Advanced Optimization Lecture Notes

Daniel Romero
Department of Information and Communication Technology
University of Agder

April 28, 2023

Contents

1 Optimization Problems **9**
- 1.1 Introduction . . . 9
  - 1.1.1 Optimization Problems . . . 9
  - 1.1.2 Solution of an Optimization Problem . . . 10
  - 1.1.3 Existence of optimal solutions . . . 11
- 1.2 Properties of Optimization Problems . . . 12
  - 1.2.1 Convexity . . . 12
  - 1.2.2 Strong Convexity . . . 16
  - 1.2.3 Lipschitz smoothness . . . 19
- 1.3 Equivalent Problems . . . 23
- 1.4 Solving Optimization Problems . . . 26

2 Optimality Conditions and Duality **27**
- 2.1 Introduction . . . 27
- 2.2 Unconstrained Optimization Problems . . . 27
  - 2.2.1 Necessary conditions when $D = 1$ . . . 27
  - 2.2.2 Sufficient conditions when $D = 1$ . . . 28
  - 2.2.3 Necessary conditions when $D \geq 1$ . . . 29
  - 2.2.4 Sufficient conditions when $D \geq 1$ . . . 29
  - 2.2.5 Closed-form Solutions . . . 29
- 2.3 Constrained Optimization Problems . . . 31
  - 2.3.1 First- and Second-order Conditions . . . 31
  - 2.3.2 Zeroth-order Conditions and Duality . . . 45

3 Introduction to Iterative Methods **59**
- 3.1 Introduction . . . 59
- 3.2 Descent Methods . . . 61
  - 3.2.1 Convergence Analysis . . . 63
- 3.3 Optimal First-order Methods . . . 66

A Mathematical Background **69**
- A.1 Linear and Affine Spaces . . . 69
- A.2 Derivatives . . . 70
- A.3 Subgradient and subdifferential . . . 72
- A.4 Mean value theorem . . . 74

Foreword

This document contains the lecture notes of the first module of the PhD course IKT719, *Advanced Optimization*, taught at the University of Agder, Norway.

This material is intended for self-study, reference, and as a guide for the lectures. The student is requested to carefully read the corresponding sections before each lecture, which proceed in a fast-paced fashion.

A number of quiz questions are included to assist in understanding and digesting the presented concepts as well as to increase their long-term persistence in the student's memory. An important part of the lectures is devoted to solve these questions. Students are encouraged to try to solve these quizzes before the lectures.

Information is organized hierarchically and presented in a schematic fashion, a format that seems well attuned to how the human brain processes information and facilitates extracting the big picture and navigating through the material.

Notation

- *iff* and "$\Leftrightarrow$" stand for *if and only if*. Similarly, "$\Rightarrow$" and "$\Leftarrow$" respectively stand for "only if" (or "then") and "if".
- With z a real number, $\lceil z \rceil$ denotes the smallest integer i satisfying $i \geq z$.
- Column vectors are denoted by boldface lowercase. Matrices by boldface uppercase letters.
- Unless otherwise stated, $||\boldsymbol{x}||$ with $\boldsymbol{x} \triangleq [x_1, \ldots, x_D]^\top \in \mathbb{R}^D$ stands for the standard Euclidean norm $||\boldsymbol{x}|| = \sqrt{x_1^2 + \ldots + x_D^2}$.
- $\text{sval}_{\max}(\boldsymbol{A})$ denotes the maximum singular value of matrix $\boldsymbol{A}$.
- $\nabla f(\boldsymbol{x})$ denotes the gradient of f evaluated at point $\boldsymbol{x}$.
- $\nabla^2 f(\boldsymbol{x})$ denotes the Hessian matrix of f evaluated at point $\boldsymbol{x}$.
- For symmetric matrices $\boldsymbol{A}$ and $\boldsymbol{B}$, the notation $\boldsymbol{A} \succeq \boldsymbol{B}$ (resp. $\boldsymbol{A} \succ \boldsymbol{B}$) means that $\boldsymbol{A} - \boldsymbol{B}$ is positive semi-definite (resp. positive definite).
- With $\phi(z) > 0$, notation "$f(z) \in \mathcal{O}(\phi(z))$ as $z \to z_0$" means that $|f(z)/\phi(z)|$ becomes bounded as $z \to z_0$. Formally, "$f(z) \in \mathcal{O}(\phi(z))$ as $z \to z_0$" if $\exists \delta, B$ such that $|f(z)| < B|\phi(z)|$ for all z satisfying $0 < |z - z_0| < \delta$.

Chapter 1

Optimization Problems

This chapter reviews optimization basics. Many of the presented concepts and terminology are used in the daily work of researchers in various engineering disciplines.

1.1 Introduction

1.1.1 Optimization Problems

[General notation]An *optimization problem* (or *program*) is denoted as

$$\underset{\boldsymbol{x}\in\mathcal{X}}{\text{minimize}}\, f(\boldsymbol{x}) \tag{1.1}$$

where

- $f : \mathcal{D} \to \mathbb{R}$ is the *objective* (or *cost*) function, where $\mathcal{D}$ is the *domain* of f. The value $f(\boldsymbol{x})$ is the cost of the decision $\boldsymbol{x}$.
- $\mathcal{X} \subset \mathcal{D}$ is the *feasible set*, or set of all possible decisions.
- $\boldsymbol{x}$ is the *decision variable*.

[Maximization problems] The *maximization* of any function $\tilde{f} : \mathcal{D} \to \mathbb{R}$ over $\mathcal{X}$ can be expressed as in (1.1) by setting e.g. $f(\boldsymbol{x}) = -\tilde{f}(\boldsymbol{x})$, $\forall \boldsymbol{x} \in \mathcal{X}$. Therefore, one can focus on minimization problems without loss of generality (w.l.o.g.).
[Examples]

- $\boldsymbol{x} = [x_1, \ldots, x_D]^\top \in \mathbb{R}_+^D$ with x_d the amount of capital invested in the d-th stock, $\mathcal{X} = \{x : \sum_{d=1}^{D} x_d \leq M\}$ with M the available capital, $f(\boldsymbol{x})$ the expected return of the investment $\boldsymbol{x}$.
- $\boldsymbol{x} = [x_1, \ldots, x_D]^\top \in \mathbb{R}_+^D$ with x_d the resistance of the d-th resistor in an electronic circuit, $\mathcal{X} \subset \mathbb{R}_+^D$ accounts for the set of available resistors that meet the circuit specifications, $f(\boldsymbol{x})$ quantifies e.g. the power consumption, out-of-band radiation, or signal-to-noise ratio.
- $\boldsymbol{x} = [x_1, \ldots, x_D]^\top \in \mathbb{R}_+^D$ with x_d the daily consumption of the d-th food ingredient, $\mathcal{X} \subset \mathbb{R}_+^D$ contains the vectors $\boldsymbol{x}$ that meet the guidelines of the World Health Organization, and $f(\boldsymbol{x})$ the price.
- Many many more...

Quiz 1.1 Provide three more use cases, possibly in your research area, where an optimization problem arises. For each example, specify (at least informally) the real-world meaning of the decision variables, objective function, and feasible set.

[Focus] In this course, the focus is on problems of the form (1.1) where $\mathcal{X}$ is a subset of the Euclidean space $\mathbb{R}^D$.
[Constrained vs. unconstrained] We typically distinguish two classes of problems:

- [unconstrained optimization] A problem is said to be *unconstrained* if $\mathcal{X} = \mathbb{R}^D$.
- [constrained optimization] A problem is said to be *constrained* if $\mathcal{X} \subsetneq \mathbb{R}^D$. In this case $\mathcal{X}$ is typically expressed as a system of equality and inequality constraints as

$$\mathcal{X} = \{\boldsymbol{x} \in \mathbb{R}^D : g_i(\boldsymbol{x}) \leq 0,\ i = 1, \ldots, M;\quad h_i(\boldsymbol{x}) = 0,\ i = 1, \ldots, N\} \tag{1.2}$$

  and the problem is written as

$$\underset{\boldsymbol{x} \in \mathbb{R}^D}{\text{minimize}} \quad f(\boldsymbol{x}) \tag{1.3a}$$
$$\text{subject to} \quad g_i(\boldsymbol{x}) \leq 0,\ i = 1, \ldots, M, \tag{1.3b}$$
$$h_i(\boldsymbol{x}) = 0,\ i = 1, \ldots, N. \tag{1.3c}$$

Quiz 1.2 Can any problem of the form (1.3) be expressed as a problem with only equality constraints ($M = 0$)?

Quiz 1.3 Can any problem of the form (1.3) be expressed as a problem with only inequality constraints ($N = 0$)?

Quiz 1.4 In view of these answers, why should we consider problems with both kinds of constraints rather than focusing just on problems with constraints of only one kind?

1.1.2 Solution of an Optimization Problem

[Minima]

- [Local minimum] $\boldsymbol{x}^* \in \mathcal{X}$ is a *local minimum* (or *local minimizer*) if $\exists \epsilon > 0$ such that $f(\boldsymbol{x}) \geq f(\boldsymbol{x}^*)$ $\forall \boldsymbol{x} \in \mathcal{X}$ with $||\boldsymbol{x} - \boldsymbol{x}^*|| < \epsilon$.
- [Strict local minimum] $\boldsymbol{x}^* \in \mathcal{X}$ is a *strict local minimum* (minimizer) if $\exists \epsilon > 0$ such that $f(\boldsymbol{x}) > f(\boldsymbol{x}^*)$ $\forall \boldsymbol{x} \in \mathcal{X}$ with $||\boldsymbol{x} - \boldsymbol{x}^*|| < \epsilon$.
- [Global minimum] $\boldsymbol{x}^* \in \mathcal{X}$ is a *global minimum* (minimizer) if $f(\boldsymbol{x}) \geq f(\boldsymbol{x}^*)$ $\forall \boldsymbol{x} \in \mathcal{X}$. The set of all global minima of a problem is denoted as

$$\arg\min_{\boldsymbol{x} \in \mathcal{X}} f(\boldsymbol{x}).$$

- [Strict global minimum] $\boldsymbol{x}^* \in \mathcal{X}$ is a *strict global minimum* if $f(\boldsymbol{x}) > f(\boldsymbol{x}^*)$ $\forall \boldsymbol{x} \in \mathcal{X}$, $\boldsymbol{x} \neq \boldsymbol{x}^*$.

[Solution] Solving an optimization problem means finding

- [opt. point] $\boldsymbol{x}^* \in \mathcal{X}$ that is a global minimizer (although sometimes one may just be interested in finding a local minimizer), in which case $\boldsymbol{x}^*$ is referred to as an *optimal point*; and/or
- [opt. value] the value $f^* \triangleq \inf_{\boldsymbol{x} \in \mathcal{X}} f(\boldsymbol{x})$, which is referred to as *optimal value*. We say that the optimum is *attained* when there exists a minimizer $\boldsymbol{x}^* \in \mathcal{X}$, in which case $f^* = f(\boldsymbol{x}^*)$.

Quiz 1.5 Can you provide an example of f and $\mathcal{X}$ for which $f^* \in \mathbb{R}$ but the optimum is not attained?

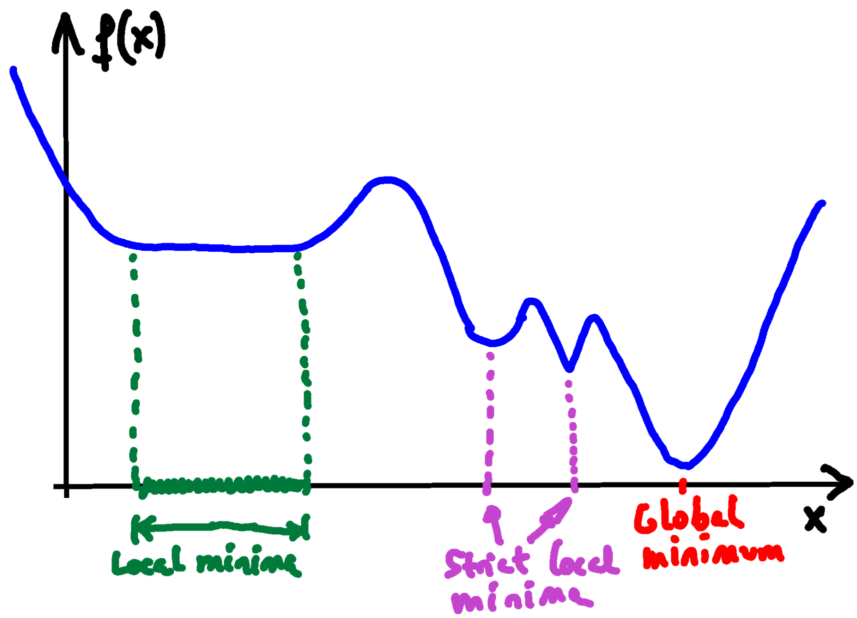


Figure 1.1: Different kinds of minima when $D = 1$.

1.1.3 Existence of optimal solutions

[Weierstrass theorem]The following result provides sufficient conditions for the existence of a global optimum.

Theorem 1.1 (Extreme value theorem, or Weierstrass theorem) *Suppose that* $f : \mathcal{X} \to \mathbb{R}$ *is continuous and (at least) one of the following conditions holds:*

- $\mathcal{X}$ *is non-empty and compact (i.e. closed and bounded).*
- $\mathcal{X}$ *is non-empty and closed, and* $f(\boldsymbol{x}) \to \infty$ *whenever* $||\boldsymbol{x}|| \to \infty$.
- *There exists* γ *such that the sublevel set* $\{\boldsymbol{x} \in \mathcal{X} : f(\boldsymbol{x}) \leq \gamma\}$ *is non-empty and compact.*

Then, there exists $\boldsymbol{x} \in \mathcal{X}$ *such that* $f(\boldsymbol{x}) = \inf_{\boldsymbol{z} \in \mathcal{X}} f(\boldsymbol{z})$.

Proof: See proof of [bertsekas1999, Prop. A8].

□

[Counterexamples]

Example 1.1 An example where the conclusion of this theorem does not hold (we say that the "minimum is not attained") is the following problem

$$\underset{x \in \mathbb{R}}{\text{minimize}} \quad x.$$

In this case, the objective function is not bounded below. But there are other examples where the conclusion does not hold and the function is bounded below, for example

$$\begin{aligned} \underset{x \in \mathbb{R}}{\text{minimize}} \quad & 1/x \\ \text{subject to} \quad & x \geq 1. \end{aligned}$$

Quiz 1.6 Why are the hypotheses of Theorem 1.1 not satisfied for the examples above?

[Relevance]It is recommendable to use Theorem 1.1 before attempting to solve an optimization problem. If none of its hypotheses hold, then it is possible that any (reasonable) iterative optimization algorithm will yield a divergent sequence of iterates; see Ch. 3.

1.2 Properties of Optimization Problems

[Motivation]No existing method can solve all problems of the form (1.1). Indeed, it is widely accepted that successfully implementing such a method would be physically impossible. For this reason, one needs to focus on special families of problems of the form (1.1); for example, one may identify that a problem is convex and use a solver for convex problems.
[convexity]Convex problems actually constitute the most important family of problems in this course. Their importance lies in the fact that a global optimum can be found in *polynomial time*.
[bounded curvature]As we will see, when the objective has a bounded *curvature*, many important algorithms converge faster and enjoy stronger convergence guarantees. Functions of this class are said to be

- [Strong convexity] strongly convex if there is a constant that lower bounds the curvature (Proposition 1.9).
- [Lipschitz smoothness] Lipschitz smooth if there is a constant that upper bounds the curvature (Proposition 1.13).

1.2.1 Convexity

[Terminology] Note that the word "convex" is used to refer to sets, functions, and problems. In each case, the definition of this word is different. One could as well have chosen different terms for each of these three cases.
[Domain] For simplicity, in the rest of the chapter, $\mathcal{D}$ is assumed to be open. When there is no loss of generality, it will also be assumed that $\mathcal{D} = \mathcal{X}$.

1.2.1.1 Convex sets

- [Def]

Definition 1.1 *A set $\mathcal{X}$ is said to be* convex *if*

$$\boldsymbol{x}, \boldsymbol{z} \in \mathcal{X} \;\Rightarrow\; \theta\boldsymbol{x} + (1-\theta)\boldsymbol{z} \in \mathcal{X} \;\forall \theta \in [0, 1].$$

- [Operations preserving convexity]Please review [boyd, Ch. 2].

1.2.1.2 Convex Functions

- [Def]

Definition 1.2 *Let $\mathcal{X}$ be a convex set. Then $f : \mathcal{X} \to \mathbb{R}$ is said to be*

– convex *if*

$$f(\theta \boldsymbol{x} + (1-\theta)\boldsymbol{z}) \leq \theta f(\boldsymbol{x}) + (1-\theta) f(\boldsymbol{z}) \quad \forall \boldsymbol{x}, \boldsymbol{z} \in \mathcal{X}, \theta \in [0,1].$$

– strictly convex[a] *if*

$$f(\theta \boldsymbol{x} + (1-\theta)\boldsymbol{z}) < \theta f(\boldsymbol{x}) + (1-\theta) f(\boldsymbol{z}) \quad \forall \boldsymbol{x}, \boldsymbol{z} \in \mathcal{X}, \theta \in [0,1].$$

[a]Not to confuse with *strongly* convex, which is defined later.

- [Operations preserving convexity]Please review [boyd, Ch. 3].
- [Properties]
  - [Linear lower bound] Convexity is equivalent to having a linear lower bound at every point. The case of differentiable functions is summarized by the following result:

Proposition 1.1 *Let $\mathcal{X} \subset \mathbb{R}^D$ be convex and let $f : \mathcal{X} \to \mathbb{R}$ be differentiable over $\mathcal{X}$.*

* *f is convex over $\mathcal{X}$ if and only if*

$$f(\boldsymbol{z}) \geq f(\boldsymbol{x}) + (\boldsymbol{z} - \boldsymbol{x})^\top \nabla f(\boldsymbol{x}), \ \forall \boldsymbol{z}, \boldsymbol{x} \in \mathcal{X}. \tag{1.4}$$

* *If (1.4) holds with strict inequality for all $\boldsymbol{z}, \boldsymbol{x} \in \mathcal{X}$ with $\boldsymbol{z} \neq \boldsymbol{x}$, then f is strictly convex over $\mathcal{X}$.*

Proof: See proof of [bertsekas1999, Prop. B.3].

□

  - [Local minima are global]The following proposition sheds light on why minimizing convex functions is so "easy", as compared to non-convex functions.

Proposition 1.2 *If $\boldsymbol{x}^*$ is a local minimum of a convex function f, then $\boldsymbol{x}^*$ is also a global minimum of f.*

Proof: We now present the proof for differentiable f. The proof for non-differentiable convex functions follows by replacing gradients with subgradients, which are guaranteed to exist at every point if the function under consideration is convex; see Appendix A.3.

If f is differentiable, then the bound (1.4) holds for all $\boldsymbol{x} \in \mathcal{X}$, in particular for $\boldsymbol{x} = \boldsymbol{x}^*$. In this case, one has $\nabla f(\boldsymbol{x}) = \nabla f(\boldsymbol{x}^*) = \mathbf{0}$ (cf. Proposition 2.4) and (1.4) becomes $f(\boldsymbol{z}) \geq f(\boldsymbol{x}^*)$, $\forall \boldsymbol{z} \in \mathcal{X}$.

□

– [Monotonic gradient] Convexity is also equivalent to having a monotonic gradient:

Proposition 1.3 *Let $\mathcal{D} \subset \mathbb{R}^D$ be convex. A continuously differentiable function $f : \mathcal{D} \to \mathbb{R}$ is convex if and only if it has a* monotonic gradient*, that is*

$$(\nabla f(\boldsymbol{z}) - \nabla f(\boldsymbol{x}))^\top (\boldsymbol{z} - \boldsymbol{x}) \geq 0 \tag{1.5}$$

for all $\boldsymbol{x}$ and $\boldsymbol{z}$.

Proof: "⇒": From convexity,

$$f(\boldsymbol{z}) \geq f(\boldsymbol{x}) + \nabla f^\top(\boldsymbol{x})(\boldsymbol{z} - \boldsymbol{x})$$
$$f(\boldsymbol{x}) \geq f(\boldsymbol{z}) + \nabla f^\top(\boldsymbol{z})(\boldsymbol{x} - \boldsymbol{z}).$$

Adding these expressions concludes the proof.

"$\Leftarrow$": Define $g(t) \triangleq f(\boldsymbol{x} + t(\boldsymbol{z} - \boldsymbol{x}))$ for $t \geq 0$ and note that $g'(t) = \nabla f^\top(\boldsymbol{x} + t(\boldsymbol{z} - \boldsymbol{x}))(\boldsymbol{z} - \boldsymbol{x})$. From gradient monotonicity (cf. (1.5)),

$$\begin{aligned} g'(t) - g'(0) &= \left[\nabla f^\top(\boldsymbol{x} + t(\boldsymbol{z} - \boldsymbol{x})) - \nabla f^\top(\boldsymbol{x})\right](\boldsymbol{z} - \boldsymbol{x}) \\ &= \frac{1}{t}\left[\nabla f^\top(\boldsymbol{y}) - \nabla f^\top(\boldsymbol{x})\right](\boldsymbol{y} - \boldsymbol{x}) \geq 0. \end{aligned} \tag{1.6}$$

where $\boldsymbol{y} \triangleq \boldsymbol{x} + t(\boldsymbol{z} - \boldsymbol{x})$. Then,

$$\begin{aligned} f(\boldsymbol{z}) &= g(1) = g(0) + \int_0^1 g'(t)dt \\ &\geq g(0) + \int_0^1 g'(0)dt = g(0) + g'(0) \\ &= f(\boldsymbol{x}) + \nabla f^\top(\boldsymbol{x})(\boldsymbol{z} - \boldsymbol{x}), \end{aligned}$$

where the inequality follows from (1.6). Convexity then follows from Proposition 1.1.

□

Mnemonic: When $D = 1$, (1.5) boils down to

$$(f'(z) - f'(x))(z - x) \geq 0. \tag{1.7}$$

If f is convex, then it follows that f' is non-decreasing; cf. Proposition 1.4. But if it is non-decreasing it clearly satisfies (1.7) for $z = x$, $z > x$, and $z < x$.

– [Curvature] For twice continuously differentiable functions, convexity is tightly related to second-order derivatives:

Proposition 1.4 *Let $\mathcal{X} \subset \mathbb{R}^D$ be a convex open set and let $f : \mathcal{X} \to \mathbb{R}$ be twice continuously differentiable over $\mathcal{X}$.*

1. *If $\nabla^2 f(\boldsymbol{x})$ is* positive semidefinite *for all $\boldsymbol{x} \in \mathcal{X}$ iff f is* convex *over $\mathcal{X}$.*
2. *If $\nabla^2 f(\boldsymbol{x})$ is* positive definite *for all $\boldsymbol{x} \in \mathcal{X}$, then f is* strictly convex *over $\mathcal{X}$.*

Proof: For the proof of the "only if" part of the statement in Part 1, see the proof of [bertsekas1999, Prop. B.4(a)]. The proof of the "if" part of the statement in Part 1 follows by extending the proof of [bertsekas1999, Prop. B.4(c)] to arbitrary open domains, rather than $\mathbb{R}^D$. For the proof of Part 2, see the proof of [bertsekas1999, Prop. B.4(b)].

□

Quiz 1.7 Show that any square matrix can be expressed as the sum of a symmetric matrix (i.e. $\boldsymbol{A} = \boldsymbol{A}^\top$) and an antisymmetric matrix (i.e. $\boldsymbol{A} = -\boldsymbol{A}^\top$).

Quiz 1.8 Let $\boldsymbol{A}$ be antisymmetric. Show that $\boldsymbol{x}^\top \boldsymbol{A} \boldsymbol{x} = 0$, $\forall \boldsymbol{x}$.

Quiz 1.9 Use the facts in the previous two quizzes to show that given a function $f(\boldsymbol{x}) \triangleq \boldsymbol{x}^\top \boldsymbol{A} \boldsymbol{x} + \boldsymbol{b}^\top \boldsymbol{x}$, where $\boldsymbol{A}$ is not necessarily symmetric, one can always find a **symmetric** matrix $\boldsymbol{A}'$ such that $f(\boldsymbol{x}) = \boldsymbol{x}^\top \boldsymbol{A}' \boldsymbol{x} + \boldsymbol{b}^\top \boldsymbol{x}$.

Thus, when dealing with quadratic functions, we can assume without loss of generality that the matrix in the quadratic term is symmetric.

Quiz 1.10 For which class of matrices $\boldsymbol{A}$ is $f(\boldsymbol{x}) \triangleq \boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}^\top \boldsymbol{x}$ a convex function?

Quiz 1.11 Provide a counterexample that shows that the converse of Part 2 of Proposition 1.4 is not true. In other words, you need to provide an example of a twice-differentiable function that is convex but its Hessian is not positive definite.

1.2.1.3 Convex problems

There are multiple definitions of "convex problem". For instance:
[general def.] Perhaps the most general definition is:

Definition 1.3 *A* convex *optimization problem is an optimization problem of the form* (1.1) *where* f *is a convex function and* $\mathcal{X}$ *is a convex set.*

Note that any problem of the form (1.3) can be expressed as in (1.1) for some $\mathcal{X}$.
[Alternative def.]An alternative definition used e.g. in [boyd] is:

Definition 1.4 *A* convex *optimization problem is an optimization problem of the form* (1.3) *where (i)* f *is convex; (ii)* $g_i(\boldsymbol{x})$ *are convex for all* i*; and (iii)* $h_i(\boldsymbol{x})$ *are linear for all* i*.*

Quiz 1.12 Find an example of optimization problem of the form (1.3) that is not convex according to Definition 1.4, yet it is convex according to Definition 1.3.

One must take into account that different books may adopt different definitions. It is also important to note that CVX sticks to an even stricter definition than Definition 1.4, where the objective and constraint functions must furthermore satisfy rules of disciplined convex programming.

1.2.2 Strong Convexity

[Overview] In words, a function is strongly convex if its curvature is greater than or equal to a positive constant. Consequently, a quadratic lower bound can be found for these functions, which strengthens convergence guarantees of many important algorithms.
[Note on multiple defs.] There are multiple definitions of strong convexity in the literature, some of them difficult to remember. However, they can be shown to be equivalent when the function is differentiable; the latter condition being necessary since some definitions only apply to differentiable functions. In these notes, we define strong convexity in Definition 1.5 and show that strongly convex functions defined in this way satisfy certain properties stated as Proposition 1.6 and Proposition 1.8. However, there are papers and books that define strong convexity using the property in Proposition 1.6 or Proposition 1.8, and show (or use the fact) that strongly convex functions defined in this way satisfy the condition in Definition 1.5. Nonetheless, our choice is more general since it does not require differentiability and may be easier to remember.

- [def.]

Definition 1.5 *Let* $\mathcal{D} \subset \mathbb{R}^D$ *be an open convex set. A function* $f : \mathcal{D} \to \mathbb{R}$ *is strongly convex if*

$$g(\boldsymbol{x}) \triangleq f(\boldsymbol{x}) - \frac{\alpha}{2}||\boldsymbol{x}||^2$$

is convex for some $\alpha > 0$*.*

In words, this definition states that a function is strongly convex if after subtracting a quadratic term (arguably the simplest strictly convex function), the result is still a convex function.

Quiz 1.13 Provide an example of a convex function that is strongly convex and another that is not.

Quiz 1.14 Suppose that f is α-strongly convex. Is it also necessarily α'-strongly convex for all $\alpha' > \alpha$? Is it also necessarily α'-strongly convex for all α' such that $0 < \alpha' < \alpha$?

- [Properties] Strong convexity is a stronger (i.e. stricter) property than convexity, and then one must expect that they also satisfy correspondingly stronger properties. The following results provide some of these properties.
  - [convexity] Of course, no reckless mind would dare to baptize the functions satisfying the condition in Definition 1.5 as "strongly convex" if those functions were not convex. Actually, they can be shown to be *strictly* convex, which is an even stronger condition. The following result settles this issue:

Proposition 1.5 *If f is strongly convex, then it is strictly convex.*

Quiz 1.15 Prove Proposition 1.5.

  - [Quadratic lower bound]Recall from Proposition 1.1 that all convex functions admit a linear lower bound as in (1.4). Strong convexity takes us a step further by guaranteeing the existence of a convex *quadratic* lower bound, as shown by the following result. See Appendix A.3 for a review of subgradients.

Proposition 1.6 *Let $\mathcal{D} \subset \mathbb{R}^D$ be an open convex set. A function $f : \mathcal{D} \to \mathbb{R}$ is strongly convex if and only if there exists $\alpha > 0$ satisfying, for every $\boldsymbol{z}$ and $\boldsymbol{x}$,*

$$f(\boldsymbol{z}) \geq f(\boldsymbol{x}) + \boldsymbol{d}_{\boldsymbol{x}}^\top (\boldsymbol{z} - \boldsymbol{x}) + \frac{\alpha}{2}||\boldsymbol{z} - \boldsymbol{x}||^2,$$

where $\boldsymbol{d}_{\boldsymbol{x}}$ is any subgradient of f at $\boldsymbol{x}$.

Proof: "$\Rightarrow$": If $g(\boldsymbol{x}) \triangleq f(\boldsymbol{x}) - \frac{\alpha}{2}||\boldsymbol{x}||^2$ is convex, then at every point $\boldsymbol{x}$ there exists a subgradient $\boldsymbol{b}_g$ (Proposition *A.5*). Therefore

$$g(\boldsymbol{z}) \geq g(\boldsymbol{x}) + \boldsymbol{b}_g^\top (\boldsymbol{z} - \boldsymbol{x}).$$

A possible choice for $\boldsymbol{b}_g$ is $\boldsymbol{b}_g = \boldsymbol{b}_f - \alpha \boldsymbol{x}$, where $\boldsymbol{b}_f$ is any subgradient of f at $\boldsymbol{x}$, resulting in

$$\begin{aligned} & f(\boldsymbol{z}) - \frac{\alpha}{2}||\boldsymbol{z}||^2 \\ & = g(\boldsymbol{z}) \geq g(\boldsymbol{x}) + (\boldsymbol{b}_f - \alpha \boldsymbol{x})^\top (\boldsymbol{z} - \boldsymbol{x}) \\ & = f(\boldsymbol{x}) - \frac{\alpha}{2}||\boldsymbol{x}||^2 + (\boldsymbol{b}_f - \alpha \boldsymbol{x})^\top (\boldsymbol{z} - \boldsymbol{x}). \end{aligned} \tag{1.8}$$

or

$$\begin{aligned} f(\boldsymbol{z}) & \geq f(\boldsymbol{x}) + \boldsymbol{b}_f^\top (\boldsymbol{z} - \boldsymbol{x}) + \frac{\alpha}{2}||\boldsymbol{z}||^2 - \frac{\alpha}{2}||\boldsymbol{x}||^2 - \alpha \boldsymbol{x}^\top (\boldsymbol{z} - \boldsymbol{x}) \\ & = f(\boldsymbol{x}) + \boldsymbol{b}_f^\top (\boldsymbol{z} - \boldsymbol{x}) + \frac{\alpha}{2}||\boldsymbol{z} - \boldsymbol{x}||^2, \end{aligned} \tag{1.9}$$

which concludes the proof.

"$\Leftarrow$": Note that (1.9) is equivalent to (1.8). Therefore, we have that

$$g(\boldsymbol{z}) \geq g(\boldsymbol{x}) + (\boldsymbol{b}_f - \alpha\boldsymbol{x})^\top(\boldsymbol{z} - \boldsymbol{x})$$

for every $\boldsymbol{x}$ and $\boldsymbol{z}$. This means that a subgradient $\boldsymbol{b}_f - \alpha\boldsymbol{x}$ can be found for g at every point, which shows that g is convex; see Proposition A.4.

□

– [How far from optimum?] Proposition 1.6 allows us to find an upper bound for the *optimality gap*, defined as $f(\boldsymbol{x}) - f^*$ for f^* the global minimum, that only depends on $\boldsymbol{d}_{\boldsymbol{x}}$ and therefore can be evaluated:

Proposition 1.7 *Let $\mathcal{D} \subset \mathbb{R}^D$ be an open convex set and let $f : \mathcal{D} \to \mathbb{R}$ be an α-strongly convex function. If $\boldsymbol{d}_{\boldsymbol{x}}$ is a subgradient at point $\boldsymbol{x}$, then*

$$f(\boldsymbol{x}) - f^* \leq \frac{1}{2\alpha}||\boldsymbol{d}_{\boldsymbol{x}}||_2^2. \tag{1.10}$$

Quiz 1.16 Prove Proposition 1.7.

– [Strictly monotonic gradient] A stronger property than Proposition 1.3 holds for strongly convex functions:[1]

Proposition 1.8 *Let $\mathcal{D} \subset \mathbb{R}^D$ be an open convex set. A continuously differentiable function $f : \mathcal{D} \to \mathbb{R}$ is strongly convex if and only if there exists $\alpha > 0$ such that*

$$(\nabla f(\boldsymbol{z}) - \nabla f(\boldsymbol{x}))^\top(\boldsymbol{z} - \boldsymbol{x}) \geq \alpha||\boldsymbol{z} - \boldsymbol{x}||^2$$

for all $\boldsymbol{x}$ and $\boldsymbol{z}$.

Proof: From Proposition 1.3, $g(\boldsymbol{x}) \triangleq f(\boldsymbol{x}) - \frac{\alpha}{2}||\boldsymbol{x}||^2$ is convex iff

$$(\nabla g(\boldsymbol{z}) - \nabla g(\boldsymbol{x}))^\top(\boldsymbol{z} - \boldsymbol{x}) \geq 0$$

or, since $\nabla g(\boldsymbol{x}) = \nabla f(\boldsymbol{x}) - \alpha\boldsymbol{x}$,

$$(\nabla f(\boldsymbol{z}) - \nabla f(\boldsymbol{x}) - \alpha(\boldsymbol{z} - \boldsymbol{x}))^\top(\boldsymbol{z} - \boldsymbol{x}) \geq 0.$$

Rearranging terms concludes the proof.

□

– [Curvature]Recall from Proposition 1.4 that convex functions have a positive semidefinite Hessian. Strong convexity provides a stronger condition: the eigenvalues of the Hessian are not just non-negative, but greater than a positive constant.

Proposition 1.9 *Let $\mathcal{D} \subset \mathbb{R}^D$ be an open convex set. A twice continuously differentiable function $f : \mathcal{D} \to \mathbb{R}$ is strongly convex if and only if there exists some $\alpha > 0$ such that*

$$\nabla^2 f(\boldsymbol{x}) \succeq \alpha\boldsymbol{I}$$

Proof: $g(\boldsymbol{x}) \triangleq f(\boldsymbol{x}) - \frac{\alpha}{2}||\boldsymbol{x}||^2$ is convex if and only if $\nabla^2 g(\boldsymbol{x}) \succeq \boldsymbol{0}$. Noting that $\nabla^2 g(\boldsymbol{x}) = \nabla^2 f(\boldsymbol{x}) - \alpha\boldsymbol{I}$ concludes the proof.

[1] The definition of strong convexity in [bertsekas1999, eq. (B.6)] is based on the following condition, which therefore only applies to differentiable functions.

□

Quiz 1.17 Let $\boldsymbol{A} = \boldsymbol{B} + \beta \boldsymbol{I}$, where $\boldsymbol{B}$ is a matrix with eigenvalues $\lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_D$, not necessarily non-negative. For which values of β is $f(\boldsymbol{x}) \triangleq \boldsymbol{x}^\top \boldsymbol{A} \boldsymbol{x} + \boldsymbol{b}^\top \boldsymbol{x}$ strongly convex?

1.2.3 Lipschitz smoothness

[Overview] As seen in Sec. 1.2.2, strong convexity is very useful since it provides a quadratic lower bound for the objective function. Another very useful property of an objective function is that of having a Lipschitz continuous gradient. Although initially these two properties may look unrelated, at the end of this section we will understand that each one can be thought of as the counterpart of the other.

- [def] In words, a function is L-Lipschitz smooth[2] if its gradient is Lipschitz continuous. Formally:

Definition 1.6 *Let $\mathcal{D} \subset \mathbb{R}^D$ be an open set. A differentiable function $f : \mathcal{D} \to \mathbb{R}$ is L-Lipschitz smooth (or has an L-Lipschitz continuous gradient) if*

$$||\nabla f(\boldsymbol{z}) - \nabla f(\boldsymbol{x})|| \leq L||\boldsymbol{z} - \boldsymbol{x}|| \quad \forall \boldsymbol{x}, \boldsymbol{z},$$

for some $L > 0$.

Quiz 1.18 Provide two examples of L-Lipschitz smooth functions. What is the value of L?

Quiz 1.19 Provide an example of a differentiable function that is not L-Lipschitz smooth.

Quiz 1.20 Suppose that f is L-Lipschitz smooth. Is it also necessarily L'-Lipschitz smooth for all $L' > L$? Is it also necessarily L'-Lipschitz smooth for all L' such that $0 < L' < L$?

- [properties]
  - [Not "too convex"] Intuitively, according to Definition 1.5, a function is strongly convex if it is not merely convex, but "sufficiently" convex. This is an informal way of saying that after subtracting a sufficiently small quadratic term, the function remains convex. In contrast, Lemma 1.1 (complemented later by Theorem 1.2) states that a Lipschitz smooth function is not "too convex", in the sense that if we subtract a sufficiently large quadratic term, the resulting function becomes *concave*.

Lemma 1.1 *Let $\mathcal{D} \subset \mathbb{R}^D$ be an open convex set. If a differentiable function $f : \mathcal{D} \to \mathbb{R}$ is L-Lipschitz smooth, then*

$$g(\boldsymbol{x}) \triangleq f(\boldsymbol{x}) - \frac{L}{2}||\boldsymbol{x}||^2$$

is concave.

[2] Be aware that the word "smooth" may be used in different contexts with different meanings; sometimes to mean that a function is differentiable, other times to mean that a function changes slowly relative to its input variable, and so on.

Proof: According to Proposition 1.3, $g(\boldsymbol{x})$ is concave iff

$$(\nabla g(\boldsymbol{z}) - \nabla g(\boldsymbol{x}))^\top (\boldsymbol{z} - \boldsymbol{x}) \leq 0$$

or, equivalently, iff

$$(\nabla f(\boldsymbol{z}) - \nabla f(\boldsymbol{x}) - L(\boldsymbol{z} - \boldsymbol{x}))^\top (\boldsymbol{z} - \boldsymbol{x}) \leq 0$$
$$(\nabla f(\boldsymbol{z}) - \nabla f(\boldsymbol{x}))^\top (\boldsymbol{z} - \boldsymbol{x}) \leq L||\boldsymbol{z} - \boldsymbol{x}||^2.$$

The proof is concluded by noting that the latter expression holds true for f any L-Lipschitz smooth function due to the Cauchy-Schwartz inequality.

□

Quiz 1.21 Is $f(x) = x^4$ L-Lipschitz smooth?

– [Quadratic upper bound] Proposition 1.6 establishes a quadratic *lower* bound for strongly convex functions. The following result provides a quadratic *upper* bound for Lipschitz smooth functions.

Proposition 1.10 *Let $\mathcal{D} \subset \mathbb{R}^D$ be an open convex set and let $f : \mathcal{D} \to \mathbb{R}$ be L-Lipschitz smooth. Then,*

$$f(\boldsymbol{z}) \leq f(\boldsymbol{x}) + \nabla f^\top(\boldsymbol{x})(\boldsymbol{z} - \boldsymbol{x}) + \frac{L}{2}||\boldsymbol{z} - \boldsymbol{x}||_2^2 \tag{1.11}$$

for all $\boldsymbol{x}$, $\boldsymbol{z}$.

Proof: Consider first the following result:

Lemma 1.2 *Let $\mathcal{D} \subset \mathbb{R}^D$ be an open convex set. Furthermore, let $f : \mathcal{D} \to \mathbb{R}$ be a differentiable function and let $g(\boldsymbol{x}) \triangleq f(\boldsymbol{x}) - \frac{L}{2}||\boldsymbol{x}||^2$ for some $L \in \mathbb{R}$. Then, g is concave iff*

$$f(\boldsymbol{z}) \leq f(\boldsymbol{x}) + \nabla f^\top(\boldsymbol{x})(\boldsymbol{z} - \boldsymbol{x}) + \frac{L}{2}||\boldsymbol{z} - \boldsymbol{x}||_2^2 \tag{1.12}$$

for all $\boldsymbol{x}$, $\boldsymbol{z}$.

Proof: From Proposition 1.1, $g(\boldsymbol{x})$ is concave iff

$$g(\boldsymbol{z}) \leq g(\boldsymbol{x}) + \nabla g(\boldsymbol{x})^\top (\boldsymbol{z} - \boldsymbol{x})$$

for all $\boldsymbol{x}$, $\boldsymbol{z}$. Combining this expression with the definition of $g(\boldsymbol{x})$ and $\nabla g(\boldsymbol{x}) = \nabla f(\boldsymbol{x}) - L\boldsymbol{x}$ concludes the proof.

□

Thus, if f is L-Lipschitz smooth, then g is concave (Lemma 1.1) and, therefore, (1.11) holds (Lemma 1.2).

□

– [How far from optimum?] Proposition 1.10 allows us to find a lower bound for the *optimality gap*, defined as $f(\boldsymbol{x}) - f^*$ for f^* the global minimum, that only depends on $\nabla f(\boldsymbol{x})$ and therefore can be evaluated:

Proposition 1.11 *Let $\mathcal{D} \subset \mathbb{R}^D$ be an open convex set. Furthermore, let $f : \mathcal{D} \to \mathbb{R}$ be a differentiable function and let $g(\boldsymbol{x}) \triangleq f(\boldsymbol{x}) - \frac{L}{2}||\boldsymbol{x}||^2$. If g is concave (which occurs for example when f is L-Lipschitz smooth), then*

$$f(\boldsymbol{x}) - f^* \geq \frac{1}{2L}||\nabla f(\boldsymbol{x})||_2^2. \tag{1.13}$$

for all $\boldsymbol{x}$.

Proof: Rearranging terms in (1.11), it follows that

$$f(\boldsymbol{x}) - f(\boldsymbol{z}) \geq -\nabla f^\top(\boldsymbol{x})(\boldsymbol{z} - \boldsymbol{x}) - \frac{L}{2}||\boldsymbol{z} - \boldsymbol{x}||_2^2. \tag{1.14}$$

Noting that $f(\boldsymbol{x}) - f^* \geq f(\boldsymbol{x}) - f(\boldsymbol{z})$ establishes that

$$f(\boldsymbol{x}) - f^* \geq -\nabla f^\top(\boldsymbol{x})(\boldsymbol{z} - \boldsymbol{x}) - \frac{L}{2}||\boldsymbol{z} - \boldsymbol{x}||_2^2 \tag{1.15}$$

for all $\boldsymbol{z}$. In particular,

$$f(\boldsymbol{x}) - f^* \geq \sup_{\boldsymbol{z}} \left[-\nabla f^\top(\boldsymbol{x})(\boldsymbol{z} - \boldsymbol{x}) - \frac{L}{2}||\boldsymbol{z} - \boldsymbol{x}||_2^2 \right] = \frac{1}{2L}||\nabla f(\boldsymbol{x})||_2^2, \tag{1.16}$$

which holds for all $\boldsymbol{x}$.

□

– [Counterpart to strong convexity] Expression (1.13) enables us to extend Lemma 1.1 to a necessary and sufficient condition upon assuming convexity:

Theorem 1.2 *Let $\mathcal{D} \subset \mathbb{R}^D$ be an open convex set. A **convex** differentiable function $f : \mathcal{D} \to \mathbb{R}$ is L-Lipschitz smooth if and only if*

$$g(\boldsymbol{x}) \triangleq f(\boldsymbol{x}) - \frac{L}{2}||\boldsymbol{x}||^2$$

is concave.

Proof: "⇒": see Lemma 1.1.
"⇐": Define

$$f_{\boldsymbol{x}}(\boldsymbol{y}) \triangleq f(\boldsymbol{y}) - \nabla f(\boldsymbol{x})^\top \boldsymbol{y}, \qquad f_{\boldsymbol{z}}(\boldsymbol{y}) \triangleq f(\boldsymbol{y}) - \nabla f(\boldsymbol{z})^\top \boldsymbol{y}.$$

These functions satisfy

$$\begin{aligned} f_{\boldsymbol{x}}(\boldsymbol{y}) &\geq f_{\boldsymbol{x}}(\boldsymbol{x}), \quad \forall \boldsymbol{y}, \\ f_{\boldsymbol{z}}(\boldsymbol{y}) &\geq f_{\boldsymbol{z}}(\boldsymbol{z}), \quad \forall \boldsymbol{y} \end{aligned} \tag{1.17}$$

because

$$\begin{aligned} & f_{\boldsymbol{x}}(\boldsymbol{y}) \geq f_{\boldsymbol{x}}(\boldsymbol{x}) \\ \Leftrightarrow\ & f(\boldsymbol{y}) - \nabla f(\boldsymbol{x})^\top \boldsymbol{y} \geq f(\boldsymbol{x}) - \nabla f(\boldsymbol{x})^\top \boldsymbol{x} \\ \Leftrightarrow\ & f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^\top(\boldsymbol{y} - \boldsymbol{x}), \end{aligned}$$

which follows from convexity of f and Proposition 1.1. Moreover, the functions

$$\begin{aligned} g_{\boldsymbol{x}}(\boldsymbol{y}) &\triangleq f_{\boldsymbol{x}}(\boldsymbol{y}) - \frac{L}{2}||\boldsymbol{y}||^2 = f(\boldsymbol{y}) - \frac{L}{2}||\boldsymbol{y}||^2 - \nabla f(\boldsymbol{x})^\top \boldsymbol{y} = g(\boldsymbol{y}) - \nabla f(\boldsymbol{x})^\top \boldsymbol{y} \\ g_{\boldsymbol{z}}(\boldsymbol{y}) &\triangleq f_{\boldsymbol{z}}(\boldsymbol{y}) - \frac{L}{2}||\boldsymbol{y}||^2 = f(\boldsymbol{y}) - \frac{L}{2}||\boldsymbol{y}||^2 - \nabla f(\boldsymbol{z})^\top \boldsymbol{y} = g(\boldsymbol{y}) - \nabla f(\boldsymbol{z})^\top \boldsymbol{y}, \end{aligned}$$

are concave since $g(\boldsymbol{y})$ is concave by hypothesis. This means that Proposition 1.11 applies for $f_{\boldsymbol{x}}(\boldsymbol{y})$ and $f_{\boldsymbol{z}}(\boldsymbol{y})$. Since (1.17) establishes that $\boldsymbol{x}$ (resp. $\boldsymbol{z}$) is the global optimum for $f_{\boldsymbol{x}}(\boldsymbol{y})$ (resp. $f_{\boldsymbol{z}}(\boldsymbol{y})$), then (1.13) reads as

$$\begin{aligned} f_{\boldsymbol{x}}(\boldsymbol{y}) - f_{\boldsymbol{x}}(\boldsymbol{x}) &\geq \frac{1}{2L}||\nabla f_{\boldsymbol{x}}(\boldsymbol{y})||_2^2 \\ f_{\boldsymbol{z}}(\boldsymbol{y}) - f_{\boldsymbol{z}}(\boldsymbol{z}) &\geq \frac{1}{2L}||\nabla f_{\boldsymbol{z}}(\boldsymbol{y})||_2^2. \end{aligned}$$

for any $\boldsymbol{y}$, and in particular:

$$f_{\boldsymbol{x}}(\boldsymbol{z}) - f_{\boldsymbol{x}}(\boldsymbol{x}) \geq \frac{1}{2L}||\nabla f_{\boldsymbol{x}}(\boldsymbol{z})||_2^2$$
$$f_{\boldsymbol{z}}(\boldsymbol{x}) - f_{\boldsymbol{z}}(\boldsymbol{z}) \geq \frac{1}{2L}||\nabla f_{\boldsymbol{z}}(\boldsymbol{x})||_2^2,$$

which in turn implies, upon substituting the definition of $f_{\boldsymbol{x}}$ and $f_{\boldsymbol{z}}$, that

$$f(\boldsymbol{z}) - f(\boldsymbol{x}) - \nabla f(\boldsymbol{x})^\top(\boldsymbol{z} - \boldsymbol{x}) \geq \frac{1}{2L}||\nabla f(\boldsymbol{z}) - \nabla f(\boldsymbol{x})||_2^2$$
$$f(\boldsymbol{x}) - f(\boldsymbol{z}) - \nabla f(\boldsymbol{z})^\top(\boldsymbol{x} - \boldsymbol{z}) \geq \frac{1}{2L}||\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{z})||_2^2.$$

Adding both expressions yields

$$(\nabla f(\boldsymbol{z}) - \nabla f(\boldsymbol{x}))^\top(\boldsymbol{z} - \boldsymbol{x}) \geq \frac{1}{L}||\nabla f(\boldsymbol{z}) - \nabla f(\boldsymbol{x})||_2^2.$$

Using the Cauchy-Schwartz inequality, it follows that f is L-Lipschitz smooth.

□

- [Counterpart to monotonic gradient] The following is the counterpart to Proposition 1.8 for Lipschitz smooth functions.

Proposition 1.12 *Let $\mathcal{D} \subset \mathbb{R}^D$ be an open set and let $f : \mathcal{D} \to \mathbb{R}$ be differentiable. If f is L-Lipschitz smooth, then*

$$(\nabla f(\boldsymbol{z}) - \nabla f(\boldsymbol{x}))^\top(\boldsymbol{z} - \boldsymbol{x}) \leq L||\boldsymbol{z} - \boldsymbol{x}||^2 \quad \forall \boldsymbol{x}, \boldsymbol{z}.$$

Quiz 1.22 Prove Proposition 1.12.

- [Curvature]Recall that Proposition 1.9 establishes a lower bound on the eigenvalues of the Hessian matrix of a strongly convex function. Correspondingly, Lipschitz smoothness provides an upper bound on those eigenvalues.

Proposition 1.13 *Let $\mathcal{D} \subset \mathbb{R}^D$ be an open convex set. A* ***convex*** *and twice differentiable function $f : \mathcal{D} \to \mathbb{R}$ is L-Lipschitz smooth if and only if*

$$\nabla^2 f(\boldsymbol{x}) \preceq L\boldsymbol{I}.$$

Proof: It follows trivially from Theorem 1.2.

□

Quiz 1.23 Let $\boldsymbol{A} = \boldsymbol{B} + \beta\boldsymbol{I}$, where $\boldsymbol{B}$ is a matrix with eigenvalues $\lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_D$, not necessarily non-negative. Given L, for which values of β is $f(\boldsymbol{x}) \triangleq \boldsymbol{x}^\top\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}^\top\boldsymbol{x}$ L-Lipschitz smooth? Conversely, given β, what is the smallest value of L so that f is L-Lipschitz smooth?

Quiz 1.24 Suppose we are given a non-convex differentiable function $f : \mathcal{D} \subset \mathbb{R}^D \to \mathbb{R}$ and we wish to determine whether it is L-Lipschitz smooth. In view of the results presented in this section, what approaches could we follow?

Quiz 1.25 Now solve the previous quiz in the scenario where f is convex.

1.3 Equivalent Problems

- [Overview]
  - [Transforming a problem] Sometimes, one is to solve an optimization problem that cannot be solved easily in its present form. However, there are often ways to transform it into a more convenient form. For example, it is desirable to express the problem in an equivalent form with a convex, strongly convex, or Lipschitz smooth objective; see Sec. 1.2.
  - [Equivalent problems]There are multiple forms of formally defining equivalence between two optimization problems. In practice, however, one may just deem two problems to be equivalent if the solution of one can be "easily" found from the solution of the other.
  - [Sec. overview]This section provides some simple rules intended to assist in finding equivalent problems with desirable properties. The proofs of these rules are straightforward and left as an exercise.
- [Transforming the objective]

  Rule 1: *Applying an increasing transformation*
  Composing the objective $f(\boldsymbol{x})$ with an increasing transformation $g(z)$ does not change the set of optimal points, yet the resulting objective $g(f(\boldsymbol{x}))$ may have desirable properties such as convexity.

Example 1.2 In statistics, it is customary to derive maximum likelihood estimators. A typical problem arising in this context is of the form

$$\underset{\mu}{\text{maximize}} \quad e^{-\sum_n (x_n - \mu)^2}$$

or, equivalently,

$$\underset{\mu}{\text{minimize}} \quad -e^{-\sum_n (x_n - \mu)^2}.$$

This objective is non-convex, but one can compose it with the increasing function $g(z) \triangleq -\log(-z)$ to obtain

$$\underset{\mu}{\text{minimize}} \quad \sum_n (x_n - \mu)^2,$$

which is a convex program; in this simple case even with closed-form solution, see Sec. 1.4.

Quiz 1.26 Obtain an equivalent convex problem to

$$\underset{\mu}{\text{minimize}} \quad \log\left(\sum_n (x_n - \mu)^2\right)$$

and indicate the transformation applied.

Quiz 1.27 What if g is just non-decreasing? Are the minimizers of $g(f(\boldsymbol{x}))$ necessarily the same as those of $f(\boldsymbol{x})$?

- [Transforming the feasible set/constraints]

  Rule 2: *Simple transformation of constraints*
  Sometimes, rewriting a constraint may turn a problem convex according to Definition 1.4 (therefore this trick is useful for using CVX).

Example 1.3 Suppose that f is convex. According to Definition 1.4, the problem

$$\begin{aligned} \underset{x_1, x_2 \in \mathbb{R}}{\text{minimize}} \quad & f(x_1, x_2) \\ \text{subject to} \quad & \frac{x_1}{x_2} - 4 \le 0 \\ & x_2 \ge 2 \end{aligned}$$

is not convex because $x_1/x_2 - 4$ is not a convex function. However, rewriting $x_1/x_2 - 4 \le 0$ as $x_1 - 4x_2 \le 0$ gives rise to a convex problem.

Quiz 1.28 How would you transform the constraint $||\boldsymbol{A}\boldsymbol{x}|| = 0$ into a valid constraint for a convex problem?

Rule 3: *Reduction of the feasible set*
Suppose that $\boldsymbol{x}$ and $\boldsymbol{z}$, with $\boldsymbol{z} \neq \boldsymbol{x}$ are both feasible and achieve the same objective, i.e., $f(\boldsymbol{x}) = f(\boldsymbol{z})$. Then one can disregard any of them to obtain an equivalent problem.

Example 1.4 Consider the problem

$$\begin{aligned} \underset{x \in \mathbb{R}}{\text{minimize}} \quad & x^2 \\ \text{subject to} \quad & x \in [-3, -1] \cup [1, 3]. \end{aligned}$$

This non-convex problem is clearly equivalent to the convex problem

$$\begin{aligned} \underset{x \in \mathbb{R}}{\text{minimize}} \quad & x^2 \\ \text{subject to} \quad & x \in [-3, -1]. \end{aligned}$$

Every solution to the second problem is also a solution to the first; moreover if x^* is a solution to the second problem, $-x^*$ is a solution to the first problem.

- [Transforming both the feasible set and objective]

Rule 4: *Precomposition*
Given the function f defined over some set $\mathcal{X}$

$$\begin{array}{ccc} \mathcal{X} & \to & \mathbb{R} \\ \boldsymbol{x} & \mapsto & f(\boldsymbol{x}), \end{array} \tag{1.18}$$

consider the problem

$$\underset{\boldsymbol{x} \in \mathcal{X}}{\text{minimize}}\, f(\boldsymbol{x}). \tag{1.19}$$

If we define another function g defined over some set $\mathcal{Z}$ such that $\mathcal{X} = \underset{g}{\text{image}}\{\mathcal{Z}\}$, that is:

$$\begin{array}{ccccc} \mathcal{Z} & \xrightarrow{g} & \mathcal{X} & \xrightarrow{f} & \mathbb{R} \\ \boldsymbol{z} & \mapsto & \boldsymbol{x} = g(\boldsymbol{z}) & \mapsto & f(\boldsymbol{x}) = f(g(u)) \end{array} \tag{1.20}$$

In that case, the problem above is equivalent to

$$\underset{\boldsymbol{z} \in \mathcal{Z}}{\text{minimize}}\, f(g(\boldsymbol{z})). \tag{1.21}$$

Example 1.5 Consider the problem

$$\begin{aligned} \underset{x\in\mathbb{R}}{\text{minimize}} \quad & f(x) \\ \text{subject to} \quad & x \geq 1. \end{aligned}$$

This problem is clearly equivalent to

$$\begin{aligned} \underset{z\in\mathbb{R}}{\text{minimize}} \quad & f(-z) \\ \text{subject to} \quad & z \leq -1, \end{aligned}$$

which follows by setting $g(z) = -z$.

Rule 5: *Exploiting equi-cost sets*
Given the function f defined over some set $\mathcal{X}$

$$\begin{array}{ccc} \mathcal{X} & \to & \mathbb{R} \\ x & \mapsto & f(x), \end{array} \tag{1.22}$$

consider the problem

$$\underset{x\in\mathcal{X}}{\text{minimize}}\, f(x). \tag{1.23}$$

Sometimes, f is a many-to-one mapping, i.e., $\exists x_1, x_2 \in \mathcal{X},\ x_1 \neq x_2 : f(x_1) = f(x_2)$. In these cases, one can further decompose $f(x)$ as $f_2(f_1(x))$ where $f_1(x)$ is a many-to-one mapping:

$$\begin{array}{ccccc} \mathcal{X} & & \xrightarrow{\qquad\qquad f \qquad\qquad} & & \mathbb{R} \\ \mathcal{X} & \xrightarrow{f_1} & \mathcal{Z} & \xrightarrow{f_2} & \mathbb{R} \\ x & \mapsto & z = f_1(x) & \mapsto & f_2(z) \end{array} \tag{1.24}$$

where $\mathcal{Z} = \underset{f_1}{\text{image}}\{\mathcal{X}\}$. Hence, (1.23) is equivalent to

$$\underset{z\in\mathcal{Z}}{\text{minimize}}\, f_2(z). \tag{1.25}$$

This trick can be used e.g. to reduce the number of variables:

Example 1.6 For a_1, a_2, b_1, and b_2 given constants, the following two problems are equivalent

$$\begin{aligned} \underset{\mathbf{x}}{\text{minimize}} \quad & x_1 + x_2 + x_3 \\ \text{subject to} \quad & a_1 \leq x_1 \leq b_1 \\ & a_2 \leq x_2 + x_3 \leq b_2 \end{aligned} \qquad \begin{aligned} \underset{\mathbf{z}}{\text{minimize}} \quad & z_1 + z_2 \\ \text{subject to} \quad & a_1 \leq z_1 \leq b_1 \\ & a_2 \leq z_2 \leq b_2. \end{aligned} \tag{1.26}$$

It can also be used to turn a non-convex problem into a convex one:

Example 1.7 The following two problems are equivalent

$$\underset{x\in\mathbb{R}}{\text{minimize}} \quad \sin^2(x) + 3\sin(x) - 5 \qquad \begin{aligned} \underset{z\in\mathbb{R}}{\text{minimize}} \quad & z^2 + 3z - 5 \\ \text{subject to} \quad & -1 \leq z \leq 1. \end{aligned} \tag{1.27}$$

Quiz 1.29 Suppose that $\boldsymbol{A} \in \mathbb{R}^{N \times D}$ is full row rank and consider the problem

$$\underset{\boldsymbol{x} \in \mathbb{R}^D}{\text{minimize}}\, f(\boldsymbol{A}\boldsymbol{x}).$$

Obtain a simpler equivalent problem.

Rule 6: *Moving complexity from the objective to the constraints*
Suppose that one is given the problem in (1.1) where f is somehow difficult to handle as an objective function. However, the following equivalent problem

$$\begin{aligned} \underset{t,\boldsymbol{x}}{\text{minimize}} \quad & t \\ \text{subject to} \quad & t \geq f(\boldsymbol{x}) \\ & \boldsymbol{x} \in \mathcal{X} \end{aligned}$$

may be easier to solve.

Example 1.8 If $\mathcal{X}$ is convex and $f(\boldsymbol{x}) = \boldsymbol{v}^\top \boldsymbol{A}^{-1}(\boldsymbol{x})\boldsymbol{v}$ with $\boldsymbol{A}(\boldsymbol{x}) \succ \boldsymbol{0}\ \forall \boldsymbol{x} \in \mathcal{X}$, then the constraint $t \geq f(\boldsymbol{x})$ becomes

$$\begin{bmatrix} t & \boldsymbol{v}^\top \\ \boldsymbol{v} & \boldsymbol{A}(\boldsymbol{x}) \end{bmatrix} \succeq \boldsymbol{0}$$

which yields a semidefinite program (SDP) [boyd]. If the function $\boldsymbol{A}(\boldsymbol{x})$ is linear, then one can resort to off-the-shelf solvers like SeDuMi or SDPT3.

1.4 Solving Optimization Problems

[Overview]Broadly speaking, most optimization problems cannot be solved by simple inspection. For this reason, one generally adopts one of the following approaches.

- [analytical approach] In certain problems, a solution can be obtained in closed form upon applying certain optimality conditions, such as the ones described in Ch. 2.
  - [Benefits] Evaluating the resulting closed-form expression is typically the most computationally efficient method to find the solution.
  - [How to find it] In many cases, closed-form solutions are obtained by applying the first-order necessary conditions for optimality; see Ch. 2. For differentiable objectives, these conditions turn an optimization problem into that of solving a system of (typically non-linear) equations:

    $$\text{optimization problem} \xrightarrow{\text{first-order conditions}} \text{system of equations}$$

    A closed-form solution to the optimization problem is found therefore if one can solve this system of equations in closed form.
- [Numerical approach] When a closed-form solution cannot be found or when it involves high computational complexity, one typically resorts to iterative algorithms, as described in Ch. 3 and the rest of the course.

Chapter 2

Optimality Conditions and Duality

This chapter summarizes some of the most important optimality conditions for constrained optimization problems. No algorithms are provided, but the material here lays the theoretical fundations of the most common algorithms and can be used to find the closed-form solution of certain optimization problems.

2.1 Introduction

[Motivation]Optimality conditions

- enable us to guarantee that a point is optimal (sufficient conditions).
- indicate if a point is not optimal (necessary conditions).
- narrow down the list of candidate optimal points (necessary conditions).
- sometimes lead to closed-form solutions → typically the most efficient algorithm!
- lie at the heart of the most important optimization algorithms.
- are useful to analyze convergence and performance of optimization algorithms.

2.2 Unconstrained Optimization Problems

2.2.1 Necessary conditions when $D = 1$

To gain intuition, let us first look at the case where $\mathcal{X} = \mathbb{R}$.

Proposition 2.1 *Let $f : \mathbb{R} \to \mathbb{R}$ be continuously differentiable and let x be a local minimum. Then, $f'(x^*) = 0$.*

Proof: On the one hand,

$$f'(x^*) = \lim_{x \downarrow x^*} \frac{f(x) - f(x^*)}{x - x^*} \geq 0.$$

On the other,

$$f'(x^*) = \lim_{x \uparrow x^*} \frac{f(x) - f(x^*)}{x - x^*} \leq 0,$$

then, it follows that $f'(x^*) = 0$.

□

[interpretation]Proposition 2.1 embodies the observation that a function cannot be increasing or decreasing at a local minimum, since in that case one would obtain a lower objective value by moving to the left or right, respectively.
[stationary points]This result, together with its extension in Proposition 2.4, motivates the term *stationary point* to refer to those points $\boldsymbol{x}$ for which $\nabla f(\boldsymbol{x}) = \mathbf{0}$. Stationary points can be classified as maxima, minima, or saddle points, the latter being defined as those stationary points that are neither maxima nor minima.

Proposition 2.2 *Let $f : \mathbb{R} \to \mathbb{R}$ be twice continuously differentiable and let x^* be a local minimum. Then, $f''(x^*) \geq 0$.*

Proof: From the mean-value theorem (Proposition *A.7*), it follows that

$$f(x) = f(x^*) + f'(x^*)(x - x^*) + \frac{1}{2} f''(\eta_x)(x - x^*)^2,$$

or

$$\frac{1}{2} f''(\eta_x) = \frac{f(x) - f(x^*) - f'(x^*)(x - x^*)}{(x - x^*)^2}.$$

Due to Proposition 2.1,

$$\frac{1}{2} f''(\eta_x) = \frac{f(x) - f(x^*)}{(x - x^*)^2}.$$

Thus

$$\frac{1}{2} f''(x^*) = \lim_{x \to x^*} \frac{1}{2} f''(\eta_x) = \lim_{x \to x^*} \frac{f(x) - f(x^*)}{(x - x^*)^2} \geq 0.$$

□

[interpretation] Since f'' is the derivative of f', one has that $f'(x)$ is increasing wherever $f''(x) > 0$. Together with Proposition 2.1, this fact then implies that, within a neighborhood of x^*, it holds that $f'(x) < 0$ for $x < x^*$ and $f'(x) > 0$ for $x > x^*$. Thus, whereas Proposition 2.1 standalone establishes that f can be neither increasing nor decreasing at x^*, Proposition 2.2 goes one step further by requiring that, within a neighborhood of x^*, function f must be decreasing for $x < x^*$ and increasing for $x > x^*$.

2.2.2 Sufficient conditions when $D = 1$

Proposition 2.3 *Let $f : \mathbb{R} \to \mathbb{R}$ be twice continuously differentiable and suppose that there exists $x^* \in \mathbb{R}$ such that $f'(x^*) = 0$ and $f''(x^*) > 0$. Then, x^* is a strict local minimum of f.*

Proof: See Proposition 2.6.

□

Clearly, this proposition also implies that if $f'(x^*) = 0$ and $f''(x^*) < 0$, then x^* is a strict local maximum.

Quiz 2.1 Thus, a stationary point x^* is a maximum if $f''(x^*) < 0$ and a minimum if $f''(x^*) > 0$. What happens if $f''(x^*) = 0$? Can we say that x^* is a saddle point in that case? Hint: consider the stationary points of $f(x) = x^4$ and $f(x) = -x^4$.

2.2.3 Necessary conditions when $D \geq 1$

Proposition 2.4 *Let $f : \mathbb{R}^D \to \mathbb{R}$ be continuously differentiable and let $\boldsymbol{x}^*$ be a local minimum. Then, $\nabla f(\boldsymbol{x}^*) = \mathbf{0}$.*

Proof: See [bertsekas1999, Prop. 1.1.1].

□

Proposition 2.5 *Let $f : \mathbb{R}^D \to \mathbb{R}$ be twice continuously differentiable and let $\boldsymbol{x}^*$ be a local minimum. Then, $\nabla^2 f(x^*) \succeq \mathbf{0}$.*

Proof: See [bertsekas1999, Prop. 1.1.1].

□

2.2.4 Sufficient conditions when $D \geq 1$

Proposition 2.6 *Let $f : \mathbb{R}^D \to \mathbb{R}$ be twice continuously differentiable and suppose that there exists $\boldsymbol{x}^* \in \mathbb{R}^D$ such that $\nabla f(\boldsymbol{x}^*) = \mathbf{0}$ and $\nabla^2 f(\boldsymbol{x}^*) \succ \mathbf{0}$. Then, $\boldsymbol{x}^*$ is a strict local minimum of f.*

Proof: See [bertsekas1999, Prop. 1.1.3].

□

[Convex Problems]

The same argument in the proof of Proposition 1.2 shows that any point $\boldsymbol{x} \in \mathcal{X}$ satisfying $\nabla f(\boldsymbol{x}) = \mathbf{0}$ is necessarily a global optimum of f. In other words, any stationary point of a convex function is a global optimum. Therefore, when minimizing convex objectives, it is not necessary to look at the second-order conditions. This argument further supports the claim that minimizing convex functions is easier than minimizing non-convex functions.

2.2.5 Closed-form Solutions

[Overview] The above optimality conditions can be used to find the closed-form solution of certain problems. When f is differentiable, one may try to obtain such a solution by solving the (generally non-linear) system of equations $\nabla f(\boldsymbol{x}) = \mathbf{0}$.

[Alternative approach to find closed-form solutions] However, instead of directly solving $\nabla f(\boldsymbol{x}) = \mathbf{0}$, sometimes it is convenient to first fix some variables and optimize w.r.t. the others. The obtained minimizers are, therefore, functions of the fixed variables. This approach is illustrated by the following example:

Example 2.1 Consider the differentiable 2D problem:

$$\underset{x,y}{\text{minimize}} \quad f(x, y) \tag{2.1}$$

- The standard approach would involve finding the solution $(x^\star, y^\star)$ of

$$\begin{cases} \frac{\partial}{\partial x} f(x, y) = 0 \\ \frac{\partial}{\partial y} f(x, y) = 0. \end{cases}$$

- An alternative approach is as follows:
  - First solve $\frac{\partial}{\partial x} f(x, y) = 0$ w.r.t. x for a fixed y $\rightarrow$ the resulting minimizer $x^\star(y)$ obviously depends on the fixed y.
  - Substituting $x^\star(y)$ in the objective yields the equivalent problem

$$\underset{y}{\text{minimize}} \quad f(x^\star(y), y).$$

  - Find the optimum $y^\star$ by invoking the first-order optimality conditions:

$$\frac{\partial}{\partial y} f(x^\star(y), y) = 0 \tag{2.2}$$

  - The minimizer of (2.1) is $(x^\star(y^\star), y^\star)$.

It can be shown that both approaches are the same just by using the chain rule, which states that

$$\frac{\partial}{\partial y} f(g_1(y), g_2(y)) = \frac{\partial}{\partial \tau} f(\tau, g_2(y))\Big|_{\tau=g_1(y)} \frac{\partial}{\partial y} g_1(y) + \frac{\partial}{\partial \tau} f(g_1(y), \tau)\Big|_{\tau=g_2(y)} \frac{\partial}{\partial y} g_2(y).$$

Applying this chain rule to (2.2) yields

$$\begin{aligned} \frac{\partial}{\partial y} f(x^\star(y), y) &= \frac{\partial}{\partial \tau} f(\tau, y)\Big|_{\tau=x^\star(y)} \frac{\partial}{\partial y} x^\star(y) + \frac{\partial}{\partial y} f(x, y)\Big|_{x=x^\star(y)} \\ &= \frac{\partial}{\partial y} f(x, y)\Big|_{x=x^\star(y)} \end{aligned}$$

Therefore both approaches are equivalent.

[Closed-form solution for a subset of variables]The above example can actually be generalized to arbitrary (not necessarily differentiable) functions:

Proposition 2.7 *Consider the function $f : \mathbb{R}^D \rightarrow \mathbb{R}$ and assume that $g_{\boldsymbol{x}_0}(\boldsymbol{z}) \triangleq f(\boldsymbol{x}_0, \boldsymbol{z})$ attains its minimum for all $\boldsymbol{x}_0$ (see e.g. Theorem 1.1). It then holds that*

$$\inf_{\boldsymbol{x},\boldsymbol{z}} f(\boldsymbol{x}, \boldsymbol{z}) = \inf_{\boldsymbol{x}} f(\boldsymbol{x}, \boldsymbol{z}^\star(\boldsymbol{x})) \tag{2.3}$$

where the function $\boldsymbol{z}^\star(\boldsymbol{x})$ returns the optimal $\boldsymbol{z}^\star$ for each $\boldsymbol{x}$, that is,

$$\boldsymbol{z}^\star(\boldsymbol{x}) \in \arg\min_{\boldsymbol{z}} f(\boldsymbol{x}, \boldsymbol{z}).$$

Quiz 2.2 Prove Proposition 2.7. Hint: prove the more general result $\inf_{\boldsymbol{x},\boldsymbol{z}} f(\boldsymbol{x},\boldsymbol{z}) = \inf_{\boldsymbol{x}}[\inf_{\boldsymbol{z}} f(\boldsymbol{x},\boldsymbol{z})]$.

This result establishes that one can first optimize w.r.t. a subset of variables and then w.r.t. the rest. If a closed-form expression can be obtained for $\boldsymbol{z}^*(\boldsymbol{x})$, then the minimization in the right-hand side of (2.3) may be easier than the minimization in the left-hand side. In practice, this means that one may solve a problem by optimizing in closed-form w.r.t. a subset of variables and numerically w.r.t. the rest.

Although Proposition 2.7 was stated for unconstrained optimization, it carries over to constrained optimization. Note that some technicalities arise since the feasible values for $\boldsymbol{z}$ may depend on the selected $\boldsymbol{x}$ and vice versa.

2.3 Constrained Optimization Problems

[Motivation] The precedent theory characterized the minimizers of unconstrained problems. Since this theory cannot directly accommodate *constrained* problems, an extension or generalization is required.

[Overview] To this end, the most natural approach would arguably be, given a constrained problem, find an "equivalent" *unconstrained* problem with the same solution. Unfortunately, we will see that this is not always possible.

[Lagrange multipliers] Most of the ensuing theory will rely on the notion of Lagrange multipliers, which are certain quantities (scalars, vectors, matrices, and so on) that

- [unconstrained] if they exist and they are known, they allow us to obtain an equivalent *unconstrained* problem with the same solution as the original constrained problem.
- [certificate] together with an optimality condition, may be used as a certificate that a given point is actually a solution.
- [properties] may allow us to prove properties of the solution.[1]
- [sensitivity] provide information of how *sensitive* the solution is to perturbations in the constraints.
- [dual] are, in certain cases, the solution to the (Lagrange) dual problem.
- [algorithms] inspire many optimization algorithms.

2.3.1 First- and Second-order Conditions

[overview] When the objective and constraint functions satisfy certain differentiability conditions, it is possible to use their derivatives to narrow down the set of candidate minimizers or to prove that a given point is actually a minimizer.

2.3.1.1 Problems with Equality Constraints

[General form]An equality-constrained optimization problem has the form:

$$\underset{\boldsymbol{x}\in\mathbb{R}^D}{\text{minimize}} \quad f(\boldsymbol{x}) \tag{2.4a}$$

$$\text{subject to} \quad h_n(\boldsymbol{x}) = 0,\ n = 1,\ldots,N. \tag{2.4b}$$

[1] For instance, one can use Lagrange multipliers to prove Snell's refraction law from the principle that the light propagates through the fastest route; see [bertsekas1999, Example 3.1.4].

First-order Necessary Conditions

Linear constraints

[strategy] The strategy in this section to obtain necessary optimality conditions for constrained problems comprises two steps:

- (i) transform a constrained problem into an unconstrained problem.
- (ii) apply the optimality conditions for unconstrained problems presented in Sec. 2.2.

[Problem form]A linearly-constrained problem can be expressed as

$$\underset{\boldsymbol{x}\in\mathbb{R}^D}{\text{minimize}} \quad f(\boldsymbol{x}) \tag{2.5a}$$

$$\text{subject to} \quad \boldsymbol{a}_n^\top \boldsymbol{x} = b_n, \; n = 1, \ldots, N \tag{2.5b}$$

or, by defining $\boldsymbol{A} \triangleq [\boldsymbol{a}_1, \ldots, \boldsymbol{a}_N]^\top$ and $\boldsymbol{b} \triangleq [b_1, \ldots, b_N]^\top$, as

$$\underset{\boldsymbol{x}\in\mathbb{R}^D}{\text{minimize}} \quad f(\boldsymbol{x}) \tag{2.6a}$$

$$\text{subject to} \quad \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}. \tag{2.6b}$$

It will be assumed that the rows of $\boldsymbol{A}$ are linearly independent, which entails no loss of generality to a practical extent.

Quiz 2.3 Why does this assumption entail no loss of generality?

[Unconstrained equivalent] Proposition A.2 implies that problem (2.6) is equivalent to the unconstrained problem

$$\underset{\boldsymbol{z}\in\mathbb{R}^{D-N}}{\text{minimize}} \quad f(\boldsymbol{F}\boldsymbol{z} + \boldsymbol{g}), \tag{2.7}$$

which results from the change of variable $\boldsymbol{x} = \boldsymbol{F}\boldsymbol{z} + \boldsymbol{g}$ with

- $\boldsymbol{F}$ any matrix satisfying $\mathcal{R}\{\boldsymbol{F}\} = \mathcal{N}\{\boldsymbol{A}\}$ and
- $\boldsymbol{g}$ any vector satisfying $\boldsymbol{A}\boldsymbol{g} = \boldsymbol{b}$.

This trick always allows one to get rid of equality constraints, even if there are additional (possibly nonlinear) (in)equality constraints.

Quiz 2.4 In view of this trick, why should we still care about problems with linear equality constraints?

[Necessary optimality conditions]Since (2.7) is unconstrained, one can apply Proposition 2.4 to conclude that at an optimum $\boldsymbol{x}^* \triangleq \boldsymbol{F}\boldsymbol{z}^* + \boldsymbol{g}$ it holds that

$$\boldsymbol{0} = \nabla_{\boldsymbol{z}} f(\boldsymbol{F}\boldsymbol{z}^* + \boldsymbol{g}) = \boldsymbol{F}^\top \left[\nabla_{\boldsymbol{x}} f(\boldsymbol{x})\right]_{\boldsymbol{x}=\boldsymbol{F}\boldsymbol{z}^*+\boldsymbol{g}} = \boldsymbol{F}^\top \nabla f(\boldsymbol{x}^*)$$

This implies that $\nabla_{\boldsymbol{x}} f(\boldsymbol{x}^*) \in \mathcal{N}\{\boldsymbol{F}^\top\} = \mathcal{R}^\perp\{\boldsymbol{F}\} = \mathcal{N}^\perp\{\boldsymbol{A}\} = \mathcal{R}\{\boldsymbol{A}^\top\}$, where the first and third equalities follow from Proposition A.1 and the second from the definition of $\boldsymbol{F}$. Thus, $\exists \boldsymbol{\xi} \in \mathbb{R}^N$ such that $\nabla_{\boldsymbol{x}} f(\boldsymbol{x}^*) = \boldsymbol{A}^\top \boldsymbol{\xi}$ or, equivalently by setting $\boldsymbol{\lambda}^* = -\boldsymbol{\xi}$,

$$\exists \boldsymbol{\lambda}^* \in \mathbb{R}^N \text{ such that} \nabla_{\boldsymbol{x}} f(\boldsymbol{x}^*) + \boldsymbol{A}^\top \boldsymbol{\lambda}^* = \boldsymbol{0}. \tag{2.8}$$

[Interpretation]Recalling that $\boldsymbol{A} \triangleq [\boldsymbol{a}_1, \ldots, \boldsymbol{a}_N]^\top$ reveals that, at any local optimum $\boldsymbol{x}^*$, there exists some $\boldsymbol{\xi} \triangleq [\xi_1, \ldots, \xi_N]^\top$ such that

$$\nabla_{\boldsymbol{x}} f(\boldsymbol{x}^*) = \boldsymbol{A}^\top \boldsymbol{\xi} = \sum_{n=1}^{N} \xi_n \boldsymbol{a}_n.$$

In words: the gradient is a linear combination of the *gradients* of the equality constraint functions in (2.5). This leads to the following question:

Question 2.1 Must the gradient of the objective at an optimal point be a linear combination of the gradients of the equality constraint functions when the latter functions are non-linear?

Non-linear constraints

[Intuition] Fig. 2.1 provides some insight into Question 2.1 for the case of a single non-linear constraint. The key observation is that a point $\boldsymbol{x}$ cannot be optimal unless $\nabla f(\boldsymbol{x})$ is perpendicular to the curve $\mathcal{S} \triangleq \{\boldsymbol{x} : h_1(\boldsymbol{x}) = 0\}$, since in that case the objective value would decrease by moving on $\mathcal{S}$ in the direction that makes the greatest angle with the gradient. Since $\nabla h_1(\boldsymbol{x})$ is perpendicular to $\mathcal{S}$ for each $\boldsymbol{x}$, the statement "$\nabla f(\boldsymbol{x})$ is perpendicular to $\mathcal{S}$" is equivalent to "$\nabla f(\boldsymbol{x})$ is parallel to $\nabla h_1(\boldsymbol{x})$", which in turn is equivalent to "there exists λ_1 such that $\nabla f(\boldsymbol{x}) = -\lambda_1 \nabla h_1(\boldsymbol{x})$".

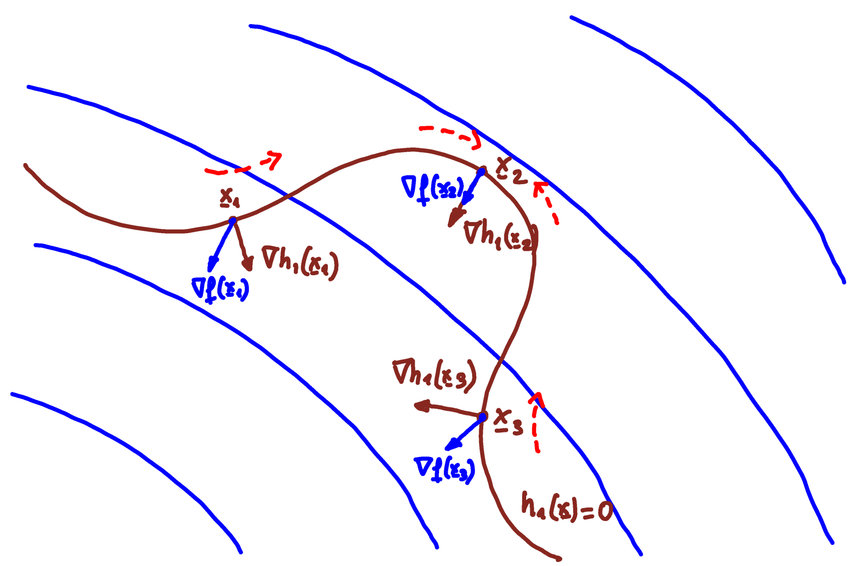


Figure 2.1: Blue curves are contour lines of f. Blue arrows represent the gradient of f at three locations $\boldsymbol{x}_1$, $\boldsymbol{x}_2$, and $\boldsymbol{x}_3$. The brown curve denotes the set of points $\boldsymbol{x}$ such that $h_1(\boldsymbol{x}) = 0$, i.e., the feasible set. When $\nabla f(\boldsymbol{x})$ is not perpendicular to the brown curve, one can move from $\boldsymbol{x}$ to another point on the curve and obtain a lower value of f. For example, in $\boldsymbol{x}_1$, the direction of the gradient indicates that the objective increases if one moves opposite to the red arrow. The same observation applies to $\boldsymbol{x}_3$. However, since $\nabla f(\boldsymbol{x}_2)$ is perpendicular to the brown curve at $\boldsymbol{x}_2$, the objective will not decrease by moving away from $\boldsymbol{x}_2$ along this curve. Thus, $\boldsymbol{x}_2$ is the minimizer.

[Toy example] One may corroborate the above intuition by considering the following example.

Example 2.2 In this example $f(\boldsymbol{x}) = x_1 + x_2$ and $h_1(\boldsymbol{x}) \triangleq h(\boldsymbol{x}) = x_1^2 + x_2^2 - 1$:

$$\begin{aligned} \underset{\boldsymbol{x}\in\mathbb{R}^2}{\text{minimize}} \quad & x_1 + x_2 \\ \text{subject to} \quad & x_1^2 + x_2^2 = 1 \end{aligned}$$

The minimum $\boldsymbol{x}^* = 2^{-1/2}[-1,-1]^\top$ is obtained by inspection in Fig. 2.2. It is observed that $\nabla f(\boldsymbol{x}^*) = [1,1]^\top$ and $\nabla h(\boldsymbol{x}^*) = \sqrt{2}[-1,-1]^\top$ are co-linear and therefore linearly dependent.

Thus, Example 2.2 shows that the answer to Question 2.1 is affirmative at least in some cases.

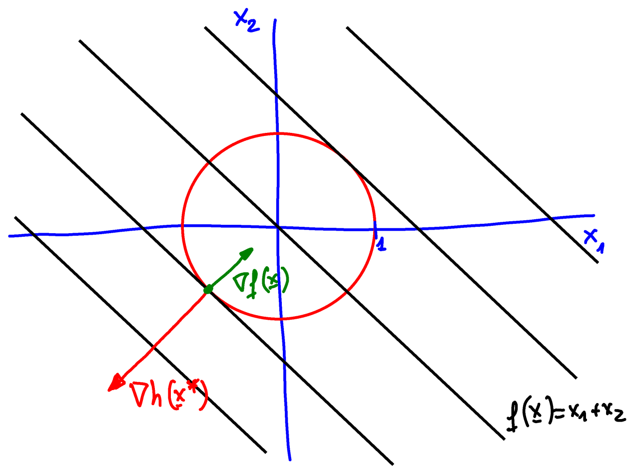


Figure 2.2: The gradient vectors of the objective and constraints are linearly dependent in Example 2.2.

[Necessary conditions]Indeed, the answer to Question 2.1 is affirmative in a broad range of cases. Before stating in which cases this occurs, the following definition is required:

Definition 2.1 *A point $\boldsymbol{x}_0 \in \mathbb{R}^D$ is regular if the vectors $\nabla h_1(\boldsymbol{x}_0), \ldots, \nabla h_N(\boldsymbol{x}_0)$ are linearly independent.*

Quiz 2.5 Are all $\boldsymbol{x} \in \mathbb{R}^D$ regular for the problem in (2.6)?

Quiz 2.6 Can there be any regular point for (2.4) if $N > D$?

Quiz 2.7 Consider (2.4) and suppose that the function $h_1(\boldsymbol{x})$ is continuously differentiable and attains its unconstrained minimum for $\boldsymbol{x} = \boldsymbol{x}_0$. Can $\boldsymbol{x}_0$ be a regular point of the problem (2.4)?

Proposition 2.8 *Consider the problem in (2.4).*
If

- $f(\boldsymbol{x})$ *and* $\{h_n(\boldsymbol{x})\}_{n=1}^N$ *are continuously differentiable.*
- $\boldsymbol{x}^*$ *is regular.*
- $\boldsymbol{x}^*$ *is a local optimum.*

Then,

$$\exists \boldsymbol{\lambda}^* \triangleq [\lambda_1^*, \ldots, \lambda_N^*]^\top \in \mathbb{R}^N \ : \ \nabla f(\boldsymbol{x}^*) + \sum_{n=1}^N \lambda_n^* \nabla h_n(\boldsymbol{x}^*) = \mathbf{0}. \tag{2.9}$$

Proof: See [bertsekas1999, Sec. 3.1].

□

[Lagrange multipliers] The scalars $\lambda_1^*, \ldots, \lambda_N^*$ are referred to as *Lagrange multipliers.*
[Non-existence example]It turns out that when the point is not regular, then the answer to Question 2.1 may be negative or, in other words, there may exist no Lagrange multipliers.

Example 2.3 In this example $f(\boldsymbol{x}) = x_1 + x_2$, $h_1(\boldsymbol{x}) = x_1^2 + x_2^2 - 1$, and $h_2(\boldsymbol{x}) = x_1 - 1$:

$$\begin{aligned} \underset{\boldsymbol{x}\in\mathbb{R}^2}{\text{minimize}} \quad & x_1 + x_2 \\ \text{subject to} \quad & x_1^2 + x_2^2 = 1 \\ & x_1 = 1. \end{aligned}$$

In this degenerate scenario, the minimum is at $\boldsymbol{x}^* = [1, 0]^\top$ since this is the only feasible point. There exist no Lagrange multipliers since $\nabla f(\boldsymbol{x}^*)$ is linearly independent of $\nabla h_1(\boldsymbol{x}^*)$ and $\nabla h_2(\boldsymbol{x}^*)$; see Fig. 2.3.

However, it is also possible that Lagrange multipliers exist for points that are not regular.

Quiz 2.8 Provide an example of optimization problem where the optimum satisfies (2.9) despite being not regular:

[Recap]To sum up,

- When a local optimum is regular, Lagrange multipliers exist.
- When a local optimum is not regular, Lagrange multipliers may not exist.

[Lagrangian]

- [def] A notational trick to simplify the statement of some results such as Proposition 2.8 is to define the *Lagrangian* function of the problem in (2.4):

$$\mathcal{L}(\boldsymbol{x}; \boldsymbol{\lambda}) \triangleq f(\boldsymbol{x}) + \sum_{n=1}^N \lambda_n h_n(\boldsymbol{x}) = f(\boldsymbol{x}) + \boldsymbol{h}^\top(\boldsymbol{x})\boldsymbol{\lambda} \tag{2.10}$$

where $\boldsymbol{\lambda} \triangleq [\lambda_1, \ldots, \lambda_N]^\top$ and $\boldsymbol{h}(\boldsymbol{x}) \triangleq [h_1(\boldsymbol{x}), \ldots, h_N(\boldsymbol{x})]^\top$. Thus, Proposition 2.8 equivalently states that, if $\boldsymbol{x}^*$ is regular and a local optimum, then

$$\exists \boldsymbol{\lambda}^* \ : \ \mathbf{0} = \nabla_{\boldsymbol{x}} \mathcal{L}(\boldsymbol{x}^*; \boldsymbol{\lambda}^*) = \nabla_{\boldsymbol{x}} f(\boldsymbol{x}^*) + \nabla_{\boldsymbol{x}} \boldsymbol{h}(\boldsymbol{x}^*)\boldsymbol{\lambda}^*$$

(recall the definition of gradient matrix in (A.5)). Obviously, $\boldsymbol{x}^*$ also satisfies $\nabla_{\boldsymbol{\lambda}} \mathcal{L}(\boldsymbol{x}^*; \boldsymbol{\lambda}) = \boldsymbol{h}(\boldsymbol{x}^*) = \mathbf{0}$ since it is feasible.

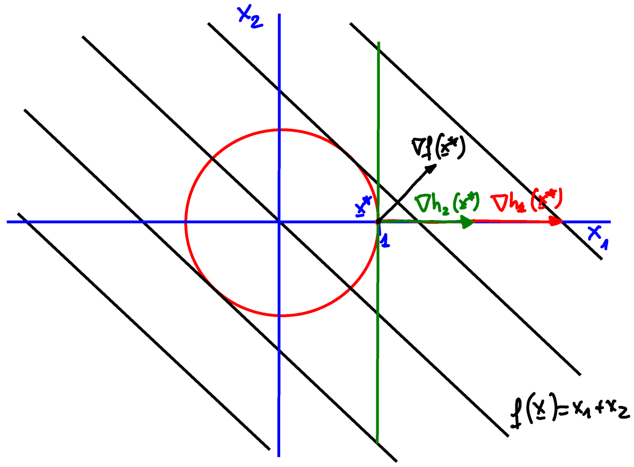


Figure 2.3: In Example 2.3 there exist no Lagrange multipliers.

- [Interpretation → Known multipliers] Suppose for simplicity that there exists only one vector $\boldsymbol{\lambda}^*$ for which the system of (generally non-linear) equations

$$\nabla f(\boldsymbol{x}) + \sum_{n=1}^{N} \lambda_n^* \nabla h_n(\boldsymbol{x}) = \mathbf{0} \tag{2.11}$$

has at least one solution in $\boldsymbol{x}$, and suppose that a *genie* has given us such a $\boldsymbol{\lambda}^*$. Now consider the unconstrained problem

$$\underset{\boldsymbol{x}\in\mathbb{R}^D}{\text{minimize}} \quad \mathcal{L}(\boldsymbol{x};\boldsymbol{\lambda}^*). \tag{2.12}$$

To some extent, the unconstrained problem (2.12) "behaves" in the same way as the constrained problem (2.4), as seen next.

From Proposition 2.4, it follows that the candidate solutions of (2.12) are precisely those vectors $\boldsymbol{x}$ satisfying (2.11). Let $\boldsymbol{x}_0$ be one of these vectors. From Proposition 2.8, if $\boldsymbol{x}_0$ is regular and feasible, then it will be a candidate solution of (2.4).

Thus, to find the minimizer of (2.4) with the theory developed in this section and under the aforementioned genie assumption, one would need to consider the following candidates

  - The candidate minimizers of (2.12) that are regular and feasible for (2.4).
  - All points that are feasible but not regular.

[Unknown multipliers] Without a genie, finding suitable Lagrange multipliers is not necessarily easy. Sometimes they can be found by solving the system of (possibly non-linear) equations $\nabla_{\boldsymbol{x},\boldsymbol{\lambda}}\mathcal{L}(\boldsymbol{x};\boldsymbol{\lambda}) = \mathbf{0}$. A different approach is by solving another optimization problem, as discussed in Sec. 2.3.2.

[Examples]

- [Quadratic problem with linear constraints]

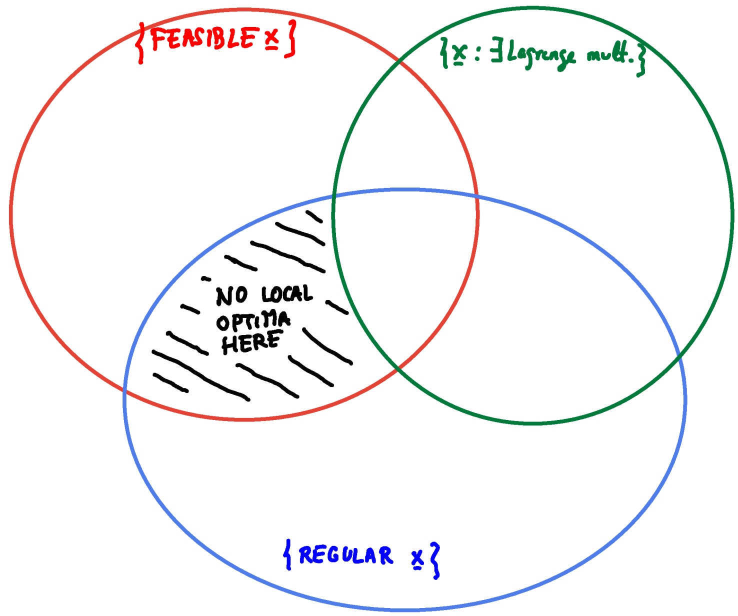


Figure 2.4: Proposition 2.8 excludes points that are feasible and regular for which there exist no Lagrange multipliers.

Quiz 2.9 Consider the problem

$$\begin{aligned} \underset{\boldsymbol{x}\in\mathbb{R}^D}{\text{minimize}} \quad & \boldsymbol{x}^\top \boldsymbol{Q}\boldsymbol{x} \\ \text{subject to} \quad & \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}. \end{aligned}$$

where $\boldsymbol{Q}$ is positive definite. Show that the minimizer of this problem is given by

$$\boldsymbol{x}^* = \boldsymbol{Q}^{-1}\boldsymbol{A}^\top(\boldsymbol{A}\boldsymbol{Q}^{-1}\boldsymbol{A}^\top)^{-1}\boldsymbol{b}.$$

Note that $\boldsymbol{A}\boldsymbol{Q}^{-1}\boldsymbol{A}^\top$ is always invertible. Why?

- [Quadratic function on the unit circle]

Quiz 2.10 Consider an arbitrary symmetric matrix $\boldsymbol{Q}$ and define the function $f : \mathcal{C} \to \mathbb{R}$, where $f(\boldsymbol{x}) \triangleq \boldsymbol{x}^\top \boldsymbol{Q}\boldsymbol{x}$ and $\mathcal{C}$ is the unit circle $\mathcal{C} \triangleq \{\boldsymbol{x} : ||\boldsymbol{x}||_2 = 1\}$. Obtain the maximum and the minimum of f. Prove your claims.

Second-order Necessary Conditions

[Feasible variations]Before introducing these conditions, define

$$\mathcal{V}(\boldsymbol{x}) \triangleq \{\boldsymbol{d} :\ \boldsymbol{d}^\top \nabla h_n(\boldsymbol{x}) = 0\ \forall n\} \tag{2.13}$$

as the subspace of first-order *feasible* variations at point $\boldsymbol{x}$. Two alternative ways (yet the same in essence) of thinking of this set

- [Small variations]Informally, if $\boldsymbol{x}$ is feasible and $\boldsymbol{d} \in \mathcal{V}(\boldsymbol{x})$ is such that $||\boldsymbol{d}||$ is small, then $h_n(\boldsymbol{x} + \boldsymbol{d}) \approx h_n(\boldsymbol{x}) + \nabla h_n^\top(\boldsymbol{x})\ \boldsymbol{d} = 0$, which means that $\boldsymbol{x} + \boldsymbol{d}$ is "approximately" feasible.

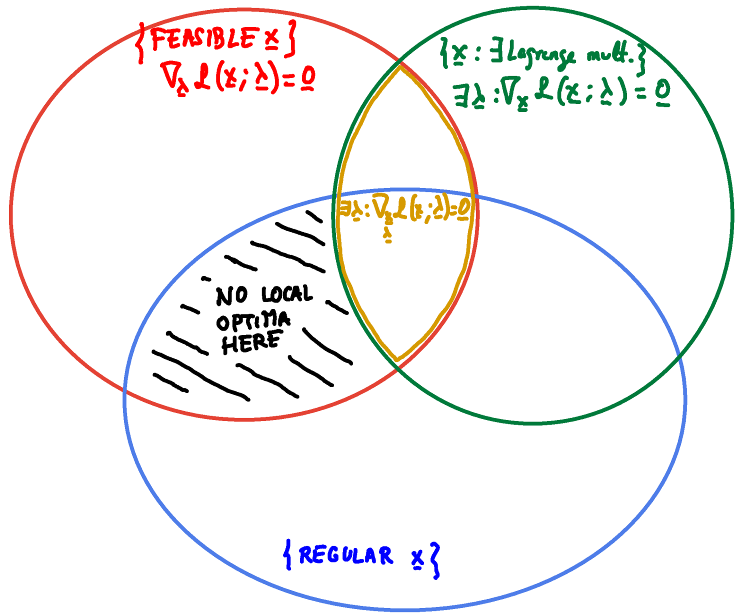


Figure 2.5: To find optimal points, one needs to check all feasible irregular points as well as all feasible regular $\boldsymbol{x}$ that satisfy $\nabla_{\boldsymbol{x},\boldsymbol{\lambda}}\mathcal{L}(\boldsymbol{x};\boldsymbol{\lambda}) = \mathbf{0}$ for some $\boldsymbol{\lambda}$. Thus, Proposition 2.8 narrows down the search for regular optima.

- [Taylor]If all constraints are replaced with their first-order Taylor approximations at a feasible $\boldsymbol{x}$, then $\boldsymbol{x} + \boldsymbol{d}$ is also feasible.

[Conditions]The following result provides a stronger conclusion than Proposition 2.8 when a stronger hypothesis holds: the objective and constraint functions are *twice* continuously differentiable:

Proposition 2.9 *Consider the problem in* (2.4).
If

- $f(\boldsymbol{x})$ *and* $\{h_n(\boldsymbol{x})\}_{n=1}^N$ *are* twice *continuously differentiable.*
- $\boldsymbol{x}^*$ *is regular.*
- $\boldsymbol{x}^*$ *is a local minimum.*

Then, $\exists \boldsymbol{\lambda}^* \triangleq [\lambda_1^*, \ldots, \lambda_N^*]^\top \in \mathbb{R}^N$ *such that*

$$\nabla f(\boldsymbol{x}^*) + \sum_{n=1}^N \lambda_n^* \nabla h_n(\boldsymbol{x}^*) = \mathbf{0} \tag{2.14a}$$

and

$$\boldsymbol{d}^\top \left[\nabla^2 f(\boldsymbol{x}^*) + \sum_{n=1}^N \lambda_n^* \nabla^2 h_n(\boldsymbol{x}^*)\right] \boldsymbol{d} \geq 0 \quad \forall \boldsymbol{d} \in \mathcal{V}(\boldsymbol{x}^*). \tag{2.14b}$$

Quiz 2.11 Does Proposition 2.9 generalize a result seen previously for unconstrained problems?

Quiz 2.12 Express (2.14) in terms of the Lagrangian.

Sufficient Conditions

[Second-order sufficient condition]

Proposition 2.10 *Consider the problem in (2.4).*
If

- $f(\boldsymbol{x})$ *and* $\{h_n(\boldsymbol{x})\}_{n=1}^N$ *are* twice *continuously differentiable.*
- *There exist* $\boldsymbol{x}^*$ *and* $\boldsymbol{\lambda}^*$ *satisfying*
  - *[feasibility]* $h_n(\boldsymbol{x}^*) = 0$, $\forall n$, *and*
  - *[first-order]* condition (2.14a), *and*
  - *[second-order]* condition

$$\boldsymbol{d}^\top \left[\nabla^2 f(\boldsymbol{x}^*) + \sum_{n=1}^N \lambda_n^* \nabla^2 h_n(\boldsymbol{x}^*) \right] \boldsymbol{d} > 0 \quad \forall \boldsymbol{d} \in \mathcal{V}(\boldsymbol{x}^*) \backslash \{\boldsymbol{0}\}.$$

Then,

- *[minimum]* $\boldsymbol{x}^*$ *is a strict local minimum*
- *[local bound]* $\gamma > 0$, $\epsilon > 0$ *such that*

$$f(\boldsymbol{x}) \geq f(\boldsymbol{x}^*) + \frac{\gamma}{2}\|\boldsymbol{x} - \boldsymbol{x}^*\|^2 \quad \forall \boldsymbol{x} : \begin{cases} h_n(\boldsymbol{x}) = 0, \ \forall n \\ \|\boldsymbol{x} - \boldsymbol{x}^*\| < \epsilon \end{cases} \tag{2.15}$$

Proof: This is [bertsekas1999, Prop. 3.2.1]. See proof therein.

□

Quiz 2.13 Does Proposition 2.10 generalize a result seen previously for unconstrained problems?

Quiz 2.14 Obtain the closed-form solution of

$$\begin{aligned} \underset{\boldsymbol{x}}{\text{minimize}} \quad & -(x_1x_2 + x_2x_3 + x_1x_3) \\ \text{subject to} \quad & x_1 + x_2 + x_3 = 3. \end{aligned}$$

This can be thought of as the problem of finding the parallelepiped with largest area given the sum of the edge lengths.

Sensitivity Analysis

[Overview] This section presents an informal interpretation of Lagrange multipliers. Specifically, their magnitudes indicate how much the optimal value changes upon loosening or tightening the constraints.
[Example] Consider a problem where each value of $\boldsymbol{x}$ corresponds to a way of allocating the resources of a company and the constraint $h_n(\boldsymbol{x}) = 0$ corresponds to the requirement that a certain expense $e(\boldsymbol{x})$ must be equal to a target amount, say e_0, which can be imposed by setting $h_n(\boldsymbol{x}) = e(\boldsymbol{x}) - e_0$. In this scenario, knowing the impact of e_0 on f^* may play a critical role in selecting e_0.
[Solution to the perturbed problem]

- [Problem] Given $\boldsymbol{u} \triangleq [u_1, \ldots, u_N]^\top$, consider the problem

$$\begin{aligned} \underset{\boldsymbol{x}\in\mathbb{R}^D}{\text{minimize}} \quad & f(\boldsymbol{x}) \\ \text{subject to} \quad & h_n(\boldsymbol{x}) = u_n, \; n = 1, \ldots, N. \end{aligned} \tag{2.16}$$

- [Solution] Under certain conditions, there exists an open ball $\mathcal{S}$ centered at $\mathbf{0}$ such that for each $\boldsymbol{u} \in \mathcal{S}$ the problem (2.16) has a solution $(\boldsymbol{x}^*(\boldsymbol{u}), \boldsymbol{\lambda}^*(\boldsymbol{u}))$.
- [Opt. solution]Let $p(\boldsymbol{u}) \triangleq f(\boldsymbol{x}^*(\boldsymbol{u}))$.
- [Sensitivity]Then, it can be shown that

$$\boxed{\nabla_{\boldsymbol{u}} p(\boldsymbol{u}) = -\boldsymbol{\lambda}^*(\boldsymbol{u}).} \tag{2.17}$$

- [Interpretation]

Quiz 2.15 What happens with the optimal value of (2.16) when $|\lambda_n^*|$ is large for some n? Which information does the sign of λ_n^* carry?

Quiz 2.16 Suppose that $f(\boldsymbol{x})$ has units of money and $h_1(\boldsymbol{x}) \triangleq i(\boldsymbol{x}) - t$, where $i(\boldsymbol{x})$ equals the number of items of some resource and t its target value. For instance, $i(\boldsymbol{x})$ can be the number of cars to be purchased, the number of employees to be hired, or number of transistors to be integrated in a circuit. How can λ_1 be interpreted? What are its units?

[Sensitivity theorem] A rigorous statement of this idea can be found in [bertsekas1999, Prop. 3.2.2].

2.3.1.2 Problems with Inequality Constraints

[General form]An inequality-constrained optimization problem has the form:

$$\underset{\boldsymbol{x}\in\mathbb{R}^D}{\text{minimize}} \quad f(\boldsymbol{x}) \tag{2.18a}$$
$$\text{subject to} \quad h_n(\boldsymbol{x}) = 0, \; n = 1, \ldots, N \tag{2.18b}$$
$$g_m(\boldsymbol{x}) \leq 0, \; m = 1, \ldots, M. \tag{2.18c}$$

More compactly, (2.18) can be expressed as

$$\underset{\boldsymbol{x}\in\mathbb{R}^D}{\text{minimize}} \quad f(\boldsymbol{x}) \tag{2.19a}$$
$$\text{subject to} \quad \boldsymbol{h}(\boldsymbol{x}) = \mathbf{0} \tag{2.19b}$$
$$\boldsymbol{g}(\boldsymbol{x}) \leq \mathbf{0}, \tag{2.19c}$$

where $\boldsymbol{g}(\boldsymbol{x}) \triangleq [g_1(\boldsymbol{x}), \ldots, g_M(\boldsymbol{x})]^\top$ and the last inequality must be understood entry-wise.
[Equality constraints vs. inequality constraints] There are certain differences between how we handle equality and inequality constraints. Before delving into this issue, note the following:

- [Active vs. inactive] The inequality constraint $g_m(\boldsymbol{x}) \leq 0$ is said to be *active* at point $\boldsymbol{x}$ if $g_m(\boldsymbol{x}) = 0$. Otherwise, it is said to be inactive.

Quiz 2.17 Suppose that the m-th inequality constraint is inactive at a *local* optimum $\boldsymbol{x}^*$. If this constraint is removed, is $\boldsymbol{x}^*$ still a local optimum?

Quiz 2.18 Suppose that the m-th inequality constraint is inactive at a *global* optimum $\boldsymbol{x}^*$. If this constraint is removed, is $\boldsymbol{x}^*$ still a global optimum?

Quiz 2.19 Suppose that the m-th inequality constraint is active at a *local* optimum $\boldsymbol{x}^*$. If the inequality in that constraint is replaced with an equality, is $\boldsymbol{x}^*$ still a local optimum?

- [Intuition complementary slackness] Let $\boldsymbol{x}^*$ be a local minimizer of (2.18) and assume for simplicity that the vectors $\{\nabla h_n(\boldsymbol{x}^*)\}_{n=1}^N \cup \{\nabla g_m(\boldsymbol{x}^*)\}_{m=1}^M$ are linearly independent. In view of the answers to Quizzes 2.17 and 2.19, it follows from Proposition 2.8 that there exist Lagrange multipliers $\{\lambda_n^*\}$ and $\{\mu_m^*\}$ such that

$$\nabla f(\boldsymbol{x}^*) + \sum_{n=1}^{N} \lambda_n^* \nabla h_n(\boldsymbol{x}^*) + \sum_{m} \mu_m^* \nabla g_m(\boldsymbol{x}^*) = \boldsymbol{0}, \tag{2.20}$$

where the second sum spans only over the active inequality constraints. Equivalently, it may span over all inequality constraints if the multipliers associated with inactive constraints are set to zero. If this is the case, one has that $\mu_m^* g_m(\boldsymbol{x}^*) = 0$ for all inactive constraints. But, clearly, this same equation also holds for all active inequality constraints. Therefore, one finds that $\mu_m^* g_m(\boldsymbol{x}^*) = 0$ for all inequality constraints at a local minimizer. This condition is called *complementary slackness.*[2]

- [Sign of multipliers]Another difference regarding inequality and equality constraints is the sign of the associated Lagrange multiplier. Recall the following:

Proposition 2.11 *For an arbitrary function f and arbitrary sets $\mathcal{A}$ and $\mathcal{B}$, it holds that*

$$\mathcal{A} \subset \mathcal{B} \quad \Rightarrow \quad \inf_{\boldsymbol{x}\in\mathcal{B}} f(\boldsymbol{x}) \leq \inf_{\boldsymbol{x}\in\mathcal{A}} f(\boldsymbol{x}).$$

Quiz 2.20 Prove Proposition 2.11. You may first assume that both infima are attained. Next you may try to come up with a proof without such an assumption. Recall the definition of infimum.

Quiz 2.21 Can a global minimum value of a problem increase when we remove one or several constraints?

Quiz 2.22 Suppose that the first inequality constraint in (2.18) is replaced with $g_1(\boldsymbol{x}) \leq u_1$, where u_1 is a constant. Let μ_1 denote the Lagrange multiplier associated with such a constraint.

1. What can we say about the solution of the resulting optimization problem for different values of u_1?
2. If this constraint is active at a certain local minimizer $\boldsymbol{x}$, what can we say about the sign of μ_1? To answer this question, you may recall (2.17) and your answer to Quiz 2.19.

These observations constitute valuable intuition but were stated in informal terms. The next section enables us to apply this knowledge in practice by formalizing the previous claims.

[2]This condition also holds for equality constraints, i.e. $\mu_n^* h_n(\boldsymbol{x}^*) = 0$, but it is never explicitly used since it is a consequence of $\boldsymbol{x}^*$ being feasible.

First-order Necessary Conditions

[Regular points]When there are inequality constraints, the notion of regular point is generalized as follows: $\boldsymbol{x}$ is regular if the vectors

$$\{\nabla h_n(\boldsymbol{x}),\ \forall n\} \cup \{\nabla g_m(\boldsymbol{x}),\ \forall m \text{ such that } g_m(\boldsymbol{x}) = 0\}$$

are linearly independent. Thus, the gradients of inactive inequality constraints are not relevant to determine whether $\boldsymbol{x}$ is regular.
[KKT]The following are the well-known Karush-Kuhn-Tucker (KKT) first-order necessary conditions.

Proposition 2.12 *Consider the problem in (2.18).*
If

- $f(\boldsymbol{x})$, $\{h_n(\boldsymbol{x})\}_{n=1}^N$, *and* $\{g_m(\boldsymbol{x})\}_{m=1}^M$ *are continuously differentiable.*
- $\boldsymbol{x}^*$ *is regular.*
- $\boldsymbol{x}^*$ *is a local minimum.*

Then, $\exists \boldsymbol{\lambda}^* \triangleq [\lambda_1^*, \ldots, \lambda_N^*]^\top \in \mathbb{R}^N$ *and* $\exists \boldsymbol{\mu}^* \triangleq [\mu_1^*, \ldots, \mu_M^*]^\top \in \mathbb{R}^M$ *such that*

- *[Dual feasibility[a]]* $\mu_m^* \geq 0$, $\forall m = 1, \ldots, M$.
- *[Complementary slackness]* $\mu_m^*\ g_m(\boldsymbol{x}^*) = 0$, $\forall m = 1, \ldots, M$.
- *[Stationarity]*

$$\nabla f(\boldsymbol{x}^*) + \sum_{n=1}^{N} \lambda_n^* \nabla h_n(\boldsymbol{x}^*) + \sum_{m=1}^{M} \mu_m^* \nabla g_m(\boldsymbol{x}^*) = \mathbf{0}. \tag{2.21}$$

[a] The reason why this condition is termed *dual feasibility* will become clear in Sec. 2.3.2.

Proof: See proof of [bertsekas1999, Prop. 3.3.1].

□

[Lagrangian] Again, these conditions can be expressed more conveniently in terms of the Lagrangian, which for a problem of the form (2.18) is given by

$$\begin{aligned}\mathcal{L}(\boldsymbol{x}; \boldsymbol{\lambda}, \boldsymbol{\mu}) &\triangleq f(\boldsymbol{x}) + \sum_{n=1}^{N} \lambda_n h_n(\boldsymbol{x}) + \sum_{m=1}^{M} \mu_m g_m(\boldsymbol{x}) \\ &= f(\boldsymbol{x}) + \boldsymbol{\lambda}^\top \boldsymbol{h}(\boldsymbol{x}) + \boldsymbol{\mu}^\top \boldsymbol{g}(\boldsymbol{x}).\end{aligned}$$

Quiz 2.23 Express (2.21) in terms of the Lagrangian.

Quiz 2.24 Compare Proposition 2.12 and Proposition 2.8.

Quiz 2.25 Recall the genie interpretation of Proposition 2.8. Can you extend it to accommodate inequality constraints using Proposition 2.12?

Second-order Necessary Conditions

[Feasible variations]Before introducing these conditions, generalize the subspace of first-order feasible variations introduced in (2.13) to accommodate inequality constraints:

$$\begin{aligned}\mathcal{V}(\boldsymbol{x}) \triangleq &\{\boldsymbol{d} : \boldsymbol{d}^\top \nabla h_n(\boldsymbol{x}) = 0 \ \forall n\} \\ &\cap \{\boldsymbol{d} : \boldsymbol{d}^\top \nabla g_m(\boldsymbol{x}) = 0 \ \forall m \text{ such that } g_m(\boldsymbol{x}) = 0\}.\end{aligned}$$

In words, the first-order feasible directions must be orthogonal to the gradients of the equality constraint functions and to the gradients of the *active* inequality constraint functions.
[KKT]

Proposition 2.13 *Consider the problem in* (2.18).
If

- $f(\boldsymbol{x})$, $\{h_n(\boldsymbol{x})\}_{n=1}^N$, *and* $\{g_m(\boldsymbol{x})\}_{m=1}^M$ *are twice continuously differentiable.*
- $\boldsymbol{x}^*$ *is regular.*
- $\boldsymbol{x}^*$ *is a local minimum.*

Then, $\exists \boldsymbol{\lambda}^* \triangleq [\lambda_1^*, \ldots, \lambda_N^*]^\top \in \mathbb{R}^N$ *and* $\exists \boldsymbol{\mu}^* \triangleq [\mu_1^*, \ldots, \mu_M^*]^\top \in \mathbb{R}^M$ *such that*

- *[Dual feasibility]* $\mu_m^* \geq 0$, $\forall m = 1, \ldots, M$.
- *[Complementary slackness]* $\mu_m^* \ g_m(\boldsymbol{x}^*) = 0$, $\forall m = 1, \ldots, M$.
- *[Stationarity]*

$$\nabla_{\boldsymbol{x}} \mathcal{L}(\boldsymbol{x}^*; \boldsymbol{\lambda}^*, \boldsymbol{\mu}^*) = \boldsymbol{0}. \tag{2.22}$$

- *[Conditionally positive semi-definite Hessian]*

$$\boldsymbol{d}^\top \nabla_{\boldsymbol{x}}^2 \mathcal{L}(\boldsymbol{x}^*; \boldsymbol{\lambda}^*, \boldsymbol{\mu}^*) \boldsymbol{d} \geq 0 \quad \forall \boldsymbol{d} \in \mathcal{V}(\boldsymbol{x}^*). \tag{2.23}$$

Proof: See proof of [bertsekas1999, Prop. 3.3.1].

□

Quiz 2.26 Does (2.23) hold when the problem is convex? Use Definition 1.4.

Sufficient Conditions

[Second-order sufficient conditions]

Proposition 2.14 *Consider the problem in (2.18).*
If

- *[differentiability]* $f(\boldsymbol{x})$, $\{h_n(\boldsymbol{x})\}_{n=1}^{N}$, *and* $\{g_m(\boldsymbol{x})\}_{m=1}^{M}$ *are twice continuously differentiable.*
- *[Sufficient conditions] There exist* $\boldsymbol{x}^*, \boldsymbol{\lambda}^*, \boldsymbol{\mu}^*$ *such that*
  - *[Primal feasibility]* $\boldsymbol{h}(\boldsymbol{x}^*) = \mathbf{0}$, $\boldsymbol{g}(\boldsymbol{x}^*) \leq \mathbf{0}$.
  - *[Dual feasibility]* $\mu_m^* \geq 0$, $\forall m = 1, \dots, M$.
  - *[Complementary slackness]* $\mu_m^* \; g_m(\boldsymbol{x}^*) = 0$, $\forall m = 1, \dots, M$.
  - *[Stationarity]*

$$\nabla_{\boldsymbol{x}} \mathcal{L}(\boldsymbol{x}^*; \boldsymbol{\lambda}^*, \boldsymbol{\mu}^*) = \mathbf{0}. \tag{2.24}$$

  - *[Conditionally positive definite Hessian]*

$$\boldsymbol{d}^\top \nabla_{\boldsymbol{x}}^2 \mathcal{L}(\boldsymbol{x}^*; \boldsymbol{\lambda}^*, \boldsymbol{\mu}^*) \boldsymbol{d} > 0 \quad \forall \boldsymbol{d} \in \mathcal{V}(\boldsymbol{x}^*). \tag{2.25}$$

- *[Positive active multipliers]* $\mu_m^* > 0$ *for all* m *such that* $g_m(\boldsymbol{x}^*) = 0$.

Then, $\boldsymbol{x}^*$ *is a strict local minimum.*

Proof: See proof of [bertsekas1999, Prop. 3.3.2].

□

Sensitivity Analysis

[Overview] The informal sensitivity analysis from Sec. 2.3.1.1 easily generalizes to the case with inequality constraints.
[Explanation]

- [Problem] Given $\boldsymbol{u} \triangleq [u_1, \dots, u_N]^\top$ and $\boldsymbol{v} \triangleq [v_1, \dots, v_M]^\top$, consider the problem

$$\begin{aligned} \underset{\boldsymbol{x} \in \mathbb{R}^D}{\text{minimize}} \quad & f(\boldsymbol{x}) \\ \text{subject to} \quad & h_n(\boldsymbol{x}) = u_n, \; n = 1, \dots, N, \\ & g_m(\boldsymbol{x}) \leq v_m, \; m = 1, \dots, M. \end{aligned} \tag{2.26}$$

- [Solution] Under certain conditions, there exists an open ball $\mathcal{S}$ centered at $[\boldsymbol{u}^\top, \boldsymbol{v}^\top]^\top = \mathbf{0}$ such that for each $[\boldsymbol{u}^\top, \boldsymbol{v}^\top]^\top \in \mathcal{S}$ the problem (2.26) has a solution $(\boldsymbol{x}^*(\boldsymbol{u}, \boldsymbol{v}), \boldsymbol{\lambda}^*(\boldsymbol{u}, \boldsymbol{v}), \boldsymbol{\mu}^*(\boldsymbol{u}, \boldsymbol{v}))$.
- [Opt. solution]Let $p(\boldsymbol{u}, \boldsymbol{v}) \triangleq f(\boldsymbol{x}^*(\boldsymbol{u}, \boldsymbol{v}))$.
- [Sensitivity]Then, it can be shown that

$$\nabla_{\boldsymbol{u}} p(\boldsymbol{u}, \boldsymbol{v}) = -\boldsymbol{\lambda}^*(\boldsymbol{u}) \tag{2.27a}$$
$$\nabla_{\boldsymbol{v}} p(\boldsymbol{u}, \boldsymbol{v}) = -\boldsymbol{\mu}^*(\boldsymbol{v}) \tag{2.27b}$$

- [Interpretation]

Quiz 2.27 If the m-th inequality constraint is inactive, then one expects that the optimal value does not change with sufficiently small variations of v_m. Is this phenomenon captured by (2.27)?

[Sensitivity theorem]A rigorous statement of this analysis can be found in [bertsekas1999, Prop. 3.3.3].

2.3.1.3 Generic Convex Feasible Sets

[Overview] This section presents a further first-order condition of interest that inspired multiple optimization methods. It basically states that there is no feasible descent direction at a minimizer.
[necessary and sufficient conditions]

Proposition 2.15 *Let $\mathcal{X}$ be a non-empty convex set and let $f : \mathbb{R}^D \to \mathbb{R}$ be continuously differentiable*[a] *over $\mathcal{X}$.*

- *If $\boldsymbol{x}^*$ is a local minimum of f over $\mathcal{X}$, then*

$$\nabla^\top f(\boldsymbol{x}^*)(\boldsymbol{x} - \boldsymbol{x}^*) \geq 0 \quad \forall \boldsymbol{x} \in \mathcal{X}. \tag{2.28}$$

- *If f is convex and $\boldsymbol{x}^*$ satisfies (2.28), then $\boldsymbol{x}^*$ is a global minimum of f over $\mathcal{X}$.*

[a] The domain of f need not actually be $\mathbb{R}^D$; it suffices that it is an open set containing $\mathcal{X}$

Proof: See [bertsekas1999, Prop. 2.1.2].

□

[Interpretation] To interpret this result, just note that (2.28) essentially says that any *feasible* direction (i.e., a direction leading to a feasible point from $\boldsymbol{x}^*$), is an *increase* direction since it makes an angle smaller than 90 degrees with the gradient; more on this later in the course.
[stationary points]Generalizing the notion of stationarity defined in Sec. 2.2, a point $\boldsymbol{x}^*$ is said to be a *stationary* point of the constrained problem

$$\begin{aligned} \underset{\boldsymbol{x} \in \mathbb{R}^D}{\text{minimize}} \quad & f(\boldsymbol{x}) \\ \text{subject to} \quad & \boldsymbol{x} \in \mathcal{X} \end{aligned} \tag{2.29}$$

if it satisfies (2.28). See [bertsekas1999, p. 194].

2.3.2 Zeroth-order Conditions and Duality

[Overview] Recall our genie interpretations of Proposition 2.8 and Proposition 2.12. They essentially state that (at least part of) the candidate minimizers of a constrained optimization problem can be found as the candidate minimizers of the Lagrangian seen as a function of $\boldsymbol{x}$. This suggests the following question:

Question 2.2 Can we find the minimizers of a *constrained* problem as the *unconstrained* minimizers of the Lagrangian function.

In this section, we will see that sometimes this is indeed the case. And remarkably, no differentiability assumptions or derivatives will be required!
[Zeroth-order sufficient condition] The hypotheses of the following result can be thought of as a derivative-free version of the KKT conditions.

Proposition 2.16 *Consider the problem*

$$\underset{\boldsymbol{x}\in\mathbb{R}^D}{\text{minimize}} \quad f(\boldsymbol{x}) \tag{2.30a}$$

$$\text{subject to} \quad \boldsymbol{h}(\boldsymbol{x}) = \boldsymbol{0} \tag{2.30b}$$

$$\boldsymbol{g}(\boldsymbol{x}) \leq \boldsymbol{0}. \tag{2.30c}$$

If $\boldsymbol{x}^*,\ \boldsymbol{\lambda}^*,\ \boldsymbol{\mu}^*$ *are such that*

- *[Primal feasibility]* $\boldsymbol{h}(\boldsymbol{x}^*) = \boldsymbol{0},\ \boldsymbol{g}(\boldsymbol{x}^*) \leq \boldsymbol{0}$.
- *[Dual feasibility]* $\mu_m^* \geq 0,\ \forall m = 1, \ldots, M$.
- *[Complementary slackness]* $\mu_m^*\ g_m(\boldsymbol{x}^*) = 0,\ \forall m = 1, \ldots, M$.
- *[Lagrangian minimizer]*

$$\boldsymbol{x}^* \in \underset{\boldsymbol{x}\in\mathbb{R}^D}{\arg\min}\, \mathcal{L}(\boldsymbol{x}; \boldsymbol{\lambda}^*, \boldsymbol{\mu}^*). \tag{2.31}$$

Then, $\boldsymbol{x}^*$ *is a global minimum.*

Proof:

$$f(\boldsymbol{x}^*) = f(\boldsymbol{x}^*) + \boldsymbol{\lambda}^{*\top}\boldsymbol{h}(\boldsymbol{x}^*) + \boldsymbol{\mu}^{*\top}\boldsymbol{g}(\boldsymbol{x}^*) \tag{2.32a}$$

$$= \inf_{\boldsymbol{x}} [\ f(\boldsymbol{x}) + \boldsymbol{\lambda}^{*\top}\boldsymbol{h}(\boldsymbol{x}) + \boldsymbol{\mu}^{*\top}\boldsymbol{g}(\boldsymbol{x})\ ] \tag{2.32b}$$

$$\leq \inf_{\boldsymbol{h}(\boldsymbol{x})=\boldsymbol{0},\ \boldsymbol{g}(\boldsymbol{x})\leq\boldsymbol{0}} [\ f(\boldsymbol{x}) + \boldsymbol{\lambda}^{*\top}\boldsymbol{h}(\boldsymbol{x}) + \boldsymbol{\mu}^{*\top}\boldsymbol{g}(\boldsymbol{x})\ ] \tag{2.32c}$$

$$\leq \inf_{\boldsymbol{h}(\boldsymbol{x})=\boldsymbol{0},\ \boldsymbol{g}(\boldsymbol{x})\leq\boldsymbol{0}} f(\boldsymbol{x}). \tag{2.32d}$$

□

Quiz 2.28 Explain why each equality and inequality in (2.32) holds.

Quiz 2.29 The proof is complete since it can be easily shown that all inequalities in (2.32) hold with equality. How can this be shown?

[Existence?] Note that there may exist no values of $\boldsymbol{x}^*,\ \boldsymbol{\lambda}^*,\ \boldsymbol{\mu}^*$ satisfying the hypotheses of Proposition 2.16. Of course, this does not mean that there exists no minimizer.

[Finding $\boldsymbol{x}^*,\ \boldsymbol{\lambda}^*,\ \boldsymbol{\mu}^*$]

- [Naive application of Proposition 2.16] Let us try to devise a method to find a minimizer without the aid of the genie. To this end, Proposition 2.16 suggests the following approach:

Procedure 2.1

1. Choose (somehow) $\boldsymbol{\lambda}^*$, $\boldsymbol{\mu}^*$ such that $\boldsymbol{\mu}^* \geq \mathbf{0}$.
2. Obtain $\boldsymbol{x}^*$ from (2.31).
3. Check if $\boldsymbol{x}^*$ and $\boldsymbol{\mu}^*$ satisfy all of the following conditions

$$\boldsymbol{h}(\boldsymbol{x}^*) = \mathbf{0}, \ \boldsymbol{g}(\boldsymbol{x}^*) \leq \mathbf{0} \tag{2.33a}$$
$$\mu_m^* \, g_m(\boldsymbol{x}^*) = 0, \ \forall m = 1, \ldots, M \tag{2.33b}$$

4. If not, go back to 1 and choose a new pair $\boldsymbol{\lambda}^*$, $\boldsymbol{\mu}^*$.

- [Lower bound of f^*] Obviously, one cannot carry out this procedure by simple trial and error: a smarter approach for choosing $\boldsymbol{\lambda}^*$, $\boldsymbol{\mu}^*$ in step 1 is required. To this end, consider the following two observations:
  - If $\boldsymbol{x}^*$, $\boldsymbol{\lambda}^*$, $\boldsymbol{\mu}^*$ satisfy the hypotheses of Proposition 2.16, then
$$f(\boldsymbol{x}^*) = \inf_{\boldsymbol{h}(\boldsymbol{x})=\mathbf{0}, \ \boldsymbol{g}(\boldsymbol{x})\leq\mathbf{0}} f(\boldsymbol{x}) = \inf_{\boldsymbol{x}\in\mathbb{R}^D} \mathcal{L}(\boldsymbol{x}; \boldsymbol{\lambda}^*, \boldsymbol{\mu}^*). \tag{2.34}$$
  - For any $\boldsymbol{\lambda}$, $\boldsymbol{\mu}$ satisfying $\boldsymbol{\mu} \geq \mathbf{0}$,
$$\inf_{\boldsymbol{h}(\boldsymbol{x})=\mathbf{0}, \ \boldsymbol{g}(\boldsymbol{x})\leq\mathbf{0}} f(\boldsymbol{x}) \geq \inf_{\boldsymbol{x}\in\mathbb{R}^D} \mathcal{L}(\boldsymbol{x}; \boldsymbol{\lambda}, \boldsymbol{\mu}). \tag{2.35}$$

Quiz 2.30 Prove (2.35).

- [Dual function and problem] Clearly, from (2.34) and (2.35), it follows that if $\boldsymbol{x}^*$, $\boldsymbol{\lambda}^*$, $\boldsymbol{\mu}^*$ satisfy the hypotheses of Proposition 2.16, then $\boldsymbol{\lambda}^*$, $\boldsymbol{\mu}^*$ maximize
$$q(\boldsymbol{\lambda}, \boldsymbol{\mu}) \triangleq \inf_{\boldsymbol{x}\in\mathbb{R}^D} \mathcal{L}(\boldsymbol{x}; \boldsymbol{\lambda}, \boldsymbol{\mu}). \tag{2.36}$$
subject to $\boldsymbol{\mu} \geq \mathbf{0}$. The function in (2.36) is named *dual function*, and the aforementioned problem, that is
$$\underset{\boldsymbol{\lambda},\boldsymbol{\mu}}{\text{maximize}} \quad q(\boldsymbol{\lambda}, \boldsymbol{\mu}) \tag{2.37a}$$
$$\text{subject to} \quad \boldsymbol{\mu} \geq \mathbf{0}, \tag{2.37b}$$
is the *dual problem* of (2.30). Conversely, (2.30) is referred to as the *primal problem* of (2.37).

Quiz 2.31 Show that the dual problem is always a convex problem. In other words, you need to prove that the dual function is always concave, regardless of whether the primal problem is convex or not.

- [Refined procedure] Thus, the smart approach to address Step 1 in Procedure 2.1 is to solve (2.37). Steps 2 and 3 should be performed with the resulting maximizers.

[Duality] Expressions (2.34) and (2.35) can be rewritten in terms of q as follows:

- [Strong duality] If the hypotheses of Proposition 2.16 are satisfied for some ($\boldsymbol{x}^*$, $\boldsymbol{\lambda}^*$, $\boldsymbol{\mu}^*$), then
$$\inf_{\boldsymbol{h}(\boldsymbol{x})=\mathbf{0}, \ \boldsymbol{g}(\boldsymbol{x})\leq\mathbf{0}} f(\boldsymbol{x}) = q(\boldsymbol{\lambda}^*, \boldsymbol{\mu}^*), \tag{2.38}$$
and we say that *strong duality* holds. In other words, we say that a problem has the *strong duality property* when the global minimum of the primal equals the global maximum of the dual. There are a number of *constraint qualifications* that guarantee that strong duality holds; see e.g. [boyd, Ch. 5].

- [Weak duality] Clearly, for any $\boldsymbol{\lambda}$, $\boldsymbol{\mu}$ satisfying $\boldsymbol{\mu} \geq \mathbf{0}$,

$$\inf_{\boldsymbol{h}(\boldsymbol{x})=\mathbf{0},\ \boldsymbol{g}(\boldsymbol{x})\leq\mathbf{0}} f(\boldsymbol{x}) \geq q(\boldsymbol{\lambda}, \boldsymbol{\mu}). \tag{2.39}$$

  This condition is referred to as *weak duality* and always holds. Another interesting expression arises upon maximizing both sides of (2.39) with respect to the *dual variables* $\boldsymbol{\lambda}, \boldsymbol{\mu}$:

$$f^* \triangleq \inf_{\boldsymbol{h}(\boldsymbol{x})=\mathbf{0},\ \boldsymbol{g}(\boldsymbol{x})\leq\mathbf{0}} f(\boldsymbol{x}) \geq \sup_{\boldsymbol{\lambda},\boldsymbol{\mu}\geq\mathbf{0}} q(\boldsymbol{\lambda}, \boldsymbol{\mu}) \triangleq q^* \tag{2.40}$$

  where f^* and q^* are respectively referred to as the *primal optimal* and *dual optimal* values.

[Solving the dual] Solving the dual (2.37) may appear easier than solving the primal (2.30) since the constraints of (2.37) are simple. However, this is not necessarily the case since the dual objective function q is oftentimes complicated and part of its complexity is moved to the constraints; see Sec. 1.3.
[Key message] To sum up, when strong duality holds, the primal solution can be found by first solving the dual and then solving the unconstrained problem (2.31).

2.3.2.1 Zeroth-Order Perspective of the Lagrangian

[Overview] Observe that we have looked at the Lagrangian from two different perspectives:

- [First- and second-order] Before Sec. 2.3.2, we mainly used the Lagrangian to summarize first- and second-order conditions, i.e., equalities and inequalities involving derivatives. Recall that the intuition behind setting the gradient of the Lagrangian equal to zero was to "align" the gradients of the objective and those of the equality and active inequality constraints; cf. Fig. 2.1.

- [Zeroth-order] In Sec. 2.3.2, we did not consider derivatives: we just looked at the *minimizers* of the Lagrangian. It remains, therefore, to look at the intuition why minimizing the Lagrangian makes sense.

[Recap zeroth-order] To this end, let us start by reflecting on the main idea presented in the previous section, which can be summarized as follows:

1. To each pair $(\boldsymbol{\lambda}, \boldsymbol{\mu})$, one can associate the following unconstrained problem:

$$\underset{\boldsymbol{x}\in\mathbb{R}^D}{\text{minimize}} \quad \mathcal{L}(\boldsymbol{x}; \boldsymbol{\lambda}, \boldsymbol{\mu}). \tag{2.41}$$

2. A solution $\boldsymbol{x}_{\mathcal{L}}(\boldsymbol{\lambda}, \boldsymbol{\mu})$ of this problem may or may not coincide with the solution of

$$\underset{\boldsymbol{x}\in\mathbb{R}^D}{\text{minimize}} \quad f(\boldsymbol{x}) \tag{2.42a}$$
$$\text{subject to} \quad \boldsymbol{h}(\boldsymbol{x}) = \mathbf{0} \tag{2.42b}$$
$$\boldsymbol{g}(\boldsymbol{x}) \leq \mathbf{0} \tag{2.42c}$$

3. If

   (a) there is strong duality

   (b) AND $(\boldsymbol{\lambda}, \boldsymbol{\mu})$ are properly chosen[3],

   then

[3] Since there is strong duality, this must be possible so long as the maximum of the dual problem is attained. For simplicity, disregard the case where this does not occur.

(a) the hypotheses of Proposition 2.16 are satisfied for at least one minimizer $\boldsymbol{x}_{\mathcal{L}}(\boldsymbol{\lambda}, \boldsymbol{\mu})$ of (2.42) (there may be other minimizers that do not satisfy these hypotheses; cf. examples below).

4. If

(a) there is not strong duality (just weak duality),

then

(a) the hypotheses of Proposition 2.16 are never satisfied, regardless of $(\boldsymbol{\lambda}, \boldsymbol{\mu})$.

Thus, it is natural to wonder:

Question 2.3 What is the relation between the solutions of (2.41) and (2.42) when strong duality does not hold?

Question 2.4 What is the relation between the solutions of (2.41) and (2.42) when strong duality holds but $(\boldsymbol{\lambda}, \boldsymbol{\mu})$ have been improperly selected?

Question 2.5 How can we know if the multipliers $\boldsymbol{\lambda}, \boldsymbol{\mu}$ are too low or too high?

Question 2.6 Is it possible that the minimizer of (2.41) coincides with the minimizer of (2.42) when strong duality does NOT hold?

Let us answer these questions while we acquire some intuition by means of three carefully selected examples.

Example 2.4 Consider the problem

$$\underset{x\in\mathbb{R}}{\text{minimize}} \quad f(x) \tag{2.43a}$$

$$\text{subject to} \quad g(x) \le 0, \tag{2.43b}$$

where functions f and g are depicted on Fig. 2.6a. Note the following:

- The solution of (2.43) is clearly $x^* = -2$.
- Fig. 2.6b plots the Lagrangian $\mathcal{L}(x;\mu)$ as a function of x for five non-negative different values of μ. Some of these values are less than μ^* and others are greater, where $\mu^* \triangleq \arg\max_\mu q(\mu) \triangleq \arg\max_\mu \inf_{x\in\mathbb{R}} \mathcal{L}(x;\mu)$.
- Since $\mathcal{L}(x;\mu) = f(x) + \mu g(x)$ and $g(x)$ is a straight line that crosses 0 at $x = -1$, one can think of each curve of Fig. 2.6b as a version of $\mathcal{L}(x;\mu)$ "rotated" around $x = -1$. Note that
  - all curves meet when $x = -1$.
  - the greater μ, the greater the rotation angle.
- Crosses in Fig. 2.6b indicate the locations of the global minima of each function.
  - when $\mu < \mu^*$ (too low rotation angles), the minimum of the Lagrangian is on the right valley, which is in the infeasible region ($g(x) > 0$).
  - When $\mu > \mu^*$ (too high rotation angles), the minimum of the Lagrangian is on the left valley, specifically inside the strictly feasible region ($g(x) < 0$).
  - when $\mu = \mu^*$ ("correct" rotation angle), there is a feasible minimum and an infeasible minimum.
- By definition, μ^* is the value of μ for which the global minimum of $\mathcal{L}(x;\mu)$ with respect to x is largest. Observe that this is the case in Fig. 2.6b since there are no crosses above the red ones.
- Fig. 2.6c compares the solution x^* of (2.43) with the *feasible* minimizer of the Lagrangian when $\mu = \mu^*$, which can be denoted as $x_{\mathcal{L}}(\mu^*)$. Observe that
  - strong duality does not hold,
  - $x^* \neq x_{\mathcal{L}}(\mu^*)$

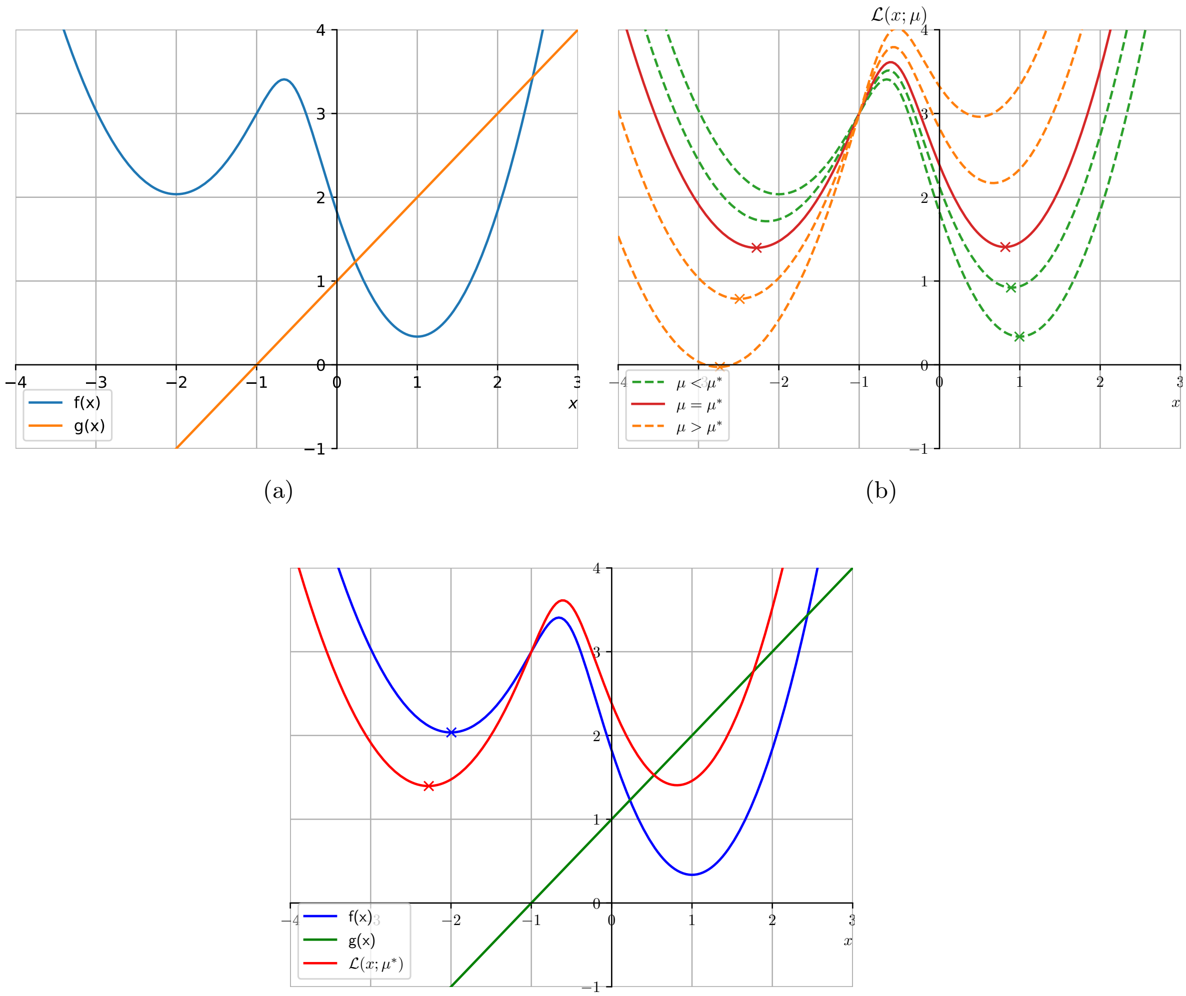
Quiz 2.32 On Fig. 2.6c, indicate the duality gap and $|x^* - x_{\mathcal{L}}(\mu^*)|$.

[interpretation] From this example, we can now extract our "zeroth-order intuition": since $\boldsymbol{\mu} \ge 0$, one can think of the second term in $\mathcal{L}(\boldsymbol{x};\boldsymbol{\mu}) = f(\boldsymbol{x}) + \mu g(\boldsymbol{x})$ as a term that penalizes infeasible points (since $g(\boldsymbol{x}) > 0$ for infeasible $\boldsymbol{x}$) and that "benefits" (strictly) feasible points (since $g(\boldsymbol{x}) \le 0$ for feasible $\boldsymbol{x}$). For a sufficiently large μ, this effect is (at least generally[4]) large enough to ensure that the minimizers of the Lagrangian (if they exist) are in the feasible set.

Quiz 2.33 Provide an example of a problem of the form (2.43) (you can simply plot f and g) where $x_{\mathcal{L}}(0)$ is infeasible and $x_{\mathcal{L}}(\mu)$ does not exist for $\mu > 0$.

[Same solution] In Example 2.4, there is no strong duality and we observed that the minimizer of the Lagrangian was different from the solution of the primal problem. One may wonder whether

[4] An example where this is not the case is $f(x) = -e^x$ and $h(x) = x$. In this case, one clearly has that $x^* = 0$, but the Lagrangian is never minimized in the feasible region. However, this is a somewhat degenerate problem, as the dual function is not defined. For simplicity, let us disregard this kind of problems.

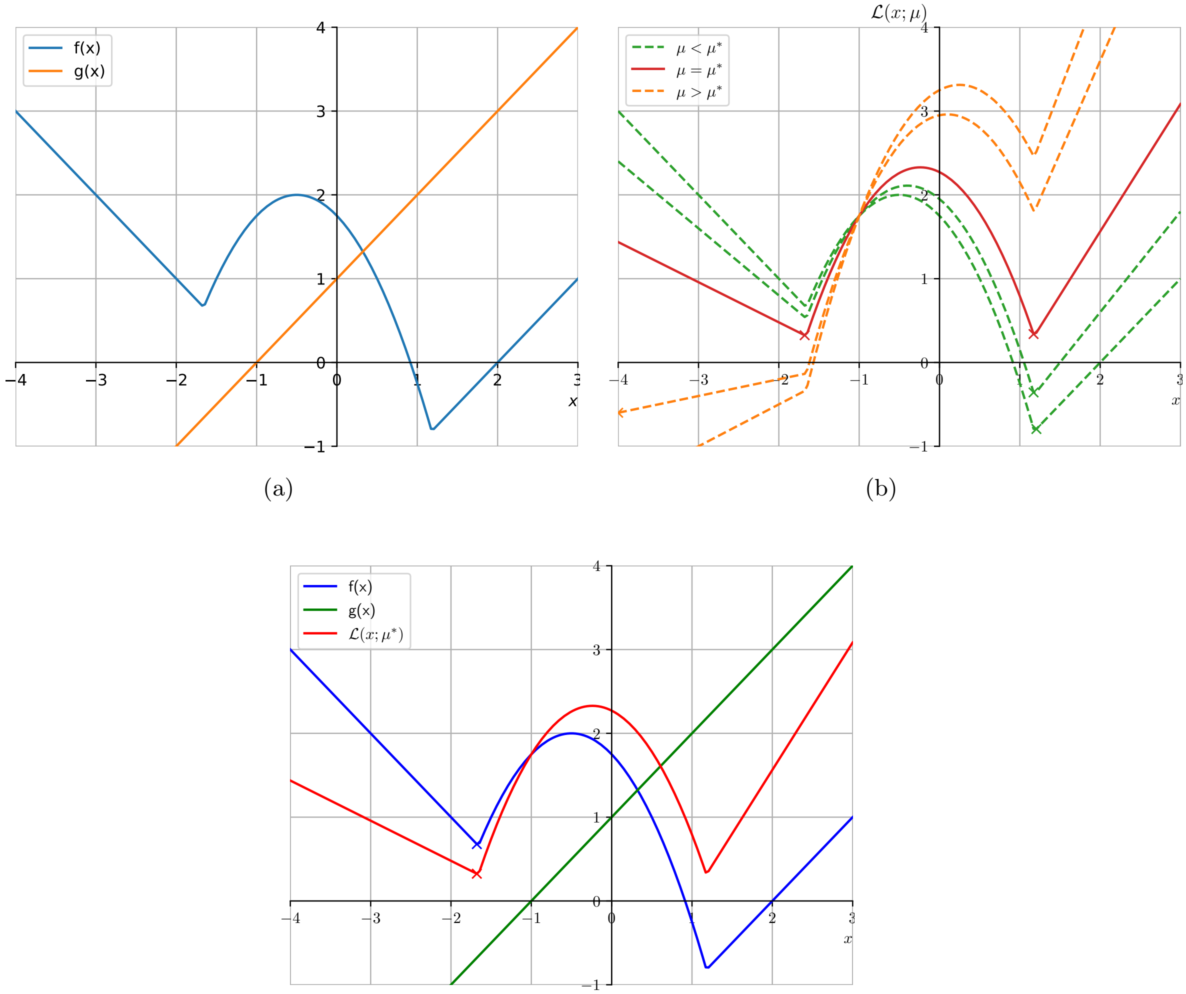


(c)

Figure 2.6: Graphical illustration of Example 2.4. (a) Objective and constraint functions. (b) Lagrangian as a function of x for different values of the Lagrange multiplier. (c) Lagrangian for the optimal Lagrange multiplier along with the objective and constraint functions.

this is necessarily the case or whether it is possible that, despite the fact that there is no strong duality, both (2.41) and (2.42) can have the same solution (Question 2.6). The following example answers this question:

Example 2.5 Consider the problem (2.43) where functions f and g are depicted on Fig. 2.7a. Clearly, from Fig. 2.7c, there is no strong duality. However, x^* and $x_{\mathcal{L}}(\mu^*)$ coincide.



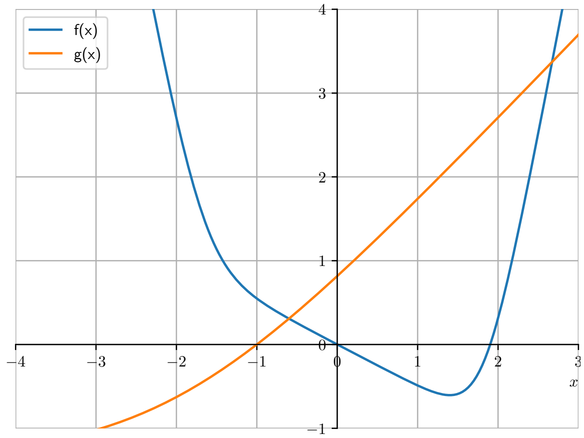
(a) (b) (c)

Figure 2.7: Graphical illustration of Example 2.5. (a) Objective and constraint functions. (b) Lagrangian as a function of x for different values of the Lagrange multiplier. (c) Lagrangian for the optimal Lagrange multiplier along with the objective and constraint functions.

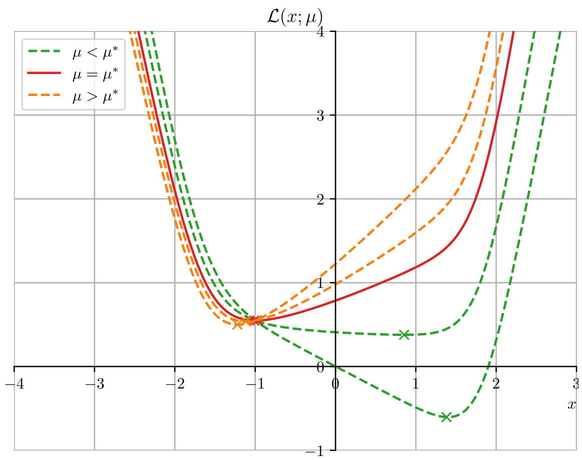
Quiz 2.34 Since there is no strong duality in Example 2.5, at least one of the hypotheses of Proposition 2.16 is violated for the pair $(x_{\mathcal{L}}(\mu^*), x^*)$. Which one is it?

[Not necessary] Thus, it follows that the hypotheses of Proposition 2.16 are sufficient but not necessary.
[Strong duality] Now let us look at a problem with strong duality.

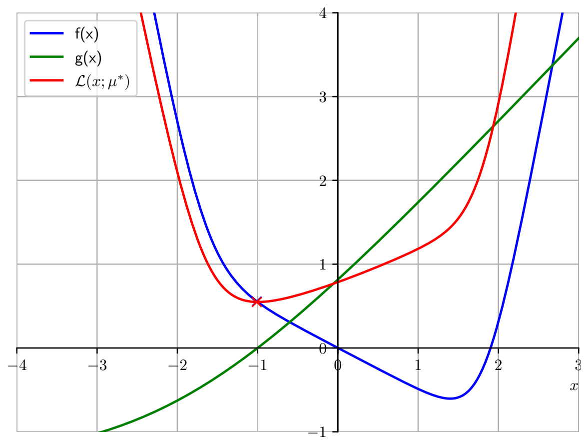
Example 2.6 Consider the problem (2.43) where functions f and g are depicted on Fig. 2.8a. Clearly, from Fig. 2.8c, there is strong duality. To illustrate that the observations of this section apply even if the constraints are non-linear, here g is non-linear, but the same conclusions apply when g is as in Examples 2.4 and 2.5.



(a)



(b)



(c)

Figure 2.8: Graphical illustration of Example 2.6. (a) Objective and constraint functions. (b) Lagrangian as a function of x for different values of the Lagrange multiplier. (c) Lagrangian for the optimal Lagrange multiplier along with the objective and constraint functions.

[key intuition] By carefully digesting these examples, one can answer the following question:

Quiz 2.35 Suppose that, in Example 2.6, we start from $\mu = 0$ and gradually increase μ. We observe that, at some point, which we can call the transition point, $x_{\mathcal{L}}(\mu)$ becomes feasible and primal optimal (MINIMAL). If we keep increasing μ beyond this point, then $x_{\mathcal{L}}(\mu)$ remains feasible but it becomes suboptimal. So far, this is expected because g acts as a penalty and the optimum is in the boundary of the feasible set. However, how can you explain that $\mathcal{L}(x_{\mathcal{L}}(\mu); \mu)$ as a function of μ becomes MAXIMIZED precisely at the same transition point?

[Finding multipliers] Recall that Sec. 2.3.2 introduced Procedure 2.1 to illustrate how having the *right* multipliers effectively turns a constrained problem into an unconstrained one. But, there, the way of choosing the multipliers in the first step was not specified. Now that we understand the intuition behind the choice of μ, we can think of an iterative approach to find the optimal multiplier of the problem

$$\underset{\boldsymbol{x}\in\mathbb{R}^D}{\text{minimize}} \quad f(\boldsymbol{x}) \tag{2.44a}$$

$$\text{subject to} \quad g(\boldsymbol{x}) \leq 0. \tag{2.44b}$$

To this end, let

$$\mathcal{M}_{\mathcal{L}}(\mu) \triangleq \underset{\boldsymbol{x}\in\mathbb{R}^D}{\arg\min}\, \mathcal{L}(\boldsymbol{x}; \mu) \tag{2.45}$$

be the set of unconstrained minimizers of the Lagrangian and consider the following (informal) iteration:

Procedure 2.2

1. Set $\mu = 0$.
2. Repeat
   (a) If there is $x_{\mathcal{L}} \in \mathcal{M}_{\mathcal{L}}(\mu)$ that satisfies the hypotheses of Proposition 2.16, return.
   (b) Elseif $g(x_{\mathcal{L}}) > 0\ \forall x_{\mathcal{L}} \in \mathcal{M}_{\mathcal{L}}(\mu)$, increase μ (we are penalizing infeasible points too little).
   (c) Elseif $g(x_{\mathcal{L}}) < 0\ \forall x_{\mathcal{L}} \in \mathcal{M}_{\mathcal{L}}(\mu)$, decrease[a] μ (we are promoting feasible points too much).

[a] Of course, without making it negative.

Quiz 2.36 Why there is no "Elseif" for the case where $g(x_{\mathcal{L}}) = 0$ for some $x_{\mathcal{L}} \in \mathcal{M}_{\mathcal{L}}(\mu)$?

Quiz 2.37 What if $\mu > 0$ and $\mathcal{M}_{\mathcal{L}}(\mu)$ contains both points with $g(x_{\mathcal{L}}) > 0$ and points with $g(x_{\mathcal{L}}) < 0$?

[Intuition] Procedure 2.2, which is presented merely for didactical purposes, provides the ultimate piece of intuition behind Lagrange duality by essentially establishing the correspondence between the dual problem and a search for the optimal weight of a penalty term.
[Equality constraints] Observe that, in this section, we focused on problems with a single inequality constraint. Problems with equality constraints can be interpreted in similar terms, yet the rules become more complicated.

2.3.2.2 Partial dualization

[Overview] Sometimes, one can "dualize" only some of the constraints if that simplifies the dual or if it results in strong duality. To this end, some of the constraints are absorbed generically as $\boldsymbol{x}^* \in \mathcal{Z}$ and treated *implicitly*. The following result extends Proposition 2.16 to this setup.

Proposition 2.17 *Consider the problem*

$$\begin{aligned} \underset{\boldsymbol{x}\in\mathbb{R}^D}{\text{minimize}} \quad & f(\boldsymbol{x}) && (2.46a)\\ \text{subject to} \quad & \boldsymbol{x} \in \mathcal{Z} && (2.46b)\\ & \boldsymbol{h}(\boldsymbol{x}) = \boldsymbol{0} && (2.46c)\\ & \boldsymbol{g}(\boldsymbol{x}) \leq \boldsymbol{0} && (2.46d) \end{aligned}$$

If $\boldsymbol{x}^*$, $\boldsymbol{\lambda}^*$, $\boldsymbol{\mu}^*$ *are such that*

- *[Primal feasibility]* $\boldsymbol{h}(\boldsymbol{x}^*) = \boldsymbol{0}$, $\boldsymbol{g}(\boldsymbol{x}^*) \leq \boldsymbol{0}$.
- *[Dual feasibility]* $\mu_m^* \geq 0$, $\forall m = 1, \ldots, M$.
- *[Complementary slackness]* $\mu_m^*\, g_m(\boldsymbol{x}^*) = 0$, $\forall m = 1, \ldots, M$.
- *[Lagrangian minimizer]*

$$\boldsymbol{x}^* \in \arg\min_{\boldsymbol{x}\in\mathcal{Z}} \mathcal{L}(\boldsymbol{x}; \boldsymbol{\lambda}^*, \boldsymbol{\mu}^*). \tag{2.47}$$

Then, $\boldsymbol{x}^*$ *is a global minimizer.*

Quiz 2.38 Prove Proposition 2.17.

[Splitting complexity] Clearly, applying partial dualization simplifies the dual problem but solving (2.47) becomes more difficult.

2.3.2.3 Geometric interpretation

[Overview] To understand the above theorems and duality, it is instructive to think in terms of the geometric interpretation of this section.
[Problem] Let us start by dualizing only a single inequality constraint; the rest of constraints become implicit by properly choosing $\mathcal{Z}$. The resulting problem reads as

$$\begin{aligned} \underset{\boldsymbol{x}\in\mathbb{R}^D}{\text{minimize}} \quad & f(\boldsymbol{x}) && (2.48a)\\ \text{subject to} \quad & \boldsymbol{x} \in \mathcal{Z} && (2.48b)\\ & g(\boldsymbol{x}) \leq 0. && (2.48c) \end{aligned}$$

Consider the following set

$$\mathcal{S} = \{(g(\boldsymbol{x}), f(\boldsymbol{x})) : \ \boldsymbol{x} \in \mathcal{Z}\} \subset \mathbb{R}^2. \tag{2.49}$$

We are therefore mapping each $\boldsymbol{x} \in \mathcal{Z}$ to a point in $\mathbb{R}^2$.

Now choose a value for the Lagrange multiplier $\mu_0 \geq 0$ and a point $\boldsymbol{x}_0 \in \mathcal{Z}$. For such a combination of values of the primal and dual variables, the Lagrangian takes the value $l_0 \triangleq \mathcal{L}(\boldsymbol{x}_0; \mu_0) = f(\boldsymbol{x}_0) + \mu_0 g(\boldsymbol{x}_0)$. But potentially many other points $\boldsymbol{x}$ may attain the same Lagrangian value, specifically, $\{\boldsymbol{x} \in \mathcal{Z} : \; f(\boldsymbol{x}) + \mu_0 g(\boldsymbol{x}) = l_0\}$. When applying the aforementioned mapping $\boldsymbol{x} \mapsto T(\boldsymbol{x}) = (g(\boldsymbol{x}), f(\boldsymbol{x}))$ to any of these points, the result must lie on the set

$$H_0 \triangleq \{(\gamma, \phi) : \; \phi + \mu_0 \gamma = l_0\}, \tag{2.50}$$

which is a straight line with slope $-\mu_0 \leq 0$ and intercept l_0. This is illustrated in Fig. 2.9.

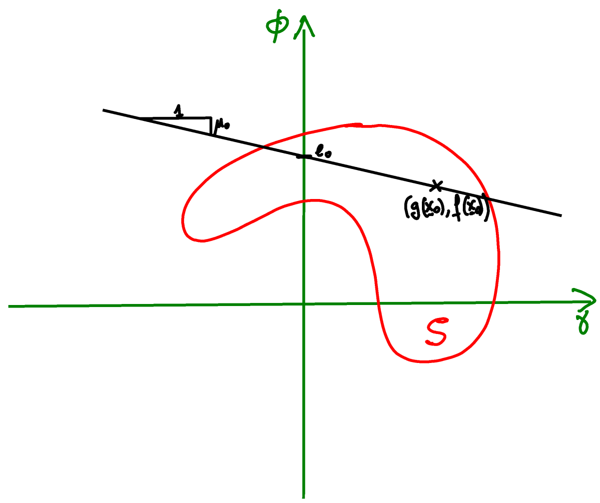


Figure 2.9: Illustration of $\mathcal{S}$.

Quiz 2.39 Locate the feasible set and the minimizer of (2.48) in Fig. 2.9.

Now, for the same fixed μ_0, choose the $\boldsymbol{x}_0 \in \mathcal{Z}$ that leads to the smallest l_0. Such an $\boldsymbol{x}_0$ will be denoted as $\boldsymbol{x}^*(\mu_0)$ since it depends on μ_0. Observe that the resulting intercept is now

$$l_0(\mu_0) \triangleq \mathcal{L}(\boldsymbol{x}^*(\mu_0); \mu_0) = \inf_{\boldsymbol{x} \in \mathcal{Z}} \mathcal{L}(\boldsymbol{x}; \mu_0) = q(\mu_0). \tag{2.51}$$

and the corresponding straight line

$$H(\mu_0) \triangleq \{(\gamma, \phi) : \; \phi + \mu_0 \gamma = l_0(\mu_0)\}. \tag{2.52}$$

In words, the smallest intercept we can obtain for a given μ_0 is given precisely by the dual function. This is illustrated in Fig. 2.10.

To sum up, for each μ_0, there are many straight lines of the form (2.50), one for each possible value of l_0. However, for each μ_0, there is a single straight line (2.52), the intercept of it being determined by (2.51).

Now, among all the straight lines $H(\mu)$ with $\mu \geq 0$, choose the one with highest intercept and let μ^* be the corresponding value of μ. The intercept of $H(\mu^*)$ is therefore

$$l_0(\mu^*) = \sup_{\mu} \; l_0(\mu) = \sup_{\mu} \; q(\mu) = q^*. \tag{2.53}$$

This is illustrated in Fig. 2.11.

Quiz 2.40 Does strong duality hold in Fig. 2.11? If the answer is negative, can you indicate the duality gap $f^* - q^*$ on the figure?

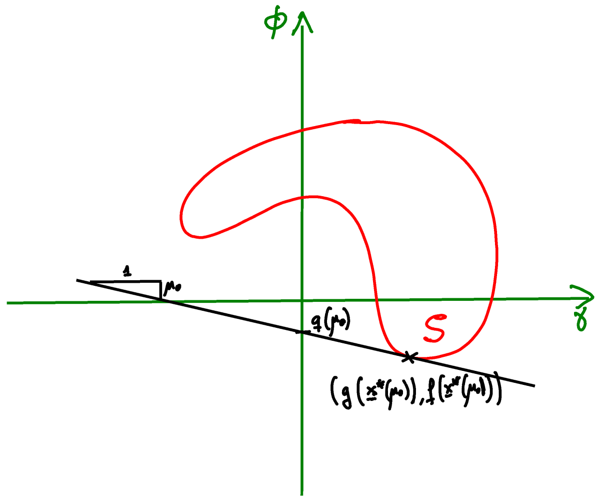


Figure 2.10: Geometric interpretation of the dual function.

Quiz 2.41 Can you indicate on the figure where the minimizers of $\mathcal{L}(\boldsymbol{x}; \mu^*)$ with respect to $\boldsymbol{x} \in \mathcal{Z}$ are mapped to on the (γ, ϕ) plane? Are all of them primal feasible?

Quiz 2.42 How does the set $\mathcal{S}$ look like when $\mu^* = 0$ and strong duality holds? Provide at least one graphical example.

Quiz 2.43 How does the set $\mathcal{S}$ look like when $\mu^* = 0$ and strong duality does not hold? Provide at least one graphical example.

Quiz 2.44 How does the set $\mathcal{S}$ look like when $\mu^* > 0$ and strong duality does not hold? Provide at least one graphical example.

Quiz 2.45 How does the set $\mathcal{S}$ look like when $\mu^* > 0$ and strong duality holds? Provide at least one graphical example.

Quiz 2.46 Are your previous answers consistent with the notion of complementary slackness introduced earlier in this chapter?

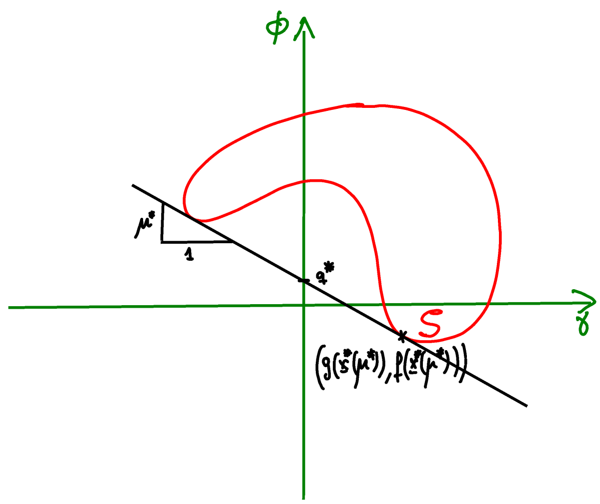


Figure 2.11: Illustration of q^*.

Chapter 3

Introduction to Iterative Methods

[Overview] This chapter reviews first- and second-order methods for minimizing differentiable functions. Most results are proven in [nesterov2004].

3.1 Introduction

[motivation]When a closed-form solution cannot be found or when obtaining such a solution is computationally too expensive, one resorts to iterative methods.
[Def.] An iterative method is an algorithm that generates a sequence of *iterates* $\boldsymbol{x}^{(0)}, \boldsymbol{x}^{(1)}, \ldots$ These algorithms are typically designed so that $\boldsymbol{x}^{(i)} \to \boldsymbol{x}^*$ as $i \to \infty$.
[Examples]

- [Gradient descent] To minimize $f(\boldsymbol{x}) : \mathbb{R}^D \to \mathbb{R}$,
  - Initialize $\boldsymbol{x}^{(0)}$
  - For $i = 1, 2, \ldots$
    * $\boldsymbol{x}^{(i)} = \boldsymbol{x}^{(i-1)} - \eta^{(i)} \nabla f(\boldsymbol{x}^{(i-1)})$
    * Quit if convergence criterion satisfied.

  The sequence of step sizes $\eta^{(i)}$ will be discussed later in the course.

  Quiz 3.1 Now you are familiar with optimality conditions, propose two simple convergence criteria that are satisfied if (and only if) $\boldsymbol{x}^{(i)}$ is somehow close to a stationary point. One may depend on $\nabla f(\boldsymbol{x}^{(i)})$ and the other on the iterates $\boldsymbol{x}^{(i)}$.

- [Alternating minimization] This method is also called *block-coordinate minimization* or the *non-linear Gauss-Seidel method*. To minimize $f(\boldsymbol{x}_1, \boldsymbol{x}_2) : \mathbb{R}^D \to \mathbb{R}$, with optimization variable $\boldsymbol{x} \triangleq [\boldsymbol{x}_1^\top, \boldsymbol{x}_2^\top]^\top$,
  - Initialize $\boldsymbol{x}_2^{(0)}$
  - For $i = 1, 2, \ldots$
    * $\boldsymbol{x}_1^{(i)} \in \arg\min_{\boldsymbol{x}_1} f(\boldsymbol{x}_1, \boldsymbol{x}_2^{(i-1)})$
    * $\boldsymbol{x}_2^{(i)} \in \arg\min_{\boldsymbol{x}_2} f(\boldsymbol{x}_1^{(i)}, \boldsymbol{x}_2)$
    * Quit if convergence criterion satisfied.

Quiz 3.2 Obtain the sequences $x_1^{(i)}$ and $x_2^{(i)}$ when $f(x_1, x_2) = a_{11}x_1^2 + a_{22}x_2^2 + 2a_{12}x_1x_2 + b_1x_1 + b_2x_2$.

- [Further methods] Many other iterative methods will be discussed in the rest of the course.

[Convergence analysis]

- [Motivation] To apply an iterative algorithm in a practical scenario, characterizing its convergence is of utmost importance for the following reasons:
  - If convergence is not guaranteed theoretically, the algorithm may not converge for certain problem instances. This means that a device that uses an optimization algorithm could malfunction or hang in a real scenario despite having worked properly in all lab tests.
  - Convergence results help us decide which algorithm to use, for instance based on which is the fastest method for a given problem.
  - Convergence results help us approximate the time needed to solve an optimization problem with a given algorithm.
  - Convergence results help us determine whether a given optimization method can be used for a given problem with the available computational resources.
- [classes of convergence analysis]
  - [Convergence to stationary points]
    * Sanity check.
    * Minimal requirement of any reasonable algorithm.
    * Does not provide global characterization of the behavior of the algorithm.
  - [Iteration complexity analysis]
    * Quantifies number of iterations to approximate an optimum: e.g. for given ϵ, find an upper bound for the minimum i such that $f(\boldsymbol{x}^{(i)}) - f(\boldsymbol{x}^*) < \epsilon$.
    * Typically global analysis (arbitrary initialization $\boldsymbol{x}^{(0)}$).
  - [Asymptotic convergence rates]
    * Typically local analysis ($\boldsymbol{x}^{(0)}$ already close to $\boldsymbol{x}^*$).
    * Provides convergence rate of error sequences, e.g.
      · $e^{(i)} \triangleq ||\boldsymbol{x}^{(i)} - \boldsymbol{x}^*||$, or
      · $e^{(i)} \triangleq f(\boldsymbol{x}^{(i)}) - f(\boldsymbol{x}^*)$.
    * [Classes of convergence] For sequences $e^{(i)}$ converging to 0, we distinguish several forms of convergence depending on how they compare to the well-known geometric progression β^i, $i = 0, 1, \ldots$ where $\beta \in (0, 1)$.
      · [Linear convergence]

Definition 3.1 *The sequence* $e^{(i)}$ *is said to converge* linearly *(or* geometrically*) if there exist* $q > 0$ *and* $\beta \in (0,1)$ *such that*

$$e^{(i)} \leq q\beta^i \quad \forall i. \tag{3.1}$$

If no such q *and* β *exist, then the sequence* $e^{(i)}$ *is said to converge* sublinearly.

The name *linear* owes to the fact that the logarithm of $e^{(i)}$ in (3.1) decreases linearly: $\log e^{(i)} < \log q + i \log \beta \quad \forall i$.

Quiz 3.3 To get some sense of what linear convergence is, suppose that there exist $q > 0$ and $\beta \in (0, 1)$ such that (3.1) holds with equality for all i. In that case, what is the relation between $e^{(i)}$ and $e^{(i-1)}$?

A sufficient condition for linear convergence is [bertsekas1999, Sec. 1.3.1].

$$\exists \beta \in (0,1) \text{ such that } \limsup_{i\to\infty} \frac{e^{(i+1)}}{e^{(i)}} \leq \beta.$$

· [Superlinear convergence]

Definition 3.2 *The sequence $e^{(i)}$ converges* superlinearly *if for every $\beta \in (0, 1)$ there exists a $q > 0$ such that*

$$e^{(i)} \leq q\beta^i \quad \forall i.$$

A sufficient condition for superlinear convergence is [bertsekas1999, Sec. 1.3.1].

$$\limsup_{i\to\infty} \frac{e^{(i+1)}}{e^{(i)}} = 0.$$

Quiz 3.4 Suppose that $e^{(i)}$ converges superlinearly. Does it necessarily converge linearly as well?

Quiz 3.5 Characterize the convergence of $e^{(i)} = 1/i$.

3.2 Descent Methods

[Goal]We would like to devise optimization methods to generate sequences of iterates $\boldsymbol{x}^{(0)}, \boldsymbol{x}^{(1)}, \ldots$, such that $f(\boldsymbol{x}^{(0)}), f(\boldsymbol{x}^{(1)}), \ldots$ is decreasing and converges to $f(\boldsymbol{x}^*)$.
[Descent lemma]To find such methods, it is useful to consider the following result, known as the *descent lemma*.

Proposition 3.1 *Let $f : \mathbb{R}^D \to \mathbb{R}$ be continuously differentiable. Suppose that*

$$||\nabla f(\boldsymbol{x} + t\boldsymbol{z}) - \nabla f(\boldsymbol{x})|| \leq L||t\boldsymbol{z}||, \quad \forall t \in [0, 1], \tag{3.2}$$

for some $L \in \mathbb{R}$, $\boldsymbol{x}$, $\boldsymbol{z} \in \mathbb{R}^D$. Then

$$f(\boldsymbol{x} + \boldsymbol{z}) \leq f(\boldsymbol{x}) + \boldsymbol{z}^\top \nabla f(\boldsymbol{x}) + \frac{L}{2}||\boldsymbol{z}||^2.$$

Proof: See proof of [bertsekas1999, Prop. A.24].

□

Quiz 3.6 Provide a sufficient condition for (3.2) to hold.

[Objective decrease]The following result establishes that moving a sufficiently small distance along a direction that makes an angle greater than 90 degrees with the gradient produces a decrease in the objective value.

Proposition 3.2 *Let $f : \mathbb{R}^D \to \mathbb{R}$ be continuously differentiable, let $\boldsymbol{x}$ be such that $\nabla f(\boldsymbol{x}) \neq \mathbf{0}$, and let $\boldsymbol{d}$ be such that $\boldsymbol{d}^\top \nabla f(\boldsymbol{x}) < 0$. Then, if (3.2) holds when $\boldsymbol{z} = \eta \boldsymbol{d}$ $\forall \eta > 0$, there exists $\delta > 0$ such that*

$$f(\boldsymbol{x} + \eta \boldsymbol{d}) < f(\boldsymbol{x}), \quad \forall \eta \in (0, \delta).$$

Quiz 3.7 Use Proposition 3.1 to prove Proposition 3.2.

[Descent methods] Proposition 3.2 suggests iterative methods of the form:

- Initialize $\boldsymbol{x}^{(0)}$
- For $i = 0, 1, \ldots$
  - $\boldsymbol{x}^{(i+1)} = \boldsymbol{x}^{(i)} + \eta^{(i)} \boldsymbol{d}^{(i)}$
  - Quit if convergence criterion satisfied.

where

- $\boldsymbol{d}^{(i)}$ is a descent direction at iteration i, i.e., it satisfies $(\boldsymbol{d}^{(i)})^\top \nabla f(\boldsymbol{x}^{(i)}) < 0$.
- $\eta^{(i)} > 0$ is sufficiently small; more information later.

This algorithm is termed *descent* method since $f(\boldsymbol{x}^{(i+1)}) < f(\boldsymbol{x}^{(i)})$ $\forall i$, unless we reach a stationary point.
[Descent directions]

- [Typical form] Some descent methods select the direction $\boldsymbol{d}^{(i)} = -\boldsymbol{D}^{(i)} \nabla f(\boldsymbol{x}^{(i)})$ where $\boldsymbol{D}^{(i)}$ is a positive definite matrix.

  Quiz 3.8 Show that the aforementioned direction is a descent direction.

- [Steepest descent method]
  - [Direction] The popular steepest descent method adopts $\boldsymbol{D}^{(i)} = \boldsymbol{I}$ or, equivalently, $\boldsymbol{d}^{(i)} = -\nabla f(\boldsymbol{x}^{(i)})$.
  - [Features]
    - No memory.
    - Relatively slow convergence → may zig-zag.
    - Convergence related to condition number.
- [Newton's method]
  - [Direction]Newton's method sets $\boldsymbol{d}^{(i)} = -[\nabla^2 f(\boldsymbol{x}^{(i)})]^{-1} \nabla f(\boldsymbol{x}^{(i)})$.
  - [Features]
    - No memory.
    - Fast convergence.
    - Insensitive to condition number.
    - Potentially unstable and prone to numerical issues.

Quiz 3.9 How many iterations does Newton's method need to minimize a convex quadratic function?

[Step-size sequences]

- [Fixed sequences] The entire sequence is selected before executing the algorithm. Thus, $\eta^{(i)}$ does not depend on $\{\boldsymbol{x}^{(i')}\}_{i'}$ or $\{f(\boldsymbol{x}^{(i')})\}_{i'}$.
  - [Constant step size] This is the simplest sequence: $\eta^{(i)} = \eta, \quad \forall i$.
  - [Diminishing step size] This is a sequence satisfying $\eta^{(i)} \to 0$. For convergence, it is typically necessary that it also satisfies the so-called *infinite travel condition*:

$$\sum_{i=0}^{\infty} \eta^{(i)} = \infty.$$

Quiz 3.10 Which of the sequences $\eta^{(i)} = 1/(i+1)$, $\eta^{(i)} = 1/(i+1)^2$, and $\eta^{(i)} = 1/\sqrt{i+1}$ satisfy the infinite travel condition?

- [Line search]
  - [Minimization rule]

$$\eta^{(i)} \in \arg\min_{\eta > 0} f(\boldsymbol{x}^{(i)} + \eta \boldsymbol{d}^{(i)}).$$

  - [Limited rule] Fix $\eta_{\max}$ and set

$$\eta^{(i)} \in \arg\min_{\eta \in [0, \eta_{\max}]} f(\boldsymbol{x}^{(i)} + \eta \boldsymbol{d}^{(i)}).$$

  - [Armijo rule (backtracking)]
    Fix $\sigma \in (0, 1/2)$, $s > 0$, and $\beta \in (0, 1)$. At every iteration i:
    * Set $\eta^{(i)} = s$.
    * Repeat
      · $\eta^{(i)} \leftarrow \beta \eta^{(i)}$

      until $f(\boldsymbol{x}^{(i)} + \eta^{(i)} \boldsymbol{d}^{(i)}) \leq f(\boldsymbol{x}^{(i)}) + \sigma \eta^{(i)} \nabla^\top f(\boldsymbol{x}^{(i)}) \boldsymbol{d}^{(i)}$.

    It can be shown that the number of iterations of this repeat loop is finite. Backtracking is perhaps the most common rule for selecting the step size in descent methods.

Quiz 3.11 What are the strengths and weaknesses of each of these line search rules?

3.2.1 Convergence Analysis

3.2.1.1 First-order Methods

[Convergence]The next result establishes the sublinear convergence of steepest descent when the objective is not necessarily strongly convex.

Theorem 3.1 *Let f be convex and L-Lipschitz smooth. If $\boldsymbol{x}^*$ is a global optimum of f and $\boldsymbol{x}^{(0)}, \boldsymbol{x}^{(1)}, \ldots$ is the sequence of iterates of steepest descent with constant step size η satisfying $0 < \eta < 2/L$. Then*

$$f(\boldsymbol{x}^{(i)}) - f(\boldsymbol{x}^*) \leq \left[\frac{||\boldsymbol{x}^{(0)} - \boldsymbol{x}^*||^2}{\eta(2-\eta L)}\right]\frac{1}{i}, \ i \geq 1. \tag{3.3}$$

Proof: This result will be proven in a homework.

□

- [sublinear] Since (3.3) is an upper bound, the algorithm may converge linearly for a specific problem instance. However, there exist non-strongly convex objectives for which the convergence of this method is sublinear [nesterov2004, Th. 2.1.7].
- [Lowest upper bound] Minimizing the right-hand side of (3.3) with respect to η yields the bound

$$f(\boldsymbol{x}^{(i)}) - f(\boldsymbol{x}^*) \leq \frac{L||\boldsymbol{x}^{(0)} - \boldsymbol{x}^*||^2}{i}, \ i \geq 1. \tag{3.4}$$

  for $\eta = 1/L$.
- [Iteration complexity]
  - [Goal] Number of iterations required to find an ϵ-optimal solution, i.e., an iterate $\boldsymbol{x}^{(i)}$ satisfying $f(\boldsymbol{x}^{(i)}) - f(\boldsymbol{x}^*) \leq \epsilon$.
  - [Error sequence] For every convex L-Lipschitz smooth function f and initialization $\boldsymbol{x}^{(0)}$, there exists a sequence $e^{(i)}_{f,\boldsymbol{x}^{(0)}} \triangleq f(\boldsymbol{x}^{(i)}) - f(\boldsymbol{x}^*)$. To simplify the explanation, assume that the algorithm is always initialized with the same $\boldsymbol{x}^{(0)}$, yet the arguments here immediately carry over to the case where this assumption is not satisfied. Therefore, the dependence on the initialization can be omitted from the notation, i.e., $e^{(i)}_{f,\boldsymbol{x}^{(0)}}$ becomes $e^{(i)}_f$.
  - [Required number of iterations] For each f, let i^{req}_f denote the smallest non-negative integer such that $e^{(i)}_f \leq \epsilon$ for all $i \geq i^{\text{req}}_f$. This integer is guaranteed to exist since (3.4) implies that $e^{(i)}_f$ converges to 0 as $i \to \infty$.
  - [Bound sequence] Similarly, let $e^{(i)}_{\text{B}} \triangleq L||\boldsymbol{x}^{(0)} - \boldsymbol{x}^*||^2/i$ denote the upper bound on $e^{(i)}_f$ implied by (3.4). Since $e^{(i)}_{\text{B}}$ converges to 0, one can find the smallest non-negative integer $i^{\text{req}}_{\text{B}}$ such that $e^{(i)}_{\text{B}} \leq \epsilon$ for all $i \geq i^{\text{req}}_{\text{B}}$.
  - [Maximum required number of iterations] Since $e^{(i)}_f \leq e^{(i)}_{\text{B}}$ for all i, it readily follows that, for all f,

$$i^{\text{req}}_f \leq i^{\text{req}}_{\text{B}} = \left\lceil \frac{L||\boldsymbol{x}^{(0)} - \boldsymbol{x}^*||^2}{\epsilon} \right\rceil.$$

- [linear convergence for strongly convex objectives]When the objective is strongly convex, then steepest descent converges linearly. For simplicity, this claim will be proven only for minimizing a quadratic objective

$$f(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}^\top \boldsymbol{x},$$

where $\boldsymbol{A} \succ \mathbf{0}$. Then, steepest descent iterates as

$$\boldsymbol{x}^{(i+1)} = \boldsymbol{x}^{(i)} - \eta^{(i)} \nabla f(\boldsymbol{x}^{(i)}) = \boldsymbol{x}^{(i)} - \eta^{(i)} (\boldsymbol{A}\boldsymbol{x}^{(i)} + \boldsymbol{b}).$$

$$\begin{aligned} ||\boldsymbol{x}^{(i+1)} - \boldsymbol{x}^*|| &= ||\boldsymbol{x}^{(i)} - \eta^{(i)} \nabla f(\boldsymbol{x}^{(i)}) - [\boldsymbol{x}^* - \eta^{(i)} \nabla f(\boldsymbol{x}^*)]|| \\ &= ||\boldsymbol{x}^{(i)} - \eta^{(i)} (\boldsymbol{A}\boldsymbol{x}^{(i)} + \boldsymbol{b}) - [\boldsymbol{x}^* - \eta^{(i)} (\boldsymbol{A}\boldsymbol{x}^* + \boldsymbol{b})]|| \\ &= ||(\boldsymbol{I} - \eta^{(i)} \boldsymbol{A})(\boldsymbol{x}^{(i)} - \boldsymbol{x}^*)|| \\ &\leq \mathrm{sval}_{\max}(\boldsymbol{I} - \eta^{(i)} \boldsymbol{A}) \; ||\boldsymbol{x}^{(i)} - \boldsymbol{x}^*|| \\ &= \max\{|1 - \eta^{(i)} \lambda_{\max}(\boldsymbol{A})|, |1 - \eta^{(i)} \lambda_{\min}(\boldsymbol{A})|\} \; ||\boldsymbol{x}^{(i)} - \boldsymbol{x}^*||. \end{aligned}$$

This bound is minimized for the constant step size

$$\eta^{(i)} = \eta = \frac{2}{\lambda_{\min}(\boldsymbol{A}) + \lambda_{\max}(\boldsymbol{A})},$$

in which case one obtains

$$||\boldsymbol{x}^{(i+1)} - \boldsymbol{x}^*|| \leq \frac{\lambda_{\max}(\boldsymbol{A}) - \lambda_{\min}(\boldsymbol{A})}{\lambda_{\max}(\boldsymbol{A}) + \lambda_{\min}(\boldsymbol{A})} \; ||\boldsymbol{x}^{(i)} - \boldsymbol{x}^*|| = \left[1 - \frac{2}{1+\kappa}\right] \; ||\boldsymbol{x}^{(i)} - \boldsymbol{x}^*||, \tag{3.5}$$

where $\kappa := \lambda_{\max}(\boldsymbol{A})/\lambda_{\min}(\boldsymbol{A})$ is called the *condition number*.

Quiz 3.12 What is the influence of κ on the convergence of the algorithm?

Quiz 3.13 Does (3.5) establish linear convergence? why?

Quiz 3.14 Use (3.5) to show that the required number of iterations needed to attain a relative error $||\boldsymbol{x}^{(i+1)} - \boldsymbol{x}^*||/||\boldsymbol{x}^{(0)} - \boldsymbol{x}^*||$ less than ϵ is

$$\mathcal{O}(\kappa \log(1/\epsilon))$$

as $\kappa \to \infty$ and $\epsilon \to 0$. You can use the fact that $\log(1+z) \approx z$ for small z.

3.2.1.2 Second-order Methods

[Newton's method]The next result establishes superlinear convergence for Newton's algorithm and alike.

Theorem 3.2 *Let f be twice continuously differentiable. Let $\boldsymbol{x}^{(i+1)} = \boldsymbol{x}^{(i)} + \eta^{(i)} \boldsymbol{d}^{(i)}$ where*

- *$\eta^{(i)}$ is chosen by backtracking with $s = 1$.*
- *The direction $\boldsymbol{d}^{(i)}$ satisfies*

$$\lim_{i\to\infty} \frac{||\boldsymbol{d}^{(i)} + [\nabla^2 f(\boldsymbol{x}^*)]^{-1} \nabla f(\boldsymbol{x}^{(i)})||}{||\nabla f(\boldsymbol{x}^{(i)})||} = 0.$$

- *$\boldsymbol{x}^{(i)} \to \boldsymbol{x}^*$, where $\nabla f(\boldsymbol{x}^*) = \mathbf{0}$ and $\nabla^2 f(\boldsymbol{x}^*) \succ \mathbf{0}$.*

Then,

$$\lim_{i\to\infty} \frac{||\boldsymbol{x}^{(i+1)} - \boldsymbol{x}^*||}{||\boldsymbol{x}^{(i)} - \boldsymbol{x}^*||} = 0.$$

Proof: See proof of [bertsekas1999, Prop. 1.3.2].

□

3.3 Optimal First-order Methods

[Nesterov's method] Let f be L-Lipschitz smooth. Nesterov's method iterates as follows:

- Initialize: $\boldsymbol{x}^{(0)}$, $\boldsymbol{x}^{(1)}$.
- For $i = 1, 2, \ldots$
  - $\boldsymbol{z}^{(i+1)} = \boldsymbol{x}^{(i)} + t^{(i)}(\boldsymbol{x}^{(i)} - \boldsymbol{x}^{(i-1)})$.
  - $\boldsymbol{x}^{(i+1)} = \boldsymbol{z}^{(i+1)} - \frac{1}{L}\nabla f(\boldsymbol{z}^{(i+1)})$.

where

$$t^{(i)} := \frac{a^{(i)} - 1}{a^{(i+1)}}, \quad i \geq 1, \qquad a^{(i)} := \frac{1}{2}\left(1 + \sqrt{1 + 4(a^{(i-1)})^2}\right). \tag{3.6}$$

[convergence]

Theorem 3.3 *The above Nesterov method satisfies*

$$f(\boldsymbol{x}^{(i)}) - f(\boldsymbol{x}^*) \leq \frac{2L||\boldsymbol{x}^{(0)} - \boldsymbol{x}^*||^2}{(i+1)^2}, \ \forall i \geq 1. \tag{3.7}$$

Proof: See [nesterov2004].

□

- [sublinear] Since (3.7) is just an upper bound, the algorithm may converge linearly for a specific problem instance. However, there exist non-strongly convex objectives for which the convergence of this method is sublinear [nesterov2004, Th. 2.1.7].
- [iteration complexity] Note that one can obtain a simpler bound of the same order by exploiting the fact that $(i+1)^2 \geq i^2$:

$$f(\boldsymbol{x}^{(i)}) - f(\boldsymbol{x}^*) \leq \frac{2L||\boldsymbol{x}^{(0)} - \boldsymbol{x}^*||^2}{i^2}, \ \forall i \geq 1. \tag{3.8}$$

  Thus, the maximum number of iterations required to attain an ϵ-optimal solution is at most

$$\sqrt{2L||\boldsymbol{x}^{(0)} - \boldsymbol{x}^*||^2/\epsilon}.$$

- [linear convergence for strongly convex objectives] Let f be α-strongly convex. Recall from Ch. 1 that this implies that

$$f(\boldsymbol{x}) - f(\boldsymbol{x}^*) \geq \frac{\alpha}{2}||\boldsymbol{x} - \boldsymbol{x}^*||^2, \quad \forall \boldsymbol{x}. \tag{3.9}$$

After $i_0 = \sqrt{8\kappa}$ iterations, where $\kappa := L/\alpha$ is the condition number[1], it follows from (3.9) and (3.8) that

$$f(\boldsymbol{x}^{(i_0)}) - f(\boldsymbol{x}^*) \leq \frac{\alpha||\boldsymbol{x}^{(0)} - \boldsymbol{x}^*||^2}{4} \leq \frac{1}{2}\left[f(\boldsymbol{x}^{(0)}) - f(\boldsymbol{x}^*)\right]. \tag{3.10}$$

Then, one can *restart* the algorithm by setting $\boldsymbol{x}^{(0)} \leftarrow \boldsymbol{x}^{(i_0)}$ and then the error halves every i_0 iterations. This implies that the maximum number of iterations required to attain an ϵ-optimal solution is

$$\mathcal{O}(\sqrt{\kappa}\log(1/\epsilon)).$$

[1]This definition of condition number generalizes the one in Sec. 3.2.1.1, which only applies to quadratic functions

Appendix A

Mathematical Background

A.1 Linear and Affine Spaces

Definition A.1 *The null space of a matrix* $\boldsymbol{A} \in \mathbb{R}^{M \times N}$ *is*

$$\mathcal{N}\{\boldsymbol{A}\} \triangleq \{\boldsymbol{x} \in \mathbb{R}^N : \boldsymbol{A}\boldsymbol{x} = \boldsymbol{0}\}$$

Definition A.2 *The column space, column span, or range of* $\boldsymbol{A} \triangleq [\boldsymbol{a}_1, \ldots, \boldsymbol{a}_N] \in \mathbb{R}^{M \times N}$ *is*

$$\mathcal{R}\{\boldsymbol{A}\} \triangleq \left\{\boldsymbol{x} \in \mathbb{R}^M : \boldsymbol{x} = \sum_{n=1}^{N} \alpha_n \boldsymbol{a}_n, \alpha_n \in \mathbb{R}\right\} = \left\{\boldsymbol{x} \in \mathbb{R}^M : \boldsymbol{x} = \boldsymbol{A}\boldsymbol{\alpha}, \boldsymbol{\alpha} \in \mathbb{R}^M\right\}$$

Quiz A.1 Show that $\mathcal{N}\{\boldsymbol{A}\}$ and $\mathcal{R}\{\boldsymbol{A}\}$ are subspaces.

Definition A.3 *The orthogonal complement of a subspace* $\mathcal{A} \subset \mathbb{R}^N$ *is defined as*

$$\mathcal{A}^{\perp} = \{\boldsymbol{x} \in \mathbb{R}^D : \boldsymbol{x}^{\top}\boldsymbol{v} = 0 \ \forall \boldsymbol{v} \in \mathcal{A}\}. \tag{A.1}$$

Proposition A.1

$$\mathcal{N}(\boldsymbol{A}) = \mathcal{R}^{\perp}(\boldsymbol{A}^{\top}) \tag{A.2}$$

Quiz A.2 Prove Proposition A.1.

Equivalently, we may write $\mathcal{N}^{\perp}(\boldsymbol{A}) = \mathcal{R}(\boldsymbol{A}^{\top})$.

The following proposition provides two alternative representations of an affine set:

Proposition A.2 *Let* $\boldsymbol{A} \in \mathbb{R}^{M \times N}$ *be a matrix with full row rank, where* $M \leq N$*, and let* $\boldsymbol{b} \in \mathbb{R}^M$*. Then,*

$$\{\boldsymbol{x} \in \mathbb{R}^N : \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}\} = \{\boldsymbol{F}\boldsymbol{\alpha} + \boldsymbol{g}, \ \boldsymbol{\alpha} \in \mathbb{R}^{N-M}\} \tag{A.3}$$

where $\boldsymbol{F} \in \mathbb{R}^{N \times N-M}$ *is such that* $\mathcal{R}\{\boldsymbol{F}\} = \mathcal{N}\{\boldsymbol{A}\}$ *and* $\boldsymbol{g}$ *satisfies* $\boldsymbol{A}\boldsymbol{g} = \boldsymbol{b}$.

Proof: Consider the difference of two vectors $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ in the left-hand side set of (A.3). Clearly,

$$\boldsymbol{A}(\boldsymbol{x}_1 - \boldsymbol{x}_2) = \boldsymbol{0}.$$

Therefore, $\boldsymbol{x}_1 - \boldsymbol{x}_2 \in \mathcal{N}\{\boldsymbol{A}\}$, which means that

$$\boldsymbol{x}_1 = \boldsymbol{x}_2 + \mathcal{N}\{\boldsymbol{A}\}.$$

The result clearly follows by making $\boldsymbol{g} = \boldsymbol{x}_2$ for any particular solution $\boldsymbol{x}_2$. ∎

□

Quiz A.3 Given $\boldsymbol{A}$, which MATLAB commands could be used to compute $\boldsymbol{F}$?

Note that in order for the requirement $\mathcal{R}\{\boldsymbol{F}\} = \mathcal{N}\{\boldsymbol{A}\}$ to hold, we just need two conditions

1. The matrix $[\boldsymbol{A}^\top, \boldsymbol{F}] \in \mathbb{R}^{N \times N}$ is full rank
2. $\boldsymbol{A}\boldsymbol{F} = \boldsymbol{0}$.

Typically, one may take $\boldsymbol{F}$ with orthonormal columns.

A.2 Derivatives

[Notation for partial derivatives] Let $f : \mathbb{R}^N \to \mathbb{R}$ and $\boldsymbol{x} \triangleq [x_1, \ldots, x_N]^\top$. Conventional notation for first- and second-order partial derivatives is [borden]:

$$D_m f(\boldsymbol{x}) = \frac{\partial f(\boldsymbol{x})}{\partial x_m}$$
$$D_{mn} f(\boldsymbol{x}) = \frac{\partial^2 f(\boldsymbol{x})}{\partial x_n \partial x_m} = \frac{\partial}{\partial x_n}\left[\frac{\partial}{\partial x_m} f(\boldsymbol{x})\right].$$

[Definitions]

- [Gradient]For $f : \mathbb{R}^N \to \mathbb{R}$ a a differentiable function, the gradient vector is

$$\nabla f(\boldsymbol{x}) \triangleq \begin{bmatrix} \frac{\partial}{\partial x_1} f(\boldsymbol{x}) \\ \frac{\partial}{\partial x_2} f(\boldsymbol{x}) \\ \vdots \\ \frac{\partial}{\partial x_N} f(\boldsymbol{x}) \end{bmatrix}. \tag{A.4}$$

  For $\boldsymbol{f} : \mathbb{R}^N \to \mathbb{R}^M$, where $\boldsymbol{f}(\boldsymbol{x}) = [f_1(\boldsymbol{x}), \ldots, f_M(\boldsymbol{x})]^\top$, the gradient *matrix* is

$$\nabla \boldsymbol{f}(\boldsymbol{x}) \triangleq [\nabla f_1(\boldsymbol{x}), \ldots, \nabla f_M(\boldsymbol{x})]. \tag{A.5}$$

  Thus, the gradient is an $N \times M$ matrix with entries

$$[\nabla \boldsymbol{f}(\boldsymbol{x})]_{n,m} \triangleq \frac{\partial f_m(\boldsymbol{x})}{\partial x_n}. \tag{A.6}$$

- [Jacobian] Let $\boldsymbol{f} : \mathbb{R}^{\mathbb{N}} \to \mathbb{R}^M$ with $\boldsymbol{f}(\boldsymbol{x}) = [f_1(\boldsymbol{x}), \ldots, f_N(\boldsymbol{x})]^\top$. The Jacobian is an $M \times N$ matrix with entries

$$[\boldsymbol{J}_{\boldsymbol{f}}(\boldsymbol{x})]_{m,n} \triangleq \frac{\partial f_m(\boldsymbol{x})}{\partial x_n} \tag{A.7}$$

  It follows that $\boldsymbol{J}_{\boldsymbol{f}}(\boldsymbol{x}) = \nabla^\top \boldsymbol{f}(\boldsymbol{x})$.

- [Hessian]For $f : \mathbb{R}^N \to \mathbb{R}$ a a twice differentiable function, the Hessian is an $N \times N$ matrix with entries

$$\left[\nabla^2 f(\boldsymbol{x})\right]_{n,m} = D_{nm} f(\boldsymbol{x}) = \frac{\partial}{\partial x_m}\left[\frac{\partial}{\partial x_n}\ f(\boldsymbol{x})\right] \tag{A.8}$$

Note that, according to Schwartz's theorem, $D_{nm} f(\boldsymbol{x}) = D_{mn} f(\boldsymbol{x})$ whenever the second-order partial derivatives are continuous at $\boldsymbol{x}$. Thus, $\nabla^2 f(\boldsymbol{x})$ is symmetric at those points $\boldsymbol{x}$ for which this condition holds.

The Hessian matrix is the **Jacobian of the gradient**.

[Chain rule]

- [For Jacobians]

Proposition A.3 *Suppose that* $\boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{g}(\boldsymbol{v}(\boldsymbol{x}))$*:*

$$\begin{array}{ccccc} \mathbb{R}^N & \xrightarrow{\boldsymbol{v}(\boldsymbol{x})} & \mathbb{R}^K & \xrightarrow{\boldsymbol{g}(\boldsymbol{v})} & \mathbb{R}^M \\ \boldsymbol{x} & & \boldsymbol{v}(\boldsymbol{x}) & & \boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{g}(\boldsymbol{v}(\boldsymbol{x})) \end{array} \tag{A.9}$$

Then,

$$\boldsymbol{J_f}(\boldsymbol{x}) = \boldsymbol{J_g}(\boldsymbol{v})|_{\boldsymbol{v}=\boldsymbol{v}(\boldsymbol{x})} \boldsymbol{J_v}(\boldsymbol{x}). \tag{A.10}$$

*Proof:*From the well-known chain rule for partial derivatives

$$\begin{aligned} \frac{\partial}{\partial x_n} f_m(\boldsymbol{x}) &= \frac{\partial}{\partial x_n} g_m(\boldsymbol{v}(\boldsymbol{x})) \\ &= \sum_k \frac{\partial}{\partial v_k} g_m(\boldsymbol{v})|_{\boldsymbol{v}=\boldsymbol{v}(\boldsymbol{x})}\ \frac{\partial}{\partial x_n} v_k(\boldsymbol{x}). \end{aligned}$$

In terms of the previous definitions, this expression becomes

$$\left[\boldsymbol{J_f}(\boldsymbol{x})\right]_{m,n} = \sum_k \left[\boldsymbol{J_g}(\boldsymbol{v})|_{\boldsymbol{v}=\boldsymbol{v}(\boldsymbol{x})}\right]_{m,k}\ \left[\boldsymbol{J_v}(\boldsymbol{x})\right]_{k,n}. \tag{A.11}$$

Finally, expressing (A.11) in matrix form yields (A.10).

□

- [For gradients]Let $\boldsymbol{f}$ be as in (A.9). Just by transposing (A.10), it follows that

$$\nabla \boldsymbol{f}(\boldsymbol{x}) = \nabla \boldsymbol{v}(\boldsymbol{x}) \nabla \boldsymbol{g}(\boldsymbol{v})|_{\boldsymbol{v}=\boldsymbol{v}(\boldsymbol{x})}. \tag{A.12}$$

- [Relation to scalar functions]To connect the above expressions with how we think when we compute derivatives of scalar functions in a single variable, suppose that $f(x) = \sin^3(x^2)$. Following (A.10) corresponds to proceeding outside-in, i.e., one would start by writing $f'(x) = [3\sin^2(x^2)] \cdot [\cos(x^2)] \cdot [2x]$. On the other hand, following (A.12) corresponds to proceeding inside-out, i.e., one would write $f'(x) = [2x] \cdot [\cos(x^2)] \cdot [3\sin^2(x^2)]$. Of course, in this scalar example, both expressions are the same, but note that order matters with functions of multiple variables since gradients and Jacobians are matrices.

- [Example]

Example A.1 Let $f(\boldsymbol{x}) = ||\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}||^2$. Then, by setting $\boldsymbol{v}(\boldsymbol{x}) = \boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}$ and $\boldsymbol{g}(\boldsymbol{v}) = ||\boldsymbol{v}||^2$, we find that

$$\begin{aligned}\nabla g(\boldsymbol{v}) &= \nabla_{\boldsymbol{v}} ||\boldsymbol{v}||^2 = 2\boldsymbol{v} \\ J_{\boldsymbol{v}}(\boldsymbol{x}) &= -\boldsymbol{A} \\ \nabla f(\boldsymbol{x}) &= -2\boldsymbol{A}^\top(\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}).\end{aligned}$$

Quiz A.4 Obtain the gradient and Hessian (when applicable) of

$$\begin{aligned}f(\boldsymbol{x}) &= \boldsymbol{a}^\top \boldsymbol{x} + b \\ f(\boldsymbol{x}) &= \boldsymbol{x}^\top \boldsymbol{a} + b \\ \boldsymbol{f}(\boldsymbol{x}) &= \boldsymbol{A}\boldsymbol{x} + \boldsymbol{b} \\ f(\boldsymbol{x}) &= \boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x} \\ f(\boldsymbol{x}, \boldsymbol{z}) &= \boldsymbol{x}^\top \boldsymbol{A}_1 \boldsymbol{x} + \boldsymbol{z}^\top \boldsymbol{A}_2 \boldsymbol{x} + \boldsymbol{z}^\top \boldsymbol{A}_3 \boldsymbol{z} \\ f(\boldsymbol{x}) &= \cos(\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x} + \boldsymbol{a}^\top \boldsymbol{x})\end{aligned}$$

A.3 Subgradient and subdifferential

This section is merely for reference. Its contents will be covered later in the course.

Definition A.4 *Let $f : \mathbb{R}^D \to \mathbb{R}$ be a function (not necessarily convex). A subgradient of f at $\boldsymbol{x}$ is any vector $\boldsymbol{g}_{\boldsymbol{x}}$ such that*

$$f(\boldsymbol{z}) \geq f(\boldsymbol{x}) + \boldsymbol{g}_{\boldsymbol{x}}^\top(\boldsymbol{z} - \boldsymbol{x}) \quad \textit{for all } \boldsymbol{z}.$$

A subdifferential $\partial f(\boldsymbol{x})$ is the set of all subgradients of f at $\boldsymbol{x}$. There are rules to compute subdifferentials that are called *weak subgradient calculus* if the purpose is to find at least a subgradient and *strong subgradient calculus* if the purpose is to find ALL subgradients. The latter is typically more difficult, but it is required by certain optimization algorithms.

Proposition A.4 *If a function has at least one subgradient at every point, then it is convex.*

Proof: Consider the points $\boldsymbol{x}$, $\boldsymbol{z}$ and $\boldsymbol{y} \triangleq \alpha\boldsymbol{x} + (1-\alpha)\boldsymbol{z}$. Since we have a subgradient $\boldsymbol{g}_{\boldsymbol{y}}$ at $\boldsymbol{y}$,

$$\begin{aligned}f(\boldsymbol{x}) &\geq f(\boldsymbol{y}) + \boldsymbol{g}_{\boldsymbol{y}}^\top(\boldsymbol{x} - \boldsymbol{y}) \\ f(\boldsymbol{z}) &\geq f(\boldsymbol{y}) + \boldsymbol{g}_{\boldsymbol{y}}^\top(\boldsymbol{z} - \boldsymbol{y}).\end{aligned}$$

Multiplying the first equation by α, the second by $(1-\alpha)$ and adding the result yields

$$\alpha f(\boldsymbol{x}) + (1-\alpha) f(\boldsymbol{z}) \geq f(\boldsymbol{y}) + \alpha \boldsymbol{g}_{\boldsymbol{y}}^\top(\boldsymbol{x} - \boldsymbol{y}) + (1-\alpha)\boldsymbol{g}_{\boldsymbol{y}}^\top(\boldsymbol{z} - \boldsymbol{y}) = f(\boldsymbol{y}),$$

which shows that f is convex.

□

Proposition A.5 *If a function is convex, then there is at least a subgradient at every point in the interior of its domain.*

Proof: The proof is based on choosing $\boldsymbol{x}_0$ in the interior of the domain of f and constructing a subgradient at this point by exploiting the fact that if f is convex, its epigraph is convex and therefore has a supporting hyperplane [boyd].

If f is convex, then the epigraph $\text{epi}f = \{[\boldsymbol{x}^\top, t]^\top : t \geq f(\boldsymbol{x})\}$ is a convex set. Thus, it has a supporting hyperplane at every boundary point, i.e., for every boundary point $[\boldsymbol{x}_0^\top, f(\boldsymbol{x}_0)]^\top$ there exists a vector $[\boldsymbol{a}^\top, b]^\top$ different from $\mathbf{0}$ satisfying

$$[\boldsymbol{a}^\top b]\left(\begin{bmatrix} \boldsymbol{x} \\ t \end{bmatrix} - \begin{bmatrix} \boldsymbol{x}_0 \\ f(\boldsymbol{x}_0) \end{bmatrix}\right) \leq 0 \quad \forall[\boldsymbol{x}^\top, t]^\top \in \text{epi}f$$

or, equivalently,

$$\boldsymbol{a}^\top(\boldsymbol{x} - \boldsymbol{x}_0) + b(t - f(\boldsymbol{x}_0)) \leq 0.$$

b cannot be positive as in that case the previous expression would not hold for t large enough. It cannot be 0 either since in that case setting $\boldsymbol{x} = \boldsymbol{x}_0 + \epsilon\boldsymbol{a}$ with $\epsilon > 0$ would also contradict the previous equation. Note that setting $\boldsymbol{x} = \boldsymbol{x}_0 + \epsilon\boldsymbol{a}$ for sufficiently small ϵ is allowed since $[\boldsymbol{x}_0^\top + \epsilon\boldsymbol{a}^\top, t]^\top$ is in the epigraph for sufficiently large t. This is because $\boldsymbol{x}_0$ is in the interior of the domain of f. Thus, b is negative and we obtain

$$\boldsymbol{a}^\top(\boldsymbol{x} - \boldsymbol{x}_0) - |b|(t - f(\boldsymbol{x}_0)) \leq 0.$$

$$\frac{1}{|b|}\boldsymbol{a}^\top(\boldsymbol{x} - \boldsymbol{x}_0) - (t - f(\boldsymbol{x}_0)) \leq 0.$$

Setting $t = f(\boldsymbol{x})$, which is possible since $[\boldsymbol{x}^\top, f(\boldsymbol{x})]^\top$ is in the epigraph, yields

$$f(\boldsymbol{x}) \geq f(\boldsymbol{x}_0) + \frac{1}{|b|}\boldsymbol{a}^\top(\boldsymbol{x} - \boldsymbol{x}_0)$$

so $\boldsymbol{a}/|b|$ is a subgradient.

□

The following result shows that the subgradient plays a role similar to the gradient, which is known to point in the steepest ascent direction.

Lemma A.1 *Let $\mathcal{C}_{\boldsymbol{x}_0} \triangleq \{\boldsymbol{x} : f(\boldsymbol{x}) \leq f(\boldsymbol{x}_0)\}$ be a sublevel set of a convex function f and let $\boldsymbol{g}$ be a subgradient at $\boldsymbol{x}_0$. Then $\boldsymbol{g}$ defines a supporting hyperplane for $\mathcal{C}_{\boldsymbol{x}_0}$:*

$$\boldsymbol{g}^\top(\boldsymbol{x} - \boldsymbol{x}_0) \leq 0 \quad \forall\boldsymbol{x} \in \mathcal{C}_{\boldsymbol{x}_0}.$$

Proof: Due to convexity of f, one has that $f(\boldsymbol{x}) \geq f(\boldsymbol{x}_0) + \boldsymbol{g}^\top(\boldsymbol{x} - \boldsymbol{x}_0)$. Since $f(\boldsymbol{x}) \leq f(\boldsymbol{x}_0)$, it follows that

$$0 \geq f(\boldsymbol{x}) - f(\boldsymbol{x}_0) \geq \boldsymbol{g}^\top(\boldsymbol{x} - \boldsymbol{x}_0).$$

□

Let $\partial f(\boldsymbol{x}_0)$ represent the set of all subgradients of f at $\boldsymbol{x}_0$. This set is called *subdifferential* of f at $\boldsymbol{x}_0$. The following provides an optimality condition that can be applied to an arbitrary function, be it differentiable or not.

Proposition A.6 *$\boldsymbol{x}_0$ is a global minimum of f if and only if $\mathbf{0} \in \partial f(\boldsymbol{x}_0)$.*

Proof: $\mathbf{0} \in \partial f(\boldsymbol{x}_0) \Leftrightarrow f(\boldsymbol{x}) \geq f(\boldsymbol{x}_0) + \mathbf{0}^\top(\boldsymbol{x} - \boldsymbol{x}_0)\ \forall\boldsymbol{x}$

□

Quiz A.5 Does this result generalize any of the results in Ch. 2?

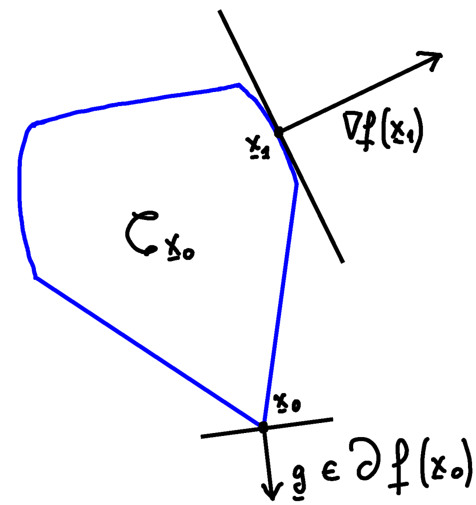


Figure A.1: A subgradient at $\boldsymbol{x}_0$ defines a supporting hyperplane for the sublevel set $\mathcal{C}_{\boldsymbol{x}_0}$.

A.4 Mean value theorem

Proposition A.7

- *If $f : \mathbb{R} \to \mathbb{R}$ is continuously differentiable, then for every a and $b > a$ there exists $\eta \in [a, b]$ such that*

$$\frac{f(b) - f(a)}{b - a} = f'(\eta)$$

- *If $f : \mathbb{R}^D \to \mathbb{R}$ is continuously differentiable, then for every $\boldsymbol{a}$ and $\boldsymbol{b}$ there exists $\alpha \in [0, 1]$ such that*

$$f(\boldsymbol{b}) - f(\boldsymbol{a}) = \nabla f^\top(\alpha \boldsymbol{a} + (1 - \alpha)\boldsymbol{b})(\boldsymbol{b} - \boldsymbol{a})$$

- *If $f : \mathbb{R}^D \to \mathbb{R}$ is twice continuously differentiable, then for every $\boldsymbol{a}$ and $\boldsymbol{b}$ there exists $\alpha \in [0, 1]$ such that*

$$f(\boldsymbol{b}) = f(\boldsymbol{a}) + \nabla f^\top(\boldsymbol{a})(\boldsymbol{b} - \boldsymbol{a}) + \frac{1}{2}(\boldsymbol{b} - \boldsymbol{a})^\top \nabla^2 f(\boldsymbol{\eta})(\boldsymbol{b} - \boldsymbol{a})$$

  for $\boldsymbol{\eta} = \alpha \boldsymbol{a} + (1 - \alpha)\boldsymbol{b}$.

See also [bertsekas1999, Props. A.22 and A.23].

Bibliography

[bertsekas1999] Dimitri P Bertsekas. *Nonlinear programming*. Athena scientific Belmont.

[borden] R. S. Borden. *A Course in Advanced Calculus*. Dover Pubns.

[boyd] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge, UK: Cambridge University Press.

[nesterov2004] Y. Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*. Springer Science & Business Media.