

LECTURE NOTES / 2026-2027

Numerical Optimization

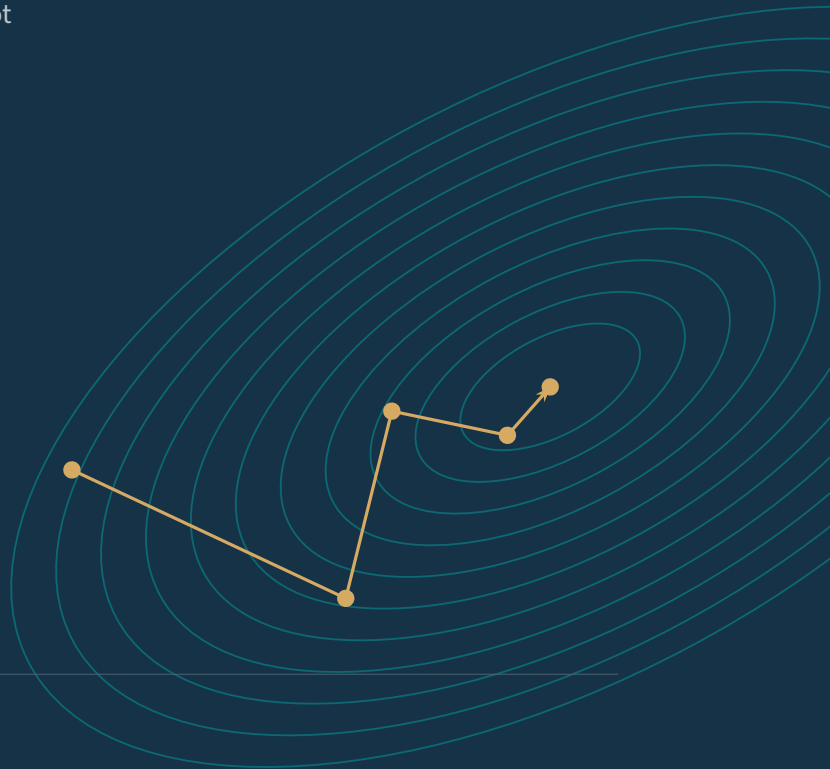
Convexity / Algorithms / Duality

WITH EXERCISES AND WORKED SOLUTIONS

Luca Nenna

Based on the lecture notes of Quentin Mérigot

Université Paris-Saclay / CNRS
Laboratoire de mathématiques d'Orsay, 91405 Orsay, France
luca.nenna@universite-paris-saclay.fr



Contents

Introduction	ii
1 Existence, uniqueness, and convexity	1
1.1 Existence	2
1.2 Elements of differential calculus	3
1.3 Convexity and necessary and sufficient optimality conditions	4
1.4 Strict convexity and uniqueness of the minimizer	7
Exercises: Quadratic functions and least squares	8
2 Gradient descent with exact line search	11
2.1 Descent methods	11
2.2 Gradient descent with exact line search	12
Exercises: Exact line search for a quadratic function	16
3 Gradient descent with exact line search II	17
3.1 Strong convexity and stability of the minimizer	17
3.2 Convergence rate of gradient descent with exact line search	19
3.3 A sufficient condition for strong convexity	20
4 Preconditioned gradient descent with backtracking	21
4.1 Choosing the step by backtracking	21
4.2 Convergence of preconditioned gradient descent with backtracking	23
5 Newton's method	26
5.1 Pure Newton's method	26
5.2 Damped Newton's method	29
Exercises: Logistic regression	32
6 Projection and projected gradient descent	40
6.1 Projection onto a closed convex set	41
6.2 Optimality conditions for constrained optimization	43
Exercises: Projections and constrained minimization	47
7 Optimization with inequality constraints	53
7.1 The penalty method	54
7.2 The Karush–Kuhn–Tucker theorem	56
Exercises: Karush–Kuhn–Tucker conditions	61
8 Lagrangian duality	69
8.1 The dual problem and weak duality	70
8.2 Strong duality	72
8.3 Uzawa's algorithm	73

Introduction

Motivation In many applications, the problem we wish to solve is naturally formulated as an optimization problem:

- In the least-squares method, an overdetermined and/or inconsistent linear system $Ax = b$ (for example, one whose equations contradict each other) is replaced by the optimization problem

$$\min_{x \in \mathbb{R}^N} \|Ax - b\|^2. \quad (1)$$

A minimizer x^* satisfies the equations $Ax = b$ as closely as possible in the least-squares sense. Conversely, when $Ax = b$ has several solutions, we can select one by solving

$$\min_{x \in K} \|x\|^2, \quad K = \{x \in \mathbb{R}^N \mid Ax = b\}. \quad (2)$$

- If $\bar{x} \in \mathbb{R}^N$ is a noisy sampled one-dimensional signal ($\bar{x}_i$ may, for example, be a measurement taken at time t_i), we can denoise it by solving the following optimization problem, where $\lambda > 0$ is a parameter:

$$\min_{x \in \mathbb{R}^N} \|x - \bar{x}\|^2 + \lambda \sum_{1 \leq i \leq N-1} |x_{i+1} - x_i|^2. \quad (3)$$

The aim is to find a vector x^* that balances two requirements: it should be close to $\bar{x}$ (the role of the first term $\|x - \bar{x}\|^2$), and it should be smooth in the sense that successive values x_i and x_{i+1} are close to each other (the role of the second term $\sum_i |x_{i+1} - x_i|^2$).

- In finance, consider portfolio optimization. Given N assets, we seek the fraction $x_i \geq 0$ of the portfolio invested in asset i . Investing the entire portfolio imposes the constraint $\sum_{1 \leq i \leq N} x_i = 1$. This leads to *constrained* optimization problems of the form

$$\min_{x \in \Delta} f(x), \quad \Delta = \left\{ x \in \mathbb{R}^N \mid x_i \geq 0 \text{ for all } i, \quad \sum_i x_i = 1 \right\}. \quad (4)$$

A typical objective is $f(x) = \frac{1}{\varepsilon} |\langle c|x \rangle - r|^2 + \langle x|Qx \rangle$, with $\varepsilon > 0$. Here $c \in \mathbb{R}^N$ contains the expected asset returns, and the first term penalizes deviations of the expected portfolio return $\langle c|x \rangle = \sum_i c_i x_i$ from the target r . The second term measures risk: Q is the symmetric positive semidefinite covariance matrix of the asset returns, and $\langle x|Qx \rangle$ is the variance of the portfolio return.

- In machine learning, many problems can be formulated as optimization problems. Classification using logistic regression or support vector machines provides examples.

For further examples, see the book by Boyd and Vandenberghe, available free of charge online: <https://web.stanford.edu/~boyd/cvxbook/>.

Constrained and unconstrained problems Optimization problems can be divided into two broad classes. In unconstrained optimization, we minimize a function on $\mathbb{R}^d$ or on its *open* domain of definition; problems (1), (3), and logistic regression are of this type. In constrained optimization, we minimize over points of $\mathbb{R}^N$ satisfying equality or inequality constraints; problems (2), (4), and support vector machines are of this type.

Convexity The main focus of this course is *convex* optimization, in which both the objective function and the feasible set are convex. Some of the preliminary results also apply to nonconvex problems. The fundamental reason for emphasizing convexity is that, in a convex optimization problem, every local minimizer is *automatically* a global minimizer.

Existence, uniqueness, and convexity

In this first part, we study the problem of minimizing a function $f \in C^1(\mathbb{R}^d)$:

$$\inf_{x \in \mathbb{R}^d} f(x). \quad (\text{P})$$

Some results below require only continuity, as stated explicitly.

- Definition 1.** (i) The *infimum*, or *optimal value*, of (P) is $\inf_{\mathbb{R}^d} f \in \mathbb{R} \cup \{-\infty\}$.
(ii) A *global minimizer* (or simply a *minimizer*) of (P) is a point $x^* \in \mathbb{R}^d$ such that $f(x^*) = \inf_{\mathbb{R}^d} f$. We denote the set of minimizers, which may be empty, by

$$\arg \min_{\mathbb{R}^d} f = \{x \in \mathbb{R}^d \mid f(x) = \inf_{\mathbb{R}^d} f\}.$$

- (iii) A *minimizing sequence* for (P) is a sequence $(x^{(k)})_{k \geq 0}$ in $\mathbb{R}^d$ such that $\lim_{k \rightarrow +\infty} f(x^{(k)}) = \inf_{x \in \mathbb{R}^d} f(x)$, allowing the limit $-\infty$.

Remark 1. Problem (P) may have no minimizer: consider, for example, $f(x) = \exp(x)$ on $\mathbb{R}$.

Lemma 1. Problem (P) has a minimizing sequence.

Proof. Write $m = \inf_{\mathbb{R}^d} f$. If m is finite, the definition of the infimum provides, for every integer $k \geq 1$, a point $x^{(k)} \in \mathbb{R}^d$ such that

$$m \leq f(x^{(k)}) \leq m + \frac{1}{k}.$$

Hence $f(x^{(k)}) \rightarrow m$. If $m = -\infty$, choose $x^{(k)}$ such that $f(x^{(k)}) \leq -k$; then $f(x^{(k)}) \rightarrow -\infty$. In either case, $x^{(0)}$ can be chosen arbitrarily. ■

1.1 Existence

Proposition 2. Let $f \in C^0(\mathbb{R}^d)$. Suppose there exists $x_0 \in \mathbb{R}^d$ such that the sublevel set $S = \{x \in \mathbb{R}^d \mid f(x) \leq f(x_0)\}$ is compact. Then problem (P) has a global minimizer.

Proof. If $f(x_0) = \inf_{\mathbb{R}^d} f$, then x_0 is already a minimizer. Otherwise, let $(x^{(k)})_{k \geq 0}$ be a minimizing sequence. Since $f(x^{(k)}) \rightarrow \inf_{\mathbb{R}^d} f < f(x_0)$, we have $f(x^{(k)}) \leq f(x_0)$, and hence $x^{(k)} \in S$, for all sufficiently large k . Compactness gives a subsequence $(x^{(\sigma(k))})_{k \geq 0}$ converging to a point $x^\infty \in S$. By continuity and the minimizing property,

$$f(x^\infty) = \lim_{k \rightarrow +\infty} f(x^{(\sigma(k))}) = \inf_{\mathbb{R}^d} f.$$

In particular, this infimum is finite, and x^∞ minimizes f . ■

Corollary 3. Let $f \in C^0(\mathbb{R}^d)$ satisfy

$$\lim_{\|x\| \rightarrow +\infty} f(x) = +\infty,$$

meaning that

$$\forall L \in \mathbb{R}, \exists R \geq 0, \quad \|x\| \geq R \implies f(x) \geq L.$$

Then (P) has a global minimizer. A function satisfying this condition is called coercive.

Proof. Choose any $x_0 \in \mathbb{R}^d$ and let $S = \{x \in \mathbb{R}^d \mid f(x) \leq f(x_0)\}$. It suffices to show that S is compact. It is closed by continuity of f . If it were unbounded, we could choose $x^{(k)} \in S$ with $\|x^{(k)}\| \geq k$ for every $k \geq 1$. Then $\|x^{(k)}\| \rightarrow +\infty$, so coercivity would imply $f(x^{(k)}) \rightarrow +\infty$, contradicting $f(x^{(k)}) \leq f(x_0)$. Thus S is closed and bounded in $\mathbb{R}^d$, and therefore compact. ■

Corollary 4. Let $f \in C^0(\mathbb{R}^d)$ satisfy

$$\forall x \in \mathbb{R}^d, \quad f(x) \geq C \|x\|^p + D,$$

where $C > 0$, $D \in \mathbb{R}$, and $p > 0$. Then (P) has a minimizer.

Proof. Choose any $x_0 \in \mathbb{R}^d$ and let $S = \{x \in \mathbb{R}^d \mid f(x) \leq f(x_0)\}$. This set is closed by continuity. For every $x \in S$,

$$C \|x\|^p + D \leq f(x) \leq f(x_0), \quad \|x\|^p \leq E := \frac{f(x_0) - D}{C}.$$

Here $E \geq 0$ by the assumed lower bound at x_0 . Thus S lies in the ball centered at 0 with radius $E^{1/p}$. It is compact, and Proposition 2 applies. ■

1.2 Elements of differential calculus

Definition 2 (Differentiability). Let V, W be real normed vector spaces, let $\Omega \subset V$ be open, and let $f : \Omega \rightarrow W$. The function f is *differentiable* at $x_0 \in \Omega$ if there exists a continuous linear map $L \in \mathcal{L}(V, W)$ such that

$$f(x_0 + h) = f(x_0) + L(h) + o(\|h\|) \quad (h \rightarrow 0).$$

This unique linear map is denoted by $d_{x_0}f \in \mathcal{L}(V, W)$ and is called the *differential* of f at x_0 .

The function f belongs to $C^1(\Omega)$ if it is differentiable at every point of Ω and the map

$$\begin{aligned} df : \Omega &\longrightarrow (\mathcal{L}(V, W), \|\cdot\|_{\text{op}}) \\ x_0 &\longmapsto d_{x_0}f \end{aligned}$$

is continuous. Here the operator norm is $\|L\|_{\text{op}} = \sup_{\|h\|=1} \|L(h)\|$.

Definition 3 (Directional derivative). Let $f : \Omega \subset V \rightarrow W$, with Ω open, and let $x_0 \in \Omega$ and $h \in V$. When it exists, the limit

$$\lim_{t \rightarrow 0} \frac{f(x_0 + th) - f(x_0)}{t}$$

is called the *directional derivative* of f at x_0 in the direction h .

Remark 2. If f is differentiable at x_0 , it has a directional derivative in every direction $h \in V$, and

$$d_{x_0}f(h) = \lim_{t \rightarrow 0} \frac{f(x_0 + th) - f(x_0)}{t}.$$

The converse is false in general. For example, consider

$$f(x, y) = \begin{cases} y^2/x, & x \neq 0, \\ y, & x = 0. \end{cases}$$

At $(0, 0)$ its directional derivative in direction (a, b) is b^2/a if $a \neq 0$, and b if $a = 0$. Thus every directional derivative exists. However, $f(s^2, s) = 1$ for $s \neq 0$, whereas $f(0, 0) = 0$: the function is not even continuous at the origin, and therefore is not differentiable there.

Definition 4. Let $f \in C^1(\mathbb{R}^d)$. For the usual Euclidean inner product, there is a unique vector $\nabla f(x_0) \in \mathbb{R}^d$, called the *gradient* of f at x_0 , such that

$$d_{x_0}f(h) = \langle \nabla f(x_0), h \rangle \quad \text{for all } h \in \mathbb{R}^d.$$

If $(e_i)_{1 \leq i \leq d}$ is the standard basis of $\mathbb{R}^d$, then

$$\nabla f(x) = \left(\frac{\partial f}{\partial e_i}(x) \right)_{1 \leq i \leq d}, \quad \frac{\partial f}{\partial e_i}(x) = \lim_{\varepsilon \rightarrow 0} \frac{f(x + \varepsilon e_i) - f(x)}{\varepsilon}.$$

From now on, we take $V = \mathbb{R}^d$ and $W = \mathbb{R}$.

Example 1. Consider $f(x) = \|x\|^2$ on $\mathbb{R}^d$. Expanding the squared norm gives $f(x+v) = \|x\|^2 + \langle 2x|v \rangle + \|v\|^2$. Hence $\nabla f(x) = 2x$, in agreement with coordinatewise differentiation.

A function $f : \Omega \subset V \rightarrow W$ is *twice differentiable* at x_0 if it is differentiable in a neighborhood of x_0 and df is differentiable at x_0 . Its second differential $d_{x_0}^2 f$ belongs to $\mathcal{L}(V, \mathcal{L}(V, W))$ and can be viewed as a continuous bilinear map. We write $d_{x_0}^2 f(h, k)$ for the directional derivative at x_0 of $x \mapsto d_x f(h)$ in direction k .

Definition 5. Let $f \in C^2(\Omega)$, where $\Omega \subseteq \mathbb{R}^d$ is open. The *Hessian* of f at $x_0 \in \Omega$ is the matrix of the bilinear form $d_{x_0}^2 f$ in the standard basis. In particular,

$$D^2 f(x) = \left(\frac{\partial^2 f}{\partial e_i \partial e_j}(x) \right)_{1 \leq i, j \leq d},$$

where $(e_i)_{1 \leq i \leq d}$ is the standard basis. Since f is C^2 , this matrix is symmetric.

1.3 Convexity and necessary and sufficient optimality conditions

Convexity

Definition 6. A function $f : \mathbb{R}^d \rightarrow \mathbb{R}$ is *convex* if

$$\forall x, y \in \mathbb{R}^d, \forall t \in [0, 1], \quad f((1-t)x + ty) \leq (1-t)f(x) + tf(y).$$

Exercise 1. Let $f : \mathbb{R}^d \rightarrow \mathbb{R}$ be convex. Show that, for every $C \in \mathbb{R}$, the sublevel set $\{x \in \mathbb{R}^d \mid f(x) \leq C\}$ is convex. Deduce that the set of minimizers of (P) is a closed convex set, possibly empty. For closedness, use the fact that a finite convex function on all of $\mathbb{R}^d$ is continuous.

Proposition 5. Let $f \in C^1(\mathbb{R}^d)$. The following are equivalent:

- (i) f is convex on $\mathbb{R}^d$.
- (ii) For every $x, y \in \mathbb{R}^d$, the function $g : t \in [0, 1] \mapsto f((1-t)x + ty)$ is convex.
- (iii) For every $x, y \in \mathbb{R}^d$, $f(y) \geq f(x) + \langle y - x | \nabla f(x) \rangle$.
- (iv) For every $x, y \in \mathbb{R}^d$, $\langle \nabla f(x) - \nabla f(y) | x - y \rangle \geq 0$.

Lemma 6. Let $f \in C^1(\mathbb{R}^d)$, $x, v \in \mathbb{R}^d$, and define $x_t = x + tv$ and $g(t) = f(x_t)$. Then

$$g'(t) = \langle \nabla f(x_t) | v \rangle. \quad (1.1)$$

Proof. This is the chain rule applied to $t \mapsto x + tv \mapsto f(x + tv)$. ■

Proof of Proposition 5. (i) $\iff$ (ii). This follows directly from the definition, by restricting f to each segment.

(ii) $\implies$ (iii). Let $x, y \in \mathbb{R}^d$ and $g(t) = f(x_t)$, where $x_t = (1-t)x + ty = x + t(y-x)$. For $0 < t \leq 1$, convexity gives

$$g(t) \leq (1-t)g(0) + tg(1), \quad g(1) \geq g(0) + \frac{g(t) - g(0)}{t}.$$

Letting $t \downarrow 0$ and using Lemma 6, we get $f(y) \geq f(x) + \langle \nabla f(x) | y - x \rangle$.

(iii) $\implies$ (iv). Add (iii) to the same inequality with x and y interchanged.

(iv) $\implies$ (ii). Again set $g(t) = f(x_t)$ with $x_t = x + t(y-x)$. For $t > s$, Lemma 6 and (iv) give

$$\begin{aligned} g'(t) - g'(s) &= \langle \nabla f(x_t) - \nabla f(x_s) | y - x \rangle \\ &= \frac{1}{t-s} \langle \nabla f(x_t) - \nabla f(x_s) | x_t - x_s \rangle \geq 0. \end{aligned}$$

Thus g' is nondecreasing, so g is convex. ■

Remark 3. The definition and Proposition 5 also apply on an open convex domain $\Omega \subseteq \mathbb{R}^d$, with all points restricted to Ω . Every segment used in the proof then lies in Ω .

Lemma 7 (Taylor's theorem with Lagrange remainder). Let $f \in C^2(\mathbb{R}^d)$, $x, v \in \mathbb{R}^d$, and $x_t = x + tv$. For every $t \geq 0$, there exists $s \in [0, t]$ such that

$$f(x_t) = f(x) + t \langle \nabla f(x) | v \rangle + \frac{t^2}{2} \langle D^2 f(x_s) v | v \rangle.$$

Lemma 8. Let $\Omega \subseteq \mathbb{R}^d$ be open, $f \in C^2(\Omega)$, $x \in \Omega$, and $v \in \mathbb{R}^d$. Set $x_t = x + tv$ and $g(t) = f(x_t)$ for those t such that $x_t \in \Omega$. Then

$$g''(t) = \langle D^2 f(x_t) v | v \rangle. \quad (1.2)$$

Proof. Write $(e_k)_k$ for the standard basis. In coordinates,

$$g(t) = f \left(x + \sum_k t v_k e_k \right),$$

$$g'(t) = \sum_i v_i \frac{\partial f}{\partial e_i} \left(x + \sum_k t v_k e_k \right) = \langle \nabla f(x_t) | v \rangle,$$

and therefore

$$g''(t) = \sum_i \sum_j v_i v_j \frac{\partial^2 f}{\partial e_j \partial e_i} \left(x + \sum_k t v_k e_k \right) = \langle D^2 f(x_t) v | v \rangle. \quad \blacksquare$$

For symmetric matrices, $A \geq 0$ means that A is positive semidefinite, that is, $\langle Av | v \rangle \geq 0$ for every v . We write $A \leq B$ if $B - A \geq 0$. The matrix A is *positive definite* if $\langle Av | v \rangle > 0$ for every nonzero v . In particular, $m\text{Id} \leq A \leq M\text{Id}$ means $m \|v\|^2 \leq \langle Av | v \rangle \leq M \|v\|^2$ for every v .

Proposition 9. Let $f \in C^2(\mathbb{R}^d)$. If $D^2 f(x) \geq 0$ for all $x \in \mathbb{R}^d$, then f is convex.

Proof of Proposition 9. Let $x, y \in \mathbb{R}^d$ and $g(t) = f((1-t)x + ty)$. Then $g''(t) = \langle D^2 f((1-t)x + ty)(y-x), y-x \rangle \geq 0$. Taylor's theorem gives, for some $s \in (0, 1)$,

$$g(1) = g(0) + g'(0) + \frac{1}{2}g''(s) \geq g(0) + g'(0).$$

Thus $f(y) \geq f(x) + \langle \nabla f(x), y-x \rangle$, and Proposition 5 implies that f is convex. ■

A necessary optimality condition

Theorem 10 (Fermat). *Let $f \in C^1(\mathbb{R}^d)$ and let x^* be a minimizer of (P). Then*

$$\nabla f(x^*) = 0. \quad (\text{Euler's equation})$$

The same conclusion holds at any local minimizer of a differentiable function on an open domain.

Remark 4. The *converse* is false: for $f(x) = x^3$ on $\mathbb{R}$, the point 0 satisfies $f'(0) = 0$ but is not a minimizer, even locally.

Proof. If x^* is a local minimizer in an open domain, then for each standard basis vector e_i there is $\varepsilon_0 > 0$ such that $f(x^* + \varepsilon e_i) \geq f(x^*)$ whenever $|\varepsilon| \leq \varepsilon_0$. Hence

$$\frac{f(x^* + \varepsilon e_i) - f(x^*)}{\varepsilon} \geq 0 \quad (0 < \varepsilon \leq \varepsilon_0).$$

Passing to the limit yields $\frac{\partial f}{\partial e_i}(x^*) \geq 0$. Similarly,

$$\frac{f(x^* + \varepsilon e_i) - f(x^*)}{\varepsilon} \leq 0 \quad (-\varepsilon_0 \leq \varepsilon < 0),$$

so $\frac{\partial f}{\partial e_i}(x^*) \leq 0$. Therefore every partial derivative vanishes. ■

A sufficient optimality condition

Theorem 11. *Let $\Omega \subseteq \mathbb{R}^d$ be open and convex, and let $f \in C^1(\Omega)$ be convex. A point $x^* \in \Omega$ minimizes f on Ω if and only if $\nabla f(x^*) = 0$. For $\Omega = \mathbb{R}^d$, this characterizes the minimizers of (P).*

Proof. The forward implication is Fermat's theorem. Conversely, if $\nabla f(x^*) = 0$, the first-order convexity inequality gives

$$\forall y \in \Omega, \quad f(y) \geq f(x^*) + \langle y - x^*, \nabla f(x^*) \rangle = f(x^*).$$

Thus x^* minimizes f on Ω . ■

1.4 Strict convexity and uniqueness of the minimizer

Definition 7. A function $f : \mathbb{R}^d \rightarrow \mathbb{R}$ is *strictly convex* if

$$\forall x, y \in \mathbb{R}^d, x \neq y, \forall t \in (0, 1), \quad f((1-t)x + ty) < (1-t)f(x) + tf(y).$$

Proposition 12. A strictly convex function $f : \mathbb{R}^d \rightarrow \mathbb{R}$ has at most one minimizer on $\mathbb{R}^d$.

Proof. Suppose, for a contradiction, that f has two distinct minimizers x^* and y^* . Strict convexity gives

$$f\left(\frac{x^* + y^*}{2}\right) < \frac{f(x^*) + f(y^*)}{2} = f(x^*),$$

contradicting the fact that x^* is a minimizer. ■

Remark 5. This proposition does not assert the existence of a minimizer.

Exercises and worked solutions

Quadratic functions and least squares

These exercises translate TD 1, Exercises 0, 1, 2, and 4. Exercise 3, on exact line search, follows Chapter 2. Each solution uses the same part labels as its question.

We study

$$f(x) = \frac{1}{2} \langle x | Qx \rangle + \langle b | x \rangle, \quad x \in \mathbb{R}^N, \quad (1.3)$$

where Q is symmetric and $b \in \mathbb{R}^N$. Positive definiteness is assumed only where stated. Recall that $Q \geq 0$ means $\langle x | Qx \rangle \geq 0$ for every x , whereas $Q > 0$ means $\langle x | Qx \rangle > 0$ for every $x \neq 0$.

TD 1 / Exercise 0

An explicit example

Let $Q = \text{diag}(\lambda_1, \lambda_2)$ and $b = 0$, with $\lambda_1, \lambda_2 \in \mathbb{R}$.

- Show that convexity of f implies $\lambda_1, \lambda_2 \geq 0$.
- Show that f is bounded below if and only if $\lambda_1, \lambda_2 \geq 0$.
- Show that the origin is a critical point, but is a global minimizer if and only if $\lambda_1, \lambda_2 \geq 0$.

Worked solution

Here $f(x) = \frac{1}{2}(\lambda_1 x_1^2 + \lambda_2 x_2^2)$.

- Convexity applied to e_i and $-e_i$ gives $0 = f(0) \leq [f(e_i) + f(-e_i)]/2 = \lambda_i/2$, for $i = 1, 2$.
- If both coefficients are nonnegative, $f \geq 0$. If $\lambda_i < 0$, then $f(te_i) = \lambda_i t^2/2 \rightarrow -\infty$ as $|t| \rightarrow \infty$.
- The gradient is $(\lambda_1 x_1, \lambda_2 x_2)$, so it vanishes at the origin for all choices of coefficients. Since $f(0) = 0$, the origin minimizes f exactly when $f \geq 0$, which is the condition in (b). If a coefficient is zero, the minimizer need not be unique.

TD 1 / Exercise 1

Convexity of a quadratic function

Assume only that Q is symmetric.

- Write f in coordinates and show that $\nabla f(x) = Qx + b$ and $D^2 f(x) = Q$.
- Deduce that f is convex if and only if $Q \geq 0$.
- For $x, y \in \mathbb{R}^N$ and $t \in [0, 1]$, prove

$$\begin{aligned} & f((1-t)x + ty) - (1-t)f(x) - tf(y) \\ &= -\frac{t(1-t)}{2} \langle Qx | x \rangle - \frac{t(1-t)}{2} \langle Qy | y \rangle + t(1-t) \langle Qx | y \rangle \\ &= -\frac{t(1-t)}{2} \langle Q(x-y) | x-y \rangle. \end{aligned}$$

Conclude that f is strictly convex if and only if $Q > 0$.

- Now assume $Q > 0$.
 - Show that $q_Q(x) = \langle x | Qx \rangle$ attains a minimum $m > 0$ on the unit sphere $K = \{x \in \mathbb{R}^N : \|x\| = 1\}$.
 - Deduce that $q_Q(x) \geq m$ on K and $q_Q(x) \geq m \|x\|^2$ on $\mathbb{R}^N$.
 - Prove that f attains its minimum on $\mathbb{R}^N$.

Worked solution

- (a) Matrix multiplication gives

$$f(x) = \frac{1}{2} \sum_{i,j=1}^N Q_{ij} x_i x_j + \sum_{i=1}^N b_i x_i.$$

Differentiating with respect to x_k and using $Q_{ik} = Q_{ki}$,

$$\frac{\partial f}{\partial x_k}(x) = \frac{1}{2} \sum_j Q_{kj} x_j + \frac{1}{2} \sum_i Q_{ik} x_i + b_k = \sum_j Q_{kj} x_j + b_k.$$

A second differentiation gives $\partial_\ell \partial_k f = Q_{k\ell}$.

- (b) The second-order characterization of convexity on $\mathbb{R}^N$ gives the claim, since the Hessian is the constant matrix Q . Alternatively, the first-order remainder is

$$f(y) - f(x) - \langle \nabla f(x) | y - x \rangle = \frac{1}{2} \langle y - x | Q(y - x) \rangle,$$

which is nonnegative for all x, y exactly when $Q \geq 0$.

- (c) Expand the quadratic term at $(1-t)x + ty$. Symmetry makes the two mixed terms equal, while the linear terms cancel. The coefficients of $\langle Qx|x \rangle$ and $\langle Qy|y \rangle$ are both $-t(1-t)/2$, yielding the displayed identities. If $Q > 0$, the remainder is strictly negative for $x \neq y$ and $0 < t < 1$. Conversely, strict convexity applied to $x = h \neq 0$, $y = 0$, and $t = 1/2$ forces $\langle Qh|h \rangle > 0$.
- (d) (i) The unit sphere is nonempty and compact, and q_Q is continuous, so it attains its minimum at some $u \in K$. Positive definiteness gives $m = q_Q(u) > 0$.
- (ii) The first inequality is the definition of m . If $x \neq 0$, apply it to $x/\|x\|$ and use homogeneity: $q_Q(x) = \|x\|^2 q_Q(x/\|x\|) \geq m \|x\|^2$. It also holds for $x = 0$.
- (iii) Cauchy–Schwarz gives

$$f(x) \geq \frac{m}{2} \|x\|^2 - \|b\| \|x\| \longrightarrow +\infty \quad \text{as } \|x\| \rightarrow \infty.$$

Thus f is continuous and coercive. Its minimum is attained by the existence theorem of Chapter 1.

TD 1 / Exercise 2

Characterizing the minimizer

Assume $Q > 0$. Show that f has a unique minimizer x^* , characterized by $Qx^* + b = 0$.

Worked solution

Existence follows from Exercise 1(d), and strict convexity gives uniqueness. The first-order optimality condition is $\nabla f(x^*) = Qx^* + b = 0$. Conversely, this condition suffices because f is convex. Since Q is invertible,

$$x^* = -Q^{-1}b, \quad f(x) = f(x^*) + \frac{1}{2} \langle x - x^* | Q(x - x^*) \rangle.$$

The last term is positive for $x \neq x^*$, also proving uniqueness directly.

TD 1 / Exercise 4

Application to least squares

Let $A \in \mathbb{R}^{M \times N}$ and $y \in \mathbb{R}^M$. Consider

$$\min_{x \in \mathbb{R}^N} f(x), \quad f(x) = \frac{1}{2} \|Ax - y\|^2.$$

- (a) Find Q, b, c such that $f(x) = \frac{1}{2} \langle Qx | x \rangle + \langle b | x \rangle + c$.
- (b) Compute $\nabla f(x)$ and $D^2 f(x)$.
- (c) Show that Q is symmetric positive semidefinite and that x^* minimizes f if and only if $A^T A x^* = A^T y$.
- (d) If A is injective, show that $Q > 0$ and that there is a unique minimizer.

Worked solution

- (a) Expanding the squared norm gives

$$f(x) = \frac{1}{2} x^T A^T A x - x^T A^T y + \frac{1}{2} y^T y, \quad Q = A^T A, \quad b = -A^T y, \quad c = \frac{1}{2} \|y\|^2.$$

- (b) Hence $\nabla f(x) = A^T (Ax - y)$ and $D^2 f(x) = A^T A$.
- (c) We have $(A^T A)^T = A^T A$ and $\langle x | A^T A x \rangle = \|Ax\|^2 \geq 0$. Thus f is convex, and the necessary and sufficient condition $\nabla f(x^*) = 0$ is exactly the normal equation $A^T A x^* = A^T y$. This equation always has a solution: indeed, $\text{Ker}(A^T A) = \text{Ker} A$, so $\text{Im}(A^T A) = (\text{Ker} A)^\perp = \text{Im}(A^T)$, which contains $A^T y$. All minimizers have the form $x^* + z$ with $z \in \text{Ker} A$.
- (d) Injectivity means $Ax \neq 0$ whenever $x \neq 0$. Therefore $\langle x | Qx \rangle = \|Ax\|^2 > 0$, so $Q > 0$. Exercise 2 gives the unique minimizer

$$x^* = (A^T A)^{-1} A^T y.$$

Gradient descent with exact line search

We wish to solve numerically the problem of minimizing a function $f \in C^0(\mathbb{R}^d)$. In general, computing a minimizer, or even the optimal value of (P), exactly is unrealistic, so we seek an *approximation*. Our aim is to construct a sequence $(x^{(k)})_{k \geq 0}$ satisfying one of the following properties:

- (a) The sequence is minimizing for (P), that is, $\lim_{k \rightarrow +\infty} f(x^{(k)}) = \inf_{\mathbb{R}^d} f$.
- (b) The sequence converges to a minimizer of f on $\mathbb{R}^d$.

This chapter focuses on the second property, which, by continuity of f , implies the first.

2.1 Descent methods

Terminology A *descent method* is an iterative procedure intended to construct a sequence satisfying (a) or (b). Typically, it has the form

$$\begin{cases} d^{(k)} = \dots & \text{descent direction,} \\ t^{(k)} = \dots & \text{step size,} \\ x^{(k+1)} = x^{(k)} + t^{(k)} d^{(k)}. \end{cases}$$

The term *descent* means that $f(x^{(k+1)}) \leq f(x^{(k)})$; this property alone does not guarantee convergence to the optimal value. We will consider the following choices:

- (a) The descent direction may be the negative gradient, $d^{(k)} = -\nabla f(x^{(k)})$, giving *gradient descent*. If f is C^2 and the Hessian at the current iterate is positive definite, another choice is $d^{(k)} = -[D^2 f(x^{(k)})]^{-1} \nabla f(x^{(k)})$, giving *Newton's method*. Positive definiteness guarantees that this is a descent direction whenever the gradient is nonzero.
- (b) The step size may be constant ($t^{(k)} = \tau$), optimal as in (2.1), or chosen by more elaborate rules designed to guarantee convergence.

Definition 8. Let $f \in C^0(\mathbb{R}^d)$. A vector v is a *descent direction* at $x \in \mathbb{R}^d$ if

$$\exists \tau > 0, \forall t \in (0, \tau], \quad f(x + tv) < f(x).$$

Exercise 2. Show that if $f \in C^1(\mathbb{R}^d)$ and $\langle v | \nabla f(x) \rangle < 0$, then v is a descent direction.

If f is differentiable at x , then $f(x + tv) = f(x) + t\langle \nabla f(x)|v \rangle + o(t)$ as $t \downarrow 0$. We therefore seek a direction that makes $\langle \nabla f(x)|v \rangle$ as small as possible *among unit vectors*. Its magnitude can be absorbed into the step size, leading to the choice $d^{(k)} = -\nabla f(x^{(k)})$.

Lemma 13. Let $f \in C^1(\mathbb{R}^d)$ and $x_0 \in \mathbb{R}^d$. Then

$$\min_{\|v\|=1} d_{x_0} f(v) = -\|\nabla f(x_0)\|.$$

If $\nabla f(x_0) \neq 0$, the unique minimizer is $v = -\nabla f(x_0)/\|\nabla f(x_0)\|$.

Proof. Cauchy–Schwarz gives

$$d_{x_0} f(v) = \langle \nabla f(x_0)|v \rangle \geq -\|\nabla f(x_0)\| \|v\|.$$

When the gradient is nonzero, equality holds precisely when v is a nonnegative multiple of $-\nabla f(x_0)$. The condition $\|v\| = 1$ then gives the stated vector. If the gradient vanishes, the differential is zero and every unit vector is a minimizer. ■

2.2 Gradient descent with exact line search

Definition 9. Let $f \in C^1(\mathbb{R}^d)$. Starting from $x^{(0)} \in \mathbb{R}^d$, gradient descent with *exact line search* is defined by

$$\begin{cases} d^{(k)} = -\nabla f(x^{(k)}), \\ t^{(k)} \in \arg \min_{t \geq 0} f(x^{(k)} + td^{(k)}), \\ x^{(k+1)} = x^{(k)} + t^{(k)}d^{(k)}, \end{cases} \quad (2.1)$$

whenever the line-search minimum is attained. If $\nabla f(x^{(k)}) = 0$, we stop, or equivalently set $t^{(k)} = 0$ and keep all subsequent iterates equal to $x^{(k)}$.

Remark 6. By construction, the iterates satisfy

$$\begin{aligned} f(x^{(k+1)}) &\leq f(x^{(k)}), \\ \langle \nabla f(x^{(k+1)})|\nabla f(x^{(k)}) \rangle &= 0. \end{aligned}$$

To prove the second identity, set $g(t) = f(x^{(k)} + td^{(k)})$. If $d^{(k)} \neq 0$, then $g'(0) = -\|\nabla f(x^{(k)})\|^2 < 0$, so any minimizing step satisfies $t^{(k)} > 0$. Fermat's theorem therefore gives

$$\begin{aligned} 0 &= g'(t^{(k)}) = \langle \nabla f(x^{(k)} + t^{(k)}d^{(k)})|d^{(k)} \rangle \\ &= -\langle \nabla f(x^{(k+1)})|\nabla f(x^{(k)}) \rangle. \end{aligned}$$

If $d^{(k)} = 0$, the identity is immediate.

Remark 7. Implementing this algorithm requires computing an optimal step size at each iteration, which is itself a one-dimensional optimization problem. An exact solution is available only in a limited number of cases. In general, other step-size rules are preferable.

Example 2. Let $f(x) = \frac{1}{2}\langle Qx|x\rangle + \langle b|x\rangle$, where Q is symmetric positive definite. Then f is strictly convex and $\nabla f(x) = Qx + b$. Given $x^{(k)} \in \mathbb{R}^d$, set $d^{(k)} = -\nabla f(x^{(k)})$. If $d^{(k)} \neq 0$, the function $g(t) = f(x_t)$, with $x_t = x^{(k)} + td^{(k)}$, is a strictly convex quadratic and has a unique minimizer. Since $g'(t) = \langle \nabla f(x_t)|d^{(k)}\rangle$,

$$\begin{aligned} g'(t^{(k)}) = 0 &\iff \langle Q(x^{(k)} + t^{(k)}d^{(k)}) + b|d^{(k)}\rangle = 0 \\ &\iff t^{(k)}\langle Qd^{(k)}|d^{(k)}\rangle - \langle d^{(k)}|d^{(k)}\rangle = 0. \end{aligned}$$

Thus $t^{(k)} = \langle d^{(k)}|d^{(k)}\rangle / \langle Qd^{(k)}|d^{(k)}\rangle > 0$. In summary, the algorithm becomes

$$\begin{cases} d^{(k)} = -(Qx^{(k)} + b), \\ t^{(k)} = \frac{\langle d^{(k)}|d^{(k)}\rangle}{\langle Qd^{(k)}|d^{(k)}\rangle}, \\ x^{(k+1)} = x^{(k)} + t^{(k)}d^{(k)}, \end{cases} \quad d^{(k)} \neq 0. \quad (2.2)$$

If $d^{(k)} = 0$, the current point is the unique minimizer $-Q^{-1}b$, and the algorithm stops.

From now on, we assume that

$$S = \{x \in \mathbb{R}^d \mid f(x) \leq f(x^{(0)})\} \text{ is compact.} \quad (2.3)$$

By Proposition 2, this guarantees that f has a minimizer on $\mathbb{R}^d$.

Lemma 14. Under assumption (2.3), the minimum defining each line-search step is attained.

Proof. Inductively, the preceding descent steps give $x^{(k)} \in S$. Set $g(t) = f(x^{(k)} + td^{(k)})$ for $t \geq 0$. If $d^{(k)} = 0$, then g is constant. Otherwise consider

$$S_g = \{t \geq 0 \mid g(t) \leq f(x^{(k)})\}.$$

This set is nonempty because $0 \in S_g$, and it is closed by continuity. Since S is bounded, there exists $R \geq 0$ such that $\|x\| \leq R$ for every $x \in S$. If $t \in S_g$, then $x^{(k)} + td^{(k)} \in S$, and hence

$$t\|d^{(k)}\| \leq \|x^{(k)} + td^{(k)}\| + \|x^{(k)}\| \leq R + \|x^{(k)}\|.$$

Thus $0 \leq t \leq (R + \|x^{(k)}\|)/\|d^{(k)}\|$. The continuous function g attains its minimum on the compact set S_g . Since $g(t) > g(0)$ outside S_g , this is also its minimum on $[0, +\infty)$. ■

Theorem 15. Let $f \in C^2(\mathbb{R}^d)$ satisfy:

- (i) The sublevel set $S = \{x \in \mathbb{R}^d \mid f(x) \leq f(x^{(0)})\}$ is compact.
- (ii) There exists $M > 0$ such that $D^2f(x) \leq MId$ for every $x \in S$.
- (iii) The function f is strictly convex.

Then the iterates of (2.1) converge to the unique global minimizer of f on $\mathbb{R}^d$.

We will use the following proposition:

Proposition 16. *Let $(x_k)_{k \geq 1}$ be a bounded sequence in $\mathbb{R}^d$ with a unique cluster point $\bar{x}$.^a Then $x_k \rightarrow \bar{x}$ as $k \rightarrow +\infty$.*

^aA point $\bar{x}$ is a cluster point of $(x_k)_{k \geq 1}$ if some subsequence $(x_{\sigma(k)})_{k \geq 1}$ converges to $\bar{x}$.

Proof of Theorem 15. Existence follows from compactness of S , and uniqueness from strict convexity. Denote the minimizer by x^* . If any iterate has zero gradient, it equals x^* by Theorem 11, and the result follows from the stopping convention. We may therefore assume that all gradients are nonzero.

Fix $k \geq 0$ and set

$$a = \nabla f(x^{(k)}), \quad x_t = x^{(k)} - ta, \quad g(t) = f(x_t), \quad t_* = t^{(k)} > 0.$$

Exact line search gives $f(x^{(k+1)}) = g(t_*) = \min_{t \geq 0} g(t)$. The set S is convex because f is convex, and both $x_0 = x^{(k)}$ and $x_{t_*} = x^{(k+1)}$ belong to S . Consequently, $x_t \in S$ for $0 \leq t \leq t_*$. On this interval,

$$g'(0) = -\|a\|^2, \quad g'(t_*) = 0, \quad g''(t) = \langle D^2 f(x_t) a | a \rangle \leq M \|a\|^2.$$

Integrating the last inequality yields

$$\|a\|^2 = g'(t_*) - g'(0) = \int_0^{t_*} g''(s) ds \leq M t_* \|a\|^2.$$

Since $a \neq 0$, we obtain $t_* \geq 1/M$. In particular, the entire trial segment $0 \leq t \leq 1/M$ lies in S , where the Hessian bound is available.

Taylor's theorem gives, for $0 \leq t \leq t_*$ and some $s \in [0, t]$,

$$\begin{aligned} g(t) &= g(0) + t g'(0) + \frac{t^2}{2} g''(s) \\ &= f(x^{(k)}) - t \|a\|^2 + \frac{t^2}{2} \langle D^2 f(x_s) a | a \rangle \\ &\leq f(x^{(k)}) + \left(\frac{M}{2} t^2 - t \right) \|a\|^2. \end{aligned}$$

The right-hand side is minimized at $t = 1/M$, now known to be admissible. Therefore

$$f(x^{(k+1)}) = g(t_*) \leq g(1/M) \leq f(x^{(k)}) - \frac{1}{2M} \|\nabla f(x^{(k)})\|^2,$$

or equivalently

$$\|\nabla f(x^{(k)})\|^2 \leq 2M(f(x^{(k)}) - f(x^{(k+1)})). \quad (2.4)$$

Summing from $k = 0$ to K gives

$$\begin{aligned} \sum_{k=0}^K \|\nabla f(x^{(k)})\|^2 &\leq 2M(f(x^{(0)}) - f(x^{(K+1)})) \\ &\leq 2M(f(x^{(0)}) - \inf_{\mathbb{R}^d} f). \end{aligned}$$

Hence the series of squared gradient norms converges, and $\|\nabla f(x^{(k)})\| \rightarrow 0$.

Finally, all iterates lie in the compact set S , so the sequence is bounded and has a cluster point. By Proposition 16, it suffices to prove that every cluster point equals x^* . Let $x^{(\sigma(k))} \rightarrow \bar{x} \in S$. By continuity of the gradient,

$$\nabla f(\bar{x}) = \lim_{k \rightarrow +\infty} \nabla f(x^{(\sigma(k))}) = 0.$$

Convexity and Theorem 11 show that $\bar{x}$ is a minimizer. Uniqueness gives $\bar{x} = x^*$, proving convergence. ■

Exercises and worked solutions

Exact line search for a quadratic function

TD 1 / Exercise 3

Gradient descent with an optimal step

Let $Q \in \mathbb{R}^{N \times N}$ be symmetric positive definite and

$$f(x) = \frac{1}{2} \langle Qx | x \rangle + \langle b | x \rangle.$$

Starting from $x^{(0)} \in \mathbb{R}^N$, exact-line-search gradient descent uses

$$d^{(k)} = -\nabla f(x^{(k)}), \quad t^{(k)} \in \arg \min_{t \in \mathbb{R}} f(x^{(k)} + t d^{(k)}), \quad x^{(k+1)} = x^{(k)} + t^{(k)} d^{(k)}.$$

If $d^{(k)} = 0$, the method stops: $x^{(k)}$ is already the minimizer.

- (a) Fix $x \in \mathbb{R}^N$ and a *nonzero* vector $v \in \mathbb{R}^N$.
 - (i) Show that $g(t) = f(x + tv)$ is a degree-two polynomial with positive leading coefficient.
 - (ii) Find its unique minimizer $t^* \in \mathbb{R}$.
- (b) Deduce, whenever $d^{(k)} \neq 0$, the formulas

$$\begin{cases} d^{(k)} = -(Qx^{(k)} + b), \\ t^{(k)} = \frac{\|d^{(k)}\|^2}{\langle d^{(k)} | Q d^{(k)} \rangle}, \\ x^{(k+1)} = x^{(k)} + t^{(k)} d^{(k)}. \end{cases}$$

Worked solution

- (a) (i) Expansion using the symmetry of Q yields

$$g(t) = f(x) + t \langle Qx + b | v \rangle + \frac{t^2}{2} \langle Qv | v \rangle.$$

Since $v \neq 0$ and $Q > 0$, the coefficient $\frac{1}{2} \langle Qv | v \rangle$ is strictly positive.

- (ii) Differentiating gives $g'(t) = \langle Qx + b | v \rangle + t \langle Qv | v \rangle$. Its unique zero is

$$t^* = -\frac{\langle Qx + b | v \rangle}{\langle Qv | v \rangle},$$

which is the unique global minimizer since $g'' > 0$.

- (b) Substitute $v = d^{(k)} = -(Qx^{(k)} + b)$ in (a). The numerator becomes $\|d^{(k)}\|^2$, and the denominator is positive, so $t^{(k)} > 0$. Thus minimizing over all $t \in \mathbb{R}$ gives the same step as minimizing over $t \geq 0$ in this descent direction. If $d^{(k)} = 0$, the line-search objective is constant and every step is optimal; the quotient would be $0/0$ and must not be used.

Gradient descent with exact line search II

3.1 Strong convexity and stability of the minimizer

Motivation. Let $f \in C^0(\mathbb{R}^d)$ attain its minimum at $x^* \in \mathbb{R}^d$. Proposition 12 shows that a strictly convex function has at most one minimizer. Equivalently,

$$f(x) \leq f(x^*) \implies x = x^*.$$

In practice, our algorithms usually compute approximations rather than the exact minimizer. For instance, they may produce a sequence $x^{(k)}$ satisfying $f(x^{(k)}) \leq f(x^*) + \varepsilon_k$, where $\varepsilon_k \geq 0$ and $\varepsilon_k \rightarrow 0$. We would like to conclude that $x^{(k)}$ is close to x^* , with a bound of the form

$$\|x^{(k)} - x^*\| \leq C \varepsilon_k^\alpha$$

for some exponent $\alpha > 0$ and constant $C > 0$. We will prove such a bound with $\alpha = 1/2$ when f is *strongly convex*.

Definition 10. A subset $X \subseteq \mathbb{R}^d$ is *convex* if it contains every line segment joining two of its points, that is,

$$\forall x, y \in X, \quad \forall t \in [0, 1], \quad (1-t)x + ty \in X.$$

Definition 11. A function $f : \mathbb{R}^d \rightarrow \mathbb{R}$ is *m-strongly convex* on a convex set $X \subseteq \mathbb{R}^d$, where $m > 0$, if the restriction of $f - \frac{m}{2} \|\cdot\|^2$ to X is convex.

Exercise 3. Show that an *m-strongly convex* function is strictly convex.

Proposition 17. Let $f \in C^1(\mathbb{R}^d)$, let $X \subseteq \mathbb{R}^d$ be convex, and let $m > 0$. Then $(i) \implies (ii) \iff (iii) \iff (iv)$, where

- (i) $f \in C^2(\mathbb{R}^d)$ and $\forall x \in X, D^2 f(x) \geq m \text{Id}$;
- (ii) f is *m-strongly convex* on X ;
- (iii) $\forall x, y \in X, \quad f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{m}{2} \|x - y\|^2$;
- (iv) $\forall x, y \in X, \quad \langle \nabla f(y) - \nabla f(x), y - x \rangle \geq m \|x - y\|^2$.

Proof. Apply the characterizations of convexity in Propositions 5 and 9 to $g(x) = f(x) - \frac{m}{2} \|x\|^2$. Those proofs use only segments in X , so the same characterizations apply to the restriction to X . Use $\nabla g(x) = \nabla f(x) - mx$ and, when f is twice continuously differentiable, $D^2g(x) = D^2f(x) - m\text{Id}$. ■

Example 3. The function $t \mapsto \exp(t)$ is not strongly convex on $\mathbb{R}$, but it is m -strongly convex on every interval $[a, b]$, with $m = \min_{t \in [a, b]} \exp(t) = \exp(a)$. Likewise, $t \mapsto t^4$ is not strongly convex on $\mathbb{R}$, but it is strongly convex on every closed interval that does not contain the origin.

Corollary 18. Let $f \in C^1(\mathbb{R}^d)$ have a minimizer $x^* \in \mathbb{R}^d$. Suppose that f is m -strongly convex on a convex set S containing x^* . Then, for every $x \in S$,

- (i) $\|x - x^*\|^2 \leq \frac{2}{m}(f(x) - f(x^*));$
- (ii) $\|x - x^*\| \leq \frac{1}{m} \|\nabla f(x)\|;$
- (iii) $f(x) - f(x^*) \leq \frac{1}{2m} \|\nabla f(x)\|^2.$

Remark 8. Part (i) immediately gives uniqueness of the minimizer in S : if $f(x) = f(x^*)$, then $x = x^*$. It also gives a quantitative stability statement: if $f(x) \leq f(x^*) + \varepsilon$, then

$$\|x - x^*\| \leq \sqrt{2\varepsilon/m}.$$

Similarly, part (ii) shows that an approximate first-order optimality condition gives an approximation of the minimizer: $\|\nabla f(x)\| \leq \varepsilon$ implies $\|x - x^*\| \leq \varepsilon/m$.

Example 4. For $f(t) = t^4$, the minimizer is $x^* = 0$. The inequality in (i) would read $t^2 \leq \frac{2}{m}t^4$. No $m > 0$ makes this inequality hold for every $t \in \mathbb{R}$, or even throughout a neighborhood of 0.

Proof of Corollary 18. Since x^* is an unconstrained minimizer, $\nabla f(x^*) = 0$. Proposition 17(iii), applied at x^* , gives

$$f(x) \geq f(x^*) + \langle \nabla f(x^*) | x - x^* \rangle + \frac{m}{2} \|x - x^*\|^2 = f(x^*) + \frac{m}{2} \|x - x^*\|^2,$$

which proves (i). For (ii), use Proposition 17(iv) and Cauchy–Schwarz:

$$m \|x - x^*\|^2 \leq \langle \nabla f(x) | x - x^* \rangle \leq \|\nabla f(x)\| \|x - x^*\|.$$

If $x \neq x^*$, divide by $\|x - x^*\|$; otherwise the claim is immediate.

For (iii), Proposition 17(iii) gives, for $x, y \in S$,

$$f(y) \geq Q(x, y) := f(x) + \langle \nabla f(x) | y - x \rangle + \frac{m}{2} \|y - x\|^2.$$

Fix x and minimize $z \mapsto Q(x, z)$ over all of $\mathbb{R}^d$. Its unique minimizer is $z^* = x - \frac{1}{m} \nabla f(x)$, and hence

$$\forall y \in S, \quad f(y) \geq \inf_{z \in \mathbb{R}^d} Q(x, z) = f(x) - \frac{1}{2m} \|\nabla f(x)\|^2.$$

Taking $y = x^*$ proves the claim. The auxiliary minimizer z^* need not belong to S . ■

3.2 Convergence rate of gradient descent with exact line search

Definition 12 (Linear convergence). Let $(u_k)_{k \geq 0}$ converge to u^* . The sequence converges linearly to u^* if there exist $0 \leq \kappa < 1$ and $k_0 \in \mathbb{N}$ such that

$$\forall k \geq k_0, \quad \|u_{k+1} - u^*\| \leq \kappa \|u_k - u^*\|.$$

Theorem 19. Let $f \in C^2(\mathbb{R}^d)$ satisfy

- (i) the sublevel set $S = \{x \in \mathbb{R}^d \mid f(x) \leq f(x^{(0)})\}$ is compact and convex;
- (ii) there exists $M > 0$ such that $\forall x \in S, D^2 f(x) \leq M \text{Id}$;
- (iii) there exists $m > 0$ such that $\forall x \in S, D^2 f(x) \geq m \text{Id}$.

Then the iterates of algorithm (2.1) converge to the unique global minimizer x^* of f on $\mathbb{R}^d$. Moreover, with $c = M/m \geq 1$,

$$f(x^{(k+1)}) - f(x^*) \leq \left(1 - \frac{1}{c}\right) (f(x^{(k)}) - f(x^*)). \quad (3.1)$$

Remark 9. Thus the function values converge linearly to $f(x^*)$. The convexity assumption on S is essential when Hessian bounds are imposed only on S : positive curvature in separate components of a sublevel set does not rule out several local minima. Convexity of f on $\mathbb{R}^d$ is one sufficient condition for S to be convex.

Remark 10. If $m \text{Id} \leq D^2 f(x) \leq M \text{Id}$ on the relevant convex region, we say that the condition number of f is bounded by $c = M/m$. Inequality (3.1) explains why this ratio matters for the convergence rate. Write $\Delta_0 = f(x^{(0)}) - f(x^*)$. If $c > 1$ and $0 < \varepsilon < \Delta_0$, it suffices to take

$$k \geq \left\lceil \frac{\log(\Delta_0/\varepsilon)}{\log(1/(1-1/c))} \right\rceil$$

to guarantee $f(x^{(k)}) - f(x^*) \leq \varepsilon$. Ignoring integer rounding, this bound is asymptotic to

$$c \log(\Delta_0/\varepsilon) \quad \text{as } c \rightarrow +\infty.$$

For a fixed error reduction, the iteration bound therefore grows proportionally to the condition number when c is large. If $\Delta_0 \leq \varepsilon$, no iteration is needed; if $c = 1$, (3.1) gives the minimum value after at most one iteration.

Proof of Theorem 19. Compactness of S gives a minimizer x^* on S . It is a global minimizer, since outside S the function is larger than $f(x^{(0)})$. The lower Hessian bound and convexity of S imply strong convexity on S , so x^* is unique. Exact line search is well defined and keeps every iterate in S .

The descent estimate (2.4) uses only compactness and convexity of S and the upper Hessian bound along the descent segment, so it applies here. Combining it with Corollary 18(iii) gives

$$2m(f(x^{(k)}) - f(x^*)) \leq \|\nabla f(x^{(k)})\|^2 \leq 2M(f(x^{(k)}) - f(x^{(k+1)})).$$

Rearranging proves (3.1). In particular, the function values converge to $f(x^*)$. Corollary 18(i) then gives

$$\|x^{(k)} - x^*\|^2 \leq \frac{2}{m} (f(x^{(k)}) - f(x^*)) \longrightarrow 0.$$

■

3.3 A sufficient condition for strong convexity

Proposition 20. *Let $f \in C^2(\mathbb{R}^d)$ and let $S \subseteq \mathbb{R}^d$ be a nonempty compact set such that $D^2 f(x)$ is positive definite for every $x \in S$. Then there exist $M \geq m > 0$ such that*

$$\forall x \in S, \quad m \text{Id} \leq D^2 f(x) \leq M \text{Id}.$$

Proof. The set $K = \{(x, v) \in S \times \mathbb{R}^d \mid \|v\| = 1\}$ is compact. The continuous function $(x, v) \mapsto \langle D^2 f(x) v | v \rangle$ attains its minimum m and maximum M on K . Its minimum is strictly positive: at a minimizing pair (x_m, v_m) , positive definiteness and $\|v_m\| = 1$ give

$$m = \langle D^2 f(x_m) v_m | v_m \rangle > 0.$$

For $w \neq 0$, applying the bounds to $v = w/\|w\|$ yields

$$m \|w\|^2 \leq \langle D^2 f(x) w | w \rangle \leq M \|w\|^2.$$

The same inequalities are immediate for $w = 0$. ■

This proposition gives a version of Theorem 19 that does not require explicit constants.

Corollary 21. *Let $f \in C^2(\mathbb{R}^d)$ satisfy*

- (i) *the sublevel set $S = \{x \in \mathbb{R}^d \mid f(x) \leq f(x^{(0)})\}$ is compact and convex;*
- (ii) *$D^2 f(x)$ is positive definite for every $x \in S$.*

Then the iterates of algorithm (2.1) converge to the unique global minimizer x^* of f on $\mathbb{R}^d$, and the sequence of function values converges linearly to $f(x^*)$.

Preconditioned gradient descent with backtracking

One drawback of gradient descent with exact line search is that choosing the step requires solving a one-dimensional optimization problem at every iteration. This chapter introduces a simple rule for selecting the step that also gives convergence guarantees. We will also allow directions other than $d^{(k)} = -\nabla f(x^{(k)})$, in preparation for Newton's method in the next chapter. We consider directions of the form

$$d^{(k)} = -B^{(k)}\nabla f(x^{(k)}), \quad (4.1)$$

where $B^{(k)}$ is symmetric positive definite. At a nonstationary point, this guarantees

$$\langle \nabla f(x^{(k)}) | d^{(k)} \rangle = -\langle \nabla f(x^{(k)}) | B^{(k)} \nabla f(x^{(k)}) \rangle < 0,$$

so $d^{(k)}$ is a descent direction. If the gradient vanishes, we stop the algorithm, or equivalently keep all subsequent iterates equal to the current point.

Remark 11. A descent method with directions given by (4.1) is called *preconditioned gradient descent*, and $B^{(k)}$ is called the *preconditioner*.

4.1 Choosing the step by backtracking

Naive backtracking

Since we want a descent method, a natural first attempt is

$$t^{(k)} = \max\{2^{-\ell} \mid \ell \in \mathbb{N}, \quad f(x^{(k)} + 2^{-\ell} d^{(k)}) \leq f(x^{(k)})\}.$$

Here and below, the candidate steps start with $\ell = 0$. In other words, given $x^{(k)}$ and $d^{(k)}$, test $1, 1/2, 1/4, \dots$ and stop as soon as the function value does not increase. Despite its simplicity, this rule need not converge. For $f(x) = x^2$ on $\mathbb{R}$, $x^{(0)} = 1$, and $d^{(k)} = -f'(x^{(k)}) = -2x^{(k)}$, the step $t^{(k)} = 1$ is always accepted. Thus $x^{(k)} = (-1)^k$, which does not converge to the minimizer.

Armijo backtracking

The idea is to require a sufficient decrease, measured against the first-order prediction. This leads to the *Armijo condition* (4.2).

Lemma 22. Let $f \in C^1(\mathbb{R}^d)$, $x, v \in \mathbb{R}^d$, with $\langle \nabla f(x)|v \rangle < 0$, and let $0 < \alpha < 1/2$. There exists $t_0 > 0$ such that, for every $t \in [0, t_0]$,

$$f(x + tv) \leq f(x) + \alpha t \langle \nabla f(x)|v \rangle. \quad (4.2)$$

Proof. As $t \rightarrow 0^+$,

$$f(x + tv) = f(x) + \alpha t \langle \nabla f(x)|v \rangle + (1 - \alpha)t \langle \nabla f(x)|v \rangle + o(t).$$

The coefficient of t in the last two terms is strictly negative, so their sum is nonpositive for sufficiently small positive t . The condition also holds at $t = 0$. ■

Definition 13 (Armijo step). Let $f \in C^1(\mathbb{R}^d)$. Given $x^{(k)}$ and a descent direction $d^{(k)}$, the *Armijo step* with parameters $0 < \alpha < 1/2$ and $0 < \beta < 1$ is

$$t^{(k)} = \max\{\beta^\ell \mid \ell \in \mathbb{N}, f(x^{(k)} + \beta^\ell d^{(k)}) \leq f(x^{(k)}) + \alpha \beta^\ell \langle \nabla f(x^{(k)})|d^{(k)} \rangle\}. \quad (4.3)$$

At a stationary point, set $d^{(k)} = 0$ and $t^{(k)} = 1$.

Remark 12. The preceding lemma ensures that the following procedure terminates:

```

armijo_step( $x^{(k)}, d^{(k)}$ ) :
     $t \leftarrow 1$ 
    while  $f(x^{(k)} + t d^{(k)}) > f(x^{(k)}) + \alpha t \langle \nabla f(x^{(k)})|d^{(k)} \rangle$  :
         $t \leftarrow \beta t$ 
    return  $t$ .
```

When f is defined only on an open set Ω , a trial point outside Ω is rejected. Equivalently, use the value $+\infty$ there when testing the condition. Since the current point belongs to Ω , sufficiently small steps lie in the domain and the same termination argument applies.

Definition 14 (Descent method with an Armijo step). Preconditioned gradient descent with Armijo backtracking is the iterative method

$$\begin{cases} d^{(k)} = -B^{(k)} \nabla f(x^{(k)}), \\ t^{(k)} \text{ is given by (4.3),} \\ x^{(k+1)} = x^{(k)} + t^{(k)} d^{(k)}, \end{cases}$$

where $B^{(k)} = (A^{(k)})^2$ and $A^{(k)}$ is symmetric positive definite. Every symmetric positive definite matrix $B^{(k)}$ has such a square root.

4.2 Convergence of preconditioned gradient descent with backtracking

Theorem 23. Let $\Omega \subseteq \mathbb{R}^d$ be open and convex, let $x^{(0)} \in \Omega$, and let $f \in C^2(\Omega)$ satisfy

- (i) the sublevel set $S = \{x \in \Omega \mid f(x) \leq f(x^{(0)})\}$ is compact and convex;
- (ii) there exist $0 < \lambda \leq \Lambda$ such that

$$\forall x \in S, \quad \forall k \in \mathbb{N}, \quad \lambda \text{Id} \leq A^{(k)} D^2 f(x) A^{(k)} \leq \Lambda \text{Id}.$$

Then the iterates in Definition 14 converge to the unique global minimizer x^* of f on Ω . Moreover, with $c = 2\alpha\lambda \min(1, \beta/\Lambda) \in (0, 1)$,

$$f(x^{(k+1)}) - f(x^*) \leq (1 - c)(f(x^{(k)}) - f(x^*)).$$

Remark 13. This analysis suggests choosing $B^{(k)}$ so that the transformed Hessians have a condition number Λ/λ close to 1. The overall scale also matters because the trial step starts at 1. For example, for $f(x_1, x_2) = Kx_1^2 + x_2^2$, with $K > 0$, the Hessian is $H = 2 \text{diag}(K, 1)$. The ideal choice is

$$A = \text{diag}((2K)^{-1/2}, 2^{-1/2}), \quad B = A^2 = H^{-1},$$

for which $AHA = \text{Id}$. More generally, when the Hessian is positive definite, taking $B^{(k)} = [D^2 f(x^{(k)})]^{-1}$ gives Newton's method.

Lemma 24. Let $f \in C^2(\Omega)$, where Ω is open, and let $S = \{x \in \Omega \mid f(x) \leq f(x^{(0)})\}$ be compact. Suppose $x^{(k)} \in S$, $B^{(k)} = (A^{(k)})^2$, with $A^{(k)}$ symmetric positive definite, and

$$\forall x \in S, \quad A^{(k)} D^2 f(x) A^{(k)} \leq \Lambda \text{Id}, \quad \Lambda > 0.$$

Then the Armijo step satisfies

$$t^{(k)} \geq \min\left(1, \frac{\beta}{\Lambda}\right).$$

Remark 14. Every trial step $t \leq 1/\Lambda$ satisfies the Armijo condition. Thus the number of step reductions is at most

$$N = \max\left\{0, \left\lceil \frac{\log \Lambda}{\log(1/\beta)} \right\rceil\right\},$$

and at most $N + 1$ trial steps are evaluated, including $t = 1$.

Proof. Write $A = A^{(k)}$, $g = \nabla f(x^{(k)})$, $d = -A^2 g$, $q = \|Ag\|^2$, and $x_t = x^{(k)} + td$. If $g = 0$, then $d = 0$ and $t^{(k)} = 1$, so the claim is immediate. Assume $g \neq 0$; then $q > 0$ and $\langle g|d \rangle = -q$.

Whenever the segment $\{x_s : 0 \leq s \leq t\}$ is contained in S , Taylor's theorem and the Hessian bound give

$$\begin{aligned} f(x_t) &= f(x^{(k)}) + t\langle g|d \rangle + \frac{t^2}{2} \langle D^2 f(x_s) d|d \rangle \quad \text{for some } s \in [0, t], \\ \langle D^2 f(x_s) d|d \rangle &= \langle A D^2 f(x_s) A (Ag) | Ag \rangle \leq \Lambda q, \end{aligned}$$

and therefore

$$f(x_t) \leq f(x^{(k)}) - tq + \frac{\Lambda t^2}{2}q.$$

We must justify applying this bound to $t = 1/\Lambda$, since the Hessian bound is assumed only on S . Initially the ray strictly decreases f . Let $\tau > 0$ be its first exit time from S , that is, the largest time for which $x_s \in S$ for every $0 \leq s \leq \tau$. Compactness of S , together with $d \neq 0$, ensures that τ is finite and $x_\tau \in S \subseteq \Omega$. Continuity gives $f(x_\tau) = f(x^{(0)}) \geq f(x^{(k)})$. Applying the bound at τ yields

$$0 \leq f(x_\tau) - f(x^{(k)}) \leq -\tau q + \frac{\Lambda \tau^2}{2}q,$$

hence $\tau \geq 2/\Lambda$. In particular, every $0 \leq t \leq 1/\Lambda$ stays in S and satisfies

$$f(x_t) \leq f(x^{(k)}) - \frac{t}{2}q \leq f(x^{(k)}) - \alpha tq.$$

Thus all such steps satisfy Armijo. If the first accepted step is 1, the claimed bound holds. Otherwise the preceding rejected step is $t^{(k)}/\beta > 1/\Lambda$, giving $t^{(k)} > \beta/\Lambda$. ■

Lemma 25. Let $X \subseteq \Omega \subseteq \mathbb{R}^d$, where X is convex and Ω is open. Let $f \in C^2(\Omega)$, and suppose that $x^* \in X$ minimizes f on X . If A is symmetric positive definite and

$$\forall x \in X, \quad AD^2f(x)A \geq \lambda \text{Id}, \quad \lambda > 0,$$

then, for every $x \in X$,

$$2\lambda(f(x) - f(x^*)) \leq \|A\nabla f(x)\|^2.$$

Proof. Consider $g(y) = f(Ay)$ on $A^{-1}\Omega$. Its gradient and Hessian are

$$\nabla g(y) = A\nabla f(Ay), \quad D^2g(y) = AD^2f(Ay)A.$$

Hence g is λ -strongly convex on the convex set $A^{-1}X$. For $y = A^{-1}x$ and $y^* = A^{-1}x^*$, the quadratic lower bound gives

$$\begin{aligned} g(y^*) &\geq g(y) + \langle \nabla g(y), y^* - y \rangle + \frac{\lambda}{2} \|y^* - y\|^2 \\ &\geq g(y) - \frac{1}{2\lambda} \|\nabla g(y)\|^2. \end{aligned}$$

The last step completes the square, as in the proof of Corollary 18(iii). Substituting the expressions for g and ∇g proves the claim. ■

Proof of Theorem 23. The minimum on the compact set S is attained and is a global minimum on Ω . The lower transformed Hessian bound, for example with $k = 0$, implies

$$D^2f(x) \geq \lambda(A^{(0)})^{-2} \geq \frac{\lambda}{\|A^{(0)}\|^2} \text{Id} \quad (x \in S),$$

where the matrix norm is the Euclidean operator norm. Thus f is m -strongly convex on S for $m = \lambda/\|A^{(0)}\|^2 > 0$, and its global minimizer x^* is unique. Since Ω is open, $\nabla f(x^*) = 0$.

Armijo backtracking is well defined and keeps every iterate in S . By its sufficient-decrease condition and Lemma 24,

$$\begin{aligned} f(x^{(k+1)}) &\leq f(x^{(k)}) + \alpha t^{(k)} \langle \nabla f(x^{(k)}) | d^{(k)} \rangle \\ &= f(x^{(k)}) - \alpha t^{(k)} \|A^{(k)} \nabla f(x^{(k)})\|^2 \\ &\leq f(x^{(k)}) - \varepsilon \|A^{(k)} \nabla f(x^{(k)})\|^2, \quad \varepsilon = \alpha \min(1, \beta/\Lambda). \end{aligned}$$

The preceding lemma, with $X = S$ and $A = A^{(k)}$, now yields

$$f(x^{(k+1)}) - f(x^*) \leq (1 - 2\lambda\varepsilon)(f(x^{(k)}) - f(x^*)).$$

Moreover, $0 < c = 2\lambda\varepsilon \leq 2\alpha\beta\lambda/\Lambda < 1$. Hence the function values converge to $f(x^*)$. Finally, strong convexity and $\nabla f(x^*) = 0$ imply

$$\|x^{(k)} - x^*\|^2 \leq \frac{2}{m} (f(x^{(k)}) - f(x^*)) \longrightarrow 0,$$

which proves convergence of the iterates. ■

Newton's method

5.1 Pure Newton's method

Constructing the iterates

Let $f \in C^2(\mathbb{R}^d)$, and suppose that $D^2 f(x) > 0$ for every $x \in \mathbb{R}^d$. Newton's method is based on the second-order Taylor expansion at the current iterate: $f(x^{(k)} + d) = g(d) + o(\|d\|^2)$, where

$$g(d) = f(x^{(k)}) + \langle \nabla f(x^{(k)}) | d \rangle + \frac{1}{2} \langle D^2 f(x^{(k)}) d | d \rangle.$$

Since the Hessian is positive definite, this quadratic model is strictly convex and has a unique minimizer $d^{(k)}$, characterized by

$$\nabla g(d^{(k)}) = D^2 f(x^{(k)}) d^{(k)} + \nabla f(x^{(k)}) = 0.$$

Thus $d^{(k)} = -[D^2 f(x^{(k)})]^{-1} \nabla f(x^{(k)})$.

Definition 15. The iterates of pure Newton's method are given by

$$\begin{cases} d^{(k)} = -[D^2 f(x^{(k)})]^{-1} \nabla f(x^{(k)}), \\ t^{(k)} = 1, \\ x^{(k+1)} = x^{(k)} + t^{(k)} d^{(k)}. \end{cases}$$

The word “pure” refers to the full step $t^{(k)} = 1$, in contrast to the damped Newton method introduced in Section 5.2.

Remark 15. The vector $d^{(k)}$ is the *Newton direction*. At a nonstationary point, it is a descent direction because

$$\langle \nabla f(x^{(k)}) | d^{(k)} \rangle = -\langle \nabla f(x^{(k)}) | [D^2 f(x^{(k)})]^{-1} \nabla f(x^{(k)}) \rangle < 0.$$

Nevertheless, the full step can increase the function value. Pure Newton's method is therefore not necessarily a descent method.

Remark 16. In dimension $d = 1$, set $h = f'$. Newton's iteration becomes

$$x^{(k+1)} = x^{(k)} - \frac{f'(x^{(k)})}{f''(x^{(k)})} = x^{(k)} - \frac{h(x^{(k)})}{h'(x^{(k)})}.$$

This is the classical Newton iteration for finding a zero of h . Here, since f is strictly convex, a zero of f' is its unique global minimizer.

Local quadratic convergence

Definition 16 (Quadratic convergence). Let $(u_k)_{k \geq 0}$ converge to u^* . The sequence converges *quadratically* to u^* if there exist $\gamma > 0$ and $k_0 \in \mathbb{N}$ such that

$$\forall k \geq k_0, \quad \|u_{k+1} - u^*\| \leq \gamma \|u_k - u^*\|^2.$$

Theorem 26. Let $f \in C^3(\mathbb{R}^d)$ satisfy

- (i) f has a minimizer x^* on $\mathbb{R}^d$;
- (ii) $D^2 f(x) > 0$ for every $x \in \mathbb{R}^d$.

There exists $r > 0$ such that, for every $x^{(0)} \in B(x^*, r)$, pure Newton's method is well defined for all $k \geq 0$ and converges quadratically to x^* .

Remark 17. Suppose the quadratic estimate holds from $k = 0$, and define the scaled error $\varepsilon_k = \gamma \|x^{(k)} - x^*\|$. Then

$$\varepsilon_{k+1} \leq \varepsilon_k^2, \quad \varepsilon_k \leq \varepsilon_0^{2^k}.$$

When $0 < \varepsilon_0 < 1$, the error bound decreases very rapidly. For example, if $\varepsilon_0 = 1/2$ and the target is $\varepsilon_k \leq \eta = 10^{-15}$, then $\eta \geq 2^{-50}$, so it suffices that $2^k \geq 50$. Six iterations are enough for this bound on the scaled error.

Proof. Strict convexity makes x^* unique, and $\nabla f(x^*) = 0$. Fix $R > 0$ and set $K = \overline{B(x^*, R)}$. By Proposition 20, there exists $m > 0$ such that $D^2 f(x) \geq m \text{Id}$ on K . Since $f \in C^3$, its third derivative is bounded on this compact convex set. Thus there is a constant $L > 0$ such that

$$\forall x, x' \in K, \quad \|D^2 f(x) - D^2 f(x')\| \leq L \|x - x'\|,$$

with the Euclidean operator norm for matrices. We may choose L strictly positive even when the Hessian is constant.

For $x \in K$, let $x_t = x^* + t(x - x^*)$. The fundamental theorem of calculus gives

$$\begin{aligned} \nabla f(x) &= \nabla f(x^*) + \int_0^1 D^2 f(x_t)(x - x^*) dt \\ &= D^2 f(x)(x - x^*) + R(x, x^*), \end{aligned}$$

where

$$R(x, x^*) = \int_0^1 (D^2 f(x_t) - D^2 f(x))(x - x^*) dt.$$

Since $\|x_t - x\| = (1 - t) \|x - x^*\|$,

$$\|R(x, x^*)\| \leq L \|x - x^*\|^2 \int_0^1 (1 - t) dt = \frac{L}{2} \|x - x^*\|^2.$$

Define the Newton map

$$N(x) = x - [D^2 f(x)]^{-1} \nabla f(x).$$

For $x \in K$,

$$N(x) - x^* = -[D^2 f(x)]^{-1} R(x, x^*),$$

and hence

$$\|N(x) - x^*\| \leq \|[D^2 f(x)]^{-1}\| \|R(x, x^*)\| \leq \frac{L}{2m} \|x - x^*\|^2.$$

Set $\gamma = L/(2m)$ and choose

$$0 < r < \min\{R, 1/\gamma\}.$$

If $\|x - x^*\| < r$, then $\|N(x) - x^*\| \leq \gamma \|x - x^*\|^2 < r$. Therefore N maps $B(x^*, r)$ into itself, and induction shows that all iterates remain in this ball and satisfy

$$\|x^{(k+1)} - x^*\| \leq \gamma \|x^{(k)} - x^*\|^2.$$

Finally, $\varepsilon_k = \gamma \|x^{(k)} - x^*\|$ satisfies $\varepsilon_k \leq \varepsilon_0^{2^k}$ with $\varepsilon_0 < 1$, so $x^{(k)} \rightarrow x^*$. ■

Remark 18. The local argument only requires a stationary point x^* with $D^2 f(x^*) > 0$ and a locally Lipschitz Hessian. By continuity, the Hessian is uniformly positive definite on a sufficiently small closed ball around x^* . The same proof then applies on that ball. We will use this local form for damped Newton's method.

Invariance under affine changes of variables

Another useful property of Newton's method is its invariance under invertible affine changes of variables. Its iterates correspond exactly after a change of coordinates, including a translation.

Proposition 27. Let $f \in C^2(\mathbb{R}^d)$ have an invertible Hessian at every point. Let A be an invertible square matrix, $b \in \mathbb{R}^d$, and define $g : \mathbb{R}^d \rightarrow \mathbb{R}$ by $g(y) = f(Ay + b)$. Consider

$$\begin{aligned} x^{(k+1)} &= x^{(k)} - [D^2 f(x^{(k)})]^{-1} \nabla f(x^{(k)}), \\ y^{(k+1)} &= y^{(k)} - [D^2 g(y^{(k)})]^{-1} \nabla g(y^{(k)}), \\ \tilde{x}^{(k)} &= Ay^{(k)} + b. \end{aligned}$$

If $x^{(0)} = \tilde{x}^{(0)}$, then $x^{(k)} = \tilde{x}^{(k)}$ for every $k \in \mathbb{N}$.

Proof. Write $z = Ay + b$. Taylor expansion gives

$$\begin{aligned} g(y + v) &= f(z + Av) \\ &= f(z) + \langle \nabla f(z) | Av \rangle + \frac{1}{2} \langle D^2 f(z) Av | Av \rangle + o(\|Av\|^2) \\ &= g(y) + \langle A^T \nabla f(z) | v \rangle + \frac{1}{2} \langle A^T D^2 f(z) Av | v \rangle + o(\|v\|^2). \end{aligned}$$

Consequently,

$$\nabla g(y) = A^T \nabla f(Ay + b), \quad D^2 g(y) = A^T D^2 f(Ay + b) A.$$

The Hessian of g is invertible, and

$$\begin{aligned} y^{(k+1)} &= y^{(k)} - (A^T D^2 f(Ay^{(k)} + b) A)^{-1} A^T \nabla f(Ay^{(k)} + b) \\ &= y^{(k)} - A^{-1} [D^2 f(Ay^{(k)} + b)]^{-1} \nabla f(Ay^{(k)} + b). \end{aligned}$$

Multiplying by A and adding b yields

$$\tilde{x}^{(k+1)} = \tilde{x}^{(k)} - [D^2 f(\tilde{x}^{(k)})]^{-1} \nabla f(\tilde{x}^{(k)}).$$

The two sequences therefore satisfy the same recurrence and have the same initial point. ■

Failure of global convergence

We now give an explicit example for which pure Newton's method fails to converge from some initial points. Consider $f(x) = \sqrt{1+x^2}$ on $\mathbb{R}$. Its derivatives are

$$f'(x) = \frac{x}{\sqrt{1+x^2}}, \quad f''(x) = \frac{1}{(1+x^2)^{3/2}}.$$

Thus f is strictly convex and is $(1+r^2)^{-3/2}$ -strongly convex on $[-r, r]$. Its unique global minimizer is $x^* = 0$.

The Newton direction solves

$$f''(x^{(k)})d^{(k)} = -f'(x^{(k)}),$$

so

$$d^{(k)} = -x^{(k)}(1 + (x^{(k)})^2), \quad x^{(k+1)} = x^{(k)} + d^{(k)} = -(x^{(k)})^3.$$

In particular, $|x^{(k)}| = |x^{(0)}|^{3^k}$, and there are three cases:

- If $|x^{(0)}| < 1$, the iterates converge to 0, with cubic convergence unless they are already at 0.
- If $|x^{(0)}| > 1$, their absolute values tend to $+\infty$.
- If $|x^{(0)}| = 1$, the iterates alternate between 1 and -1 and do not converge.

5.2 Damped Newton's method

A simple modification of pure Newton's method gives global convergence under suitable assumptions.

Definition 17. *Damped Newton's method* is the iteration

$$\begin{cases} d^{(k)} = -[D^2 f(x^{(k)})]^{-1} \nabla f(x^{(k)}), \\ t^{(k)} = \text{armijo_step}(x^{(k)}, d^{(k)}), \\ x^{(k+1)} = x^{(k)} + t^{(k)} d^{(k)}, \end{cases}$$

where the Armijo rule uses $0 < \alpha < 1/2$ and $0 < \beta < 1$, as in the preceding chapter. At a stationary point, the iterates remain fixed.

Damped Newton's method is a preconditioned gradient method with

$$B^{(k)} = [D^2 f(x^{(k)})]^{-1}.$$

Theorem 28. Let $f \in C^3(\mathbb{R}^d)$ satisfy

- (i) the sublevel set $S = \{x \in \mathbb{R}^d \mid f(x) \leq f(x^{(0)})\}$ is compact and convex;
- (ii) $D^2 f(x) > 0$ for every $x \in S$.

Then damped Newton's method converges to the unique global minimizer x^* of f on $\mathbb{R}^d$. Moreover, there exist $k_0 \in \mathbb{N}$ and $\gamma > 0$ such that $\gamma \|x^{(k_0)} - x^*\| < 1$ and

$$\forall k \geq k_0, \quad \|x^{(k+1)} - x^*\| \leq \gamma \|x^{(k)} - x^*\|^2.$$

Remark 19. Armijo backtracking automatically accepts full Newton steps near the solution. The method has a global linear convergence guarantee for the function values, supplied by the preconditioned-gradient analysis, and eventually enters a regime of quadratic convergence for the iterates.

The proof has two parts. First, we verify the assumptions of Theorem 23. Then we show that the full step is accepted once the iterates are sufficiently close to x^* .

Lemma 29. Under the assumptions of Theorem 28, let m, M satisfy

$$0 < m \leq M, \quad m\text{Id} \leq D^2 f(x) \leq M\text{Id} \quad (x \in S).$$

Each Newton preconditioner has a symmetric positive definite square root

$$A^{(k)} = [D^2 f(x^{(k)})]^{-1/2}, \quad (A^{(k)})^2 = B^{(k)}.$$

Moreover,

$$\forall k \in \mathbb{N}, \quad \forall x \in S, \quad \frac{m}{M}\text{Id} \leq A^{(k)} D^2 f(x) A^{(k)} \leq \frac{M}{m}\text{Id}.$$

Proof. Proposition 20 supplies m and M . Armijo backtracking keeps the iterates in S , so all the required Hessians are positive definite. If

$$D^2 f(x^{(k)}) = P^T \text{diag}(\mu_1, \dots, \mu_d) P,$$

with P orthogonal and $\mu_i > 0$, define

$$A^{(k)} = P^T \text{diag}(\mu_1^{-1/2}, \dots, \mu_d^{-1/2}) P.$$

This is symmetric positive definite and squares to $B^{(k)}$. The Hessian bounds at $x^{(k)}$ imply

$$\frac{1}{M} \|v\|^2 \leq \|A^{(k)} v\|^2 = \langle [D^2 f(x^{(k)})]^{-1} v | v \rangle \leq \frac{1}{m} \|v\|^2.$$

For any $x \in S$, we also have

$$m \|A^{(k)} v\|^2 \leq \langle A^{(k)} D^2 f(x) A^{(k)} v | v \rangle \leq M \|A^{(k)} v\|^2.$$

Combining the inequalities gives the result, with $\lambda = m/M$ and $\Lambda = M/m$. ■

By Theorem 23, the iterates therefore converge to the unique global minimizer x^* .

Lemma 30. *Under the assumptions of Theorem 28, there exists $k_1 \in \mathbb{N}$ such that $t^{(k)} = 1$ for every $k \geq k_1$.*

Proof. Choose a closed ball K around x^* on which $D^2 f(x) \geq m_* \text{Id}$ for some $m_* > 0$, and on which the Hessian is Lipschitz with a constant $L_* > 0$. Such a ball exists because $D^2 f(x^*) > 0$ and $f \in C^3$. Since $x^{(k)} \rightarrow x^*$ and $\nabla f(x^*) = 0$, the Newton directions $d^{(k)}$ tend to 0. Consequently the whole segment from $x^{(k)}$ to $x^{(k)} + d^{(k)}$ lies in K for all sufficiently large k .

Write $x = x^{(k)}$, $d = d^{(k)}$, and $H = D^2 f(x)$. Taylor's theorem with a Lipschitz Hessian gives

$$\begin{aligned} f(x + d) &\leq f(x) + \langle \nabla f(x) | d \rangle + \frac{1}{2} \langle H d | d \rangle + \frac{L_*}{6} \|d\|^3 \\ &= f(x) - \frac{1}{2} \langle H d | d \rangle + \frac{L_*}{6} \|d\|^3, \end{aligned}$$

because $Hd = -\nabla f(x)$. Since $\langle Hd | d \rangle \geq m_* \|d\|^2$, the condition

$$\|d\| \leq \frac{6(1/2 - \alpha)m_*}{L_*}$$

implies

$$f(x + d) \leq f(x) - \alpha \langle Hd | d \rangle = f(x) + \alpha \langle \nabla f(x) | d \rangle.$$

The right-hand side is exactly the Armijo bound for $t = 1$. Because $d^{(k)} \rightarrow 0$ and $\alpha < 1/2$, this condition holds eventually, so backtracking accepts the first trial step. ■

Completion of the proof of Theorem 28. We have proved global convergence and eventual acceptance of $t^{(k)} = 1$. The local argument in the proof of Theorem 26 applies around x^* , since its Hessian is positive definite there. Choose $k_0 \geq k_1$ so that $x^{(k_0)}$ lies in the ball from that argument. All subsequent iterates are pure Newton iterates, satisfy the quadratic estimate, and have $\gamma \|x^{(k_0)} - x^*\| < 1$, as required. ■

Exercises and worked solutions

Logistic regression

The model and the optimization problem

This tutorial studies logistic regression, a standard method for binary classification. Let $x_a \in \mathbb{R}^d$, $a \in A = \{1, \dots, n\}$, be a finite data set, with $n, d \geq 1$. Its labels partition A into $A_0 \sqcup A_1$: set $y_a = 0$ for $a \in A_0$ and $y_a = 1$ for $a \in A_1$. We seek a function whose values approximate these labels. Logistic regression uses

$$u_w(x) = \sigma(\langle w|x \rangle), \quad \sigma(t) = \frac{e^t}{1 + e^t} = \frac{1}{1 + e^{-t}}, \quad w \in \mathbb{R}^d.$$

Unlike an unrestricted nonconstant polynomial, this model always takes values strictly between 0 and 1. Its output is interpreted as an estimated probability of label 1. The scalar $\langle w|x \rangle$ is the *linear score*; $u_w(x)$ is its sigmoid transform.

For a regularization parameter $\gamma \geq 0$, consider

$$\min_{w \in \mathbb{R}^d} F(w), \quad F(w) = \sum_{a \in A} f_a(w) + \frac{\gamma}{2} \|w\|^2, \quad (5.1)$$

where

$$f_a(w) = \begin{cases} -\log(1 - \sigma(\langle w|x_a \rangle)), & a \in A_0, \\ -\log(\sigma(\langle w|x_a \rangle)), & a \in A_1. \end{cases}$$

The quadratic term penalizes large coefficients. Once a minimizer w^* has been found, the trained function u_{w^*} can be applied to new data; the usual classification rule assigns label 1 when its value is at least $1/2$. We will determine when a finite minimizer exists and when the algorithms studied in the course converge to it.

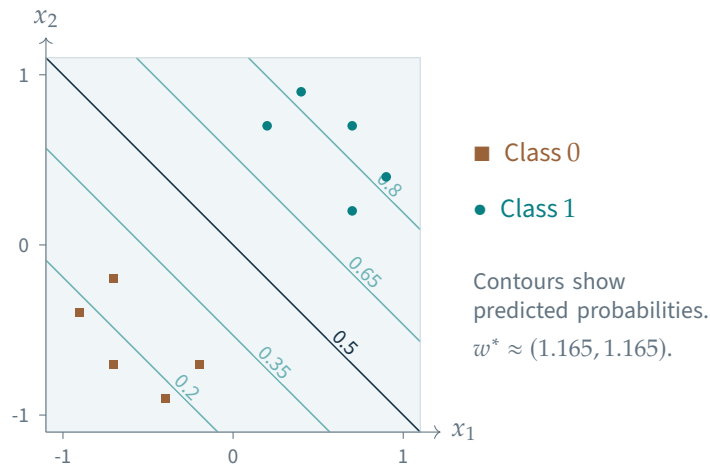


Figure 5.1: Illustrative synthetic data and level sets of the fitted probability $u_{w^*}(x) = \sigma(\langle w^*|x \rangle)$ with $\gamma = 1$. Squares represent class 0 and disks class 1.

TD 2 / Q0

The sigmoid and the statistical interpretation

Show that $0 < \sigma(t) < 1$ for every $t \in \mathbb{R}$, that $\sigma(t) \rightarrow 0$ as $t \rightarrow -\infty$, and that $\sigma(t) \rightarrow 1$ as $t \rightarrow +\infty$. Interpret problem (5.1) when $\gamma = 0$.

Worked solution

Since $e^t > 0$ and $1 + e^t > e^t$, the sigmoid lies in $(0, 1)$. The expression $e^t/(1 + e^t)$ gives the limit at $-\infty$, while $1/(1 + e^{-t})$ gives the limit at $+\infty$.

Write $p_a = \sigma(\langle w | x_a \rangle)$. If the labels are modeled as conditionally independent Bernoulli variables with success probabilities p_a , their likelihood is

$$\mathcal{L}(w) = \prod_{a \in A} p_a^{y_a} (1 - p_a)^{1 - y_a}.$$

For $\gamma = 0$,

$$F(w) = -\log \mathcal{L}(w) = -\sum_{a \in A} [y_a \log p_a + (1 - y_a) \log(1 - p_a)].$$

Minimizing F is therefore maximum likelihood estimation. Each loss encourages a large probability for the observed label. This does not guarantee an exact fit or the existence of a finite minimizer: for one observation $x_1 = 1, y_1 = 1$, the loss $\log(1 + e^{-w})$ decreases to 0 as $w \rightarrow +\infty$, but never attains 0.

The regularized problem: existence and uniqueness

Throughout Questions Q1–Q6, assume $\gamma > 0$.

TD 2 / Q1

Existence of a minimizer

Show that $f_a \geq 0$ for every $a \in A$. Deduce that $F(w) \geq \frac{\gamma}{2} \|w\|^2$ and that F has a minimizer on $\mathbb{R}^d$.

Worked solution

Both $\sigma(\langle w | x_a \rangle)$ and $1 - \sigma(\langle w | x_a \rangle)$ lie in $(0, 1)$, so their negative logarithms are positive. Consequently,

$$F(w) \geq \frac{\gamma}{2} \|w\|^2 \longrightarrow +\infty \quad \text{as } \|w\| \rightarrow +\infty.$$

The function F is continuous and coercive. More explicitly, its nonempty sublevel set $\{w : F(w) \leq F(0)\}$ is closed and contained in the ball of radius $\sqrt{2F(0)/\gamma}$. It is compact, so F attains a minimum there. Outside this sublevel set, $F > F(0)$; the minimum is therefore global.

TD 2 / Q2

Convexity and uniqueness

- (a) If $g : \mathbb{R} \rightarrow \mathbb{R}$ is convex and $x \in \mathbb{R}^d$, show that $w \mapsto g(\langle w|x \rangle)$ is convex.
- (b) Let $g(t) = -\log(1 - \sigma(t))$. Show that $\sigma'(t) = \sigma(t)(1 - \sigma(t))$, then that $g''(t) \geq 0$. Deduce convexity of g and of f_a for $a \in A_0$. Establish the corresponding conclusion for $a \in A_1$.
- (c) Deduce that F is γ -strongly convex and that $D^2F(w) \geq \gamma \text{Id}$ for every $w \in \mathbb{R}^d$.
- (d) Show that the minimizer of F is unique. Denote it by w^* .

Worked solution

- (a) For $w, z \in \mathbb{R}^d$ and $\theta \in [0, 1]$, linearity and convexity give

$$\begin{aligned} g(\langle (1 - \theta)w + \theta z | x \rangle) &= g((1 - \theta)\langle w | x \rangle + \theta\langle z | x \rangle) \\ &\leq (1 - \theta)g(\langle w | x \rangle) + \theta g(\langle z | x \rangle). \end{aligned}$$

- (b) Direct differentiation yields

$$\sigma'(t) = \frac{e^t}{(1 + e^t)^2} = \sigma(t)(1 - \sigma(t)) > 0.$$

Since $1 - \sigma(t) = 1/(1 + e^t)$, we have $g(t) = \log(1 + e^t)$, $g'(t) = \sigma(t)$, and $g''(t) = \sigma'(t) > 0$. Thus g is convex, and part (a) applies. For label 1, the scalar loss is

$$-\log \sigma(t) = \log(1 + e^t) - t = g(t) - t;$$

subtracting a linear function preserves convexity.

- (c) The sum $\sum_a f_a$ is convex, so $F - \frac{\gamma}{2} \|\cdot\|^2$ is convex: this is precisely γ -strong convexity. All the terms are smooth; their Hessians are positive semidefinite, hence $D^2F(w) \geq \gamma \text{Id}$.
- (d) Strong convexity implies strict convexity, which gives at most one minimizer. Question Q1 supplies existence, so w^* is unique.

Derivatives and convergence of gradient descent

TD 2 / Q3

Computing the gradient

Show that

$$\begin{aligned} \nabla f_a(w) &= (\sigma(\langle w | x_a \rangle) - y_a) x_a, \\ \nabla F(w) &= \sum_{a \in A} (\sigma(\langle w | x_a \rangle) - y_a) x_a + \gamma w. \end{aligned} \tag{5.2}$$

Worked solution

The two scalar losses from Question Q2 have the common expression

$$f_a(w) = \log(1 + e^{\langle w | x_a \rangle}) - y_a \langle w | x_a \rangle.$$

Its scalar derivative with respect to $\langle w | x_a \rangle$ is $\sigma(\langle w | x_a \rangle) - y_a$, while the gradient of the score is x_a . The chain rule gives the first formula. Summing it and using $\nabla(\frac{\gamma}{2} \|w\|^2) = \gamma w$ gives the second.

TD 2 / Q4

Computing the Hessian

- (a) For a column vector $x \in \mathbb{R}^d$, show that xx^T is a symmetric positive semidefinite matrix with entries $x_i x_j$, and that $\langle xx^T v | v \rangle = \langle x | v \rangle^2$.
- (b) Given $(x, y) \in \mathbb{R}^d \times \mathbb{R}$, let $G(w) = (\sigma(\langle w | x \rangle) - y)x$. Show that

$$\frac{\partial G_i}{\partial w_j}(w) = \sigma(\langle w | x \rangle)(1 - \sigma(\langle w | x \rangle))x_i x_j.$$

With $s_a(w) = \sigma(\langle w | x_a \rangle)$, deduce the formulas

$$\begin{aligned} D^2 f_a(w) &= s_a(w)(1 - s_a(w))x_a x_a^T, \\ D^2 F(w) &= \sum_{a \in A} s_a(w)(1 - s_a(w))x_a x_a^T + \gamma \text{Id}, \end{aligned} \quad (5.3)$$

Worked solution

- (a) Matrix multiplication gives $(xx^T)_{ij} = x_i x_j = x_j x_i$. Moreover,

$$\langle xx^T v | v \rangle = v^T x x^T v = (x^T v)^2 = \langle x | v \rangle^2 \geq 0.$$

Thus the matrix is positive semidefinite; it is not in general positive definite, since every vector orthogonal to x lies in its kernel.

- (b) Since $G_i(w) = (\sigma(\sum_j w_j x_j) - y)x_i$, the chain rule gives $\partial G_i / \partial w_j = \sigma'(\langle w | x \rangle)x_j x_i$. Substitute the derivative from Question Q2. This is the Jacobian of each gradient in (5.2); adding the derivative γId of γw yields (5.3).

TD 2 / Q5

Convergence of gradient descent

- (a) For any initial point $w^{(0)} \in \mathbb{R}^d$, show that $S = \{w : F(w) \leq F(w^{(0)})\}$ is compact.
- (b) Using (5.3), show that $\langle D^2 f_a(w)v | v \rangle \leq \|x_a\|^2 \|v\|^2$ for every v . Deduce that $\langle D^2 F(w)v | v \rangle \leq M \|v\|^2$, where

$$M = \gamma + \sum_{a \in A} \|x_a\|^2.$$

- (c) Deduce that $\gamma \text{Id} \leq D^2 F(w) \leq M \text{Id}$ for every w .
- (d) Conclude that gradient descent with exact line search, and gradient descent with Armijo backtracking, converge to w^* .

Worked solution

- (a) Continuity makes S closed. Question Q1 gives $\|w\|^2 \leq 2F(w^{(0)})/\gamma$ on S , so S is bounded and therefore compact. It is nonempty and convex, since it is a sublevel set of a convex function.
- (b) For every $s \in [0, 1]$, $s(1 - s) = 1/4 - (s - 1/2)^2 \leq 1/4$. Cauchy–Schwarz thus gives the stronger estimate

$$\langle D^2 f_a(w)v | v \rangle = s_a(w)(1 - s_a(w))\langle x_a | v \rangle^2 \leq \frac{1}{4} \|x_a\|^2 \|v\|^2.$$

Summing yields the requested bound and, in fact, the sharper one with $\widehat{M} = \gamma + \frac{1}{4} \sum_a \|x_a\|^2 \leq M$.

- (c) The lower bound is Question Q2(c), and the upper bound is part (b). These are inequalities of quadratic forms, which are exactly the stated symmetric-matrix inequalities.
- (d) The hypotheses of Theorem 19 hold with $m = \gamma$ and the compact convex set S . Exact line search therefore converges to w^* , with

$$F(w^{(k+1)}) - F(w^*) \leq (1 - \gamma/M)(F(w^{(k)}) - F(w^*)).$$

For Armijo, fix $0 < \alpha < 1/2$ and $0 < \beta < 1$ and use the direction $-\nabla F(w^{(k)})$. Theorem 23 applies with preconditioner $B^{(k)} = \text{Id}$, $\lambda = \gamma$, and $\Lambda = M$. It gives convergence and a contraction factor $1 - c$, where $c = 2\alpha\gamma \min(1, \beta/M) > 0$. In both cases, convergence of the iterates follows also from

$$\|w^{(k)} - w^*\|^2 \leq \frac{2}{\gamma}(F(w^{(k)}) - F(w^*)).$$

At a zero gradient, keep subsequent iterates fixed at w^* .

Newton's method for the regularized problem

TP 2 / Q6

Lipschitz continuity of the Hessian and Newton convergence

For a square matrix E , let $\|E\|_{\text{op}} = \sup_{\|v\|=1} \|Ev\|$ be the Euclidean operator norm. For a symmetric matrix the correct identity is

$$\|E\|_{\text{op}} = \sup_{\|v\|=1} |\langle Ev | v \rangle|.$$

Set $g_a(w) = \sigma(\langle w | x_a \rangle)(1 - \sigma(\langle w | x_a \rangle))$.

- (a) For $E = D^2F(w_0) - D^2F(w_1)$, prove that

$$|\langle Ev | v \rangle| \leq \left(\sum_{a \in A} |g_a(w_0) - g_a(w_1)| \|x_a\|^2 \right) \|v\|^2.$$

- (b) Show that g_a is C^1 and Lipschitz on the sublevel set S from Question Q5. Give a possible Lipschitz constant.
- (c) Deduce that $w \mapsto D^2F(w)$ is Lipschitz on S in operator norm.
- (d) Deduce global convergence of Newton's method with Armijo backtracking. State what is guaranteed for pure Newton's method.

Worked solution

- (a) A symmetric matrix has an orthonormal eigenbasis. Its operator norm and $\sup_{\|v\|=1} |\langle Ev | v \rangle|$ both equal its largest absolute eigenvalue. The absolute value is essential: for $E = -\text{Id}$, the norm is 1 but $\sup_{\|v\|=1} \langle Ev | v \rangle = -1$. The regularization terms cancel in the Hessian difference, so

$$\langle Ev | v \rangle = \sum_{a \in A} (g_a(w_0) - g_a(w_1)) \langle x_a | v \rangle^2.$$

Apply the triangle inequality and $\langle x_a | v \rangle^2 \leq \|x_a\|^2 \|v\|^2$ to obtain the claim.

(b) The sigmoid is smooth, hence so is g_a . With $s = \sigma(\langle w|x_a \rangle)$, differentiation gives

$$\nabla g_a(w) = s(1-s)(1-2s)x_a, \quad \|\nabla g_a(w)\| \leq \frac{1}{4}\|x_a\|.$$

Integrating the derivative along the segment between w_0 and w_1 yields

$$|g_a(w_0) - g_a(w_1)| \leq \frac{1}{4}\|x_a\|\|w_0 - w_1\|.$$

Thus $\tilde{L}_a = \|x_a\|/4$ works on all of $\mathbb{R}^d$, and in particular on S .

(c) Take the supremum over unit vectors in part (a), and use part (b). We obtain the global estimate

$$\|D^2F(w_0) - D^2F(w_1)\|_{\text{op}} \leq L\|w_0 - w_1\|, \quad L = \frac{1}{4} \sum_{a \in A} \|x_a\|^3. \quad (5.4)$$

(d) The Newton direction is obtained by solving

$$D^2F(w^{(k)})d^{(k)} = -\nabla F(w^{(k)}).$$

This system is invertible because $D^2F \geq \gamma \text{Id}$. With $0 < \alpha < 1/2$ and $0 < \beta < 1$, Armijo chooses the first $t^{(k)} \in \{1, \beta, \beta^2, \dots\}$ satisfying

$$F(w^{(k)} + t^{(k)}d^{(k)}) \leq F(w^{(k)}) + \alpha t^{(k)}\langle \nabla F(w^{(k)})|d^{(k)} \rangle.$$

All iterates remain in the compact convex set S . The function is C^∞ , and its Hessian is positive definite on S . Theorem 28 therefore gives convergence from any initial point, eventual acceptance of $t^{(k)} = 1$, and eventual quadratic convergence.

Pure Newton takes $t^{(k)} = 1$ from the outset. Its guaranteed convergence is local, as in Theorem 26. In fact, its Newton map N satisfies

$$\|N(w) - w^*\| \leq \frac{L}{2\gamma} \|w - w^*\|^2,$$

by integrating the Hessian difference along $[w^*, w]$ and using $\|[D^2F(w)]^{-1}\|_{\text{op}} \leq 1/\gamma$. For $L > 0$, an initial error strictly less than $2\gamma/L$ therefore ensures quadratic convergence. If $L = 0$, all $x_a = 0$, so $F(w) = n \log 2 + \frac{\gamma}{2}\|w\|^2$ and a full Newton step reaches $w^* = 0$. Global convergence from the preceding hypotheses is established for the damped method; the local estimate alone does not establish it for full steps from arbitrary initial points.

The unregularized problem

From now on, set $\gamma = 0$. Existence and uniqueness no longer follow from the quadratic penalty. Assume that the class-0 feature vectors positively span $\mathbb{R}^d$, meaning

$$\forall z \in \mathbb{R}^d, \quad z = \sum_{a \in A_0} \lambda_a x_a \quad \text{for some coefficients } \lambda_a \geq 0. \quad (5.5)$$

The coefficients are nonnegative, not necessarily strictly positive. This is a sufficient assumption for the conclusions below. Fix $w^{(0)} \in \mathbb{R}^d$, set $C = F(w^{(0)})$, and let $S = \{w : F(w) \leq C\}$.

TD 2 / Q7

Existence without regularization

- (a) Show that $w \in S$ implies $f_a(w) \leq C$ for every a . Deduce that, for $a \in A_0$,

$$-\log \left(\frac{1}{1 + e^{\langle w|x_a \rangle}} \right) \leq C, \quad \langle w|x_a \rangle \leq C.$$

- (b) Show that assumption (5.5) implies

$$\kappa := \min_{\|z\|=1} \max_{a \in A_0} \langle z|x_a \rangle > 0. \quad (5.6)$$

Hint: the maximum defines a continuous function on the unit sphere. If its minimum were nonpositive, express a minimizing unit vector as a nonnegative combination of the x_a .

- (c) Deduce that for every $w \in \mathbb{R}^d$ there is $a \in A_0$ such that $\langle w|x_a \rangle \geq \kappa \|w\|$.
 (d) Conclude that S is compact and that F has a minimizer.

Worked solution

- (a) Since all losses are nonnegative and $\gamma = 0$, $f_a(w) \leq \sum_b f_b(w) = F(w) \leq C$. For $a \in A_0$, this is $\log(1 + e^{\langle w|x_a \rangle}) \leq C$. Since $\log(1 + e^t) \geq t$, it implies $\langle w|x_a \rangle \leq C$. Equivalently, monotonicity of $-\log$ gives $1/(1 + e^{\langle w|x_a \rangle}) \geq e^{-C}$ and hence $e^{\langle w|x_a \rangle} \leq e^C - 1 < e^C$.
 (b) The function $\phi(z) = \max_{a \in A_0} \langle z|x_a \rangle$ is continuous, as a maximum of finitely many continuous functions. The unit sphere is compact, so its minimum is attained at some z_* with $\|z_*\| = 1$. If $\phi(z_*) \leq 0$, then $\langle z_*|x_a \rangle \leq 0$ for every $a \in A_0$. By (5.5), write $z_* = \sum_{a \in A_0} \lambda_a x_a$, $\lambda_a \geq 0$. Taking the inner product with z_* gives the contradiction

$$1 = \|z_*\|^2 = \sum_{a \in A_0} \lambda_a \langle z_*|x_a \rangle \leq 0.$$

Therefore $\kappa = \phi(z_*) > 0$.

- (c) For $w \neq 0$, apply (5.6) to $z = w/\|w\|$ and multiply by $\|w\|$. The finite maximum is attained by an index in A_0 . For $w = 0$, any index in the nonempty set A_0 works.
 (d) Combining (a) and (c) gives $\kappa \|w\| \leq C$ on S . Thus S is bounded, and it is closed by continuity of F . It is nonempty, so compactness gives a minimum there, which is global since $F > C$ outside S . In fact, the same argument for arbitrary w yields

$$F(w) \geq \max_{a \in A_0} \log(1 + e^{\langle w|x_a \rangle}) \geq \kappa \|w\|.$$

This also proves coercivity directly.

TD 2 / Q8

Positive Hessian and strong convexity on sublevel sets

Using (5.3) and (5.5), show that $D^2F(w)$ is symmetric positive definite at every $w \in \mathbb{R}^d$. Distinguish this conclusion from global strong convexity.

Worked solution

For every $w, v \in \mathbb{R}^d$,

$$\langle D^2F(w)v|v \rangle = \sum_{a \in A} s_a(w)(1 - s_a(w)) \langle x_a|v \rangle^2.$$

Every coefficient $s_a(w)(1 - s_a(w))$ is strictly positive. If this sum is zero, then $\langle x_a|v \rangle = 0$ for every a , in particular for every $a \in A_0$. Assumption (5.5) implies that the x_a , $a \in A_0$, span $\mathbb{R}^d$, so $v = 0$. This proves positive definiteness. Along any nontrivial line segment, the second derivative is positive; hence F is strictly convex. Together with Question Q7, this gives a unique global minimizer w^* .

Positive definiteness at each point does not provide a uniform positive lower eigenvalue on all of $\mathbb{R}^d$. Indeed, for any unit vector v , set $b_a = \langle v|x_a \rangle$. Then

$$\langle D^2F(tv)v|v \rangle = \sum_{a \in A} \sigma(tb_a)(1 - \sigma(tb_a))b_a^2 \longrightarrow 0 \quad (t \rightarrow +\infty).$$

Terms with $b_a = 0$ vanish, and all other terms tend to zero by Question Q0. Thus no constant $m > 0$ can satisfy $D^2F(w) \geq m\text{Id}$ for all w : this finite-data unregularized objective is not globally strongly convex.

On the compact convex set S , however, the continuous Rayleigh quotient has a strictly positive minimum. Proposition 20 gives constants $M_S \geq m_S > 0$ with

$$m_S \text{Id} \leq D^2F(w) \leq M_S \text{Id} \quad (w \in S).$$

Consequently F is m_S -strongly convex on this sublevel set.

TD 2 / Q9; TP 2 / Q9

Convergence without regularization

Under assumption (5.5), deduce convergence of gradient descent with Armijo backtracking and of Newton's method with Armijo backtracking for $\gamma = 0$.

Worked solution

For every initial point, Question Q7 gives a compact sublevel set S . It is convex because every loss is convex. Question Q8 gives uniform bounds $m_S \text{Id} \leq D^2F \leq M_S \text{Id}$ on S . With fixed $0 < \alpha < 1/2$ and $0 < \beta < 1$, all hypotheses of Theorem 23 hold for the identity preconditioner. Thus Armijo gradient descent converges to the unique global minimizer w^* , and its function values converge linearly.

For Newton, $F \in C^\infty$ and its Hessian is positive definite on S . The Newton direction is therefore well defined; Armijo keeps the iterates in S . Theorem 28 gives global convergence from the chosen initial point, eventual acceptance of full steps, and eventual quadratic convergence. The Lipschitz estimate (5.4) also remains valid when $\gamma = 0$, since the quadratic term canceled from the Hessian difference.

These conclusions use strong convexity on the particular compact sublevel set, not global strong convexity. Pure Newton has the local guarantee of Theorem 26; the global statement above concerns Newton with Armijo damping.

Projection and projected gradient descent

We now consider constrained optimization problems of the form

$$\min_{x \in K} f(x), \quad (6.1)$$

where $K \subseteq \mathbb{R}^d$ is a nonempty closed convex set and $f : \mathbb{R}^d \rightarrow \mathbb{R}$ is convex.

In constrained optimization, even when f is differentiable,

$$x^* \in \arg \min_K f \not\Rightarrow \nabla f(x^*) = 0.$$

Example 5. Take $K = [1, 2] \subseteq \mathbb{R}$ and $f(x) = x^2$. The minimum of f over K is attained at $x^* = 1$, although $f'(x^*) = 2 \neq 0$.

Before designing optimization algorithms, we must therefore identify the appropriate optimality conditions.

Proposition 31 (Existence). *The constrained optimization problem (6.1) has a solution if either of the following conditions holds:*

- (i) K is nonempty and compact, and f is continuous;
- (ii) K is nonempty and closed, f is continuous, and $\lim_{\|x\| \rightarrow +\infty} f(x) = +\infty$.

Proposition 32 (Uniqueness). *If f is strictly convex and K is convex, then the constrained optimization problem (6.1) has at most one solution.*

6.1 Projection onto a closed convex set

Theorem 33 (Characterization of the projection). *Let $K \subseteq \mathbb{R}^d$ be nonempty, closed, and convex. For every $x \in \mathbb{R}^d$, the minimization problem*

$$\min_{q \in K} \|q - x\|^2 \quad (6.2)$$

has a unique minimizer p . Moreover, p is characterized by

$$p \in K \quad \text{and} \quad \forall q \in K, \quad \langle x - p | p - q \rangle \geq 0. \quad (6.3)$$

Definition 18 (Projection). *Let $K \subseteq \mathbb{R}^d$ be nonempty, closed, and convex, and let $x \in \mathbb{R}^d$. The unique minimizer of (6.2) is called the *projection of x onto K* and is denoted by $p_K(x)$.*

Proof of Theorem 33. The function $f(q) = \|x - q\|^2$ is continuous, coercive, and strictly convex. The preceding existence and uniqueness propositions therefore give a unique minimizer on K .

Suppose that p minimizes (6.2). For $q \in K$, let $q_t = (1-t)p + tq$, where $t \in [0, 1]$. Convexity of K gives $q_t \in K$. Since $f(q_t) \geq f(p)$,

$$\forall t \in (0, 1), \quad \frac{f(q_t) - f(p)}{t} \geq 0.$$

Letting $t \rightarrow 0^+$ yields

$$\langle \nabla f(p) | q - p \rangle = 2\langle p - x | q - p \rangle \geq 0,$$

which is (6.3).

Conversely, suppose that p satisfies (6.3). By convexity of f ,

$$\forall q \in K, \quad f(q) \geq f(p) + \langle q - p | \nabla f(p) \rangle = f(p) + 2 \underbrace{\langle q - p | p - x \rangle}_{\geq 0} \geq f(p).$$

Thus p minimizes f over K . ■

Corollary 34. *Let $K \subseteq \mathbb{R}^d$ be nonempty, closed, and convex, and let $x, y \in \mathbb{R}^d$. Then*

- (i) $\langle x - y | p_K(x) - p_K(y) \rangle \geq \|p_K(x) - p_K(y)\|^2$;
- (ii) $\|p_K(x) - p_K(y)\| \leq \|x - y\|$.

Proof. Applying the preceding theorem with $p = p_K(x)$ and $q = p_K(y)$ gives

$$\langle x - p_K(x) | p_K(x) - p_K(y) \rangle \geq 0.$$

Interchanging x and y gives

$$\langle y - p_K(y) | p_K(y) - p_K(x) \rangle \geq 0.$$

Adding these inequalities proves (i). Assertion (ii) follows from (i) and the Cauchy–Schwarz inequality. ■

Example 6. Let $e \in \mathbb{R}^d$ be nonzero and $K = \{\lambda e \mid \lambda \in \mathbb{R}\}$. Then

$$p_K(x) = \frac{\langle x|e \rangle}{\|e\|^2} e.$$

Example 7. Let $A \in \mathbb{R}^{m \times d}$, $b \in \mathbb{R}^m$, and assume that $K = \{q \in \mathbb{R}^d \mid Aq = b\}$ is nonempty. For $x \in \mathbb{R}^d$, (6.3) characterizes its projection by

$$\begin{aligned} p = p_K(x) &\iff p \in K \text{ and } \forall q \in K, \langle x - p | p - q \rangle \geq 0 \\ &\iff p \in K \text{ and } \forall v \in \text{Ker} A, \langle x - p | v \rangle = 0 \\ &\iff p \in K \text{ and } x - p \in (\text{Ker} A)^\perp. \end{aligned}$$

Moreover, $(\text{Ker} A)^\perp = \text{Im} A^T$, since

$$\begin{aligned} v \in \text{Ker} A &\iff \forall w \in \mathbb{R}^m, \langle Av | w \rangle = 0 \\ &\iff \forall w \in \mathbb{R}^m, \langle v | A^T w \rangle = 0 \\ &\iff v \in (\text{Im} A^T)^\perp. \end{aligned}$$

Hence $\text{Ker} A = (\text{Im} A^T)^\perp$, and taking orthogonal complements gives the claimed identity.

Proposition 35. If $\text{Ker} A^T = \{0\}$ and $K = \{q \mid Aq = b\}$, then

$$p_K(x) = x - A^T(AA^T)^{-1}(Ax - b).$$

Proof. The assumption implies that A is surjective, so K is nonempty, and AA^T is positive definite: for $w \neq 0$, $\langle AA^T w | w \rangle = \|A^T w\|^2 > 0$. By the preceding characterization, $p = p_K(x)$ if and only if $Ap = b$ and $x - p \in \text{Im} A^T$. Thus $p = x - A^T w$ for some w satisfying

$$A(x - A^T w) = b, \quad \text{that is,} \quad AA^T w = Ax - b.$$

Therefore $w = (AA^T)^{-1}(Ax - b)$, which gives the formula for p . ■

Example 8. Let $K = \{q \in \mathbb{R}^d \mid q_i \geq 0 \text{ for } 1 \leq i \leq d\}$. When $d = 1$, $p_K(x) = \max(x, 0) =: x^+$. To prove this, we again use (6.3). The point $p = x^+ \in K$ is the projection of x if

$$\forall q \in K, \quad (x - x^+)(x^+ - q) \geq 0.$$

If $x \geq 0$, then $x - x^+ = 0$. If $x < 0$, then $x - x^+ < 0$ and $x^+ - q = -q \leq 0$, so the inequality also holds. More generally:

Proposition 36. Let $K = \{q \in \mathbb{R}^d \mid q_i \geq 0 \text{ for } 1 \leq i \leq d\}$. Then, for every $x \in \mathbb{R}^d$,

$$p_K(x) = (x_1^+, \dots, x_d^+).$$

Example 9. Let $K = K_1 \times \dots \times K_\ell \subseteq \mathbb{R}^d$, where each $K_i \subseteq \mathbb{R}^{d_i}$ is nonempty, closed, and convex, and $d = d_1 + \dots + d_\ell$. Then

$$p_K(\underbrace{x_1}_{\in \mathbb{R}^{d_1}}, \dots, \underbrace{x_\ell}_{\in \mathbb{R}^{d_\ell}}) = (p_{K_1}(x_1), \dots, p_{K_\ell}(x_\ell)).$$

Example 10. Let $K = \{x \in \mathbb{R}^d \mid \|x\| \leq 1\}$, where $\|\cdot\|$ denotes the Euclidean norm. Then

$$p_K(x) = \begin{cases} x & \text{if } \|x\| \leq 1, \\ \frac{x}{\|x\|} & \text{otherwise.} \end{cases}$$

The case $x \in K$ is immediate. If $x \notin K$, set $p = x/\|x\|$. For every $q \in K$,

$$\langle x - p, p - q \rangle = \left(1 - \frac{1}{\|x\|}\right) \langle x, p - q \rangle.$$

The first factor is positive. By Cauchy–Schwarz and $\langle x, p \rangle = \|x\|$,

$$\langle x, p - q \rangle \geq \|x\| - \|x\| \|q\| \geq 0.$$

The characterization (6.3) therefore gives $p_K(x) = p$.

6.2 Optimality conditions for constrained optimization

Necessary and sufficient optimality conditions

In this section, we prove several equivalent optimality conditions for constrained optimization.

A point x^* can minimize a differentiable function f over a compact set K even when $\nabla f(x^*) \neq 0$. To design correct optimization algorithms, we need an appropriate necessary and sufficient optimality condition.

Theorem 37. Let $K \subseteq \mathbb{R}^d$ be nonempty, closed, and convex, and let $f \in C^1(\mathbb{R}^d)$ be convex. For a point $x^* \in K$, the following statements are equivalent:

- (i) $x^* \in \arg \min_K f$;
- (ii) $\forall y \in K, \langle x^* - y, -\nabla f(x^*) \rangle \geq 0$ (variational formulation);
- (iii) $\exists \tau > 0, p_K(x^* - \tau \nabla f(x^*)) = x^*$ (geometric formulation);
- (iv) $\forall \tau > 0, p_K(x^* - \tau \nabla f(x^*)) = x^*$.

Remark 20. This theorem does not assert existence or uniqueness. It only characterizes the optimality of a feasible point for a constrained optimization problem.

Proof. (ii) $\implies$ (i): By convexity of f ,

$$\forall y \in \mathbb{R}^d, \quad f(y) \geq f(x^*) + \langle y - x^*, \nabla f(x^*) \rangle.$$

For $y \in K$, the inner product is nonnegative by (ii), so

$$\forall y \in K, \quad f(y) \geq f(x^*).$$

Since $x^* \in K$, it minimizes f over K .

(i) $\implies$ (ii): Let $y \in K$ and $y_t = (1 - t)x^* + ty$ for $t \in [0, 1]$. Convexity gives $y_t \in K$. Since x^* minimizes f over K ,

$$\forall t \in (0, 1), \quad \frac{f(x^* + t(y - x^*)) - f(x^*)}{t} \geq 0.$$

Taking the limit as $t \rightarrow 0^+$ yields $\langle \nabla f(x^*) | y - x^* \rangle \geq 0$.

To prove the equivalence of (ii), (iii), and (iv), fix $\tau > 0$. Applying the projection characterization to $x^* - \tau \nabla f(x^*)$ gives

$$\begin{aligned} p_K(x^* - \tau \nabla f(x^*)) = x^* &\iff \forall y \in K, \langle -\tau \nabla f(x^*) | x^* - y \rangle \geq 0 \\ &\iff \forall y \in K, \langle -\nabla f(x^*) | x^* - y \rangle \geq 0. \end{aligned}$$

Here we used $x^* \in K$ and $\tau > 0$. The final condition is independent of τ , proving both equivalences. ■

Example 11. Let $K = \mathbb{R}_+^d$. We have shown that

$$p_K(x) = (\max(x_1, 0), \dots, \max(x_d, 0)).$$

Let $f \in C^1(\mathbb{R}^d)$ be convex and $x^* \in K$. By the preceding theorem,

$$\begin{aligned} x^* \in \arg \min_K f &\iff \forall t > 0, p_K(x^* - t \nabla f(x^*)) = x^* \\ &\iff \forall t > 0, \forall i, \quad \max\left(x_i^* - t \frac{\partial f}{\partial x_i}(x^*), 0\right) = x_i^*. \end{aligned}$$

This is equivalent to the coordinatewise conditions

$$\begin{cases} \frac{\partial f}{\partial x_i}(x^*) = 0 & \text{if } x_i^* > 0, \\ \frac{\partial f}{\partial x_i}(x^*) \geq 0 & \text{if } x_i^* = 0. \end{cases}$$

The projected gradient algorithm

Definition 19. Let $f \in C^1(\mathbb{R}^d)$ and let $K \subseteq \mathbb{R}^d$ be nonempty, closed, and convex. Given $x^{(0)} \in \mathbb{R}^d$, the projected gradient algorithm with constant step size $\tau > 0$ is defined by

$$x^{(k+1)} = p_K(x^{(k)} - \tau \nabla f(x^{(k)})). \quad (6.4)$$

In particular, $x^{(k)} \in K$ for every $k \geq 1$.

Remark 21. To implement this method, we must be able to compute the projection onto K . The method is especially useful when that projection has an explicit formula or can be computed efficiently.

Theorem 38. Let $K \subseteq \mathbb{R}^d$ be nonempty, closed, and convex. Suppose that $f \in C^2(\mathbb{R}^d)$ satisfies

$$\exists 0 < m \leq M \quad \text{such that} \quad \forall x \in \mathbb{R}^d, \quad m \text{Id} \leq D^2 f(x) \leq M \text{Id}. \quad (6.5)$$

Then the sequence defined by (6.4) converges to the unique minimizer $x^* \in \arg \min_K f$, provided that

$$0 < \tau < \frac{2m}{M^2}.$$

We can write $x^{(k+1)} = F(x^{(k)})$, where $F(x) = p_K(x - \tau \nabla f(x))$. If $x^{(k)}$ converges to some $\bar{x} \in \mathbb{R}^d$, continuity of F gives

$$\bar{x} = F(\bar{x}) = p_K(\bar{x} - \tau \nabla f(\bar{x})).$$

By Theorem 37, $\bar{x}$ then minimizes f over K . It remains to prove that the iteration converges. We use the contraction mapping theorem.

Theorem 39 (Contraction mapping theorem). *Let $F : \mathbb{R}^d \rightarrow \mathbb{R}^d$ be a contraction: there exists $\kappa \in [0, 1)$ such that*

$$\forall x, y \in \mathbb{R}^d, \quad \|F(x) - F(y)\| \leq \kappa \|x - y\|.$$

Then

- (i) F has a unique fixed point x^* ;
- (ii) for every $x^{(0)} \in \mathbb{R}^d$, the sequence defined by $x^{(k+1)} = F(x^{(k)})$ converges to x^* .

We also need the following lemma.

Lemma 40. *Let $f \in C^2(\mathbb{R}^d)$ satisfy (6.5), and let $\tau > 0$. The mapping $G : x \mapsto x - \tau \nabla f(x)$ satisfies*

$$\|G(x) - G(y)\|^2 \leq (1 - 2\tau m + M^2\tau^2) \|x - y\|^2.$$

Proof. Set $v = y - x$ and $x_t = x + tv$. Applying the fundamental theorem of calculus to $h(t) = \nabla f(x_t)$ gives

$$\nabla f(y) - \nabla f(x) = \int_0^1 D^2 f(x_t) v \, dt.$$

Consequently,

$$\begin{aligned} \|\nabla f(y) - \nabla f(x)\|^2 &= \left\| \int_0^1 D^2 f(x_t) v \, dt \right\|^2 \\ &\leq \int_0^1 \|D^2 f(x_t) v\|^2 \, dt \\ &\leq \int_0^1 \|D^2 f(x_t)\|^2 \|v\|^2 \, dt \\ &\leq M^2 \|v\|^2. \end{aligned}$$

Here the Euclidean operator norm of a symmetric matrix equals the largest absolute value of its eigenvalues, which is at most M for $D^2 f(x_t)$. Moreover,

$$\begin{aligned} \langle \nabla f(y) - \nabla f(x) | y - x \rangle &= \left\langle \int_0^1 D^2 f(x_t) v \, dt | v \right\rangle \\ &= \int_0^1 \langle D^2 f(x_t) v | v \rangle \, dt \\ &\geq m \|v\|^2. \end{aligned}$$

Therefore, for all $x, y \in \mathbb{R}^d$,

$$\begin{aligned} \|G(x) - G(y)\|^2 &= \|x - y - \tau(\nabla f(x) - \nabla f(y))\|^2 \\ &= \|x - y\|^2 - 2\tau \langle x - y | \nabla f(x) - \nabla f(y) \rangle + \tau^2 \|\nabla f(x) - \nabla f(y)\|^2 \\ &\leq (1 - 2\tau m + M^2\tau^2) \|x - y\|^2. \end{aligned}$$

■

Proof of Theorem 38. Set $G(x) = x - \tau \nabla f(x)$ and $F = p_K \circ G$, so that $x^{(k+1)} = F(x^{(k)})$. Since p_K is 1-Lipschitz,

$$\|F(x) - F(y)\| = \|p_K(G(x)) - p_K(G(y))\| \leq \|G(x) - G(y)\|.$$

The lemma therefore gives

$$\|F(x) - F(y)\|^2 \leq (1 - 2\tau m + M^2\tau^2) \|x - y\|^2.$$

Let $\kappa = \sqrt{1 - 2\tau m + M^2\tau^2}$. The quantity under the square root is nonnegative because it equals $(1 - \tau m)^2 + \tau^2(M^2 - m^2)$. Furthermore,

$$\kappa^2 < 1 \iff \tau(M^2\tau - 2m) < 0 \iff 0 < \tau < \frac{2m}{M^2}.$$

Under this step-size condition, F is a contraction and has a unique fixed point x^* . By Theorem 37, a point minimizes f over K if and only if it is a fixed point of F . Thus the constrained problem has exactly one minimizer, and the contraction mapping theorem shows that the iteration $x^{(k+1)} = F(x^{(k)})$ converges to it. ■

Exercises and worked solutions

Projections and constrained minimization

These exercises use the course results, especially the characterization of projection onto a nonempty closed convex set, to compute projections and derive constrained optimality conditions. They also provide formulas for numerical implementations. Throughout, $n \geq 1$ and all projections use the Euclidean norm.

I. Projection onto a half-space

TD 3 / I.Q1

Projection onto a half-space

Let $v \in \mathbb{R}^n$ be nonzero and let $K = \{x \in \mathbb{R}^n \mid \langle v|x \rangle \leq 0\}$. Prove that the projection of $y \in \mathbb{R}^n$ onto K is

$$p_K(y) = \begin{cases} y, & \text{if } \langle v|y \rangle \leq 0, \\ y - \frac{\langle v|y \rangle}{\|v\|^2} v, & \text{otherwise.} \end{cases}$$

Worked solution

The set K is nonempty, closed, and convex. If $\langle v|y \rangle \leq 0$, then $y \in K$ and its distance to K is zero, so $p_K(y) = y$.

Otherwise, write $\alpha = \langle v|y \rangle / \|v\|^2 > 0$ and set $p = y - \alpha v$. Then $\langle v|p \rangle = 0$, so $p \in K$. For every $z \in K$,

$$\langle y - p | p - z \rangle = \alpha \langle v | p - z \rangle = -\alpha \langle v | z \rangle \geq 0.$$

The projection characterization, Theorem 33, gives $p = p_K(y)$.

II. Minimization with coordinatewise lower bounds

Fix $f \in \mathbb{R}^n$ and let

$$K = \{x \in \mathbb{R}^n \mid x_i \geq f_i \text{ for } 1 \leq i \leq n\}.$$

We study the minimization of $J \in C^1(\mathbb{R}^n)$ over K . Here e_i denotes the i th standard basis vector.

TD 3 / II.Q1

Projection onto the obstacle set

Show that

$$p_K(x) = (\max(x_i, f_i))_{1 \leq i \leq n}.$$

Worked solution

The set K is nonempty, closed, and convex. Set $p_i = \max(x_i, f_i)$, so $p \in K$. For any $z \in K$, consider $(x_i - p_i)(p_i - z_i)$. If $x_i \geq f_i$, the first factor is zero. If $x_i < f_i$, then $p_i = f_i$, and both factors are nonpositive.

Thus

$$\langle x - p | p - z \rangle = \sum_{i=1}^n (x_i - p_i)(p_i - z_i) \geq 0.$$

The projection characterization proves $p = p_K(x)$.

TD 3 / II.Q2

Necessary optimality conditions

Let $J \in C^1(\mathbb{R}^n)$ and $x^* \in \arg \min_K J$.

(a) If $x_i^* > f_i$, set $\varepsilon_0 = x_i^* - f_i$. Show that $x_\varepsilon = x^* + \varepsilon e_i \in K$ for every $\varepsilon \in [-\varepsilon_0, +\infty)$ and that $J(x_\varepsilon) \geq J(x^*)$. Deduce that $\partial J / \partial x_i(x^*) = 0$.

(b) If $x_i^* = f_i$, show that $x_\varepsilon = x^* + \varepsilon e_i \in K$ for every $\varepsilon \geq 0$. Deduce that $\partial J / \partial x_i(x^*) \geq 0$.

Conclude that

$$x^* \in \arg \min_K J \implies \nabla J(x^*) = \lambda \in \mathbb{R}_+^n, \quad \lambda_i(x_i^* - f_i) = 0 \quad (1 \leq i \leq n).$$

Worked solution

(a) Only coordinate i changes, and $x_i^* + \varepsilon \geq x_i^* - \varepsilon_0 = f_i$ when $\varepsilon \geq -\varepsilon_0$. Thus $x_\varepsilon \in K$, and optimality gives $J(x_\varepsilon) \geq J(x^*)$. For $\varepsilon > 0$, the difference quotient $[J(x^* + \varepsilon e_i) - J(x^*)]/\varepsilon$ is nonnegative; for $-\varepsilon_0 \leq \varepsilon < 0$, it is nonpositive. The two limits exist and equal $\partial J / \partial x_i(x^*)$, which must therefore be zero.

(b) For $\varepsilon \geq 0$, the changed coordinate is $f_i + \varepsilon \geq f_i$, and all other coordinates remain feasible. Thus $J(x^* + \varepsilon e_i) \geq J(x^*)$. Dividing by $\varepsilon > 0$ and letting $\varepsilon \downarrow 0$ gives $\partial J / \partial x_i(x^*) \geq 0$.

Set $\lambda_i = \partial J / \partial x_i(x^*)$. Each coordinate is nonnegative. If $x_i^* > f_i$, then $\lambda_i = 0$; if $x_i^* = f_i$, then $x_i^* - f_i = 0$. Hence $\lambda_i(x_i^* - f_i) = 0$ in both cases.

TD 3 / II.Q3

Sufficiency for a convex objective

Assume that $J \in C^1(\mathbb{R}^n)$ is convex. Let $\bar{x} \in K$ satisfy

$$\nabla J(\bar{x}) = \lambda \in \mathbb{R}_+^n, \quad \lambda_i(\bar{x}_i - f_i) = 0 \quad (1 \leq i \leq n).$$

Prove that $\bar{x} \in \arg \min_K J$.

Worked solution

For every $z \in K$, convexity gives

$$\begin{aligned} J(z) &\geq J(\bar{x}) + \langle \nabla J(\bar{x}) | z - \bar{x} \rangle \\ &= J(\bar{x}) + \sum_{i=1}^n \lambda_i(z_i - \bar{x}_i) - \sum_{i=1}^n \lambda_i(\bar{x}_i - f_i) \\ &= J(\bar{x}) + \sum_{i=1}^n \lambda_i(z_i - f_i) \geq J(\bar{x}). \end{aligned}$$

The last inequality uses $\lambda_i \geq 0$ and $z_i \geq f_i$. Since $\bar{x} \in K$, it minimizes J on K .

TD 3 / II.Q4**The discrete obstacle problem**

Consider the energy

$$J(x) = \frac{1}{2} \sum_{i=0}^n (x_{i+1} - x_i)^2, \quad x \in \mathbb{R}^n,$$

with the convention $x_0 = x_{n+1} = 0$.

- (a) Compute $\nabla J(x)$.
- (b) Using Questions II.Q2 and II.Q3, prove that $\bar{x} \in \arg \min_K J$ if and only if $\bar{x} \in K$ and, for every $1 \leq i \leq n$,

$$\begin{cases} \bar{x}_i = \frac{\bar{x}_{i-1} + \bar{x}_{i+1}}{2}, & \text{if } \bar{x}_i > f_i, \\ \bar{x}_i \geq \frac{\bar{x}_{i-1} + \bar{x}_{i+1}}{2}, & \text{if } \bar{x}_i = f_i. \end{cases}$$

Worked solution

- (a) The coordinate x_i appears in the two summands $\frac{1}{2}(x_i - x_{i-1})^2$ and $\frac{1}{2}(x_{i+1} - x_i)^2$. Hence

$$(\nabla J(x))_i = (x_i - x_{i-1}) - (x_{i+1} - x_i) = 2x_i - x_{i-1} - x_{i+1}.$$

This formula includes $i = 1$ and $i = n$ using the prescribed boundary values; for $n = 1$, it gives $J'(x_1) = 2x_1$.

- (b) The energy J is convex because it is a sum of squares of linear functions. Questions II.Q2 and II.Q3 therefore show that a feasible point $\bar{x}$ is a minimizer exactly when

$$\begin{cases} 2\bar{x}_i - \bar{x}_{i-1} - \bar{x}_{i+1} = 0, & \text{if } \bar{x}_i > f_i, \\ 2\bar{x}_i - \bar{x}_{i-1} - \bar{x}_{i+1} \geq 0, & \text{if } \bar{x}_i = f_i. \end{cases}$$

Dividing by 2 gives the required characterization.

In fact the minimizer exists and is unique. For any nonzero h , with $h_0 = h_{n+1} = 0$,

$$\langle D^2 J h | h \rangle = \sum_{i=0}^n (h_{i+1} - h_i)^2 > 0 :$$

equality would force every h_i to equal $h_0 = 0$. Thus J is a positive definite quadratic form, so it is strictly convex and coercive. The set K is nonempty, closed, and convex.

III. Minimization on the unit simplex

The *unit simplex* is

$$\Delta = \left\{ x \in \mathbb{R}_+^n \mid \sum_{i=1}^n x_i = 1 \right\}.$$

It appears in probability, geometry, and finance. Its coordinates can represent probabilities or the fractions of a portfolio invested in different assets: nonnegativity excludes negative holdings, and the sum constraint allocates the whole portfolio.

For $y \in \mathbb{R}^n$, we seek its projection onto Δ :

$$\min_{x \in \Delta} \|x - y\|^2. \quad (6.6)$$

For $\lambda \in \mathbb{R}$, define $x^\lambda \in \mathbb{R}^n$ by $x_i^\lambda = \max(y_i - \lambda, 0)$. The objective is to prove that the solution of (6.6) has this form for a suitable $\lambda = \lambda^*$.

TD 3 / III.Q1

Existence and uniqueness

Show that problem (6.6) has a unique solution.

Worked solution

The set Δ is nonempty because $(1, 0, \dots, 0) \in \Delta$. It is closed, convex, and bounded, since $0 \leq x_i \leq 1$ on Δ . Thus it is compact. The continuous function $x \mapsto \|x - y\|^2$ attains its minimum there, and its strict convexity makes that minimizer unique. Equivalently, this follows directly from the projection theorem.

TD 3 / III.Q2

Finding a feasible threshold

Set $g(\lambda) = \sum_{i=1}^n x_i^\lambda$. Show that g is continuous, that $g(\lambda) \rightarrow +\infty$ as $\lambda \rightarrow -\infty$, and that $g(\lambda) \rightarrow 0$ as $\lambda \rightarrow +\infty$. Deduce that some $\lambda^* \in \mathbb{R}$ satisfies $g(\lambda^*) = 1$, and therefore $x^{\lambda^*} \in \Delta$.

Worked solution

Each function $\lambda \mapsto \max(y_i - \lambda, 0)$ is continuous, so their sum g is continuous. For $\lambda < \min_i y_i$,

$$g(\lambda) = \sum_{i=1}^n y_i - n\lambda \longrightarrow +\infty \quad (\lambda \rightarrow -\infty).$$

For $\lambda \geq \max_i y_i$, all terms vanish, so $g(\lambda) = 0$. The intermediate value theorem gives λ^* with $g(\lambda^*) = 1$. The coordinates of x^{λ^*} are nonnegative and sum to 1, so this vector belongs to Δ .

The threshold is also unique. If $a < b < \max_i y_i$, choose j with $y_j > b$. The j th summand strictly decreases from a to b , while all the others are nonincreasing. Thus $g(a) > g(b)$. Since any threshold with $g(\lambda) = 1$ lies below $\max_i y_i$, there can be only one.

TD 3 / III.Q3

Proving the simplex projection formula

Let $\bar{x} = x^{\lambda^*}$ and write $\lambda = \lambda^*$.

(a) Explain why it suffices to prove that, for every $z \in \Delta$,

$$\langle y - \bar{x} | \bar{x} - z \rangle = \sum_{i=1}^n (y_i - \bar{x}_i)(\bar{x}_i - z_i) \geq 0.$$

(b) If $\bar{x}_i > 0$, prove that $\bar{x}_i = y_i - \lambda$ and

$$(y_i - \bar{x}_i)(\bar{x}_i - z_i) = \lambda(\bar{x}_i - z_i).$$

(c) If $\bar{x}_i = 0$, prove that $y_i \leq \lambda$ and

$$(y_i - \bar{x}_i)(\bar{x}_i - z_i) \geq \lambda(\bar{x}_i - z_i).$$

(d) Using $\bar{x}, z \in \Delta$, deduce the inequality in (a) and conclude that $p_\Delta(y) = x^\lambda$.

Worked solution

- (a) We already know that $\bar{x} \in \Delta$. Since Δ is nonempty, closed, and convex, Theorem 33 states exactly that $\bar{x} = p_\Delta(y)$ if and only if the displayed inequality holds for every $z \in \Delta$.
- (b) If $\max(y_i - \lambda, 0) > 0$, then $\bar{x}_i = y_i - \lambda$, so $y_i - \bar{x}_i = \lambda$. Multiplication by $\bar{x}_i - z_i$ gives the asserted equality.
- (c) If $\max(y_i - \lambda, 0) = 0$, then $y_i \leq \lambda$. As $z_i \geq 0$, multiplication by $-z_i \leq 0$ reverses this inequality. Therefore

$$(y_i - \bar{x}_i)(\bar{x}_i - z_i) = -y_i z_i \geq -\lambda z_i = \lambda(\bar{x}_i - z_i).$$

- (d) Combining (b) and (c), then summing over all coordinates, gives

$$\langle y - \bar{x} | \bar{x} - z \rangle \geq \lambda \sum_{i=1}^n (\bar{x}_i - z_i) = \lambda(1 - 1) = 0.$$

Part (a) now proves $p_\Delta(y) = (\max(y_i - \lambda^*, 0))_{i=1}^n$.

TD 3 / III.Q4

Projection onto the unit ℓ^1 ball

Let $B = \{x \in \mathbb{R}^n \mid \sum_{i=1}^n |x_i| \leq 1\}$. We now seek the projection of $y \in \mathbb{R}^n$ onto B .

- (a) Show that if $y \in \mathbb{R}_+^n$, then $p_B(y) \in \mathbb{R}_+^n$.
- (b) Deduce that if $y \in \mathbb{R}_+^n \setminus B$, then $p_B(y) = p_\Delta(y)$.
- (c) For $y \notin B$, write $y_i = \varepsilon_i \tilde{y}_i$, where $\tilde{y}_i = |y_i|$ and $\varepsilon_i \in \{-1, 1\}$. Prove that

$$(p_B(y))_i = \varepsilon_i (p_\Delta(\tilde{y}))_i.$$

Worked solution

The set B is nonempty, closed, and convex, so its Euclidean projection exists and is unique.

- (a) Suppose $y_i \geq 0$ for every i , and set $p = p_B(y)$. If $p_i < 0$, replace this coordinate by 0, leaving all others unchanged, to obtain $q \in B$. Then

$$\|p - y\|^2 - \|q - y\|^2 = (p_i - y_i)^2 - y_i^2 = p_i^2 - 2p_i y_i > 0,$$

contradicting minimality of p . Thus every $p_i \geq 0$.

- (b) By (a), $p = p_B(y)$ is nonnegative. We claim that $\sum_i p_i = 1$. If instead $s = \sum_i p_i < 1$, let $r = \sum_i y_i > 1$ and choose $0 < t < (1 - s)/(r - s)$. The point $z_t = (1 - t)p + ty$ is nonnegative and satisfies

$$\sum_i (z_t)_i = s + t(r - s) < 1,$$

so $z_t \in B$. But $p \neq y$ and $\|z_t - y\| = (1 - t)\|p - y\| < \|p - y\|$, a contradiction. Hence $p \in \Delta$. Because $\Delta \subseteq B$, the point p also minimizes the distance to y over Δ , and uniqueness gives $p = p_\Delta(y)$.

- (c) Let $D = \text{diag}(\varepsilon_1, \dots, \varepsilon_n)$, choosing either sign when $y_i = 0$. Then $D^T D = \text{Id}$, $D^2 = \text{Id}$, $DB = B$, and $y = D\tilde{y}$. For every $x \in B$,

$$\|x - y\| = \|Dx - \tilde{y}\|.$$

The change of variables $x \mapsto Dx$ is a bijection of B , so uniqueness of the projection gives $p_B(y) = Dp_B(\tilde{y})$. Since $\tilde{y} \in \mathbb{R}_+^n \setminus B$, part (b) yields $p_B(\tilde{y}) = p_\Delta(\tilde{y})$, proving the formula. Thus, if $\sum_i |y_i| > 1$, the projection is obtained by thresholding absolute values and restoring their signs:

$$(p_B(y))_i = \varepsilon_i \max(|y_i| - \lambda^*, 0), \quad \sum_{i=1}^n \max(|y_i| - \lambda^*, 0) = 1.$$

Here $\lambda^* > 0$, since the threshold sum at $\lambda = 0$ is $\sum_i |y_i| > 1$ and it is nonincreasing. In particular, a zero coordinate of y remains zero, so its chosen sign is immaterial. If $y \in B$, its projection is simply y .

Optimization with inequality constraints

In this chapter, we study optimization problems $\min_K f$ in which the feasible set is defined by inequalities. Given convex functions $c_1, \dots, c_\ell \in C^1(\mathbb{R}^d)$, we set

$$K = \{x \in \mathbb{R}^d \mid c_1(x) \leq 0, \dots, c_\ell(x) \leq 0\}.$$

Each inequality $c_i(x) \leq 0$ is called a *constraint*, and c_i is its *constraint function*. We often write the problem as

$$P := \min_{c_1(x) \leq 0, \dots, c_\ell(x) \leq 0} f(x).$$

Lemma 41. *If $c_1, \dots, c_\ell : \mathbb{R}^d \rightarrow \mathbb{R}$ are continuous and convex, then*

$$K = \{x \in \mathbb{R}^d \mid c_i(x) \leq 0 \text{ for } 1 \leq i \leq \ell\}$$

is closed and convex.

Proof. First, let $(x_n)_n$ be a sequence in K converging to x . For every i , $c_i(x_n) \leq 0$. Continuity gives $c_i(x) \leq 0$, so $x \in K$. Hence K is closed.

Next, let $x, y \in K$ and $t \in [0, 1]$. By convexity of each c_i ,

$$c_i((1-t)x + ty) \leq (1-t)c_i(x) + tc_i(y) \leq 0.$$

Thus $(1-t)x + ty \in K$, proving that K is convex. ■

Example 12 (Simplex). Consider the set

$$K = \left\{ x \in \mathbb{R}^d \mid x_i \geq 0 \text{ for every } i, \sum_{i=1}^d x_i = 1 \right\},$$

which represents the probability measures on $\{1, \dots, d\}$. It has the required form with

$$\begin{aligned} c_i(x) &= -x_i & \text{for } 1 \leq i \leq d, \\ c_{d+1}(x) &= \sum_{i=1}^d x_i - 1, & c_{d+2}(x) &= -\left(\sum_{i=1}^d x_i - 1\right). \end{aligned}$$

7.1 The penalty method

The idea is to approximate the constrained problem $\min_K f$, where $K = \{x \mid c_i(x) \leq 0 \text{ for every } i\}$, by an unconstrained problem $\min_{\mathbb{R}^d} f_\varepsilon$. The function f_ε is the sum of f and terms that become large when a constraint is violated. For $\varepsilon > 0$, define

$$P_\varepsilon := \min_{x \in \mathbb{R}^d} f_\varepsilon(x), \quad f_\varepsilon(x) := f(x) + \frac{1}{\varepsilon} \sum_{i=1}^{\ell} \max(c_i(x), 0)^2.$$

A point at which $c_i(x) > 0$ incurs a positive *penalty*. For a fixed infeasible point, this penalty tends to $+\infty$ as $\varepsilon \rightarrow 0^+$. More precisely,

$$\forall x \in \mathbb{R}^d, \quad \lim_{\varepsilon \rightarrow 0^+} \frac{1}{\varepsilon} \sum_{i=1}^{\ell} \max(c_i(x), 0)^2 = \begin{cases} 0 & \text{if } c_i(x) \leq 0 \text{ for every } i, \\ +\infty & \text{otherwise.} \end{cases}$$

The next proposition makes this intuition precise.

Proposition 42. *Assume that K is nonempty, that $c_1, \dots, c_\ell$ are continuous and convex, and that $f \in C^0(\mathbb{R}^d)$ is strictly convex and coercive:*

$$\lim_{\|x\| \rightarrow +\infty} f(x) = +\infty.$$

Then

- (i) P and P_ε , for each $\varepsilon > 0$, have unique minimizers, denoted by x^* and x_ε^* , respectively;
- (ii) $\lim_{\varepsilon \rightarrow 0^+} x_\varepsilon^* = x^*$.

Proof. (i) Since $f_\varepsilon \geq f$, the function f_ε is coercive. It is also continuous, so it attains its minimum on $\mathbb{R}^d$. The function f likewise attains its minimum on the nonempty closed set K . Strict convexity of f and convexity of K give uniqueness for P .

To prove uniqueness for P_ε , we show that f_ε is strictly convex. Let $x, y \in \mathbb{R}^d$, $t \in [0, 1]$, and $x_t = (1-t)x + ty$. Convexity of c_i gives

$$c_i(x_t) \leq (1-t)c_i(x) + tc_i(y),$$

hence

$$\begin{aligned} \max(c_i(x_t), 0) &\leq \max((1-t)c_i(x) + tc_i(y), 0) \\ &\leq (1-t) \max(c_i(x), 0) + t \max(c_i(y), 0). \end{aligned}$$

All terms are nonnegative, so we may square this inequality. By convexity of $r \mapsto r^2$,

$$\begin{aligned} \max(c_i(x_t), 0)^2 &\leq [(1-t) \max(c_i(x), 0) + t \max(c_i(y), 0)]^2 \\ &\leq (1-t) \max(c_i(x), 0)^2 + t \max(c_i(y), 0)^2. \end{aligned}$$

Thus every penalty term is convex, and f_ε is strictly convex as the sum of a strictly convex function and convex functions.

- (ii) Since $x^* \in K$,

$$P_\varepsilon \leq f_\varepsilon(x^*) = f(x^*) = P.$$

Therefore

$$f(x_\varepsilon^*) + \frac{1}{\varepsilon} \sum_{i=1}^{\ell} \max(c_i(x_\varepsilon^*), 0)^2 = P_\varepsilon \leq P.$$

In particular, $f(x_\varepsilon^*) \leq P$, so every x_ε^* belongs to the compact sublevel set $\{x \mid f(x) \leq P\}$. Continuity and coercivity also imply that f is bounded below; choose $m \in \mathbb{R}$ such that $f \geq m$. For every i ,

$$\begin{aligned} \frac{1}{\varepsilon} \max(c_i(x_\varepsilon^*), 0)^2 &\leq \frac{1}{\varepsilon} \sum_{j=1}^{\ell} \max(c_j(x_\varepsilon^*), 0)^2 \\ &\leq P - f(x_\varepsilon^*) \leq P - m. \end{aligned}$$

Consequently,

$$\max(c_i(x_\varepsilon^*), 0)^2 \leq \varepsilon(P - m).$$

Consider any sequence $\varepsilon_n \rightarrow 0^+$ for which $x_{\varepsilon_n}^*$ converges to a cluster point $\bar{x}$. By continuity, $\max(c_i(\bar{x}), 0)^2 = 0$ for every i , hence $\bar{x} \in K$. Also, $f(x_{\varepsilon_n}^*) \leq P$, so $f(\bar{x}) \leq P$. Thus $\bar{x}$ minimizes f over K , and uniqueness gives $\bar{x} = x^*$. Since the family $(x_\varepsilon^*)_{\varepsilon>0}$ is bounded and every cluster point as $\varepsilon \rightarrow 0^+$ equals x^* , the whole family converges to x^* . ■

Lemma 43. Let $p(t) = \max(t, 0)^2$ for $t \in \mathbb{R}$. Then $p \in C^1(\mathbb{R})$ and $p'(t) = 2 \max(t, 0)$.

Proof. Since $p(t) = O(t^2)$ near zero, p is differentiable at zero with $p'(0) = 0$. For $t > 0$, $p(t) = t^2$ and $p'(t) = 2t$; for $t < 0$, $p(t) = 0$ and $p'(t) = 0$. Thus

$$p'(t) = \begin{cases} 0 & \text{if } t < 0, \\ 0 & \text{if } t = 0, \\ 2t & \text{if } t > 0. \end{cases}$$

Hence p' is continuous and $p'(t) = 2 \max(t, 0)$. ■

Proposition 44. Suppose that $f \in C^1(\mathbb{R}^d)$ is strictly convex and that $c_1, \dots, c_\ell \in C^1(\mathbb{R}^d)$ are convex. For $\varepsilon > 0$,

$$x_\varepsilon^* \in \arg \min_{\mathbb{R}^d} f_\varepsilon \iff \nabla f(x_\varepsilon^*) + \frac{2}{\varepsilon} \sum_{i=1}^{\ell} \max(c_i(x_\varepsilon^*), 0) \nabla c_i(x_\varepsilon^*) = 0. \quad (7.1)$$

Proof. As shown above, f_ε is strictly convex. The preceding lemma also gives $f_\varepsilon \in C^1(\mathbb{R}^d)$. The unconstrained optimality condition is therefore

$$x_\varepsilon^* \in \arg \min_{\mathbb{R}^d} f_\varepsilon \iff \nabla f_\varepsilon(x_\varepsilon^*) = 0.$$

Writing $f_\varepsilon(x) = f(x) + \varepsilon^{-1} \sum_{i=1}^{\ell} p(c_i(x))$ and applying the chain rule, we obtain

$$\begin{aligned} \nabla f_\varepsilon(x) &= \nabla f(x) + \frac{1}{\varepsilon} \sum_{i=1}^{\ell} p'(c_i(x)) \nabla c_i(x) \\ &= \nabla f(x) + \frac{2}{\varepsilon} \sum_{i=1}^{\ell} \max(c_i(x), 0) \nabla c_i(x). \end{aligned}$$

This proves the equivalence. ■

7.2 The Karush–Kuhn–Tucker theorem

The remainder of this chapter develops the Karush–Kuhn–Tucker (KKT) theorem, which characterizes optimality for convex optimization with inequality constraints. We begin with affine constraints, for which no additional constraint qualification is needed.

Theorem 45 (Karush–Kuhn–Tucker: affine constraints). *Let $f \in C^1(\mathbb{R}^d)$ be convex, and let $c_i(x) = \langle a_i | x \rangle - b_i$ for $1 \leq i \leq \ell$. Set*

$$K = \{x \in \mathbb{R}^d \mid c_i(x) \leq 0 \text{ for } 1 \leq i \leq \ell\}.$$

A point x^* minimizes f over K if and only if there exists $\lambda \in \mathbb{R}^\ell$ satisfying

$$\begin{cases} -\nabla f(x^*) = \sum_{i=1}^{\ell} \lambda_i \nabla c_i(x^*), \\ x^* \in K, \\ \lambda_i \geq 0 & \text{for } 1 \leq i \leq \ell, \\ \lambda_i c_i(x^*) = 0 & \text{for } 1 \leq i \leq \ell. \end{cases}$$

Remark 22. By analogy with the Lagrange multiplier theorem for equality constraints, the vector λ is called a vector of *Lagrange multipliers*.

Remark 23. The four conditions in the theorem have standard names:

- $\nabla f(x^*) + \sum_i \lambda_i \nabla c_i(x^*) = 0$ is *stationarity*;
- $x^* \in K$ is *primal feasibility*;
- $\lambda_i \geq 0$ is *dual feasibility*;
- $\lambda_i c_i(x^*) = 0$ is *complementary slackness*.

All four conditions must be checked when applying the KKT theorem.

We first prove necessity by passing to the limit in the optimality condition (7.1) for penalized problems.

Necessity in Theorem 45. Let x^* minimize f over K . First assume that f is strictly convex and coercive. Set $\varepsilon_N = 1/N$ and let x_N^* be the unique minimizer of the penalized problem

$$\min_{x \in \mathbb{R}^d} \left(f(x) + \frac{1}{\varepsilon_N} \sum_{i=1}^{\ell} \max(c_i(x), 0)^2 \right).$$

Existence and convergence $x_N^* \rightarrow x^*$ follow from Proposition 42. The optimality condition gives

$$\nabla f(x_N^*) + \frac{2}{\varepsilon_N} \sum_{i=1}^{\ell} \max(c_i(x_N^*), 0) \nabla c_i(x_N^*) = 0.$$

Let $I = \{i \in \{1, \dots, \ell\} \mid c_i(x^*) < 0\}$ be the set of inactive constraints. By continuity and finiteness of I , there is N_0 such that

$$\forall N \geq N_0, \quad \forall i \in I, \quad c_i(x_N^*) < 0.$$

Since $\nabla c_i = a_i$, for $N \geq N_0$ we have

$$\begin{aligned} -\nabla f(x_N^*) &= \frac{2}{\varepsilon_N} \sum_{i=1}^{\ell} \max(c_i(x_N^*), 0) a_i \\ &= \frac{2}{\varepsilon_N} \sum_{i \notin I} \max(c_i(x_N^*), 0) a_i \in C, \end{aligned}$$

where

$$C = \left\{ \sum_{i=1}^{\ell} \lambda_i a_i \mid \lambda \in \mathbb{R}_+^{\ell}, \quad \lambda_i = 0 \text{ for every } i \in I \right\}.$$

This finitely generated cone is closed by Lemma 46. Since ∇f is continuous and $x_N^* \rightarrow x^*$, it follows that $-\nabla f(x^*) \in C$. Thus there exists $\lambda \in \mathbb{R}_+^{\ell}$ such that

$$-\nabla f(x^*) = \sum_{i=1}^{\ell} \lambda_i a_i = \sum_{i=1}^{\ell} \lambda_i \nabla c_i(x^*),$$

with $\lambda_i = 0$ whenever $i \in I$. If $i \notin I$, feasibility of x^* gives $c_i(x^*) = 0$. Hence $\lambda_i c_i(x^*) = 0$ for every i , proving all four conditions.

To remove the extra assumptions of strict convexity and coercivity, define

$$\tilde{f}(x) = f(x) + \|x - x^*\|^2.$$

This function is strictly convex, and x^* is its unique minimizer on K . Convexity of f gives the affine lower bound

$$\tilde{f}(x) \geq f(x^*) + \langle \nabla f(x^*), x - x^* \rangle + \|x - x^*\|^2,$$

which tends to $+\infty$ as $\|x\| \rightarrow +\infty$. Thus $\tilde{f}$ is coercive, so the preceding argument applies to it. We obtain $\lambda \in \mathbb{R}_+^{\ell}$ satisfying

$$-\nabla \tilde{f}(x^*) = \sum_{i=1}^{\ell} \lambda_i \nabla c_i(x^*), \quad \lambda_i c_i(x^*) = 0 \quad \text{for every } i.$$

Finally, $\nabla \tilde{f}(x^*) = \nabla f(x^*)$, giving the desired multipliers for f . ■

Lemma 46. Let $a_1, \dots, a_{\ell} \in \mathbb{R}^d$. The finitely generated cone

$$C = \left\{ \sum_{i=1}^{\ell} \lambda_i a_i \mid \lambda \in \mathbb{R}_+^{\ell} \right\}$$

is closed. The cone generated by an empty family is understood to be $\{0\}$.

Proof. Suppose first that $a_1, \dots, a_{\ell}$ are linearly independent. If $v_n = \sum_i \lambda_i^n a_i \in C$ converges to v , then v belongs to their span, which is a closed finite-dimensional subspace. Coordinates in this basis depend continuously on the vector, so $\lambda_i^n \rightarrow \lambda_i \geq 0$ and $v = \sum_i \lambda_i a_i \in C$. Thus C is closed in this case.

In general, consider a representation $v = \sum_{i \in S} \lambda_i a_i$, where $\lambda_i > 0$ for $i \in S$. If the vectors indexed by S are linearly dependent, choose a nonzero relation $\sum_{i \in S} \mu_i a_i = 0$, reversing its sign if necessary so that some $\mu_i > 0$. Set

$$t = \min_{\substack{i \in S \\ \mu_i > 0}} \frac{\lambda_i}{\mu_i} > 0, \quad \lambda'_i = \lambda_i - t \mu_i.$$

If $\mu_i > 0$, the choice of t gives $\lambda'_i \geq 0$; if $\mu_i \leq 0$, then $\lambda'_i \geq \lambda_i > 0$. At least one coefficient becomes zero, while

$$v = \sum_{i \in S} \lambda'_i a_i.$$

Repeating this reduction expresses v as a nonnegative combination of a linearly independent subset of the generators (or as the empty sum when $v = 0$).

Therefore C is the union, over the finitely many linearly independent subsets of $\{a_1, \dots, a_\ell\}$, of their generated cones, together with $\{0\}$. Each of these cones is closed by the first part of the proof. A finite union of closed sets is closed. ■

Sufficiency in Theorem 45. Suppose that $x^* \in \mathbb{R}^d$ and $\lambda \in \mathbb{R}^\ell$ satisfy the four KKT conditions. Convexity and stationarity give

$$\begin{aligned} \forall x \in \mathbb{R}^d, \quad f(x) &\geq f(x^*) + \langle x - x^* | \nabla f(x^*) \rangle \\ &= f(x^*) - \sum_{i=1}^{\ell} \lambda_i \langle x - x^* | \nabla c_i(x^*) \rangle. \end{aligned} \quad (7.2)$$

Each c_i is convex (in particular, every affine function is convex), so

$$c_i(x) \geq c_i(x^*) + \langle x - x^* | \nabla c_i(x^*) \rangle.$$

Equivalently,

$$\forall x \in \mathbb{R}^d, \quad -\langle x - x^* | \nabla c_i(x^*) \rangle \geq c_i(x^*) - c_i(x). \quad (7.3)$$

Combining (7.2) and (7.3), using $\lambda_i \geq 0$, yields

$$f(x) \geq f(x^*) + \sum_{i=1}^{\ell} \lambda_i (c_i(x^*) - c_i(x)).$$

Complementary slackness gives

$$\forall x \in \mathbb{R}^d, \quad f(x) \geq f(x^*) - \sum_{i=1}^{\ell} \lambda_i c_i(x).$$

For $x \in K$, $c_i(x) \leq 0$ and $\lambda_i \geq 0$, so $\lambda_i c_i(x) \leq 0$. Therefore

$$\forall x \in K, \quad f(x) \geq f(x^*).$$

Since $x^* \in K$, this proves $x^* \in \arg \min_K f$. ■

The Karush–Kuhn–Tucker theorem: general convex constraints

Definition 20. For $x \in K$, the set

$$A(x) = \{i \in \{1, \dots, \ell\} \mid c_i(x) = 0\}$$

is called the *active set* at x .

Definition 21. The constraints $c_i(x) \leq 0$ are said to satisfy the *constraint qualification* at $x \in K$ if there exists a direction $v \in \mathbb{R}^d$ such that, for every $i \in A(x)$, either

$$\langle \nabla c_i(x) | v \rangle < 0,$$

or

$$\langle \nabla c_i(x) | v \rangle = 0 \quad \text{and} \quad c_i \text{ is affine.}$$

This gives the following version of the KKT theorem.

Theorem 47 (Karush–Kuhn–Tucker: convex constraints). *Let $f, c_1, \dots, c_\ell \in C^1(\mathbb{R}^d)$ be convex and set*

$$K = \{x \in \mathbb{R}^d \mid c_i(x) \leq 0 \text{ for } 1 \leq i \leq \ell\}.$$

Let $x^* \in K$ be a feasible point at which the constraints satisfy Definition 21. Then x^* minimizes f over K if and only if there exists $\lambda \in \mathbb{R}^\ell$ satisfying

$$\begin{cases} -\nabla f(x^*) = \sum_{i=1}^{\ell} \lambda_i \nabla c_i(x^*), \\ x^* \in K, \\ \lambda_i \geq 0 & \text{for } 1 \leq i \leq \ell, \\ \lambda_i c_i(x^*) = 0 & \text{for } 1 \leq i \leq \ell. \end{cases}$$

The KKT conditions are sufficient even without the constraint qualification; the qualification is used for necessity.

Proof. Sufficiency follows from the preceding proof, which only used convexity and differentiability of the constraint functions.

For necessity, suppose that x^* minimizes f over K , and define the cone of linearized feasible directions

$$K(x^*) = \{v \in \mathbb{R}^d \mid \langle \nabla c_i(x^*) | v \rangle \leq 0 \text{ for every } i \in A(x^*)\}.$$

Choose $\bar{v}$ as in Definition 21, and fix $v \in K(x^*)$ and $\delta > 0$. We show that $x^* + \varepsilon(v + \delta \bar{v}) \in K$ for all sufficiently small $\varepsilon > 0$. There are three cases:

- (C1) If $i \notin A(x^*)$, then $c_i(x^*) < 0$, and continuity gives $c_i(x^* + \varepsilon(v + \delta \bar{v})) < 0$ for sufficiently small $\varepsilon > 0$.
- (C2) If $i \in A(x^*)$ and $\langle \nabla c_i(x^*) | \bar{v} \rangle < 0$, differentiability gives

$$\begin{aligned} c_i(x^* + \varepsilon(v + \delta \bar{v})) &= c_i(x^*) + \varepsilon \langle \nabla c_i(x^*) | v + \delta \bar{v} \rangle + o(\varepsilon) \\ &= \underbrace{\varepsilon \langle \nabla c_i(x^*) | v \rangle}_{\leq 0} + \underbrace{\varepsilon \delta \langle \nabla c_i(x^*) | \bar{v} \rangle}_{< 0} + o(\varepsilon) < 0 \end{aligned}$$

for sufficiently small $\varepsilon > 0$.

- (C3) If $i \in A(x^*)$ and $\langle \nabla c_i(x^*) | \bar{v} \rangle = 0$, the qualification implies that c_i is affine. Hence

$$\begin{aligned} c_i(x^* + \varepsilon(v + \delta \bar{v})) &= c_i(x^*) + \varepsilon \langle \nabla c_i(x^*) | v + \delta \bar{v} \rangle \\ &= \varepsilon \langle \nabla c_i(x^*) | v \rangle \leq 0. \end{aligned}$$

Since there are only finitely many constraints, a common sufficiently small positive ε satisfies all of them. Optimality of x^* therefore gives

$$\frac{f(x^* + \varepsilon(v + \delta\bar{v})) - f(x^*)}{\varepsilon} \geq 0.$$

First let $\varepsilon \rightarrow 0^+$ to obtain

$$\langle \nabla f(x^*) | v + \delta\bar{v} \rangle \geq 0.$$

Then let $\delta \rightarrow 0^+$, giving

$$\langle \nabla f(x^*) | v \rangle \geq 0 \quad \text{for every } v \in K(x^*).$$

Apply Farkas' lemma below with $a_i = -\nabla c_i(x^*)$ for $i \in A(x^*)$ and $p = \nabla f(x^*)$. It yields nonnegative coefficients λ_i such that

$$\nabla f(x^*) = - \sum_{i \in A(x^*)} \lambda_i \nabla c_i(x^*).$$

Set $\lambda_i = 0$ for $i \notin A(x^*)$. Stationarity, dual feasibility, and complementary slackness follow; primal feasibility was assumed. If the active set is empty, the same argument uses the cone $\{0\}$ and gives $\nabla f(x^*) = 0$. ■

Lemma 48 (Farkas' lemma). *Let $a_1, \dots, a_\ell \in \mathbb{R}^d$ and define*

$$\tilde{C} = \{v \in \mathbb{R}^d \mid \langle a_i | v \rangle \geq 0 \text{ for every } i\}, \quad C = \left\{ \sum_{i=1}^{\ell} \lambda_i a_i \mid \lambda \in \mathbb{R}_+^\ell \right\}.$$

For every $p \in \mathbb{R}^d$,

$$[\langle p | v \rangle \geq 0 \text{ for every } v \in \tilde{C}] \iff p \in C.$$

Proof. If $p = \sum_i \lambda_i a_i \in C$ and $v \in \tilde{C}$, then

$$\langle p | v \rangle = \sum_i \lambda_i \langle a_i | v \rangle \geq 0.$$

Conversely, suppose that $p \notin C$. The cone C is nonempty, convex, and closed by Lemma 46, so let $q = p_C(p)$. The projection characterization gives

$$\langle p - q | z - q \rangle \leq 0 \quad \text{for every } z \in C.$$

Taking $z = 0$ and $z = 2q$ yields $\langle p - q | q \rangle = 0$. It follows that $\langle p - q | z \rangle \leq 0$ for every $z \in C$. Therefore $v = q - p$ satisfies $\langle a_i | v \rangle \geq 0$ for every i , so $v \in \tilde{C}$, whereas

$$\langle p | v \rangle = \langle q + (p - q) | q - p \rangle = -\|p - q\|^2 < 0.$$

This contradicts the proposed nonnegativity condition and proves the converse. ■

Exercises and worked solutions

Karush–Kuhn–Tucker conditions

Part I. Applying the KKT conditions

KKT reminder. Let $J \in C^1(\mathbb{R}^d)$ be convex, let $c_i(x) = \langle a_i | x \rangle - b_i$ for $1 \leq i \leq k$, and set $K = \{x \in \mathbb{R}^d \mid c_i(x) \leq 0 \text{ for all } i\}$. Theorem 45 states that

$$x^* \in \arg \min_K J \iff \exists \lambda \in \mathbb{R}^k : \begin{cases} \nabla J(x^*) + \sum_{i=1}^k \lambda_i \nabla c_i(x^*) = 0, & \text{stationarity,} \\ c_i(x^*) \leq 0 \quad (1 \leq i \leq k), & \text{primal feasibility,} \\ \lambda_i \geq 0 \quad (1 \leq i \leq k), & \text{dual feasibility,} \\ \lambda_i c_i(x^*) = 0 \quad (1 \leq i \leq k), & \text{complementary slackness.} \end{cases} \quad (7.4)$$

TD 4 / Part I / Q1

A strictly convex quadratic problem

Consider the problem of minimizing J over $K = \{x \in \mathbb{R}^2 \mid c_1(x) \leq 0, c_2(x) \leq 0\}$, where

$$J(x) = (x_1 - 2)^2 + 2(x_2 - 1)^2, \quad c_1(x) = x_1 + 4x_2 - 3, \quad c_2(x) = -x_1 + x_2.$$

- Write the system of equations and inequalities in $x^* \in \mathbb{R}^2$ and $\lambda \in \mathbb{R}^2$ that characterizes optimality by the KKT theorem.
- Find every solution of this system by considering the four cases

$$(\lambda_1, \lambda_2) = (0, 0), \quad \lambda_1 > 0, \lambda_2 = 0, \quad \lambda_1 = 0, \lambda_2 > 0, \quad \lambda_1 > 0, \lambda_2 > 0.$$

Deduce the minimizer x^* .

Worked solution

- The Hessian is $\text{diag}(2, 4)$, so J is strictly convex. It is coercive, and the closed feasible set contains the origin; hence there is a unique minimizer. Since

$$\nabla J(x) = \begin{pmatrix} 2x_1 - 4 \\ 4x_2 - 4 \end{pmatrix}, \quad \nabla c_1 = \begin{pmatrix} 1 \\ 4 \end{pmatrix}, \quad \nabla c_2 = \begin{pmatrix} -1 \\ 1 \end{pmatrix},$$

the necessary and sufficient KKT conditions are

$$\begin{cases} 2x_1 - 4 + \lambda_1 - \lambda_2 = 0, \\ 4x_2 - 4 + 4\lambda_1 + \lambda_2 = 0, \\ x_1 + 4x_2 \leq 3, \quad x_2 \leq x_1, \\ \lambda_1 \geq 0, \quad \lambda_2 \geq 0, \\ \lambda_1(x_1 + 4x_2 - 3) = 0, \quad \lambda_2(-x_1 + x_2) = 0. \end{cases}$$

- We solve each multiplier case.
 - If $\lambda_1 = \lambda_2 = 0$, stationarity gives $x = (2, 1)$. Since $c_1(2, 1) = 3 > 0$, this point is infeasible.

- If $\lambda_1 > 0$ and $\lambda_2 = 0$, stationarity gives $x_1 = 2 - \lambda_1/2$ and $x_2 = 1 - \lambda_1$. Complementary slackness requires

$$0 = c_1(x) = 3 - \frac{9}{2}\lambda_1.$$

Thus $\lambda_1 = 2/3$ and $x = (5/3, 1/3)$. Here $c_2(x) = -4/3 < 0$, so all conditions hold.

- If $\lambda_1 = 0$ and $\lambda_2 > 0$, complementary slackness gives $x_1 = x_2 = t$. Stationarity becomes $2t - 4 - \lambda_2 = 0$ and $4t - 4 + \lambda_2 = 0$. Hence $t = 4/3$ and $\lambda_2 = -4/3$, contradicting positivity.
- If both multipliers are positive, both constraints are active: $x_1 = x_2 = 3/5$. Stationarity then gives

$$\lambda_1 - \lambda_2 = \frac{14}{5}, \quad 4\lambda_1 + \lambda_2 = \frac{8}{5},$$

so $\lambda_1 = 22/25$ and $\lambda_2 = -48/25$. This also violates dual feasibility.

Consequently, the only KKT pair is

$$x^* = \left(\frac{5}{3}, \frac{1}{3}\right), \quad \lambda^* = \left(\frac{2}{3}, 0\right), \quad J(x^*) = 1.$$

TD 4 / Part I / Q2

A nonconvex quadratic: candidates and global optimality

Consider

$$J(x) = (x_1 - 2)^2 + 2(x_2 - 1)^2 + 5x_1x_2,$$

$$c_1(x) = -x_1, \quad c_2(x) = x_1 - 2x_2, \quad K = \{x \mid c_1(x) \leq 0, c_2(x) \leq 0\}.$$

- Prove that this problem has at least one minimizer.
- Check whether J is convex, and write the KKT conditions that are *necessary* at a minimizer. Explain why these conditions alone do not guarantee optimality in this example.
- Find every KKT pair by considering the same four multiplier cases as in Q1. Compare the objective values of the candidate points.
- Determine all global minimizers and justify your conclusion.

Worked solution

- The feasible set is

$$K = \{(x_1, x_2) \mid x_1 \geq 0, x_2 \geq x_1/2\}.$$

It is closed and nonempty. On K , $x_1x_2 \geq 0$, so

$$J(x) \geq (x_1 - 2)^2 + 2(x_2 - 1)^2 \longrightarrow +\infty \quad \text{as } \|x\| \rightarrow +\infty, x \in K.$$

Thus J is coercive on K . By continuity, it attains its minimum there. Global coercivity on all of $\mathbb{R}^2$ is not needed.

- We have

$$\nabla J(x) = \begin{pmatrix} 2x_1 + 5x_2 - 4 \\ 5x_1 + 4x_2 - 4 \end{pmatrix}, \quad D^2J = \begin{pmatrix} 2 & 5 \\ 5 & 4 \end{pmatrix}.$$

The determinant is $8 - 25 = -17 < 0$, so the Hessian is indefinite. The objective is not convex, and the sufficient direction of the convex KKT theorem does not apply.

KKT conditions remain necessary for a differentiable objective with these affine constraints. Indeed, at a local minimizer the derivative of J is nonnegative in every linearized feasible direction; for affine constraints these are feasible directions for sufficiently small positive steps. Farkas' lemma then gives nonnegative multipliers for the active constraints. Equivalently, here the direction $(1, 1)$ strictly decreases both constraint functions, so the constraint qualification holds at every feasible point.

The necessary system is

$$\begin{cases} 2x_1 + 5x_2 - 4 - \lambda_1 + \lambda_2 = 0, \\ 5x_1 + 4x_2 - 4 - 2\lambda_2 = 0, \\ x_1 \geq 0, & x_1 - 2x_2 \leq 0, \\ \lambda_1 \geq 0, & \lambda_2 \geq 0, \\ \lambda_1 x_1 = 0, & \lambda_2 (x_1 - 2x_2) = 0. \end{cases}$$

(c) The four cases give:

- If $\lambda_1 = \lambda_2 = 0$, solve $2x_1 + 5x_2 = 4$ and $5x_1 + 4x_2 = 4$. This gives

$$x = \left(\frac{4}{17}, \frac{12}{17} \right).$$

Both constraints are strictly satisfied, so this is a KKT point.

- If $\lambda_1 > 0$ and $\lambda_2 = 0$, then $x_1 = 0$. The second stationarity equation gives $x_2 = 1$, and the first gives $\lambda_1 = 1$. This pair is admissible.
- If $\lambda_1 = 0$ and $\lambda_2 > 0$, then $x_2 = x_1/2$. The equations become

$$\frac{9}{2}x_1 - 4 + \lambda_2 = 0, \quad 7x_1 - 4 - 2\lambda_2 = 0.$$

Hence $x_1 = 3/4$, $x_2 = 3/8$, and $\lambda_2 = 5/8$. This pair is admissible.

- If both multipliers are positive, both constraints are active, so $x = (0, 0)$. Stationarity gives $\lambda_2 = -2$ and $\lambda_1 = -6$, which are inadmissible.

Thus all KKT pairs and objective values are

x	λ	$J(x)$
$(4/17, 12/17)$	$(0, 0)$	$70/17$
$(0, 1)$	$(1, 0)$	4
$(3/4, 3/8)$	$(0, 5/8)$	$15/4$

In particular, KKT conditions alone are not sufficient: the feasible KKT point $(0, 1)$ has a larger objective value than $(3/4, 3/8)$.

(d) A global minimizer exists by (a), and every global minimizer must be one of the three KKT points found in (c). Since

$$\frac{15}{4} < 4 < \frac{70}{17},$$

the unique global minimizer and minimum value are

$$x^* = \left(\frac{3}{4}, \frac{3}{8} \right), \quad \min_K J = \frac{15}{4}.$$

This conclusion uses existence, necessity, and an exhaustive comparison of candidates, rather than a convexity argument.

TD 4 / Part I / Q3

Matrix form of the KKT conditions

For vectors, write $x \geq 0$ when every coordinate is nonnegative, and $x \geq y$ when $x - y \geq 0$. Assume the hypotheses of the affine KKT reminder above. Let $A \in \mathcal{M}_{k,d}(\mathbb{R})$ have rows a_i^T , and let $b = (b_1, \dots, b_k)^T$.

- (a) Show that $K = \{x \in \mathbb{R}^d \mid Ax \leq b\}$.
- (b) Show that the KKT theorem can be written as

$$x^* \in \arg \min_K J \iff \exists \lambda \in \mathbb{R}^k : \begin{cases} -\nabla J(x^*) = A^T \lambda, \\ Ax^* \leq b, \quad \lambda \geq 0, \\ \lambda_i(\langle a_i | x^* \rangle - b_i) = 0 \quad (1 \leq i \leq k). \end{cases} \quad (7.5)$$

Worked solution

- (a) By the definition of matrix multiplication, $(Ax - b)_i = \langle a_i | x \rangle - b_i = c_i(x)$. Therefore the k scalar inequalities $c_i(x) \leq 0$ are exactly $Ax \leq b$.
- (b) Because $\nabla c_i = a_i$,

$$\sum_{i=1}^k \lambda_i \nabla c_i = A^T \lambda.$$

Substituting this identity and $c_i(x^*) = (Ax^* - b)_i$ into (7.4) gives (7.5). Under primal and dual feasibility, complementary slackness can also be written $\langle \lambda | Ax^* - b \rangle = 0$: every summand is nonpositive, so the sum vanishes exactly when every summand vanishes.

TD 4 / Part I / Q4

Equality and inequality constraints together

Let $J \in C^1(\mathbb{R}^d)$ be convex and

$$K = \{x \in \mathbb{R}^d \mid Ax \leq b, Ex = f\}, \quad A \in \mathcal{M}_{k,d}(\mathbb{R}), \quad E \in \mathcal{M}_{\ell,d}(\mathbb{R}),$$

where $b \in \mathbb{R}^k$ and $f \in \mathbb{R}^\ell$ are fixed vectors. Show that

$$x^* \in \arg \min_K J \iff \exists (\lambda, \mu) \in \mathbb{R}^k \times \mathbb{R}^\ell : \begin{cases} -\nabla J(x^*) = A^T \lambda + E^T \mu, \\ Ax^* \leq b, \quad Ex^* = f, \\ \lambda \geq 0, \\ \lambda_i(\langle a_i | x^* \rangle - b_i) = 0 \quad (1 \leq i \leq k). \end{cases} \quad (7.6)$$

Worked solution

Replace the equality $Ex = f$ by the two vector inequalities $Ex - f \leq 0$ and $-Ex + f \leq 0$. The affine KKT theorem gives multipliers $\lambda \geq 0$, $\alpha \geq 0$, and $\beta \geq 0$ with

$$-\nabla J(x^*) = A^T \lambda + E^T \alpha - E^T \beta.$$

Set $\mu = \alpha - \beta \in \mathbb{R}^\ell$. Complementary slackness for these equality constraints is automatic because both $Ex^* - f$ and $-Ex^* + f$ are zero. The remaining conditions are precisely (7.6).

Conversely, given λ, μ satisfying (7.6), write $\mu = \mu^+ - \mu^-$, where $\mu_j^+ = \max(\mu_j, 0)$ and $\mu_j^- = \max(-\mu_j, 0)$. Then $\alpha = \mu^+$ and $\beta = \mu^-$ are nonnegative multipliers for the two inequalities representing each equality, and they satisfy all affine KKT conditions. Hence x^* is a minimizer. In particular, equality multipliers μ have no sign restriction.

Part II. Proving KKT for one nonlinear constraint

Let $J, c \in C^1(\mathbb{R}^d)$ be convex, and assume that $K = \{x \in \mathbb{R}^d \mid c(x) \leq 0\}$ is nonempty. We study

$$x^* \in \arg \min_K J \iff \exists \lambda \in \mathbb{R} : \begin{cases} c(x^*) \leq 0, \\ \lambda \geq 0, \\ \lambda c(x^*) = 0, \\ -\nabla J(x^*) = \lambda \nabla c(x^*). \end{cases} \quad (7.7)$$

The sufficient direction holds under these assumptions. For the necessary direction we will impose the constraint qualification below.

TD 4 / Part II / Q1

Sufficiency of the KKT conditions

Suppose that x^* and λ satisfy the four conditions on the right-hand side of (7.7).

- (a) Prove optimality when $\lambda = 0$.
- (b) Suppose $\lambda > 0$, so that $c(x^*) = 0$. Prove that, for every $x \in K$,

$$\begin{aligned} J(x) &\geq J(x^*) + \langle \nabla J(x^*) | x - x^* \rangle, \\ 0 &\geq c(x) - c(x^*) \geq \langle \nabla c(x^*) | x - x^* \rangle. \end{aligned}$$

- (c) Conclude that $J(x) \geq J(x^*)$ for every $x \in K$.

Worked solution

- (a) If $\lambda = 0$, stationarity gives $\nabla J(x^*) = 0$. Convexity implies $J(x) \geq J(x^*) + \langle \nabla J(x^*) | x - x^* \rangle = J(x^*)$ for every $x \in \mathbb{R}^d$. Since $x^* \in K$, it is also optimal over K .
- (b) The first inequality is the supporting-hyperplane inequality for the differentiable convex function J . Similarly, convexity of c gives $c(x) - c(x^*) \geq \langle \nabla c(x^*) | x - x^* \rangle$. Since $c(x) \leq 0$ and $c(x^*) = 0$, the difference is nonpositive.
- (c) By stationarity and (b),

$$\langle \nabla J(x^*) | x - x^* \rangle = -\lambda \langle \nabla c(x^*) | x - x^* \rangle \geq 0.$$

Combining this with the lower bound for J proves $J(x) \geq J(x^*)$ on K . Feasibility of x^* completes the proof.

Assumptions for the penalty argument. From now on, assume the constraint qualification

$$\forall x \in \mathbb{R}^d, \quad c(x) = 0 \implies \nabla c(x) \neq 0. \quad (7.8)$$

For Q3–Q5, suppose also that J is strictly convex and coercive: $J(x) \rightarrow +\infty$ as $\|x\| \rightarrow +\infty$. For $\varepsilon > 0$, define

$$P = \min_{x \in K} J(x), \quad P_\varepsilon = \min_{x \in \mathbb{R}^d} J_\varepsilon(x),$$

$$J_\varepsilon(x) = J(x) + \frac{1}{\varepsilon} \max(c(x), 0)^2.$$

The nonemptiness of K remains part of the assumptions.

TD 4 / Part II / Q2

Qualification for a ball constraint

Verify (7.8) for $c(x) = \|x\|^2 - 1$ and identify K .

Worked solution

Here $\nabla c(x) = 2x$. If $c(x) = 0$, then $\|x\| = 1$, and therefore $\|\nabla c(x)\| = 2 \neq 0$. The qualification holds. Moreover,

$$K = \{x \in \mathbb{R}^d \mid \|x\| \leq 1\}$$

is the closed Euclidean unit ball.

TD 4 / Part II / Q3

Existence and uniqueness of the minimizers

Justify the existence and uniqueness of the minimizer x^* of P and of the minimizer x_ε^* of P_ε for every $\varepsilon > 0$.

Worked solution

The function c is continuous and convex, so K is closed and convex. It is nonempty by assumption. Continuity and coercivity of J give existence of a minimizer on K , and strict convexity gives uniqueness.

The function $p(t) = \max(t, 0)^2$ is convex and nondecreasing on $\mathbb{R}$. Since c is convex, $p \circ c$ is convex: for $t \in [0, 1]$,

$$\begin{aligned} p(c((1-t)x + ty)) &\leq p((1-t)c(x) + tc(y)) \\ &\leq (1-t)p(c(x)) + tp(c(y)). \end{aligned}$$

Consequently, $J_\varepsilon = J + \varepsilon^{-1}p \circ c$ is strictly convex. It is continuous, and $J_\varepsilon \geq J$ makes it coercive. It therefore has a unique minimizer on $\mathbb{R}^d$.

TD 4 / Part II / Q4

Stationarity of the penalized problem

Show that x_ε^* satisfies

$$\nabla J(x_\varepsilon^*) + \lambda_\varepsilon \nabla c(x_\varepsilon^*) = 0, \quad \lambda_\varepsilon = \frac{2}{\varepsilon} \max(c(x_\varepsilon^*), 0). \quad (7.9)$$

For Q5 we will also use a sequence $\varepsilon_i \rightarrow 0^+$ such that $x_{\varepsilon_i}^* \rightarrow x^*$. The convergence needed here follows from the penalty method; its proof is included below.

Worked solution

The function $p(t) = \max(t, 0)^2$ is continuously differentiable with $p'(t) = 2 \max(t, 0)$, including at $t = 0$. The chain rule gives

$$\nabla J_\varepsilon(x) = \nabla J(x) + \frac{2}{\varepsilon} \max(c(x), 0) \nabla c(x).$$

Since x_ε^* is an unconstrained minimizer, $\nabla J_\varepsilon(x_\varepsilon^*) = 0$, proving (7.9). Notice that $\lambda_\varepsilon \geq 0$.

Convergence needed for Q5. Let $m = \min_{\mathbb{R}^d} J$, which is finite by continuity and coercivity. Since $x^* \in K$, optimality of x_ε^* gives

$$J(x_\varepsilon^*) + \frac{1}{\varepsilon} \max(c(x_\varepsilon^*), 0)^2 = J_\varepsilon(x_\varepsilon^*) \leq J_\varepsilon(x^*) = J(x^*).$$

In particular, x_ε^* lies in the fixed compact sublevel set $\{x \mid J(x) \leq J(x^*)\}$, and

$$0 \leq \max(c(x_\varepsilon^*), 0)^2 \leq \varepsilon (J(x^*) - m).$$

For any sequence $\varepsilon_i \rightarrow 0^+$, take a convergent subsequence of $(x_{\varepsilon_i}^*)$. Its limit $\bar{x}$ satisfies $c(\bar{x}) \leq 0$ and $J(\bar{x}) \leq J(x^*)$ by continuity. Thus $\bar{x}$ is a constrained minimizer, and uniqueness implies $\bar{x} = x^*$. Every cluster point is therefore x^* ; compactness yields the stronger conclusion

$$\lim_{\varepsilon \rightarrow 0^+} x_\varepsilon^* = x^*.$$

This in particular supplies the sequence required in Q5.

TD 4 / Part II / Q5

Necessity and bounded multipliers

Prove the necessary direction in (7.7).

- (a) Suppose $c(x^*) < 0$. Show that $\lambda_{\varepsilon_i} = 0$ for all sufficiently large i . Use (7.9) to obtain $-\nabla J(x^*) = 0 = \lambda \nabla c(x^*)$ with $\lambda = 0$.
- (b) Suppose $c(x^*) = 0$. Use the qualification and (7.9) to show that $(\lambda_{\varepsilon_i})_i$ is bounded. Deduce the existence of a multiplier satisfying every condition in (7.7).

Worked solution

- (a) Since $x_{\varepsilon_i}^* \rightarrow x^*$ and c is continuous, $c(x_{\varepsilon_i}^*) < 0$ for all sufficiently large i . Hence $\lambda_{\varepsilon_i} = 0$, and the stationarity equation becomes $\nabla J(x_{\varepsilon_i}^*) = 0$. Taking the limit yields $\nabla J(x^*) = 0$. With $\lambda = 0$, dual feasibility and complementary slackness hold, and $x^* \in K$ is already known.
- (b) The qualification gives $a := \|\nabla c(x^*)\| > 0$. Continuity of ∇c and $x_{\varepsilon_i}^* \rightarrow x^*$ imply

$$\|\nabla c(x_{\varepsilon_i}^*)\| \geq \frac{a}{2}$$

for all sufficiently large i . Similarly, $\|\nabla J(x_{\varepsilon_i}^*)\| \leq M$ for some finite M . Taking norms in (7.9) and using $\lambda_{\varepsilon_i} \geq 0$ gives

$$0 \leq \lambda_{\varepsilon_i} = \frac{\|\nabla J(x_{\varepsilon_i}^*)\|}{\|\nabla c(x_{\varepsilon_i}^*)\|} \leq \frac{2M}{a}.$$

The finitely many earlier terms are finite, so the whole sequence is bounded. Pass to a subsequence with $\lambda_{\varepsilon_i} \rightarrow \lambda \geq 0$. Taking the limit in stationarity gives

$$-\nabla J(x^*) = \lambda \nabla c(x^*).$$

Since $c(x^*) = 0$, complementary slackness is automatic, and all four KKT conditions hold.

Removing the extra assumptions on the objective. The preceding proof used strict convexity and coercivity to control the penalized problems. For a general convex $J \in C^1(\mathbb{R}^d)$ and an existing minimizer $x^* \in K$, set

$$\tilde{J}(x) = J(x) + \|x - x^*\|^2.$$

This function is strictly convex and has x^* as its unique constrained minimizer. The bound

$$\tilde{J}(x) \geq J(x^*) + \langle \nabla J(x^*), x - x^* \rangle + \|x - x^*\|^2$$

shows that it is coercive. Applying the proved necessity result to $\tilde{J}$ yields a multiplier, and $\nabla \tilde{J}(x^*) = \nabla J(x^*)$ gives the same KKT conditions for J . Thus, under the qualification, the equivalence (7.7) holds for every convex differentiable objective.

Lagrangian duality

Motivation: computing Lagrange multipliers

We again consider a constrained optimization problem:

$$P := \inf_{x \in K} f(x), \quad K = \{x \in \mathbb{R}^d \mid c_i(x) \leq 0, i = 1, \dots, \ell\}, \quad (8.1)$$

where

- (i) $f \in C^1(\mathbb{R}^d)$ is convex;
- (ii) $c_i(x) = \langle a_i | x \rangle - b_i$ is affine.

When the infimum is attained, we write it as a minimum. The following proposition shows that knowing the Lagrange multipliers supplied by the Karush–Kuhn–Tucker theorem (Theorem 45) allows us to recover a constrained minimizer by solving an unconstrained problem.

Proposition 49. *Let x^* be a minimizer of (8.1), and let $\lambda \in \mathbb{R}_+^\ell$ be associated KKT multipliers. Then*

$$x^* \in \arg \min_{x \in \mathbb{R}^d} f_\lambda(x), \quad f_\lambda = f + \sum_{i=1}^{\ell} \lambda_i c_i. \quad (8.2)$$

Proof. The KKT theorem gives

$$-\nabla f(x^*) = \sum_{i=1}^{\ell} \lambda_i \nabla c_i(x^*), \quad \lambda_i \geq 0.$$

The function f_λ is convex, since it is the sum of the convex function f and affine functions. Stationarity gives $\nabla f_\lambda(x^*) = 0$, which characterizes a global minimizer of this differentiable convex function on $\mathbb{R}^d$. ■

How can these multipliers be computed in practice? We will see that they solve a second optimization problem, called the *dual problem*. This leads to Uzawa's algorithm, a simple method for optimization with inequality constraints.

8.1 The dual problem and weak duality

Define the *Lagrangian*

$$L : \mathbb{R}^d \times \mathbb{R}^\ell \rightarrow \mathbb{R}, \quad L(x, \lambda) = f(x) + \sum_{i=1}^{\ell} \lambda_i c_i(x). \quad (8.3)$$

The constrained problem can be written as an infimum in x of a supremum in λ . This minimax formulation is valid even when no saddle point exists.

Proposition 50 (Minimax formulation). *Let $f, c_1, \dots, c_\ell : \mathbb{R}^d \rightarrow \mathbb{R}$ be arbitrary functions, let $K = \{x \mid c_i(x) \leq 0 \text{ for all } i\}$, and let L be defined by (8.3). Then, with the convention $\inf \emptyset = +\infty$,*

$$P = \inf_{x \in K} f(x) = \inf_{x \in \mathbb{R}^d} \sup_{\lambda \in \mathbb{R}_+^\ell} L(x, \lambda). \quad (8.4)$$

Proof. First observe that, for each i ,

$$\sup_{\lambda_i \geq 0} \lambda_i c_i(x) = \begin{cases} 0 & \text{if } c_i(x) \leq 0, \\ +\infty & \text{if } c_i(x) > 0. \end{cases} \quad (8.5)$$

Indeed, in the first case every product is nonpositive, and equality is attained at $\lambda_i = 0$. In the second case, taking $\lambda_i = n \rightarrow +\infty$ makes the product tend to $+\infty$. Since the coordinates of λ can be chosen independently,

$$\begin{aligned} \sup_{\lambda \in \mathbb{R}_+^\ell} \sum_{i=1}^{\ell} \lambda_i c_i(x) &= \sum_{i=1}^{\ell} \sup_{\lambda_i \geq 0} \lambda_i c_i(x) \\ &= \begin{cases} 0 & \text{if } x \in K, \\ +\infty & \text{if } x \notin K. \end{cases} \end{aligned}$$

Consequently,

$$\sup_{\lambda \in \mathbb{R}_+^\ell} L(x, \lambda) = \begin{cases} f(x) & \text{if } x \in K, \\ +\infty & \text{otherwise.} \end{cases}$$

Taking the infimum over $x \in \mathbb{R}^d$ proves (8.4). ■

Definition 22 (Dual problem). The dual problem associated with (8.4) is

$$D := \sup_{\lambda \in \mathbb{R}_+^\ell} \inf_{x \in \mathbb{R}^d} L(x, \lambda) = \sup_{\lambda \in \mathbb{R}_+^\ell} g(\lambda), \quad g(\lambda) := \inf_{x \in \mathbb{R}^d} L(x, \lambda). \quad (8.6)$$

The function g is called the *dual function*; it may take the value $-\infty$.

Remark 24. The dual problem is obtained by interchanging the infimum and supremum in the primal minimax formulation. The dual function is concave because it is a pointwise infimum of affine functions of λ .

Example 13 (Computing a dual problem). Consider $f(x) = \frac{1}{2} \|x\|^2$ on $\mathbb{R}^d$ and $c_i(x) = \langle a_i | x \rangle - b_i$, for $i = 1, \dots, \ell$. Let $A \in \mathcal{M}_{\ell, d}(\mathbb{R})$ have rows a_i^T , and let $b = (b_1, \dots, b_\ell)^T$. Then $(Ax - b)_i = \langle a_i | x \rangle - b_i$, so

$$L(x, \lambda) = \frac{1}{2} \|x\|^2 + \sum_{i=1}^{\ell} \lambda_i (\langle a_i | x \rangle - b_i) = \frac{1}{2} \|x\|^2 + \langle \lambda | Ax - b \rangle.$$

For any $\lambda \in \mathbb{R}^\ell$, the function $f_\lambda = L(\cdot, \lambda)$ is strictly convex, and

$$\nabla f_\lambda(x) = x + A^T \lambda.$$

Its unique minimizer is therefore $x_\lambda = -A^T \lambda$. Substitution gives

$$\begin{aligned} g(\lambda) &= L(x_\lambda, \lambda) \\ &= \frac{1}{2} \|-A^T \lambda\|^2 + \langle \lambda | -AA^T \lambda - b \rangle \\ &= -\frac{1}{2} \|A^T \lambda\|^2 - \langle \lambda | b \rangle. \end{aligned}$$

Thus the dual problem is

$$D = \sup_{\lambda \in \mathbb{R}_+^\ell} \left(-\frac{1}{2} \|A^T \lambda\|^2 - \langle \lambda | b \rangle \right).$$

Remark 25. The dual problem can have far fewer variables than the primal problem when $\ell \ll d$. Its constraints also have a particularly simple form: $\lambda_i \geq 0$.

Proposition 51 (Weak duality). Let $f, c_1, \dots, c_\ell$ be real-valued functions on $\mathbb{R}^d$. The primal and dual values satisfy

$$D \leq P. \quad (8.7)$$

Remark 26. The inequality $D \leq P$ is called *weak duality*; equality $D = P$ is called *strong duality*. Weak duality always holds. Strong duality requires additional assumptions.

Proof. For every $x \in K$ and every $\lambda \in \mathbb{R}_+^\ell$,

$$g(\lambda) = \inf_{\tilde{x} \in \mathbb{R}^d} L(\tilde{x}, \lambda) \leq L(x, \lambda) = f(x) + \sum_{i=1}^{\ell} \lambda_i c_i(x) \leq f(x).$$

Taking the supremum over $\lambda \geq 0$, then the infimum over $x \in K$, gives $D \leq P$. If $K = \emptyset$, then $P = +\infty$ and the inequality is immediate. ■

8.2 Strong duality

Theorem 52 (Strong duality). *Let $f \in C^1(\mathbb{R}^d)$ be convex, let $c_i(x) = \langle a_i | x \rangle - b_i$ for $i = 1, \dots, \ell$, and let $K = \{x \in \mathbb{R}^d \mid c_i(x) \leq 0 \text{ for all } i\}$.*

- (i) *If the primal problem (8.1) has a minimizer x^* , then $P = D$, and the dual supremum is attained.*
- (ii) *If x^* solves the primal problem and $\lambda^* \in \mathbb{R}_+^\ell$ maximizes g , then*

$$x^* \in \arg \min_{x \in \mathbb{R}^d} L(x, \lambda^*).$$

Proof. (i) The affine-constraint KKT theorem gives $\lambda \in \mathbb{R}^\ell$ such that

$$\begin{cases} \nabla f(x^*) + \sum_{i=1}^\ell \lambda_i \nabla c_i(x^*) = 0, \\ x^* \in K, \\ \lambda_i \geq 0 & i = 1, \dots, \ell, \\ \lambda_i c_i(x^*) = 0 & i = 1, \dots, \ell. \end{cases}$$

The first condition implies $\nabla f_\lambda(x^*) = 0$ for the convex function $f_\lambda = L(\cdot, \lambda)$, hence x^* minimizes f_λ on $\mathbb{R}^d$. Complementary slackness therefore yields

$$\begin{aligned} D &\geq g(\lambda) = L(x^*, \lambda) \\ &= f(x^*) + \sum_{i=1}^\ell \lambda_i c_i(x^*) = f(x^*) = P. \end{aligned}$$

Together with weak duality, this gives $D = P$. Since $g(\lambda) = D$, the dual optimum is attained at this multiplier vector.

- (ii) Let x^* and λ^* be arbitrary primal and dual optimizers. By (i),

$$\begin{aligned} f(x^*) = P = D = g(\lambda^*) &= \inf_{x \in \mathbb{R}^d} L(x, \lambda^*) \\ &\leq L(x^*, \lambda^*) \\ &= f(x^*) + \sum_{i=1}^\ell \lambda_i^* c_i(x^*) \leq f(x^*). \end{aligned} \tag{8.8}$$

The last inequality follows from feasibility and $\lambda_i^* \geq 0$. All inequalities must be equalities. In particular,

$$\sum_{i=1}^\ell \lambda_i^* c_i(x^*) = 0, \quad \inf_{x \in \mathbb{R}^d} L(x, \lambda^*) = L(x^*, \lambda^*),$$

which proves (ii). ■

Remark 27 (Saddle points). Under the preceding hypotheses, a primal–dual optimal pair satisfies

$$L(x^*, \lambda) \leq L(x^*, \lambda^*) \leq L(x, \lambda^*) \quad \text{for all } x \in \mathbb{R}^d, \lambda \in \mathbb{R}_+^\ell.$$

Such a pair is a *saddle point*. Unlike the minimax identity (8.4), the existence of a saddle point includes attainment and strong duality.

8.3 Uzawa's algorithm

Notation and assumptions Vector inequalities are interpreted coordinatewise. Let $A \in \mathcal{M}_{\ell,d}(\mathbb{R})$, $b \in \mathbb{R}^\ell$, and

$$K = \{x \in \mathbb{R}^d \mid Ax \leq b\} = \{x \in \mathbb{R}^d \mid c_i(x) \leq 0, i = 1, \dots, \ell\},$$

where $c_i(x) = \langle a_i | x \rangle - b_i$ and a_i^T is the i th row of A . Assume $K \neq \emptyset$ and that the primal minimum is attained. Under the hypotheses of Theorem 52,

$$P = \min_{x \in K} f(x) = \min_{x \in \mathbb{R}^d} (f(x) + \langle \lambda^* | Ax - b \rangle),$$

where $\lambda^* \in \mathbb{R}_+^\ell$ is a dual maximizer and

$$D = \max_{\lambda \in \mathbb{R}_+^\ell} g(\lambda), \quad g(\lambda) = \inf_{x \in \mathbb{R}^d} L(x, \lambda), \quad L(x, \lambda) = f(x) + \langle \lambda | Ax - b \rangle.$$

Uzawa's algorithm applies projected gradient ascent to this dual problem. To write it as a gradient method, the dual function must be differentiable and the inner minimizations must be solvable; convexity of f alone does not guarantee these properties. They hold, in particular, in the positive-definite quadratic case below, where we will compute g and its gradient explicitly.

Starting from any $\lambda^{(0)} \in \mathbb{R}_+^\ell$ (typically $\lambda^{(0)} = 0$), the method takes the form

$$\begin{cases} x^{(k)} \in \arg \min_{x \in \mathbb{R}^d} L(x, \lambda^{(k)}), \\ \gamma^{(k)} = \nabla g(\lambda^{(k)}), \\ \lambda^{(k+1)} = \mathbf{p}_{\mathbb{R}_+^\ell}(\lambda^{(k)} + \tau \gamma^{(k)}). \end{cases}$$

Remark 28. Because we maximize g , the update uses gradient *ascent*. Projection onto $\mathbb{R}_+^\ell$ replaces every negative coordinate by zero.

Uzawa's algorithm for a quadratic objective

From now on, let $Q \in \mathcal{S}_d(\mathbb{R})$ be positive definite and $e \in \mathbb{R}^d$. Consider

$$P = \min_{x \in K} f(x), \quad f(x) = \frac{1}{2} \langle x | Qx \rangle + \langle e | x \rangle, \quad K = \{x \in \mathbb{R}^d \mid Ax \leq b\} \neq \emptyset. \quad (8.9)$$

Lemma 53. The dual function g is a differentiable concave quadratic on $\mathbb{R}^\ell$, with

$$\begin{cases} g(\lambda) = -\frac{1}{2} \langle x_\lambda | Qx_\lambda \rangle - \langle \lambda | b \rangle, \\ \nabla g(\lambda) = Ax_\lambda - b, \end{cases}$$

where

$$x_\lambda = -Q^{-1}(e + A^T \lambda)$$

is the unique minimizer of $L(\cdot, \lambda)$ on $\mathbb{R}^d$.

Proof. The Lagrangian is

$$\begin{aligned} L(x, \lambda) &= f(x) + \sum_{i=1}^{\ell} \lambda_i (\langle a_i | x \rangle - b_i) \\ &= \frac{1}{2} \langle x | Qx \rangle + \langle e | x \rangle + \langle \lambda | Ax - b \rangle. \end{aligned}$$

For fixed λ , the Hessian in x is Q , so $f_\lambda = L(\cdot, \lambda)$ is strictly convex. Its gradient vanishes exactly when

$$Qx + e + A^T \lambda = 0,$$

which gives the unique minimizer $x_\lambda = -Q^{-1}(e + A^T \lambda)$. Using this identity,

$$\begin{aligned} g(\lambda) &= L(x_\lambda, \lambda) \\ &= \frac{1}{2} \langle x_\lambda | Q x_\lambda \rangle + \langle e + A^T \lambda | x_\lambda \rangle - \langle \lambda | b \rangle \\ &= -\frac{1}{2} \langle x_\lambda | Q x_\lambda \rangle - \langle \lambda | b \rangle \\ &= -\frac{1}{2} \langle Q^{-1}(e + A^T \lambda) | e + A^T \lambda \rangle - \langle \lambda | b \rangle \\ &= -\frac{1}{2} \langle A Q^{-1} A^T \lambda | \lambda \rangle - \langle A Q^{-1} e | \lambda \rangle - \frac{1}{2} \langle Q^{-1} e | e \rangle - \langle \lambda | b \rangle. \end{aligned}$$

The matrix $AQ^{-1}A^T$ is symmetric positive semidefinite: for any $v \in \mathbb{R}^\ell$,

$$\langle A Q^{-1} A^T v | v \rangle = \langle Q^{-1} A^T v | A^T v \rangle \geq 0.$$

Hence g is concave. Differentiating its explicit expression yields

$$\nabla g(\lambda) = -A Q^{-1} A^T \lambda - A Q^{-1} e - b = A x_\lambda - b.$$

■

Corollary 54. For the constrained quadratic problem (8.9), Uzawa's algorithm is

$$\begin{cases} x^{(k)} = -Q^{-1}(e + A^T \lambda^{(k)}), \\ \gamma^{(k)} = A x^{(k)} - b, \\ \lambda^{(k+1)} = \max(\lambda^{(k)} + \tau \gamma^{(k)}, 0), \end{cases} \quad (8.10)$$

where the maximum is taken coordinatewise.

Convergence of Uzawa's algorithm

We conclude by proving convergence of the primal iterates generated by Uzawa's algorithm.

Theorem 55. Assume $K \neq \emptyset$ in (8.9). Let $(x^{(k)}, \lambda^{(k)})$ be generated by (8.10), starting from any $\lambda^{(0)} \in \mathbb{R}_+^\ell$. Let $m > 0$ be the smallest eigenvalue of Q and let $\|A\|$ be the Euclidean operator norm of A . If $A \neq 0$, assume

$$0 < \tau < \frac{2m}{\|A\|^2}.$$

If $A = 0$, any $\tau > 0$ is allowed. Then

$$\lim_{k \rightarrow \infty} x^{(k)} = x^*,$$

where x^* is the unique minimizer of (8.9).

Proof. The function f is continuous, strictly convex, and coercive because Q is positive definite. Since K is nonempty and closed, it has a unique minimizer x^* on K . Strong duality

gives a dual maximizer λ^* , which need not be unique. Since $g \in C^1(\mathbb{R}^\ell)$ is concave, the projected optimality condition (Theorem 37, applied to $-g$) gives

$$\lambda^* = \mathbf{p}_{\mathbb{R}_+^\ell}(\lambda^* + \tau \nabla g(\lambda^*)) \quad \text{for every } \tau > 0.$$

By strong duality and Lemma 53,

$$x^* = x_{\lambda^*} = -Q^{-1}(e + A^T \lambda^*), \quad \nabla g(\lambda^*) = Ax^* - b.$$

Set $u_k = \lambda^{(k)} - \lambda^*$ and $v_k = x^{(k)} - x^*$. Nonexpansiveness of projection implies

$$\begin{aligned} \|\lambda^{(k+1)} - \lambda^*\|^2 &= \|\mathbf{p}_{\mathbb{R}_+^\ell}(\lambda^{(k)} + \tau(Ax^{(k)} - b)) \\ &\quad - \mathbf{p}_{\mathbb{R}_+^\ell}(\lambda^* + \tau(Ax^* - b))\|^2 \\ &\leq \|u_k + \tau Av_k\|^2 \\ &= \|u_k\|^2 + 2\tau \langle u_k | Av_k \rangle + \tau^2 \|Av_k\|^2. \end{aligned}$$

Subtracting the two stationarity equations gives $A^T u_k = -Qv_k$. Consequently,

$$\langle u_k | Av_k \rangle = \langle A^T u_k | v_k \rangle = -\langle Qv_k | v_k \rangle \leq -m \|v_k\|^2.$$

It follows that

$$\|\lambda^{(k+1)} - \lambda^*\|^2 \leq \|\lambda^{(k)} - \lambda^*\|^2 - (2\tau m - \tau^2 \|A\|^2) \|x^{(k)} - x^*\|^2.$$

Define $r_k = \|\lambda^{(k)} - \lambda^*\|^2$ and $\delta = 2\tau m - \tau^2 \|A\|^2 > 0$. Then

$$0 \leq \delta \|x^{(k)} - x^*\|^2 \leq r_k - r_{k+1}.$$

The nonnegative sequence (r_k) is nonincreasing, hence convergent, so $r_k - r_{k+1} \rightarrow 0$. Therefore $x^{(k)} \rightarrow x^*$. This proof also covers $A = 0$, for which $\delta = 2\tau m > 0$ (and in fact every $x^{(k)} = -Q^{-1}e = x^*$). ■