $\boldsymbol{x}_{k+1}$ that obeys the equation above.

This is exactly what the proximal operator does. If f is differentiable, then

$$\begin{aligned} \boldsymbol{x}_{k+1} &= \arg\min_{\boldsymbol{x}\in\mathbb{R}^N} \left(f(\boldsymbol{x}) + \frac{1}{2\alpha_k}\|\boldsymbol{x} - \boldsymbol{x}_k\|_2^2 \right) \\ &\Updownarrow \\ \mathbf{0} &= \nabla f(\boldsymbol{x}_{k+1}) + \frac{1}{\alpha_k}(\boldsymbol{x}_{k+1} - \boldsymbol{x}_k). \end{aligned} \tag{3}$$

So the proximal point method can be interpreted as a backward Euler discretization for gradient flow.

Note that we assumed the differentiability of f above purely for illustration; we can compute the prox operator whether or not f has a gradient.

Example: Least squares

Suppose we want to solve the standard least-squares problem

$$\underset{\boldsymbol{x}\in\mathbb{R}^N}{\text{minimize}} \, \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_2^2.$$

When $\boldsymbol{A}$ has full column rank, we know that the solution is given by $\widehat{\boldsymbol{x}}_{\text{ls}} = (\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A})^{-1}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{y}$. However, we also know that when $\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}$ is not well-conditioned, this inverse can be unstable to compute, and iterative descent methods (gradient descent and conjugate gradients) can take many iterations to converge.

Consider the proximal point iteration (with fixed $\alpha_k = \alpha$) for solving this problem:

$$\boldsymbol{x}_{k+1} = \underset{\boldsymbol{x}\in\mathbb{R}^N}{\arg\min} \left(\frac{1}{2}\|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_2^2 + \frac{1}{2\alpha}\|\boldsymbol{x} - \boldsymbol{x}_k\|_2^2 \right).$$

Here we have the closed form solution

$$\begin{aligned} \boldsymbol{x}_{k+1} &= (\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A} + \delta\mathbf{I})^{-1}(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{y} + \delta\boldsymbol{x}_k), \quad \delta = \frac{1}{\alpha} \\ &= \boldsymbol{x}_k + (\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A} + \delta\mathbf{I})^{-1}\boldsymbol{A}^{\mathrm{T}}(\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}_k). \end{aligned}$$

Now each step is equivalent to solving a least-squares problem, but this problem can be made well-conditioned by choosing δ (i.e., α)

appropriately. The iterations above will converge to $\widehat{\boldsymbol{x}}_{\text{ls}}$ for any value of α; as we decrease α (increase δ), the number of iterations to get within a certain accuracy of $\widehat{\boldsymbol{x}}_{\text{ls}}$ increases, but the least-squares problems involved are all very well conditioned. For α very small, we are back at gradient descent (with step size α).

This is actually a well-known technique in numerical linear algebra called *iterative refinement*.

Proximal gradient algorithms

Recall the core update equation for the proximal point method:

$$\boldsymbol{x}_{k+1} = \text{prox}_{\alpha_k f}(\boldsymbol{x}_k) = \arg\min_{\boldsymbol{x} \in \mathbb{R}^N} \left(f(\boldsymbol{x}) + \frac{1}{2\alpha} \|\boldsymbol{x} - \boldsymbol{x}_k\|_2^2 \right).$$

Suppose that we did not wish to fully solve this problem at each iteration. If f is differentiable, we could approximate this update by replacing $f(\boldsymbol{x})$ with its linear approximation $f(\boldsymbol{x}_k) + \langle \boldsymbol{x} - \boldsymbol{x}_k, \nabla f(\boldsymbol{x}_k) \rangle$. This would yield the update

$$\begin{aligned}
\boldsymbol{x}_{k+1} &= \arg\min_{\boldsymbol{x} \in \mathbb{R}^N} \left(f(\boldsymbol{x}_k) + \langle \boldsymbol{x} - \boldsymbol{x}_k, \nabla f(\boldsymbol{x}_k) \rangle + \frac{1}{2\alpha_k} \|\boldsymbol{x} - \boldsymbol{x}_k\|_2^2 \right) \\
&= \arg\min_{\boldsymbol{x} \in \mathbb{R}^N} \left(\frac{\alpha_k}{2} \|\nabla f(\boldsymbol{x}_k)\|_2^2 + \langle \boldsymbol{x} - \boldsymbol{x}_k, \nabla f(\boldsymbol{x}_k) \rangle + \frac{1}{2\alpha_k} \|\boldsymbol{x} - \boldsymbol{x}_k\|_2^2 \right) \\
&= \arg\min_{\boldsymbol{x} \in \mathbb{R}^N} \left(\frac{1}{2\alpha_k} \|\boldsymbol{x} - \boldsymbol{x}_k + \alpha_k \nabla f(\boldsymbol{x}_k)\|_2^2 \right) \\
&= \boldsymbol{x}_k - \alpha_k \nabla f(\boldsymbol{x}_k).
\end{aligned}$$

Thus, taking a linear approximation of f, the proximal method simply reduces to standard gradient descent. (Note that the first equality

above comes from the fact that the presence/absence of $f(\boldsymbol{x}_k)$ and $\|\nabla f(\boldsymbol{x}_k)\|_2^2$ does not affect what the minimizer is, as $\boldsymbol{x}_k$ is fixed.)

Where this starts getting interesting is when we encounter optimization problems where the objective function can be broken into the sum two parts, i.e.,

$$f(\boldsymbol{x}) = g(\boldsymbol{x}) + h(\boldsymbol{x}),$$

where both g and h are convex, but g is smooth (differentiable) and h is a non-smooth function for which there is a fast proximal operator. Such optimization problems quite a bit more often than you might expect.

The **proximal gradient** algorithm is the result of applying the proximal point method to minimize the approximation of f where we take a linear approximation to the smooth component g. Using the same argument as above, this results in the update rule[1]

$$\begin{aligned}
\boldsymbol{x}_{k+1} &= \arg\min_{\boldsymbol{x}\in\mathbb{R}^N} \left(g(\boldsymbol{x}_k) + \langle \boldsymbol{x} - \boldsymbol{x}_k, \nabla g(\boldsymbol{x}_k)\rangle + h(\boldsymbol{x}) + \frac{1}{2\alpha_k}\|\boldsymbol{x} - \boldsymbol{x}_k\|_2^2 \right) \\
&= \arg\min_{\boldsymbol{x}\in\mathbb{R}^N} \left(h(\boldsymbol{x}) + \frac{1}{2\alpha_k}\|\boldsymbol{x} - \boldsymbol{x}_k + \alpha_k \nabla g(\boldsymbol{x}_k)\|_2^2 \right) \\
&= \text{prox}_{\alpha_k h}\left(\boldsymbol{x}_k - \alpha_k \nabla g(\boldsymbol{x}_k)\right).
\end{aligned}$$

This is also called *forward-backward splitting*, with the "forward" referring to the gradient step, and the "backward" to the proximal step. (The prox step is still making progress, just like the gradient step; the forward and backward refer to the interpretations of gradient descent and the proximal algorithm as forward and backward Euler discretizations, respectively.)

[1] Again, the second line comes from removing $g(\boldsymbol{x}_k)$ and adding a multiple of $\|\nabla g(\boldsymbol{x}_k)\|_2^2$ and then completing the square.

Example: The LASSO

Recall our friend the LASSO:

$$\underset{\boldsymbol{x}\in\mathbb{R}^N}{\text{minimize}} \ \frac{1}{2}\|\boldsymbol{y}-\boldsymbol{A}\boldsymbol{x}\|_2^2+\tau\|\boldsymbol{x}\|_1.$$

We take

$$g(\boldsymbol{x})=\frac{1}{2}\|\boldsymbol{y}-\boldsymbol{A}\boldsymbol{x}\|_2^2, \quad \text{so} \quad \nabla g(\boldsymbol{x})=\boldsymbol{A}^{\mathrm{T}}(\boldsymbol{A}\boldsymbol{x}-\boldsymbol{y}),$$

and

$$h(\boldsymbol{x})=\tau\|\boldsymbol{x}\|_1.$$

The prox operator for the ℓ_1 norm is:

$$\begin{aligned}\text{prox}_{\alpha h}(\boldsymbol{z}) &= \underset{\boldsymbol{x}\in\mathbb{R}^N}{\arg\min}\left(\tau\|\boldsymbol{x}\|_1+\frac{1}{2\alpha}\|\boldsymbol{x}-\boldsymbol{z}\|_2^2\right)\\ &= T_{\tau\alpha}(\boldsymbol{z}),\end{aligned}$$

where $T_{\tau\alpha}$ is the soft-thresholding operator

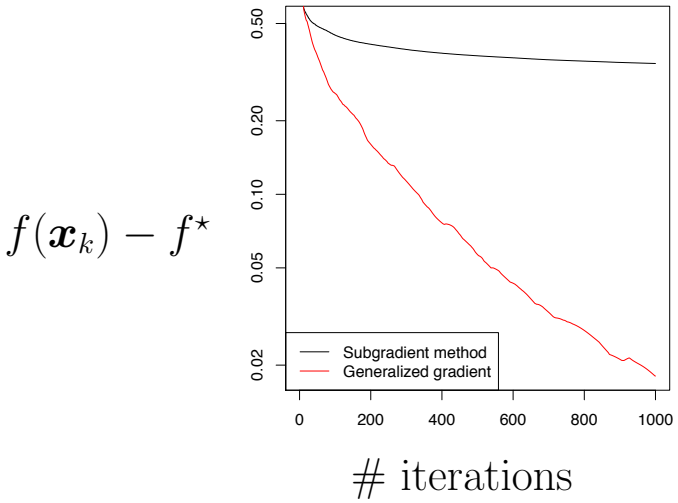
$$\left(T_{\tau\alpha}(\boldsymbol{z})\right)_i=\begin{cases}z_i-\tau\alpha, & z_i\geq\tau\alpha,\\ 0, & |z_i|\leq\tau\alpha,\\ z_i+\tau\alpha, & z_i\leq-\tau\alpha.\end{cases}$$

Hence, the gradient step requires an application of $\boldsymbol{A}$ and $\boldsymbol{A}^{\mathrm{T}}$, and the proximal step simply requires a soft-thresholding operation. The iteration looks like

$$\boldsymbol{x}_{k+1}=T_{\tau\alpha_k}\left(\boldsymbol{x}_k+\alpha_k\boldsymbol{A}^{\mathrm{T}}(\boldsymbol{y}-\boldsymbol{A}\boldsymbol{x}_k)\right).$$

This is also called the *iterative soft thresholding algorithm*, or ISTA.

Here is a comparison[2] of a typical run for ISTA versus the subgradient method. ISTA absolutely crushes the subgradient method.



[2]This is taken from the lecture notes of Geoff Gordon and Ryan Tibshirani; "generalized gradient" in the legend means ISTA.

Convergence of the proximal gradient method

The convergence analysis of the proximal gradient method is extremely similar to what we did for gradient descent. In fact, gradient descent is a special case of the proximal gradient method (when $h(\boldsymbol{x}) = 0$), and our analysis will recover the same result. We will assume that g is L-smooth, but we will make no assumptions on h aside from convexity. As before, we will use a fixed "step size", $\alpha_k = 1/L$ for all k. We will $\boldsymbol{x}^\star$ denote any minimizer of f.

The general structure of the argument is as follows:

1. Using the L-smoothness of g as well as the first-order characterization of convexity, we can establish that

$$f(\boldsymbol{x}_{k+1}) \leq f(\boldsymbol{z}) + \langle \boldsymbol{x}_k - \boldsymbol{z}, \boldsymbol{d}_k \rangle - \frac{1}{2L}\|\boldsymbol{d}_k\|_2^2 \tag{4}$$

   for all $\boldsymbol{z} \in \mathbb{R}^N$ where $\boldsymbol{d}_k := L(\boldsymbol{x}_k - \boldsymbol{x}_{k+1})$.

2. From (4) we can conclude, by setting $\boldsymbol{z} = \boldsymbol{x}_k$, that

$$f(\boldsymbol{x}_{k+1}) \leq f(\boldsymbol{x}_k) - \frac{1}{2L}\|\boldsymbol{d}_k\|_2^2 \leq f(\boldsymbol{x}_k),$$

   and thus $f(\boldsymbol{x}_k)$ is non-increasing at every step.

3. From (4) we can also conclude, by setting $\boldsymbol{z} = \boldsymbol{x}^\star$, that

$$f(\boldsymbol{x}_{k+1}) \leq f(\boldsymbol{x}^\star) + \langle \boldsymbol{x}_k - \boldsymbol{x}^\star, \boldsymbol{d}_k \rangle - \frac{1}{2L}\|\boldsymbol{d}_k\|_2^2.$$

   By exactly the same argument as we have seen in the analysis of both gradient descent and Nesterov's method, we can show that this bound is equivalent to

$$f(\boldsymbol{x}_{k+1}) - f(\boldsymbol{x}^\star) \leq \frac{L}{2}\left(\|\boldsymbol{x}_k - \boldsymbol{x}^\star\|_2^2 - \|\boldsymbol{x}_{k+1} - \boldsymbol{x}^\star\|_2^2\right).$$

4. This yields a telescopic sum, and hence by an identical argument to that used in analyzing gradient descent, we arrive at the bound

$$f(\boldsymbol{x}_k) - f(\boldsymbol{x}^\star) \leq \frac{L}{2k}\|\boldsymbol{x}_0 - \boldsymbol{x}^\star\|_2^2.$$

Thus, the proximal gradient algorithm exhibits the same convergence rate as gradient descent: $O(1/k)$. This is remarkable when considering that it holds for *any* h. This result is in fact a kind of "master result" for the convergence rate of many different algorithms:

- gradient descent (take $h(\boldsymbol{x}) = 0$),
- the proximal point method (take $g(\boldsymbol{x}) = 0$),
- the proximal gradient method.

The work above gives a unified analysis for all three of these, showing that they all exhibit $O(1/k)$ convergence.

Note that the only novelty in the analysis above compared to that of gradient descent is the derivation of (4). To establish this inequality, we first note that

$$\begin{aligned} f(\boldsymbol{x}_{k+1}) &= g(\boldsymbol{x}_{k+1}) + h(\boldsymbol{x}_{k+1}) \\ &\leq g(\boldsymbol{x}_k) - \frac{1}{L}\langle \boldsymbol{d}_k, \nabla g(\boldsymbol{x}_k)\rangle + \frac{1}{2L}\|\boldsymbol{d}_k\|_2^2 + h(\boldsymbol{x}_{k+1}), \end{aligned} \tag{5}$$

where the inequality follows directly from the definition of L-smoothness. We now use two facts to get an upper bound on this expression. First, note that from the first-order characterization of convexity,

$$g(\boldsymbol{z}) \geq g(\boldsymbol{x}_k) + \langle \boldsymbol{z} - \boldsymbol{x}_k, \nabla g(\boldsymbol{x}_k)\rangle. \tag{6}$$

Second, since

$$\begin{aligned}\boldsymbol{x}_{k+1} &= \text{prox}_{h/M}\left(\boldsymbol{x}_k - \frac{1}{L}\nabla g(\boldsymbol{x}_k)\right) \\ &= \arg\min_{\boldsymbol{x}}\left(h(\boldsymbol{x}) + \frac{L}{2}\left\|\boldsymbol{x} - \boldsymbol{x}_k + \frac{1}{L}\nabla g(\boldsymbol{x}_k)\right\|_2^2\right),\end{aligned}$$

we know

$$\mathbf{0} \in \partial h(\boldsymbol{x}_{k+1}) - \boldsymbol{d}_k + \nabla g(\boldsymbol{x}_k) \quad \Rightarrow \quad \boldsymbol{d}_k - \nabla g(\boldsymbol{x}_k) \in \partial h(\boldsymbol{x}_{k+1}).$$

Thus

$$h(\boldsymbol{z}) \geq h(\boldsymbol{x}_{k+1}) + \langle \boldsymbol{z} - \boldsymbol{x}_{k+1}, \boldsymbol{d}_k - \nabla g(\boldsymbol{x}_k)\rangle. \tag{7}$$

We combine (6) and (7) back into (5) to obtain

$$\begin{aligned}f(\boldsymbol{x}_{k+1}) &\leq g(\boldsymbol{z}) + \langle \boldsymbol{x}_k - \boldsymbol{z}, \nabla g(\boldsymbol{x}_k)\rangle - \frac{1}{L}\langle \boldsymbol{d}_k, \nabla g(\boldsymbol{x}_k)\rangle + \frac{1}{2L}\|\boldsymbol{d}_k\|_2^2 \\ &\quad + h(\boldsymbol{z}) - \left\langle \boldsymbol{z} - \boldsymbol{x}_k + \frac{1}{L}\boldsymbol{d}_k, \boldsymbol{d}_k - \nabla g(\boldsymbol{x}_k)\right\rangle \\ &= f(\boldsymbol{z}) + \langle \boldsymbol{x}_k - \boldsymbol{z}, \boldsymbol{d}_k\rangle + \frac{1}{L}\left(\frac{L}{2L} - 1\right)\|\boldsymbol{d}_k\|_2^2 \\ &\leq f(\boldsymbol{z}) + \langle \boldsymbol{x}_k - \boldsymbol{z}, \boldsymbol{d}_k\rangle - \frac{1}{2L}\|\boldsymbol{d}_k\|_2^2,\end{aligned}$$

which establishes (4).

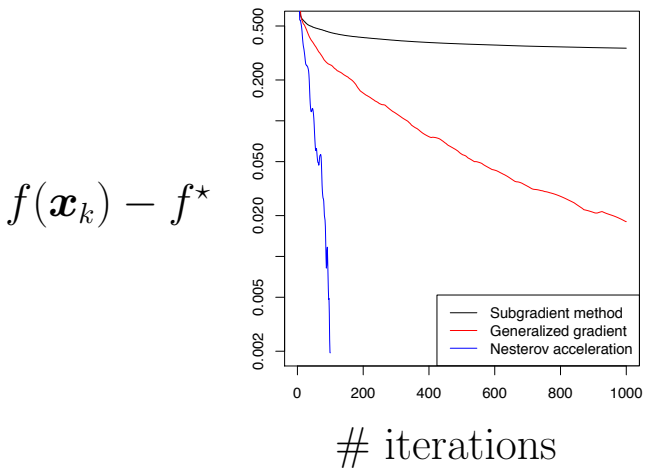
Accelerated proximal gradient

We can accelerate the proximal gradient method in exactly the same way we accelerated gradient descent – in fact, the Nesterov's method for gradient descent is simply a special case as that for the proximal gradient algorithm. The accelerated iteration is

$$\begin{aligned} \boldsymbol{p}_k &= \frac{k-1}{k+2}\left(\boldsymbol{x}_k - \boldsymbol{x}_{k-1}\right) \\ \boldsymbol{x}_{k+1} &= \operatorname{prox}_{\alpha_k h}\left(\boldsymbol{x}_k + \boldsymbol{p}_k - \alpha_k \nabla g(\boldsymbol{x}_k + \boldsymbol{p}_k)\right). \end{aligned}$$

Again, the computations here are in general no more involved than for the non-accelerated version, but the number of iterations can be significantly lower. We will not prove it here (see [BT09] for an analysis), but adding in the momentum term results in convergence rate of $O(1/k^2)$ using a similar argument as before.

The numerical performance can also be dramatically better. Here are typical runs[3] for the LASSO, which compares the standard proximal gradient method (ISTA) to its accelerated version (FISTA):



[3] Again, this example comes from Gordon and Tibshirani; as before "generalized gradient" means ISTA, and "Nesterov acceleration" means FISTA.

References

[BT09] A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM J. Imaging Sci.*, 2(1):183–202, 2009.

[PB14] N. Parikh and S. Boyd. Proximal algorithms. *Foundations and Trends in Optimization*, 1(3):123–231, 2014.

Reading materials

Lemma

∇f is Lipschitz with constant L if and only if

$$f(y) \leq f(x) + \nabla f(x)^T(y-x) + \frac{L}{2}\|y-x\|_2^2, \text{ for all } x, y.$$

Proof. Suppose ∇f is Lipschitz with constant L. Consider $g(t) = f(x + t(y-x))$. Then $g'(t) = \nabla f(x + t(y-x))^T(y-x)$.
Then

$$\begin{aligned}
&f(y) - f(x) - \nabla f(x)^T(y-x) \\
&= g(1) - g(0) - \nabla f(x)^T(y-x) \\
&= \int_0^1 \nabla f(x + t(y-x))^T(y-x) - \nabla f(x)^T(y-x)dt \\
&= \int_0^1 (\nabla f(x + t(y-x)) - \nabla f(x))^T(y-x)dt \\
&\leq \int_0^1 \|\nabla f(x + t(y-x)) - \nabla f(x)\|_2\|y-x\|_2 dt \\
&\leq \int_0^1 Lt||y-x||_2^2 dt \\
&= \frac{L}{2}||y-x||_2^2.
\end{aligned}$$

Conversely, suppose $f(y) \leq f(x) + \nabla f(x)^T(y-x) + \frac{L}{2}\|y-x\|_2^2, \text{ for all } x, y$. Consider the function $\phi_x(z) := f(z) - \nabla f(x)^T z$.
ϕ_x is convex and $\nabla \phi_x(z) = \nabla f(z) - \nabla f(x)$.
Since, $f(z) \leq f(y) + \nabla f(y)^T(z-y) + \frac{L}{2}||z-y||^2$, we have

$$f(z) - \nabla f(x)^T z \leq f(y) - \nabla f(x)^T y + (\nabla f(y) - \nabla f(x))^T(z-y) + \frac{L}{2}||z-y||^2$$

That is

$$\phi_x(z) \leq \phi_x(y) + \nabla \phi_x(y)^T(z-y) + \frac{L}{2}||z-y||^2$$

We minimized both sides over z. The left hand side is minimized at $z = x$.
The right hand side is minimized at $z = -\frac{1}{L}\nabla \phi_x(y) + y$. Hence,

$$\begin{aligned}
f(x) - \nabla f(x)^T x = \phi_x(x) &\leq \phi_x(y) + \nabla \phi_x(y)^T(-\frac{1}{L}\nabla \phi_x(y)) + \frac{L}{2}\|\frac{1}{L}\nabla \phi_x(y)\|^2 \\
&= f(y) - \nabla f(x)^T y - \frac{1}{2L}\|\nabla f(y) - \nabla f(x)\|^2
\end{aligned}$$

So

$$f(y) - f(x) - \nabla f(x)^T(y-x) \geq \frac{1}{2L}\|\nabla f(y) - \nabla f(x)\|^2$$

Interchange the role of x, y, we get

$$f(x) - f(y) - \nabla f(y)^T(x-y) \geq \frac{1}{2L}\|\nabla f(y) - \nabla f(x)\|^2$$

Adding the two inequalities, we get

$$(\nabla f(x) - \nabla f(y))^T(x-y) \geq \frac{1}{L}\|\nabla f(x) - \nabla f(y)\|^2$$

Hence, we have

$$\begin{aligned}
\|\nabla f(y) - \nabla f(x)\|^2 &\leq L(\nabla f(y) - \nabla f(x))^T(y-x) \\
&\leq L\|\nabla f(y) - \nabla f(x)\|\|y-x\|
\end{aligned}$$

□

Reading materials

Lemma

Suppose f is μ-strongly convex. Then

$$2\mu(f(x) - f^*) \leq \|\nabla f(x)\|_2^2.$$

Proof. Since f is μ-strongly convex,

$$f(y) \geq f(x) + \nabla f(x)^T(y - x) + \frac{\mu}{2}\|y - x\|_2^2.$$

We minimize both sides with respect to y. Taking gradient on the right hand side, we note that the minimizer is $x - \frac{1}{\mu}\nabla f(x)$.
Therefore,

$$f^* = \inf_y f(y) \geq \inf_y \left\{ f(x) + \nabla f(x)^T(y - x) + \frac{\mu}{2}\|y - x\|_2^2 \right\} = f(x) - \frac{1}{2\mu}\|\nabla f(x)\|_2^2.$$

Hence,

$$f^* \geq f(x) - \frac{1}{2\mu}\|\nabla f(x)\|_2^2$$

□