Optimization. Applications in image processing

Mila NIKOLOVA

CMLA (CNRS-UMR 8536)–ENS de Cachan, 61 av. du Président Wilson, 94235 Cachan Cedex, France
nikolova@cmla.ens-cachan.fr
http://mnikolova.perso.math.cnrs.fr
Textbook and Slides: http://mnikolova.perso.math.cnrs.fr/Courses.html

December 2016

Contents

Objectives: obtain a good knowledge of the most important optimization methods, provide tools to help reading the literature, conceive and analyze new methods. At the implementation stage, numerical methods are always on $\mathbb{R}^n$.

Main References: [2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13]

Notations and abbreviations

- Abbreviations:
  - n.v.s.—normed vector space (V).
  - l.s.c.—lower semi-continuous (for a function).
  - w.r.t.—with respect to
  - resp.—respectively
- dim(.)—dimension.
- $\langle .\,,. \rangle$ inner product ($\equiv$ scalar product) on a n.v.s. V.
- $\lfloor a \rfloor$ denotes the integer part of $a \in \mathbb{R}$.
- $\mathcal{O}$ is an open subset (arbitrary).
- $\mathcal{O}(U)$ denotes an open subset of V containing $U \subset V$.
- $\mathrm{int}(U)$ stands for the interior of U.
- $B(u,r)$ is the open ball centered at u with radius r. The relevant closed ball is $\overline{B(u,r)}$.
- $\mathbb{N}$ is the set of non negative integers.
- $\mathbb{R}^q_+ = \{v \in \mathbb{R}^q : v[i] \geqslant 0, 1 \leqslant i \leqslant q\}$ for any positive integer q.
- $\overline{\mathbb{R}} = \mathbb{R} \cup \{+\infty\}$.
- $(e_i)_{i=1}^n$ is the canonical basis of $\mathbb{R}^n$.
- For an $n \times n$ matrix B:
  - B^T is its transpose and B^H its conjugate transpose.
  - $\lambda_i(B)$—an eigenvalue of B.
  - $\lambda_{\min}(B)$ (resp. $\lambda_{\max}(B)$)—the minimal (resp. the maximal) eigenvalue of B.
  - Remind: the spectral radius of B is $\stackrel{\text{def}}{=} \max_{1 \leqslant i \leqslant n} \big|\lambda_i(B)\big|$.
  - $B \succ 0$—B is positive definite, $B \succeq 0$—B is positive semi-definite.
- Id is the identity operator.
- $\mathrm{diag}(b[1], \cdots, b[n])$ is an $n \times n$ diagonal matrix whose diagonal entries are $b[i]$, $1 \leqslant i \leqslant n$.
- For $f : V_1 \times V_2 \times \cdots V_m \to Y$ where V_j, $j \in \{1, 2, \cdots, m\}$ and Y are n.v.s., $D_j f(u_1, \cdots, u_m)$ is the differential of f at $u = (u_1, \cdots, u_m)$ with respect to u_j.
- $\mathcal{L}(V;Y)$, for V and Y n.v.s., is the set of all linear continuous applications from V to Y.
- For $A \in \mathcal{L}(V;Y)$, its adjoint is A^*.
- $\mathrm{Isom}(V;Y) \subset \mathcal{L}(V;Y)$ are all bijective applications that have a continuous inverse application.
- o-function—satisfies $\lim_{t\to\infty} o(t)/t = 0$
- O-function—satisfies $\lim_{t\to\infty} O(t)/t = K$ where K is a constant.

Chapter 1

Generalities

1.1 What is mathematical optimization?

Optimization models the goal of solving a task in the "best way". Saying "best" implies a viewpoint and possible compromises.

Examples:

• Running a business: to maximize profit, minimize loss, maximize efficiency and/or minimize risk.

• Design: minimize the weight of a bridge/truss, and maximize the strength, within the design constraints

• Planning: select a flight route to minimize time and/or fuel consumption of an airplane, while respecting the paths of other airplanes and other safety conditions.

Optimization is an essential tool in life, business, applied sciences.

A concrete example:

sun hats c_1 euro sun hats with logo p_1 euro

umbrellas c_2 euro umbrellas with logo p_2 euro

available money for investment k euro

free storage room b m^3 storage room for rent r m^3 for d euro / m^3

one sun hat s_1 m^3 one umbrella s_2 m^3

---

actions: x_1 hats ordered x_2 umbrellas ordered x_3 space rented

range constraints: $x_1 \geqslant 0 \quad x_2 \geqslant 0 \quad 0 \leqslant x_3 \leqslant r$

storage constraint: $s_1x_1 + s_2x_2 \leqslant b + x_3$

investment constraint: $c_1x_1 + c_2x_2 + dx_3 \leqslant k$

profit expression: $(p_1 - c_1)x_1 + (p_2 - c_2)x_2 - dx_3$

Goal: maximize the profit expression.

Formal definition: to minimize (or maximize) a real objective function by deciding the values of free variables from within an allowed set.

- The objective function represents the whole range of (known) possible choices
- The objective function should allows comparison between the different choices.

Last few decades: astonishing improvements in computer hardware and software, which motivated great leap in optimization modeling, algorithm designs, and implementations.

1.2 Optimization problems

Many tasks are formulated as the minimization / maximization of an objective function (energy, criterion) whose solution is the sought after object (a signal, an image).

General Form : find a solution $\hat{u} \in V$ such that

$$(P) \qquad \hat{u} \in U \ \text{ and } \ F(\hat{u}) = \inf_{u \in U} F(u) = -\sup_{u \in U} \big(-F(u) \big)$$

- $F : V \to \mathbb{R}$ functional (objective function, criterion, energy) to minimize.
- V—real Hilbert space, if not explicitly specified.
- $U \subset V$ constraint set (feasible set), supposed nonempty and closed.

If $U = V$ then the problem is called unconstrained. It is easier to solve.

Otherwise, the problem is called constrained. Important constraints:

$$\text{equality constraints} \qquad U = \{u \in V : g_i(u) = 0,\ i = 1, \dots, p\}, \quad p < n \tag{1.1}$$

$$\text{inequality constraints} \qquad U = \{u \in V : h_i(u) \leqslant 0,\ i = 1, \dots, q\}, \quad q \in \mathbb{N} \tag{1.2}$$

Minimizer $\hat{u}$ and minimum $F(\hat{u})$.

Relative (local) minimum (resp. minimizer) and minimum global minimum (resp. minimizer).

A "stationary point" (where the derivative is zero) is also used as a solution, but it can be a local maximum, a local minimum, or a saddle point.

1.2.1 Standard Form for quadratic functional

$$F(u) = \frac{1}{2} B(u,u) - c(u), \quad \text{for} \quad B \in \mathcal{L}(V \times V\,; \mathbb{R}) \ \text{ and } \ c \in \mathcal{L}(V; \mathbb{R}), \tag{1.3}$$

where B is bilinear and symmetric (i.e. $B(u,v) = B(v,u),\ \forall u, v \in V$).

If V is equipped with an inner product $\langle ., . \rangle$, we can write down (Riesz's representation theorem)

$$F(u) = \frac{1}{2} \langle Bu, u \rangle - \langle c, u \rangle \tag{1.4}$$

If $V = \mathbb{R}^n$ in (1.4), $B \in \mathbb{R}^{n \times n}$ is a symmetric matrix (i.e. $B = B^T$) and $c \in \mathbb{R}^n$.

1.2.2 Ill-posed inverse problems

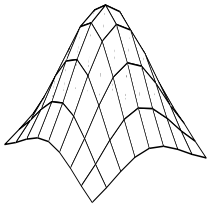
Out-of-focus picture: $\boldsymbol{v = a * u_o}$ + noise $= \boldsymbol{A u_o}$ + noise

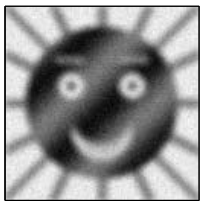
$\boldsymbol{A}$ is ill-conditioned $\equiv$ (nearly) noninvertible

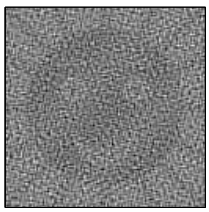
Least-squares solution: $\boldsymbol{\hat{u} = \arg\min_u \left\{ \|Au - v\|^2 \right\}}$

Tikhonov regularization: $\boldsymbol{\hat{u} \overset{\text{def}}{=} \arg\min_u \left\{ \|Au - v\|^2 + \beta \sum_i \|D_i u\|^2 \right\}}$ for $\boldsymbol{\{D_i\} \approx \nabla,\ \beta > 0}$


Original $\boldsymbol{u_o}$


Blur $\boldsymbol{a}$


Data $\boldsymbol{v}$


$\hat{\boldsymbol{u}}$: Least-squares


$\hat{\boldsymbol{u}}$: Tikhonov

1.2.3 Regularized objective functionals on $\mathbb{R}^n$

$$\text{Solution } \hat{u} \begin{cases} \nearrow & \text{close to data production model } \Psi(u,v) \quad \text{(data-fidelity)} \\ \searrow & \text{coherent with priors and desiderata } \Phi(u) \quad \text{(prior)} \end{cases}$$

Combining models: $\hat{u} \stackrel{\text{def}}{=} \arg\min_u \; F(u)$ where

$$F(u) = \Psi(u,v) + \beta\Phi(u), \quad \beta > 0 \tag{1.5}$$

$$\Phi(u) = \sum_i \varphi(\|\mathrm{D}_i u\|) \tag{1.6}$$

$\mathrm{D}_i u$—discrete approximation of the gradient or the Laplacian of the image or signal at i, or finite differences, $\|.\|$ is usually the ℓ_2 or the ℓ_1 norm[1], $\beta > 0$ parameter and $\varphi : \mathbb{R}_+ \to \mathbb{R}_+$ increasing function, e.g.

$$\varphi(t) \;\Big\|\; t^\alpha,\ 0 \leqslant \alpha \leqslant 2 \;\Big|\; \sqrt{t^2+\alpha} \;\Big|\; \frac{\alpha\, t}{\alpha\, t+1} \;\Big|\; \min\{t^2, \alpha\} \;\Big|\; \ln(\alpha t+1) \;\Big|\; (\cdots) \qquad \alpha > 0$$

Ψ data-fitting term for data $v \in \mathbb{R}^m$, usually $A \in \mathbb{R}^{m\times n}$ and

$$\Psi(u,v) = \|Au - v\|_p^p, \quad p \in \{1,2\} \tag{1.7}$$

or another function (the discrepancy).

One can have

$$\mathrm{D}_i = e_i^T, \quad \forall i \in \{1, \cdots, n\},$$

then $\Phi(u) = \sum_i \varphi(\,|u[i]|\,)$.

Differences between neighboring pixels play an important role in signal processing and in imaging. If u is an one-dimensional signal, first-order differences yield

$$\|\mathrm{D}_i u\| = \Big|u[i] - u[i-1]\Big|, \quad i = 2, \cdots, n.$$

Let u be an image of size $M \times N$ — columnwize rearranged into a n-length vector, $n = MN$:

$$u = [u_{1,1}, \ldots, u_{M,1}, u_{1,2}, \ldots, u_{M,2}, \ldots, u_{1,N}, \ldots, u_{M,N}]^T \simeq [u_1, \ldots, u_p]^T$$

Often

$$\|\mathrm{D}_i u\| = \Big((u[i] - u[i-m])^2 + (u[i] - u[i-1])^2\Big)^{1/2} \tag{1.8}$$

[1]For $p \in [1, +\infty[$ the ℓ_p norm of a vector v is $\|v\|_p = \left(\sum_i v[i]^p\right)^{\frac{1}{p}}$ while $\|v\|_\infty = \max_i \big|v[i]\big|$.

or

$$\|\mathrm{D}_k u\| = \Big|u[i] - u[i-m]\Big| \quad \text{and} \quad \|\mathrm{D}_{k+1} u\| = \Big|u[i] - u[i-1]\Big|. \tag{1.9}$$

D_i can also be any other linear mapping.

Table of Contents

1.3 Optimization algorithms

Usually the solution $\hat{u}$ is defined implicitly and cannot be computed in an explicit way, in one step.

1.3.1 Iterative minimization methods

Construct a sequence $(u_k)_{k\in\mathbb{N}}$ initialized by u_0 converging to $\hat{u}$—a solution of (P):

$$u_{k+1} = G(u_k), \quad k = 0, 1, \ldots, \tag{1.10}$$

where $G : V \to U$ is iterative scheme (often defined implicitly). The solution $\hat{u}$ can be local (relative) if (P) is nonconvex. The choice of u_0 (= the initialization) can be crucial if (P) is nonconvex.

G is constructed using information on (P), e.g. $F(u_k)$, $\nabla F(u_k)$, $g_i(u_k)$, $h_i(u_k)$, $\nabla g_i(u_k)$ and $\nabla h_i(u_k)$ (or subgradients instead of ∇ for nonsmooth functions). By (1.10), the iterates of G are

$$\begin{aligned} u_1 &= G(u_0); \\ u_2 &= G(u_1) = G \circ G(u_0) \stackrel{\text{def}}{=} G_2(u_0); \\ &\cdots \\ u_k &= \underbrace{G \circ \cdots \circ G}_{k \text{ times}}\ (u_0) \stackrel{\text{def}}{=} G_k(u_0) = G \circ G_{k-1}(u_0) \end{aligned}$$

Key questions:

- Given an iterative method G, determine if there is convergence;
- If convergence, to what kind of accumulation point?
- Given two iterative methods G_1 and G_2 choose the faster one.

Definition 1 *$\hat{u}$ is a fixed point for G if $G(\hat{u}) = \hat{u}$.*

Let X be a metric space equipped with distance d.

Definition 2 *$G : X \to X$ is a contraction if there exists $\gamma \in (0, 1)$ such that*

$$d(G(u_1), G(u_2)) \leqslant \gamma d(u_1, u_2), \quad \forall u_1, u_2 \in X$$

$\Rightarrow G$ is Lipschitzian[2] $\Rightarrow$ uniformly continuous.

Theorem 1 (Fixed point theorem) *Let X be complete and $G : X \to X$ a contraction. Then G admits a unique fixed point, $\hat{u} = G(\hat{u})$. (see [14, p.141]).*

Theorem 2 (Fixed point theorem-bis) *Let X be complete and $G : X \to X$ be such that $\exists k_0$ for which G_{k_0} is a contraction. Then G admits a unique fixed point. (see [14, p.142]).*

Note that in the latter case G is not necessarily a contraction.

[2] A function $f : V \to X$ is ℓ-Lipschitz continuous if $\forall (u, v) \in V \times V$, we have $\|f(u) - f(v)\| \leqslant \ell \|u - v\|$.

1.3.2 Local convergence rate

In general, a nonlinear problem (P) cannot be solved exactly in a finite number of iterations.
Goal: attach to G precise indicators of the asymptotic rate of convergence of u_k towards $\hat{u}$. Refs. [11, 10, 5, 4]

Here $V = \mathbb{R}^n$, $\|\cdot\| = \|\cdot\|_2$ and we simply assume that u_k converges to $\hat{u}$.

Q-convergence studies the quotient $Q_k \stackrel{\text{def}}{=} \dfrac{\|u_{k+1} - \hat{u}\|}{\|u_k - \hat{u}\|}$, $k \in \mathbb{N}$.

$$Q = \limsup_{k\to\infty} Q_k$$

- If $Q < \infty$ then $\forall \varepsilon > 0$ $\exists k_0$ such that

$$\|u_{k+1} - \hat{u}\| \leqslant (Q + \varepsilon)\|u_k - \hat{u}\|, \quad \forall k \geqslant k_0.$$

  $\equiv$ Bound on the error at iteration $k+1$ in terms of the error at iteration k. (Crucial if $Q < 1$.)

- $0 < Q < 1$: Q-linear convergence;
- $Q = 0$: Q-superlinear convergence (called also superlinear convergence);
- In particular, $Q_k = O(\|u_k - \hat{u}\|_2)$: Q-quadratic convergence.
- Compare how G_1 and G_2 converge towards the same $\hat{u}$:
  If $Q(G_1, \hat{u}) < Q(G_2, \hat{u})$ then G_1 is faster than G_2 in the sense of Q.

R-convergence studies the rate of the root $R_k \stackrel{\text{def}}{=} \|u_k - \hat{u}\|^{1/k}$, $k \in \mathbb{N}$.

$$R = \limsup_{k\to\infty} R_k$$

- $0 < R < 1$: R-linear convergence; this means geometric or exponential convergence since

$$\forall \varepsilon \in (0, 1 - R) \ \ \exists k_0 \text{ such that } R_k \leqslant (R + \varepsilon), \ \forall k \geqslant k_0 \iff \|u_k - \hat{u}\| \leqslant (R + \varepsilon)^k$$

- $R = 0$: R-superlinear convergence.
- G_1 is faster than G_2 in the sense of R if $R(G_1, \hat{u}) < R(G_2, \hat{u})$

Remark 1 Sublinear convergence if $Q_k \to 1$ or $R_k \to 1$. Convergence is too slow, choose another G.

Lemma 1 *Let $G : U \subset \mathbb{R}^n \to \mathbb{R}^n$, $\|.\|$ be any norm on $\mathbb{R}^n$, $\exists B(\hat{u}, \delta) \subset U$ and $\exists \gamma \in [0, 1)$ such that*

$$\|G(u) - \hat{u}\| \leqslant \gamma \|u - \hat{u}\|, \quad \forall u \in B(\hat{u}, \delta)$$

Then $\forall u_0 \in B(\hat{u}, \delta)$ the iterates given by (1.10) remain in $B(\hat{u}, \delta)$ and converge to $\hat{u}$.

Proof. Let $u_0 \in B(\hat{u}, \delta)$, then $\|u_1 - \hat{u}\| = \|G(u_0) - \hat{u}\| \leqslant \gamma\|u_0 - \hat{u}\| < \|u_0 - \hat{u}\|$, hence $u_1 \in B(\hat{u}, \delta)$. Using induction, $u_k \in B(\hat{u}, \delta)$ and $\|u_k - \hat{u}\| \leqslant \gamma^k \|u_0 - \hat{u}\|$. Thus $\lim_{k\to\infty} u_k = \hat{u}$. □

Theorem 3 (Ostrowski) *[11, p.300] Let $G : U \subset \mathbb{R}^n \to \mathbb{R}^n$ be differentiable at $\hat{u} \in \mathrm{int}(U)$ and has a fixed point $G(\hat{u}) = \hat{u} \in U$. If*

$$\sigma \stackrel{\mathrm{def}}{=} \max_{1 \leqslant i \leqslant n} \big|\lambda_i\big(\nabla G(\hat{u})\big)\big| < 1, \tag{1.11}$$

then $\exists \mathcal{O}(\hat{u})$ such that $\forall u_0 \in \mathcal{O}(\hat{u})$ we have $u_k \in \mathcal{O}(\hat{u})$ and $u_k \to \hat{u}$.

Proof. $\nabla G(\hat{u})$ is not necessarily symmetric and positive definite. Condition (1.11) says that $\forall \varepsilon > 0$ there is an induced matrix norm $\|.\|$ on $\mathbb{R}^{n \times n}$ such that[3]

$$\|\nabla G(\hat{u})\| \leqslant \sigma + \varepsilon. \tag{1.12}$$

Let us choose ε such that

$$\varepsilon < \frac{1}{2}(1 - \sigma).$$

The differentiability of G at $\hat{u}$ implies that $\exists \delta > 0$ so that $B(\hat{u}, \delta) \subset U$ and

$$\|G(u) - G(\hat{u}) - \nabla G(\hat{u})(u - \hat{u})\| \leqslant \varepsilon \|u - \hat{u}\|, \quad \forall u \in B(\hat{u}, \delta) \tag{1.13}$$

Using that $G(\hat{u}) = \hat{u}$, and combining (1.12) and (1.13), we get

$$\begin{aligned} \|G(u) - \hat{u}\| &= \|G(u) - G(\hat{u}) - \nabla G(\hat{u})(u - \hat{u}) + \nabla G(\hat{u})(u - \hat{u})\| \\ &\leqslant \|G(u) - G(\hat{u}) - \nabla G(\hat{u})(u - \hat{u})\| + \|\nabla G(\hat{u})\| \, \|u - \hat{u}\| \\ &< (2\varepsilon + \sigma)\|u - \hat{u}\| < \|u - \hat{u}\|, \quad \forall u \in B(\hat{u}, \delta) \end{aligned}$$

The conclusion follows from the observation that $2\varepsilon + \sigma < 1$ and Lemma 1. □

Theorem 5 (linear convergence) *Under the conditions of Theorem 3, $\max_{1 \leqslant i \leqslant n} |\lambda_i(\nabla G(\hat{u}))| = R$, where R is the root convergence factor.*

Proof—see [11, p.301].

Illustration of the role of convergence conditions

[3]Let B be an $n \times n$ real matrix. Its spectral radius is defined as its largest in absolute value eigenvalue,

$$\sigma(B) \stackrel{\mathrm{def}}{=} \max_{1 \leqslant i \leqslant n} \big|\lambda_i(B)\big|.$$

Given a vector norm on $\mathbb{R}^n$, say $\|.\|$, the corresponding induced matrix norm on the space of all $n \times n$ matrices reads

$$\|B\| \stackrel{\mathrm{def}}{=} \sup\{\|Bu\| : u \in \mathbb{R}^n, \|u\| \leqslant 1\}.$$

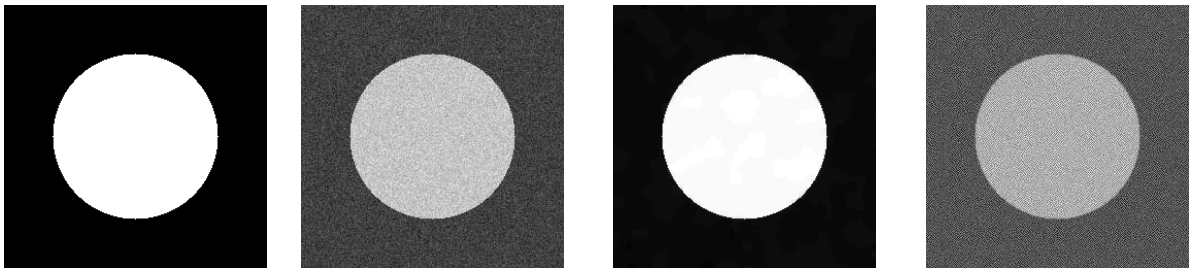
Theorem 4 *[5, p. 18]*

(1) Let B be an $n \times n$ matrix and $\| \, . \, \|$ any matrix norm. Then

$$\sigma(B) \leqslant \| \, B \, \| \, .$$

(2) Given a matrix B and a number $\varepsilon > 0$, there exists at least one induced matrix norm $\| \, . \, \|$ such that

$$\| \, B \, \| \leqslant \sigma(B) + \varepsilon.$$



(a) Original (b) Observed noisy (c) Correctly restored (d) No convergence

Figure 1.1: Convergence conditions satisfied in (c), not satisfied in (d).

1.4 Analysis of optimization problems

References : [15, 3, 16, 5, 17, 9, 18].

1.4.1 Remainders

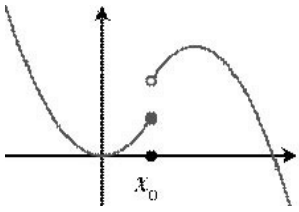
Definition 3 *Let $F : V \to \ ]-\infty, +\infty]$ where V is a real topological space.*

- *The* domain *of F is the set* $\mathrm{dom}F = \{u \in V \ \mid \ F(u) < +\infty\}$.
- *The* epigraph *of F is the set* $\mathrm{epi}F = \{(u, \lambda) \in V \times \mathbb{R} \ \mid \ F(u) \leqslant \lambda\}$.

Definition 4 *A function F on a real n.v.s. V is* proper *if $F : V \to \ ]-\infty, +\infty]$ and if it is not identically equal to $+\infty$.*

Definition 5 *$F : V \to \mathbb{R}$ is coercive if* $\lim_{\|u\| \to \infty} F(u) = +\infty$.

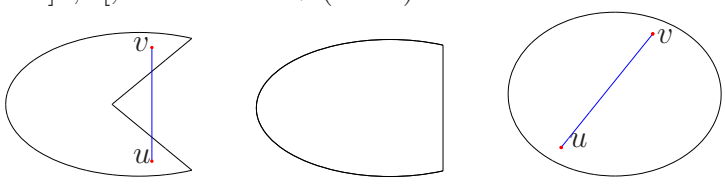
Definition 6 *$F : V \to \ ]-\infty, +\infty]$ for V a real topological space is* lower semi-continuous (l.s.c.) *if*
$\forall \lambda \in \mathbb{R}$ the set $\left\{u \in V \ \mid \ F(u) \leqslant \lambda\right\}$ is is closed in V.



F is l.s.c.

If F is l.s.c., then $-F$ is upper semi-continuous.
If F is continuous, then it is l.s.c. and upper semi-continuous.

Definition 7 (Convex subset) *Let V be any real vector space. A subset $U \subset V$ is convex if $\forall u, v \in U$ and $\theta \in]0, 1[$, we have $\theta u + (1-\theta)v \in U$.*



U nonconvex *U non strictly convex* *U strictly convex*

Remind that F can be convex but not coercive.

Definition 8 (Convex function) *Let V be any real vector space. A proper function $F : U \subset V \to \mathbb{R}$ is convex if $\forall u, v \in U$ and $\theta \in]0, 1[$*

$$F(\theta u + (1-\theta)v) \leqslant \theta F(u) + (1-\theta)F(v)$$

F is strictly convex when the inequality is strict whenever $u \neq v$.

Property 1 *Important properties [16, p. 8]:*

- *F is l.s.c. if and only if $\forall u \in V$ and $\forall \varepsilon > 0$ there is a neighborhood $\mathcal{O}$ of u such that*

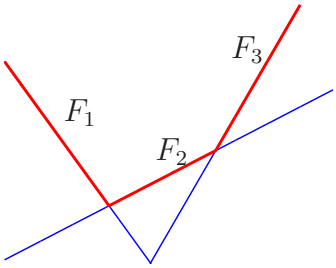
$$F(v) \geqslant F(u) - \varepsilon, \quad \forall v \in \mathcal{O} \ .$$

- *If F is l.s.c.[4] and $u_k \to u$ as $k \to \infty$ then $\liminf_{k\to\infty} F(u_k) \geqslant F(u)$.*

- *If $(F_i)_{i\in I}$, where I is an index set, is a family of l.s.c. functions then the superior envelop of $(F_i)_{i\in I}$ is l.s.c. In words, the function F defined by*

$$F(u) = \sup_{i\in I} F_i(u)$$

is l.s.c.



- *If $(F_i)_{i\in I}$ is a family of l.s.c. convex functions then the superior envelop F of $(F_i)_{i\in I}$ is l.s.c. and convex.*

Definition 9 (Principle of minimization) Consider the problem: minimise $F(u)$ subject to $u \in U$.

- Set of feasible solutions $= \mathrm{dom} F \cap U$;
- Optimal value (it is unique): $\theta = \inf\{F(u) \ : \ u \in U\}$;
- Set of (global) minimizers: $\{u \in U \ : \ F(u) = \theta\}$;
- Local minimizer $\hat{u}$ – there is a neighborhood $\mathcal{O} \cap U$ such that $F(\hat{u}) \leqslant F(u)\ \forall u \in \mathcal{O} \cap U$;
- Strict local minimizer $\hat{u}$ – there is $\mathcal{O} \cap U$ such that $F(\hat{u}) < F(u)\ \forall u \in \mathcal{O} \cap U$ with $u \neq \hat{u}$;
- Isolated local minimizer $\hat{u}$ – there is $\mathcal{O} \cap U$ such that $\hat{u}$ is the only local minimizer in $\mathcal{O} \cap U$.

All isolated local minimizers are strict.

1.4.2 Existence, uniqueness of the solution

Theorem 6 (Existence) *Let $U \subset \mathbb{R}^n$ be non-empty and closed, $F : \mathbb{R}^n \to \mathbb{R}$ l.s.c. and proper. If U is not bounded, we suppose that F is coercive. Then $\exists \hat{u} \in U$ such that $F(\hat{u}) = \inf_{u\in U} F(u)$.*

Note that F can be non-convex; moreover $\hat{u}$ is not necessarily unique.

Proof. Two parts:

(a) U bounded $\Rightarrow$ U is compact, since F is l.s.c., Weierstrass theorem[5] yields the result.

(b) U not necessarily bounded. Choose $u_0 \in U$.

F coercive $\Rightarrow \exists r > 0$ such that

$$\|u\| > r \quad \Rightarrow \quad F(u_0) < F(u)$$

[4] If F is upper semi-continuous and $u_k \to u$ as $k \to \infty$ then $\limsup_{k\to\infty} F(u_k) \leqslant F(u)$.

[5] Weierstrass theorem: Let V be a n.v.s. and $K \subset V$ a compact. If $F : V \to \mathbb{R}$ is l.s.c. on K, then F achieves a minimum on K. If $F : V \to \mathbb{R}$ is upper semi continuous on K, then F achieves a maximum on K. If F is continuous on K, then F achieves a minimum and a maximum on K. See e.g. [9, p. 40]

Then $\hat{u} \in \widetilde{U} \stackrel{\text{def}}{=} \overline{B(u_0, r)} \cap U$ which is compact. The conclusion is obtained as in (a). □

Let us underline that the conditions in the theorem are only strong sufficient conditions. Much weaker existence conditions can be found e.g. in [15, p. 96].

- Alternative proof using minimizing sequences.
- The theorem extends to separable Hilbert spaces under additional conditions, see e.g. [5, 19].

Optimization problems are often called "feasible" when U and F are convex and the conditions of Theorem 6 are met. Remind that F can be convex but not coercive (Fig. 1.2(b)).

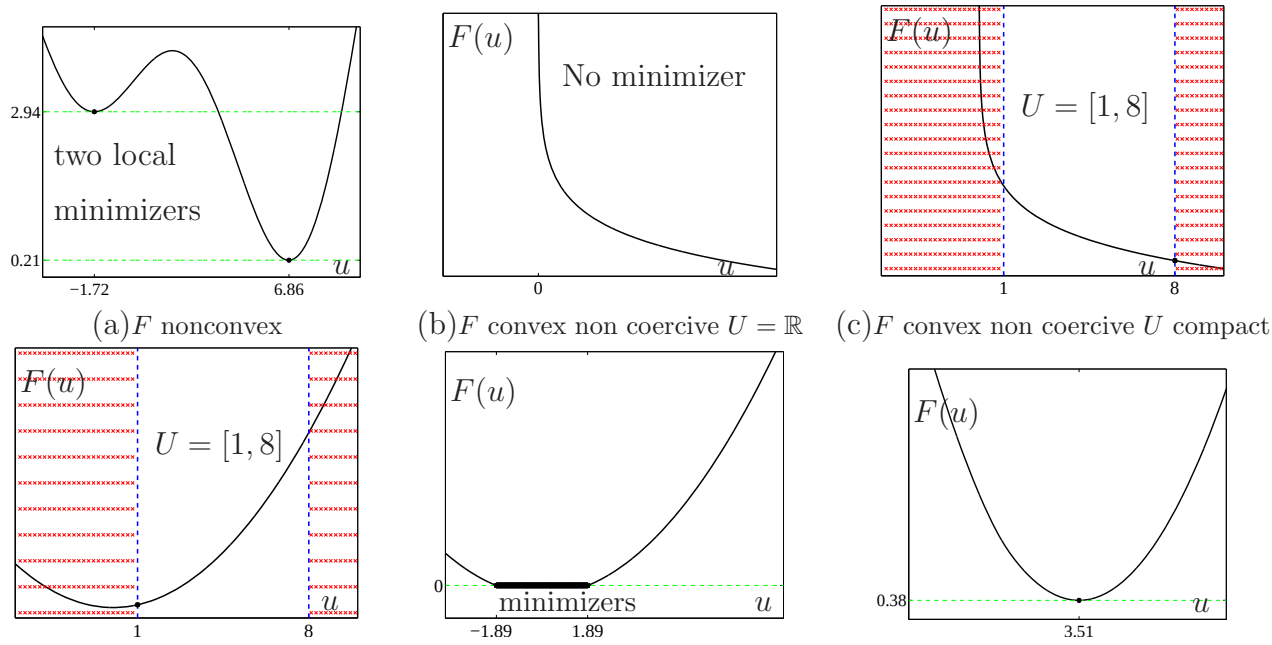


Figure 1.2: Illustration of Definition 8 and Theorem 7. Minimizers are depicted with a thick point.

Theorem 7 *For $U \subset V$ convex, let $F : U \to \mathbb{R}$ be proper, convex and l.s.c.*

1. *If F has a relative (local) minimum at $\hat{u} \in U$, it is a (global) minimum w.r.t. U. [5]*
2. *The set of minimizers $\widehat{U} = \left\{ \hat{u} \in U : F(\hat{u}) = \inf_{u \in U} F(u) \right\}$ is convex and closed. [17, p. 35]*
3. *If F is strictly convex, then F admits at most one minimum and the latter is strict.*
4. *In addition, suppose that either F is coercive or U is bounded. Then $\widehat{U} \neq \varnothing$. [17, p. 35]*

1.4.3 Convex conjugates, Separation theorem and Equivalent optimization problems

Given an optimization problem defined by $F : V \to \mathbb{R}$ and $U \subset V$,

$$(P) \qquad \text{find} \quad \hat{u} \quad \text{such that} \quad F(\hat{u}) = \inf_{u \in U} F(u) \qquad \Leftrightarrow \qquad \text{find} \quad \hat{u} = \arg\inf_{u \in U} F(u)$$

there are many different optimization problems that are in some way equivalent to (P). For instance

- $\hat{w} = \arg\min_{X} \mathcal{F}(w)$ where $\mathcal{F} : W \to \mathbb{R}$, $X \subset W$ and there is $f : W \to V$ such that $\hat{u} = f(\hat{w})$;
- $(\hat{u}, \hat{b}) = \arg\min_{u \in \mathcal{U}} \max_{b \in X} \mathcal{F}(u, b)$ where $\mathcal{F} : V \times W \to \mathbb{R}$, $\mathcal{U} \subset V$, $X \subset W$ and $\hat{u}$ solves (P);

Some of these equivalent optimization problems are easier to solve than the original (P). We shall see such reformulations in what follows. The way enabling to recover them is in general an open question. In many cases, such a reformulation (found by intuition and proven mathematically) is valid only for a particular problem (P).

There are several duality principles in optimization theory relating a problem expressed in terms of vectors in an n.v.s. V to a problem expressed in terms of hyperplanes in the n.v.s.

Definition 10 *A hyperplane (affine) is a set of the form*

$$[h = \alpha] \stackrel{\text{def}}{=} \{w \in V \ \mid \ \langle h, w \rangle = \alpha\}$$

where $h : V \to \mathbb{R}$ is a linear nonzero functional and $\alpha \in \mathbb{R}$ is a constant.

A milestone is the Hahn-Banach theorem, stated below in its geometric form [16, 9].

Definition 11 *Let $U \subset V$ and $K \subset V$. The hyperplane $[h = \alpha]$ separates K and U*

- nonstrictly *if $\langle h, w \rangle \leqslant \alpha$, $\forall w \in U$ and $\langle h, w \rangle \geqslant \alpha$, $\forall w \in K$;*
- strictly *if there exists $\varepsilon > 0$ such that $\langle h, w \rangle - \varepsilon \leqslant \alpha$, $\forall w \in U$ and $\langle h, w \rangle + \varepsilon \geqslant \alpha$, $\forall w \in K$.*

Theorem 8 (Hahn-Banach theorem, geometric form) *Let $U \subset V$ and $K \subset V$ be convex, nonempty and disjoint sets.*

(i) If U is open, there is a closed hyperplane[6] $[h = \alpha]$ that separates K and U nonstrictly;

(ii) If U is closed and K is compact[7], then there exists a closed hyperplane $[h = \alpha]$ that separates K and U in a strict way.

Definition 10 and Theorem 8 are illustrated in Fig. 1.3 where $K = \{u\}$. Note that U and K in Definition 11 are not needed to be convex.

Remark 2 *An example of equivalent optimization problems is seen in Fig. 1.3:*
the shortest distance from a point u to a convex closed set U (i.e., the orthogonal projection of u on U) is equal to the maximum of the distances from the point u to a hyperplane separating the point u from the set U.

One systematic approach to derive equivalent optimization problems is centered about the interrelation between an n.v.s. V and its dual.

[6]The hyperplane $[h = \alpha]$ is closed iff h is continuous [16, p. 4], [9, p. 130]. If $V = \mathbb{R}^n$ any hyperplane is closed.
[7]Let us remind that if K is a subset of a finite dimensional space, it is compact iff K is closed and bounded.

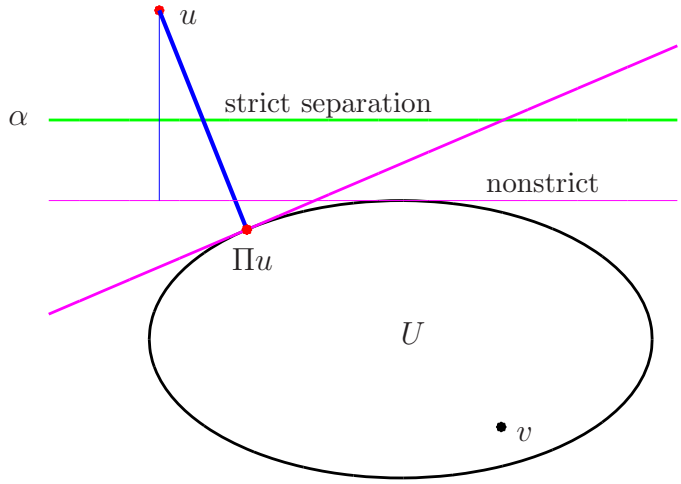


Figure 1.3: Green line: $[h = \alpha] = \{w \mid \langle h, w\rangle = \alpha\}$ separates strictly $\{u\}$ and U: for $v \in U$ we have $\langle h, v\rangle < \alpha$ whereas $\langle h, u\rangle > \alpha$. The orthogonal projection of u onto U is Πu (red dots). Lines in magenta provide nonstrict separation between u and U.

Definition 12 *The dual $V^\star$ of a n.v.s. V is composed of all continuous linear functionals on V, i.e.,*

$$V^\star \stackrel{\text{def}}{=} \{f : V \to \mathbb{R} \mid f \text{ linear and continuous}\}$$

$V^\star$ is endowed with the norm

$$\|f\|_{V^\star} = \sup_{u\in V,\ \|u\|\leqslant 1} |f(u)|$$

The n.v.s. V is reflexive *if $V^{\star\star} = V$ where $V^{\star\star}$ is the dual of $V^\star$.*

If $V = \mathbb{R}^n$ then $V^\star = \mathbb{R}^n$ (see e.g. [9, p. 107]). The dual of the Hilbert space[8] L_2 (resp. ℓ_2) is yet again L_2 (resp. ℓ_2)—see e.g. [9, pp. 107-109]. Clearly, all these spaces are reflexive as well.

Definition 13 *Let F defined on a real n.v.s. V be proper. The function $F^\star : V^\star \to \overline{\mathbb{R}}$ given by*

$$\boxed{F^\star(v) \stackrel{\text{def}}{=} \sup_{u\in V} \big\{\, \langle u, v\rangle - F(u)\big\}} \tag{1.14}$$

is called the convex conjugate *or the* polar *function of F.*

The Fenchel-Young inequality:

$$u \in V, v \in V^\star \quad \Rightarrow \quad \langle u, v\rangle \leqslant F(u) + F^\star(v) \tag{1.15}$$

The application $v \mapsto \langle u, v\rangle - F(u)$ is convex and continuous, hence l.s.c., for any fixed $u \in V$. According to Property 1, $F^\star$ is convex and l.s.c. (as being a superior convex envelop).

We can repeat the process in (1.14), thereby leading to the bi-conjugate $F^{\star\star} : V \to \overline{\mathbb{R}}$:

$$F^{\star\star}(u) \stackrel{\text{def}}{=} \sup_{v\in V^\star} \big\{\, \langle u, v\rangle - F^\star(v)\big\}$$

One can note that $F^{\star\star}(u) \leqslant F(u)$, $\forall u \in V$.

[8] Remind that $\mathbb{R}^n$ is a Hilbert space as well.

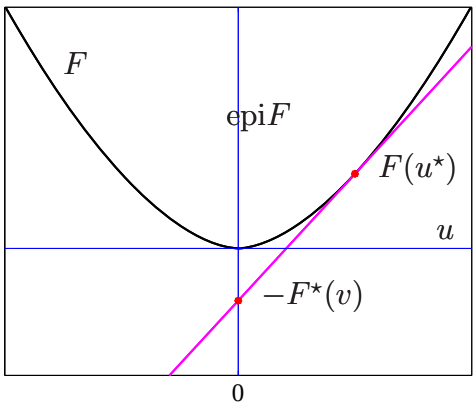


Figure 1.4: The convex conjugate at v determines an hyperplane that separates nonstrictly epi$F$ and its complement in $\mathbb{R}^n$.

Theorem 9 (Fenchel-Moreau) *Let F, defined on a real n.v.s. V, be proper and* convex. *Then*

$$F^{\star\star} = F$$

For the proof, see [16, p. 10] or [18, 17].

The theorem below is an important tool to equivalently reformulate optimization problems.

Theorem 10 (Fenchel-Rockafellar) *Let Φ and Ψ be convex on V. Assume that there exists u_0 such that $\Phi(u_0) < +\infty$, $\Psi(u_0) < +\infty$ and Φ is* continuous *at u_0. Then*

$$\inf_{u\in V}\big(\Phi(u)+\Psi(u)\big) = \max_{v\in V^\star}(-\Phi^\star(-v)-\Psi^\star(v)\big)$$

The proof can be found in [16, p. 11] or in [9, p. 201] (where the assumptions on Φ and Ψ are slightly different).

Example 1 *V is a real reflexive n.v.s. Let $U \subset V$ and $K \subset V^\star$ be convex compact nonempty subsets.*

$$\Psi(u) \stackrel{\text{def}}{=} \sup_{v\in K}\langle v,u\rangle = \max_{v\in K}\langle v,u\rangle \leqslant \max_{v\in K}\|v\|\,\|u\| \tag{1.16}$$

where Schwarz inequality was used. Note that we can replace sup *with* max *in* (1.16) *because K is compact and $v \mapsto \langle v,u\rangle$ is continuous (Weierstrass theorem). Clearly $\max_{v\in K}\|v\|$ is finite and Ψ is continuous and convex.*

- *Let $w \in V^\star \setminus K$. The set $\{w\}$ being compact, Hahn-Banach theorem 8 tells us that there exists $h \in V^{\star\star} = V$ enabling us to separate strictly $\{w\}$ and $K \subset V^\star$, that is*

$$\langle v,h\rangle < \langle w,h\rangle\,, \quad \forall v \in K$$

and the constant $c \stackrel{\text{def}}{=} \inf_{v\in K}\big(\langle w,h\rangle - \langle v,h\rangle\big)$ meets $c > 0$. Then

$$\langle w,h\rangle - \langle v,h\rangle \geqslant c > 0, \quad \forall\, v \in K \quad \Leftrightarrow \quad \langle w,h\rangle - \sup_{v\in K}\langle v,h\rangle \geqslant c > 0$$

Using that $\alpha h \in V$ for any $\alpha > 0$, we have

$$\begin{aligned}\Psi^\star(w) = \sup_{u\in V}\big(\langle w,u\rangle - \Psi(u)\big) &\geqslant \sup_{\alpha>0}\big(\langle w,\alpha h\rangle - \Psi(\alpha h)\big) = \sup_{\alpha>0}\big(\alpha\langle w,h\rangle - \sup_{v\in K}\langle v,\alpha h\rangle\big)\\ &= \sup_{\alpha>0}\alpha\big(\langle w,h\rangle - \sup_{v\in K}\langle v,h\rangle\big) \geqslant c\sup_{\alpha>0}\alpha = +\infty. \qquad (1.17)\end{aligned}$$

- *Let now $w \in K \subset V^\star$. Then $\langle w,u\rangle - \max_{v\in K}\langle v,u\rangle \leqslant 0$, $\forall u \in V$, hence*

$$\Psi^\star(w) = \sup_{u\in V}\big(\langle w,u\rangle - \Psi(u)\big) = \sup_{u\in V}\big(\langle w,u\rangle - \max_{v\in K}\langle v,u\rangle\big) = 0 \qquad (1.18)$$

where the upper bound 0 is always reached for $u = 0$.

- *Combining (1.17) and (1.18) shows that the convex conjugate of Ψ reads*

$$\Psi^\star(v) = \begin{cases} +\infty & \text{if} \quad v \notin K \\ 0 & \text{if} \quad v \in K \end{cases}$$

- *Set*

$$\Phi(u) = \begin{cases} +\infty & \text{if} \quad u \notin U \\ 0 & \text{if} \quad u \in U \end{cases}$$

By the definition of Ψ in (1.16) and the one of Φ, the Fenchel-Rockafellar Theorem 10 yields

$$\min_{u\in U}\big(\max_{v\in K}\langle u,v\rangle\big) = \min_{u\in U}\Psi(u) = \min_{u\in V}\big(\Phi(u)+\Psi(u)\big) = \max_{v\in V^\star}\big(-\Phi^\star(-v) - \Psi^\star(v)\big) \qquad (1.19)$$

The maximum in the right side cannot be reached if $v \notin K$ because in this case $-\Psi^\star(v) = -\infty$. Hence $\max_{v\in V^\star}\big(-\Phi^\star(-v) - \Psi^\star(v)\big) = \max_{v\in K}\big(-\Phi^\star(-v)\big)$. Furthermore

$$-\Phi^\star(-v) = -\sup_{u\in V}\big(-\langle v,u\rangle - \Phi(u)\big) = \inf_{u\in V}\big(\langle v,u\rangle + \Phi(u)\big) = \inf_{u\in U}\langle v,u\rangle = \min_{u\in U}\langle v,u\rangle$$

It follows that

$$\max_{v\in V^\star}\big(-\Psi^\star(v) - \Phi^\star(-v)\big) = \max_{v\in K}\big(\min_{u\in U}\langle v,u\rangle\big) \qquad (1.20)$$

Combining (1.19) and (1.20) shows that $\min_{u\in U}\big(\max_{v\in K}\langle u,v\rangle\big) = \max_{v\in K}\big(\min_{u\in U}\langle v,u\rangle\big)$.

In Example 1 we have proven the fundamental Min-Max theorem used often in classical game theory. The precise statement is given below.

Theorem 11 (Min-Max) *Let V be a reflexive real n.v.s. Let $U \subset V$ and $K \subset V^\star$ be convex, compact nonempty subsets. Then*

$$\min_{u\in U}\max_{v\in K}\langle u,v\rangle = \max_{v\in K}\min_{u\in U}\langle v,u\rangle$$

Chapter 2

Unconstrained differentiable problems

2.1 Preliminaries (regularity of F)

Remainders (see e.g. [20, 21, 5]):

1. For V and Y real n.v.s., $f : V \to Y$ if differentiable at $u \in V$ if $\exists Df(u) \in \mathcal{L}(V, Y)$ (linear continuous application from V to Y) such that

$$f(u+v) = f(u) + Df(u)v + \|v\|\varepsilon(v) \quad \text{where} \quad \lim_{v\to 0} \varepsilon(v) = 0.$$

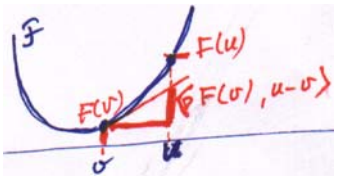
2. If $Y = \mathbb{R}$ (i.e. $f : V \to \mathbb{R}$) and the norm on V is derived from an inner product $\langle ., . \rangle$, then $\mathcal{L}(V; \mathbb{R})$ is identified to V (via a canonical isomorphism). Then $\nabla f(u) \in V$—the gradient of f at u is defined by

$$\langle \nabla f(u), v \rangle = Df(u)v, \ \forall v \in V.$$

   Note that if $V = \mathbb{R}^n$ we have $Df(u) = (\nabla f(u))^T$.

3. Under the conditions given above in 2., if f is twice differentiable, we can identify the second differential $D^2 f$ and the Hessian $\nabla^2 f$ via $D^2 f(u)(v, w) = \langle \nabla^2 f(u)v, w \rangle$.

Property 2 *Let $U \subset V$ be convex. Suppose $F : V \to \mathbb{R}$ is differentiable in $\mathcal{O}(U)$*



1. *F convex on U $\Leftrightarrow$ $F(v) \geqslant F(u) + \langle \nabla F(u),\ v - u \rangle$, $\forall u, v \in U$.*

2. *F strictly convex on U $\Leftrightarrow$ $F(v) > F(u) + \langle \nabla F(u),\ v - u \rangle$, $\forall u, v \in U,\ v \neq u$.*

A proof of statement 1 can be found in Appendix, p. 111.

Property 3 *Let $U \subset V$ be convex. Suppose F twice differentiable in $\mathcal{O}(U)$.*

1. *F convex on U $\Leftrightarrow$ $\langle \nabla^2 F(u)(v-u), (v-u) \rangle \geqslant 0$, $\forall u, v \in U$.*

2. *If $\langle \nabla^2 F(u)(v-u), (v-u) \rangle > 0$, $\forall u, v \in U,\ v \neq u$ then F is strictly convex on U.*

Definition 14 *F is called* strongly convex *or equivalently* elliptic *if $F \sim \mathcal{C}^1$ and if $\exists\ \mu > 0$ such that*

$$\langle \nabla F(v) - \nabla F(u),\ v - u \rangle \geqslant \mu \|v - u\|^2, \ \forall v, u \in V.$$

F in (1.4) is strongly convex if and only if $B \succ 0$.

Property 4 *Suppose $F : V \to \mathbb{R}$ is differentiable in V.*

1. *$F : V \to \mathbb{R}$ strongly convex $\Rightarrow$ F is strictly convex, coercive and*
   $F(v) - F(u) \geqslant \langle \nabla F(u), v - u \rangle + \frac{\mu}{2}\|v - u\|^2, \quad \forall v, u \in V.$
2. *Suppose F twice differentiable in V:*
   F strongly convex $\Leftrightarrow$ $\langle \nabla^2 F(u)v, v \rangle \geqslant \mu \|v\|^2, \quad \forall v \in V.$

Remark 3 *If $F : \mathbb{R}^n \to \mathbb{R}$ is twice differentiable, then $\nabla^2 F(u)$ (known as the* Hessian *of F at u) is an $n \times n$* symmetric matrix*, $\forall u \in \mathbb{R}^n$. Indeed, $\big(\nabla^2 F(u)\big)[i, j] = \frac{\partial^2 F(u)}{\partial u[i] \partial u[j]} = \big(\nabla^2 F(u)\big)[j, i]$.*

2.2 Gauss-Seidel method (one coordinate at a time)

Consider the minimization of a coercive proper convex function $F : \mathbb{R}^n \to \mathbb{R}$.

For each $k \in \mathbb{N}$, the algorithm involves n steps: for $i = 1, \ldots, n$, $u_{k+1}[i]$ is updated according to

$$\begin{aligned} & F(u_{k+1}[1], \ldots, u_{k+1}[i-1], u_{k+1}[i], u_k[i+1], \ldots, u_k[n]) \\ = \inf_{\rho \in \mathbb{R}} \quad & F(u_{k+1}[1], \ldots, u_{k+1}[i-1], \quad \rho, \quad u_k[i+1], \ldots, u_k[n]) \end{aligned}$$

Theorem 12 *F is $\mathcal{C}^1$, strictly convex and coercive $\Rightarrow$ $(u_k)_{k \in \mathbb{N}}$ converges to $\hat{u}$. (see [6])*

Remark 4 Differentiability is essential. E.g. apply the method to

$$F(u[1], u[2]) = u[1]^2 + u[2]^2 - 2\big(u[1] + u[2]\big) + 2\big|u[1] - u[2]\big| \quad \text{with} \quad u_0 = (0, 0).$$

Note that $\inf_{\mathbb{R}^2} F(u) = F(1, 1) = -2$.

Remark 5 *Extension for $F(u) = \|Au - v\|^2 + \sum \lambda_i |u[i]|$, see [6].*

2.3 First-order (Gradient) methods

They use $F(u)$ and $\nabla F(u)$ and rely on $F(u_k - \rho_k d_k) \approx F(u_k) - \rho_k \langle \nabla F(u_k), d_k \rangle$. Different tasks:

- Choose a descent direction $(-d_k)$ - such that $F(u_k - \rho d_k) < F(u_k)$ for $\rho > 0$ small enough;
- Line search : find ρ_k such that $F(u_k - \rho_k d_k) - F(u_k) < 0$ is sufficiently negative.
- Stopping rule: introduce an error function which measures the quality of an approximate solution $\hat{u}$, e.g. choose a norm $\|.\|$, $\tau > 0$ (small enough) and test:
  - Iterates: if $\|u_{k+1} - u_k\| \leqslant \tau$, then stop ;
  - Gradient: if $\|\nabla F(u_k)\| \leqslant \tau$, then stop ;
  - Objective: if $F(u_k) - F(u_{k+1}) \leqslant \tau$, then stop ;

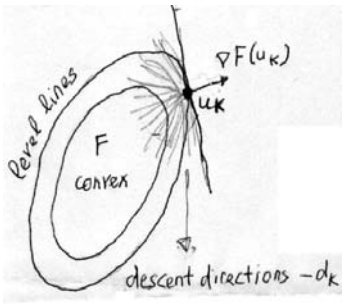
or a combination of these.

Descent direction $-\boldsymbol{d_k}$ for F at $u_k \Leftrightarrow F(u_k - \rho d_k) < F(u_k), \quad \rho > 0$ (small enough)
F differentiable:

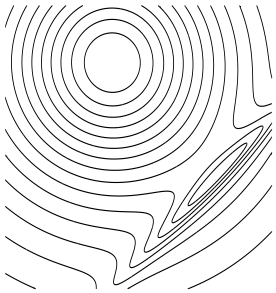
$$\langle \nabla F(u_k), d_k \rangle > 0 \quad \Rightarrow \quad -\boldsymbol{d_k} \text{ is a descent direction for } F \text{ at } u_k \tag{2.1}$$

F convex and differentiable:

$$\langle \nabla F(u_k), d_k \rangle > 0 \quad \Leftrightarrow \quad -\boldsymbol{d_k} \text{ is a descent direction for } F \text{ at } u_k \tag{2.2}$$



$F : \mathbb{R}^2 \to \mathbb{R}$ smooth strictly convex
one minimizer



$F : \mathbb{R}^2 \to \mathbb{R}$ smooth nonconvex
two minimizers

Using a first-order expansion,

$$F(u_k - \rho d_k) = F(u_k) - \rho \langle \nabla F(u_k), d_k \rangle + \rho \|d_k\| \, \varepsilon(\rho d_k) \tag{2.3}$$

$$\text{where} \qquad \varepsilon(\rho d_k) \to 0 \ \text{ as } \ \rho \to 0 \tag{2.4}$$

Then (2.3) can be rewritten as

$$F(u_k) - F(u_k - \rho d_k) = \rho \Big(\langle \nabla F(u_k), d_k \rangle - \|d_k\| \, \varepsilon(\rho d_k) \Big).$$

Consider that $\langle \nabla F(u_k), d_k \rangle > 0$. By (2.4), there is $\overline{\rho} > 0$ such that $\langle \nabla F(u_k), d_k \rangle > \|d_k\| \, \varepsilon(\rho d_k)$, $0 < \rho < \overline{\rho}$. Then $-d_k$ is a descent direction since $F(u_k) - F(u_k - \rho d_k) > 0$, $0 < \rho < \overline{\rho}$.

Note that some methods construct $\rho_k d_k$ in an automatic way.

2.3.1 The steepest descent method

Motivation : make $F(u_k) - F(u_{k+1})$ as large as possible. In (2.3) we have $\langle \nabla F(u_k), d_k \rangle \leqslant \|F(u_k)\| \, \|d_k\|$ (Schwarz inequality) where the equality is reached if $d_k \propto \nabla F(u_k)$.

The steepest descent method = Gradient with optimal stepsize is defined by

$$F\left(u_k - \rho_k \nabla F(u_k)\right) = \inf_{\rho \in \mathbb{R}} F(u_k - \rho \nabla F(u_k)) \tag{2.5}$$

$$u_{k+1} = u_k - \rho_k \nabla F(u_k), \quad k \in \mathbb{N}. \tag{2.6}$$

Theorem 13 *If $F : \mathbb{R}^n \to \mathbb{R}$ is strongly convex, $(u_k)_{k \in \mathbb{N}}$ given by (2.5)-(2.6) converges to the unique minimizer $\hat{u}$ of F.*

Proof. Suppose that $\nabla F(u_k) \neq 0$ (otherwise $\hat{u} = u_k$). Proof in 5 steps.

(a) Denote

$$f_k(\rho) = F\Big(u_k - \rho\nabla F(u_k)\Big), \quad \rho > 0.$$

f_k is coercive, strictly convex, hence it admits a unique minimizer $\rho \overset{\text{def}}{=} \rho(u_k)$ and the latter solves the equation $f_k'(\rho) = 0$. Thus

$$f_k'(\rho) = -\left\langle \nabla F\big(u_k - \rho\nabla F(u_k)\big)\,,\, \nabla F(u_k)\right\rangle = 0 \tag{2.7}$$

Since

$$u_{k+1} = u_k - \rho\nabla F(u_k) \quad \Leftrightarrow \quad \nabla F(u_k) = \frac{1}{\rho}(u_k - u_{k+1}) \tag{2.8}$$

the left hand side of (2.8) inserted into (2.7) yields

$$\langle \nabla F(u_{k+1}), \nabla F(u_k)\rangle = 0, \tag{2.9}$$

i.e. two consecutive directions are orthogonal, while the right hand side equation of (2.8) inserted into (2.7) shows that

$$\langle \nabla F(u_{k+1}), u_k - u_{k+1}\rangle = 0.$$

Using the last equation and the assumption that F is strongly convex,

$$F(u_k) - F(u_{k+1}) \geqslant \langle \nabla F(u_{k+1}), u_k - u_{k+1}\rangle + \frac{\mu}{2}\|u_k - u_{k+1}\|^2 = \frac{\mu}{2}\|u_k - u_{k+1}\|^2 \tag{2.10}$$

(b) Since $\big(F(u_k)\big)_{k\in\mathbb{N}}$ is decreasing and bounded from below by $F(\hat{u})$ we deduce that $\big(F(u_k)\big)_{k\in\mathbb{N}}$ converges, hence

$$\lim_{k\to\infty}\big(F(u_k) - F(u_{k+1})\big) = 0$$

Inserting this result into (2.10) shows that [1]

$$\lim_{k\to\infty}\|u_k - u_{k+1}\| = 0 \tag{2.11}$$

(c) Using (2.9) allows us to write down

$$\|\nabla F(u_k)\|^2 = \langle \nabla F(u_k), \nabla F(u_k)\rangle - \langle \nabla F(u_k), \nabla F(u_{k+1})\rangle = \langle \nabla F(u_k), \nabla F(u_k) - \nabla F(u_{k+1})\rangle$$

By Schwarz's inequality,

$$\|\nabla F(u_k)\|^2 \leqslant \|\nabla F(u_k)\|\ \|\nabla F(u_k) - \nabla F(u_{k+1})\|$$

hence

$$\|\nabla F(u_k)\| \leqslant \|\nabla F(u_k) - \nabla F(u_{k+1})\| \tag{2.12}$$

(d) The facts that $\big(F(u_k)\big)_{k\in\mathbb{N}}$ is decreasing and that F is coercive implies that $\exists\ r > 0$ such that $u_k \in \overline{B(0,r)}$, $\forall k \in \mathbb{N}$. Since $F \sim \mathcal{C}^1$, ∇F is uniformly continuous on the compact $\overline{B(0,r)}$. Using (2.11), $\forall\varepsilon > 0$ there are $\eta > 0$ and $k_0 \in \mathbb{N}$ such that

$$\|u_k - u_{k+1}\| < \eta \quad \Rightarrow \quad \|\nabla F(u_k) - \nabla F(u_{k+1})\| < \varepsilon, \quad \forall k \geqslant k_0.$$

[1] Remind that (2.11) does not mean that the sequence (u_k) converges!
Consider $u_k = \sum_{i=0}^{k}\frac{1}{k+1}$. It is well known that $(u_k)_{k\in\mathbb{N}}$ diverges. Nevertheless, $u_{k+1} - u_k = \frac{1}{k+2} \to 0$ as $k \to \infty$.

Consequently,

$$\lim_{k\to\infty} \|\nabla F(u_k) - \nabla F(u_{k+1})\| = 0.$$

Combining this result with (2.12) shows that

$$\lim_{k\to\infty} \nabla F(u_k) = 0. \tag{2.13}$$

(e) Using that F is strongly convex, that $\nabla F(\hat{u}) = 0$ and Schwarz's inequality,

$$\mu\|u_k - \hat{u}\|^2 \leqslant \langle \nabla F(u_k) - \nabla F(\hat{u}) \,,\, u_k - \hat{u}\rangle = \langle \nabla F(u_k) \,,\, u_k - \hat{u}\rangle \leqslant \|\nabla F(u_k)\| \, \|u_k - \hat{u}\|.$$

Thus we have a bound on the error at iteration k

$$\|u_k - \hat{u}\| \leqslant \frac{1}{\mu}\|\nabla F(u_k)\| \to 0 \quad \text{as} \quad k \to \infty$$

where the convergence result is due to (2.13). □

Remark 6 *Note the role of the assumption that $V = \mathbb{R}^n$ is of finite dimension in this proof.*

Quadratic strictly convex problem. Consider $F : \mathbb{R}^n \to \mathbb{R}$ of the form:

$$F(u) = \frac{1}{2}\langle Bu, u\rangle - \langle c, u\rangle \,, \quad B \succ 0, \quad B = B^T. \tag{2.14}$$

The full algorithm: for any $k \in \mathbb{N}$, do

$$\begin{aligned} d_k &= Bu_k - c \\ \rho_k &= \frac{\|d_k\|^2}{\langle Bd_k, d_k\rangle} \\ u_{k+1} &= u_k - \rho_k d_k \end{aligned}$$

$\Leftrightarrow$ a method to solve a linear system $Bu = c$ when $B \succ 0$ and $B^T = B$.

Theorem 14 *The statement of Theorem 13 holds true if F is $\mathcal{C}^1$, strictly convex and coercive.*

Proof. (Sketch.) $F(u_k)$ is decreasing, bounded from below. Then $\exists\theta > 0$ such that $u_k \in B(0,\theta)$, $\forall k \in \mathbb{N}$, i.e. $(u_k)_{k\in\mathbb{N}}$ is bounded. Hence there exists a convergent subsequence $\left(u_{k_j}\right)_{j\in\mathbb{N}}$; let us denote

$$\overline{u} = \lim_{j\to\infty} u_{k_j}.$$

∇F being continuous and using (2.13), $\lim_{j\to\infty} \nabla F(u_{k_j}) = \nabla F(\overline{u}) = 0$. Remind that F admits a unique minimizer $\hat{u}$ and the latter satisfies $\nabla F(\hat{u}) = 0$. It follows that $\hat{u} = \overline{u}$. □

Note that we do not have any bound on the error $\|u_k - \hat{u}\|$ as in the proof of Theorem 14.

- Q-linear convergence of $\big(F(u_k) - F(\hat{u})\big)_{k\in\mathbb{N}}$ towards zero in a neighborhood of $\hat{u}$, under additional conditions—see [4, p. 33].

- Steepest descent method can be very-very bad: the sequence of iterates is subject to zigzags, the step-size can decrease consderably.
  However, the steepest descent method serves as a basis for all the methods actually used.

Remark 7 Consider

$$F(u[1], u[2]) = \frac{1}{2}\big(\alpha_1 u[1]^2 + \alpha_2 u[2]^2\big), \quad 0 < \alpha_1 < \alpha_2.$$

Clearly, $\hat{u} = 0$ and

$$\nabla F(u) = \begin{bmatrix} \alpha_1 u[1] \\ \alpha_2 u[2] \end{bmatrix}$$

Initialize the steepest descent method with $u_0 \neq 0$. Iterations read

$$u_{k+1} = u_k - \rho \nabla F(u_k) = \begin{bmatrix} u_k[1] - \rho\alpha_1 u_k[1] \\ u_k[2] - \rho\alpha_2 u_k[2] \end{bmatrix}$$

In order to get $u_{k+1} = 0$ we need $\rho\alpha_1 = 1$ and $\rho\alpha_2 = 1$ which is impossible ($\alpha_1 \neq \alpha_2$). Finding the solution $\hat{u} = 0$ needs an infinite number of iterations.

2.3.2 Gradient with variable step-size

V—Hilbert space, $\|u\| = \sqrt{\langle u, u\rangle}$, $\forall u \in V$.

$$\underline{u_{k+1} = u_k - \rho_k \, \nabla F(u_k), \quad \rho_k > 0, \ \forall k \in \mathbb{N}} \tag{2.15}$$

Theorem 15 *Let $F : V \to \mathbb{R}$ be differentiable in V. Suppose $\exists \mu$ and $\exists M$ such that $0 < \mu < M$, and*

(i) $\langle \nabla F(u) - \nabla F(v), \ u - v\rangle \geqslant \mu\|u - v\|^2, \quad \forall (u, v) \in V^2$ *(i.e. F is strongly convex);*

(ii) $\|\nabla F(u) - \nabla F(v)\| \leqslant M\|u - v\|, \quad \forall (u, v) \in V^2$

Consider the iteration (2.15) *where*

$$\begin{aligned} \textit{where} \quad & \frac{\mu}{M^2} - \zeta \leqslant \rho_k \leqslant \frac{\mu}{M^2} + \zeta, \ \forall k \in \mathbb{N} \\ \textit{and} \quad & \zeta \in \Big]0, \frac{\mu}{M^2}\Big[\end{aligned}$$

Then $(u_k)_{k\in\mathbb{N}}$ in (2.15) *converges to the unique minimizer $\hat{u}$ of F and*

$$\|u_{k+1} - \hat{u}\| \leqslant \gamma^k \|u_0 - \hat{u}\|, \quad \textit{where} \ \ \gamma = \sqrt{\zeta^2 M^2 - \frac{\mu^2}{M^2} + 1} < 1.$$

Proof. See the proof of Theorem 31, p. 49 for $\Pi_U = \mathrm{Id}$. □

If F is twice differentiable, condition (ii) becomes $\sup_{u\in V} \big\|\nabla^2 F(u)\big\| \leqslant M$.

Remark 8 Denote by $\hat{u}$ the fixed point of

$$G(u) = u - \rho \nabla F(u).$$

$\nabla G(u) = I - \rho\nabla^2 F(u)$. By Theorem 3, p. 12, convergence is ensured if $\max_i \Big|\lambda_i\big(\nabla^2 F(u)\big)\Big| < \dfrac{1}{\rho}$.

Quadratic strictly convex problem (2.14)**, p. 24.** Convergence is improved—it is ensured if

$$\frac{1}{\lambda_{\max}(B)} - \zeta \leqslant \rho_k \leqslant \frac{1}{\lambda_{\max}(B)} + \zeta \quad \text{with} \quad \zeta \in \left]0,\ \frac{1}{\lambda_{\max}(B)}\right[\ .$$

Proof. See the proof of Proposition 17, p. 52.

2.4 Line search

2.4.1 Introduction

Merit function $f : \mathbb{R} \to \mathbb{R}$

$$f(\rho) = F(u_k - \rho d_k), \quad \rho > 0,$$

where $-d_k$ is a descent direction, e.g. $d_k = \nabla F(u_k)$ and in any case $\langle \nabla F(u_k), d_k \rangle > 0$.

The goal is to choose $\rho > 0$ such that f is decreased enough. If the previous ρ_{k-1} was good one may prefer $\rho \approx \rho_{k-1}$.

Line search is of crucial importance since it is done at each iteration.

We fix at iteration k and drop indexes (when possible). Thus we write

$$f(\rho) = F(u - \rho d), \quad \text{then} \quad f'(\rho) = -\langle \nabla F(u - \rho d), d \rangle .$$

In particular,

$$f'(0) = -\langle \nabla F(u), d \rangle < 0$$

since $-d$ is a descent direction, see (2.1), p. 22.

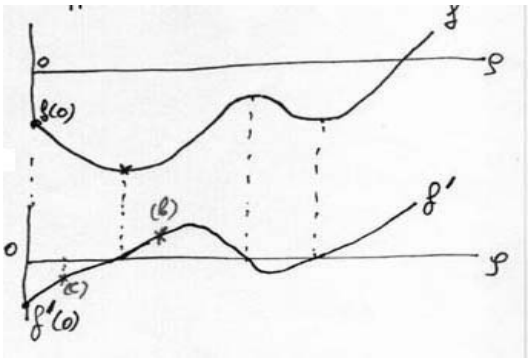
Usually, line search is a subprogram where f can be evaluated only point-wise.

General scheme with 3 possible exists:

(a) $f'(\rho) = 0$—ρ seems to minimize f (true is f convex)

(b) $f'(\rho) > 0$—f seems to have a minimum for a smaller ρ

(c) $f'(\rho) < 0$—f seems to have a minimum for a larger ρ



2.4.2 Schematic algorithm for line-search

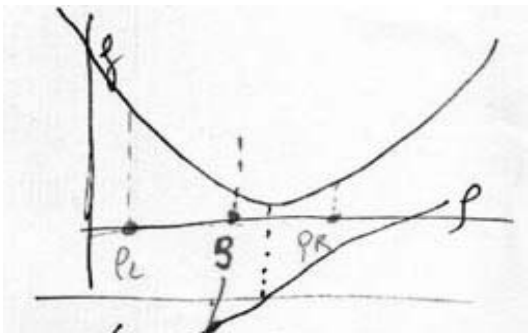
ρ_L–a too small ρ; ρ_R–a too large ρ;

– step (0). Initialize $\rho_L = 0$, ρ_R such that $f'(\rho_R) > 0$ and $\rho > 0$.

An initial ρ_R can be found using extrapolation.

– step (1). Test $\rho > 0$:

- if $f'(\rho) = 0$ then stop.
- if $f'(\rho) < 0$, set $\rho_L = \rho$ and go to Step 2;



  - if $f'(\rho) > 0$, set $\rho_R = \rho$ and go to Step 2.

– step (2). Compute a new $\rho \in]\rho_L, \rho_R[$. Loop to step (1).

For some functions F, ρ can be calculated explicitly—algorithms are faster.

Fletcher's initialization—assume that f is locally quadratic, then take $\rho = 2\frac{F(u_k)-F(u_{k-1})}{f'(0)} > 0$.

Once a ρ_R is found, a line-search algorithm is a sequence of interpolations that reduce the bracket $[\rho_L, \rho_R]$, i.e. ρ_L increases $\leqslant$ ρ_R decreases.

Property 5 *Each $[\rho_L, \rho_R]$ contains a $\hat{\rho}$ such that $f'(\hat{\rho}) = 0$. Infinite interpolations entail $\rho_R - \rho_L = 0$.*

Historically: one has tried to find $\hat{\rho}$ such that $f'(\hat{\rho}) = 0$. Such a $\hat{\rho}$ is called optimal stepsize. Looking for this $\hat{\rho}$ is not a good strategy in practice.

Interpolation methods. There are many choices. For instance:

- Bissection method: $\rho = \dfrac{\rho_L + \rho_R}{2}$
- Polynomial fitting; fit a polynomial that coincides with the points ρ_i already tested and compute a ρ that minimizes this new function explicitly. E.g. use a 2^{nd} or 3^{rd} order polynomial.

Precautions:

- Attention to roundoff errors.
- Avoid infinite loops—impose emergency exits.
- Attention when programming the computation of f'.
- Mathematical proofs need assumptions on f. It may happen that they are not satisfied or that they exclude the roundoff errors.
- Line-search is time consuming!

2.4.3 Modern line-search methods

Arguments:

- Devise tolerant stopping rules—we minimize F and not f!
- Striving to minimize f along the current direction is useless.

Intuitions:

- If f is quadratic: $f(\rho) = \frac{1}{2}c\rho^2 + f'(0)\rho + f(0)$ for $c > 0$, then $\hat{\rho} = -f'(0)/c > 0$;
- If f is affine: $f(\rho) = f(0) + \rho f'(0)$ (etc).

Goal: predict the decrease of f with respect to $f(0)$.

Wolfe's conditions

Here f is the usual merit function, $f(\rho) = F(u_k - \rho d)$, $\rho > 0$. Goals at each iteration:

- decrease f enough (hence u_{k+1} will not be too far from u_k);

- increase f' enough (hence u_{k+1} will not be too close to u_k).

Choose two coefficients $0 < c_0 < c_1 < 1$, e.g. $c_0 < 1/2$ and $c_1 > 1/2$. Do the following 3 tests:

1. $\left\{ \begin{array}{ll} (a) & f(\rho) \leqslant f(0) + c_0\ \rho f'(0) \\ (b) & f'(\rho) \geqslant c_1 f'(0) \end{array} \right. \quad \Rightarrow \quad$ terminate

2. $f(\rho) > f(0) + c_0\ \rho f'(0) \quad \Rightarrow \quad$ set $\rho_R = \rho$ (extrapolation step)

3. $\left\{ \begin{array}{ll} (a) & f(\rho) \leqslant f(0) + c_0\ \rho f'(0) \\ (b) & f'(\rho) < c_1 f'(0) \end{array} \right. \quad \Rightarrow \quad$ set $\rho_L = \rho$ (interpolation step)

Theorem 16 *([4, p. 45]) Suppose that f is $\mathcal{C}^1$ and bounded from below. Then Wolfe's line-search* terminates *(i.e. the number of the line-search iterations is finite).*

Wolfe's line search can be combined with any kind of descent direction $-d$. However, line search is helpless if $-d_k$ is too orthogonal to $\nabla F(u_k)$. The angle θ_k between the direction and the gradient is crucial. Put

$$\cos\theta_k = \frac{\langle \nabla F(u_k), d_k \rangle}{\|\nabla F(u_k)\|\ \|d_k\|} \tag{2.16}$$

We can say that $-d_k$ is a "definite" descent direction if $\cos\theta_k > 0$ is large enough.

Theorem 17 *If $\nabla F(u)$ is Lipschitz-continuous with constant ℓ on $\{u : F(u) \leqslant F(u_0)\}$ and the minimization algorithm uses Wolfe's rule, then*

$$r(\cos\theta_k)^2 \|\nabla F(u_k)\|^2 \leqslant F(u_k) - F(u_{k+1}), \quad \forall k \in \mathbb{N}, \tag{2.17}$$

where the constant $r > 0$ is independent of k.

Theorem 18 *Let $F \sim \mathcal{C}^1$ be bounded from below. Assume that the iterative scheme is such that* (2.17) *holds true for a constant $r > 0$ independent of k. If the series*

$$\sum_{k=0}^{\infty} (\cos\theta_k)^2 \quad \textit{diverges} \tag{2.18}$$

then $\lim_{k\to\infty} \nabla F(u_k) = 0$.

The proofs of Theorems 17 and 18 are given in Appendix 7.2, p. 111 and 7.3, p. 112.

Property (2.18) depends on the way the direction is computed.

Remark 9 *Wolfe's rule needs to compute the value of $f'(\rho) = -\langle \nabla F(u_k - \rho d_k), d_k \rangle$ at each cycle of the line-search. Computing $\nabla F(u_k - \rho d_k)$ is costly when compared to the computation of $F(u_k - \rho d_k)$.*

Other line-search methods avoid the computation of $\nabla F(u_k - \rho d_k)$.

Armijo's methods

Choose $c \in (0, 1)$. Tests for ρ:

1. $f(\rho) \leqslant f(0) + c\rho f'(0) \quad \Rightarrow \quad$ terminate;

2. $f(\rho) > f(0) + c\rho f'(0) \quad \Rightarrow \quad$ set $\rho_R = \rho$.

It is easy to see that this line-search terminates as well. The interest of the method is that it does not need to compute $f'(\rho)$. It ensures that ρ is not too large. However, it can be dangerous since it never increases ρ; thus it relies on the initial step-size ρ_0. More details: [4].

Majorize-Minimize Line Search [22]

The goal is that ρ is an update of the previous ρ_k.

$$\rho_{k+1} \approx \arg\min \ \mathrm{f}(\rho; \rho_k)$$

for

$$\mathrm{f}(\rho; \rho_k) = f(\rho_k) + (\rho - \rho_k) f'(\rho_k) + \frac{1}{2} b_k (\rho - \rho_k)^2$$

where $b_k > 0$ ensures that

$$\mathrm{f}(\rho; \rho_k) \geqslant f(\rho), \quad \forall\, \rho \geqslant 0, \quad \forall\, \rho \geqslant 0.$$

Update:

$$\rho_{k+1} = \rho_k - \theta f'(\rho_k)/b_k$$

where $\theta \in (0, 2)$ is a fixed relaxation parameter. For details see [23, 22].

- Interest: line search is done in one iteration;
- Danger: ρ_k can become arbitrarily small as k increases.

Other Line-Searches

We can evoke e.g. the Goldstein and Price method. More details: e.g. [10, 4].

In practice it often had appeared that taking a good constant step-size yields a faster convergence.

2.5 Hints to solve linear systems

2.5.1 Condition number

Definition 15 *The condition number of an $n \times n$ matrix A, denoted by* $\mathrm{cond}(A)$*, is defined by*

$$\mathrm{cond}(A) \stackrel{\text{def}}{=} \|A\|_2 \ \|A^{-1}\|_2 = \left(\frac{\max_{1 \leqslant i \leqslant n} \left|\lambda_i(A^T A)\right|}{\min_{1 \leqslant i \leqslant n} \left|\lambda_i(A^T A)\right|} \right)^{1/2}.$$

where λ_i are the eigenvalues of $(\cdot)$.

The condition number gives a bound on how inaccurate the solution $A^{-1}v$ is before the effects of round-off error.

Solving $Au = v$ under perturbations [2]:

$$A(u + \delta u) = v + \delta v \quad \Rightarrow \quad \frac{\|\delta u\|}{\|u\|} \leqslant \mathrm{cond}(A) \frac{\|\delta v\|}{\|v\|}$$

[2] From $v = Au$, $\|v\| \leqslant \|A\| \|u\|_2$ or equivalently, $1/\|u\| \leqslant \|A\|/\|v\|$. By $A\delta u = \delta v$ we have $\|\delta u\| \leqslant \|A^{-1}\| \, \|\delta v\|$.

2.5.2 Preconditioning

Preconditioning of circulant matrices

An $n \times n$ matrix C is said to be circulant if its $(j+1)$th row is a cyclic right shift of the jth row, $0 \leqslant j \leqslant n-1$:

$$C = \begin{bmatrix} c_0 & c_{n-1} & c_{n-2} & \dots & c_2 & c_1 \\ c_1 & c_0 & c_{n-1} & \dots & & c_2 \\ \vdots & c_1 & c_0 & \ddots & \ddots & \vdots \\ c_{n-2} & & \ddots & \ddots & & c_{n-1} \\ c_{n-1} & c_{n-2} & c_{n-1} & \dots & c_1 & c_0 \end{bmatrix}$$

Note that C is determined by n components only, c_i for $0 \leqslant i \leqslant n-1$.

Circulant matrices are diagonalized by the discrete Fourier matrix F (see e.g. [24, p. 73]):

$$\mathrm{F}[p,q] = \frac{1}{\sqrt{n}} \exp\left(\frac{2\pi i\, pq}{n}\right), \quad 0 \leqslant p,q \leqslant n-1, \; i \stackrel{\text{def}}{=} \sqrt{-1}$$

We have

$$\mathrm{F}C\mathrm{F}^H = \Lambda = \mathrm{diag}(\lambda_1(C), \cdots, \lambda_n(C)), \quad \text{where} \quad \lambda_p(C) = \sum_{j=0}^{n-1} c_j \exp\left(\frac{2\pi i\, jp}{n}\right), \quad p = 0, \dots, n-1$$

where F^H is the conjugate transpose of F. $\Lambda(C)$ can be obtained in $O(n \log n)$ operations using the fast Fourier transform (FFT) of the first column of C.

Note that $\mathrm{F}^{-1} = \mathrm{F}^H$. Solving

$$Cu = v$$

amounts to solve

$$\Lambda \widetilde{u} = \widetilde{v} \quad \text{where} \quad \widetilde{v} = \mathrm{F}v; \quad \text{then} \quad u = \mathrm{F}^H \widetilde{u}.$$

Here $\mathrm{F}v$ and $\mathrm{F}^H\widetilde{u}$ are computed using the FFT and the inverse FFT, respectively.

For preconditioning of block-circulant matrices for image restoration, see [25, p. 938].

Preconditioning

The speed of algorithms can be substantially improved using preconditioning; see, e.g. [26, 27, 10, 28].

Instead of solving $Bu = c$, we solve the preconditioned system

$$P^{-1}Bu = P^{-1}c,$$

where the $n \times n$ matrix P is called the preconditioner. P is chosen according to the following criteria:

- P should be constructed within $O(n \log n)$ operations;
- $Pv = w$ should be solved in $O(n \log n)$ operations;
- The eigenvalues of $P^{-1}B$ should be clustered (i.e. well concentrated near 1);
- If $B \succ 0$ then $P^{-1}B \succ 0$.

Toeplitz Systems

An $n \times n$ Toeplitz matrix A is defined using $2n-1$ scalars, say a_p, $1-n \leqslant p \leqslant n-1$. It is constant along its diagonals:

$$A = \begin{bmatrix} a_0 & a_{-1} & \dots & a_{-n+2} & a_{-n+1} \\ a_1 & a_0 & a_{-1} & \dots & a_{-n+2} \\ \vdots & a_1 & a_0 & \ddots & \vdots \\ a_{n-2} & & \ddots & \ddots & a_{-1} \\ a_{n-1} & a_{n-2} & \dots & a_1 & a_0 \end{bmatrix} \tag{2.19}$$

Toeplitz matrices are important since they arise in deblurring, recursive filtering, auto-regressive (AR) modeling and many others.

Circulant preconditioners

We emphasize that the use of circulant matrices as preconditioners for Toeplitz systems allows the use of FFT throughout the computations, and FFT is highly parallelizable and has been efficiently implemented on multiprocessors.

Given an integer n, we denote by $\mathcal{C}_n$ the set of all circulant $n \times n$ matrices.

Below we sketch several classical circulant preconditioners.

- Strang's preconditioner [29]

  It is the first circulant preconditioner that was proposed in the literature. P is defined by to be the matrix that copies the central diagonals of A and reflects them around to complete the circulant requirement. For A given by (2.19), the diagonals p_j of the Strang preconditioner $P = [p_{m-\ell}]_{0 \leqslant m,\ell < n}$ are given by

$$p_j = \begin{cases} a_j & 0 < j \leqslant \lfloor n/2 \rfloor \\ a_{j-n} & \lfloor n/2 \rfloor < j < n \\ p_{n+j} & 0 < -j < n \end{cases}$$

  P (of size $n \times n$) satisfies [3]

$$\|P - A\|_1 = \min_{C \in \mathcal{C}_n} \|C - A\|_1 \quad \text{and} \quad \|P - A\|_\infty = \min_{C \in \mathcal{C}_n} \|C - A\|_\infty.$$

- T. Chan's preconditioner for A as in (2.19) [30]

  For A as in (2.19), P is defined by

$$\|P - A\|_F = \min_{C \in \mathcal{C}_n} \|C - A\|_F$$

[3] Remind that for an $n \times n$ matrix B with elements $B[i,j]$ (i for row, j for column)

$$\|B\|_1 = \max_{1 \leqslant j \leqslant n} \sum_{i=1}^{n} \big|B[i,j]\big|$$

$$\|B\|_\infty = \max_{1 \leqslant i \leqslant n} \sum_{j=1}^{n} \big|B[i,j]\big|$$

where $\|.\|_F$ denotes the Frobenius norm[4]. The jth diagonals of P are shown to be equal to

$$p_j = \begin{cases} \frac{(n-j)a_j + j a_{j-n}}{n} & 0 \leqslant j < n \\ p_{n+j} & 0 < -j < n \end{cases}$$

- Preconditioners by embedding [31]

  For A as in (2.19), let B be such that the matrix below is $2n \times 2n$ circulant:

  $$\begin{bmatrix} A & B^H \\ B & A \end{bmatrix}$$

  R. Chan's circulant preconditioner P is defined as $P = A + B$.

- Preconditioners by minimization of norms

  Tyrtyshnikov's circulant preconditioner is defined by

  $$\|\mathrm{Id} - P^{-1}A\|_F = \min_{C \in \mathcal{C}_n} \|\mathrm{Id} - C^{-1}A\|_F.$$

  It has some nice properties and is called the superoptimal circulant preconditioner. See [32].

- Other circulant preconditioners are derived from kernel functions.

Various types of preconditioners are known for different classes of matrices.

2.6 Second-order methods

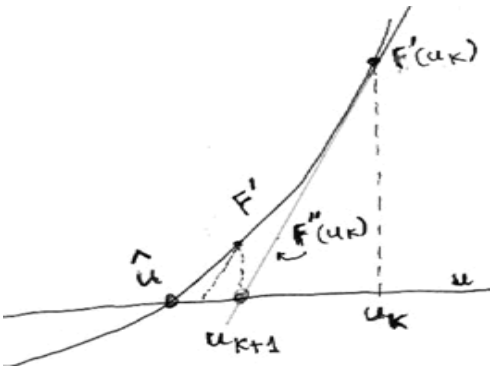
Based on $F(u_{k+1}) - F(u_k) \approx \langle \nabla F(u_k), u_{k+1} - u_k \rangle + \frac{1}{2} \langle \nabla^2 F(u_k)(u_{k+1} - u_k), u_{k+1} - u_k \rangle$

Crucial approach to derive a descent direction at each iteration of a scheme minimizing a smooth objective F.

2.6.1 Newton's method

Example 2 *Let $F : \mathbb{R} \to \mathbb{R}$, find $\hat{u}$ such that $F'(\hat{u}) = \nabla F(\hat{u}) = 0$.*
We choose u_{k+1} such that:

$$\frac{\nabla F(u_k)}{u_k - u_{k+1}} = \nabla^2 F(u_k) = F''(u_k)$$



Newton's method: $$u_{k+1} = u_k - \left(\nabla^2 F(u_k)\right)^{-1} \nabla F(u_k) \tag{2.20}$$

[4]The Frobenius norm of an $n \times n$ matrix B is defined by

$$\|B\|_F = \left(\sum_{i=1}^{n} \sum_{j=1}^{n} \left| B[i,j] \right|^2 \right)^{1/2}.$$

Theorem 19 *Let V be a Hilbert space. Assume that $\nabla^2 F(u_k)$ is continuous and invertible near $\hat{u}$, say $\mathcal{O} \ni \hat{u}$, and that*[5] $\left\| \left(\nabla^2 F(u)\right)^{-1} \right\|_{\mathcal{L}(V)}$ *is bounded above on $\mathcal{O}$. The Newton's method converges towards $\hat{u}$ and the convergence is Q-superlinear.*

Proof. Using Taylor expansion for ∇F about u_k:

$$0 = \nabla F(\hat{u}) = \nabla F(u_k) + \nabla^2 F(u_k)\,(\hat{u} - u_k) + \|\hat{u} - u_k\|\varepsilon(\hat{u} - u_k) = 0,$$

where $\varepsilon(\hat{u} - u_k) \to 0$ as $\|\hat{u} - u_k\| \to 0$. By Newton's method (2.20),

$$\nabla F(u_k) + \nabla^2 F(u_k)\,(u_{k+1} - u_k) = 0$$

Subtracting this equation from the previous one yields

$$\nabla^2 F(u_k)\,(\hat{u} - u_{k+1}) + \|\hat{u} - u_k\|\varepsilon(\hat{u} - u_k) = 0.$$

Since $\nabla^2 F(u_k)$ is invertible, the latter equation is equivalent to

$$\hat{u} - u_{k+1} = -\left(\nabla^2 F(u_k)\right)^{-1}\|\hat{u} - u_k\|\varepsilon(\hat{u} - u_k).$$

From the assumptions, the constant M below is finite:

$$M \stackrel{\text{def}}{=} \sup_{u \in \mathcal{O}(\hat{u})} \left\| \left(\nabla^2 F(u)\right)^{-1} \right\|_{\mathcal{L}(V)}.$$

Then

$$\|u_{k+1} - \hat{u}\| \leqslant M\|\hat{u} - u_k\|\varepsilon(\hat{u} - u_k)$$

and $Q_k = \frac{\|u_{k+1} - \hat{u}\|}{\|\hat{u} - u_k\|} \leqslant M\varepsilon(\hat{u} - u_k) \to 0$ as $\|\hat{u} - u_k\| \to 0$. Hence the Q-superlinearity of the convergence. □

Comments

- Under the conditions of the theorem, if in addition F is $\mathcal{C}^3$ then convergence is Q-quadratic.
- Direction and step-size are automatically chosen ($-\left(\nabla^2 F(u_k)\right)^{-1}\nabla F(u_k)$);
- If F is nonconvex, $-\left(\nabla^2 F(u_k)\right)^{-1}\nabla F(u_k)$ may not be a descent direction, hence the $\hat{u}$ found in this way is not necessarily a minimizer if F;
- Convergence is very fast;
- Computing $\left(\nabla^2 F(u_k)\right)^{-1}$ can be difficult, i.e. solving the equation $\nabla^2 F(u_k)z = \nabla F(u_k)$ with respect to z can require a considerable computational effort;

  Numerically it can be very unstable thus entailing a violent divergence;
- Preconditioning can help if $\nabla^2 F(u_k)$ has a favorable structure, see [27];
- In practice—efficient to get a high-precision $\hat{u}$ if we are close enough to the sought-after $\hat{u}$.

[5] Remind that $\mathcal{L}(V) \equiv \mathcal{L}(V, V)$ and that for $B \in \mathcal{L}(V)$, we have $\|B\|_{\mathcal{L}(V)} = \sup_{u \in V \setminus \{0\}} \frac{\|Bu\|_V}{\|u\|_V}$, where $\|\cdot\|_V$ is the norm on V. If $V = \mathbb{R}^n$, $\|B\|_{\mathcal{L}(V)}$ is just the induced matrix norm.

2.6.2 General quasi-Newton Methods

Approximation : $\big(\mathcal{H}_k(u_k)\big)^{-1} \approx \Big(\nabla^2 F(u_k)\Big)^{-1}$

$$u_{k+1} = u_k - \big(\mathcal{H}_k(u_k)\big)^{-1}\nabla F(u_k) \tag{2.21}$$

Possible to keep $\mathcal{H}_k^{-1}(u_k)$ constant for several consecutive iterations.

Sufficient conditions for convergence:

Theorem 20 *$F : U \subset V \to \mathbb{R}$ twice differentiable in $\mathcal{O}(U)$, V complete n.v.s. Suppose there exist constants $r > 0$, $M > 0$ and $\gamma \in]0,1[$ such that $B \stackrel{\text{def}}{=} \overline{B(u_0, r)} \subset U$ and*

1. $\sup\limits_{k\in\mathbb{N}} \sup\limits_{u\in B} \|\mathcal{H}_k^{-1}(u)\|_{\mathcal{L}(V)} \leqslant M$
2. $\sup\limits_{k\in\mathbb{N}} \sup\limits_{u,v\in B} \|\nabla^2 F(u) - \mathcal{H}_k(v)\|_{\mathcal{L}(V)} \leqslant \dfrac{\gamma}{M}$
3. $\|\nabla F(u_0)\|_V \leqslant \dfrac{r}{M}(1-\gamma)$

Then the sequence generated by (2.21) satisfies:

(i) $u_k \in B, \quad \forall k \in \mathbb{N}$;

(ii) $\lim\limits_{k\to\infty} u_k \to \hat{u}$; *moreover, $\hat{u}$ is the unique zero of $\nabla F(u)$ on B;*

(iii) The convergence is geometric with

$$\|u_k - \hat{u}\|_V \leqslant \frac{\gamma^k}{(1-\gamma)}\|u_1 - u_0\|_V.$$

Proof. We will use that iteration (2.21) is equivalent to

$$\mathcal{H}_k(u_k)\big(u_{k+1} - u_k\big) + \nabla F(u_k) = 0, \quad \forall k \in \mathbb{N}. \tag{2.22}$$

There are several steps.

(a) We will show 3 preliminary results:

$$\|u_{k+1} - u_k\| \leqslant M\|\nabla F(u_k)\|, \quad \forall k \in \mathbb{N}; \tag{2.23}$$

$$u_{k+1} \in B, \quad \forall k \in \mathbb{N}\ ; \tag{2.24}$$

$$\|\nabla F(u_{k+1})\| \leqslant \frac{\gamma}{M}\|u_{k+1} - u_k\|. \tag{2.25}$$

We start with $k = 0$. Iteration (2.21) yields (2.23) and (2.24) for $k = 1$:

$$\underline{\|u_1 - u_0\| \leqslant}\|\mathcal{H}_0^{-1}(u_0)\|\ \|\nabla F(u_0)\| \leqslant \underline{M\ \|\nabla F(u_0)\|} \leqslant M\frac{r}{M}(1-\gamma) < r \quad \Rightarrow \quad \underline{u_1 \in B}$$

Using (2.22) for $k = 0$,

$$\nabla F(u_1) = \nabla F(u_1) - \nabla F(u_0) - \mathcal{H}_0(u_0)\,\big(u_1 - u_0\big)$$

Consider the application $u \mapsto \nabla F(u) - \mathcal{H}_0(u_0)u$. Its gradient is $\nabla^2 F(u) - \mathcal{H}_0(u_0)$. By the generalized mean-value theorem [6] and assumption 2 we obtain (2.25) for $k = 1$:

$$\begin{aligned} \|\nabla F(u_1) - \mathcal{H}_0(u_0)u_1 - \nabla F(u_0) + \mathcal{H}_0(u_0)u_0\| & \\ = \|\nabla F(u_1)\| &\leqslant \sup_{u\in B} \|\nabla^2 F(u) - \mathcal{H}_0(u_0)\|\, \|u_1 - u_0\| \\ &\leqslant \frac{\gamma}{M}\|u_1 - u_0\|. \end{aligned}$$

Suppose that (2.23)-(2.25) hold till $(k-1)$ inclusive, which means we have

$$\begin{aligned} &\|u_k - u_{k-1}\| \leqslant M\|\nabla F(u_{k-1})\| \\ &u_k \in B; \\ &\|\nabla F(u_k)\| \leqslant \frac{\gamma}{M}\|u_k - u_{k-1}\| \end{aligned} \tag{2.26}$$

We check if these hold for k as well. Using (2.26) and assumption 1, iteration (2.21) yields (2.23):

$$\underline{\|u_{k+1} - u_k\| \leqslant} \|\mathcal{H}_k^{-1}(u_k)\|\, \|\nabla F(u_k)\| \leqslant \underline{M\, \|\nabla F(u_k)\|} \leqslant M\frac{\gamma}{M}\, \|u_k - u_{k-1}\| = \gamma\|u_k - u_{k-1}\|$$

Thus we have established that

$$\|u_{k+1} - u_k\| \leqslant \gamma\|u_k - u_{k-1}\| \leqslant \cdots \leqslant \gamma^k\|u_1 - u_0\| \tag{2.27}$$

Using the triangular inequality, (2.23) for $k = 0$ and assumption 3, we get (2.24) for the actual k:

$$\begin{aligned} \|u_{k+1} - u_0\| &\leqslant \|u_{k+1} - u_k\| + \|u_k - u_{k-1}\| + \cdots + \|u_1 - u_0\| = \sum_{i=0}^{k} \|u_{i+1} - u_i\| \\ &\leqslant \left(\sum_{i=0}^{k} \gamma^i\right) \|u_1 - u_0\| \leqslant \left(\sum_{i=0}^{\infty} \gamma^i\right) \|u_1 - u_0\| = \frac{1}{1-\gamma}\|u_1 - u_0\| \\ &\leqslant \frac{M\|\nabla F(u_0)\|}{1-\gamma} \leqslant r \quad \Rightarrow \quad \underline{u_{k+1} \in B.} \end{aligned}$$

Using (2.22) yet again

$$\nabla F(u_{k+1}) = \nabla F(u_{k+1}) - \nabla F(u_k) - \mathcal{H}_k(u_k)\left(u_{k+1} - u_k\right)$$

Applying in a similar way the generalized mean-value theorem to the application $u \mapsto \nabla F(u) - \mathcal{H}_k(u_k)u$ and using assumption 2 entails (2.25):

$$\underline{\|\nabla F(u_{k+1})\|} \leqslant \sup_{u\in B} \|\nabla^2 F(u) - \mathcal{H}_k(u_k)\|\, \|u_{k+1} - u_k\| \underline{\leqslant \frac{\gamma}{M}\|u_{k+1} - u_k\|.}$$

[6] Generalized mean-value theorem. Let $f : U \subset V \to W$ and $a \in U$, $b \in U$ such that the segment $[a, b] \in U$. Assume f is continuous on $[a, b]$ and differentiable on $]a, b[$. Then

$$\|f(b) - f(a)\|_W \leqslant \sup_{u\in]a,b[} \|\nabla f(u)\|_{\mathcal{L}(V,W)}\|b - a\|_V$$

(b) Let us prove the existence of a zero of ∇F in B. Using (2.27), $\forall k \in \mathbb{N},\ \forall j \in \mathbb{N}$ we have

$$\|u_{k+j} - u_k\| \leqslant \sum_{i=0}^{j-1} \|u_{k+i+1} - u_{k+i}\| \leqslant \sum_{i=0}^{j-1} \gamma^{k+i} \|u_1 - u_0\| \leqslant \gamma^k \|u_1 - u_0\| \sum_{i=0}^{\infty} \gamma^i = \frac{\gamma^k}{1-\gamma} \|u_1 - u_0\|, \tag{2.28}$$

hence $(u_k)_{k\in\mathbb{N}}$ is a Cauchy sequence. The latter, combined with the fact that $B \subset V$ is complete shows that $\exists\ \hat{u} \in B$ such that $\lim\limits_{k\to\infty} u_k = \hat{u}$.

(c) Since ∇F is continuous on $\mathcal{O}(U) \supset B$ and using (2.25), we obtain the existence result:

$$\|\nabla F(\hat{u})\| = \lim_{k\to\infty} \|\nabla F(u_k)\| \leqslant \frac{\gamma}{M} \lim_{k\to\infty} \|u_{k+1} - u_k\| = 0$$

(d) Uniqueness of $\hat{u}$ in B.
Suppose $\exists \overline{u} \in B$, $\overline{u} \neq \hat{u}$ such that $\nabla F(\overline{u}) = 0 = \nabla F(\hat{u})$. We can hence write

$$\overline{u} - \hat{u} = -\mathcal{H}_0^{-1}(u_0) \Big(\nabla F(\overline{u}) - \nabla F(\hat{u}) - \mathcal{H}_0(u_0)\,(\overline{u} - \hat{u}) \Big)$$

$$\underline{\|\overline{u} - \hat{u}\|} \leqslant \|\mathcal{H}_0^{-1}(u_0)\| \sup_{u\in B} \|\nabla^2 F(u) - \mathcal{H}_0(u_0)\|\ \|\overline{u} - \hat{u}\| \leqslant M \frac{\gamma}{M}\ \|\overline{u} - \hat{u}\| \underline{< \|\overline{u} - \hat{u}\|}$$

This is impossible, hence $\hat{u}$ is unique in B.

(e) Geometric convergence: using (2.28),

$$\|\hat{u} - u_k\| = \lim_{j\to\infty} \|u_{k+j} - u_k\| \leqslant \frac{\gamma^k}{1-\gamma} \|u_1 - u_0\|.$$

□

Iteration (2.21) is quite general. In particular, Newton's method (2.20) corresponds to $\mathcal{H}_k(u_k) = \nabla^2 F(u_k)$ while (variable) stepsize Gradient descent to $\mathcal{H}_k(u_k) = \frac{1}{\rho_k} I$.

2.6.3 Generalized Weiszfeld's method (1937)

It is a Quasi-Newton method (with linearization of the gradient); see [33, 34, 35]

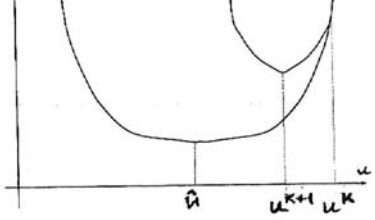
Assumptions: $F : \mathbb{R}^n \to \mathbb{R}$ is twice continuously differentiable, strictly convex, bounded from below and coercive. (F can be constrained to U—a nonempty convex polyhedral set.)

F is approximated from above by a quadratic function $\mathcal{F}$ of the form

$$\mathcal{F}(u, v) = F(v) + \langle u - v, \nabla F(v) \rangle + \frac{1}{2} \langle u - v,\ \mathcal{H}(v)(u - v) \rangle$$

under the assumptions that $\underline{\forall u \in \mathbb{R}^n}$

- $\mathcal{F}(u, v) \geqslant F(u)$, for any fixed v ;
- $\mathcal{H} : \mathbb{R}^n \to \mathbb{R}^{n\times n}$ is continuous and symmetric ;
- $0 < \mu_0 \leqslant \lambda_i\big(\mathcal{H}(u)\big) \leqslant \mu_1 < \infty,\ 1 \leqslant i \leqslant n.$



Note that $\mathcal{F}(u,u) = F(u)$. The iteration for this method reads

$$u_{k+1} = \arg\min_u \mathcal{F}(u, u_k), \quad k \in \mathbb{N}.$$

For u_k fixed, $\mathcal{F}$ is $\mathcal{C}^2$, coercive, bounded below and strictly convex. The minimizer u_{k+1} exists and satisfies the quasi-Newton linear equation

$$0 = \nabla_1 \mathcal{F}(u_{k+1}, u_k) = \nabla F(u_k) + \mathcal{H}(u_k)(u_{k+1} - u_k).$$

Here $\mathcal{H}(.)$ can be seen as an approximation of $\nabla^2 F(.)$.

Recently – increase interest in Weiszfeld's approach [36].

2.6.4 Half-quadratic regularization

Minimize $F : \mathbb{R}^n \to \mathbb{R}$ of the form

$$F(u) = \|Au - v\|_2^2 + \beta \sum_{i=1}^{r} \varphi(\|\mathrm{D}_i u\|_2), \tag{2.29}$$

where $\mathrm{D}_i \in \mathbb{R}^{s\times n}$ for $s \in \{1,2\}$: $s = 2$—e.g. discrete gradients, see (1.8) and $s = 1$—e.g. finite differences, see (1.9), p. 9. Let us denote

$$\mathrm{D} \stackrel{\text{def}}{=} \begin{bmatrix} D_1 \\ D_2 \\ \vdots \\ D_r \end{bmatrix} \in \mathbb{R}^{rs\times n}. \quad \text{Assumption:} \quad \ker(A^TA) \cap \ker(D^TD) = \{0\}.$$

Furthermore, $\varphi : \mathbb{R}_+ \to \mathbb{R}$ is a convex edge-preserving potential function (see [2]), e.g.

φ	φ'	φ''
$\sqrt{t^2+\alpha}$	$\dfrac{t}{\sqrt{\alpha+t^2}}$	$\dfrac{\alpha}{(\sqrt{\alpha+t^2})^3}$
$\alpha \log\left(\cosh\left(\dfrac{t}{\alpha}\right)\right)$	$\tanh\left(\dfrac{t}{\alpha}\right)$	$\left(1-\tanh^2\left(\dfrac{t}{\alpha}\right)\right)\dfrac{1}{\alpha}$
$\lvert t\rvert - \alpha\log\left(1+\dfrac{\lvert t\rvert}{\alpha}\right)$	$\dfrac{t}{\alpha+\lvert t\rvert}$	$\dfrac{\alpha}{(\alpha+\lvert t\rvert)^2}$
$\begin{cases} t^2/2 & \text{if } \lvert t\rvert \leqslant \alpha, \\ \alpha\lvert t\rvert - \alpha^2/2 & \text{if } \lvert t\rvert > \alpha, \end{cases}$	$\begin{cases} t & \text{if } \lvert t\rvert \leqslant \alpha, \\ \alpha\mathrm{sign}(t) & \text{if } \lvert t\rvert > \alpha, \end{cases}$	$\begin{cases} 1 & \text{if } \lvert t\rvert \leqslant \alpha, \\ 0 & \text{if } \lvert t\rvert > \alpha, \end{cases}$
$\lvert t\rvert^\alpha,\ 1 < \alpha \leqslant 2$	$\alpha\lvert t\rvert^{\alpha-1}\mathrm{sign}(t)$	$\alpha(\alpha-1)\lvert t\rvert^{\alpha-2}$

where $\alpha > 0$ is a parameter. Note that F is differentiable with

$$\nabla F(u) = 2A^TAu - 2A^Tv + \beta \sum_{i=1}^{r} \mathrm{D}_i^T \frac{\varphi'(\|\mathrm{D}_i u\|)}{\|\mathrm{D}_i u\|} \mathrm{D}_i u \tag{2.30}$$

However φ' is (nearly) bounded, so φ'' is close to zero on large regions; if A is not well conditioned (a common case in practice), $\nabla^2 F$ has nearly null regions and convergence of usual methods can be extremely slow. Newton's method reads $u_{k+1} = u_k - (\nabla^2 F(u_k)))^{-1}\nabla F(u_k)$. E.g., for $s = 1$ we have $\mathrm{D}_i u \in \mathbb{R}$, so $\|\mathrm{D}_i u\|_2 = |\mathrm{D}_i u|$ and $\nabla^2 F(u) = 2A^TA + \beta \sum_{i=1}^{r} \varphi''(|\mathrm{D}_i u|)\mathrm{D}_i^T\mathrm{D}_i$.

For the class of F considered here, $\nabla^2 F(u)$ has typically a bad condition number so $(\nabla^2 F(u)^{-1})$ is difficult (unstable) or impossible to obtain numerically. Newton's method is practically unfeasible.

Half-quadratic regularization, two forms multiplicative ("×") and additive ("+") introduced in [37] and [25], respectively. Main idea: using an auxiliary variable b, construct an augmented criterion $\mathcal{F}(u,b)$ which is quadratic in u and separable in b, such that

$$(\hat{u},\hat{b}) = \arg\min_{u,b} \mathcal{F}(u,b) \quad \Rightarrow \quad \hat{u} = \arg\min_{u} F(u)$$

$$\text{Alternate minimization:} \quad \forall k \in \mathbb{N}, \quad \left\{ \begin{array}{rcl} u_{k+1} & = & \arg\min_{u} \mathcal{F}(u, b_k) \\ b_{k+1}[i] & = & \arg\min_{b} \mathcal{F}(u_{k+1}, b), \quad 1 \leqslant i \leqslant r \ . \end{array} \right.$$

The construction of these augmented criteria rely on Theorem 9 (Fenchel-Moreau, p. 18).

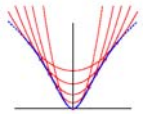
Multiplicative form.

References [38, 39, 40, 2, 41, 42].
Assumptions (easily satisfied) on φ in order to guarantee global convergence:

(a) $t \to \varphi(t)$ is convex and increasing on $\mathbb{R}_+$, $\varphi \not\equiv 0$ and $\varphi(0) = 0$
(b) $t \to \varphi(\sqrt{t})$ is concave on $\mathbb{R}_+$,
(c) φ is $\mathcal{C}^1$ on $\mathbb{R}_+$ and $\varphi'(0) = 0$,
(d) $\varphi''(0^+) > 0$,
(e) $\lim_{t\to\infty} \varphi(t)/t^2 = 0$.

Proposition 1 *We have*

$$\psi(b) = \sup_{t\in\mathbb{R}} \left\{ -\frac{1}{2}bt^2 + \varphi(t) \right\} \quad \Leftrightarrow \quad \varphi(t) = \min_{b\geqslant 0} \left(\frac{1}{2}t^2 b + \psi(b) \right) \qquad (2.31)$$



Moreover, $\hat{b} \stackrel{\text{def}}{=} \left\{ \begin{array}{ll} \dfrac{\varphi'(t)}{t} & \text{if} \quad t > 0 \\ \varphi''(0^+) & \text{if} \quad t = 0 \end{array} \right.$ *is the unique point yielding* $\varphi(t) = \frac{1}{2}t^2\hat{b} + \psi(\hat{b})$.

The proof of the proposition is outlined in Appendix 7.4, p. 113.

Based on Proposition 1 we minimize w.r.t. (u,b) an $\mathcal{F}$ of the form given below

$$\mathcal{F}(u,b) = \|Au - v\|_2^2 + \beta \sum_{i=1}^{r} \left(\frac{b[i]}{2} \|\mathrm{D}_i u\|_2^2 + \psi(b[i]) \right).$$

Note that $(u,b) \mapsto \mathcal{F}(u,b)$ is nonconvex. Iterates for $k \in \mathbb{N}$ read as [7]:

$$u_{k+1} = (H(b_k))^{-1} 2A^T v, \qquad H(b) = 2A^T A + \beta \sum_{i=1}^{r} b[i] \mathrm{D}_i^T \mathrm{D}_i \qquad (2.32)$$

$$b_{k+1}[i] = \frac{\varphi'(\|\mathrm{D}_i u_{k+1}\|_2)}{\|\mathrm{D}_i u_{k+1}\|_2}, \quad 1 \leqslant i \leqslant r. \qquad (2.33)$$

Combining the expression for ∇F in (2.30) along with $\nabla_u \mathcal{F}$ on can see that

[7] Note that $\nabla_u \mathcal{F} = H(b)u - 2A^T v$

Theorem 21 ([41]) *Let F be of the form (2.29) where $\ker A \cap \ker D = \{0\}$, all assumptions (a)-(e) (p. 38) hold and $\varphi''(0) > 0$. Let $\hat{u}$ denote the unique minimizer of F. The sequence $\{u_k\}$ generated by (2.32)-(2.33) satisfies $u_k \to \hat{u}$ as $k \to \infty$. Furthermore, the iterates satisfy*

$$u_{k+1} = u_k - (\mathcal{H}(u_k))^{-1} \nabla F(u_k) \quad \text{with} \quad \mathcal{H}(u) = H\left(\left(\frac{\varphi'(\|\mathrm{D}_i u\|_2)}{\|\mathrm{D}_i u\|_2}\right)_{i=1}^{r}\right)$$

where $H(\cdot)$ is given in (2.32) and $\mathcal{H}(u) \succ 0$, $\forall\, u \in \mathbb{R}^n$.

Comments

- u_k in (2.32) can be computed efficiently using CG (see § 2.7) if direct inversion is heavy.
- Interpretation : b_k edge variable (≈ 0 if "discontinuity")
- $\mathcal{F}$ is nonconvex w.r.t. (u, b). However, under (a)-(e) the minimum is unique and convergence is ensured, $u_k \to \hat{u}$ (see [41, 42]).
- By Theorem 21, this is a Quasi-Newton method. Compare with $\mathcal{H}$ in Newton method (2.20).
- Iterates amount to the relaxed fixed point iteration described next.

Theorem 22 ([43]) *Let F be as in (2.29) and all assumptions in Theorem 21 hold. Put $\nabla F(u)$ into the form*

$$\nabla F(u) = L(u)u - z$$

where $z := 2A^T y$ is independent of u. Then $L(u) \in \mathbb{R}^{n \times n}$ is invertible and the iterates given by

$$u_{k+1} = (L(u_k))^{-1} z \tag{2.34}$$

are equal to the iterates given by (2.32) for any k.

The scheme in (2.34) is known as the relaxed fixed point iteration. It amounts to linearize ∇F at each iteration. By the last theorem, the multiplicative algorithm (2.32)-(2.33) produces the same vectors u_k as the scheme in (2.34). Observe that the relaxed fixed point iteration in (2.34) is much simpler than (2.32)-(2.33) – there is no need to compute and to store the auxiliary vector b whose size typically is close to $2n$.

Non-convex objectives The multiplicative for has been used to solve non-convex problems since the beginning. Local convergence results were established e.g. in [39]. Recently, the set of simple conditions given below were considered in [44]:

$$\begin{array}{ll} (a) & t \to \varphi(t) \ \text{ is increasing and nonconstant on } \ \mathbb{R}_{\geqslant 0} \\ (b) & t \to \varphi(t) \ \text{ is } \ \mathcal{C}^1 \ \text{ on } \mathbb{R}_{>0} \ \text{ and continuous at zero} \\ (c) & t \to t^{-1}\varphi'(t) \ \text{ is decreasing and bounded on } \ \mathbb{R}_{>0}. \end{array}$$

The boundedness assumption in (c) implies that $\varphi'(0) = 0$, hence the objective is smooth. Under these conditions, monotone convergence to an isolated minimizer is established in [44]. Nonconvex functions φ satisfying the above conditions are, e.g.,

$$\varphi(t) = \ln(1 + t^2) \qquad \text{and} \qquad \varphi(t) = \frac{t^2}{1 + t^2}$$

A detailed analysis of the convergence process is exhibited in [44].

Additive form.

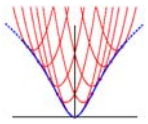
References: [45, 40, 41, 42].
Usual assumptions on φ to ensure convergence:

$$\begin{array}{ll}(a) & t \to \varphi(t) \text{ is convex and increasing on } \mathbb{R}_+, \quad \varphi \not\equiv 0 \text{ and } \varphi(0) = 0\\ (b) & t \to t^2/2 - \varphi(t) \text{ is convex,}\\ (c) & \varphi'(0) = 0,\\ (d) & \lim_{t\to\infty} \varphi(t)/t^2 < 1/2.\end{array}$$

Notice that φ is differentiable on $\mathbb{R}_+$. Indeed, (a) implies that $\varphi'(t^-) \leqslant \varphi'(t^+)$, for all t. If for some t the latter inequality is strict, (b) cannot be satisfied. If follows that $\varphi'(t^-) = \varphi'(t^+) = \varphi'(t)$.

In order to avoid additional notations, we write $\varphi(\|.\|)$ and $\psi(\|.\|)$ next.

Proposition 2 *For $\|.\| = \|.\|_2$ and $b \in \mathbb{R}^s$, $t \in \mathbb{R}^s$, $s \in \{1, 2\}$, we have:*



$$\psi(\|b\|) = \max_{t\in\mathbb{R}^s} \left\{ -\frac{1}{2}\|b - t\|^2 + \varphi(\|t\|) \right\} \quad \Leftrightarrow \quad \varphi(\|t\|) = \min_{b\in\mathbb{R}^s} \left(\frac{1}{2}\|t - b\|^2 + \psi(\|b\|) \right) \tag{2.35}$$

Moreover, $\hat{b} \stackrel{\text{def}}{=} t - \varphi'(\|t\|)\dfrac{t}{\|t\|}$ *is the unique point yielding* $\varphi(\|t\|) = \frac{1}{2}\|t - \hat{b}\|^2 + \psi(\|\hat{b}\|)$.

For the proof – see Appendix 7.5

Based on Proposition 2, we consider

$$\mathcal{F}(u, b) = \|Au - v\|_2^2 + \beta \sum_{i=1}^{r} \left(\frac{1}{2} \|\mathrm{D}_i u - b_i\|_2^2 + \psi(b_i) \right), \quad b_i \in \mathbb{R}^s \tag{2.36}$$

and the iterations read, $\forall k \in \mathbb{N}$,

$$u_{k+1} = \mathcal{H}^{-1} \left(2A^T v + \beta \sum_{i=1}^{r} \mathrm{D}_i^T b_{i_k} \right), \qquad \mathcal{H} = 2A^T A + \beta \sum_{i=1}^{r} \mathrm{D}_i^T \mathrm{D}_i \tag{2.37}$$

$$b_{i_{k+1}} = \mathrm{D}_i u_k - \varphi'(\|\mathrm{D}_i u_k\|_2) \frac{\mathrm{D}_i u_k}{\|\mathrm{D}_i u_k\|_2}, \quad 1 \leqslant i \leqslant r. \tag{2.38}$$

Theorem 23 ([41]) *Let F be of the form (2.29) where $\ker A \cap \ker D = \{0\}$, all assumptions (a)-(d) (p. 40) hold and $\varphi''(0) > 0$. Let $\hat{u}$ denote the minimizer of F. The sequence $\{u_k\}$ generated by (2.37)-(2.38) satisfies $u_k \to \hat{u}$ as $k \to \infty$. Furthermore, the iterates satisfy*

$$u_{k+1} = u_k - \mathcal{H}^{-1} \nabla F(u_k)$$

where $\mathcal{H}$ is given in (2.37) and $\mathcal{H} \succ 0$.

Comments

- By Theorem 23, (2.37)-(2.38) is a Quasi-Newton method. Compare with $\mathcal{H}$ in Newton method (2.20).
- If the inverse $\mathcal{H}^{-1} \in \mathbb{R}^{n\times n}$ cannot be stored, it has to be evaluated at each iteration. Preconditioning of $\mathcal{H}$ (see § 2.5.2, p. 30) is usually easy and accelerates the algorithm.

Comparison between the two forms.

- " $\times$ "-form: less iterations but each iteration is expensive;
- " $+$ "-form: more iterations but each iteration is cheaper and it is possible to precondition $\mathcal{H}$.
- For both forms, convergence is faster than Gradient, DFP and BFGS (§ 2.6.5), and non-linear CG (§ 2.7.2) for functions of the form (2.29).
- Overall, the " $+$ "-form is faster than the " $\times$ "-form in (2.32)-(2.33). For details, see [41].
- We are not aware about comparisons between the " $+$ "-form and the relaxed fixed point iteration (Theorem 22) which yields the same iterates u_k as the " $\times$ "-form .

2.6.5 Standard quasi-Newton methods

References: [4, 46, 10]. Here $F : \mathbb{R}^n \to \mathbb{R}$ with $\|u\| = \sqrt{\langle u, u\rangle}$, $\forall u \in \mathbb{R}^n$.

$\mathcal{H}_k \approx \nabla^2 F(u_k) \succ 0$ and $M_k \approx \left(\nabla^2 F(u_k)\right)^{-1} \succ 0$, $\forall k \in \mathbb{N}$.

$$\begin{aligned} \Delta_k &\stackrel{\text{def}}{=} u_{k+1} - u_k \\ g_k &\stackrel{\text{def}}{=} \nabla F(u_{k+1}) - \nabla F(u_k) \end{aligned}$$

Definition 16 *Quasi-Newton (or secant) equation:*

$$M_{k+1} g_k = \Delta_k \tag{2.39}$$

Interpretation: the mean value H of $\nabla^2 F$ between u_k and u_{k+1} satisfies $g_k = H\Delta_k$. So (2.39) forces M_{k+1} to have the same action as H^{-1} on g_k (subspace of dimension one).

There are infinitely many quasi-Newton matrices satisfying Definition 16.

General quasi-Newton scheme

0. Fix u_0, tolerance $\varepsilon \gtrapprox 0$, $M_0 \succ 0$ such that $M_0 = M_0^T$. If $\|\nabla F(u_0)\| \leqslant \varepsilon$ stop, else go to 1.

   For $k = 1, 2, \ldots$, do:

1. $d_k = M_k \nabla F(u_k)$
2. Line search along $-d_k$ to find ρ_k (e.g. Wolfe, $\rho_0 = 1$)
3. $u_{k+1} = u_k - \rho_k d_k$
4. if $\|\nabla F(u_{k+1})\| \leqslant \varepsilon$ stop else go to step 5
5. $M_{k+1} = M_k + C_k \succ 0$ so that $M_{k+1} = M_{k+1}^T$ and (2.39) holds; then loop to step 1

$C_k \succeq 0$ is a correction matrix. For stability and simplicity, it should be "minimal" in some sense. Various choices for C_k can be done.

DFP (Davidon-Fletcher-Powel)

Historically the first (1959). Apply the General quasi-Newton scheme along with

$$\begin{aligned} M_{k+1} &= M_k + C_k \\ C_k &= \frac{\Delta_k \Delta_k^T}{g_k^T \Delta_k} - \frac{M_k g_k \, g_k^T M_k}{g_k^T M_k g_k} \end{aligned}$$

Each of the matrices composing C_k is of rank $\leqslant 1$, hence $\operatorname{rank} C_k \leqslant 2$, $\forall k \in \mathbb{N}$. Note that $C_k = C_k^T$.

M_{k+1} is symmetric and satisfies the quasi-Newton equation (2.39):

$$M_{k+1} g_k = M_k g_k + \Delta_k \frac{\Delta_k^T g_k}{g_k^T \Delta_k} - M_k g_k \frac{g_k^T M_k g_k}{g_k^T M_k g_k} = M_k g_k + \Delta_k - M_k g_k = \Delta_k$$

BFGC (Broyden-Fletcher-Goldfarb-Shanno)

Proposed in 1970. Apply the General quasi-Newton scheme along with

$$\begin{aligned} M_{k+1} &= M_k + C_k \\ C_k &= -\frac{\Delta_k g_k^T M_k + M_k g_k \Delta_k^T}{g_k^T \Delta_k} + \left(1 + \frac{g_k^T M_k g_k}{g_k^T \Delta_k}\right) \frac{\Delta_k \Delta_k^T}{g_k^T \Delta_k} \end{aligned}$$

Obviously M_{k+1} satisfies (2.39) as well. BFGC is often preferred to DFP.

Comment: One can approximate $\nabla^2 F$ using $\mathcal{H}_{k+1} = \mathcal{H}_k + \widetilde{C}_k$ where $\mathcal{H}_k$ must satisfy the so called "dual" quasi-Newton equation, $\mathcal{H}_{k+1} \Delta_k = g_k$, $\forall k$ (see [4, p. 55]).

Theorem 24 ([4], p. 58.) *Let $M_0 \succ 0$ (resp. $\mathcal{H}_0 \succ 0$). Then $g_k^T \Delta_k > 0$ is a necessary and sufficient condition for DFP and BFGS formulae to give $M_k \succ 0$ (resp. $\mathcal{H}_k \succ 0$), $\forall k \in \mathbb{N}$.*

Remark 10 *It can be shown that DFP and BFGC formulae are mutually dual.* See [4, p. 56].

Theorem 25 ([4], p. 58.) *Let F be convex, bounded from below and ∇F Lipschitzian on $\{u : F(u) \leqslant F(u_0)\}$. Then the BFGS algorithm with Wolfe's line-search and $M_0 \succ 0$, $M_0 = M_0^T$ yields*

$$\liminf_{k \to \infty} |\nabla F(u_k)| = 0.$$

- Locally - Q superlinear convergence;
- Drawback: we have to store $n \times n$ matrices;
- DFP, BFGC available in Matlab Optimization toolbox.

Recently the BFGS minimization method was extended in [47] to handle nonsmooth, not necessarily convex problems. The Matlab package HANSO developed by the authors is freely available[8].

[8] http://www.cs.nyu.edu/overton/software/hanso/

2.7 Subspace methods

Main idea:

$$u_{k+1} = u_k - \mathrm{D}_k R_k$$

where D_k is a subspace of descent directions and R_k is a multidimensional stepsize.

> In general it is difficult to construct a multidimensional stepsize method combining suitable convergence properties and low computational cost.

2.7.1 Linear Conjugate Gradient method (CG), 1952

Due to Hestenes and Stiefel, 1952, still contains a lot of material for research.

Preconditioned CG is an important technique to solve large systems of equations.

For $B = B^T$ and $B \succ 0$, powerful method to minimize F given below for n very large

$$F(u) = \frac{1}{2}\langle Bu, u\rangle - \langle c, u\rangle, \quad u \in \mathbb{R}^n$$

or equivalently to solve $Bu = c$. (Remind that $B \succ 0$ implies $\lambda_{\min}(B) > 0$.) Then F is strongly convex. By the definition of F, we have a very useful relation:

$$\nabla F(u - v) = B(u - v) - c = \nabla F(u) - Bv, \quad \forall u, v \in \mathbb{R}^n. \tag{2.40}$$

Main idea : at each iteration, compute u_{k+1} such that

$$\begin{aligned} F(u_{k+1}) &= \inf_{u \in u_k + H_k} F(u) \\ H_k &= \mathrm{Span}\{\nabla F(u_i), 0 \leqslant i \leqslant k\} \end{aligned}$$

u_{k+1} minimizes F over an affine subspace (and not only along one direction, as in Gradient methods.)

Theorem 26 *The CG method converges after n iterations at most and it provides the unique exact minimiser $\hat{u}$ obeying $F(\hat{u}) = \min_{u \in \mathbb{R}^n} F(u)$.*

Proof. Define the subspace

$$H_k = \left\{ g(\alpha) = \sum_{i=0}^{k} \alpha[i] \nabla F(u_i) : \alpha[i] \in \mathbb{R},\ 0 \leqslant i \leqslant k \right\} \tag{2.41}$$

H_k is a closed convex set. Set $f(\alpha) \stackrel{\text{def}}{=} F\big(u_k - g(\alpha)\big)$. Note that f is convex and coercive.

$$F(u_{k+1}) = \inf_{\alpha \in \mathbb{R}^{k+1}} F\big(u_k - g(\alpha)\big) = \inf_{\alpha \in \mathbb{R}^{k+1}} f(\alpha) = f(\alpha_k).$$

α_k is the unique solution of $\nabla f(\alpha) = 0$. By (2.41), $\partial g(\alpha)/\partial \alpha[i] = \nabla F(u_i)$. Then

$$\frac{\partial f(\alpha)}{\partial \alpha[i]} = 0 = -\langle \nabla F(u_k - g(\alpha)), \nabla F(u_i)\rangle = -\langle \nabla F(u_{k+1}), \nabla F(u_i)\rangle, \quad 0 \leqslant i \leqslant k. \tag{2.42}$$

Hence for $0 \leqslant k \leqslant n-1$ we have

$$\underline{\langle \nabla F(u_{k+1}), \nabla F(u_i)\rangle = 0, \quad 0 \leqslant i \leqslant k} \tag{2.43}$$

$$\Rightarrow \quad \underline{\langle \nabla F(u_{k+1}), u\rangle = 0, \quad \forall u \in H_k} \tag{2.44}$$

It follows that $\{\nabla F(u_i)\}$ are linearly independent. Conclusions about convergence:

- If $\nabla F(u_k) = 0$ then $\hat{u} = u_k$ (terminate).
- If $\nabla F(u_k) \neq 0$ then $\dim H_k = k + 1$.

  Suppose that $\nabla F(u_{n-1}) \neq 0$, then $H_{n-1} = \mathbb{R}^n$. By (2.44),

$$\langle \nabla F(u_n), u \rangle = 0, \quad \forall u \in \mathbb{R}^n \quad \Rightarrow \quad \nabla F(u_n) = 0 \quad \Rightarrow \quad \hat{u} = u_n.$$

The proof is complete. □

Remark 11 *In practice, numerical errors can require a higher number of iterations.*

The derivation of the CG algorithm is outlined in Appendix 7.6 on p. 114

CG Algorithm:

0. Initialization: $d_{-1} = 0$, $\|\nabla F(u_{-1})\| = 1$ and $u_0 \in \mathbb{R}^n$. Then for $k = 0, 1, 2, \cdots$ do:
1. if $\nabla F(u_k) = 0$ then terminate; else go to step 2
2. $\xi_k = \dfrac{\|\nabla F(u_k)\|^2}{\|\nabla F(u_{k-1})\|^2}$ (where $\nabla F(u_k) = Bu_k - c$, $\forall k$)
3. $d_k = \nabla F(u_k) + \xi_k \, d_{k-1}$ (by step 0, we have $d_0 = \nabla F(u_0)$)
4. $\rho_k = \dfrac{\langle \nabla F(u_k), d_k \rangle}{\langle B d_k, d_k \rangle}$ (where $B = \nabla^2 F(u_k)$, $\forall k$)
5. $u_{k+1} = u_k - \rho_k d_k$; then loop to step 1.

Main Properties :

- $\forall k$, $\langle \nabla F(u_{k+1}), \nabla F(u_i) \rangle = 0$ if $0 \leqslant i \leqslant k$
- $\forall k$, $\langle B d_{k+1}, d_i \rangle = 0$ if $0 \leqslant i \leqslant k$— directions are conjugated w.r.t. $\nabla^2 F = B$
- since $B \succ 0$ it follows that $\underline{(d_k)_{k=1}^{\widetilde{n}} \text{ are linearly independent}}$ where $\widetilde{n} \leqslant n$ is such that $\nabla F(u_{\widetilde{n}}) = 0$.
- CG does orthogonalization: $D^T B D$ is diagonal where $D = [d_1, \ldots, d_n]$ (the directions).

Theorem 27 ([10], p. 114) *If B has only r different eigenvalues, then the CG method converges after r iterations at most.*

Convergence is faster if the eigenvalues of B are "concentrated".

2.7.2 Non-quadratic Functionals (non-linear CG)

References: [10, 4]

Nonlinear variants of the CG are well studied and have proved to be quite successful in practice. However, in general there is no guarantee to converge in a finite number of iterations.

Main idea: cumulate past information when choosing the new descent direction.

Fletcher-Reeves method (FR)

step 0. Initialization: $d_{-1} = 0$, $\|\nabla F(u_{-1})\| = 1$ and $u_0 \in \mathbb{R}^n$.

$\forall k \in \mathbb{N}$:

1. Compute $\nabla F(u_k)$; if $\|\nabla F(u_k)\| \leqslant \varepsilon$ then stop, otherwise go to step 2
2. $\xi_k = \dfrac{\|\nabla F(u_k)\|^2}{\|\nabla F(u_{k-1})\|^2}$
3. $d_k = \nabla F(u_k) + \xi_k \, d_{k-1}$
4. Test: if $\langle \nabla F(u_k), d_k \rangle < 0$ ($-d_k$ is not a descent direction) set $d_k = \nabla F(u_k)$
5. Line-search along $-d_k$ to obtain $\rho_k > 0$
6. $u_{k+1} = u_k - \rho_k d_k$, then loop to step 1.

C. Comments (see [4, p. 72])

- The new direction d_k still involves memory from previous directions ("Markovian property").
- Conjugacy has little meaning in the non-quadratic case.
- FR is heavily based on the locally quadratic shape of F (only step 4 is new w.r.t. CG).
- If the line-search is exact, the new direction $-d_k$ is a descent direction.

Polak-Ribière (PR) method

Remark 12 *There are many variants of the FR method that differ from each other mainly in the choice of the parameter ξ_k. An important variant is the PR method.*

By the mean-value formula there is $\widetilde{\rho} \in]0, \rho_{k-1}]$ such that along the line $\mathrm{span}(d_{k-1})$ we have

$$\nabla F(u_k) = \nabla F(u_{k-1}) - \rho_{k-1} B_k \, d_{k-1} \tag{2.45}$$

$$\text{for } B_k = \nabla^2 F(u_{k-1} - \widetilde{\rho} d_{k-1}) \tag{2.46}$$

Note that $B_k = B_k^T$.

Main idea: choose a ξ_k in the FR method that conjugates d_k and d_{k-1} w.r.t. B_k, i.e. that yields

$$\langle B_k d_{k-1} \, , d_k \rangle = 0.$$

However, B_k is unknown. By way of compromise, the result given below will be used.

Lemma 2 *Let B_k be given by* (2.45)-(2.46) *and consider the FR methods where*

(i) $\xi_k = \dfrac{\langle \nabla F(u_k) - \nabla F(u_{k-1}), \nabla F(u_k) \rangle}{\|\nabla F(u_{k-1})\|^2}$ *(the PR formula)*

(ii) *ρ_{k-1} and ρ_{k-2} are optimal.*

Then $\langle B_k d_{k-1} \, , d_k \rangle = 0$.

The proof is given in Appendix 7.7, p. 116

PR method

Apply the FR method, p. 45, where ξ_k in step 2 is replaced by the PR formula

$$\xi_k = \frac{\langle \nabla F(u_k) - \nabla F(u_{k-1}), \nabla F(u_k) \rangle}{\|\nabla F(u_{k-1})\|^2}.$$

Comparison between FR and PR (see [4, p. 75])

- FR converges globally (F convex, coercive).
- There exists a counter-example where PR does not converge
- PR converges if F is locally strongly convex
- PR converges much faster than FR (so the latter is rarely used in practice)
- Relation with quasi-Newton

Chapter 3

Constrained optimization

3.1 Preliminaries

References : [48, 5, 4], also [6, 49, 12].

3.2 Optimality conditions

Theorem 28 (Euler (in)equalities) *Let V be a real n.v.s., $U \subset V$ convex, $F : \mathcal{O}(U) \to \mathbb{R}$ proper (see the definition on p. 13) and differentiable at $\hat{u} \in U$.*

1. *If F admits at $\hat{u}$ a minimum w.r.t. U, then*

$$\langle \nabla F(\hat{u}), u - \hat{u} \rangle \geqslant 0, \quad \forall u \in U \tag{3.1}$$

2. *Assume in addition that F is convex.* Necessary and sufficient condition (NSC) for a minimum*:*

$$F(\hat{u}) = \min_{u \in U} F(u) \quad \Leftrightarrow \quad (3.1) \text{ holds.}$$

Moreover, if F is strictly convex, the minimizer $\hat{u}$ is unique.

3. *If U is open, then : (3.1) $\iff$ $\nabla F(\hat{u}) = 0$.*

Proof. We have $F(\hat{u}) \leqslant F(\hat{u} + v)$ for any v satisfying $\hat{u} + v \in U$. Consider an arbitrary v such that

$$u = \hat{u} + v \in U.$$

Since U is convex,

$$\theta \in [0, 1] \quad \Rightarrow \quad \theta u + (1 - \theta)\hat{u} = \theta(\hat{u} + v) + (1 - \theta)\hat{u} = \hat{u} + \theta v \ \in U.$$

The first-order expansion of F about $\hat{u} + \theta v$ reads

$$F(\hat{u} + \theta v) - F(\hat{u}) = \theta \left\langle \nabla F(\hat{u}) \, , \, v \right\rangle + \theta \|v\| \varepsilon(\theta v). \tag{3.2}$$

Suppose that $\langle \nabla F(\hat{u}) \, , \, v \rangle < 0$ and note that v is fixed. Since $\varepsilon(\theta v) \to 0$ as $\theta \to 0$, we can find θ small enough such that $\langle \nabla F(\hat{u}) \, , \, v \rangle + \|v\| \left| \varepsilon(\theta v) \right| < 0$, hence $F(\hat{u}) > F(\hat{u} + \theta v)$ by (3.2), i.e. there is no minimum at $\hat{u}$. It follows that necessarily

$$\underline{\langle \nabla F(\hat{u}) \, , \, u - \hat{u} \rangle} = \langle \nabla F(\hat{u}) \, , \, v \rangle \underline{\geqslant 0}.$$

Statement 2. The necessity of (3.1) follows from statement 1. To show its sufficiency, combine (3.1) with the convexity of F, Property 2, statement 1 (p. 20): $F(u) - F(\hat{u}) \geqslant \langle \nabla F(\hat{u}), u - \hat{u} \rangle$, $\forall u \in U$. If F is strictly convex, the uniqueness of $\hat{u}$ follows from (3.1) and Property 2-2, namely $F(u) - F(\hat{u}) > \langle \nabla F(\hat{u}), u - \hat{u} \rangle$, $\forall u \in U$, $u \neq \hat{u}$.

Statement 3 follows directly from statement 1. □

3.2.1 Projection theorem

Theorem 29 (Projection) *Let V be real Hilbert space, $U \subset V$ nonempty, convex and closed, and $\|u\| = \sqrt{\langle u, u \rangle}$, $\forall u \in V$. Given $v \in V$, the following statements are equivalent:*

1. *$\exists \hat{u} = \Pi_U v \in U$ unique such that $\|v - \hat{u}\| = \inf_{u \in U} \|v - u\|$, where Π_U is the projection onto U;*
2. *$\hat{u} \in U$ and $\langle v - \hat{u}, u - \hat{u} \rangle \leqslant 0$, $\forall u \in U$.*

Classical proof [14, p.391] or [9]. Shorter proof using Theorem 28—see below.

Proof. If $v \in U$, then $\hat{u} = v$ and $\Pi_U = \mathrm{Id}$. Consider next that $u \in V \setminus U$ and

$$F(u) = \frac{1}{2}\|v - u\|^2.$$

F is clearly $\mathcal{C}^\infty$, strictly convex and coercive. The projection $\hat{u}$ of v onto U solves the problem:

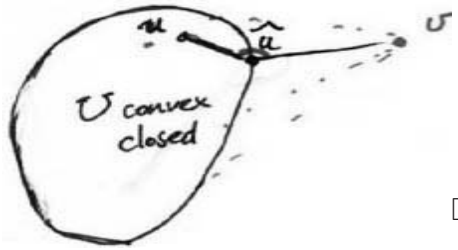
$$F(\hat{u}) = \inf_{u \in U} F(u) = \min_{u \in U} F(u).$$

By Theorem 28-2, such an $\hat{u}$ exists and it is unique; it satisfies $\langle \nabla F(\hat{u}), u - \hat{u} \rangle \geqslant 0$, $\forall u \in U$. Noticing that

$$\nabla F(u) = u - v,$$

we get

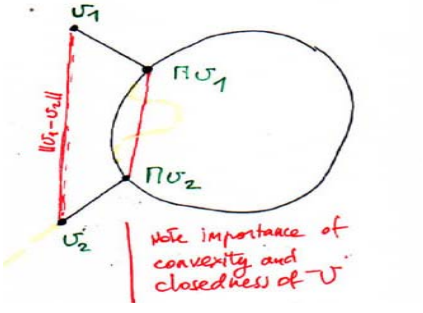
$$\langle \hat{u} - v, u - \hat{u} \rangle \geqslant 0, \quad \forall u \in U \quad \Leftrightarrow \quad \langle v - \hat{u}, u - \hat{u} \rangle \leqslant 0, \quad \forall u \in U.$$



□

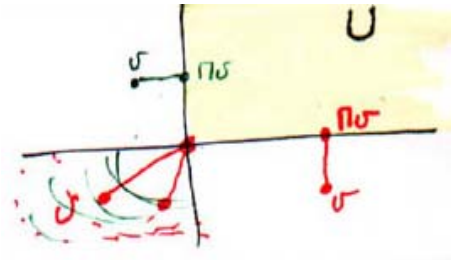
Property 6 *The application $\Pi_U : V \to U$ satisfies*

1. *$v - \Pi_U v = 0 \ \Leftrightarrow \ v \in U$*
2. *$\|\Pi_U v_1 - \Pi_U v_2\| \leqslant \|v_1 - v_2\|, \ \forall v_1, v_2 \in V$.*
   (Π_U is uniformly continuous, Lipschitz)



3. *Π_U linear $\Leftrightarrow$ U is a sub-vector space. Then statement 2 reads $(v - \Pi_U v) \perp u$, $\forall u \in U$.*

A typical constraint is $U = \{u \in \mathbb{R}^n \ ; \ u[i] \geqslant 0, \ \forall i\}$
Then U is not a vector subspace, Π_U is nonlinear



3.3 General methods

3.3.1 Gauss-Seidel method under Hyper-cube constraint

Consider the minimization of a coercive proper convex function $F : \mathbb{R}^n \to \mathbb{R}$ under a hyper-cube constraint:

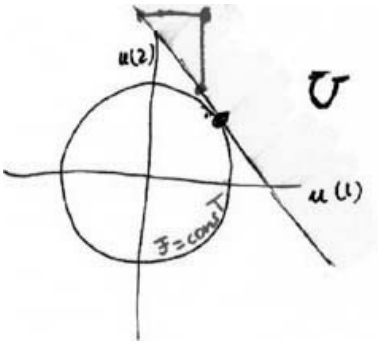
$$U = \{u \in \mathbb{R}^n : a_i \leqslant u[i] \leqslant b_i,\ 1 \leqslant i \leqslant n\} \ \text{ with } \ a_i \in [-\infty, \infty[,\ b_i \in]-\infty, \infty],\ 1 \leqslant i \leqslant n. \quad (3.3)$$

Iterations: $\forall k \in \mathbb{N},\ \ \forall i = 1, \ldots, n$

$$\begin{aligned} & F(u_{k+1}[1], \ldots, u_{k+1}[i-1], u_{k+1}[i], u_k[i+1], \ldots, u_k[n]) \\ = \inf_{a_i \leqslant \rho \leqslant b_i} \ & F(u_{k+1}[1], \ldots, u_{k+1}[i-1], \quad \rho, \quad u_k[i+1], \ldots, u_k[n]) \end{aligned}$$

Theorem 30 *If F is strongly convex and U is of the form (3.3), then (u_k) converges to the unique $\hat{u}$ such that $F(\hat{u}) = \min_{u \in U} F(u)$.*

Remark 13 The method cannot be extended to a general U. E.g., consider $F(u) = (u[1]^2 + u[2]^2)$ and $U = \{u \in \mathbb{R}^2 : u[1] + u[2] \geqslant 2\}$, and initialize with $u_0[1] \neq 1$ or $u_0[2] \neq 1$.



The algorithm is blocked at the boundary of U at a point different from the solution.

3.3.2 Gradient descent with projection and varying step-size

V—Hilbert space, $U \subset V$ convex, closed, non empty, $F : V \to \mathbb{R}$ is convex and differentiable in V. Here again, $\|u\| = \sqrt{\langle u, u \rangle}$, $\forall u \in V$.

Motivation:

$$\begin{aligned} & \hat{u} \in U \ \text{ and } \ F(\hat{u}) = \inf_{u \in U} F(u) \\ \Leftrightarrow \ & \hat{u} \in U \ \text{ and } \ \rho \langle \nabla F(\hat{u}), u - \hat{u} \rangle \geqslant 0,\ \forall u \in U,\ \rho > 0 \quad \text{(the NSC for a minimum)} \\ \Leftrightarrow \ & \hat{u} \in U \ \text{ and } \ \langle \hat{u} - \rho \nabla F(\hat{u}) - \hat{u}, u - \hat{u} \rangle \leqslant 0, \ \ \forall u \in U, \ \ \rho > 0 \\ \Leftrightarrow \ & \hat{u} = \Pi_U\big(\hat{u} - \rho \nabla F(\hat{u})\big),\ \rho > 0. \end{aligned}$$

In words, $\hat{u}$ is the fixed point of the application

$$G(u) = \Pi_U\big(u - \rho \nabla F(u)\big), \ \ \text{where } \ \Pi_U \ \text{ is the projection operator onto } U$$

Theorem 31 *Let $F : V \to \mathbb{R}$ be differentiable in V and $U \subset V$ a nonempty convex and closed subset. Suppose that $\exists \mu$ and $\exists M$ such that $0 < \mu < M$, and*

(i) $\langle \nabla F(u) - \nabla F(v),\ u - v \rangle \geqslant \mu \|u - v\|^2, \ \ \forall (u, v) \in V^2$ *(i.e. F is strongly convex);*

(ii) $\|\nabla F(u) - \nabla F(v)\| \leqslant M \|u - v\|, \ \ \forall (u, v) \in V^2$

For Π_U the projection operator onto U, consider

$$\begin{aligned} G_k(u) &\overset{\text{def}}{=} \Pi_U\big(u - \rho_k \nabla F(u)\big) \\ \text{where} \quad & \frac{\mu}{M^2} - \zeta \leqslant \rho_k \leqslant \frac{\mu}{M^2} + \zeta, \ \forall k \in \mathbb{N} \qquad (3.4) \\ \text{and} \quad & \zeta \in \left]0, \frac{\mu}{M^2}\right[\quad \text{(fixed)} \qquad (3.5) \\ u_{k+1} &= G_k(u_k), \quad \forall k \in \mathbb{N}. \qquad (3.6) \end{aligned}$$

Then $(u_k)_{k\in\mathbb{N}}$ in (3.6) converges to the unique minimizer $\hat{u}$ of F and

$$\|u_{k+1} - \hat{u}\| \leqslant \gamma^k \|u_0 - \hat{u}\|, \quad \text{where} \quad \gamma = \sqrt{\zeta^2 M^2 - \frac{\mu^2}{M^2} + 1} < 1. \qquad (3.7)$$

Remark 14 *If we fix $\zeta \gtrapprox 0$, γ is nearly optimal but the range for ρ_k is decreased; and vice-versa, for $\zeta \lessapprox \frac{\mu}{M^2}$, the range for ρ_k is nearly maximal $\left]0, \frac{2\mu}{M^2}\right[$ but $\gamma \lessapprox 1$.*

Ideally, one would wish the largest range for ρ_k and the least value for γ...

Proof. By its strongly convexity, F admits a unique minimizer. Let $u \in V$ and $v \in V$.

$$\begin{aligned} \|G_k(u) - G_k(v)\|^2 &= \|\Pi_U\big(u - \rho_k \nabla F(u)\big) - \Pi_U\big(v - \rho_k \nabla F(v)\big)\|^2 \\ &\leqslant \|u - v - \rho_k\big(\nabla F(u) - \nabla F(v)\big)\|^2 \\ &= \|u - v\|^2 - 2\rho_k \langle \nabla F(u) - \nabla F(v), u - v\rangle + \rho_k^2 \|\nabla F(u) - \nabla F(v)\|^2 \\ &\leqslant \|u - v\|^2 - 2\rho_k \mu \|u - v\|^2 + \rho_k^2 M^2 \|u - v\|^2 \\ &= (\rho_k^2 M^2 - 2\rho_k \mu + 1)\, \|u - v\|^2 \\ &= f(\rho_k)\, \|u - v\|^2, \end{aligned}$$

where f is convex and quadratic and reads as

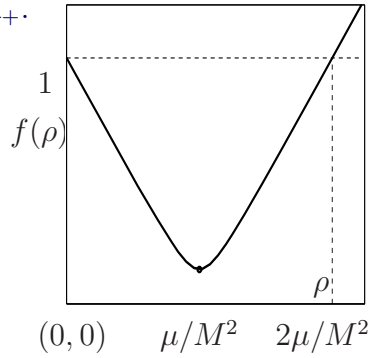
$$f(\rho) \overset{\text{def}}{=} \rho^2 M^2 - 2\rho\mu + 1.$$

Since $0 < \mu < M$, the discriminant of f is negative, $4\mu^2 - 4M^2 < 0$, hence

$$f(\rho) > 0, \ \forall \rho \geqslant 0.$$

Then $\sqrt{f(\rho)}$ is real and positive. It is easy to check that $\rho \to \sqrt{f(\rho)}$ is strictly convex on $\mathbb{R}_+$ when $0 < \mu < M$ and that it admits a unique minimizer on $\mathbb{R}_+$.
More precisely:

- $\sqrt{f(0)} = \sqrt{f\left(\frac{2\mu}{M^2}\right)} = 1$
- $\arg\min_\rho \sqrt{f(\rho)} = \frac{\mu}{M^2}$
- $0 < \sqrt{f\left(\frac{\mu}{M^2}\right)} < 1$



For any ζ as given in (3.5), we check that

$$\sqrt{f(\frac{\mu}{M^2}-\zeta)}=\sqrt{f(\frac{\mu}{M^2}+\zeta)}=\sqrt{\zeta^2M^2-\frac{\mu^2}{M^2}+1}=\gamma,$$

where the last equality comes from (3.7). By (3.5) yet again,

$$\underline{\zeta^2M^2-\frac{\mu^2}{M^2}+1<\frac{\mu^2}{M^2}}-\frac{\mu^2}{M^2}+1\leqslant\underline{1}\quad\Rightarrow\quad\gamma<1.$$

Hence for any ρ_k as specified by (3.4)-(3.5)

$$\sqrt{f(\rho_k)}\leqslant\gamma<1.$$

It follows that $\underline{G_k \text{ is a contraction, } \forall k\in\mathbb{N}}$ since $\|G_k(u)-G_k(v)\|\leqslant\gamma\|u-v\|$, $\gamma<1$, for any $(u,v)\in V^2$, hence $G_k(\hat{u})=\hat{u}$ for all $k\in\mathbb{N}$. We can write down

$$\|u_{k+1}-\hat{u}\|=\|G_k(u_k)-G_k(\hat{u})\|\leqslant\gamma\|u_k-\hat{u}\|\leqslant\cdots\leqslant\gamma^k\|u_0-\hat{u}\|$$

where $\gamma<1$ is given in (3.7). □

Remark 15 *If $U=V$ then $\Pi_U=\mathrm{Id}$.*

Remark 16 (About the Assumptions) If $\exists\nabla^2F$ then $\nabla^2F\succ0$ because F is strongly convex. Then we have $M=\lambda_{\max}(\nabla^2F)\geqslant\mu=\lambda_{\min}(\nabla^2F)>0$. Note that most usually, $M>\mu$.

Taking into account the structure of the problem can increase the convergence speed.

Constrained quadratic strictly convex problem: $F(u)=\langle Bu,u\rangle-\langle c,u\rangle$, $u\in\mathbb{R}^n$, $B=B^T$, $B\succ0$ (hence $\lambda_{\min}(B)>0$), $U\subset\mathbb{R}^n$ nonempty, closed and convex. Using that $\nabla F(u)=Bu-c$, iterations are

$$u_{k+1}=\Pi_U\big(u_k-\rho_k(Bu_k-c)\big),\quad k\in\mathbb{N}.$$

Since $\nabla^2F(u)=B,\ \forall u\in\mathbb{R}^n$, Theorem 31 ensures convergence if

$$\frac{\lambda_{\min}(B)}{\big(\lambda_{\max}(B)\big)^2}-\zeta\leqslant\rho_k\leqslant\frac{\lambda_{\min}(B)}{\big(\lambda_{\max}(B)\big)^2}+\zeta,\quad\zeta\in\left]0,\frac{\lambda_{\min}(B)}{\big(\lambda_{\max}(B)\big)^2}\right[\tag{3.8}$$

Drawback: $\lambda_{\min}(B)/\big(\lambda_{\max}(B)\big)^2$ can be very-very small.

For any $u,v\in\mathbb{R}^n$,

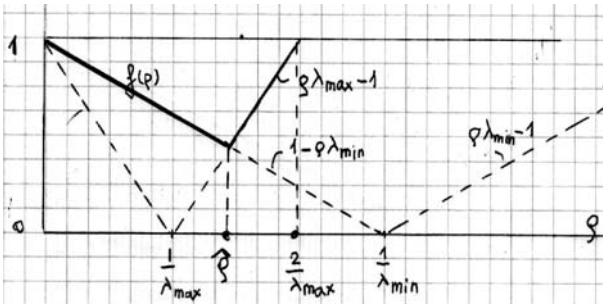
$$\big\|\Pi_U\big(u-\rho_k(Bu-c)\big)-\Pi_U\big(v-\rho_k(Bv-c)\big)\big\|_2\leqslant\|(u-v)-\rho_kB(u-v)\|_2\leqslant\|\mathrm{Id}-\rho_kB\|_2\ \|u-v\|_2.$$

Remind that for any square matrix B, if $Bv=\lambda_i(B)v$ then $(\mathrm{Id}-B)v=v-\lambda_i(B)v=(1-\lambda_i(B))v$. Since $\mathrm{Id}-\rho B$, $\rho>0$, is symmetric, its spectral radius is

$$f(\rho)=\max_{1\leqslant i\leqslant n}\big|\lambda_i(\mathrm{Id}-\rho B)\big|=\max\Big\{\big|1-\rho\lambda_{\min}(B)\big|\,,\,\big|1-\rho\lambda_{\max}(B)\big|\Big\}$$

We have f convex and

- $f(\rho) > 0$ if $\lambda_{\min}(B) < \lambda_{\max}(B)$;
- $\arg\min_{\rho} f(\rho) = \dfrac{2}{\lambda_{\min}(B) + \lambda_{\max}(B)} < 1$
- $f(0) = f\left(\dfrac{2}{\lambda_{\max}(B)}\right) = 1$



Convergence is ensured if $0 < \rho_k < \dfrac{2}{\lambda_{\max}(B)}$. This bound for ρ_k is much better than (3.8) that was established in Theorem 31 in a more general case.

Remark 17 *Difficulty : in general Π_U has no an explicit form. In such a case we can use a penalization method, or another iterative method (see later sections).*

3.3.3 Penalty (barrier) methods

Main idea : Replace the constrained minimization of $F : \mathbb{R}^n \to \mathbb{R}$ by an unconstrained problem.

Construct $\mathcal{G} : \mathbb{R}^n \to \mathbb{R}$ continuous, convex, $\mathcal{G}(u) \geqslant 0,\ \forall u \in \mathbb{R}^n$ and such that

$$\mathcal{G}(u) = 0 \ \Leftrightarrow\ u \in U \tag{3.9}$$

$\forall \omega > 0$ define

$$(P_\omega) \qquad\qquad \mathcal{F}_\omega(u) = F(u) + \omega \mathcal{G}(u)$$

We will consider that $\omega \to +\infty$.

Theorem 32 *Let F be continuous, coercive and strictly convex, and U convex, defined by* (3.9)*. Then*

(1) $\forall \omega > 0$, $\exists u_\omega$ unique such that $\mathcal{F}_\omega(u_\omega) = \inf_{u \in \mathbb{R}^n} \mathcal{F}_\omega(u)$;

(2) $\lim_{\omega \to +\infty} u_\omega = \hat{u}$ where $\hat{u}$ is the unique solution of $F(\hat{u}) = \inf_{u \in U} F(u)$.

The proof can be found e.g. in [50, p.205]. The idea is very intuitive, not always good.

Convex programming Problem : U given by (1.2), i.e.

$$U = \{u \in \mathbb{R}^n \ \mid\ h_i(u) \leqslant 0,\ 1 \leqslant i \leqslant q\}$$

with $h_i : \mathbb{R}^n \to \mathbb{R}, i = 1, \ldots, q$ convex functions. Consider

$$\begin{aligned} \mathcal{G}(u) &= \sum_{i=1}^{q} g_i(u), \\ g_i(u) &= \max\{h_i(u), 0\} \end{aligned}$$

G clearly satisfies the requirements for continuity, convexity and (3.9). Let us check the last point (3.9). If for all $1 \leqslant i \leqslant q$ we have $h_i(u) \leqslant 0$, i.e. $u \in U$, then $g_i(u) = 0$ for all $1 \leqslant i \leqslant q$ and thus $G(u) = 0$. Note that $G(u) \geqslant 0$ for all $u \in \mathbb{R}^n$. If $u \notin U$, there is at least one index i such that $h_i(u) > 0$ which leads to $g_i(u) > 0$ and $G(u) > 0$. Thus $G(u) = 0$ shows that $u \in U$.

Comments Note that $\mathcal{G}$ is not necessarily differentiable.

Difficulty : construct "good" functions $\mathcal{G}$ (differentiable, convex). This constitutes the main limitation of penalization methods.

3.4 Equality constraints

The main idea is to get rid off the constraints.

3.4.1 Lagrange multipliers

Here V_1, V_2 and Y are n.v.s. Optimality conditions are based on the Implicit functions theorem.

Theorem 33 (Implicit functions theorem) *Let $\mathcal{G} : \mathcal{O} \subset V_1 \times V_2 \to Y$ be $\mathcal{C}^1$ in $\mathcal{O}$, where V_2 and Y are* complete *(i.e. Banach spaces). Suppose that $\hat{u} = (\hat{u}_1, \hat{u}_2) \in \mathcal{O}$ and $v \in Y$ are such that $\mathcal{G}(\hat{u}_1, \hat{u}_2) = v$ and $D_2\mathcal{G}(\hat{u}_1, \hat{u}_2) \in \mathrm{Isom}(V_2; Y)$.*

Then $\exists \mathcal{O}_1 \subset V_1$, $\exists \mathcal{O}_2 \subset V_2$ such that $(\hat{u}_1, \hat{u}_2) \in \mathcal{O}_1 \times \mathcal{O}_2 \subset \mathcal{O}$ and there is a unique continuous function $g : \mathcal{O}_1 \subset V_1 \to V_2$, called an implicit function*, such that*

$$\{(u_1, u_2) \in \mathcal{O}_1 \times \mathcal{O}_2 : \mathcal{G}(u_1, u_2) = v\} = \{(u_1, u_2) \in \mathcal{O}_1 \times V_2 : u_2 = g(u_1)\}.$$

Moreover g is differentiable at $\hat{u}_1$ and

$$Dg(\hat{u}_1) = -\left(D_2\mathcal{G}(\hat{u}_1, \hat{u}_2)\right)^{-1} D_1\mathcal{G}(\hat{u}_1, \hat{u}_2). \tag{3.10}$$

Proof. See [21, p. 30], [20, p. 176]

Let us prove (3.10). Note that g is differentiable at $\hat{u}_1$:

$$\mathcal{G}(u_1, g(u_1)) - v = 0 \in Y, \quad \forall u_1 \in \mathcal{O}_1.$$

Since $\mathcal{O}_1$ is open, differentiation of both sides of this identity w.r.t. $u_1 = \hat{u}_1$ yields

$$D_1\mathcal{G}(\hat{u}_1, g(\hat{u}_1)) + D_2\mathcal{G}(\hat{u}_1, g(\hat{u}_1))Dg(\hat{u}_1) = 0.$$

Example 3 *$V_1 = V_2 = \mathbb{R}$, $G(u_1, u_2) = u_1^2 - u_2 = 0 = v \in Y = \mathbb{R}$. Then $D_2G(u) = -1 \in \mathrm{Isom}(\mathbb{R}, \mathbb{R})$ and $u_2 = g(u_1) = u_1^2$. Thus $G(u_1, u_2) = 0 = G(u_1, g(u_1))$, $\forall u_1 \in \mathbb{R}$.*

Theorem 34 (Necessary condition for a constrained minimum) *Let $\mathcal{O} \subset V = V_1 \times V_2$ be open where V_1 and V_2 are real n.v.s., and V_2 be complete. Consider that $\mathcal{G} : \mathcal{O} \to V_2$ is $\mathcal{C}^1$ and set*

$$U = \{u \in \mathcal{O} : \mathcal{G}(u_1, u_2) = 0\}.$$

Suppose that $F : V \to \mathbb{R}$ is differentiable at $\hat{u} \in U$ and that $D_2\mathcal{G}(\hat{u}) \in \mathrm{Isom}(V_2, V_2)$. If F has a relative minimum w.r.t. U at $\hat{u}$, then there is an application $\lambda(\hat{u}) \in \mathcal{L}(V_2, \mathbb{R})$ such that

$$DF(\hat{u}) + \lambda(\hat{u})D\mathcal{G}(\hat{u}) = 0. \tag{3.11}$$

Proof. By the Implicit Functions Theorem 33, there are two open sets $\mathcal{O}_1 \in V_1$ and $\mathcal{O}_2 \in V_2$, with $\hat{u} \in \mathcal{O}_1 \times \mathcal{O}_2 \subset \mathcal{O}$, and a continuous application $g : \mathcal{O}_1 \to \mathcal{O}_2$ such that

$$(\mathcal{O}_1 \times \mathcal{O}_2) \cap U = \{(u_1, u_2) \in (\mathcal{O}_1 \times V_2) \ \mid \ u_2 = g(u_1)\}$$

and

$$Dg(\hat{u}_1) = -(D_2\mathcal{G}(\hat{u}))^{-1} D_1\mathcal{G}(\hat{u}). \tag{3.12}$$

Define

$$\mathcal{F}(u_1) \stackrel{\text{def}}{=} F\big(u_1, g(u_1)\big), \quad \forall u_1 \in \mathcal{O}_1.$$

Then $\mathcal{F}(\hat{u}_1) = \inf_{u_1 \in \mathcal{O}_1} \mathcal{F}(u_1)$ entails $F(\hat{u}) = \inf_{u \in U} F(u)$ for $\hat{u} = \big(\hat{u}_1, g(\hat{u}_1)\big)$. Since $\mathcal{O}_1$ is open, $\hat{u}_1$ satisfies

$$\begin{aligned} 0 &= D\mathcal{F}(\hat{u}_1) = D_1F\big(\hat{u}_1, g(\hat{u}_1)\big) + D_2F\big(\hat{u}_1, g(\hat{u}_1)\big)\, \underline{Dg(\hat{u}_1)} \\ &= D_1F\big(\hat{u}_1, g(\hat{u}_1)\big) - D_2F\big(\hat{u}_1, g(\hat{u}_1)\big)\, (D_2\mathcal{G}(\hat{u}))^{-1} D_1\mathcal{G}(\hat{u}), \end{aligned}$$

where the last equality comes from (3.12). Using that $\hat{u} = \big(\hat{u}_1, g(\hat{u}_1)\big)$, we can hence write down

$$\begin{aligned} D_1F(\hat{u}) &= D_2F(\hat{u})\, (D_2\mathcal{G}(\hat{u}))^{-1} D_1\mathcal{G}(\hat{u}) \\ D_2F(\hat{u}) &= D_2F(\hat{u})\, (D_2\mathcal{G}(\hat{u}))^{-1} D_2\mathcal{G}(\hat{u}) \end{aligned}$$

where the second equality is an obvious identity. We have

$$DF(\hat{u}) + \lambda(\hat{u}) D\mathcal{G}(\hat{u}) = 0$$

as claimed in (3.11) by setting

$$\lambda(\hat{u}) = -D_2F(\hat{u})\, (D_2\mathcal{G}(\hat{u}))^{-1}.$$

Since $D_2F(\hat{u}) \in \mathcal{L}(V_2\,; \mathbb{R})$ and $(D_2\mathcal{G}(\hat{u}))^{-1} \in \mathcal{L}(V_2\,; V_2)$, we have $\lambda(\hat{u}) \in \mathcal{L}(V_2\,; \mathbb{R})$. □

The most usual case arising in practice is considered next.

Theorem 35 *Let $\mathcal{O} \subset \mathbb{R}^n$ be open and $g_i : \mathcal{O} \to \mathbb{R}$, $1 \leqslant i \leqslant p$ be $\mathcal{C}^1$-functions in $\mathcal{O}$.*

$$U = \{u \in \mathcal{O} : g_i(u) = 0, 1 \leqslant i \leqslant p\} \subset \mathcal{O}$$

Let $Dg_i(\hat{u}) \in \mathcal{L}(\mathbb{R}^n, \mathbb{R})$, $1 \leqslant i \leqslant p$ be linearly independent and $F : \mathcal{O} \to \mathbb{R}$ differentiable at $\hat{u}$. If $\hat{u} \in U$ solves the problem $\inf_{u \in U} F(u)$ then there exist p real numbers $\lambda_i(\hat{u}) \in \mathbb{R}$, $1 \leqslant i \leqslant p$, uniquely defined, such that

$$DF(\hat{u}) + \sum_{i=1}^{p} \lambda_i(\hat{u})\ Dg_i(\hat{u}) = 0. \tag{3.13}$$

$\lambda_i(\hat{u}),\ 1 \leqslant i \leqslant p$ are called Lagrange multipliers. The constraint set U as also known as feasible set.

The problem considered in the last theorem is a particular case of Theorem 34.

Proof. Put $\mathcal{G} \stackrel{\text{def}}{=} (g_1, \cdots, g_p)$. Then

$$D\mathcal{G}(\hat{u}) = \begin{bmatrix} Dg_1(\hat{u}) \\ \cdots \\ Dg_p(\hat{u}) \end{bmatrix} = \begin{bmatrix} \frac{\partial g_1}{\partial \hat{u}_1} & \cdots & \frac{\partial g_1}{\partial \hat{u}_p} & \cdots & \frac{\partial g_1}{\partial \hat{u}_n} \\ \cdots & & & & \\ \frac{\partial g_p}{\partial \hat{u}_1} & \cdots & \frac{\partial g_p}{\partial \hat{u}_p} & \cdots & \frac{\partial g_p}{\partial \hat{u}_n} \end{bmatrix}$$

Since $Dg_i(\hat{u})$ are linearly independent, $\text{rank}(D\mathcal{G}(\hat{u})) = p \leqslant n$, so we can assume that the first $p \times p$ submatrix of $D\mathcal{G}(\hat{u})$ is invertible. If $p = n$, $\hat{u}$ is uniquely determined by $\mathcal{G}$. Consider next that $p < n$. Let $\{e_1, \cdots e_n\}$ be the canonical basis of $\mathbb{R}^n$. Define

$$\begin{aligned} V_1 &\stackrel{\text{def}}{=} \text{span}\{e_{p+1}, \cdots, e_n\} \\ V_2 &\stackrel{\text{def}}{=} \text{span}\{e_1, \cdots, e_p\} \end{aligned}$$

Redefine $\mathcal{G}$ in the following way:

$$\begin{aligned} \mathcal{G} : V_1 \times V_2 &\to V_2 \\ (u_1, u_2) &\to \mathcal{G}(u) = \sum_{i=1}^{p} g_i(u) e_i. \end{aligned}$$

Since the first $p \times p$ submatrix of $D\mathcal{G}(\hat{u})$ is invertible, $D_2\mathcal{G}(\hat{u}) \in \text{Isom}(V_2, V_2)$. Noticing that the elements of V_2 belong to $\mathbb{R}^p$, Theorem 34 shows that there exists $\lambda(\hat{u}) \in \mathbb{R}^p$ (i.e. real numbers $\lambda_i(\hat{u})$, $1 \leqslant i \leqslant p$) such that

$$DF(\hat{u}) + \lambda(\hat{u}) D\mathcal{G}(u) = 0 \quad \Leftrightarrow \quad DF(\hat{u}) + \sum_{i=1}^{p} \lambda_i(\hat{u}) Dg_i(\hat{u}) = 0$$

The uniqueness of $\lambda(\hat{u})$ is due to the fact that $\text{rank}(D\mathcal{G}(\hat{u})) = p$. □

Remark 18 *Since $F : \mathcal{O} \in \mathbb{R}^n \to \mathbb{R}$ and $g_i : \mathcal{O} \in \mathbb{R}^n \to \mathbb{R}$, $1 \leqslant i \leqslant p$, we can identify differentials with gradients using the scalar product on $\mathbb{R}^\ell$, $\ell \in \{p, n\}$. By introducing the Lagrangian function*

$$L(u, \lambda) := F(u) + \sum_{i=1}^{p} \lambda_i \, g_i(u)$$

(3.13) *can be rewritten as $\nabla_u L(u, \lambda) = 0$, i.e.,*

$$\nabla_u L(u, \lambda) = \nabla F(\hat{u}) + \sum_{i=1}^{p} \lambda_i \, \nabla g_i(u) = 0 \quad n \text{ equations} \tag{3.14}$$

The numbers λ_i, $i \in \{1, 2, \cdots, p\}$ are called Lagrange multipliers. *They obey $\nabla_\lambda L(u, \lambda) = 0$, i.e.,*

$$\nabla_{\lambda_i} L(u, \lambda) = g_i(u) = 0, \quad 1 \leqslant i \leqslant p \qquad p \text{ equations} \tag{3.15}$$

We have $n + p$ unknowns, $\hat{u} \in \mathbb{R}^n$ and $\lambda \in \mathbb{R}^p$, and we have a system of $n + p$ (nonlinear) equations. (3.14) *an* (3.15) *are known as the Karush-Kuhn-Tucker (KKT) conditions for equality constraints. They have been found by Louis Lagrange in 18th century.*

These are necessary conditions. Complete the resolution of the problem by analyzing if $\hat{u}$ indeed is a minimizer. In particular, if F is convex and continuous, and if U is convex and $\{\nabla g_i\}$ linearly independent on V, then (3.14)-(3.15) yields the minimum of F.

Example 4 *$F(u) = u_1 + u_2$ and $U = \{u \in \mathbb{R}^2 \mid h(u) = u_1^2 + u_2^2 - 2 = 0\}$. Then $\hat{u} = (-1, -1)$.*

3.4.2 Application to linear systems

Quadratic function under affine constraints

For $B \in \mathbb{R}^{n\times n}$, $B = B^T$, $B \succ 0$, and $c \in \mathbb{R}^n$, consider

$$\begin{array}{rcl} \text{minimize} \quad F(u) & = & \frac{1}{2}\langle Bu, u\rangle - \langle c, u\rangle \quad \text{subject to} \;\; u \in U \\ \text{where} \;\; U & = & \{u \in \mathbb{R}^n : Au = v\} \quad \text{with} \quad A \in \mathbb{R}^{p\times n},\; \text{rank}A = p < n. \end{array} \tag{3.16}$$

(3.14) yields

$$\begin{array}{rcl} Bu + A^T\lambda & = & c \\ Au & = & v \end{array} \tag{3.17}$$

Projection onto U in (3.16) The projection $\hat{u} = \Pi_U(w)$ of $w \in (\mathbb{R}^n \setminus U)$ onto U satisfies

$$\hat{u} = \arg\min_{u\in U} \frac{1}{2}\|u - w\|^2 = \arg\min_{u\in U} \left(\frac{1}{2}\langle u, u\rangle - \langle w, u\rangle\right).$$

(3.17) yields

$$\begin{array}{rcl} u + A^T\lambda & = & w \\ Au & = & v \end{array}$$

Using that AA^T is invertible (see (3.16))

$$Au + AA^T\lambda = Aw \quad \Rightarrow \quad \lambda = (AA^T)^{-1}A(w - u) = (AA^T)^{-1}(Aw - v)$$

Then

$$u + A^T\lambda = u + A^T(AA^T)^{-1}(Aw - v) = w$$

We deduce

$$\hat{u} = \Pi_U(w) = \left(I - A^T(AA^T)^{-1}A\right)w + A^T(AA^T)^{-1}v \tag{3.18}$$

- If $p = 1$, then $A \in \mathbb{R}^{1\times n}$ is a row-vector, $v \in \mathbb{R}$ and $AA^T = \|A\|_2^2$.
- If $v = 0$, then $U = \ker A$ is a vector sub-space and Π_U is linear

$$\Pi_U = I - A^T(AA^T)^{-1}A \tag{3.19}$$

"Solving" a linear system: $Au = v \in \mathbb{R}^m$ where $u \in \mathbb{R}^n$

- $\text{rank}A = m \leqslant n$: $F(u) = \|u\|^2$, $U = \{u \in \mathbb{R}^n : Au = v\}$
  $\hat{u} = A^T(AA^T)^{-1}v$ (the minimum norm solution). Compare with (3.19).
- $\text{rank}A = n \leqslant m$: $F(u) = \|Au - v\|^2$, $U = \mathbb{R}^n$
  $\hat{u} = (A^TA)^{-1}A^Tv$ (the least-squares (LS) solution)

These solutions are usually unstable—remind the LS solution on p. 8 and Example **??**.

Generalized inverse, pseudoinverse — if A has linearly dependent rows or columns

SVD (singular value decomposition) of $A \in \mathbb{R}^{m\times n}$:

$$A = Q_1 \Sigma Q_2$$

- $Q_1 Q_1^T = Q_1^T Q_1 = I_m, \quad Q_2 Q_2^T = Q_2^T Q_2 = I_n$ (orthonormal matrices)
- the columns of Q_1 = eigenvectors of AA^T
- the columns of Q_2 = eigenvectors of de $A^T A$
- Σ, $m \times n$, diagonal, $\Sigma[i,i],\ 1 \leqslant i \leqslant r$ singular values $= \sqrt{\text{eigenvalues} \neq 0}$ of AA^T and $A^T A$, $r = \text{rank} A$

The pseudo-inverse $A^\dagger = Q_2^T \Sigma^\dagger Q_1^T$ where $\Sigma^\dagger[i,j] = \begin{cases} \dfrac{1}{\Sigma[i,i]} & \text{if } i = j \text{ and } 1 \leqslant i \leqslant r \\ 0 & \text{else} \end{cases}$

It can be shown that this $A^\dagger$ corresponds to

$$U = \left\{ u \in \mathbb{R}^n : \|Au - v\| = \inf_{w\in\mathbb{R}^n} \|Aw - v\| \right\}$$

$$\hat{u} \ : \ \|\hat{u}\| = \inf_{u\in U} \|u\| \quad \Rightarrow \quad \hat{u} = A^\dagger v$$

See [26, 51].

The SVD plays a key role in many fields – tool for analysis and computation..

Remark 19 *Generalized inverse—in a similar way when $A : V_1 \to V_2$ is compact and V_1, V_2 are Hilbert spaces. (see e.g. [51].)*

3.4.3 Inexact quadratic penalty for equality constraints

Consider the constraint minimisation problem

$$\text{minimize} \ \ F(u) \quad \text{subject to} \quad u \in U := \{u \in \mathcal{O} : g_i(u) = 0, \ \ 1 \leqslant i \leqslant p\}\,. \tag{3.20}$$

Following the penalty approach, one choose $\mathcal{G}$ in (3.9)

$$\mathcal{G}(u) = \sum_{i=1}^{p} g_i^2(u). \tag{3.21}$$

By the penalty approach (subsection 3.3.3) and Theorem 32, one could track the minimizer of

$$\mathcal{F}_\omega(u) := F(u) + \frac{\omega}{2} \sum_{i=1}^{p} g_i^2(u) \quad \text{for} \quad \omega \to \infty \tag{3.22}$$

A more subtle approach, linking penalty and Lagrangian multipliers, is stated next.

Theorem 36 *Let $\{\tau_k\}$ be a sequence of tolerance parameters satisfying $\tau_k \to 0$ and let $\{\omega_k\} \to +\infty$. For any k, find an approximate minimizer u_k of $\mathcal{F}_{\omega_k}(\cdot)$ satisfying $\|\nabla_u \mathcal{F}_{\omega_k}(u_k)\| \leqslant \tau_k$. Let $\hat{u}$ be a limit point of the sequence $\{u_k\}$. Then $\hat{u}$ is a stationary point of $\mathcal{G}$ in* (3.21)*. Furthermore, if $\{\nabla g_i(\hat{u})\}_{i=1}^p$*

are linearly independent, then $\hat{u} \in U$ *and* $\hat{u}$ *is a KKT point for problem* (3.20) *where for any infinite subsequence* k_j *such that* $\lim_{k_j \to \infty} u_{k_j} = \hat{u}$*, the vector* λ *given by*

$$\lim_{k_j \to \infty} \omega_{k_j} g_i(u_{k_j}) = \lambda_i \quad 1 \leqslant i \leqslant p \tag{3.23}$$

satisfies the KKT conditions (3.14)-(3.15) *for the constrained problem* (3.20).

Proof. By differentiating $\mathcal{F}_{\omega_k}$ in (3.22) we obtain

$$\nabla \mathcal{F}_{\omega_k}(u_k) = \nabla F(u_k) + \omega_k \sum_{i=1}^{p} g_i(u_k) \nabla g_i(u_k), \tag{3.24}$$

so from the termination criterion given by τ_k, we have that

$$\|\nabla \mathcal{F}_{\omega_k}(u_k)\| = \left\| \nabla F(u_k) + \omega_k \sum_{i=1}^{p} g_i(u_k) \nabla g_i(u_k) \right\| \leqslant \tau_k \tag{3.25}$$

By rearranging this expression and using the inequality $\|a\| - \|b\| \leqslant \|a + b\|$, we obtain

$$\left\| \sum_{i=1}^{p} g_i(u_k) \nabla g_i(u_k) \right\| \leqslant \frac{1}{\omega_k} \left(\tau_k + \|\nabla F(u_k)\| \right)$$

Let $\hat{u}$ be a limit point of the sequence of iterates. Then there is a subsequence k_j such that $\lim_{k_j \to \infty} u_{k_j} = \hat{u}$. Taking such a limit, the right-hand-side approaches zero. Thus,

$$\sum_{i=1}^{p} g_i(\hat{u}) \nabla g_i(\hat{u}) = 0. \tag{3.26}$$

Hence $\hat{u}$ is a stationary point of the function $\mathcal{G}$ in (3.21).

Consider next that the constraint gradients $\{\nabla g_i(\hat{u})\}_{i=1}^{p}$ are linearly independent at $\hat{u}$. Then (3.26) shows that $g_i(\hat{u}) = 0$, $1 \leqslant k \leqslant p$, i.e., $\hat{u} \in U$. Hence the second KKT condition in (3.15) is satisfied. We want to check the first KKT condition (3.14) as well, and to prove the limit (3.23).

Let $G(u)$ denote the matrix of constraint gradients (the Jacobian), that is

$$G(u)^T = [\nabla g_i(u)]_{i=1}^{p}$$

and let λ^k denote the vector $[\omega_k g_i(u_k)]_{i=1}^{p}$. By (3.24) and (3.25) one has

$$G(u_k)^T \lambda^k = \nabla \mathcal{F}_{\omega_k}(u_k) - \nabla F(u_k), \quad \|\nabla \mathcal{F}_{\omega_k}(u_k)\| \leqslant \tau_k. \tag{3.27}$$

Taking the limits for a subsequence k_j with $\lim_{k_j \to \infty} u_{k_j} = \hat{u}$ we conclude that

$$\nabla F(\hat{u}) + G(\hat{u})^T \lambda = 0$$

so that λ as defined in the theorem satisfies also the first KKT condition (3.14). □

The additional condition $\tau_k \to 0$ for inexact solving the intermediate steps of the penalty method improves the convergence. The quantities $\omega_k g_i(u_k)$ are estimates at iteration k of the Lagrange parameter λ. This fact underlines the "Augmented Lagrangian Methods" (ALM) considered next.

3.4.4 Augmented Lagrangian method

These are a class of algorithms for solving constrained optimization problems. An additional term is designed to mimic a Lagrange multiplier based on Theorem 36. The main advantage is that penalization does not need to go to infinity and that thus, ill-conditioning is avoided.

Let λ denote the true Lagrange parameter. From Theorem 36 we have $g_i(u_k) \approx -\frac{1}{\omega_k}(\lambda_i - \lambda_i^k), \forall\, i$. The augmented Lagrangian function L_A includes this explicit estimate of the Lagrange multipliers λ in the objective.

$$L_A(u, \lambda; \omega) = F(u) + \sum_{i=1}^{p} \lambda_i g_i(u) + \frac{\omega}{2} \sum_{i=1}^{p} \big(g_i(u)\big)^2 \tag{3.28}$$

L_A is a combination of the Lagrangian function and the quadratic penalty function. The algorithms fixes the penalty parameter $\omega_k > 0$ and at the kth iteration fixes λ at the current iterate λ^k and performs minimization of $L_A(u, \lambda)$ with respect to u.

Augmented Lagrangian Algorithm – Equality Constraints [10]

Given $\omega_0 > 0$, $\tau > 0$, starting points λ^0 and u_0
for $k = 0, 1, 2, \ldots$

- Find an approximate minimizer u_k of $L_A(u, \lambda^k; \omega_k)$ starting at u_k^S
  and terminate when $\|\nabla_u L_A(u, \lambda^k; \omega_k)\| \leqslant \tau_k$;
- If convergence is satisfied - stop;
- Otherwise
  Update $\lambda^{k+1} = \lambda^k + \omega_k g(u_k)$;
  Choose $\omega_{k+1} \geqslant \omega_k$;
  Set starting point for the next iteration to $u_{k+1}^S = u_k$;
  Select tolerance τ_{k+1}

end (for).

Theorem 37 *([10, p. 517]) Let $\hat{u}$ be a minimizer of (3.20) such that $\{\nabla g_i(\hat{u})\}_{i=1}^p$ are linearly independent. Then there is a threshold value $\bar{\omega}$ such that for all $\omega > \bar{\omega}$ $\hat{u}$ is a strict local minimizer of $L_A(u, \lambda; \omega)$.*

Convergence of ALM can be assured without increasing ε indefinitely. Ill conditioning is therefore less of a problem than in penalty methods.

3.5 Inequality constraints

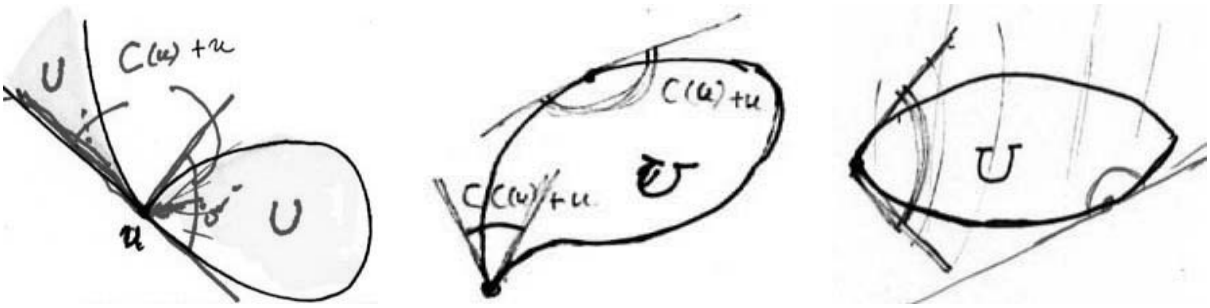
Set of constraints: $U \subset V$, $U \neq \varnothing$, where V is a real n.v.s.

3.5.1 Abstract optimality conditions

Definition 17 *Let V be a n.v.s., $U \subset V$, $\neq \varnothing$. The cone of all* feasible directions *at $u \in U$ reads*

$$C(u) = \{0\} \cup \left\{ v \in V \setminus \{0\} : \exists (u_k)_{k \geqslant 0}, u \neq u_k \in U, \forall k \geqslant 0,\ \lim_{k\to\infty} u_k = u,\ \lim_{k\to\infty} \frac{u_k - u}{\|u_k - u\|} = \frac{v}{\|v\|} \right\}$$

$C(u)$ is a cone with vertex 0, not necessarily convex. $C(u)$ is the closure of the set of all directions such that starting from u, we can reach another point $v \in U$.



Lemma 3 ([5]) *$C(u)$ is closed, $\forall u \in U$.*

Lemma 4 *If U is convex then $U \subset u + C(u)$, for all $u \in U$.*

Theorem 38 (Necessary condition for constrained minimum) *V real n.v.s., $U \subset V$, $U \neq \varnothing$ (arbitrary) and $F : \mathcal{O}(U) \to \mathbb{R}$ differentiable at $\hat{u} \in U$. If F admits at $\hat{u} \in U$ a constrained minimum, then*

$$\langle \nabla F(\hat{u}),\ u - \hat{u} \rangle \geqslant 0, \quad \forall u \in \{\hat{u} + C(\hat{u})\}.$$

Proof. —see e.g. [5].

From Theorem 28, p. 47, if U is convex, then $\langle \nabla F(\hat{u}), u - \hat{u} \rangle \geqslant 0,\ \forall u \in U$.

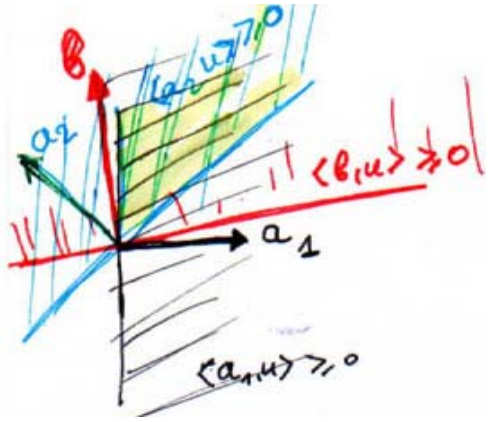
3.5.2 Farkas-Minkowski (F-M) theorem

Theorem 39 *Let V be a Hilbert space, I—a finite set of indexes, $a_i \in V,\ \forall i \in I$ and $b \in V$. Then*

1. $\{u \in V : \langle a_i, u \rangle \geqslant 0, \forall i \in I\} \subset \{u \in V : \langle b, u \rangle \geqslant 0\}$

if and only if

2. $\exists \lambda_i \geqslant 0, \forall i \in I \ \mid \ b = \sum_{i \in I} \lambda_i a_i$



If $\{a_i, i \in I\}$ are linearly independent,
then $\{\lambda_i, i \in I\}$ are determined in a unique way.

The proof is given in Appendix 7.8 on p. 117

Remark 20 *Link with : $g_1, \ldots, g_p$ linearly independent and $[\langle g_i, u \rangle = 0, \forall i \Rightarrow \langle f, u \rangle = 0]$ then $\exists \lambda_1, \ldots, \lambda_p$ such that $f = \sum_{i=1}^{p} \lambda_i g_i$.*

3.5.3 Constraint qualification

$\boxed{U = \{u \in \mathcal{O} : h_i(u) \leqslant 0,\ 1 \leqslant i \leqslant q\}}$ where $\mathcal{O} \subset V$ (open), V is a n.v.s., $h_i : V \to \mathbb{R}$, $1 \leqslant i \leqslant q$, $q \in \mathbb{N}$ (finite). We always assume that $U \neq \varnothing$.

How to describe $C(u)$ in an easier way?

Definition 18 *The set of* active constraints *at $u \in U$ is defined by:*

$$\boxed{I(u) = \{1 \leqslant i \leqslant q : h_i(u) = 0\}}$$

Since $h_i(u) \leqslant 0$, $\forall u \in U$, the figures on p. 59 (below Definition 17) suggest that in some cases, if $i \in I(u)$, then h_i reaches at u its maximum w.r.t. U, hence $\langle \nabla h_i(u), v - u \rangle \leqslant 0$ for some $v \in U$. (We say "some" because U can be nonconvex.) This observation underlies the definition of Q:

$$Q(u) = \{v \in V : \langle \nabla h_i(u), v \rangle \leqslant 0,\ \forall i \in I(u)\} \qquad \text{(convex cone)} \tag{3.29}$$

This cone is much more practical that C, even though it still depends on u. The same figures show that in some cases $Q(u) = C(u)$. Definition 19 below is constructed in such a way so that these cones are equal in the most important cases.

Definition 19 *The constraints are <u>qualified</u> at $u \in U$ if <u>one</u> of the following conditions hold:*

- $\exists w \in V \setminus \{0\}$ *such that* $\forall i \in I(u)$, $\langle \nabla h_i(u), w \rangle \leqslant 0$, *where the inequality is <u>strict</u> if h_i is not affine;*
- *h_i is affine, $\forall i \in I(u)$.*

Naturally, $Q(u) = V$ if $I(u) = \varnothing$.

Lemma 5 ([5]) *Let h_i for $i \in I(u)$ be differentiable at u* $\quad \Rightarrow \quad$ $C(u) \subset Q(u)$

Theorem 40 ([5]) *Assume that*

(i) *h_i differentiable at u, $\forall i \in I(u)$*
(ii) *h_i continuous at u, $\forall i \in \{1, \cdots, q\} \setminus I(u)$* $\quad \Rightarrow \quad$ $\boxed{C(u) = Q(u)}$
(iii) *Constraints are qualified at u*

Example 5 $U = \{v \in \mathbb{R}^n : \langle a_i, v \rangle \leqslant b_i, 1 \leqslant i \leqslant q\} \neq \varnothing$ (then the constraints are qualified $\forall v \in U$) and we have $C(u) = Q(u) = \{v \in \mathbb{R}^n : \langle a_i, v \rangle \leqslant 0, \forall i \in I(u)\}$

3.5.4 Kuhn & Tucker Relations

Putting together all previous results leads to one of the most important statements in Optimization theorey—the Kuhn-Tucker (KT) relations, stated below.

Theorem 41 (Necessary conditions for a minimum) *V—Hilbert space, $\mathcal{O} \subset V$ open, $h_i : \mathcal{O} \to \mathbb{R}$, $\forall i$*

$$U = \{u \in \mathcal{O} : h_i(u) \leqslant 0,\ 1 \leqslant i \leqslant q\} \tag{3.30}$$

(i) h_i *differentiable at* $\hat{u}$, $\forall i \in I(\hat{u})$
(ii) h_i *continuous at* $\hat{u}$, $\forall i \in \{1, \cdots, q\} \setminus I(\hat{u})$
(iii) *Constraints are qualified at* $\hat{u} \in U$
(iv) $F : \mathcal{O} \to \mathbb{R}$ *differentiable at* $\hat{u}$
(v) F *has a relative minimum at* $\hat{u}$ *w.r.t.* U

Then

$$\exists \lambda_i(\hat{u}) \geqslant 0,\ 1 \leqslant i \leqslant q \ \text{ such that } \ \nabla F(\hat{u}) + \sum_{i=1}^{q} \lambda_i(\hat{u}) \nabla h_i(\hat{u}) = 0, \quad \sum_{i=1}^{q} \lambda_i(\hat{u}) h_i(\hat{u}) = 0\ . \tag{3.31}$$

$\lambda_i(\hat{u})$ are called Generalized Lagrange Multipliers

Proof. Assumptions (i), (ii) and (iii) are as in Theorem 40, hence

$$C(\hat{u}) = Q(\hat{u}).$$

By Theorem 38 (p. 60), or equivalently by Theorem 28 (p. 47)—since $Q(\hat{u})$ is convex,

$$\langle \nabla F(\hat{u}), v \rangle \geqslant 0,\ \forall v \stackrel{\text{def}}{=} u - \hat{u} \in Q(\hat{u})\ . \tag{3.32}$$

The definition of Q at $\hat{u}$—see (3.29) (p. 61)—can be rewritten as

$$Q(\hat{u}) = \{v \in V : -\langle \nabla h_i(\hat{u}), v \rangle \geqslant 0,\ \forall i \in I(\hat{u})\}.$$

Combining this with (3.32), we can write down:

$$Q(\hat{u}) = \big\{v \in V\ :\ -\langle \nabla h_i(\hat{u}), v \rangle \geqslant 0,\ \forall i \in I(\hat{u})\big\} \subset \big\{v \in V\ :\ \langle \nabla F(\hat{u}), v \rangle \geqslant 0\big\}.$$

By the F-M Theorem 39 (p. 60),

$$\exists \lambda_i \geqslant 0,\ i \in I(\hat{u}) \ \text{ such that } \ \nabla F(\hat{u}) = -\sum_{i \in I(\hat{u})} \lambda_i(\hat{u}) \nabla h_i(\hat{u}). \tag{3.33}$$

From the definition of $I(\hat{u})$, we see that $\sum_{i \in I(\hat{u})} \lambda_i(\hat{u}) h_i(\hat{u}) = 0$. Fixing $\lambda_i(\hat{u}) = 0$ whenever $i \in I \setminus I(\hat{u})$ leads to the last equation in (3.31). □

Remarks

- $\lambda_i(\hat{u}) \geqslant 0, i \in I(\hat{u})$ are defined in a unique way if and only if the set $\{\nabla h_i(\hat{u}),\ i \in I(\hat{u})\}$ is linearly independent (recall the Farkas-Minkowski lemma).
- The KT relation (3.31) depend on $\hat{u}$—difficult to exploit directly.
- (3.31) yields a system of (often nonlinear) equations and inequations—not easy to solve.

3.6 Convex inequality constraint problems

3.6.1 Adaptation of previous results

Lemma 6 *If $\mathcal{O}$ is convex and $h_i : \mathcal{O} \subset V \to \mathbb{R},\ 1 \leqslant i \leqslant q$ are convex $\Rightarrow$ U in (3.30) (namely $U = \{u \in \mathcal{O} : h_i(u) \leqslant 0,\ 1 \leqslant i \leqslant q\}$) is convex.*

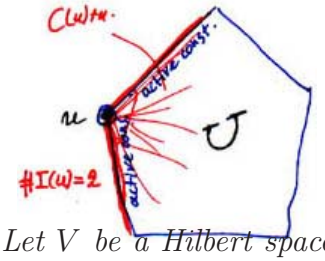
Proof. Straightforward (apply the basic definition for convexity). □

Definition 20 Convex constraints $h_i : \mathcal{O} \subset V \to \mathbb{R}$, $1 \leqslant i \leqslant q$ are qualified *if $U \neq \varnothing$ and if*

- *Either $\exists w \in \mathcal{O}$ such that $h_i(w) \leqslant 0,\ \forall i = 1, \ldots, q$ and $h_i(w) < 0$ if h_i is* not *affine.*
- *Or h_i, $1 \leqslant i \leqslant q$, are affine.*

Emphasize that these conditions are independent of $\hat{u}$, hence they are much easier to use.

Example 6 $U = \{v \in \mathbb{R}^n \ \mid \ \langle a_i, v\rangle \leqslant b_i,\ 1 \leqslant i \leqslant q\} \neq \varnothing$.
For $u \in U$ the set of active constraints $\langle a_i, u\rangle = b_i$ has a geometric representation as seen next.



Theorem 42 (Necessary and sufficient conditions for a minimum on U) *Let V be a Hilbert space, $\mathcal{O} \subset V$ (open) and $U \supset \mathcal{O}$ convex. Suppose $\hat{u} \in U$ and that*

(i) $F : \mathcal{O} \subset V \to \mathbb{R}$ and $h_i : \mathcal{O} \subset V \to \mathbb{R}$, $1 \leqslant i \leqslant q$ are differentiable at $\hat{u}$;

(ii) $h_i : V \to \mathbb{R}$, $1 \leqslant i \leqslant q$, are convex;

(iii) Convex constraints qualification holds (Definition 20).

Then we have the following statements:

1. *If F admits at $\hat{u} \in U$ a relative minimum w.r.t. U, then*

$$\exists\ \{\lambda_i(\hat{u}) \in \mathbb{R}_+ : 1 \leqslant i \leqslant q\} \quad \text{such that}$$

$$\nabla F(\hat{u}) + \sum_{i=1}^{q} \lambda_i(\hat{u}) \nabla h_i(\hat{u}) = 0 \quad \text{and} \tag{3.34}$$

$$\sum_{i=1}^{q} \lambda_i(\hat{u}) h_i(\hat{u}) = 0\ . \tag{3.35}$$

2. *If F is convex on U and (3.34)-(3.35) holds then F has at $\hat{u}$ a constrained minimum w.r.t. U.*

Note that as before, $I := \{i\ :\ h_i(\hat{u}) = 0$.

Proof. To prove 1, we have to show that <u>if</u> the constraints are qualified in the convex sense (Definition 20) then they are qualified in the general sense (Definition 19, p. 61), for any $u \in U$, which will allows us to apply the KT Theorem 41.
To prove statement 2, we have to check that $F(\hat{u}) \leqslant F(u)$, $\forall u \in U$.
The details are outlined in Appendix on p. 119.

If $\lambda_i(\hat{u}) \in \mathbb{R}^q_+$ were known, then we would have and unconstrained minimization problem.

Example 7 $U = \{u \in \mathbb{R}^n : Au \leqslant b \in \mathbb{R}^q\}$: the constraints are qualified if and only if U non-empty.

F convex. The necessary and sufficient conditions for F to have a constrained minimum at $\hat{u} \in U$: $\exists \lambda \in \mathbb{R}^q_+$ such that $\nabla F(\hat{u}) + A^T\lambda = 0$ with $\lambda_i = 0$ if $\langle a_i, \hat{u}\rangle - b_i < 0$.

3.6.2 Lagrangian Duality

V, W any subsets

Lemma 7 *For any $L : V \times W \to \mathbb{R}$, $\forall u \in V$ and $\forall \lambda \in W$ we have:*

$$\sup_{\lambda\in W}\inf_{u\in V} L(u,\lambda) \leqslant \inf_{u\in V}\sup_{\lambda\in W} L(u,\lambda)$$

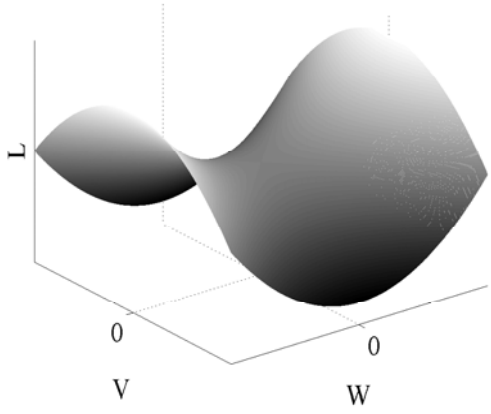
Proof. Take $u \in V$, $\lambda \in W$ arbitrary.

$$\inf_{u\in V} L(u,\lambda) \leqslant L(u,\lambda) \leqslant \sup_{\lambda\in W} L(u,\lambda), \quad \forall u \in V$$
$$\inf_{u\in V} L(u,\lambda) \leqslant \inf_{u\in V}\sup_{\lambda\in W} L(u,\lambda) \stackrel{\text{def}}{=} K$$
$$\inf_{u\in V} L(u,\lambda) \leqslant K, \quad \underline{\forall \lambda \in W} \quad \Rightarrow \quad \sup_{\lambda\in W}\inf_{u\in V} L(u,\lambda) \leqslant K = \inf_{u\in V}\sup_{\lambda\in W} L(u,\lambda)$$

The proof is complete. □

Definition 21 Saddle point *of $L : V \times W \to \mathbb{R}$ at $(\hat{u}, \hat{\lambda})$:* $\boxed{\sup_{\lambda\in W} L(\hat{u},\lambda) = L(\hat{u},\hat{\lambda}) = \inf_{u\in V} L(u,\hat{\lambda})}$

In our context: W—the space of Generalized Lagrange Multipliers



L has a saddle point at $(0, 0)$.

Theorem 43 *Let $(\hat{u}, \hat{\lambda})$ be a saddle-point for $L : V \times W \to \mathbb{R}$. Then*

$$\sup_{\lambda\in W}\inf_{u\in V} L(u,\lambda) = L(\hat{u},\hat{\lambda}) = \inf_{u\in V}\sup_{\lambda\in W} L(u,\lambda).$$

Proof. One has $\inf_{u\in V}\Big(\sup_{\lambda\in W} L(u,\lambda)\Big) \leqslant \sup_{\lambda\in W} L(\hat{u},\lambda) \stackrel{\text{Def.Saddle Pt.}}{=} \inf_{u\in V} L(u,\hat{\lambda}) \leqslant \sup_{\lambda\in W}\inf_{u\in V} L(u,\lambda)$.

Compare with Lemma 7 (p. 64) to conclude. □

For the reminder:

$$\boxed{(P) \quad \text{find} \ \ \hat{u} \ \ \text{such that} \quad F(\hat{u}) = \min_{u \in U} F(u) \quad \text{where} \ \ U = \{u : h_i(u) \leqslant 0, \ 1 \leqslant i \leqslant q\}}$$

The Lagrangian associated to (P): $\boxed{L(u, \lambda) = F(u) + \sum_{i=1}^{q} \lambda_i \, h_i(u)}$

Theorem 44 *Let $F : V \to \mathbb{R}$ and $h_i : V \to \mathbb{R}$, $1 \leqslant i \leqslant q$, where V is a Hilbert space.*

1. *$(\hat{u}, \hat{\lambda}) \in V \times \mathbb{R}^q_+$ is a saddle-point of L $\quad \Rightarrow \quad$ $\hat{u} \in U$ solves (P)*

2. *Let $\hat{u} \in U$ solve (P). Suppose also that* $\begin{cases} - & F \text{ and } h_i, \ 1 \leqslant i \leqslant q \text{ are differentiable at } \hat{u} \\ - & F \text{ and } h_i, \ 1 \leqslant i \leqslant q \text{ are convex} \\ - & \text{constraints are qualified (convex sense)} \end{cases}$

   $\Rightarrow \quad \exists \hat{\lambda} \in \mathbb{R}^q_+$ *such that $(\hat{u}, \hat{\lambda})$ is a saddle point of L*

Proof. 1). By $L(\hat{u}, \lambda) \leqslant L(\hat{u}, \hat{\lambda}), \ \forall \lambda \in \mathbb{R}^q_+$,

$$L(\hat{u}, \lambda) - L(\hat{u}, \hat{\lambda}) \leqslant 0, \quad \forall \lambda \in \mathbb{R}^q_+.$$

By the definition of L the latter reads

$$\begin{aligned} & F(\hat{u}) + \sum_{i=1}^{q} \lambda_i h_i(\hat{u}) - F(\hat{u}) - \sum_{i=1}^{q} \hat{\lambda}_i h_i(\hat{u}) \leqslant 0, \quad \forall \lambda \in \mathbb{R}^q_+ \\ \Leftrightarrow \quad & \sum_{i=1}^{q} \left(\lambda_i - \hat{\lambda}_i\right) h_i(\hat{u}) \leqslant 0, \quad \forall \lambda \in \mathbb{R}^q_+. \end{aligned} \tag{3.36}$$

Since $\hat{\lambda} \in \mathbb{R}^q_+$, for any $i \in \{1, \cdots, q\}$ apply (3.36) with

$$\lambda_i \to \infty, \ \lambda_j = \hat{\lambda}_j, \ j \neq i \quad \Rightarrow \quad h_i(\hat{u}) \leqslant 0.$$

Then $h_i(\hat{u}) \leqslant 0, \ 1 \leqslant i \leqslant q$, hence $\underline{\hat{u} \in U}$.

- Let $\lambda_i = 0, \ 1 \leqslant i \leqslant q \quad \overset{(3.36), \ \hat{u} \in U}{\Longrightarrow} \quad \sum_{i=1}^{q} \hat{\lambda}_i h_i(\hat{u}) \geqslant 0$

- $\Big[h_i(\hat{u}) \leqslant 0 \ \text{ and } \ \hat{\lambda}_i \geqslant 0\Big], \quad 1 \leqslant i \leqslant q \quad \Rightarrow \quad \sum_{i=1}^{q} \hat{\lambda}_i h_i(\hat{u}) \leqslant 0$

$$\Rightarrow \quad \underline{\sum_{i=1}^{q} \hat{\lambda}_i h_i(\hat{u}) = 0.}$$

Using that $L(\hat{u}, \hat{\lambda}) \leqslant L(u, \hat{\lambda}), \ \forall u \in U$

$$\begin{aligned} F(\hat{u}) &= F(\hat{u}) + \sum_{i=1}^{q} \hat{\lambda}_i h_i(\hat{u}) = L(\hat{u}, \hat{\lambda}) \leqslant L(u, \hat{\lambda}) \\ &= F(u) + \sum_{i=1}^{q} \hat{\lambda}_i h_i(u) \leqslant F(u), \ \forall u \in U \quad (\text{remind} \ \ \hat{\lambda}_i h_i(u) \leqslant 0, \ \forall u \in U) \end{aligned}$$

2). Let $\hat{u}$ solve (P). By KT theorem (see (3.34)-(3.35), p. 63), $\exists\ \hat{\lambda} \in \mathbb{R}^q_+$ such that

$$\nabla F(\hat{u}) + \sum_{i=1}^{q} \hat{\lambda}_i \nabla h_i(\hat{u}) = 0 \ \text{ and } \ \underbrace{\sum_{i=1}^{q} \hat{\lambda}_i h_i(\hat{u}) = 0}\,. \tag{3.37}$$

We have to check that $(\hat{u}, \hat{\lambda})$ is a saddle point of L.

$$\forall \lambda \in \mathbb{R}^q_+, \ \ \underline{L(\hat{u}, \lambda)} = F(\hat{u}) + \sum_{i=1}^{q} \lambda_i h_i(\hat{u}) \underline{\leqq} F(\hat{u}) = F(\hat{u}) + \underbrace{\sum_{i=1}^{q} \hat{\lambda}_i h_i(\hat{u})} = \underline{L(\hat{u}, \hat{\lambda})}.$$

$u \to F(u) + \sum_{i=1}^{q} \hat{\lambda}_i h_i(u) = L(u, \hat{\lambda})$ is convex and reaches its minimum at $\hat{u}$. Hence

$$\underline{L(\hat{u}, \hat{\lambda})} = F(\hat{u}) + \sum_{i=1}^{q} \hat{\lambda}_i h_i(\hat{u}) \underline{\leqslant L(u, \hat{\lambda})} = F(u) + \sum_{i=1}^{q} \hat{\lambda}_i h_i(u), \ \ \forall u \in U.$$

Hence $(\hat{u}, \hat{\lambda})$ is a saddle point of L. □

For $\lambda \in \mathbb{R}^q_+$ (called the dual variable) we define $u_\lambda \in V$ (Hilbert space) and $K : \mathbb{R}^q_+ \to \mathbb{R}$ by:

$$(P_\lambda) \qquad \begin{array}{rcl} u_\lambda \in V \ \mid \ L(u_\lambda, \lambda) & = & \inf\limits_{u \in V} L(u, \lambda) \\ K(\lambda) & \stackrel{\text{def}}{=} & L(u_\lambda, \lambda). \end{array} \tag{3.38}$$

Dual Problem:

$$(P^*) \qquad \hat{\lambda} \in \mathbb{R}^q_+ \ \mid \ K(\hat{\lambda}) = \sup_{\lambda \in \mathbb{R}^q_+} K(\lambda)$$

Lemma 8 *Assume that*

(i) $h_i : V \to \mathbb{R}$, $i = 1, \ldots, q$, *are continuous;*

(ii) $\forall \lambda \in \mathbb{R}^q_+$ (P_λ) *admits a unique solution* u_λ*;*

(iii) $\lambda \to u_\lambda$ *is continuous on* $\mathbb{R}^q_+$

Then K *in* (3.38) *is differentiable and* $\langle \nabla K(\lambda), \eta \rangle = \sum_{i=1}^{q} \eta_i \, h_i(u_\lambda), \ \forall \eta \in \mathbb{R}^q$.

The proof of this lemma is outlined in Appendix 7.9 on p. 119.

Theorem 45 *Two "reciprocal" statements.*

1. *Assume that*

$$\left\{ \begin{array}{ll} (i) & h_i : V \to \mathbb{R}, \ i = 1, \ldots, q, \text{ are continuous;} \\ (ii) & \forall \lambda \in \mathbb{R}^q_+ \ \ (P_\lambda) \text{ admits a unique solution } u_\lambda; \\ (iii) & \lambda \to u_\lambda \text{ is continuous on } \mathbb{R}^q_+; \\ (iv) & \hat{\lambda} \in \mathbb{R}^q_+ \text{ solves } (P^*) \end{array} \right. \quad \Rightarrow \quad u_{\hat{\lambda}} \text{ solves } (P)$$

2. Assume that

$$\left\{\begin{array}{ll}(i) & \textit{(P) admits a solution } \hat{u};\\ (ii) & F: V \to \mathbb{R} \textit{ and } h_i: V \to \mathbb{R},\ 1 \leqslant i \leqslant q \textit{ are convex;}\\ (iii) & F \textit{ and } h_i,\ 1 \leqslant i \leqslant q, \textit{ are differentiable at } \hat{u};\\ (iv) & \textit{(convex) constraints are qualified.}\end{array}\right. \quad \Rightarrow \quad (P^*) \textit{ admits a solution}$$

Proof. Statement 1. Let $\hat{\lambda}$ solve (P^*). Then

$$K(\hat{\lambda}) = L(u_{\hat{\lambda}}, \hat{\lambda}) = \inf_{u \in V} L(u, \hat{\lambda})$$

K is differentiable (Lemma 8 (p. 66)) and has at $\hat{\lambda}$ a maximum w.r.t. $\mathbb{R}^q_+$ (convex set), hence

$$\left\langle \nabla K(\hat{\lambda}), \lambda - \hat{\lambda} \right\rangle \underline{\leqq} 0, \quad \forall \lambda \in \mathbb{R}^q_+.$$

This, combined with Lemma 8, yields

$$\left\langle \nabla K(\hat{\lambda}), \lambda \right\rangle = \sum_{i=1}^{q} \lambda_i h_i(u_{\hat{\lambda}}) \leqslant \sum_{i=1}^{q} \hat{\lambda}_i h_i(u_{\hat{\lambda}}) = \left\langle \nabla K(\hat{\lambda}), \hat{\lambda} \right\rangle, \quad \forall \lambda \in \mathbb{R}^q_+.$$

$$\Rightarrow \quad \underbrace{L(u_{\hat{\lambda}}, \lambda)} = F(u_{\hat{\lambda}}) + \sum_{i=1}^{q} \lambda_i h_i(u_{\hat{\lambda}}) \underbrace{\leqslant} F(u_{\hat{\lambda}}) + \sum_{i=1}^{q} \hat{\lambda}_i h_i(u_{\hat{\lambda}}) = \underbrace{L(u_{\hat{\lambda}}, \hat{\lambda})}, \quad \underbrace{\forall \lambda \in \mathbb{R}^q_+}.$$

$$\Rightarrow \quad \sup_{\lambda \in \mathbb{R}^q_+} L(u_{\hat{\lambda}}, \lambda) = L(u_{\hat{\lambda}}, \hat{\lambda}).$$

Consequently $(u_{\hat{\lambda}}, \hat{\lambda})$ is a saddle point of L. By Theorem 44-1 (p. 65), $u_{\hat{\lambda}}$ solves (P).

Statement 2. Using Theorem 44-2 (p. 65), $\exists \hat{\lambda} \in \mathbb{R}^q_+$ such that $(\hat{u}, \hat{\lambda})$ is a saddle point of L.

$$L(\hat{u}, \hat{\lambda}) = \inf_{u \in V} L(u, \hat{\lambda}) = \sup_{\lambda \in \mathbb{R}^q_+} L(\hat{u}, \lambda) \quad \Leftrightarrow \quad K(\hat{\lambda}) = \sup_{\lambda \in \mathbb{R}^q_+} L(\hat{u}, \lambda).$$

The proof is complete. □

3.6.3 Uzawa's Method

Compute $\hat{\lambda}$ = a solution of (P^*)

Maximization of K using gradient with projection:

$$\begin{array}{rcll}\lambda_{k+1} & = & \Pi_+ \left(\lambda_k + \rho \nabla K(\lambda_k)\right) & \text{“}+\rho > 0\text{” because we maximize}\\ (\Pi_+ \lambda)[i] & = & \max\{\lambda[i],\ 0\}, \quad 1 \leqslant i \leqslant q, & \forall \lambda \in \mathbb{R}^q\\ \nabla K(\lambda_k)[i] & = & h_i(u_{\lambda_k}), \quad 1 \leqslant i \leqslant q & \\ u_k \overset{\text{def}}{=} u_{\lambda_k} & = & \arg\min_{u \in V} L(u, \lambda_k) & \end{array}$$

Alternate Optimization $(u_k, \lambda_k) \in V \times \mathbb{R}^q_+$. For $k \in \mathbb{N}$

$$\begin{array}{rcl} u_k & = & \arg\inf_{u \in V} \left\{ F(u) + \sum_{i=1}^{q} \lambda_k[i]\ h_i(u) \right\}\\ \lambda_{k+1}[i] & = & \max\left\{0,\ \lambda_k[i] + \rho\ h_i(u_k)\right\}, \quad 1 \leqslant i \leqslant q \end{array}$$

Theorem 46 *Assume that*

(i) $F : \mathbb{R}^n \to \mathbb{R}$ *is strongly convex with constant* $\mu > 0$

(ii) $U = \{u \in \mathbb{R}^n : Au \leqslant b\} \neq \varnothing, \quad A \in \mathbb{R}^{q\times n}, b \in \mathbb{R}^q$

(iii) $0 < \rho < \dfrac{2\mu}{\|A\|_2^2}$

Then

$$\lim_{k\to\infty} u_k = \hat{u} \quad (\textit{the unique solution of } (P))$$

If moreover $\operatorname{rank} A = q \quad \Rightarrow \quad \lim_{k\to\infty} \lambda_k = \hat{\lambda} \quad (\textit{the unique solution of } (P^*))$

Emphasize that (u_k) converges even if (λ_k) diverges.
Remind: $\|A\|_2 = \sup_u \frac{\|Au\|_2}{\|u\|_2}$, $\|.\|$ Euclidean norm $\|u\|_2 = \sqrt{\langle u, u\rangle}$.

Proof. Denote by $h : \mathbb{R}^n \to \mathbb{R}^q$ the function

$$h(u) \stackrel{\text{def}}{=} Au - b \in \mathbb{R}^q. \tag{3.39}$$

Then the constraint set U reads

$$U = \left\{u \in \mathbb{R}^n \;\middle|\; \big(h(u)\big)[i] \leqslant 0,\ 1 \leqslant i \leqslant q\right\}.$$

For $\lambda \in \mathbb{R}^q_+$,

$$L(u, \lambda) = F(u) + \langle \lambda, h(u)\rangle = F(u) + \langle \lambda,\, Au - b\rangle$$

Then $\exists \hat{\lambda} \in \mathbb{R}^q_+$ such that $(\hat{u}, \hat{\lambda})$ is a saddle point of L. The latter is defined by the system

$$\nabla F(\hat{u}) + A^T\hat{\lambda} = 0 \tag{3.40}$$

$$\left\langle h(\hat{u}), \lambda - \hat{\lambda}\right\rangle \leqslant 0, \quad \forall \lambda \in \mathbb{R}^q_+. \tag{3.41}$$

In order to proceed, one looks how to apply the projection theorem (29, p. 48). For any $\rho > 0$, (3.41) is equivalent to

$$\left\langle \hat{\lambda} - \big(\hat{\lambda} + \rho\, h(\hat{u})\big)\,,\, \lambda - \hat{\lambda}\right\rangle \geqslant 0, \quad \forall \lambda \in \mathbb{R}^q_+.$$

By the Projection theorem, $\hat{\lambda}$ is the projection of $\hat{\lambda} + \rho h(\hat{u})$ on $\mathbb{R}^q_+$, i.e.

$$\hat{\lambda} = \Pi_+\big(\hat{\lambda} + \rho\, h(\hat{u})\big).$$

Iterates solve the system for $k \in \mathbb{N}$:

$$\nabla F(u_k) + A^T\lambda_k = 0$$
$$\lambda_{k+1} = \Pi_+\big(\lambda_k + \rho\, h(u_k)\big).$$

$$\nabla F(u_k) - \nabla F(\hat{u}) + A^T(\lambda_k - \hat{\lambda}) = 0 \quad \Leftrightarrow \quad A^T(\lambda_k - \hat{\lambda}) = -\big(\nabla F(u_k) - \nabla F(\hat{u})\big) \tag{3.42}$$

Convergence of u_k:

$$\begin{aligned}
\|\lambda_{k+1}-\hat\lambda\|_2^2 &= \|\Pi_+\big(\lambda_k+\rho\,h(u_k)\big)-\Pi_+\big(\hat\lambda+\rho\,h(\hat u)\big)\|_2^2\\
&\leqslant \|\lambda_k-\hat\lambda+\rho\,A(u_k-\hat u)\|_2^2 \qquad \text{by (3.39): } \ h(u_k)-h(\hat u)=A(u_k-\hat u)\\
&= \|\lambda_k-\hat\lambda\|_2^2+2\rho\left\langle A^T(\lambda_k-\hat\lambda)\,,\,u_k-\hat u\right\rangle+\rho^2\,\|A(u_k-\hat u)\|_2^2\\
&= \|\lambda_k-\hat\lambda\|_2^2-2\rho\,\langle\nabla F(u_k)-\nabla F(\hat u)\,,\,u_k-\hat u\rangle+\rho^2\,\|A(u_k-\hat u)\|_2^2 \quad \text{(using (3.42))}\\
&\leqslant \|\lambda_k-\hat\lambda\|_2^2-\rho 2\mu\,\|u_k-\hat u\|_2^2+\rho^2\,\|A\|_2^2\,\|u_k-\hat u\|_2^2 \quad (F \text{ is strongly convex, see (i)})\\
&= \|\lambda_k-\hat\lambda\|_2^2-\rho\left(2\mu-\rho\,\|A\|_2^2\right)\,\|u_k-\hat u\|_2^2
\end{aligned}$$

Note that by (iii), we have $2\mu-\rho\,\|A\|_2^2>0$. It follows that $\|\lambda_{k+1}-\hat\lambda\|_2\leqslant\|\lambda_k-\hat\lambda\|_2$, hence $\Big(\|\lambda_{k+1}-\hat\lambda\|_2\Big)_{k\in\mathbb{N}}$ is decreasing. Being bounded from below, $\Big(\|\lambda_{k+1}-\hat\lambda\|_2\Big)_{k\in\mathbb{N}}$ converges (not necessarily to 0), that is

$$\lim_{k\to\infty}\Big(\|\lambda_{k+1}-\hat\lambda\|_2^2-\|\lambda_k-\hat\lambda\|_2^2\Big)=0$$

Then

$$0\leqslant\rho\left(2\mu-\rho\,\|A\|_2^2\right)\,\|u_k-\hat u\|^2\leqslant\|\lambda_k-\hat\lambda\|_2^2-\|\lambda_{k+1}-\hat\lambda\|_2^2\to 0\ \text{ as }\ k\to\infty.$$

Consequently, $\|u_k-\hat u\|_2\to 0$ as $k\to\infty$.

Possible convergence for $\boldsymbol{\lambda^k}$

$(\lambda^k)_{k\geqslant 0}$ bounded (because $\|\lambda^k-\hat\lambda\|$ decreasing) $\Rightarrow$ $\exists$ subsequence $\lambda^{k'}\to\hat\lambda'$ such that

$$\nabla\mathcal{F}(\hat u)+A^T\hat\lambda'=\lim_{k'\to\infty}\Big(\nabla\mathcal{F}(u^{k'})+A^T\lambda^{k'}\Big)$$

If $\mathrm{rank}A=q$, then:
$\Big[\ \mathrm{Range}(A)=\mathbb{R}^q\ \Leftrightarrow\ \ker A^T=\{0\}\ \Big]$ hence $\nabla\mathcal{F}(\hat u)+A^T\hat\lambda=0$ has a unique solution. □

Remark 21 *The method of Uzawa amounts to projected gradient maximization with respect to λ (solving the dual problem).*

3.7 Unifying framework and second-order conditions

Find $\hat u$ such that $F(\hat u)=\inf\limits_{u\in U}F(u)$ where

$$U=\left\{u\in\mathbb{R}^n:\ \begin{array}{rcl} g_i(u) &=& 0,\quad 1\leqslant i\leqslant p\\ h_i(u) &\leqslant& 0,\quad 1\leqslant i\leqslant q\end{array}\right\}\neq\varnothing$$

Associated Lagrangian:

$$L(u,\lambda,\mu)=F(u)+\sum_{i=1}^{p}\lambda_i\,g_i(u)+\sum_{i=1}^{q}\mu_i\,h_i(u),\quad \mu_i\geqslant 0,\quad 1\leqslant\mu\leqslant q.$$

3.7.1 Karush-Kuhn-Tucker Conditions (1st order)

Theorem 47 (KKT) *Let $\hat{u} \in U$ be a solution (local) of $\inf_{u\in U} F(u)$ and*

1. *$F : \mathbb{R}^n \to \mathbb{R}$, $\{g_i\}$ and $\{h_i\}$ be $\mathcal{C}^1$ on $\mathcal{O}(\hat{u})$*
2. *$\{\nabla g_i(\hat{u}), 1 \leqslant i \leqslant p \ \& \ \nabla h_i(\hat{u}), i \in I(\hat{u})\}$ linearly independent*

$\Rightarrow$ $\exists\ \hat{\lambda} \in \mathbb{R}^p$ *and* $\hat{\mu} \in \mathbb{R}^q_+$ *such that for*

$$\boxed{\begin{array}{l} \nabla_u L(\hat{u}, \hat{\lambda}, \hat{\mu}) = 0 \\ 1 \leqslant i \leqslant p: \quad g_i(\hat{u}) = 0 \ \text{ and } \ \hat{\lambda}_i g_i(\hat{u}) = 0 \\ 1 \leqslant i \leqslant q: \quad h_i(\hat{u}) \leqslant 0, \ \ \hat{\mu}_i \geqslant 0 \ \text{ and } \ \hat{\mu}_i h_i(\hat{u}) = 0 \end{array}}$$

Note that $\hat{\mu}_i > 0,\ \forall i \in I(\hat{u})$ - complementarity conditions (facilities for numerical methods). For details – see [10, p. 331].

$\hat{\lambda}$, $\hat{\mu}$ uniqueness by assumption 2.

3.7.2 Second order conditions

(To verify if $\hat{u}$ is a minimum indeed)

$$\text{Critical cone } \hat{C} = \left\{ v \in C(\hat{u}) : \begin{array}{l} \langle \nabla g_i(\hat{u}), v\rangle = 0, 1 \leqslant i \leqslant p \\ \langle \nabla h_i(\hat{u}), v\rangle = 0, \text{ if } i \in I(\hat{u}) \text{ and } \hat{\mu}_i > 0 \end{array} \right\}$$

Theorem 48 (CN) *Suppose that $F : \mathbb{R}^n \to \mathbb{R}$, $\{g_i\}$ and $\{h_i\}$ are $\mathcal{C}^2$ on $\mathcal{O}(\hat{u})$ and that all conditions of Theorem 47 hold.*

(CN) Then $\nabla^2_{uu} L(\hat{u}, \hat{\lambda}, \hat{\mu})(v, v) \succeq 0, \quad \forall v \in \hat{C}$.

(CS) If in addition $\nabla^2_{uu} L(\hat{u}, \hat{\lambda}, \hat{\mu})(v, v) \succ 0, \forall v \in \hat{C} \setminus \{0\}$ then $\hat{u}$ = strict (local) minimum of $\inf_{u\in U} F(u)$.

Lagrangian convex (non-negative curvature) for all directions in $\hat{C}$.

3.7.3 Standard forms (QP, LP)

A. Linear programming (LP) We are given: $c \in \mathbb{R}^n$, $A \in \mathbb{R}^{p\times n}$, $b \in \mathbb{R}^p$, $C \in \mathbb{R}^{q\times n}$ and $f \in \mathbb{R}^q$. Assumption: the constraint set is nonempty and it is not reduced to one point.

$$\min \langle c, u\rangle \quad \text{subject to} \quad \left\{ \begin{array}{l} Au - b = 0 \in \mathbb{R}^p \\ Cu - f \leqslant 0 \in \mathbb{R}^q \end{array} \right.$$

$L(u, \lambda, \mu) = \langle c, u\rangle + \langle \lambda, Au - b\rangle + \langle \mu, Cu - f\rangle$. Dual problem (KKT):

$$\begin{aligned} c + A^T\lambda + C^T\mu &= 0 \\ Au - b &= 0 \\ \langle \mu, Cu - f\rangle &= 0 \\ \mu &\geqslant 0 \end{aligned}$$

B. Quadratic programming (QP) We are given: $B \in \mathbb{R}^{n\times n}$ with $B \succ 0$ and $B^T = B$, $A \in \mathbb{R}^{p\times n}$, $b \in \mathbb{R}^p$, $C \in \mathbb{R}^{q\times n}$ and $f \in \mathbb{R}^q$. The same assumption as in LP.

$$F(u) = \frac{1}{2}\langle Bu, u\rangle - \langle c, u\rangle \quad \text{subject to} \quad \begin{cases} Au - b = 0 \in \mathbb{R}^p \\ Cu - f \leqslant 0 \in \mathbb{R}^q \end{cases}$$

There is exactly one solution.

Associated Lagrangian

$L(u, \lambda, \mu) = \frac{1}{2}\langle Bu, u\rangle - \langle c, u\rangle + \langle \lambda, Au - b\rangle + \langle \mu, Cu - f\rangle$. Dual problem (KKT), similarly.

N.B. Various algorithms to solve LP or QP can be found on the web. Otherwise, see next subsection.

3.7.4 Interior point methods

Constraints yield numerical difficulties. (When a constraint is satisfied, it can be difficult to realize a good decrease at the next iteration.)

Interior point methods—main idea: satisfy constraints non-strictly. Constraints are satisfied asymptotically. At each iteration one realizes an important step along the direction given by $\nabla^2 F$ even though calculations are more tricky.

Actually interior point methods are considered among the most powerful methods in presence of constraints (both linear or non-linear). See [10] (chapters 14 and 19) or [52] (chapter 1).

Sketch of example: minimize $F : \mathbb{R}^n \to \mathbb{R}$ subject to $Au \leqslant b$, $b \in \mathbb{R}^q$.

Set $y = b - Au$ then $y \geqslant 0$.

Notations: $\mathbb{1} = [1, \ldots, 1]^T$, $Y = \mathrm{diag}(y_1, \ldots, y_q)$, $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_q)$

We need to solve:

$$S(u, y, \lambda) \stackrel{\text{def}}{=} \begin{bmatrix} \nabla F(u) + A^T\lambda \\ Au - b + y \\ Y\Lambda\mathbb{1} \end{bmatrix} = 0 \quad \text{subject to} \quad y \geqslant 0,\ \lambda \geqslant 0.$$

Note that $Y\Lambda\mathbb{1} = 0$ is the complementarity condition.

Duality measure: $\mu = \dfrac{1}{q} y^T\lambda$.

Central path defined using $\tau = \sigma\mu$ for $\sigma \in [0, 1]$ fixed.

At each iteration, one derives a central path $(u_\tau, y_\tau, \lambda_\tau)$, $\tau > 0$ by solving

$$S(u_\tau, y_\tau, \lambda_\tau) = \begin{bmatrix} 0 \\ 0 \\ \tau\mathbb{1} \end{bmatrix}, y_\tau > 0, \lambda_\tau > 0$$

The step to a new point is done so that we remain in the interior of U. The complementary condition is relaxed using τ.

The new direction in the variables u, λ, y is found using Newton's method.

Often some variables can be solved explicitly and eliminated from S.

Interior point methods are polynomial time methods, and were definitely one of the main events in optimization during the last decade, in particular, in linear programming.

3.8 Nesterov's approach

For V a Hilbert space, consider $F : V \to \mathbb{R}$ convex and Lipschitz differentiable and $U \subset V$ a closed and convex domain. The problem is to solve

$$\inf_{u \in U} F(u) \tag{3.43}$$

The Lipschitz constant of ∇F is denoted by ℓ.

It is shown in 2004 by Nesterov, [53, Theorem 2.1.7], that no algorithm that uses only values $F(u_k)$ and gradients $\nabla F(u_k)$ has a better rate of convergence than $O(1/k^2)$ uniformly on all problems of the form (3.43) where k is the iterations number. The convergence rate is in term of objective function, that is $|F(u_k) - F(\hat{u})| \leqslant C/k^2$ where C is a constant proportional to $\ell \|u_0 - \hat{u}\|^2$. This result alone gives no information on the convergence rate of u_k to $\hat{u}$.

Nestorov's algorithm [54].

- ℓ is the Lipschitz constant of ∇F;
- $\|\cdot\|$ is a norm and d is a convex function such that there exists $\sigma > 0$ satisfying

$$d(u) \geqslant \frac{\sigma}{2}\|u_0 - u\|^2, \qquad \forall u \in U.$$

- For $k = 0$ set $u_0 \in U$ and $x_0 = 0$.

For $k \geqslant 1$:

1. $\eta_k = \nabla F(u_k)$
2. $z_k = \arg\min_{z \in U} \left(\langle \eta_k, z - u_k \rangle + \frac{\ell}{2}\|z - u_k\|^2 \right)$
3. $x_k = x_{k-1} + \frac{k+1}{2}\eta_k$
4. $w_k = \arg\min_{w \in U} \left(\frac{\ell d(w)}{\sigma} + \langle x_k, w \rangle \right)$
5. $u_{k+1} = \frac{2}{k+3} w_k + \frac{k+1}{k+3} z_k$

Proposition 3 ([54]) *This algorithm satisfies*

$$0 \leqslant F(u_k) - F(\hat{u}) \leqslant \frac{4\ell d(\hat{u})}{\sigma(k+1)(k+2)}$$

The rationale: similarly to CG, one computes the direction at iteration k by using the information in $\{\nabla F(u_{k-1}), \cdots, \nabla F(u_0)\}$.

For details and applications in image processing, see [55] and [56].

Y. Nesterov proposes an accelerated algorithm in [57] where the estimation of the Lipschitz constant is adaptive and improved.

Chapter 4

Non differentiable problems

Textbooks: [17, 18, 7, 8, 58, 4]. If not specified, V is a real Hilbert space.

4.1 Specificities

4.1.1 Examples

Non-differentiable functions frequently arise in practice.

Minimax problems

Minimize $F(u) = \max_{i \in I} \varphi_i(u)$ where I is a finite set of indexes and $\forall i \in I$, φ_i is convex and smooth.

Regularized objective functionals on $\mathbb{R}^n$

The minimizers of functionals of the form $F(u) = \Psi(u) + \beta\Phi(u)$ (remind (1.5)-(1.6), p. 8, and the explanations that follow) are very popular to define a restored image or signal. They are often used in learning theory, in approximation and in many other fields. Nonsmooth F have attracted a particular attention because of the specific properties of their solutions.

- Non-differentiable regularization term

$$\Phi(u) = \sum_i \varphi(\|\mathrm{D}_i u\|), \quad \text{with} \quad \varphi'(0^+) > 0$$

  Note that $\varphi'(0^+) > 0$ entails that Φ is nonsmooth.

  - For $\varphi(t) = t$ and D_i the discrete approximation of the gradient of u at i, along with $\|.\| = \|.\|_2 =$ Total Variation (TV) regularization [59];
    For $\varphi(t) = t$ and $\mathrm{D}_i \in \mathbb{R}^{1\times n} =$ median pixel regularization (approximate TV) [60];
  - Other non-differentiable (nonconvex) potential functions φ:
    * $\varphi(t) = \dfrac{t}{\alpha + t}$, see [37, 25] :
    * $\varphi(t) = t^\alpha, \quad \alpha \in]0, 1[$
  - If $u =$ the coefficients of the decomposition of the image in a wavelet basis or a frame and $\mathrm{D}_i = e_i$, $\forall i$, then $\hat{u} =$ shrinkage estimation of the coefficients [61, 62, 63, 64]). See Examples 8 (p. 74) and 8 (p. 91).

- Non-differentiable data-fitting

  Since 2002, very popular in imaging science [65, 66, 67, 68, 69].

$$F(u) = \|Au - v\|_1 + \beta\Phi(u), \quad \psi'(0^+) > 0, \tag{4.1}$$

  The first term is known as ℓ_1 data fitting.
  $A =$ an $m \times n$ matrix with rows $= a_i^T$, $1 \leqslant i \leqslant m$.

Constrained nonsmooth problems

Typical forms are

$$\text{minimize } \|u\|_1 \text{ subject to } \|Au - v\| \leqslant \tau$$

or

$$\text{minimize } \|u\|_1 \text{ subject to } Au = v$$

These are frequently used in Compression, in Coding and in Compressive sensing. The reason is that they lead to sparse solution, i.e. $\hat{u}[i] = 0$ for many indexes i.

4.1.2 Kinks

Definition 22 *A kink is a point u where $\nabla F(u)$ is not defined (in the usual sense).*

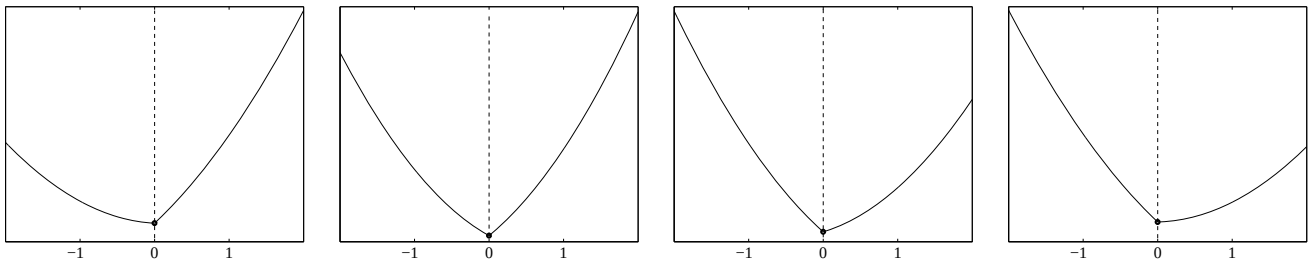
Theorem 49 *(*Rademacher, [7, p. 189]) Let $F : \mathbb{R}^n \to]-\infty, +\infty]$ be convex and $F \not\equiv +\infty$. Then the subsetset $\left\{u \in \operatorname{int}\operatorname{dom} F \ : \ \nexists \nabla F(u)\right\}$ is of Lebesgue measure zero in $\mathbb{R}^n$.

The statement extends to nonconvex functions [70, p.403] and to mappings $F : \mathbb{R}^n \to \mathbb{R}^m$ [71, p.81].

Hence F is differentiable at almost every u. However minimizers are frequently located at kinks.

Example 8 *Consider $F(u) = \dfrac{1}{2}\|u - w\|^2 + \beta|u|$ for $\beta > 0$ and $u, w \in \mathbb{R}$. The minimizer $\hat{u}$ of F reads*

$$\hat{u} = \begin{cases} w + \beta & \text{if } \ w < -\beta \\ 0 & \text{if } \ |w| \leqslant \beta \\ w - \beta & \text{if } \ w > \beta \end{cases} \qquad (\hat{u} \text{ is shrunk w.r.t. } w.)$$



$\beta = 1,\ w = -0.9$ $\beta = 1,\ w = -0.2$ $\beta = 1,\ w = 0.5$ $\beta = 1,\ w = 0.95$

4.2 Basic notions

Definition 23 $\Gamma_0(V)$ *is the class of all l.s.c. convex proper functions on V (a real Hilbert space) . (Remind Definitions 4, 6 and 8 on p. 13.)*

4.2.1 Preliminaries

Definition 24 *$f : V \to \mathbb{R} \cup \{+\infty\}$, is said to be* positively homogeneous *if* $f(\nu u) = \nu f(u),\ \forall \nu > 0, \forall u \in V$.

Definition 25 *$f : V \to \mathbb{R} \cup \{+\infty\}$, is said to be* sublinear *if it is convex and positively homogeneous.*

Lemma 9 *f is* positively homogeneous $\iff$ $f(\nu u) \leqslant \nu f(u),\ \forall \nu > 0, \forall u \in V$

Proof. ($\Rightarrow$) is obvious. Next: ($\Leftarrow$). Let $\underline{f(\nu u) \leqslant \nu f(u)},\ \forall \nu > 0, \forall u \in V$. Since $\nu u \in V$ and $\nu^{-1} > 0$

$$f(u) = f(\nu^{-1}\nu u) \leqslant \nu^{-1} f(\nu u) \quad \Leftrightarrow \quad \underline{\nu f(u) \leqslant f(\nu u)},\ \forall \nu > 0, \forall u \in V$$

hence $\underline{f(\nu u) = \nu f(u),\ \forall \nu > 0, \forall u \in V}$. □

Inequalities are usually easier to check that equalities.

Proposition 4 *f is sublinear* $\Leftrightarrow$ $f(\nu u + \mu v) \leqslant \nu f(u) + \mu f(v),\ \forall (u, v) \in V^2,\ \forall \nu > 0, \forall \mu > 0$

Proof. ($\Rightarrow$) Set $\eta = \nu + \mu$. Using the convexity of f for $1 - \frac{\mu}{\eta} = \frac{\nu}{\eta} > 0$,

$$\underbrace{f(\nu u + \mu v)}_{} = f\left(\eta\left(\frac{\nu}{\eta}u + \frac{\mu}{\eta}v\right)\right) = \eta f\left(\frac{\nu}{\eta}u + \frac{\mu}{\eta}v\right) \underbrace{\leqslant}_{} \eta\left(\frac{\nu}{\eta}f(u) + \frac{\mu}{\eta}f(v)\right) = \underbrace{\nu f(u) + \mu f(v)}_{} \tag{4.2}$$

($\Leftarrow$) Taking $\mu + \nu = 1$ in (4.2) shows that f is convex. Furthermore

$$\underbrace{f(\nu u)}_{} = f\left(\frac{\nu}{2}u + \frac{\nu}{2}u\right) \underbrace{\leqslant}_{} 2\frac{\nu}{2}f(u) = \underbrace{\nu f(u)}_{}.$$

f is positively homogeneous by Lemma 9. □

Sublinearity is stronger than convexity—it does not require $\mu + \nu = 1$—and weaker than linearity—it requires that $\mu > 0,\ \nu > 0$.

Definition 26 *Suppose that $U \subset V$ is nonempty.*

- Support function $\sigma_U(u) = \sup_{s \in U} \langle s, u \rangle \in \mathbb{R} \cup \{\infty\}$
- Indicator function $\iota_U(u) = \begin{cases} 0 & \text{if } u \in U \\ +\infty & \text{if } u \notin U \end{cases}$
- Distance from $u \in V$ to U: $\mathrm{d}_U(u) = \inf_{v \in U} \|u - v\|$.

We denote by $f^\star : V^\star \to \mathbb{R}$ the convex conjugate of the function $f : V \to \mathbb{R}$ (see (1.14), p. 17). When $U \neq \varnothing$ is convex and closed, we have

$$\iota_U^\star(u) = \sup_{v \in V} \big(\langle u, v \rangle - \iota_U(v) \big) = \sup_{v \in U} \langle u, v \rangle = \sigma_U(u) \quad \text{and} \quad \sigma_U^\star(v) = \iota_U^{\star\star}(v) = \iota_U(v) \tag{4.3}$$

where the second part results from Theorem 9 (p. 18).

Proposition 5 (p. 19, [17]) *σ_U is l.s.c. and sub-linear. Moreover*

$$\sigma_U(u) < \infty \ \ \forall u \in V \quad \Leftrightarrow \quad U \subset V \text{ is bounded.}$$

Lemma 10 *Any ℓ_p-norm is the support function of the unit ball B_q of the dual norm ℓ_q ($\frac{1}{p} + \frac{1}{q} = 1$).*

Note that this is the kind of functions we deal with in practical optimization problems.

4.2.2 Directional derivatives

Lemma 11 *Let $F : V \to \mathbb{R}$ be convex, where V is a n.v.s. The function*

$$t \to \frac{F(u+tv) - F(u)}{t}, \quad t \in \mathbb{R}_+$$

is increasing.

Proof. For any $t > 0$, choose an arbitrary $\tau \geqslant 0$. Set

$$\alpha \stackrel{\text{def}}{=} \frac{\tau}{t+\tau}, \quad \text{then} \quad 1 - \alpha = \frac{t}{t+\tau}$$

We check that

$$\alpha u + (1-\alpha)(u + (t+\tau)v) = \frac{\tau}{t+\tau}u + \frac{t}{t+\tau}u + \frac{t}{t+\tau}(t+\tau)v = \underline{u+tv}$$

Hence

$$F\big(\alpha u + (1-\alpha)(u+(t+\tau)v)\big) = F(u+tv).$$

Using the last result and the convexity of F yield

$$\begin{aligned} \underbrace{F(u+tv)} \quad &= \quad F(\alpha u + (1-\alpha)\big(u+(t+\tau)v\big) \\ &\underbrace{\leqslant} \quad \frac{\tau}{t+\tau}F(u) + \frac{t}{t+\tau}F\big(u+(t+\tau)v\big) = \underbrace{F(u) + \frac{t}{t+\tau}\Big(F\big(u+(t+\tau)v\big) - F(u)\Big)} \end{aligned}$$

It follows that

$$\frac{F(u+tv) - F(u)}{t} \leqslant \frac{F\big(u+(t+\tau)v\big) - F(u)}{t+\tau}, \quad \forall t > 0, \ \forall \tau \in \mathbb{R}_+ \ .$$

□

Definition 27 *Let $F : V \to \overline{\mathbb{R}}$ be convex and proper. The (one-sided)* directional derivative *of F at $u \in V$ along the direction of $v \in V$ reads*

$$\delta F(u)(v) = \lim_{t \searrow 0} \frac{F(u+tv) - F(u)}{t}, \quad \forall u \in V \tag{4.4}$$

$$= \inf_{t>0} \frac{F(u+tv) - F(u)}{t}, \quad \forall u \in V \tag{4.5}$$

Whenever F is convex and proper, the limit in (4.4) always exists (see [17, p. 23]). The definition in (4.4) is equivalent to (4.5) since by Lemma 11 (p. 76), the function $t \to \frac{F(u+tv)-F(u)}{t}$, $t \in \mathbb{R}_+$ goes to its unique infimum when $t \searrow 0$.

Remark 22 *For nonconvex functions, defined on more general spaces, directional derivatives can be defined only using (4.4); they do exist whenever the limit in (4.4) do exist.*

$\delta F(u)(v)$ is the right-hand side derivative. The left-hand side derivative is $-\delta F(u)(-v)$.

$$F \ \text{ convex} \quad \Rightarrow \quad -\delta F(u)(-v) \leqslant \delta F(u)(v).$$

When F is differentiable at u in the usual sense, $\delta F(u)(v) = \langle \nabla F(u), v \rangle$, $\forall v \in V$.

Proposition 6 *Let $F : V \to \overline{\mathbb{R}}$ be convex and proper. For any $u \in V$ fixed, $v \to \delta F(u)(v)$ is sublinear.*

Proof. By Definition 27—(4.4), it is obvious that $v \to \delta F(u)(v)$ is positively homogeneous, i.e. $\delta F(u)(\lambda v) = \lambda \delta F(u)(v)$, $\forall \lambda > 0$. Let us check that it is convex. Choose $\mu > 0$ and $\nu > 0$ such that $\mu + \nu = 1$, and $v \in V$ and $w \in V$.

$$\begin{aligned} F\big(u + t(\mu v + \nu w)\big) - F(u) &= F\big(\mu(u + tv) + \nu(u + tw)\big) - \big(\mu F(u) + \nu F(u)\big), \quad \forall t > 0 \\ &\leqslant \mu\big(F(u + tv) - F(u)\big) + \nu\big(F(u + tw) - F(u)\big), \quad \forall t > 0. \end{aligned}$$

Divide by $t > 0$

$$\frac{F\big(u + t(\mu v + \nu w)\big) - F(u)}{t} \leqslant \mu \frac{F(u + tv) - F(u)}{t} + \nu \frac{F(u + tw) - F(u)}{t}, \quad \forall t > 0.$$

For $t \searrow 0$ we get $\delta F(u)(\mu v + \nu w) \leqslant \mu \delta F(u)(v) + \nu \delta F(u)(w)$. □

Proposition 7 (p. 239, [7]) *If $F \in \Gamma_0(V)$ is Lipschitz with constant $\ell > 0$ on $B(u, \rho)$, $\rho > 0$ then*

$$\|z - u\| \leqslant \rho \;\Rightarrow\; |\delta F(z)(v) - \delta F(z)(w)| \leqslant \ell \|v - w\|, \quad \forall v, w \in V.$$

Proposition 8 (1^{st}-order approximation) *Let $F : \mathbb{R}^n \to \mathbb{R}$ be convex and proper, with Lipschitz constant $\ell > 0$ and $u \in \mathbb{R}^n$. Then $\forall \varepsilon > 0$, $\exists \rho > 0$ such that*

$$\|v\| < \rho \quad \Rightarrow \quad |F(u + v) - F(u) - \delta F(u)(v)| \leqslant \varepsilon \|v\|.$$

The proof of the proposition is outlined in Appendix, p. 120

4.2.3 Subdifferentials

For a given space V, the class of all subsets of V is denoted by 2^V. For a mapping T from V to 2^V one uses the notations

$$T : V \to 2^V \quad \text{and} \quad T \rightrightarrows V$$

If T is single-valued, these amount to $T : V \to V$.

Definition 28 *Let $F : V \to \overline{\mathbb{R}}$ be convex and proper. The* subdifferential *of F is the* set-valued operator *$\partial F : V \to 2^V$ whose values at $u \in V$ are given by*

$$\begin{aligned} \partial F(u) &= \{g \in V : \langle g, v \rangle \leqslant \delta F(u)(v), \forall v \in V\} && (4.6) \\ &= \{g \in V : F(z) \geqslant F(u) + \langle g, z - u \rangle, \ \forall z \in V\} && (4.7) \end{aligned}$$

A subgradient *of F at u is a selection $g \in V$ such that $g \in \partial F(u)$.*

If F is differentiable at u then $\partial F(u) = \{\nabla F(u)\}$.

Lemma 12 *The two formulations for ∂F in (4.6) and in (4.7), Definition 28, are equivalent.*

Proof. Since by Lemma 11, the function $t \to \frac{F(u+tv)-F(u)}{t}$, $t \in \mathbb{R}_+$, has its infimum for $t \searrow 0$, using the definition for δF in (4.5), we can write

$$\begin{aligned}
\partial F(u) = \{g \in V : \langle g, v\rangle \leqslant \delta F(u)(v), \forall v \in V\} &= \left\{g \in V : \langle g, v\rangle \leqslant \frac{F(u+tv)-F(u)}{t}, \forall v \in V, \forall t > 0\right\} \\
\text{(use } z = u + tv \Leftrightarrow v = (z-u)/t) &= \left\{g \in V : \frac{1}{t}\langle g, z-u\rangle \leqslant \frac{F(z)-F(u)}{t}, \forall z \in V, \forall t > 0\right\} \\
&= \{g \in V : \langle g, z-u\rangle \leqslant F(z) - F(u), \forall z \in V\}
\end{aligned}$$

The conclusion follows from the observation that when (u, v) describe V^2 and t describes $\mathbb{R}_+$, then $z = u + tv$ describes V. □

Observe that by the definition of ∂F in (4.6) (p. 77),

$$\boxed{\delta F(u)(v) = \sup\big\{\langle g, v\rangle : g \in \partial F(u)\big\} = \sigma_{\partial F(u)}(v)}$$

where σ is the support function (see Definition 26, p. 75). Moreover,

$$-\delta F(u)(-v) \leqslant \langle g, v\rangle \leqslant \delta F(u)(v), \quad \forall (g, v) \in (\partial F(u) \times V)$$

Property 7 *Let $F \in \Gamma_0(V)$. Then the set $\{\partial F(u)\}$ is closed and convex [p. 277, [72]].*

Theorem 50 *Let $F \in \Gamma_0(V)$. Then ∂F is a* monotone mapping *[58, Theorem 3.1.11]:*

$$\forall u_1, u_2 \in V \quad \Rightarrow \quad \langle g_2 - g_1, u_2 - u_1\rangle \geqslant 0, \quad \forall g_1 \in \partial F(u_1), \quad g_2 \in \partial F(u_2)$$

Theorem 51 (p. 281,[7]) *A function $F \in \Gamma_0(V)$ is strictly convex if and only if*

$$\forall u_1, u_2 \in V \quad \Rightarrow \quad \langle g_2 - g_1, u_2 - u_1\rangle > 0, \quad \forall g_1 \in \partial F(u_1), \quad g_2 \in \partial F(u_2)$$

Proposition 9 (p. 263, [73]) *Let $G \in \mathbb{R}^{m\times n}$ and $F : \mathbb{R}^n \to \mathbb{R}$ read*

$$F(u) = \|Gu\|_2$$

Then

$$\partial F(u) = \begin{cases} \dfrac{G^T G u}{\|Gu\|_2} & \text{if} \quad Gu \neq 0 \\ \\ \left\{G^T h \;\middle|\; \|h\|_2 \leqslant 1, h \in \mathbb{R}^m\right\} & \text{if} \quad Gu = 0 \end{cases} \tag{4.8}$$

Proof. Form the definition of the subdifferential (see Definition 28)

$$\partial F(u) = \left\{ g \in \mathbb{R}^n \;\middle|\; \|Gz\|_2 - \|Gu\|_2 \geqslant \langle g, z-u \rangle \,,\ \forall z \in \mathbb{R}^n \right\} \tag{4.9}$$

If $Gu \neq 0$ then F is differentiable at u and $\partial F(u) = \{\nabla F(u)\}$, hence the result [1].

Consider that $Gu = 0$. Clearly, $\|Gz\|_2 = \|G(z-u)\|_2$. After setting $w = z-u$, (4.9) equivalently reads

$$\partial F(u) = \left\{ g \in \mathbb{R}^n \;\middle|\; \|Gw\|_2 \geqslant \langle g, w \rangle \,,\ \forall w \in \mathbb{R}^n \right\} \tag{4.10}$$

Define the set

$$S \stackrel{\text{def}}{=} \left\{ G^T h \;\middle|\; \|h\|_2 \leqslant 1, h \in \mathbb{R}^m \right\}$$

If $G^T h \in S$, then Schwarz inequality yields

$$\left\langle G^T h \,, w \right\rangle = \langle h \,, Gw \rangle \leqslant \|Gw\|_2, \ \forall w \in \mathbb{R}^n$$

By (4.10), $G^T h \in \partial F(u)$ and hence $S \subseteq \partial F(u)$.

Suppose that

$$\exists\ g \in \partial F(u) \quad \text{obeying} \quad g \notin S \tag{4.11}$$

Then using the H-B separation Theorem 8 (p. 16), there exist $w \in \mathbb{R}^n$ and $\alpha \in \mathbb{R}$ so that the hyperplane $\{x \in \mathbb{R}^n \;|\; \langle w, x \rangle = \alpha\}$ separates g and S so that

$$\langle g, w \rangle > \alpha > \langle z, w \rangle \,, \quad \forall z \in S$$

Consequently,

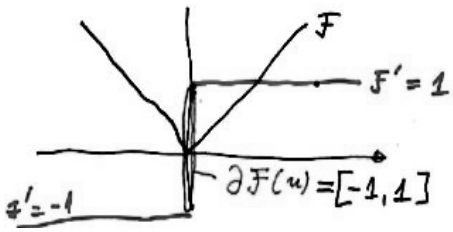
$$\langle g, w \rangle > \alpha \geqslant \sup_h \left\{ \left\langle G^T h, w \right\rangle \;\middle|\; \|h\|_2 \leqslant 1, h \in \mathbb{R}^m \right\} = \|Gw\|_2$$

A comparison with (4.10) shows that $g \notin \partial F(u)$. Hence the assumption in (4.11) is false. Consequently, $\partial F(u) = S$. □

Example 9 *Let* $F : \mathbb{R} \to \mathbb{R}$ *read* $F(u) = |u|$. *By* (4.8)

$u \neq 0 \quad \Rightarrow \quad \partial F(u) = \frac{u}{|u|} = \text{sign}(u)$

$u = 0 \quad \Rightarrow \quad \partial F(0) = \left\{ h \in \mathbb{R} \;\middle|\; |h| \leqslant 1 \right\} = [-1, +1]$



Some calculus rules ([58, 7, 74])

- $\Psi \in \Gamma_0(V)$ and $\Phi \in \Gamma_0(V)$, and $\exists \tilde{u} \in \text{dom}\Psi \cap \text{dom}\Phi$ where Ψ or Φ is continuous

$$\partial(\mu\Psi + \nu\Phi)(u) = \mu\partial\Psi(u) + \nu\partial\Phi(u), \quad \mu, \nu \geqslant 0, \quad \forall u \in V.$$

[1] Let $f \stackrel{\text{def}}{=} \|u\|_2 = \left(\sum_i u[i]^2\right)^{1/2}$. Then

$$u \neq 0 \quad \Rightarrow \quad \nabla f(u) = \frac{u}{\|u\|_2}$$

We have $F = f \circ G$. For $Gu \neq 0$, noticing that $\nabla Gu = G^T$, chain rule entails the first equation in (4.8).

- $\Psi \in \Gamma_0(V)$ and $\Phi \in \Gamma_0(V)$

$$\partial\big(\Psi(u) + \Phi(v)\big) = \partial\Psi(u) \times \partial\Phi(v), \quad \forall u \in V, \forall v \in V.$$

- $A : V \to W$ bounded linear operator, A^* its adjoint, $b \in W$ and $F \in \Gamma_0(W)$

$$\partial\big(F \circ (Au + b)\big) = A^* \circ \partial F(Au + b)$$

- $\Psi \in \Gamma_0(W)$, $\Phi \in \Gamma_0(V)$, $A : V \to W$ bounded linear operator, $0 \in \mathrm{int}\big(\mathrm{dom}(\Psi - A(\mathrm{dom}\Phi)\big)$

$$\partial(\Psi \circ A + \Phi) = A^* \circ (\partial\Psi) \circ A + \partial\Phi$$

- Let $U \subset V$ be nonempty, convex and closed. For any $u \in V \backslash U$ we have $\partial \mathrm{d}_U(u) = \left\{ \dfrac{u - \Pi_U u}{\mathrm{d}_U(u)} \right\}$. □

Continuity properties of the subdifferential of $F \in \Gamma_0(\mathbb{R}^n)$

Theorem 52 (Mean-value theorem [7], p.257.) *Let $F : \mathbb{R}^n \to \mathbb{R}$ be convex and proper. If $u \neq v$ then $\exists \theta \in (0,1)$ and $\exists g \in \partial F(\theta v + (1-\theta)u)$ such that $F(v) - F(u) = \langle g, v - u \rangle$*

Property 8 ([7]) *$\partial F(u)$ is compact for any $u \in \mathbb{R}^n$.*

Property 9 (p.282, [7]) *$u \to \partial F(u)$ is locally bounded: $B \subset \mathbb{R}^n$ bounded $\Rightarrow \partial F(B) \subset \mathbb{R}^n$ bounded.*

Moreover [p. 282, [7]]:

- $B \subset \mathbb{R}^n$ compact $\Rightarrow \partial F(B)$ compact.
- $B \subset \mathbb{R}^n$ compact and connected $\Rightarrow \partial F(B)$ compact and connected.
- $B \subset \mathbb{R}^n$ convex — in general $\partial F(B)$ is not convex.

∂F takes its values in a compact set when u itself varies in a compact set, according to

Theorem 53 (Continuity [7], p.283) *∂F is* outer semi-continuous *at any $u \in \mathbb{R}^n$:*

$$\forall \varepsilon > 0 \exists \rho > 0 \ : \ \|u - v\| \leqslant \rho \quad \Rightarrow \quad \partial F(v) \subset \partial F(u) + B(0, \varepsilon).$$

Corollary 1 (p. 283 [7]) *For $F : \mathbb{R}^n \to \mathbb{R}$ convex, the function $u \to \delta F(u)(v)$ is upper semi-continuous:*

$$u \in \mathbb{R}^n \quad \Rightarrow \quad \delta F(u)(v) = \limsup_{z \to u} \delta F(z)(v), \quad \forall v \in \mathbb{R}^n.$$

4.3 Optimality conditions

4.3.1 Unconstrained minimization problems

Necessary and sufficient condition for a global minimizer of a proper, convex function:

Theorem 54 (Fermat's rule) *Let $F \in \Gamma_0(V)$, then*

$$F(v) \geqslant F(\hat{u}), \forall v \in V \quad \Leftrightarrow \quad 0 \in \partial F(\hat{u}) \quad \Leftrightarrow \quad \delta F(\hat{u})(v) \geqslant 0, \forall v \in V.$$

Proof. Using Definition 28, (4.7) (p. 77)

$$g = 0 \in \partial F(\hat{u}) \quad \Leftrightarrow \quad F(v) \geqslant F(\hat{u}) + \langle 0, v - u \rangle, \ \forall v \in V \quad \Leftrightarrow \quad F(v) \geqslant F(\hat{u}), \ \forall v \in V$$

Using Definition 28, (4.6), namely $\partial F(u) = \{g \in V : \langle g, v \rangle \leqslant \delta F(u)(v), \forall v \in V\}$,

$$g = 0 \in \partial F(\hat{u}) \quad \Leftrightarrow \quad 0 \leqslant \delta F(\hat{u})(v), \forall v \in V$$

□

Denote the set of minimizers of F by

$$\widehat{U} = \{u \big| u \in (\partial F)^{-1}(0)\} \tag{4.12}$$

The set $\widehat{U}$ is closed and convex.

Remark 23 *If F is strictly convex and coercive, then $\widehat{U} = \{\hat{u}\}$, i.e. the minimizer is unique.*

4.3.2 General constrained minimization problems

Consider the problem

$$F(\hat{u}) = \min_{u \in U} F(u) \tag{4.13}$$

where $U \subset V$ is closed, convex and nonempty.

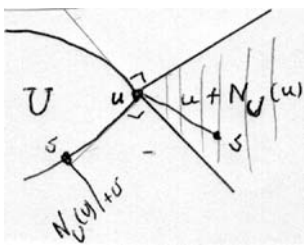
Remark 24 *For any $F \in \Gamma_0(V)$ and $U \subset V$ convex and closed and $U \neq \varnothing$, the problem in (4.13) can be redefined as an unconstrained nondifferentiable minimization problem via*

$$\mathcal{F}(u) = F(u) + \iota_U(u) \quad \Leftrightarrow \quad \arg\min_{u \in V} \mathcal{F}(u) = \arg\min_{u \in U} F(u).$$

where ι stands for indicator function (see Definition 26, p. 75).

Definition 29 *A direction v is normal to $U \subset V$ at u if*

$$\langle w - u, v \rangle \leqslant 0, \ \forall w \in U$$



Definition 30 *The* normal cone *operator for U, for any $u \in U$ reads*

$$N_U(u) = \begin{cases} \{v \in V \ : \ \langle w - u, v \rangle \leqslant 0, \ \forall w \in U\} & \text{if } u \in U \\ \varnothing & \text{if } u \notin U \end{cases} \tag{4.14}$$

$N_U(u)$ convex and closed. Note that if $u \stackrel{\text{def}}{=} \Pi_U(v)$ then $v-u \in N_U(u),\ \forall v \in V$.

Lemma 13 *For ι_U indicator function of U and N_U as given in (4.14), we have*

$$\partial \iota_U = N_U.$$

Proof. By Corollary 3 (p. 90) and using that $\iota_U^*(v) = \sup_{z\in U}\langle z, v\rangle = \sigma_U(v)$ one has

$$\partial\iota_U(u) = \{v \in V :\ \iota_U(u) + \iota_U^*(v) = \langle u, v\rangle\} = \{v \in V :\ \iota_U(u) + \sup_{z\in U}\langle z, v\rangle = \langle u, v\rangle\}$$

If $u \notin U$, obviously $\partial\iota_U(u) = \varnothing$. Consider next that $u \in U$ in which case $\iota_U(u) = 0$. We obtain:

$$\partial\iota_U(u) = \{v \in V :\ \sup_{z\in U}\langle z, v\rangle = \langle u, v\rangle\} = \{v \in V :\ \langle z, v\rangle \leqslant \langle u, v\rangle\,,\ \forall z \in U\}$$ □

Theorem 55 *Let $F \in \Gamma_0(V)$ and $U \subset V$ be closed, convex and nonempty, and $\exists \tilde{u} \in \mathrm{dom}F \cap U$ where F is continuous. Then*

$$F(\hat{u}) = \min_{u\in U} F(u) \quad\Leftrightarrow\quad \delta F(\hat{u})(v-\hat{u}) \geqslant 0,\ \forall v \in U \quad\Leftrightarrow\quad 0 \in \partial F(\hat{u}) + N_U(\hat{u})$$

Proof. $\forall v \in U$ and $t \in]0, 1]$ we have $\hat{u} + t(v-\hat{u}) \in U$.

$$\begin{aligned} F(\hat{u}) = \min_{u\in U} F(u) &\Leftrightarrow F\big(\hat{u} + t(v-\hat{u})\big) \geqslant F(\hat{u}),\ \forall t \in]0, 1], \forall v \in U \\ &\Leftrightarrow \frac{F\big(\hat{u} + t(v-\hat{u})\big) - F(\hat{u})}{t} \geqslant 0, \forall t \in]0, 1], \forall v \in U \\ &\Leftrightarrow \inf_{t\in]0,1]} \frac{F\big(\hat{u} + t(v-\hat{u})\big) - F(\hat{u})}{t} \geqslant 0, \forall t \in]0, 1], \forall v \in U \\ &\Leftrightarrow \delta F(\hat{u})(v-\hat{u}) \geqslant 0,\ \forall v \in U. \end{aligned} \tag{4.15}$$

By Lemma 13 (p. 82) we get

$$\partial F(u) + N_U(u) = \partial F(u) + \partial\iota_U(u) = \partial\big(F(u) + \iota_U(u)\big).$$

Using Theorem 54 (p. 81) and Remark 24 (p. 81),

$$F(\hat{u}) = \min_{u\in U} F(u) \quad\Leftrightarrow\quad 0 \in \partial\big(F(\hat{u}) + \iota_U(\hat{u})\big) = \partial F(\hat{u}) + N_U(\hat{u})$$ □

Remark 25 *If $U = V$, the condition of Theorem 55 reads: $\delta F(\hat{u})(v) \geqslant 0,\ \forall v \in V$.*

4.3.3 Minimality conditions under explicit constraints

Consider problem (4.13), namely $F(\hat{u}) = \min_{u\in U} F(u)$, for $h_i \in \Gamma_0(\mathbb{R}^n)$, $1 \leqslant i \leqslant q$

$$U = \left\{ u \in \mathbb{R}^n :\ \begin{array}{rcll} \langle a_i, u\rangle & = & b_i, & 1 \leqslant i \leqslant p \quad (Au = b \in \mathbb{R}^p) \\ h_i(u) & \leqslant & 0, & 1 \leqslant i \leqslant q \end{array} \right\} \neq \varnothing \tag{4.16}$$

The set of active constraints at u reads $I(u) \stackrel{\text{def}}{=} \{i : h_i(u) = 0\}$

Theorem 56 ([7], Theorem 2.1.4 (p. 305) and Proposition 2.2.1 (p. 308).) *Consider U as in* (4.16) *and $F \in \Gamma_0(\mathbb{R}^n)$ Then the following statements are equivalent:*

1. *$\hat{u}$ solves the problem $F(\hat{u}) = \min_{u \in U} F(u)$;*
2. *$\exists \lambda \in \mathbb{R}^p,\ \mu \in \mathbb{R}^q$ such that*

$$0 \in \partial F(\hat{u}) + \sum_{i=1}^{p} \lambda_i a_i + \sum_{i=1}^{q} \mu_i \partial h_i(\hat{u}), \quad \mu_i \geqslant 0 \text{ and } \mu_i h_i(\hat{u}) = 0,\ 1 \leqslant i \leqslant q \tag{4.17}$$

3. *$0 \in \partial F(\hat{u}) + N_U(\hat{u})$ where*

$$N_U(u) = \big\{ A^T\lambda + \sum_{i \in I(u)} \mu_i z_i : \lambda \in \mathbb{R}^p, \mu_i \geqslant 0, z_i \in \partial h_i(u), \forall i \in I(u) \big\} \tag{4.18}$$

Remark 26 *Note that the condition in 2, namely* (4.17) *ensures that the normal cone at $\hat{u}$ w.r.t. U reads as in* (4.18) *which entails that the constraints are qualified.*

The existence of coefficients satisfying (4.17) is called Karush-Kuhn-Tucker (KKT) conditions. The requirement $\mu_i h_i(\hat{u}) = 0,\ 1 \leqslant i \leqslant q$ is called transversality or complementarity condition (or equivalently slackness).

$(\lambda, \mu) \in \mathbb{R}^p \times \mathbb{R}^q_+$ are called Lagrange multipliers. The Lagrangian function reads

$$L(u, \lambda, \mu) = F(u) + \sum_{i=1}^{p} \lambda_i (\langle a_i, u \rangle - b_i) + \sum_{i=1}^{q} \mu_i\, h_i(u) \tag{4.19}$$

Theorem 57 ([7], Theorem 4.4.3 (p. 337).) *We posit the assumption of Theorem 56. Then the following statements are equivalent:*

1. *$\hat{u}$ solves the problem $F(\hat{u}) = \min_{u \in U} F(u)$;*
2. *$\big(\hat{u},\ (\hat{\lambda}, \hat{\mu})\big)$ is a saddle point of L in* (4.19) *over $\mathbb{R}^n \times \big(\mathbb{R}^p \times \mathbb{R}^q_+\big)$.*

4.4 Some minimization methods

Descent direction $-d$: $\exists \rho > 0$ such that $F(u - \rho d) < F(u) \quad \Leftrightarrow \quad \langle g_u, d \rangle > 0, \quad \forall g_u \in \partial F(u)$

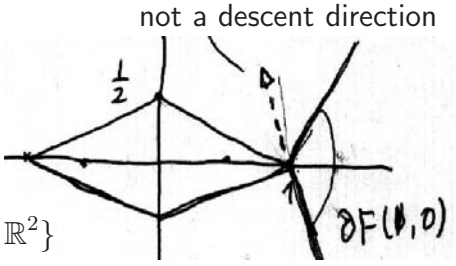
Note that $d = -g_u$ for some $g_u \in \partial F(u)$ is not necessarily a descent direction.

Example

$F(u) = |u[1]| + 2|u[2]|$

$$\delta F(1,0)(v) = \lim_{t \searrow 0} \frac{|1 + tv[1]| - 1 + 2|tv[2]|}{t} = v[1] + 2|v[2]|$$

$$\begin{aligned} \partial F(1,0) &= \{ g \in \mathbb{R}^2 \ \mid \ \langle g, v \rangle \leqslant \delta F(1,0)(v),\ \forall v \in \mathbb{R}^2 \} \\ &= \{ g \in \mathbb{R}^2 \ \mid \ g[1]v[1] + g[2]v[2] \leqslant v[1] + 2|v[2]|,\ \forall v \in \mathbb{R}^2 \} \\ &= \{1\} \times [-2, 2] \end{aligned}$$



The steepest descent method:

$$-d_k \in \arg \min_{\|d\|=1} \max_{g \in \partial F(u_k)} \langle g, d \rangle$$

is unstable and may converge to non-optimal points [7, 4].

Some difficulties inherent to the minimization of non-smooth functions

- Usually one cannot compute $\partial F(u)$, but only some elements $g \in \partial F(u)$.
- Stopping rule: $0 \in \partial F(u_k)$ is difficult to implement.
  An approximation like $\|g_k\| \leqslant \varepsilon$ can never be satisfied.
- $u_k \to \hat{u},\ g_{\hat{u}} \in \partial F(\hat{u}) \not\Rightarrow \exists\{g_k \in \partial F(u_k)\} \to g_{\hat{u}}$
- Minimizing a smooth approximation of F may generate large numerical errors while the properties of the minimizer of F and the one of its smoothed version are usually very different.
- Specialized minimization algorithms are needed.

4.4.1 Subgradient methods

Subgradient projections are significantly easier to implement than exact projections and have been used for solving a wide range of problems.

Subgradient algorithm with projection

Minimize $F : \mathbb{R}^n \to \mathbb{R}$ subject to the constraint set U (closed, convex, nonempty, possibly $= \mathbb{R}^n$)
For $k \in \mathbb{N}$

1. if $0 \in \partial F(u_k)$, stop (difficult to verify); else
2. obtain $\underline{g_k \in \partial F(u_k)}$ (easy in general);
3. (possible) line-search to find $\rho_k > 0$;
4. $u_{k+1} = \Pi_U \left(u_k - \rho_k \dfrac{g_k}{\|g_k\|} \right)$

Remark 27 $-d_k = -\dfrac{g_k}{\|g_k\|}$ is not necessarily a descent direction. This entails oscillations of the value $(F(u_k))_{k\in\mathbb{N}}$.

Theorem 58 ([75, 4]) *Suppose that $F : \mathbb{R}^n \to \mathbb{R}$ is convex and reaches its minimum at $\hat{u} \in \mathbb{R}^n$. If $(\rho_k)_{k\geqslant 0}$ satisfies*

$$\sum_k \rho_k = +\infty \quad \textit{and} \quad \sum_k \rho_k^2 < +\infty \qquad \Rightarrow \qquad \lim_{k\to\infty} u_k = \hat{u}. \tag{4.20}$$

By (4.20), ρ_k converges fast to 0 which means that the convergence of u_k is slow (sub-linear). (E.g., $\rho_k = 1/(k+1)$ for $k \in \mathbb{N}$.)

4.4.2 Gauss-Seidel method for separable non-differentiable terms

Consider that $u \in \mathbb{R}^n$ and

$$F(u) = \Psi(u) + \sum_{i=1}^{n} \beta_i \varphi_i(|u[i]|), \quad \beta_i \geqslant 0, \quad {\varphi_i}'(0^+) > 0, \quad \forall i \tag{4.21}$$

Algorithm.

$\forall k \in \mathbb{N}$, each iteration k has n steps:
$1 \leqslant i \leqslant n$, compute

$$\xi = \frac{\partial}{\partial u[i]} \Psi\left(u^{(k)}[1], \ldots, u^{(k)}[i-1],\ 0,\ \ u^{(k-1)}[i+1], \ldots, u^{(k-1)}[n]\right);$$

if $\quad |\xi| \leqslant \beta_i \varphi_i'(0^+) \quad \Rightarrow \quad u^{(k)}[i] = 0;$ $\qquad (\star)$

else $u[i]^{(k)}$ solves the equation on $\mathbb{R} \setminus \{0\}$

$$\beta_i \varphi_i'(u^{(k)}[i]) + \frac{\partial}{\partial u[i]} \Psi\left(u^{(k)}[1], \ldots, u^{(k)}[i-1],\ u^{(k)}[i],\ \ u^{(k-1)}[i+1], \ldots, u^{(k-1)}[n]\right) = 0,$$

where $\quad \operatorname{sign}\left(u^{(k)}[i]\right) = -\operatorname{sign}(\xi)$

The components located at kinks are found exactly in $(\star)$. Note that for $t \neq 0$, we have $\frac{d}{dt}\varphi_i(|t|) = \varphi_i'(|t|)\operatorname{sign}(t)$. In particular, for $\varphi_i(t) = |t|$, we have good simplifications: $\varphi_i'(0^+) = 1$ and $\varphi_i'(t) = \operatorname{sign}(t)$ if $t \neq 0$.

Theorem 59 *Let $F : \mathbb{R}^n \to \mathbb{R}$ in (4.21) be convex, proper and coercive, $\Psi \sim \mathcal{C}^1$ be convex, $\beta_i \geqslant 0$ and $\varphi_i : \mathbb{R}_+ \to \mathbb{R}$ be convex, $\mathcal{C}^1$ and $\varphi_i'(0) > 0$, $1 \leqslant i \leqslant n$.*

1. *Ψ is coercive;*

2. *$\forall \rho > 0,\ \exists \eta > 0$ such that $\|u\| \leqslant \rho,\ |t| \leqslant \rho \ \Rightarrow\ \Psi(u + te_i) - \Psi(u) \geqslant t\dfrac{\partial \Psi(u)}{\partial u_i} + t^2\eta,\ 1 \leqslant i \leqslant n.$*

Then the Gauss-Seidel *method given above converges to a minimizer $\hat{u}$ of F.*

For the proof under condition 1 see [6] and under condition 2 [66].

Comments

- Condition 2 is quite loose since it does not require that Ψ is globally coercive and is often easy to verify.
- The method cannot be extended to a non separable nonsmooth term.

Remark 28 *Note that an objective with nonsmooth data fidelity of the form (4.1), e.g. $F(u) = \sum_{i=1}^{m} \varphi(|\langle a_i, u\rangle - v_i|) + \beta\Psi(u)$, $\varphi'(0) > 0$, $m \leqslant n$, can be rewritten in the form of (4.21), provided that $\{a_i, 1 \leqslant i \leqslant m\}$ is linearly independent. Let $A \in \mathbb{R}^{n\times n}$ be an invertible matrix whose first m columns are a_i, $1 \leqslant i \leqslant m$. Setting $\tilde{v}_i = v_i$, $1 \leqslant i \leqslant m$ and $\tilde{v}_i = 0$, $m+1 \leqslant i \leqslant n$, we can apply a change of variables $z = Au - \tilde{v}$ and consider*

$$\mathcal{F}(z) = \sum_{i=1}^{n} \varphi(|z_i|) + \beta\Psi\left(A^{-1}(z - \tilde{v})\right)$$

4.4.3 Algorithms based on a reformulation of ℓ_1

For any $w \in \mathbb{R}^n$ we consider the decomposition

$$w = w^+ - w^- \quad \text{where} \quad w^+[i] := \max(w[i], 0) \geqslant 0, \quad w^-[i] = \max(-w[i], 0) \geqslant 0, \quad 1 \leqslant i \leqslant n$$

Alliney (1994) [76] exhibited that

$$\min \|w\|_1 \quad \Leftrightarrow \quad \min \sum_i \big(w^+[i] + w^-[i]\big) \quad \text{subject to} \quad w^+[i] \geqslant 0,\ w^-[i] \geqslant 0,\ 1 \leqslant i \leqslant n. \tag{4.22}$$

Full algorithms are developed in [77] in the cases given next where G is a finite differences operator (e.g., it can approximate anisotropic TV model) and $\mathbb{1}$ is the vector of all ones of appropriate size.

(a) $F(u) = \|Au - d\|_1 + \beta\|Gu\|_1$ subject to $u \geqslant 0$ (image pixels are non negative). By setting $h := Au - d$ and $w := \beta Gu$, the minimization of F can be written as a linear programming problem

$$\begin{aligned} & \min_{u,h^+,h^-,w^+,w^-} \mathbb{1}^T(h^+ + h^-) + \mathbb{1}^T(w^+ + w^-) \\ \text{subject to} \quad & Au - d = h^+ - h^- \\ & \beta Gu = w^+ - w^- \\ & u, h^+, h^-, w^+, w^- \geqslant 0 \end{aligned}$$

(b) $F(u) = \|Au-d\|_2^2+\beta\|Gu\|_1$ subject to $u \geqslant 0$. By setting $w := \beta Gu$, the problem can be written as quadratic programming problem:

$$\begin{aligned} & \min_{u,w^+,w^-} \|Au - d\|_2^2 + \mathbb{1}^T(w^+ + w^-) \\ \text{subject to} \quad & \beta Gu = w^+ - w^- \\ & u, w^+, w^- \geqslant 0 \end{aligned}$$

The solutions of these problems are characterized using the Lagrange multipliers. They are solved by interior point method with CG iterations, starting from a feasible point.

Chapter 5

Resolvent and Proximal operators

5.1 Maximal monotone and resolvent operators

5.1.1 Nonexpansive operators

For an overview – see [78] and the monograph [3]. Here V is a real Hilbert space.

Definition 31 (chap. 4.1 [3]) *An operator $T : V \to V$ is*

1. nonexpansive *if* $\|Tu - Tv\| \leqslant \|u - v\|, \quad \forall (u, v) \in V^2$
2. firmly nonexpansive *if one of the following equivalent conditions holds:*

$$\begin{array}{rcll} \|Tu - Tv\|^2 & \leqslant & \langle Tu - Tv, u - v\rangle & \\ \|Tu - Tv\|^2 & \leqslant & \|u - v\|^2 - \|(\mathrm{Id} - T)u - (\mathrm{Id} - T)v\|^2 & \forall (u, v) \in V^2 \end{array} \tag{5.1}$$

An obvious consequence is that

Lemma 14 ([78]) *$T : V \to V$ is firmly nonexpansive if and only if $(\mathrm{Id} - T)$ is firmly nonexpansive.*

If T is firmly nonexpansive, then it is nonexpansive; the converse, however, is false (e.g., $-\mathrm{Id}$). When T is Lipschitz continuous with a constant in $(0, 1)$, then T is referred to as a contraction.

The set of fixed points of T reads as

$$\mathrm{Fix}\, T := \{u \in V \ : \ u = Tu\}$$

Remark 29 *In many applications* $\mathrm{Fix}\, T$ *is not a singleton in which case the reached fixed point depends on the starting value $u^{(0)}$.*

Proposition 10 (chap. 4.3 [3]) *Let $T : V \to V$ is firmly nonexpansive. Then* $\mathrm{Fix}\, T$ *is convex, closed and*

$$\mathrm{Fix}\, T = \bigcap_{u \in V} \{v \in V \ : \ \langle v - Tu, u - Tu\rangle \leqslant 0\}$$

Definition 32 *Let $T : V \to V$ be nonexpansive. Then T is averaged with constant $\alpha \in]0, 1[$ (or α-avereged) if there exists a nonexpansive operator R such that $T := (1 - \alpha)\mathrm{Id} + \alpha R$.*

Note that if T is nonexpansive, it is not necessarily averaged; consider e.g., $T = -\mathrm{Id}$.

Proposition 11 *Let $T : V \to V$ be nonexpansive and let $\alpha \in]0,1[$. Then*

$$T \text{ is } \alpha\text{-averaged} \quad \Leftrightarrow \quad R := (1-\alpha)\mathrm{Id} + \alpha T \text{ is nonexpansive.}$$

Further, if T is α-averaged for $\alpha \in]0, \frac{1}{2}]$ then T is is firmly nonexpansive.

Corollary 2 *[chap. 4.4 [3]] $T : V \to V$ is firmly nonexpansive if and only if*

$$R := \frac{1}{2}\mathrm{Id} + \frac{1}{2}T$$

is $\frac{1}{2}$-averaged.

Then, assuming that $\mathrm{Fix}\, T \neq \varnothing$, on has

$$u = Ru \quad \Leftrightarrow \quad 2u = (\mathrm{Id} + T)u \quad \Leftrightarrow \quad u = Tu$$

Therefore, $\mathrm{Fix}\, R = \mathrm{Fix}\, T$: any fixed point of R can be obtained by the fixed point iteration (5.2) using the $\frac{1}{2}$-averaged operator in Lemma 60.

An operator $T : V \to V$ is asymptotically regular if $u_n - Tu_n \to 0$ as $n \to \infty$. This property does not imply convergence, even boundedness cannot be guaranteed. However, it is satisfied by any firmly nonexpansive operator. This fact plays an important role in the proof of the theorem below which is a generalization of the celebrated Opial's theorem (1976).

Theorem 60 *[chap. 5.2 [3]] Let T be firmly nonexpansive with $\mathrm{Fix}\, T \neq \varnothing$. Then*

(a) *the sequence (known as Picard iterates)*

$$u^{k+1} = T(u^k) \tag{5.2}$$

converges weakly to a point in $\mathrm{Fix}\, T$.

(b) *Let $(\lambda_k)_{k\in\mathbb{N}}$ be a sequence in $[0,2]$ such that $\sum_{k\in\mathbb{N}} \lambda_k(2-\lambda_k) = +\infty$. The sequence*

$$u^{k+1} = u^k + \lambda_k \left(T(u^k) - u^k\right) \tag{5.3}$$

converges weakly to a point in $\mathrm{Fix}\, T$.

5.1.2 Maximally monotone operators

Firmly nonexpansive mappings are closely related to maximally monotone operators. Recall that a mapping A on V is monotone if for any $u_1, u_2 \in V$ one has $\langle g_2 - g_1, u_2 - u_1\rangle \geqslant 0, \ \ \forall g_1 \in A(u_1), \ \ g_2 \in A(u_2)$. (This should be compared with Theorem 50, p. 78).

Definition 33 *An operator A on V is maximally monotone if there is no monotone operator that properly contains it (i.e., no enlargement of its graph is possible without destroying its monotonicity).*

Examples of maximally monotone operators: continuous linear monotone operators, (sub)differential operators of functions that are convex, lower semicontinuous, and proper (i.e. all functions in $\Gamma_0(V)$).

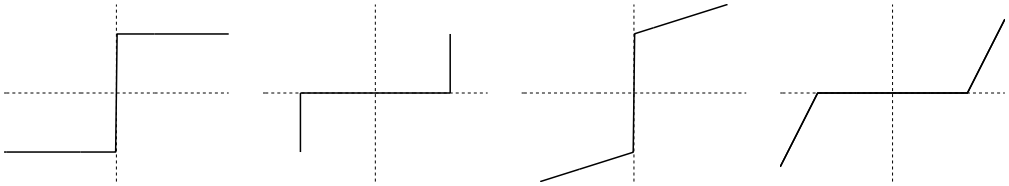
Definition 34 *Given a maximally monotone operator $A : V \rightrightarrows V$, the associated resolvent operator is*

$$J_A := (\mathrm{Id} + A)^{-1}$$

Example 10 *We consider the set-valued function*

$$A(u) := \begin{cases} \{-1\} & u < 0 \\ [-1, 1] & u = 0 \\ \{1\} & u > 0 \end{cases}$$

which is the subdifferential of $u \mapsto |u|$.



Correspondences between maximally monotone and firmly nonexpansive operators:

Theorem 61 *(Minty 1962)*

- *$T : V \to V$ is firmly nonexpansive $\quad \Rightarrow \quad$ $B := T^{-1} - \mathrm{Id}$ is maximally monotone (and $J_B = T$).*
- *$A : V \rightrightarrows V$ is maximally monotone $\quad \Rightarrow \quad$ J_A is firmly nonexpansive (and $A = J_A^{-1} - \mathrm{Id}$).*

Based on the two classes and dualizing one gets the resolvent identity (see [78, p. 123])

$$\mathrm{Id} = J_A + J_{A^{-1}}$$

A consequence of Theorem 61 is that an operator T is firmly nonexpansive if and only if it is the resolvent of a maximal monotone operator A, i.e. $T = J_A$ (see, e.g., [72, sec. 12]).

5.1.3 Resolvent operator

We can identify A with the (sub)gradient of a function in $\Gamma_0(V)$. From Fermat's rule (Theorem 54, p. 81), minimizing $F \in \Gamma_0(V)$ amounts to solving the inclusion

$$0 \in \partial F(u) \quad \Leftrightarrow \quad 0 \in \gamma \partial F(u), \ \ \forall\, \gamma > 0 \quad \Leftrightarrow \quad u \in u + \gamma \partial F(u) = (\mathrm{Id} + \gamma \partial F)(u), \ \ \forall\, \gamma > 0$$

or equivalently, to finding a solution to the fixed point equation

$$\boxed{u = (\mathrm{Id} + \gamma \partial F)^{-1}(u)} \tag{5.4}$$

where $(\mathrm{Id} + \gamma \partial F)^{-1}$ is the resolvent operator associated to ∂F, $\gamma > 0$ is a stepsize. This is a fundamental tool for finding the root of any maximal monotone operator [79, 80], such as e.g. the subdifferential of a convex function.

5.2 Moreau's conjugacy and proximal calculus

5.2.1 Conjugate dual functions theorem

Theorem 62 *[p. 277, [72]] Let $F \in \Gamma_0(V)$ and $F^\star : V^\star \to \mathbb{R}$ be its convex conjugate (see (1.14), p. 17). Then*

$$v \in \partial F(u) \quad \Leftrightarrow \quad F(u) + F^\star(v) = \langle u, v\rangle \quad \Leftrightarrow \quad u \in \partial F^\star(v) \tag{5.5}$$

Proof. By (4.7) and using that $F = F^{\star\star}$ (see Theorem 9, p. 18):

$$\begin{aligned}
\underbrace{v \in \partial F(u)} \quad &\Leftrightarrow \quad F(z) \geqslant F(u) + \langle v, z-u\rangle\,,\ \forall z \in V \\
&\Leftrightarrow \quad \langle v, z\rangle - F(z) \leqslant \langle v, u\rangle - F(u),\ \ \forall z \in V \\
&\Leftrightarrow \quad F^\star(v) = \sup_{z\in V}\Big\{\,\langle v, z\rangle - F(z)\Big\} = \langle v, u\rangle - F(u), \\
&\Leftrightarrow \quad F^\star(v) = \langle v, u\rangle - F(u) \\
\underbrace{\Leftrightarrow} \quad &\underbrace{F(u) + F^\star(v) = \langle u, v\rangle} \\
&\Leftrightarrow \quad F^\star(v) + F^{\star\star}(u) = \langle u, v\rangle \quad \Leftrightarrow \quad \underbrace{u \in \partial F^\star(v)}
\end{aligned}$$

Corollary 3 *We posit the conditions of Theorem 62. Then*

$$\boxed{\partial F(u) = \{v \in V \ :\ F(u) + F^*(v) = \langle u, v\rangle\}\ .}$$

5.2.2 Proximity operators

Proximal operators were inaugurated by Moreau in 1962 [81] as a generalization of convex projection operators.

Proposition 12 (p. 278, [72].) *Let $f \in \Gamma_0(V)$. For any $v \in V$ and $\gamma > 0$, the function $\mathcal{P}_f : V \to \mathbb{R}$ below*

$$\underline{\mathcal{P}_{\gamma f}(u) = \frac{1}{2\gamma}\|v-u\|^2 + f(u)} \tag{5.6}$$

admits a unique minimizer.

Definition 35 *The unique minimizer in Proposition 12 is denoted by*

$$\underline{\operatorname{prox}_{\gamma f} v = \arg\min_{u\in V} \mathcal{P}_{\gamma f}(u).} \tag{5.7}$$

$\operatorname{prox}_{\gamma f} v$ *is the proximal point of v with respect to γf and* $\underline{\operatorname{prox}_{\gamma f}}$ *is the* proximity operator for γf.

Remark 30 *From (5.6) and (5.7), the optimality conditions (Theorem 54, p. 81) yield*

$$\begin{aligned}
\underline{\hat{u} = \operatorname{prox}_{\gamma f} v} \ &\Leftrightarrow\ 0 \in \partial\mathcal{P}_{\gamma f}(\hat{u}) \ \Leftrightarrow\ 0 \in \hat{u} - v + \gamma\partial f(\hat{u}) \ \Leftrightarrow\ v - \hat{u} \in \gamma\partial f(\hat{u}) \\
&\Leftrightarrow\ v \in (\mathrm{Id} + \gamma\partial f)(\hat{u}) \ \Leftrightarrow\ \underline{\hat{u} = (\mathrm{Id} + \gamma\partial f)^{-1}(v)}
\end{aligned} \tag{5.8}$$

where the last equality comes from the fact that $\hat{u}$ is the unique minimizer of $\mathcal{P}_{\gamma f}$ (Proposition 12). Hence $\operatorname{prox}_{\gamma f}$ (the proximal operator of γf) coincides with the resolvent operator associated to ∂f:

$$\boxed{(I + \gamma\partial f)^{-1} = \operatorname{prox}_{\gamma f}} \tag{5.9}$$

see (5.4) *(p. 89). Observe that* $\arg\min_u\{\frac{1}{2\gamma}\|u-v\|^2+f(u)\}=\arg\min_u\{\frac{1}{2}\|u-v\|^2+\gamma f(u)\}$. *From* (5.8) *one also has*

$$v-\mathrm{prox}_{\gamma f}\,v\in\gamma\partial f(\mathrm{prox}_{\gamma f}\,v) \tag{5.10}$$

Example 11 Several examples of prox_f for $f\in\Gamma_0(V)$ – see [82]:

1. $f(u)=\langle a,u\rangle-\beta$, $a\in V$, $\beta\in\mathbb{R}$. Then $\mathrm{prox}_f\,v=v-a$ (translation)
2. $U\subset V$ is closed, convex and $\neq\varnothing$. $f(u)=\boldsymbol{\iota}_U(u)$ then $\mathrm{prox}_{\boldsymbol{\iota}_U}\,v=\Pi_U(v)$
3. $f(u)=\beta\|u\|_2^2$, $\beta\geqslant 0$. Then $\mathrm{prox}_f\,v=\dfrac{1}{1+2\beta}\,v$
4. $f(u)=\beta\|u\|_2$, $\beta\geqslant 0$. Then $\mathrm{prox}_f\,v=\max\{\|v\|_2-\beta,0\}\dfrac{v}{\|v\|_2}$ (see Lemma 15, p. 93)
5. $f(u)=\Psi(-u)$ then $\mathrm{prox}_f v=-\mathrm{prox}_\Psi(-v)$
6. $f(u)=\Psi(u-z)$, $z\in V$. Then $\mathrm{prox}_f\,v=z+\mathrm{prox}_\Psi\,(v-z)$
7. $f(u)=\Psi(u/\beta)$, $\beta\in\mathbb{R}\setminus\{0\}$. Then $\mathrm{prox}_f\,v=\beta\mathrm{prox}_{\Psi/\beta^2}\,(v/\beta)$.
8. $f(u)=\displaystyle\sum_{i=1}^n\beta_i\big|u[i]-z[i]\big|$, $\beta_i>0$, $\forall i$. Then $\big(\mathrm{prox}_f\,v\big)[i]=z[i]+T^{\beta_i}\big(v[i]-z[i]\big)$, $\forall i$, where

$$T^\beta(t)\stackrel{\text{def}}{=}\begin{cases}0 & \text{if } |t|<\beta\\ t-\beta\mathrm{sign}(t) & \text{otherwise}\end{cases}\qquad T^\beta \text{ is a soft shrinkage operator (Example 8 on p. 74).}$$

Definition 36 *The Moreau envelope or Moreau-Yoshida regularization is given by the infimal convolution:*

$${}^\gamma f(v):=\inf_{u\in V}\left\{\frac{1}{2\gamma}\|v-u\|^2+f(u)\right\}$$

5.2.3 Proximal decomposition

Relationship with the classical Projection Theorem:

Theorem 63 ([72], p. 280.) *Let* $f\in\Gamma_0(V)$ *and* $f^\star\in\Gamma_0(V)$ *be its convex conjugate. For any* $u,v,w\in V$ *we have the equivalence*

$$w=u+v\ \textit{and}\ f(u)+f^\star(v)=\langle u,v\rangle\quad\Leftrightarrow\quad u=\mathrm{prox}_f\,w,\ v=\mathrm{prox}_{f^\star}\,w.$$

Meaning: Link with the Projection theorem; see item 2 in Example 11, p. 91.

Proof. Two parts.
($\Rightarrow$) By the definition of the convex conjugate[1] $f^\star(v)=\sup_{z\in V}\big(\langle z,v\rangle-f(z)\big)$, we have

$$\langle u,v\rangle-f(u)=f^\star(v)\geqslant\langle z,v\rangle-f(z),\ \ \forall z\in V\quad\Rightarrow\quad f(z)-f(u)\geqslant\langle z-u,v\rangle,\ \ \forall z\in V \tag{5.11}$$

[1]This definition was already given in (1.14) on p. 17.

$$\begin{aligned}
\underbrace{\mathcal{P}_f(z)-\mathcal{P}_f(u)} &= \frac{1}{2}\|z-w\|^2+f(z)-\frac{1}{2}\|u-w\|^2-f(u),\ \ \forall z\in V\\
&= \frac{1}{2}\|z-u-v\|^2-\frac{1}{2}\|v\|^2+f(z)-f(u),\ \ \forall z\in V \qquad \text{set } v=w-u\\
&\underbrace{\geqslant}\ \frac{1}{2}\|z-u-v\|^2-\frac{1}{2}\|v\|^2+\langle z-u,v\rangle,\ \ \forall z\in V \qquad \text{(by (5.11))}\\
&= \frac{1}{2}\|z-u\|^2-\langle z-u,v\rangle+\frac{1}{2}\|v\|^2-\frac{1}{2}\|v\|^2+\langle z-u,v\rangle,\ \ \forall z\in V\\
&= \underbrace{\frac{1}{2}\|z-u\|^2\geqslant 0},\ \ \forall z\in V.
\end{aligned}$$

It follows that u is the unique point such that $\mathcal{P}_f(u)<\mathcal{P}_f(z)$, $\forall z\in V,\ z\neq u$. Hence $u=\operatorname{prox}_f w$.

In a similar way one establishes that $v=\operatorname{prox}_{f^\star} w$.

($\Leftarrow$) Since $u=\operatorname{prox}_f w$ for $w\in V$, we know that u is the unique minimizer of $\mathcal{P}_f$ on V for this w (see Proposition 12 on p. 90). Hence

$$\begin{aligned}
\mathcal{P}_f\big(t(z-u)+u\big) &= \frac{1}{2}\|t(z-u)+u-w\|^2+f\big(t(z-u)+u\big)\\
&\geqslant \frac{1}{2}\|u-w\|^2+f(u)=\mathcal{P}_f(u)=\frac{1}{2}\|\widetilde{v}\|^2+f(u),\ \ \forall z\in V,\ t\in\mathbb{R}
\end{aligned}$$

where we set

$$\widetilde{v}=w-u$$

Rearranging the obtained inequality yields

$$\frac{1}{2}\|\widetilde{v}\|^2-\frac{1}{2}\|t(z-u)-\widetilde{v}\|^2+f(u)\leqslant f\big(t(z-u)+u\big),\ \ \forall z\in V,\ t\in\mathbb{R}.$$

Noticing that $\frac{1}{2}\|\widetilde{v}\|^2-\frac{1}{2}\|t(z-u)-\widetilde{v}\|^2=-\frac{1}{2}t^2\|z-u\|^2+t\,\langle z-u,\widetilde{v}\rangle$, the last inequality entails that

$$-\frac{1}{2}t^2\|z-u\|^2+t\,\langle z-u,\widetilde{v}\rangle+f(u)\leqslant f\big(t(z-u)+u\big),\ \ \forall z\in V,\ t\in\mathbb{R}. \tag{5.12}$$

Remind that $t(z-u)+u=tz+(1-t)u$. Since f is convex, for $t\in]0,1[$ we have

$$f\big(t(z-u)+u\big)\leqslant tf(z)+(1-t)f(u)=f(u)+t\big(f(z)-f(u)\big),\ \ \forall t\in]0,1[,\ \forall z\in V.$$

Inserting the last inequality into (5.12) leads to

$$\begin{aligned}
&-\frac{1}{2}t^2\|z-u\|^2+t\,\langle z-u,\widetilde{v}\rangle+f(u)\leqslant t\big(f(z)-f(u)\big)+f(u),\quad \forall t\in]0,1[,\ \forall z\in V,\\
\Leftrightarrow\ &-\frac{1}{2}t^2\|z-u\|^2+t\,\langle z-u,\widetilde{v}\rangle\leqslant t\big(f(z)-f(u)\big),\quad \forall t\in]0,1[,\ \forall z\in V,\\
\Leftrightarrow\ &-\frac{1}{2}t\|z-u\|^2+\langle z-u,\widetilde{v}\rangle\leqslant f(z)-f(u),\quad \forall t\in]0,1[,\ \forall z\in V,\\
\Leftrightarrow\ &\langle z,\widetilde{v}\rangle-f(z)\leqslant\langle u,\widetilde{v}\rangle-f(u)+\frac{1}{2}t\|z-u\|^2,\quad \forall t\in]0,1[,\ \forall z\in V.
\end{aligned}$$

Letting $t\to 0$ yields

$$\langle z,\widetilde{v}\rangle-f(z)\leqslant\langle u,\widetilde{v}\rangle-f(u),\ \ \forall z\in V.$$

Then

$$\sup_{z\in V}\big(\,\langle z,\widetilde{v}\rangle-f(z)\big)=f^\star(\widetilde{v})=\langle u,\widetilde{v}\rangle-f(u)$$

Taking $v=\widetilde{v}=w-u$ achieves the proof. □

In words, it was proven that $\forall w\in V$, for a given $f\in\Gamma_0(V)$, we have a unique decomposition

$$\boxed{w = \operatorname{prox}_f w + \operatorname{prox}_{f^\star} w}$$

which provides a tool to compute prox.

5.2.4 Computing the prox of a function: the case of $\|\cdot\|_2$

Comparing (5.4) (p. 89) with (5.8) (p. 90) shows that the prox of a function yields its minimizer. By (5.6) and Definition 35, computing the prox of a function generally requires to solve a minimization problem. Finding closed-form prox operators for important classes of functions f is a lively area of research.

The result (4) in Example 11 was obtained in 2009 and is actually used in many algorithms.

Lemma 15 ([83], p. 577) *For $\alpha > 0$, $\beta > 0$ and $v \in \mathbb{R}^n$, where n is any positive integer, set*

$$f(u) = \frac{\alpha}{2}\|u - v\|_2^2 + \beta\|u\|_2$$

The unique minimizer of f is given by

$$\hat{u} = \max\left\{\|v\|_2 - \frac{\beta}{\alpha},\, 0\right\}\frac{v}{\|v\|_2} \tag{5.13}$$

where the convention $0 \cdot (0/0) = 0$ is followed.

Proof. Since f is strictly convex, bounded below, and coercive, it has unique minimizer $\hat{u}$. Using Proposition 9 (p. 78) with $G = \mathrm{Id}$, the subdifferential of f reads

$$\partial f(u) = \alpha(u - v) + \beta \begin{cases} \dfrac{u}{\|u\|_2} & \text{if} \quad u \neq 0 \\ \\ \left\{h \in \mathbb{R}^n \;\middle|\; \|h\|_2 \leqslant 1\right\} & \text{if} \quad u = 0 \end{cases}$$

According to the optimality conditions, $0 \in \partial f(\hat{u})$, hence

$$\begin{cases} \alpha(\hat{u} - v) + \beta\dfrac{\hat{u}}{\|\hat{u}\|_2} = 0 & \text{if} \quad \hat{u} \neq 0 \\ \\ \alpha v \in \left\{h \in \mathbb{R}^n \;\middle|\; \|h\|_2 \leqslant \beta\right\} & \text{if} \quad \hat{u} = 0 \end{cases}$$

From the second condition, it is obvious that

$$\hat{u} = 0 \quad \Leftrightarrow \quad \|v\|_2 \leqslant \frac{\beta}{\alpha} \tag{5.14}$$

Consider next that $\hat{u} \neq 0$ in which case $\|v\|_2 > \dfrac{\beta}{\alpha}$. One has

$$\hat{u}\left(1 + \frac{\beta}{\alpha\|\hat{u}\|_2}\right) = v$$

Noticing that $\left(1 + \frac{\beta}{\alpha\|\hat{u}\|_2}\right) > 0$, we extract [2] $\hat{u}$ from the equation above. Thus

$$\hat{u} = \left(1 - \frac{\beta}{\alpha\|v\|_2}\right)v = \left(\|v\|_2 - \frac{\beta}{\alpha}\right)\frac{v}{\|v\|_2} \quad \text{and} \quad \|v\|_2 - \frac{\beta}{\alpha} > 0 \tag{5.15}$$

[2] One derives $\|u\|_2(1 + \frac{\beta}{\alpha\|\hat{u}\|_2}) = \|v\|_2$, hence $\|u\|_2 = \|v\|_2 - \frac{\beta}{\alpha}$. Therefore

$$v = \hat{u}\left(1 + \frac{\beta}{\alpha\|v\|_2 - \beta}\right) = \hat{u}\frac{\alpha\|v\|_2}{\alpha\|v\|_2 - \beta} \quad \text{where} \quad \alpha\|v\|_2 > \beta$$

Combining (5.14) and (5.15) leads to (5.13). □

Remark 31 If $n = 1$ in Lemma 15, that is $f(u) = \frac{\alpha}{2}(u - v)^2 + \beta|u|, \quad u \in \mathbb{R}$ then (5.13) amounts to the soft-shrinkage operator in Example 8, p. 74.

Remark 32 *For $i = 1, \cdots, k$ and $u_i \in \mathbb{R}^{n_i}$ suppose that $f(u) = \sum_{i=1}^{k} f_i(u_i)$. Then*

$$\text{prox}_{\gamma f}(v) = (\text{prox}_{\gamma f_1}(v_1), \ldots, \text{prox}_{\gamma f_k}(v_k))$$

which offers a major computational simplification.

5.2.5 Contraction properties

Proposition 13 ([74], Lemma 2.4.) *For $f \in \Gamma_0(V)$, prox_f and $(\text{Id} - \text{prox}_f)$ are* firmly nonexapnsive

Proof. From the definition of the subdifferential of f in (4.7) (p. 77)

$$\partial f(u) = \{g \in V : \langle g, z - u\rangle + f(u) \leqslant f(z), \ \forall z \in V\}$$

and from (5.10) $v - \text{prox}_f v \in \partial f(\text{prox}_f v)$ for any $v \in V$. Hence

$$\left\langle v - \text{prox}_f v, \text{prox}_f z - \text{prox}_f v \right\rangle + f(\text{prox}_f v) \leqslant f(\text{prox}_f z)$$

$$\left\langle z - \text{prox}_f z, \text{prox}_f v - \text{prox}_f z \right\rangle + f(\text{prox}_f z) \leqslant f(\text{prox}_f v)$$

Adding these two inequalities yields

$$\left\langle v - \text{prox}_f v, \text{prox}_f z - \text{prox}_f v \right\rangle - \left\langle z - \text{prox}_f z, \text{prox}_f z - \text{prox}_f v \right\rangle \leqslant 0$$

$$\left\langle v - z + \text{prox}_f z - \text{prox}_f v, \text{prox}_f z - \text{prox}_f v \right\rangle \leqslant 0$$

and finally

$$\|\text{prox}_f z - \text{prox}_f v\|^2 \leqslant \left\langle z - v, \text{prox}_f z - \text{prox}_f v \right\rangle$$

The conclusion follows from Definition 31. □

Since $f \in \Gamma_0(V)$, ∂f and $\gamma \partial f$ are maximally monotone. By Proposition 13, its resolvent operator, see (5.9) (p. 90) is firmly nonexpansive. Thus the proximal point algorithm given next converges by Theorem 60.

Proximal Point Algorithm

Initialization: u^0, $\gamma > 0$. Iterates: for $k \in \mathbb{N}$

$$u^{k+1} = \text{prox}_{\gamma f}(u^k) = \arg\min_u \left\{ \frac{1}{2\gamma}\|u - u^k\|^2 + f(u) \right\} = (\text{Id} + \gamma \partial f)^{-1}(u^k)$$

Often, $(\text{Id} + \gamma \partial f)^{-1}$ cannot be calculated in closed-form.

Lemma 16 ([72]) *Let $f \in \Gamma_0(V)$. For $w \in V$ and $z \in V$ set*

$$\begin{aligned} u &= \operatorname{prox}_f w \qquad u' = \operatorname{prox}_f z \\ v &= \operatorname{prox}_{f^\star} w \qquad v' = \operatorname{prox}_{f^\star} z \end{aligned} \tag{5.16}$$

The operator prox_f *is* monotonous *in the sense that*

$$\langle u - u', v - v' \rangle \geqslant 0$$

The property of Lemma 16 reads as

$$\left\langle \operatorname{prox}_f w - \operatorname{prox}_f z \,,\, \operatorname{prox}_{f^\star} w - \operatorname{prox}_{f^\star} z \right\rangle \geqslant 0, \quad \forall w \in V,\ \forall z \in V$$

Proof. By Theorem 63 one has

$$f(u) + f^\star(v) = \langle u, v \rangle \quad \text{and} \quad f(u') + f^\star(v') = \langle u', v' \rangle \tag{5.17}$$

and that by Fenchel-Young inequality (1.15) (p. 17)

$$f^\star(v') \geqslant \langle u, v' \rangle - f(u) \quad \text{and} \quad f(u') \geqslant \langle v, u' \rangle - f^\star(v) \tag{5.18}$$

we obtain

$$\begin{array}{rcl} \langle u - u', v - v' \rangle & = & \langle u, v \rangle + \langle u', v' \rangle - \langle u, v' \rangle - \langle u', v \rangle \\ \text{use (5.17)} & = & f(u) + f^\star(v) + \underbrace{f(u')} + \underbrace{f^\star(v')} - \langle u, v' \rangle - \langle u', v \rangle \\ \text{use (5.18)} & \underbrace{\geqslant} & f(u) + f^\star(v) + \langle u, v' \rangle - f(u) + \langle v, u' \rangle - f^\star(v) - \langle u, v' \rangle - \langle u', v \rangle = \underline{0}. \end{array}$$

The proof is complete. □

Proposition 14 ([72]) *For any $w, z \in V$*

$$\|\operatorname{prox}_f w - \operatorname{prox}_f z\| \leqslant \|w - z\| \tag{5.19}$$

so that $\operatorname{prox}_f : V \to V$ *is continuous.*

Proof. We use the notations introduced in (5.16).

$$\begin{array}{rcl} \|w - z\|^2 & = & \|u + v - u' - v'\|^2 \\ & = & \|u - u'\|^2 + \|v - v'\|^2 + 2\,\langle u - u', v - v' \rangle \\ \text{use Lemma 16} & \geqslant & \|u - u'\|^2 \end{array}$$

□

Moreover, equality in (5.19) holds if and only if $\operatorname{prox}_{f^\star} w = \operatorname{prox}_{f^\star} z$.
The proposition states a kind of distance reduction.

5.3 A proximal algorithm for the ROF functional

5.3.1 Discrete approximations of the operators ∇ and div

If $u \in V = \mathbb{R}^{M\times N}$, its gradient $\nabla u \in W := V \times V$ is given by [1] forward Euler discretization

$$(\nabla u)_{ij} = ((\nabla_x u)_{ij}, (\nabla_y u)_{ij})$$

where

$$(\nabla_y u)_{ij} = \begin{cases} u_{i+1,j} - u_{i,j} & i < M \\ 0 & i = M \end{cases} \qquad (\nabla_x u)_{ij} = \begin{cases} u_{i,j+1} - u_{i,j} & j < N \\ 0 & j = N \end{cases}$$

Other choices of discretization are possible. The scalar product in W is defined by

$$\langle \zeta, \xi \rangle := \sum_{ij} \zeta_{ij}^x \xi_{ij}^x + \zeta_{ij}^y \xi_{ij}^y$$

where $\zeta = (\zeta^x, \zeta^y)$ and $\xi = (\xi^x, \xi^y)$. By analogy with the continuous setting, one must have $\mathrm{div} = -\nabla^*$ (where ∇^* is the adjoint of ∇). Thus for any $w \in W$ and $u \in V$ it holds that

$$\langle \nabla u, \xi \rangle_W = - \langle u, \mathrm{div}\xi \rangle_V$$

One easily deduces that

$$(\mathrm{div}\xi)_{ij} = \begin{cases} \xi_{i,j}^y - \xi_{i,j+1} & 1 < i < M \\ \xi_{ij}^y & i = 1 \\ -\xi_{i-1,j}^y & i = M \end{cases} + \begin{cases} \xi_{i,j}^x - \xi_{i,j-1} & 1 < j < N \\ \xi_{ij}^x & j = 1 \\ -\xi_{i,j-1}^x & j = N \end{cases}$$

For what follows, it is convenient to mention that [1]

$$\|\nabla\|^2 = \|\mathrm{div}\|^2 \lesssim 8. \tag{5.20}$$

5.3.2 ℓ_2-TV minimization (Chambolle 2004, [1])

Minimize $F(u) = \frac{1}{2}\|u - v\|_2^2 + \beta \mathrm{TV}(u)$ where

$$\mathrm{TV}(u) = \sup \Big\{ \int u(x) \mathrm{div}\xi(x) dx : \xi \in \mathcal{C}_c^1(\Omega, \mathbb{R}^2), |\xi(x)| \leqslant 1, \forall x \in \Omega \Big\} \tag{5.21}$$

Let $u \in \mathbb{R}^{n\times n}$ and $v \in \mathbb{R}^{n\times n}$. Define

$$K = \big\{ \mathrm{div}\xi \ \mid \ \xi \in \big(\mathbb{R}^{n\times n}\big)^2, \|\xi_{i,j}\|_2 \leqslant 1, 1 \leqslant i,j \leqslant n \big\} \subset \mathbb{R}^{n\times n} \tag{5.22}$$

where $\|\xi_{i,j}\|_2 = \sqrt{(\xi_{ij}^x)^2 + (\xi_{ij}^y)^2}$. Let us emphasize that K is convex and closed[3]. The discrete equivalent of the TV regularization in (5.21) reads

$$\mathrm{TV}(u) = \sup \big\{ \langle u, w \rangle : w \in K \big\} = \sup_{w\in K} \langle u, w \rangle = \sigma_K(u)$$

Its convex conjugate (see (1.14), p. 17)

$$\mathrm{TV}^*(w) = \sup_u \left(\langle u, w \rangle - \mathrm{TV}(u) \right) = \iota_K(w)$$

[3]The set K is composed out of "oscillatory" components.

where we use the fact that $\sigma_K^* = \iota_K$, see (4.3) (p. 75). Since $\mathrm{TV} \in \Gamma_0(\mathbb{R}^{n\times n})$, using Fenchel-Moreau Theorem 9 (p. 18), $\mathrm{TV}(u) = \mathrm{TV}^{**}(u)$.

Using that

$$\partial F(u) = u - v + \beta \partial \mathrm{TV}(u),$$

Theorem 54 (p. 81) implies that F has a minimum at $\hat{u}$ if and only if

$$0 \in \partial F(\hat{u}) \quad \Leftrightarrow \quad 0 \in \hat{u} - v + \beta \partial \mathrm{TV}(\hat{u}) \quad \Leftrightarrow \quad \frac{v-\hat{u}}{\beta} \in \partial \mathrm{TV}(\hat{u})$$

$$\text{(Use Theorem 62 p. 90)} \Leftrightarrow \quad \hat{u} \in \partial \mathrm{TV}^*\left(\frac{v-\hat{u}}{\beta}\right) \quad \Leftrightarrow \quad 0 \in -\frac{\hat{u}}{\beta} + \frac{1}{\beta}\partial \mathrm{TV}^*(\hat{w}) \quad \text{where} \quad \boxed{\hat{w} \stackrel{\text{def}}{=} \frac{v-\hat{u}}{\beta}}$$

$$\Leftrightarrow \quad \frac{v}{\beta} \in \underline{\frac{v}{\beta} - \frac{\hat{u}}{\beta}} + \frac{1}{\beta}\partial \mathrm{TV}^*(\hat{w}) \quad \Leftrightarrow \quad \frac{v}{\beta} \in \hat{w} + \frac{1}{\beta}\partial \mathrm{TV}^*(\hat{w})$$

$$\Leftrightarrow \quad 0 \in \hat{w} - \frac{v}{\beta} + \frac{1}{\beta}\partial \mathrm{TV}^*(\hat{w})$$

$\Leftrightarrow$ $\hat{w}$ minimizes the function $\mathcal{F}$ below

$$\mathcal{F}(\hat{w}) = \min_{w\in\mathbb{R}^{n\times n}} \left\{ \frac{1}{2}\left\| w - \frac{v}{\beta}\right\|^2 + \frac{1}{\beta}\mathrm{TV}^*(w) \right\} = \min_{w\in K} \left\{ \frac{1}{2}\left\| w - \frac{v}{\beta}\right\|_2^2 \right\}$$

$\Leftrightarrow$

$$\hat{w} = \Pi_K\left(\frac{v}{\beta}\right) \quad \Leftrightarrow \quad \hat{u} = v - \beta \Pi_K\left(\frac{v}{\beta}\right) \tag{5.23}$$

where Π_K is the orthogonal projection onto K. To calculate $\hat{u}$, we need to find $\mathrm{Div}\hat{\xi}$ that solves the nonlinear projection onto Π_K:

$$\min\left\{ \left\| \mathrm{Div}\,\xi - \frac{v}{\beta}\right\|^2 : \ \xi \in (\mathbb{R}^{n\times n})^2, \quad \|\xi[i,j]\|_2^2 - 1 \leqslant 0, \quad 1 \leqslant i,j \leqslant n \right\}$$

KT optimality conditions (Theorem 42, p. 63) $\Rightarrow$ see [1] for details.

Algorithm:

$$\xi_{k+1}[i,j] = \frac{\xi_k[i,j] + \rho\Big(\nabla\big(\mathrm{Div}\ \xi_k - v/\beta\big)\Big)[i,j]}{1 + \rho\Big|\Big(\nabla\big(\mathrm{Div}\ \xi_k - v/\beta\big)\Big)[i,j]\Big|}$$

Theorem 64 (p. 91, [1]) *For $\rho \leqslant 1/8$, convergence of* $\mathrm{Div}\ \xi_k$ *to $\hat{w} = \Pi_K(v/\beta)$ as $k \to \infty$ and*

$$\hat{u} = v - \beta\hat{w}$$

This algorithm amounts to a special instance of Bermùdez and Moreno's Algorithm (1979), see Aujol 2009 [55]. Hence convergence for $\rho \leqslant 1/4$.

Chapter 6

Splitting and penalty methods

Different splittings lead to different implementations of the proximal gradient method for the same original problem.

6.1 Proximal algorithms

Let V be a Hilbert space[1] and $\Psi \in \Gamma_0(V)$ and $\Phi \in \Gamma_0(V)$ Consider the problem

$$\text{find} \quad \hat{u} = \arg\min_u F(u), \quad F(u) = \Psi(u) + \Phi(u).$$

The astuteness of splitting methods for $F = \Psi + \Phi$ is to use separately

$$\text{prox}_{\gamma\Psi} = (\text{Id} + \gamma\partial\Psi)^{-1} \text{ and } \text{prox}_{\gamma\Phi} = (\text{Id} + \gamma\partial\Phi)^{-1}$$

Usually these are much easier to obtain than the resolvent for F, namely $\text{prox}_{\gamma F} = (\text{Id} + \gamma\partial F)^{-1}$.

Even though the literature is abundant, these can basically be systematized into three main classes:

- forward-backward, see [84, 85, 86];
- Douglas/Peaceman-Rachford, see [87];
- double-backward (little-used), see [88, 89].

A theoretical overview of all these methods can be found in [90, 91]. They are essentially based on proximal calculus.

6.1.1 Forward-Backward (FB) splitting

Forward-backward can be seen as a generalization of the classical gradient projection method for constrained convex optimization. One must assume that either Ψ or Φ is differentiable.

Proposition 15 ([74], Proposition 3.1.) *Assume that $\Psi \in \Gamma_0(V)$, $\Phi \in \Gamma_0(V)$, where V is a Hilbert space, that Ψ is differentiable and that $F = \Psi + \Phi$ is coercive. Let $\gamma \in]0, +\infty[$. Then*

$$F(\hat{u}) = \min_{u \in V} F(u) \quad \Leftrightarrow \quad \hat{u} = \text{prox}_{\gamma\Phi}\big(\hat{u} - \gamma\nabla\Psi(\hat{u})\big).$$

Moreover, $\hat{u}$ is unique if either Ψ of Φ is strictly convex.

[1] Hence $V = V^\star$.

Proof. A sequence of equivalences:

$$\begin{aligned}
F(\hat{u}) = \min_{u\in V} F(u) \quad &\Leftrightarrow \quad 0 \in \partial(\Psi+\Phi)(\hat{u}) = \partial\Psi(\hat{u}) + \partial\Phi(\hat{u}) = \partial\Phi(\hat{u}) + \{\nabla\Psi(\hat{u})\} \\
&\Leftrightarrow \quad -\nabla\Psi(\hat{u}) \in \partial\Phi(\hat{u}) \\
&\Leftrightarrow \quad -\gamma\nabla\Psi(\hat{u}) \in \gamma\partial\Phi(\hat{u}) \\
&\Leftrightarrow \quad \big(\hat{u} - \gamma\nabla\Psi(\hat{u})\big) - \hat{u} \in \gamma\partial\Phi(\hat{u}) \quad \text{(use (5.8), p. 90 identify } \big(\hat{u} - \gamma\nabla\Psi(\hat{u})\big) \text{ with } v) \\
&\Leftrightarrow \quad \hat{u} = \mathrm{prox}_{\gamma\Phi}\big(\hat{u} - \gamma\nabla\Psi(\hat{u})\big)
\end{aligned}$$

F is strictly convex if either Ψ of Φ is so, in which case $\hat{u}$ is unique. □

FB splitting relies on the fixed-point equation

$$u = \mathrm{prox}_{\gamma\Phi}\big(u - \gamma\nabla\Psi(u)\big)$$

The basic iteration consists in two steps:

$$\begin{aligned}
u_{k+\frac{1}{2}} &= u_k - \gamma\nabla\Psi(u_k) \qquad \text{(forward, explicit)} \\
u_{k+1} &= \mathrm{prox}_{\gamma\Phi}\left(u_{k+\frac{1}{2}}\right) \qquad \text{(backward, implicit)}
\end{aligned}$$

Formally, the second step amounts to solving an inclusion, hence its implicit nature.

A practical algorithm proposed by Combettes and Wajs in [74] is a more general iteration where γ is iteration-dependent, errors are allowed in the evaluation of the operators $\mathrm{prox}_{\gamma\Phi}$ and $\nabla\Psi$ and a relaxation sequence $(\lambda_k)_{k\in\mathbb{N}}$ is introduced. Admission of errors allows some tolerance in the numerical implementation of the algorithm, while the iteration-dependent parameters $(\gamma_k)_{k\in\mathbb{N}}$ and $(\lambda_k)_{k\in\mathbb{N}}$ can improve its convergence speed.

Theorem 65 ([74], Theorem 3.4.) *Assume that $\Psi \in \Gamma_0(V)$, $\Phi \in \Gamma_0(V)$, where V is a Hilbert space, that Ψ is differentiable with a $1/\beta$-Lipschitz gradient, $\beta \in]0,\infty[$, and that $F = \Psi + \Phi$ is coercive. Let the sequences $(\gamma_k)_{k\in\mathbb{N}} \subset \mathbb{R}_+$, $(\lambda_k)_{k\in\mathbb{N}} \subset \mathbb{R}_+$, $(a_k)_{k\in\mathbb{N}} \subset V$ and $(b_k)_{k\in\mathbb{N}} \subset V$ satisfy*

- $0 < \inf_{k\in\mathbb{N}} \gamma_k \leqslant \sup_{k\in\mathbb{N}} \gamma_k < 2\beta$;
- $0 < \inf_{k\in\mathbb{N}} \lambda_k \leqslant \sup_{k\in\mathbb{N}} \lambda_k \leqslant 1$;
- $\sum_{k\in\mathbb{N}} \|a_k\| < +\infty$ *and* $\sum_{k\in\mathbb{N}} \|b_k\| < +\infty$.

Consider the iteration:

$$u_{k+1} = u_k + \lambda_k\Big(\mathrm{prox}_{\gamma_k\Phi}\big(u_k - \gamma_k(\nabla\Psi(u_k) + b_k\big) + a_k - u_k\Big), \quad k \in \mathbb{N}.$$

Then

1. *$(u_k)_{k\in\mathbb{N}}$ converges weakly to $\hat{u} \in \widehat{U}$.*
2. *$(u_k)_{k\in\mathbb{N}}$ converges* strongly *to $\hat{u} \in \widehat{U}$ if one of the following conditions hold:*
   - *$\mathrm{int}\widehat{U} \neq \varnothing$;*
   - *either Ψ or Φ satisfy the condition below on $\widehat{U}$:*
     $f \in \Gamma_0(V)$; $\forall (v_k)_{k\in\mathbb{N}}$, $\forall (w_k)_{k\in\mathbb{N}}$ belonging to V and $v \in V$ and $v \in \partial f(w) \in V$ we have $\big[w_k \rightharpoonup w,\ v_k \rightharpoonup v,\ v_k \in \partial f(w_k),\ \forall k \in \mathbb{N}\big] \ \Rightarrow$ w is a strong cluster point of $(w_k)_{k\in\mathbb{N}}$.

Let us emphasize that if $V = \mathbb{R}^n$, convergence is always strong.

6.1.2 Douglas-Rachford splitting

Douglas-Rachford splitting, generalized and formalized in [87], is a much more general class of monotone operator splitting methods. A crucial property of the Douglas-Rachford splitting scheme is its high robustness to numerical errors that may occur when computing the proximity operators prox_Ψ and prox_Φ, see [90].

Proposition 16 ([92], Proposition 18.) *Assume that $\Psi \in \Gamma_0(V)$, $\Phi \in \Gamma_0(V)$ and that $F = \Psi + \Phi$ is coercive. Let $\gamma \in]0, +\infty[$. Then*

$$F(\hat{u}) = \min_{u \in V} F(u) \quad \Leftrightarrow \quad \hat{u} = \mathrm{prox}_{\gamma\Phi} z \quad \textit{where} \quad z = \Big((2\mathrm{prox}_{\gamma\Psi} - \mathrm{Id}) \circ (2\mathrm{prox}_{\gamma\Phi} - \mathrm{Id}) \Big)(z).$$

Moreover, $\hat{u}$ is unique if either Ψ or Φ is strictly convex.

Given a function f, the expression $(2\mathrm{prox}_f - \mathrm{Id})$ is also called reflection operator.

Proof. Since F is coercive, $\widehat{U} \neq \varnothing$. The following chain of equivalences yields the result.

$$F(\hat{u}) = \inf_{u \in V} F(u) \quad \Leftrightarrow \quad 0 \in \partial F(\hat{u}) \quad \Leftrightarrow \quad 0 \in (\partial\Psi + \partial\Phi)(\hat{u}) \quad \Leftrightarrow \quad 0 \in \gamma\partial\Psi(\hat{u}) + \gamma\partial\Phi(\hat{u})$$

$$\Leftrightarrow \quad \exists z \in V : \begin{cases} \hat{u} - z \in \gamma\partial\Psi(\hat{u}) \quad \Leftrightarrow \quad 2\hat{u} - z \in (\mathrm{Id} + \gamma\partial\Psi)(\hat{u}) \\ z - \hat{u} \in \gamma\partial\Phi(\hat{u}) \quad \Leftrightarrow \quad z \in (\mathrm{Id} + \gamma\partial\Phi)(\hat{u}) \quad \Leftrightarrow \quad \underline{\hat{u} = (\mathrm{Id} + \gamma\partial\Phi)^{-1}(z)} \end{cases} \qquad \text{(a)}$$

$$\Leftrightarrow \quad \exists z \in V : \begin{cases} 2(\mathrm{Id} + \gamma\partial\Phi)^{-1}(z) - z \in (\mathrm{Id} + \gamma\partial\Psi)(\hat{u}) \qquad \text{(use (a) for } \hat{u}) \\ \hat{u} = (\mathrm{Id} + \gamma\partial\Phi)^{-1}(z) \end{cases}$$

$$\Leftrightarrow \quad \exists z \in V : \begin{cases} \big(2(\mathrm{Id} + \gamma\partial\Phi)^{-1} - \mathrm{Id}\big) z \in (\mathrm{Id} + \gamma\partial\Psi)(\hat{u}) \\ \hat{u} = (\mathrm{Id} + \gamma\partial\Phi)^{-1}(z) \end{cases}$$

$$\Leftrightarrow \quad \exists z \in V : \begin{cases} \underline{\hat{u} =} (\mathrm{Id} + \gamma\partial\Psi)^{-1} \circ \big(2(\mathrm{Id} + \gamma\partial\Phi)^{-1} - \mathrm{Id}\big)(z) = \underline{\mathrm{prox}_{\gamma\Psi} \circ \big(2\mathrm{prox}_{\gamma\Phi} - \mathrm{Id}\big)(z)} & \text{(b)} \\ \hat{u} = \mathrm{prox}_{\gamma\Phi} z & \text{(c)} \end{cases}$$

$$\Leftrightarrow \quad \exists z \in V : \begin{cases} z = 2\hat{u} - (2\hat{u} - z) = 2\mathrm{prox}_{\gamma\Psi} \circ \big(2\mathrm{prox}_{\gamma\Phi} - \mathrm{Id}\big)(z) - (2\mathrm{prox}_{\gamma\Phi} - \mathrm{Id})(z) \\ \hat{u} = \mathrm{prox}_{\gamma\Phi} z \quad \text{(above: insert (b) for the first } 2\hat{u} \text{ and (c) for the second term)} \end{cases}$$

$$\Leftrightarrow \quad \exists z \in V : \begin{cases} z = (2\mathrm{prox}_{\gamma\Psi} - \mathrm{Id}) \circ (2\mathrm{prox}_{\gamma\Phi} - \mathrm{Id}\big) z \\ \hat{u} = \mathrm{prox}_{\gamma\Phi} z \end{cases}$$

F is coercive by assumption. When either Ψ of Φ is strictly convex, F is strictly convex as well. Hence F admits a unique minimizer $\hat{u}$. □

Thus DR splitting is based on the fixed point equation

$$z = (2\mathrm{prox}_{\gamma\Psi} - \mathrm{Id}) \circ (2\mathrm{prox}_{\gamma\Phi} - \mathrm{Id})(z)$$

and F reaches its minimum at $\hat{u} = \mathrm{prox}_{\gamma\Phi} z$ where z is the fixed point of the equation above.

Remark 33 *Note that the roles of Φ and Ψ can be interchanged since they share the same assumptions.*

Given a fixed scalar $\gamma > 0$ and a sequence $\lambda_k \in (0, 2)$, this class of methods can be expressed via the following recursion written in a compact form

$$z^{(k+1)} = \left(\left(1 - \frac{\lambda_k}{2} \right) \mathrm{Id} + \frac{\lambda_k}{2} (2\mathrm{prox}_{\gamma\Psi} - \mathrm{Id}) \circ (2\mathrm{prox}_{\gamma\Phi} - \mathrm{Id}) \right) z^{(k)}$$

A robust DR algorithm is proposed by Combettes and Pesquet in [92].

Theorem 66 ([92], Theorem 20.) *Assume that $\Psi \in \Gamma_0(V)$, $\Phi \in \Gamma_0(V)$ and that $F = \Psi + \Phi$ is coercive and $\gamma \in]0, \infty[$. Let the sequences $(\lambda_k)_{k\in\mathbb{N}} \subset \mathbb{R}_+$, $(a_k)_{k\in\mathbb{N}} \subset V$ and $(b_k)_{k\in\mathbb{N}} \subset V$ satisfy*

(i) $0 < \lambda_k < 2$, $\forall k \in \mathbb{N}$ *and* $\sum_{k\in\mathbb{N}} \lambda_k(2 - \lambda_k) = +\infty$;

(ii) $\sum_{k\in\mathbb{N}} \lambda_k \big(\|a_k\| + \|b_k\| \big) < +\infty$.

Consider the iteration for $k \in \mathbb{N}$:

$$\begin{aligned} z_{k+\frac{1}{2}} &= (2\mathrm{prox}_{\gamma\Phi} - \mathrm{Id}) z_k + b_k \\ z_{k+1} &= z_k + \lambda_k \Big(\mathrm{prox}_{\gamma\Psi}(2z_{k+\frac{1}{2}} - z_k) + a_k - z_{k+\frac{1}{2}} \Big) \end{aligned}$$

Then $(z_k)_{k\in\mathbb{N}}$ converges weakly to a point $\hat{z} \in V$ and

$$\hat{u} = \mathrm{prox}_{\gamma\Phi} \hat{z} = \arg\min_{u\in V} F(u).$$

When V is of finite dimension, e.g. $V = \mathbb{R}^n$, the sequence $(z_k)_{k\in\mathbb{N}}$ converges strongly.

The sequences a_k and b_k model the numerical errors when computing the proximity operators. They are under control using (ii). Conversely, if the convergence rate can be established, one can easily derive a rule on the number of inner iterations at each outer iteration k such that (ii) is verified.

This algorithm naturally inherits the splitting property of the Douglas-Rachford iteration, in the sense that the operators $\mathrm{prox}_{\gamma\Psi}$ and $\mathrm{prox}_{\gamma\Phi}$ are used in separate steps. Note that the algorithm allows for the inexact implementation of these two proximal steps via the incorporation of the error terms a_k and b_k. Moreover, a variable relaxation parameter λ_k enables an additional flexibility.

6.2 Conjugacy based primal-dual algorithms

6.2.1 A max-representation tool

Main tool: any function $h \in \Gamma_0(V)$ (i.e., l.s.c. convex proper function) satisfies $h^{**} = h$ (Theorem 9, p. 18) and thus [2] it admits the following variational max-representation [18]:

$$h(u) = \max_{z\in V^*} \big\{ \langle u, z \rangle - h^*(z) \big\}$$

[2] Here h^* is the convex conjugate of h given by $h^*(z) = \sup \langle z, u \rangle - h(u)$, see Definition 13, p. 17.

This well known and fundamental relation is in fact the key player not only for handling constraints, but also for deriving "full splitting" of most optimization problems.

Notation: V and W – real n.v.s. (e.g., $V = \mathbb{R}^{M\times N}$ and $W = \left(\mathbb{R}^{M\times N}\right)^2$). Unit balls with respect to the ℓ_∞ norm[3]

$$B_{1,\infty} := \left\{ u \in V \ : \ \|u\|_\infty := \max_{ij} |u_{ij}| \leqslant 1 \right\} \tag{6.1}$$

$$B_{2,\infty} := \left\{ x = (x', x'') \in W \ : \ \|x\|_\infty := \max_{ij} \sqrt{(x'_{i,j})^2 + (x''_{i,j})^2} \leqslant 1 \right\} \tag{6.2}$$

Example 12 *Let $h(u) := \|u\|_1$ where $u \in V = \mathbb{R}^p$.*

$$h^*(z) = \sup_{u\in\mathbb{R}^p} \left\{ \langle u, z\rangle - \|u\|_1 \right\} = \sup_{u\in\mathbb{R}^p} \sum u_i z_i - |u_i| = \sup_{u\in\mathbb{R}^p} \sum |u_i|(z_i \mathrm{sign}(u_i) - 1) = \sup_{u\in\mathbb{R}^p} \sum |u_i|(|z_i| - 1)$$

because u must satisfy $\mathrm{sign}(u_i) = \mathrm{sign}(z_i)$. If there is i such that $|z_i| > 1$ then $h^*(z) = +\infty$. Conversely, if $|z_i| \leqslant 1$ then $h^*(z) = 0$.

$$h^*(z) = \left\{ \begin{array}{ll} 0 & \mathit{if} \ \ \|z\|_\infty \leqslant 1 \\ +\infty & \mathit{if} \ \ \|z\|_\infty > 1 \end{array} \right. = \iota_{B_\infty(1)}(z)$$

Therefore

$$h(u) = \sup_{z\in\mathbb{R}^p} \left\{ \langle z, u\rangle - \iota_{B_\infty(1)}(z) \right\} = \sup_{\|z\|_\infty \leqslant 1} \langle z, u\rangle$$

We know this fact from Lemma 10 (p. 75).

$$h(Au - d) = \sup_{z\in\mathbb{R}^p} \left\{ \langle z, Au - d\rangle - h^*(z) \right\}$$

In particular,

$$\|Au - d\|_1 = \sup_{z\in\mathbb{R}^p} \left\{ \langle z, Au - d\rangle - \iota_{B_{1,\infty}}(z) \right\} \tag{6.3}$$

Remark 34 *Let $p, u \in \mathbb{R}^2$. One has*

$$\sup_p \{ p_1 u_1 + p_2 u_2 \mid \sqrt{p_1^2 + p_2^2} \leqslant 1 \} = \sqrt{u_1^2 + u_2^2}$$

Using Schwarz inequality, one has for $u \propto p$

$$\sup_{\|p\|_2 \leqslant 1} \langle p, u\rangle = \sup_{\|p\|_2 \leqslant 1} \|p\|_2 \|u\|_2 = \|u\|_2$$

For an image $u \in \mathbb{R}^{M\times N}$, the isotropic discrete form of the TV semi-norm reads as

$$\mathrm{TV}(u) = \sum_{ij} \sqrt{(\nabla_y u)_{ij}^2 + (\nabla_x u)_{ij}^2} \tag{6.4}$$

Lemma 17 *Let $u \in V := \mathbb{R}^{M\times N}$ and $W := V \times V$. Using $B_{2,\infty}$ as in (6.2) it holds*

$$\mathrm{TV}(u) = \sup_{\xi\in W} \left(\ \langle \xi, \nabla u\rangle - \iota_{B_{2,\infty}}(\xi) \ \right)$$

[3] If $V = \mathbb{C}^{M\times N}$ then $B_{1,\infty} := \left\{ u \in V \ : \ \|u\|_\infty := \max_{ij} \sqrt{\Re(u_{ij}) + \Im(u_{ij})} \leqslant 1 \right\}$.

Proof. The term on the right admits a component-wise representation in W. Thus

$$\begin{aligned}\sup_{\xi^x,\xi^y} \langle \xi, \nabla u\rangle - \iota_{B_{2,\infty}}(\xi) &= \sup_{(\xi^x,\xi^y)\in B_{2,\infty}} \sum_{ij} \xi^x_{ij}(\nabla_x u)_{ij} + \xi^y_{ij}(\nabla_y u)_{ij} \\ &= \sum_{ij} \sup_{\sqrt{(\xi^x_{ij})^2+(\xi^y_{ij})^2}\leqslant 1} \xi^x_{ij}(\nabla_x u)_{ij} + \xi^y_{ij}(\nabla_y u)_{ij} = \sum_{ij} \sqrt{(\nabla_y u)^2_{ij} + (\nabla_x u)^2_i}\end{aligned}$$

where the last equality comes from Remark 34. □

Remark 35 *The dual* $\left(\frac{1}{2}\|Au-d\|_2^2\right)^*$ *is explicit if and only if* A^TA *is invertible (easy to verify). Otherwise* $\frac{1}{2}\|Au-d\|_2^2$ *can be dualized. Using that*

$$\sup_v \langle v, z\rangle - \frac{1}{2\lambda}\|v\|_2^2 = \frac{\lambda}{2}\|z\|_2^2$$

one has

$$\frac{\lambda}{2}\|Au-d\|_2^2 = \sup_{v\in Z} \langle v, Au-d\rangle - \frac{1}{2\lambda}\|v\|^2$$

6.2.2 Elements of saddle-point formulations

Let $h : W \to \mathbb{R}$ and $g : V \to \mathbb{R}$ be proper, l.s.c. convex functions. Consider the minimization of

$$F(u) = h(Bu) + g(u) \tag{6.5}$$

where $B : V \to W$ is a linear operator with induced norm $\|B\| := \max\{\|Bu\| \ : \ u \in V, \ \ \|u\| \leqslant 1\}$. Using the fact that

$$h(Bu) = \max_{w\in W} \langle Bu, w\rangle - h^*(w) = \max_{w\in W} \langle u, B^*w\rangle - h^*(w)$$

we end up with he following saddle point problems:

$$\min_{u\in V}\max_{w\in W} \left\{\mathcal{F}(u,w) := \langle Bu, w\rangle + g(u) - h^*(w) = \langle u, B^*w\rangle + g(u) - h^*(w)\right\} \tag{6.6}$$

In the case of (6.6) consider the induced optimization problems:

$$\text{(P)} \quad \inf_{u\in V}\left\{r(u) := \sup_w \mathcal{F}(u,w)\right\} \quad \text{and} \quad \text{(D)} \quad \sup_{w\in W}\left\{q(w) = \inf_u \mathcal{F}(u,w)\right\}$$

From Lemma 7 (p. 64) one always has $\sup_{w\in W} q(w) \leqslant \inf_{u\in V} r(u)$, which is referred to as "weak duality". The inequality in the weak duality case can be strict, in which case the quantity

$$\Gamma_{\text{PD}} := \inf_{u\in V} r(u) - \sup_{w\in W} q(w) = \inf_{u\in V}\sup_{w\in W} \mathcal{F}(u,w) - \sup_{w\in W}\inf_{u\in V} \mathcal{F}(u,w)$$

is called the duality gap between the pair of problems (P) − (D).

Our standing assumption is that the convex-concave function $\mathcal{F}$ has a saddle-point (Definition 21, p. 64), i.e., there exists $\hat{u}$ and $\hat{w}$ such that

$$\mathcal{F}(\hat{u}, w) \leqslant \mathcal{F}(\hat{u}, \hat{w}) \leqslant \mathcal{F}(u, \hat{w}) \quad \forall\, u \in V, \quad \forall\, w \in W$$

The existence of a saddle-point corresponds to zero duality gap $\Gamma_{\mathrm{PD}} = 0$. According to Definition 21 (p. 64) and Theorem 43 (p. 64), $(\hat{u}, \hat{w})$ is a saddle-point of $\mathcal{F}$ if and only if $\hat{u}$ is an optimal solution of the primal problem (P), $\hat{w}$ is an optimal solution of the dual problem (D). In this case

$$\inf_{u\in V}\sup_{w\in W}\mathcal{F}(u,w) = \sup_{w\in W}\inf_{u\in V}\mathcal{F}(u,w) = \mathcal{F}(\hat{u},\hat{w})$$

where $\mathcal{F}(\hat{u}, \hat{w})$ is the saddle-point value.

From Theorem 10 (FencheL-Rockafellar, p. 18), the minimization of F in (6.5) is equivalent to

$$\max_{w\in W}\big\{-h^*(-B^*w) - g^*(w)\big\} = \min_{u\in V}\big\{h(Bu) + g(u)\big\}$$

More details can be found e.g. in [18], [15, Chapter 5]. Since we have assumed that the problem in (6.6) has at least one solution $(\hat{u}, \hat{w}) \in V \times W$, it holds that

$$B\hat{u} \in \partial h^*(\hat{u}) \quad \text{and} \quad -(B^*\hat{w}) \in \partial g^*(\hat{u})$$

We recall two basic facts:

(a) For $g \in \Gamma_0(V)$ one has by (5.4) (p. 89) and (5.9) (p. 90)

$$u^{k+1} = (\mathrm{Id} + \gamma\partial g)^{-1} = \mathrm{prox}_{\gamma g}(u^k) = \arg\min_u \left\{\frac{1}{2\gamma}\|u - u^k\|^2 + g(u)\right\}$$

(b) for a concave u.s.c. function $f : W \to \mathbb{R}$ one has

$$\begin{aligned}\arg\max_v f(v) &= \arg\min_v -f(v)\\ \max_v f(v) &= -\min_v -f(v)\end{aligned}$$

6.2.3 The context of imaging applications

Typically we have to find the minimizers of functions of the form

$$F(u) = \lambda\Psi(Au) + \beta\Phi(Hu)$$

The use of two parameters (λ, β) is redundant :

$$\arg\min_u F(u) = \arg\min_u \frac{\lambda}{\beta}\Psi(Au) + \Phi(Hu) \tag{6.7}$$

$$= \arg\min_u \Psi(Au) + \frac{\beta}{\lambda}\Phi(Hu) \tag{6.8}$$

We will set either $\lambda = 1$ or $\beta = 1$ in order to obtain better step-sizes which is closely dependent on the values of (λ, β) and also of $\|A\|$ and $\|H\|$. The choice $\lambda = 1$ or $\beta = 1$ can play a crucial role on the speed of an algorithm.

6.2.4 Full proximal primal-dual algorithm

First we focus on the saddle-point formulation in (6.6),

$$\min_{u\in V}\max_{w\in W}\left\{\mathcal{F}(u,w) := \langle Bu, w\rangle + g(u) - h^*(w)\right\} \tag{6.9}$$

Chambolle and Pock [93] propose a full proximal primal-dual algorithm based on the formulation in (6.9). It is assumed that h and g are simple in the sense that their resolvent operators

$$u = (\mathrm{Id} + \tau\partial g)^{-1}(z) = \arg\min_u \left\{\frac{1}{2\tau}\|u-z\|^2 + g(u)\right\}$$

have a closed-form representation or are easy to solve.

Algorithm (Algorithm 1, [93])

Initialization: Choose τ, $\sigma > 0$, $\tau\sigma\|B\|^2 < 1$, $\theta \in [0,1]$, $(u^0, w^0) \in (V, W)$, $\bar{u}^0 = u^0$

Iterations $n \geqslant 0$

$$\left\{\begin{array}{lcl} w^{n+1} &=& (I + \sigma\partial h^*)^{-1}(w^n + \sigma B\bar{u}^n) \\ u^{n+1} &=& (I + \tau\partial g)^{-1}(u^n - \tau B^* w^{n+1}) \\ \bar{u}^{n+1} &=& u^{n+1} + \theta(u^{n+1} - u^n) \end{array}\right. \tag{6.10}$$

Theorem 67 (Theorem 1, [93]) *Assume that problem* (6.6) *has a saddle point* $(\hat{u}, \hat{w}) \in V \times W$. *Choose* $\theta = 1$, $\tau\sigma\|B\|^2 < 1$, *and let* $(u^n, \bar{u}^n, w^n)$ *be defined by* (6.10). *Then:*

(a) *For any* n, (u^n, w^n) *remains bounded.*

(b) *Set* $u_N := \left(\sum_{n=1}^N u^n\right)/N$ *and* $w_N := \left(\sum_{n=1}^N w^n\right)/N$. *The weak cluster points of* (u_N, w_N) *are saddle-points of* (6.6).

(c) *There exists a saddle-point* (u^*, w^*) *such that* $u^n \to u^*$ *and* $w^n \to w^*$.

From (b), the sequence of iterates $(u^n)_{n\in\mathbb{N}}$ is not necessarily strictly decreasing (can oscillate).

The choice $\theta = 1$ corresponds to a simple linear extrapolation based on the current and previous iterates.

The choice $\theta = 0$ yields the classical Arrow-Hurwicz algorithm [94].

Acceleration of the algorithm is possible if h or g^* is uniformly convex [93, Sec. 5]

Application to minimize an $\ell_1 - \mathrm{TV}$ objective

$$F(u) = \lambda\|Au - d\|_1 + \beta\sum_{i,j}\sqrt{(\nabla_y u)_{ij}^2 + (\nabla_x u)_{ij}^2}$$

Using Lemma 17 (p. 102) and Example 12 (p. 102 one has to solve

$$\min_u F(u) = \min_{u\in V}\max_{\xi\in W}\max_{v\in Z} \lambda\langle Au, v\rangle - \lambda\langle d, v\rangle - \iota_{B_{1,\infty}}(v) + \beta\langle\nabla u, \xi\rangle - \iota_{B_{2,\infty}}(\xi)$$

In order to apply the algorithm in (6.10) to solve the above saddle point-problem, we identify $B = \begin{pmatrix} \lambda A \\ \beta\nabla \end{pmatrix}$, $g(u) = 0$, $w = (v, \xi)$ and

$$h^*(v,\xi) = h_1^*(v) + h_2^*(\xi) \qquad h_1^*(v) = \lambda\langle d, v\rangle + \iota_{B_{1,\infty}}(v) \qquad h_2^*(\xi) = \iota_{B_{2,\infty}}(\xi)$$

Since $\|\nabla\|_2^2 \leqslant 8$, see (5.20) one has [4] $\|B\|_2^2 \leqslant \lambda^2\|A\|^2 + \beta^2 8$. The step-sizes are determined by $\theta \in [0,1]$ $\tau > 0$, $\sigma > 0$ and $\tau\sigma\|B\|_2^2 < 1$, hence,

$$\tau\sigma\left(\lambda^2\|A\|^2 + \beta^2 8\right) < 1$$

The iterates are computed as it follows:

(a) $\xi^{(x,y)_{i,j}^{n+1}} = \dfrac{z_{i,j}^{(x,y)}}{\max\left(1, \sqrt{(z_{i,j}^x)^2 + (z_{i,j}^y)^2}\right)}$ where $z = \xi^n + \sigma\beta\nabla\bar{u}^n$

(b) $v_{i,j}^{n+1} = \dfrac{h_{i,j}}{\max(1, |h_{i,j}|)}$ where $h := \lambda\sigma(A\bar{u}^n - d) + v^n$

(c) $u^{n+1} = u^n - \tau\left(\lambda A^* v^{n+1} - \beta\mathrm{div}(\xi^{n+1})\right)$

(d) $\bar{u}^{n+1} = u^{n+1} + \theta(u^{n+1} - u^n)$

Minimization of a quadratic plus TV **objective**

$$F(u) = \frac{\lambda}{2}\|Au - d\|^2 + \beta\sum_{i,j}\sqrt{(\nabla_y u)_{ij}^2 + (\nabla_x u)_{ij}^2}$$

Using Lemma 17 one has to solve

$$\min_u F(u) = \min_{u\in V}\max_{\xi\in W}\frac{\lambda}{2}\|Au - d\|^2 + \beta\langle\nabla u, \xi\rangle - \iota_{B_{2,\infty}}(\xi)$$

Thus $h^*(\xi) = \iota_{B_{2,\infty}}(\xi)$ and $g(u) = \frac{\lambda}{2}\|Au - d\|^2$ and $B = \beta\nabla$. The step-size parameters should obey $\tau\sigma\beta^2 8 < 1$; it is more beneficial to set $\beta = 1$ and use the first formulation in (6.7).

(a) $\xi^{(x,y)_{i,j}^{n+1}} = \dfrac{z_{i,j}^{(x,y)}}{\max\left(1, \sqrt{(z_{i,j}^x)^2 + (z_{i,j}^y)^2}\right)}$ where $z = \xi^n + \sigma\beta\nabla\bar{u}^n$

(b) $u^{n+1} = \arg\min_u\left\{\dfrac{\lambda}{2}\|Au - d\|^2 - \beta\langle u, \mathrm{div}\xi\rangle + \dfrac{1}{2\tau}\|u - u^n\|^2\right\}$, hence u^{n+1} must solve

$$\tau\left(\lambda A^* A + \mathrm{Id}\right)u = u^n + \tau\left(\lambda A^* d + \beta\mathrm{div}\xi^{n+1}\right)$$

(c) $\bar{u}^{n+1} = u^{n+1} + \theta(u^{n+1} - u^n)$

Unless A is diagonal, the solution in (b) may be difficult to find (in a simple way). If Au can be written as a convolution, a circulant convolution approximation may be used to solve (b) using FFT. Note that this approximation holds only if u satisfies the periodic boundary conditions.

[4] Note that $\|B\|_2^2 = \mu_{\max}(B^T B)$ where $\mu_{\max}$ stands for maximal eigenvalue. If A corresponds to a convolution operator, $\|B\|_2^2$ can be approximated using the FFT.

A^*A is difficult to compute Now dualization with respect to $\|Au-d\|_2$ is needed. Using Remark 35 (p. 103) [5] one has

$$\frac{\lambda}{2}\|Au-d\|_2^2 = \sup_{v\in Z}\langle v, Au-d\rangle - \frac{1}{2\lambda}\|v\|^2 \tag{6.11}$$

Using (6.11) and Lemma 17 one has to solve

$$\min_u F(u) = \min_{u\in V}\max_{\xi\in W}\max_{v\in Z}\langle Au, v\rangle - \langle d, v\rangle - \frac{1}{2\lambda}\|v\|^2 + \beta\langle\nabla u, \xi\rangle - \boldsymbol{\iota}_{B_{2,\infty}}(\xi)$$

We identify

$$h^*(v,\xi) = h_1^*(v) + h_2^*(\xi) \qquad h_1^*(v) = \frac{1}{2\lambda}\|v\|^2 + \langle d, v\rangle \qquad h_2^*(\xi) = \boldsymbol{\iota}_{B_{2,\infty}}(\xi)$$

$g = 0$ and

$$B = \begin{pmatrix} A \\ \beta\nabla \end{pmatrix} \quad w = \begin{pmatrix} v \\ \xi \end{pmatrix} \quad \Rightarrow \quad \langle Bu, w\rangle = \langle Au, v\rangle + \beta\langle\nabla u, \xi\rangle \tag{6.12}$$

$\|B\|^2 \leqslant \|A\|^2 + 8\beta^2$ hence

$$\tau\sigma\|B\|^2 \leqslant \tau\sigma(\|A\|^2 + 8\beta^2) < 1$$

It seems more advantageous to fix $\beta = 1$ and use the first formulation in (6.7) (p. 104). We have

$$\mathrm{prox}_{\sigma h^*}(x,y) = \Big(\mathrm{prox}_{\sigma h_1^*}(x), \mathrm{prox}_{\sigma h_2^*}(y)\Big)$$

where

$$\mathrm{prox}_{\sigma h_1^*}(x) = \arg\min_v \frac{1}{2\sigma}||v-x||^2 + \frac{1}{2\lambda}\|v\|^2 + \langle d, v\rangle = \frac{\lambda}{\lambda+\sigma}(x-\sigma d)$$

The iterates are computed as it follows:

(a) $\xi^{(x,y)^{n+1}}_{i,j} = \dfrac{z^{(x,y)}_{i,j}}{\max\left(1, \sqrt{(z^x_{i,j})^2 + (z^y_{i,j})^2}\right)}$ where $z = \xi^n + \sigma\beta\nabla\bar{u}^n$

(b) $v^{n+1} = \dfrac{\lambda}{\lambda+\sigma}\left(v^n + \sigma(A\bar{u}^n - d)\right)$

(c) $u^{n+1} = u^n - \tau A^* v^{n+1} + \tau\beta\mathrm{div}(\xi^{n+1})$

(d) $\bar{u}^{n+1} = u^{n+1} + \theta(u^{n+1} - u^n)$

The computation in (a) is the same as in the case of the $\ell_1 - \mathrm{TV}$ objective.

6.2.5 A Proximal Alternating Predictor-Corrector Algorithm

In [95] the equivalent saddle point problem in (6.6) is considered (we write $H := B^*$)

$$\min_{u\in V}\max_{w\in W}\left\{\mathcal{F}(u,w) := \langle u, Hw\rangle + g(u) - h^*(w)\right\} \tag{6.13}$$

where

[5] By Remark 35 : $\sup_v \langle v, z\rangle - \dfrac{1}{2\lambda}\|v\|_2^2 = \dfrac{\lambda}{2}\|z\|_2^2$

- $g : V \to \mathbb{R}$ is convex continuously differentiable and its gradient ∇g is Lipschitz continuous with constant L, i.e. $\forall\, u_1, u_2 \in V$ it holds that $\|g(u_1) - g(u_2)\| \leqslant L\|u_1 - u_2\|$.
- $h^* : W \to \mathbb{R}$ is convex proper lsc (possibly nonsmooth).
- $H : W \to V$ is a linear map.

Algorithm PAPC

Initialization: $u_0 \in V$, $w^0 \in W$, $\tau L \leqslant 1$ and $\tau\sigma\|H\| \leqslant 1$.

$$\begin{aligned} \bar{u}^{k+1} &= u^k - \tau(Hw^k + \nabla g(u^k) \\ w^{k+1} &= \operatorname{prox}_{\sigma h^*}(w^k + \sigma H^* \bar{u}^{k+1}) \\ u^{k+1} &= u^k - \tau(Hw^{k+1} + \nabla g(u^k)) \end{aligned} \tag{6.14}$$

One can notice that $\bar{u}^k$ is a prediction step and that u^k is a correction.

Theorem 68 *([95]) Let $\{(\bar{u}^k, w^k, u^k)\}_{k\in\mathbb{N}}$ be the sequence generated by the PAPC algorithm with $\sigma L \leqslant 1$ and $\sigma\tau\|H\| \leqslant 1$. Then the sequence $\{(w^k, u^k)\}_{k\in\mathbb{N}}$ converges to a saddle-point $(\hat{u}, \hat{w})$ of $\mathcal{F}$ in* (6.13).

Application to minimize an $\ell_2 - \mathrm{TV}$ objective

$$F(u) = \frac{\lambda}{2}\|Au - d\|_2^2 + \beta \sum_{i,j} \sqrt{(\nabla_y u)_{ij}^2 + (\nabla_x u)_{ij}^2}$$

Using Lemma 17 (p. 102) and Remark 35 (p. 103) we have

$$\min_u F(u) = \min_{u\in V} \max_{\xi\in W} \frac{\lambda}{2}\|Au - d\|_2^2 - \beta \langle u, \operatorname{div}(\xi)\rangle - \boldsymbol{\iota}_{B_{2,\infty}}(\xi) \tag{6.15}$$

Comparing with (6.13), one has $H(w) = -\beta\mathrm{div}(w)$, $g(u) = \frac{\lambda}{2}\|Au - d\|_2^2$ and $h^*(w) = \boldsymbol{\iota}_{B_{2,\infty}}(w)$. The algorithm's constants (τ, σ) satisfy

$$\tau\frac{\lambda}{2}\|A\| \leqslant 1 \quad \text{and} \quad \sigma\tau\beta^2\|\mathrm{div}\|^2 = \sigma\tau\beta^2 8 \leqslant 1$$

The iterates are given by

(a) $\bar{u}^{k+1} = u^k - \tau(\lambda v - \beta \mathrm{div} w^k) \quad \text{where} \quad v := A^*(Au^k - d)$

(b) $w^{k+1} = \arg\min_w \left\{ \frac{1}{2\sigma}\|w - w^k\|^2 - \beta \langle \nabla\bar{u}, w\rangle + \boldsymbol{\iota}_{B_{2,\infty}}(w) \right\}$

$$w_{ij}^{k+1} = \frac{z_{i,j}^{(x,y)}}{\max\left(1, \sqrt{(z_{i,j}^x)^2 + (z_{i,j}^y)^2}\right)} \quad \text{where} \quad z = w^k + \sigma\beta\nabla\bar{u}^{k+1}$$

(c) $u^{k+1} = u^k - \tau(\lambda v - \beta \mathrm{div} w^{k+1})$

6.2.6 Alternating direction method of multipliers (ADMM)

The ADMM, dating back to the early 1980's, is a form of Augmented Lagrangian method (see subsection 3.4.3 on p. 57) that became very popular in recent years because it can handle "big-data" problems, imaging problems and learning, quite easily. It is worth emphasizing that ADMM is not an approximate version of the classical augmented Lagrangian algorithm. Excellent overviews can be found in [96, 97].

The ADMM is a (basically convex) optimization algorithm can take an advantage of the structure of the optimization problem by breaking it into smaller pieces, each of which are then easier to handle. It is relatively easy to implement and the code can run in a parallel manner. It has tendency towards slow "tail convergence" so it is well suited for problems where high accuracy is not required.

Unconstrained problem

Let $V = \mathbb{R}^n$ and $\Psi \in \Gamma_0(V)$ and $\Phi \in \Gamma_0(V)$. Consider the problem

$$\text{find} \quad \hat{u} \in \arg\min_{u\in\mathbb{R}^n} F(u), \quad F(u) = \Psi(u) + \Phi(Au) \tag{6.16}$$

where $A \in \mathbb{R}^{m\times n}$. Solving this problem is equivalent to extract the optimal $\hat{u}$ out of the solutions of

$$(\hat{u}, \hat{z}) \in \arg\min_{u,z} \big\{\mathcal{F}(u, z) = \Psi(u) + \Phi(z) \;\big|\; Au - z = 0\big\}. \tag{6.17}$$

Note that $\mathcal{F}$ is separable in (u, z). Recall that the Lagrangian for this problem is given by

$$L(u, z, \lambda) = \Psi(u) + \Phi(z) + \langle\lambda, Au - z\rangle$$

and that an optimal solution$\hat{u}, \hat{z}, \hat{\lambda})$ satisfies $\min_{u,z}\max_\lambda L(u, z, \lambda)$. Uzawa's Method (p. 67) solves such a problem.

The ADMM for this problem takes the form [97, 96, 98]

$$\begin{array}{rcl} u^{k+1} & \in & \arg\min\limits_{u\in\mathbb{R}^n} \Big\{\Psi(u) + \Phi(z^k) + \big\langle\lambda^k, Au - z^k\big\rangle + \dfrac{\omega}{2}\|Au - z^k\|^2\Big\} \\ z^{k+1} & \in & \arg\min\limits_{z\in\mathbb{R}^n} \Big\{\Psi(u^{k+1}) + \Phi(z) + \big\langle\lambda^k, Au^{k+1} - z\big\rangle + \dfrac{\omega}{2}\|Au^{k+1} - z\|^2\Big\} \\ \lambda^{k+1} & \in & \lambda^k + \omega(Au^{k+1} - z^{k+1}) \end{array} \tag{6.18}$$

There are several constant terms in the first two subproblems that can be dropped out.

Unlike the classical ALM, the ADMM essentially decouples the functions Ψ and Φ. Often, this decoupling makes it possible to exploit the individual structure of Ψ and Φ so that the first two subproblems in (6.18) can be computed in an efficient (and parallel) way.

Example 13 Consider the minimization of

$$F(u) = \|Au - v\|_2^2 + \beta\|\nabla u\|_2.$$

One do not want to tackle directly the minimization of nonsmooth functions. Similarly to (6.17), one reformulates [73]

$$\arg\min_{u,z} \big\{\mathcal{F}(u, z) = \Big\{\|Au - v\|_2^2 + \beta\|z\|_2 \;\big|\; \nabla u - z = 0\Big\}.$$

Following (6.18), the corresponding ADMM is given by

$$\begin{aligned} u^{k+1} &\in \arg\min_{u\in\mathbb{R}^n}\left\{\|Au-v\|_2^2+\langle\lambda^k,\nabla u\rangle+\frac{\omega}{2}\|\nabla u-z^k\|^2\right\}\\ z^{k+1} &\in \arg\min_{z\in\mathbb{R}^n}\left\{\beta\|z\|_2^2-\langle\lambda^k,z\rangle+\frac{\omega}{2}\|\nabla u^{k+1}-z\|^2\right\}\\ \lambda^{k+1} &\in \lambda^k+\omega(\nabla u^{k+1}-z^{k+1}) \end{aligned}$$

The minimization subproblems are easy to solve. Lemma 15 (p. 93) can be used to update z^k.

General constraint problems

The convergence of the ADMM to a minimizer $\hat{u}$ of F in (6.16) was recently extended to objectives with more than two summands in [99]. Let $V=\mathbb{R}^n$ and $\Theta_i\in\Gamma_0(\mathbb{R}^{n_i})$ for $i=1,\ldots,p$. The problem to solve reads as

$$\min\left\{\mathcal{F}(u_1,\ldots,u_p)=\sum_{i=1}^p\Theta(u_i)\;\Big|\;\sum_{i=1}^p A_iu_i=b,\quad u_i\in U_i,\quad i=1,\ldots,p\right\} \tag{6.19}$$

where $\Theta_i:\mathbb{R}^{n_i}\to\mathbb{R}$ is a closed, proper and convex function (not necessarily smooth), $A_i\in\mathbb{R}^{m\times n_i}$, $U_i\subseteq\mathbb{R}^{n_i}$ is a closed, convex and nonempty set, $b\in\mathbb{R}^m$ is given and $n_1+n_2+\ldots+n_p=n$. Extending the ADMM philosophy in (6.18) to this problem yields

$$\begin{aligned} u_1^{k+1} &\in \arg\min_{u_1\in U_1}\left\{\Theta(u_1)+\langle\lambda^k,Au_1\rangle+\frac{\omega}{2}\left\|A_1u_1+\sum_{i=2}^p A_iu_i^k-b\right\|^2\right\}\\ &\ldots\\ u_i^{k+1} &\in \arg\min_{u_i\in U_i}\left\{\Theta(u_i)+\langle\lambda^k,Au_i\rangle+\frac{\omega}{2}\left\|A_iu_i+\sum_{j=1}^{i-1}A_ju_j^k+\sum_{j=i+1}^{p}A_ju_j^k-b\right\|^2\right\}\\ &\ldots\\ u_p^{k+1} &\in \arg\min_{u_p\in U_p}\left\{\Theta(u_p)+\langle\lambda^k,Au_p\rangle+\frac{\omega}{2}\left\|A_pu_p+\sum_{j=1}^{p-1}A_ju_j^k-b\right\|^2\right\}\\ \lambda^{k+1} &= \lambda^k+\omega\left(\textstyle\sum_{i=1}^p A_iu_i^{k+1}-b\right). \end{aligned} \tag{6.20}$$

The efficiency of the scheme (6.20) has been verified empirically by some recent applications. The following theorem is extracted from [99, Theorem 4.1].

Theorem 69 *Let $\Theta_i\in\Gamma_0(\mathbb{R}^{n_i})$ be strongly convex with the modulus μ_i for $i=1,\ldots,p$. For any*

$$0<\omega<\min_{i=1}^p\left\{\frac{2\mu_i}{3(p-1)\|A_i\|^2}\right\}$$

the sequence $(u_1^k,\ldots,u_p^k)$ generated by (6.20) converges to a solution of problem (6.19).

Chapter 7

Appendix

7.1 Proof of Property 2-1, p. 20

($\Rightarrow$) Let $u, v) \in U$ and $\theta \in (0, 1)$. Using that F is convex and the fact that $\theta u+(1-\theta)v = v+\theta(u-v)$

$$F(v + \theta(u - v)) - F(v) \leqslant \theta(F(u) - F(v))$$

Dividing both sides by θ and letting $\theta \searrow 0$ yields

$$\langle \nabla F(v), u - v\rangle := \lim_{\theta \searrow 0} \frac{F(v + \theta(u - v)) - F(v)}{\theta} \leqslant F(u) - F(v)$$

($\Leftarrow$) Replacing v by $u + \theta(v - u)$ gives rise to

$$\begin{aligned} F(u) &\geqslant F(u + \theta(v - u)) + \langle \nabla F(u + \theta(v - u)), u - (u + \theta(v - u)))\rangle \\ &= F(u + \theta(v - u)) - \theta \langle \nabla F(u + \theta(v - u)), v - u\rangle \end{aligned}$$

and thus

$$(1 - \theta)F(u) \geqslant (1 - \theta)F(u + \theta(v - u)) - (1 - \theta)\theta \langle \nabla F(u + \theta(v - u)), v - u\rangle \tag{7.1}$$

Similarly

$$\begin{aligned} F(v) &\geqslant F(u + \theta(v - u)) + \langle \nabla F(u + \theta(v - u)), v - (u + \theta(v - u)))\rangle \\ &= F(u + \theta(v - u)) + (1 - \theta) \langle \nabla F(u + \theta(v - u)), v - u\rangle \end{aligned}$$

hence

$$\theta F(v) \geqslant \theta F(u + \theta(v - u)) + \theta(1 - \theta) \langle \nabla F(u + \theta(v - u)), v - u\rangle \tag{7.2}$$

Summing up (7.1) and (7.2) leads to

$$\theta F(v)+(1-\theta)F(u) \geqslant \theta F(u+\theta(v-u))+(1-\theta)F(u+\theta(v-u)) = F(u+\theta(v-u)) = F(\theta v+(1-\theta)u)$$

7.2 Proof of Theorem 17, p. 28

Note that condition 1 in Wolfe's rule corresponds to a descent scenario. By the expression for $\cos(\theta_k)$ in (2.16) and using that $u_{k+1} = u_k - \rho_k d_k$ we get

$$\rho_k f'(0) = -\rho_k \langle \nabla F(u_k), d_k\rangle = -\cos(\theta_k)\|\nabla F(u_k)\| \, \|\rho_k d_k\| = -\cos(\theta_k)\|\nabla F(u_k)\| \, \|u_k - u_{k+1}\| < 0.$$

Combining this result with condition 1(a) yields

$$0 < c_0\ \rho_k|f'(0)| = \underline{c_0\ \cos(\theta_k)\|\nabla F(u_k)\|\ \|u_k - u_{k+1}\| \leqslant} f(0) - f(\rho_k) = \underline{F(u_k) - F(u_{k+1})} \qquad (7.3)$$

Note that $F(u_k) - F(u_{k+1})$ is bounded from below by a positive number (descent is guaranteed by test 1). Reminding that $f(\rho_k) = F(u_k - \rho_k d_k) = F(u_{k+1})$, condition 1(b) reads

$$f'(\rho_k) = -\langle \nabla F(u_{k+1}), d_k\rangle \geqslant c_1 f'(0) = -c_1 \langle \nabla F(u_k), d_k\rangle$$

Add $\langle \nabla F(u_k), d_k\rangle$ to both sides of the above inequality:

$$-\langle \nabla F(u_{k+1}) - \nabla F(u_k), d_k\rangle \geqslant (1 - c_1)\langle \nabla F(u_k), d_k\rangle .$$

Using Schwarz's inequality and the definition of $\cos(\theta_k)$,

$$\|\nabla F(u_{k+1}) - \nabla F(u_k)\|\,\|d_k\| \geqslant (1 - c_1)\cos(\theta_k)\|\nabla F(u_k)\|\,\|d_k\|.$$

Division of both sides by $\|d_k\| \neq 0$ leads to

$$\|\nabla F(u_{k+1}) - \nabla F(u_k)\| \geqslant (1 - c_1)\cos(\theta_k)\|\nabla F(u_k)\|$$

Combining this result with the Lipschitz property, i.e. $\|\nabla F(u_{k+1}) - \nabla F(u_k)\| \leqslant \ell\|u_{k+1} - u_k\|$, yields

$$\ell\|u_{k+1} - u_k\| \geqslant \|\nabla F(u_{k+1}) - \nabla F(u_k)\| \geqslant (1 - c_1)\cos(\theta_k)\|\nabla F(u_k)\|$$

Multiplying both sides by $\cos(\theta_k)\|\nabla F(u_k)\|$ and using (7.3) for the underlined term below leads to

$$(1 - c_1)\big(\cos(\theta_k)\big)^2\|\nabla F(u_k)\|^2 \leqslant \frac{\ell}{c_0}\Big(\underline{c_0\ \cos(\theta_k)\|\nabla F(u_k)\|\ \|u_{k+1} - u_k\|}\Big) \leqslant \frac{\ell}{c_0}\Big(F(u_k) - F(u_{k+1})\Big).$$

For

$$r = \frac{c_0(1 - c_1)}{\ell} > 0$$

we obtain

$$r\big(\cos(\theta_k)\big)^2\|\nabla F(u_k)\|^2 \leqslant F(u_k) - F(u_{k+1}).$$

The proof is complete. □

Why in the proof we use only test 1 ?

7.3 Proof of Theorem 18, p. 28

Since $\big(F(u_k)\big)_{k\in\mathbb{N}}$ is decreasing and bounded from below, there is a finite number $\widetilde{F} \in \mathbb{R}$ such that $F(u_k) \geqslant \widetilde{F},\ \forall k \in \mathbb{N}$. Then (2.17) yields

$$\begin{aligned}\sum_{k=0}^{\infty}(\cos\theta_k)^2\|\nabla F(u_k)\|^2 &\leqslant \frac{1}{r}\sum_{k=0}^{\infty}\big(F(u_k) - F(u_{k+1})\big) = \frac{1}{r}\big(F(u_0) - F(u_1) + F(u_1) - F(u_2) + \cdots\big)\\ &\leqslant \frac{1}{r}\big(F(u_0) - \widetilde{F}\big) < +\infty. \qquad (7.4)\end{aligned}$$

If there was $\delta > 0$ such that $\|\nabla F(u_k)\| \geqslant \delta$, for an infinite number of k's, then (2.18) would yield

$$\sum_k(\cos\theta_k)^2\|\nabla F(u_k)\|^2 \geqslant \delta^2\sum_k(\cos\theta_k)^2 \to \infty$$

which contradicts (7.4). Hence the conclusion.

7.4 Proof of Proposition 1, p. 38

Put

$$\theta(t) = -\varphi(\sqrt{t}), \tag{7.5}$$

then θ is convex by (b) and continuous on $\mathbb{R}_+$ by (c). Its convex conjugate—see (1.14), p. 17—is $\theta^\star(b) = \sup_{t\geqslant 0}\{bt - \theta(t)\}$ where $b \in \mathbb{R}$. Define $\psi(b) = \theta^\star(-\frac{1}{2}b)$ which means that

$$\psi(b) = \sup_{t\geqslant 0}\left\{-\frac{1}{2}bt - \theta(t)\right\} = \sup_{t\geq 0}\left\{-\frac{1}{2}bt^2 + \varphi(t)\right\}. \tag{7.6}$$

By Theorem 9, p. 18, we have $(\theta^\star)^\star = \theta$. Calculating $(\theta^\star)^\star$ at t^2 and using (7.5) yields

$$-\varphi(t) = \theta(t^2) = \sup_{b\leqslant 0}\left\{bt^2 - \theta^\star(b)\right\} = \sup_{b\geqslant 0}\left\{-\frac{1}{2}bt^2 - \psi(b)\right\}.$$

Since $\theta(t) \leqslant 0$ on $\mathbb{R}_+$, we have $\theta^\star(b) = +\infty$ if $b > 0$ and then the supremum of $b \to bt^2 - \theta^\star(b)$ in the middle expression above necessarily corresponds to $b \leqslant 0$. Finally,

$$\varphi(t) = \inf_{b\geqslant 0}\left\{\frac{1}{2}bt^2 + \psi(b)\right\}. \tag{7.7}$$

The statement $\varphi \Rightarrow \psi$ comes from the fact that $(\theta^\star)^\star = \theta$.

Next we focus on the possibility to achieve the supremum in ψ jointly with the infimum in φ. For any $\hat{b} > 0$, define $f_{\hat{b}} : \mathbb{R}_+ \to \mathbb{R}$ by $f_{\hat{b}}(t) = \frac{1}{2}\hat{b}t + \theta(t)$, then (7.6) yields $\psi(\hat{b}) = -\inf_{t\geqslant 0} f_{\hat{b}}(t)$. Observe that $f_{\hat{b}}$ is convex by (b) with $f_{\hat{b}}(0) = 0$ by (a) and $\lim_{t\to\infty} f_{\hat{b}}(t) = +\infty$ by (e), hence $f_{\hat{b}}$ has a unique minimum reached at a $\hat{t} \geqslant 0$. According to (7.6), $\psi(\hat{b}) = -\frac{1}{2}\hat{b}\hat{t}^2 + \varphi(\hat{t})$, then equivalently the infimum in (7.7) is reached for $\hat{b}$ since $\varphi(\hat{t}) = \frac{1}{2}\hat{b}\hat{t}^2 + \psi(\hat{b})$. Notice that $\theta'(t) = -\frac{\varphi'(\sqrt{t})}{2\sqrt{t}}$ and that $f'_{\hat{b}}(t) = \frac{1}{2}\hat{b} + \theta'(t)$ is increasing on $\mathbb{R}_+$ by (b). If $f'_{\hat{b}}(0^+) \geqslant 0$, i.e. if $\hat{b} \geqslant \varphi''(0^+)$, $f_{\hat{b}}$ reaches its minimum at $\hat{t} = 0$. Otherwise, its minimum is reached for a $\hat{t} > 0$ such that $f'_{\hat{b}}(\hat{t}) = 0$, i.e. $\hat{b} = -2\theta'(\hat{t})$. In this case, $t \to -\frac{1}{2}\hat{b}t^2 + \varphi(t)$ in the last expression of (7.6) reaches its supremum for a $\hat{t}$ that satisfies $\hat{b} = -2\theta'(\hat{t}^2)$. Hence $\hat{b}$ is as given in the proposition.

7.5 Proof of Proposition 2, p. 40

By (b) and (d), $\left(\frac{1}{2}\|t\|^2 - \varphi(\|t\|)\right)$ is convex and coercive, so the maximum of ψ is reached. The formula for ψ in (2.35) is equivalent to

$$\psi(\|b\|) + \frac{1}{2}\|b\|^2 = \max_{t\in\mathbb{R}^s}\left\{\langle b, t\rangle - \left(\frac{1}{2}\|t\|^2 - \varphi(\|t\|)\right)\right\} \tag{7.8}$$

The term on the left is the convex conjugate of $\frac{1}{2}\|t\|^2 - \varphi(\|t\|)$, see (1.14), p. 17. The latter term being convex continuous and $\not\equiv \infty$, using convex bi-conjugacy (Theorem 9, p. 18), (7.8) is equivalent to

$$\frac{1}{2}\|t\|^2 - \varphi(\|t\|) = \sup_{b\in\mathbb{R}^s}\left\{\langle b, t\rangle - \left(\psi(\|b\|) + \frac{1}{2}\|b\|^2\right)\right\} \tag{7.9}$$

Noticing that the supremum in (7.9) is reached, (7.9) is equivalent to

$$\varphi(\|t\|) = \min_{b\in\mathbb{R}^s}\left\{-\langle b, t\rangle + \left(\psi(\|b\|) + \frac{1}{2}\|b\|^2\right) + \frac{1}{2}\|t\|^2\right\} = \min_{b\in\mathbb{R}^s}\left(\frac{1}{2}\|t - b\|^2 + \psi(\|b\|)\right)$$

Hence (2.35) is proven.

Using (2.35), the maximum of ψ and the minimum of φ are reached by a pair (t, b) such that

$$b - t + \varphi'(\|t\|)\frac{t}{\|t\|} = 0.$$

For t fixed, the solution for b, denoted $\hat{b}$ is unique and is as stated in the proposition.

7.6 Derivation of the CG algorithm, p. 44

Assume that $\nabla F(u_k) \neq 0$, then the optimal $g(\alpha_k) \neq 0$ for $\alpha_k \in \mathbb{R}^{k+1}$ is the unique point defined by (2.42). Then for all iterates $0 \leqslant j \leqslant k$, we have $g(\alpha_j) \neq 0$. We will express $g(\alpha_j)$ in the form

$$g(\alpha_j) = \rho_j d_j = \sum_{i=0}^{j} \alpha_j[i]\nabla F(u_i), \quad 0 \leqslant j \leqslant k, \quad d_j \in \mathbb{R}^n, \quad \rho_j \in \mathbb{R} \tag{7.10}$$

Since $g(\alpha_j) \neq 0$, we have $d_j \neq 0$ and $\rho_j \neq 0$, $0 \leqslant j \leqslant k$. We will choose d_j in such a way that successive directions are easy to compute. Iterates read

$$\underline{u_{j+1} = u_j - \rho_j d_j} = u_j - \sum_{i=0}^{j} \alpha_j[i]\nabla F(u_i), \quad 0 \leqslant j \leqslant k \tag{7.11}$$

(a) A major result. Combining (2.40) and (7.11),

$$\nabla F(u_{j+1}) = \nabla F(u_j - \rho_j d_j) = \nabla F(u_j) - \rho_j B d_j, \quad 0 \leqslant j \leqslant k. \tag{7.12}$$

For $k \geqslant 1$, combining (2.43) and (7.12) yields

$$\underline{0 =} \langle \nabla F(u_{j+1}), \nabla F(u_i)\rangle = \langle \nabla F(u_j), \nabla F(u_i)\rangle - \rho_j \langle Bd_j, \nabla F(u_i)\rangle = -\rho_j \underline{\langle Bd_j, \nabla F(u_i)\rangle}, \ 0 \leqslant i < j \leqslant k.$$

By (7.18), d_j is a linear combination of $\nabla F(u_i)$, $0 \leqslant i \leqslant j$; this, joined to (2.43) yields

$$\underline{\langle Bd_j, d_i\rangle = 0, \quad 0 \leqslant i < j \leqslant k}. \tag{7.13}$$

We will say that the directions d_j and d_i are conjugated with respect to B. Since $B \succ 0$, this result shows as well that $\underline{d_j}$ and d_i, $0 \leqslant i < j \leqslant k$ $\underline{\text{are linearly independent.}}$ Till iteration $k+1$, (7.10) can be put into a matrix form:

$$\Big[\rho_0 d_0 \vdots \rho_1 d_1 \vdots \ \cdots \ \vdots \rho_k d_k\Big] = \Big[\nabla F(u_0) \vdots \nabla F(u_1) \vdots \ \cdots \vdots \nabla F(u_k)\Big] \begin{bmatrix} \alpha_0[0] & \alpha_1[0] & \cdots & \alpha_k[0] \\ & \alpha_1[1] & \cdots & \alpha_k[1] \\ & & \ddots & \vdots \\ & & & \alpha_k[k] \end{bmatrix}$$

The linear independence of $\big(\rho_i d_i\big)_{i=0}^{j}$ and $\big(\nabla F(u_i)\big)_{i=0}^{j}$ for any $0 \leqslant j \leqslant k$ implies

$$\alpha_j[j] \neq 0, \quad 0 \leqslant j \leqslant k. \tag{7.14}$$

(b) Calculus of successive directions. By (7.12) we see that

$$-\rho_j B d_j = \nabla F(u_{j+1}) - \nabla F(u_j), \quad 0 \leqslant j \leqslant k.$$

Using this results, the fact that $B = B^T$, along with (7.13) yields

$$0 = \langle Bd_k, d_j\rangle = \langle d_k, Bd_j\rangle = \langle \rho_k d_k, -\rho_j Bd_j\rangle = \left\langle \underline{\rho_k d_k}, \nabla F(u_{j+1}) - \nabla F(u_j)\right\rangle, \quad 0 \leqslant j \leqslant k-1.$$

Introducing $\rho_k d_k$ according to (7.10) in the last term above entails

$$0 = \left\langle \sum_{i=0}^{k} \alpha_k[i]\nabla F(u_i), \nabla F(u_{j+1}) - \nabla F(u_j)\right\rangle, \quad 0 \leqslant j \leqslant k-1. \tag{7.15}$$

Using that $\alpha_j[j] \neq 0$—see (7.14)—and that $\rho_j \neq 0$ in (7.10), for all $0 \leqslant j \leqslant k$, the constants below are well defined:

$$\gamma_j[i] = \frac{\alpha_j[i]}{\rho_j \alpha_j[j]}, \quad 1 \leqslant i \leqslant j \leqslant k. \tag{7.16}$$

Using (7.11)

$$d_k = \sum_{i=0}^{k} \alpha_k[i]\nabla F(u_i) = \alpha_k[k]\left(\sum_{i=0}^{k-1} \frac{\alpha_k[i]}{\alpha_k[k]}\nabla F(u_i) + \nabla F(u_k)\right)$$

Then (7.15) is equivalent to

$$0 = \left\langle \sum_{i=0}^{k-1} \gamma_k[i]\nabla F(u_i) + \nabla F(u_k)\ ,\ \nabla F(u_{j+1}) - \nabla F(u_j)\right\rangle, \quad 0 \leqslant j \leqslant k-1.$$

Using (2.43), this equation yields:

$$\begin{aligned} j = k-1: \qquad & -\gamma_k[k-1]\ \|\nabla F(u_{k-1})\|^2 + \|\nabla F(u_k\|^2 = 0 \\ 0 \leqslant j \leqslant k-2: \qquad & -\gamma_k[j]\ \|\nabla F(u_j)\|^2 + \gamma_k[j+1]\ \|\nabla F(u_{j+1})\|^2 = 0 \end{aligned}$$

It is easy to check that the solution is

$$\gamma_k[j] = \frac{\|\nabla F(u_k\|^2}{\|\nabla F(u_j)\|^2}, \quad 0 \leqslant j \leqslant k-1. \tag{7.17}$$

Note that with the help of $\gamma_j[i]$ in (7.16), d_j in (7.10) can equivalently written down as

$$d_j = \sum_{i=0}^{j} \gamma_j[i]\nabla F(u_i), \quad 0 \leqslant j \leqslant k, \tag{7.18}$$

provided that ρ_j solves the problem $\min_{\rho} F(u_j - \rho d_j)$, $0 \leqslant j \leqslant k$. From (7.17) and (7.18),

$$\begin{aligned} \underline{d_k} &= \sum_{i=0}^{k-1} \gamma_k[i]\nabla F(u_i) + \nabla F(u_k) = \sum_{i=0}^{k-1} \frac{\|\nabla F(u_k\|^2}{\|\nabla F(u_i)\|^2}\nabla F(u_i) + \nabla F(u_k) \\ &= \nabla F(u_k) + \frac{\|\nabla F(u_k\|^2}{\|\nabla F(u_{k-1})\|^2}\left(\sum_{i=0}^{k-2} \frac{\|\nabla F(u_{k-1}\|^2}{\|\nabla F(u_i)\|^2}\nabla F(u_i) + \nabla F(u_{k-1})\right) \\ &\quad \underline{= \nabla F(u_k) + \frac{\|\nabla F(u_k\|^2}{\|\nabla F(u_{k-1})\|^2} d_{k-1}.} \end{aligned}$$

The latter result provides a simple way to compute the new d_k using the previous d_{k-1}:

$$d_0 = \nabla F(u_0) \ \text{ and } \ d_i = \nabla F(u_i) + \frac{\|\nabla F(u_i\|^2}{\|\nabla F(u_{i-1})\|^2} d_{i-1}, \quad 1 \leqslant i \leqslant k.$$

(*c*) The optimal ρ_k for d_k. The optimal step-size ρ_k is the unique solution of the problem

$$F(u_k - \rho_k d_k) = \inf_{\rho\in\mathbb{R}} F(u_k - \rho d_k).$$

ρ_k is hence the unique minimizer of f as given below,

$$f(\rho) = \frac{1}{2}\left\langle B(u_k - \rho d_k), (u_k - \rho d_k)\right\rangle - \left\langle c, (u_k - \rho d_k)\right\rangle.$$

hence the unique solution of

$$0 = f'(\rho) = \rho\left\langle Bd_k, d_k\right\rangle - \left\langle Bu_k, d_k\right\rangle + \left\langle c, d_k\right\rangle = \rho\left\langle Bd_k, d_k\right\rangle - \left\langle \nabla F(u_k), d_k\right\rangle$$

reads

$$\underline{\rho_k = \frac{\langle \nabla F(u_k), d_k\rangle}{\langle Bd_k, d_k\rangle}}.$$

All ingredients of the algorithm are established.

7.7 Proof of Lemma 2, p. 45

By (2.45), we have $-B_k\, d_{k-1} = \frac{1}{\rho_k}(\nabla F(u_k) - \nabla F(u_{k-1}))$. Then we can write

$$\frac{\langle -B_k d_{k-1}, \nabla F(u_k)\rangle}{\langle B_k d_{k-1}, d_{k-1}\rangle} = \frac{\big\langle \big(\nabla F(u_k) - \nabla F(u_{k-1})\big), \nabla F(u_k)\big\rangle}{\big\langle \big(-\nabla F(u_k) + \nabla F(u_{k-1})\big), d_{k-1}\big\rangle} = \frac{\big\langle \big(\nabla F(u_k) - \nabla F(u_{k-1})\big), \nabla F(u_k)\big\rangle}{-\underline{\langle \nabla F(u_k), d_{k-1}\rangle} + \langle \nabla F(u_{k-1}), d_{k-1}\rangle}$$

If ρ_{k-1} is optimal, i.e. if $F(u_{k-1} - \rho_k d_{k-1}) = \inf_\rho F(u_{k-1} - \rho d_{k-1})$, then

$$\langle \nabla F(u_{k-1} - \rho_{k-1} d_{k-1}), d_{k-1}\rangle = \underline{0 = \langle \nabla F(u_k), d_{k-1}\rangle}$$

Then

$$\frac{\langle -B_k d_{k-1}, \nabla F(u_k)\rangle}{\langle B_k d_{k-1}, d_{k-1}\rangle} = \frac{\big\langle \big(\nabla F(u_k) - \nabla F(u_{k-1})\big), \nabla F(u_k)\big\rangle}{\langle \nabla F(u_{k-1}), d_{k-1}\rangle}$$

If ρ_{k-2} is optimal as well, in the same way we have $\langle \nabla F(u_{k-1}), d_{k-2}\rangle = 0$. Using this result and the formula for d_{k-1} according to step 3 in PR, the denominator on the right hand side is

$$\langle \nabla F(u_{k-1}), d_{k-1}\rangle = \langle \nabla F(u_{k-1}), \nabla F(u_{k-1}) + \xi_{k-1} d_{k-2}\rangle = \|\nabla F(u_{k-1})\|^2.$$

It follows that

$$\frac{\langle -B_k d_{k-1}, \nabla F(u_k)\rangle}{\langle B_k d_{k-1}, d_{k-1}\rangle} = \xi_k$$

for ξ_k as in the Lemma. Using the expression for ξ_k established above, we find:

$$\begin{aligned}\langle B_k d_{k-1}\,,\, d_k\rangle &= \langle B_k d_{k-1}\,,\, \nabla F(u_k) + \xi_k d_{k-1}\rangle = \langle B_k d_{k-1}\,,\, \nabla F(u_k)\rangle + \xi_k \langle B_k d_{k-1}\,, d_{k-1}\rangle \\ &= \langle B_k d_{k-1}\,,\, \nabla F(u_k)\rangle - \frac{\langle B_k d_{k-1}, \nabla F(u_k)\rangle}{\langle B_k d_{k-1}, d_{k-1}\rangle}\, \langle B_k d_{k-1}\,, d_{k-1}\rangle = 0.\end{aligned}$$

Hence the result.

7.8 Proof of the Farkas-Minkowski theorem 39, p. 60

Let 2 hold. Using that $\langle a_i, u\rangle \geqslant 0, \forall i \in I$, we obtain

$$\langle b, u\rangle = \sum_{i\in I} \lambda_i \langle a_i, u\rangle \geqslant 0.$$

Thus 2 implies the inclusion stated in 1.

Define the subset

$$K \stackrel{\text{def}}{=} \left\{ \sum_{i\in I} \lambda_i a_i \in V \;\mid\; \lambda_i \geqslant 0,\ \forall i \in I \right\} \tag{7.19}$$

The proof of 1 $\Rightarrow$ 2 is aimed at showing that $b \in K$. The proof is conducted ad absurdum and is based of 3 substatements.

(a) K in (7.19) have the properties stated below.

- K is a cone with vertex at 0.

$$\alpha > 0 \;\;\Rightarrow\;\; \alpha\lambda_i \geqslant 0,\ \forall i \in I \;\;\Rightarrow\;\; \alpha \sum_{i\in I} \lambda_i a_i = \sum_{i\in I} (\alpha\lambda_i) a_i \in K.$$

- K is convex.
  Let $0 < \theta < 1$, $\lambda_i \geqslant 0,\ \forall i \in I$ and $\mu_i \geqslant 0,\ \forall i \in I$.

$$\theta \sum_{i\in I} \lambda_i a_i + (1-\theta) \sum_{i\in I} \mu_i a_i = \sum_{i\in I} \big(\theta\lambda_i + (1-\theta)\mu_i\big) a_i \in K$$

  because $\theta\lambda_i + (1-\theta)\mu_i \geqslant 0$.

- K is closed in V.

  – Suppose that $\{a_i,\ i \in I\}$ are linearly independent. Set $\mathcal{A} = \operatorname{span}\{a_i : i \in I\}$—a finite dimensional vector subspace (hence closed). Consider an arbitrary $(v_j)_{j\in\mathbb{N}} \subset K$. For any $j \in \mathbb{N}$, there is a unique set $\{\lambda_j[i] \geqslant 0,\ i \in I\}$ so that

$$v_j = \sum_{i\in I} \lambda_j[i] a_i \in K \cap \mathcal{A}.$$

  Let $\lim\limits_{j\to\infty} v_j = v \in V$ then $v \in V \cap \mathcal{A}$ which is equivalent to

$$\lim_{j\to\infty} v_j = \sum_{i\in I} \left(\lim_{j\to\infty} \lambda_j[i] \right) a_i \in K \cap \mathcal{A}.$$

  Hence K is closed in V.

  – Let $\{a_i,\ i \in I\}$ be linearly dependent. Then K is a finite union of cones corresponding to linearly independent subsets of $\{a_i\}$. These cones are closed, hence K is closed.

(b) Let $K \subset V$ be convex, closed and $K \neq \varnothing$. Let $b \in V \setminus K$. Using the Hahn-Banach theorem 8, $\exists h \in V$ and $\exists \alpha \in \mathbb{R}$ such that

$$\begin{aligned} \langle u, h\rangle &> \alpha, \quad \forall u \in K, \\ \langle b, h\rangle &< \alpha. \end{aligned}$$

By the Projection Theorem 29 (p. 48), there is a unique $c \in K$ such that

$$\|b - c\| = \inf_{u \in K} \|b - u\| > 0 \quad (> 0 \text{ because } b \notin K)$$

$c \in V$ satisfies

$$\langle u - c, b - c\rangle \leqslant 0, \quad \forall u \in K. \tag{7.20}$$

Then

$$\begin{aligned} \|c - b\|^2 > 0 \quad &\Leftrightarrow \quad \|c\|^2 - \langle c, b\rangle - \langle c, b\rangle + \|b\|^2 > 0 \\ &\Leftrightarrow \quad \|c\|^2 - \langle c, b\rangle > \langle c, b\rangle - \|b\|^2 \quad \Leftrightarrow \quad \langle c, c - b\rangle > \langle b, c - b\rangle \end{aligned}$$

Set $\underline{h = c - b}$ and α such that

$$\langle c, c - b\rangle = \langle c, h\rangle > \underline{\alpha > \langle b, h\rangle} = \langle b, c - b\rangle$$

From (7.20)

$$\langle u - c, h\rangle \geqslant 0 \quad \Rightarrow \quad \underline{\langle u, h\rangle} \geqslant \langle c, h\rangle \underline{> \alpha, \quad \forall u \in K}.$$

Remark 36 *The hyperplane* $\{u \in V : \langle h, u\rangle = \alpha\}$ *separates* strictly $\{b\}$ *and* K.

(c) Let K read as in (7.19). Then

$$b \notin K \quad \Rightarrow \quad \exists h \in V \text{ such that } \langle a_i, h\rangle \geqslant 0, \ \forall i \in I \text{ and } \langle b, h\rangle < 0.$$

Since $b \notin K$, then (b) shows that $\exists h \in V$ and $\exists \alpha \in \mathbb{R}$ such that (Theorem 8)

$$\begin{aligned} \langle u, h\rangle &> \alpha, \quad \forall u \in K, \\ \langle b, h\rangle &< \alpha. \end{aligned}$$

In particular

$$u = 0 \in K \quad \Rightarrow \quad 0 = \langle u, h\rangle > \alpha \quad \Rightarrow \quad \alpha < 0.$$

Since K is a cone

$$\langle \lambda u, h\rangle > \alpha, \ \forall \lambda > 0, \ \forall u \in K \quad \Rightarrow \quad \langle u, h\rangle > \frac{\alpha}{\lambda}, \quad \forall \lambda > 0, \ \forall u \in K.$$

Since $\alpha/\lambda \nearrow 0$ as $\lambda \nearrow \infty$ it follows that

$$\langle u, h\rangle \geqslant 0, \quad \forall u \in K.$$

Choose arbitrarily $a_i \in K$ for $i \in I$. Then $\langle a_i, h\rangle \geqslant 0, \quad \forall i \in I$.

(d) Conclusion. Statement 2 means that $b \in K$. By (c), we have

$$\text{statement 2 fails} \quad \Rightarrow \quad \text{statement 1 fails.}$$

Consequently, statement 1 implies statement 2.

7.9 Proof of Lemma 8, p. 66

Restate (P_λ):

$$K(\lambda) = \inf_{u\in V} L(u,\lambda) = L(u_\lambda,\lambda).$$

Let $\lambda, \lambda+\eta \in \mathbb{R}^q_+$. We have:

$$(a)\;\; L(u_\lambda,\lambda) \leqslant L(u_{\lambda+\eta},\lambda) \quad \text{and} \quad (b)\;\; L(u_{\lambda+\eta},\lambda+\eta) \leqslant L(u_\lambda,\lambda+\eta)$$

$$(a) \quad \Rightarrow \quad K(\lambda) \leqslant \underline{F(u_{\lambda+\eta}) + \sum_{i=1}^q (\lambda_i+\eta_i)h_i(u_{\lambda+\eta})} - \sum_{i=1}^q \eta_i h_i(u_{\lambda+\eta}) = \underline{K(\lambda+\eta)} - \sum_{i=1}^q \eta_i h_i(u_{\lambda+\eta}).$$

$$(b) \quad \Rightarrow \quad K(\lambda+\eta) \leqslant \underline{F(u_\lambda) + \sum_{i=1}^q \lambda_i h_i(u_\lambda)} + \sum_{i=1}^q \eta_i h_i(u_\lambda) = \underline{K(\lambda)} + \sum_{i=1}^q \eta_i h_i(u_\lambda).$$

$$\Rightarrow \quad \sum_{i=1}^q \eta_i h_i(u_{\lambda+\eta}) \leqslant K(\lambda+\eta) - K(\lambda) \leqslant \sum_{i=1}^q \eta_i h_i(u_\lambda).$$

Then $\exists\, \theta \in [0,1]$ such that

$$\begin{aligned} K(\lambda+\eta) - K(\lambda) &= (1-\theta)\sum_{i=1}^q \eta_i h_i(u_\lambda) + \theta \sum_{i=1}^q \eta_i h_i(u_{\lambda+\eta}) \\ &= \sum_{i=1}^q \eta_i h_i(u_\lambda) + \theta \sum_{i=1}^q \eta_i \Big(h_i(u_{\lambda+\eta}) - h_i(u_\lambda) \Big) \end{aligned}$$

Since $\lambda \to u_\lambda$ is continuous and h_i, $1 \leqslant i \leqslant q$ are continuous, for any $\lambda \in \mathbb{R}^q_+$ we can write down

$$K(\lambda+\eta) - K(\lambda) = \sum_{i=1}^q \eta_i h_i(u_\lambda) + \|\eta\|\varepsilon(\eta) \quad \text{where} \quad \lim_{\eta\to 0} \varepsilon(\eta) = 0$$

It follows that K is differentiable and that $\langle \nabla K(\lambda), \eta\rangle = \sum_{i=1}^q \eta_i\, h_i(u_\lambda)$.

7.9.1 Proof of Theorem 42, p. 63

To prove 1, we have to show that <u>if</u> the constraints are qualified in the convex sense (Definition 20) then they are qualified in the general sense (Definition 19, p. 61), for any $u \in U$, which will allows us to apply the KT Theorem 41.

- Let $w \neq \hat{u}$, $w \in U$, satisfy Definition 20. Set $\underline{v = w - \hat{u}}$. Using that $h_i(\hat{u}) = 0,\ \forall i \in I(\hat{u})$, and that h_i are convex (see Property 2-1, p. 20),

  $$i \in I(\hat{u}) \quad \Rightarrow \quad \langle \nabla h_i(\hat{u}), v\rangle = h_i(\hat{u}) + \langle \nabla h_i(\hat{u}), w-\hat{u}\rangle \leqslant h_i(w).$$

  By Definition 20, we know that $h_i(w) \leqslant 0$ and that $h_i(w) < 0$ is h_i is not affine. Then

  $$\langle \nabla h_i(\hat{u}), w-\hat{u}\rangle = \langle \nabla h_i(\hat{u}), v\rangle \leqslant 0\ ,$$

  where the inequality is strict if h_i is not affine. Hence Definition 19 is satisfied for v.

- Let $\hat{u} = w \in U$. Since $h_i(\hat{u}) < 0$ is impossible for any $i \in I(\hat{u})$, Definition 20 means that all h_i for $i \in I(\hat{u})$ are affine. Then Definition 19 is trivially satisfied.

The result in statement 1 follows from the KT relations (Theorem 41, p. 62).

To prove statement 2, we have to check that $F(\hat{u}) \leqslant F(u)$, $\forall u \in U$. Let $u \in U$ (arbitrary). The convexity of h_i (Property 2-1) and the definition of U yield the two inequalities below:

$$i \in I(\hat{u}) \quad \Rightarrow \quad \langle \nabla h_i(\hat{u}), u - \hat{u} \rangle \leqslant h_i(u) - h_i(\hat{u}) \leqslant 0, \quad \forall u \in U.$$

Using these inequalities and noticing that $\lambda_i(\hat{u}) \geqslant 0$, we can write down

$$\begin{aligned}
F(\hat{u}) &\leqslant F(\hat{u}) - \sum_{i \in I(\hat{u})} \lambda_i(\hat{u}) \big(h_i(u) - h_i(\hat{u})\big) \\
&\leqslant F(\hat{u}) - \sum_{i \in I(\hat{u})} \lambda_i(\hat{u}) \langle \nabla h_i(\hat{u}), u - \hat{u} \rangle \\
&\leqslant F(\hat{u}) - \left\langle \sum_{i=1}^{q} \lambda_i(\hat{u}) \nabla h_i(\hat{u}) \, , \, u - \hat{u} \right\rangle \qquad \big(\text{by } (3.35), \ \lambda_i(\hat{u}) = 0 \text{ if } i \notin I(\hat{u})\big) \\
&= F(\hat{u}) + \langle \nabla F(\hat{u}) \, , \, u - \hat{u} \rangle \qquad (\text{by } (3.34)) \\
&\leqslant F(u), \quad \forall u \in U \qquad (F \text{ is convex}).
\end{aligned}$$

7.10 Proof of Proposition 8, p. 77

Proof. By contradiction: let $\exists \varepsilon > 0$ and $(v_k)_{k \in \mathbb{N}}$ with

$$t_k \overset{\text{def}}{=} \|v_k\| \leqslant \frac{1}{k+1}$$

such that

$$\big|F(u + v_k) - F(u) - \delta F(u)(v_k)\big| > \varepsilon t_k = \varepsilon \|v_k\|, \quad \forall k \in \mathbb{N}.$$

$\frac{\|v_k\|}{t_k} = 1$, $\forall k \in \mathbb{N}$ $\Rightarrow$ $\frac{v_k}{t_k} \in \{w \in \mathbb{R}^n : \|w\| = 1\}$ (compact set), $\forall k \in \mathbb{N}$ $\Rightarrow$ $\exists v_{k_j}$ such that

$$\lim_{j \to \infty} \frac{v_{k_j}}{t_{k_j}} = v, \quad \|v\| = 1 \tag{7.21}$$

For $\ell > 0$ a local Lipschitz constant of F we have

$$\begin{aligned}
\varepsilon t_{k_j} &< \big|F(u + v_{k_j}) - F(u) - \delta F(u)(v_{k_j})\big| \\
&\leqslant \big|F(u + v_{k_j}) - F(u + t_{k_j} v)\big| + \big|F(u + t_{k_j} v) - F(u) - \delta F(u)(t_{k_j} v)\big| + \big|\delta F(u)(t_{k_j} v) - \delta F(u)(v_{k_j})\big| \\
&\leqslant 2\ell \|v_{k_j} - t_{k_j} v\| + \big|F(u + t_{k_j} v) - F(u) - t_{k_j} \delta F(u)(v)\big|
\end{aligned}$$

Note that by Lemma 7 (p. 77) we have $\big|\delta F(u)(t_{k_j} v) - \delta F(u)(v_{k_j})\big| \leqslant \ell \|t_{k_j} v - v_{k_j}\|$, which is used to get the last inequality.

Using that $\lim_{j \to \infty} t_{k_j} = 0$ along with the definition of δF in (4.4) (p. 76) and (7.21)

$$\varepsilon \leqslant 2\ell \lim_{j \to \infty} \left\| \frac{v_{k_j}}{t_{k_j}} - v \right\| + \lim_{j \to \infty} \left| \frac{F(u + t_{k_j} v) - F(u)}{t_{k_j}} - \delta F(u)(v) \right| = 0.$$

This contradicts the assumption that $\varepsilon > 0$.

Bibliography

[1] A. Chambolle, "An algorithm for total variation minimization and application", *Journal of Mathematical Imaging and Vision*, vol. 20, no. 1, Jan. -Mar. 2004.

[2] G. Aubert and P. Kornprobst, *Mathematical problems in image processing*, Springer-Verlag, Berlin, 2 edition, 2006.

[3] H. H. Bauschke and P. L. Combettes, *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*, 2011.

[4] J. F. Bonnans, J.-C. Gilbert, C. Lemaréchal, and C. A. Sagastizábal, *Numerical optimization (theoretical and practical aspects)*, Springer, 2003.

[5] P. G. Ciarlet, *Introduction à l'analyse numérique matricielle et à l'optimisation*, Collection mathématiques appliquées pour la maîtrise. Dunod, Paris, 5e edition, 2000.

[6] R. Glowinski, J. Lions, and R. Trémolières, *Analyse numérique des inéquations variationnelles*, vol. 1, Dunod, Paris, 1 edition, 1976.

[7] J.-B. Hiriart-Urruty and C. Lemaréchal, *Convex analysis and Minimization Algorithms, vol. I*, Springer-Verlag, Berlin, 1996.

[8] J.-B. Hiriart-Urruty and C. Lemaréchal, *Convex analysis and Minimization Algorithms, vol. II*, Springer-Verlag, Berlin, 1996.

[9] D. Luenberger, *Introduction to Linear and Nonlinear Programming*, Addison-Wesley, New York, 1 edition, 1973.

[10] J. Nocedal and S. Wright, *Numerical Optimization*, Springer, New York, 2 edition, 2006.

[11] J. Ortega and W. Rheinboldt, *Iterative Solution of Nonlinear Equations in Several Variables*, Academic Press, New York, 1970.

[12] W. Press, S. Teukolsky, W. Vetterling, and B. Flannery, *Numerical recipes, the art of scientific computing*, Cambridge Univ. Press, New York, 1992.

[13] H. H. Bauschke and P. L. Combettes, *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*, Springer, 2011.

[14] L. Schwartz, *Analyse: Topologie générale et analyse fonctionnelle*, vol. 2 of *Enseignement des sciences*, Hermann, Paris, 1993.

[15] A. Auslender and M. Teboulle, *Asymptotic Cones and Functions in Optimization and Variational Inequalities*, Springer, New York, 2003.

[16] H. Brézis, *Analyse fonctionnelle*, Collection mathématiques appliquées pour la maîtrise. Masson, Paris, 1992.

[17] I. Ekeland and R. Temam, *Convex Analysis and Variational Problems*, SIAM, Amsterdam: North Holland, 1976.

[18] R. T. Rockafellar, *Convex Analysis*, Princeton University Press, 1970.

[19] L. Ambrosio, N. Fusco, and D. Pallara, *Functions of Bounded Variation and Free Discontinuity Problems*, Oxford Mathematical Monographs. Oxford University Press, 2000.

[20] L. Schwartz, *Analyse II. Calcul différentiel et équations différentielles*, vol. 2 of *Enseignement des sciences*, Hermann, Paris, 1997.

[21] A. Avez, *Calcul différentiel*, Masson, 1991.

[22] C. Labat and J. Idier, "Convergence of conjugate gradient methods with a closed-form stepsize formula", *J. of Optimization Theory and Applications*, vol. 136, no. 1, pp. 43–60, 2008.

[23] E. Chouzenoux, J. Idier, and S. Moussaoui, "A majorizeâminimize strategy for subspace optimization applied to image restoration", *IEEE Transactions on Image Processing*, vol. 20, no. 6, pp. 1517–1528, June 2011.

[24] P. Davis, *Circulant matrices*, John Wiley, New York, 3 edition, 1979.

[25] D. Geman and C. Yang, "Nonlinear image recovery with half-quadratic regularization", *IEEE Transactions on Image Processing*, vol. IP-4, no. 7, pp. 932–946, July 1995.

[26] G. Strang, *Linear Algebra and its Applications*, Brooks / Cole, 3 edition, 1988.

[27] R. Chan and M. Ng, "Conjugate gradient methods for toeplitz systems", *SIAM Review*, vol. 38, no. 3, pp. 427–482, Sep. 1996.

[28] R. H.-F. Chan and X.-Q. Jin, *An Introduction to Iterative Toeplitz Solvers*, Cambridge University Press, 2011.

[29] G. Strang, "A proposal for toeplitz matrix calculations", *Stud. Appl. Math.*, vol. 74, 1986.

[30] T. Chan, "An optimal circulant preconditioner for toeplitz systems", *SIAM J. Sci. Stat. Comput.*, , no. 9, pp. 766â–771, 1988.

[31] R. Chan, "Circulant preconditioners for hermitian toeplitz systems", *SIAM J. Matrix Anal. Appl.*, , no. 10, pp. 542â–550, 1989.

[32] E. Tyrtyshnikov, "Optimal and superoptimal circulant preconditioners", *SIAM J. Matrix Anal. Appl*, , no. 13, pp. 459–â473, 1992.

[33] E. Weiszfeld, "Sur le point pour lequel la somme des distances de n points donnÃ©s est minimum", *TÃ´hoku Math. J*, , no. 43, pp. 355–386, 1937.

[34] H. E. Voss and U. Eckhardt, "Linear convergence of generalized Weiszfeld's method", *Computing*, vol. 25, no. 3, pp. 243–251, 1980.

[35] T. Chan and P. Mulet, "On the convergence of the lagged diffusivity fixed point method in total variation image restoration", *SIAM Journal on Numerical Analysis*, vol. 36, no. 2, pp. 354–367, 1999.

[36] A. Beck and S. Sabach, "Weiszfeld's method: Old and new results", *J Optim Theory Appl*, vol. 164, no. 1, pp. 1–40, 2015.

[37] D. Geman and G. Reynolds, "Constrained restoration and recovery of discontinuities", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-14, no. 3, pp. 367–383, Mar. 1992.

[38] P. Charbonnier, L. Blanc-Féraud, G. Aubert, and M. Barlaud, "Deterministic edge-preserving regularization in computed imaging", *IEEE Transactions on Image Processing*, vol. 6, no. 2, pp. 298–311, Feb. 1997.

[39] A. H. Delaney and Y. Bresler, "Globally convergent edge-preserving regularized reconstruction: an application to limited-angle tomography", *IEEE Transactions on Image Processing*, vol. 7, pp. 204–221, Feb. 1998.

[40] J. Idier, "Convex half-quadratic criteria and auxiliary interacting variables for image restoration", *IEEE Transactions on Image Processing*, vol. 10, no. 7, pp. 1001–1009, July 2001.

[41] M. Nikolova and M. Ng, "Analysis of half-quadratic minimization methods for signal and image recovery", *SIAM Journal on Scientific Computing*, vol. 27, no. 3, pp. 937–966, 2005.

[42] M. Allain, J. Idier, and Y. Goussard, "On global and local convergence of half-quadratic algorithms", *IEEE Transactions on Image Processing*, vol. 15, no. 5, pp. 1130–1142, 2006.

[43] M. Nikolova and R. Chan, "The equivalence of half-quadratic minimization and the gradient linearization iteration", *IEEE Transactions on Image Processing*, vol. 16, no. 6, pp. 1623–1627, June 2007.

[44] M. C. Robini and Y. Zhu, "Generic half-quadratic optimization for image reconstruction", *SIAM Journal on Imaging Sciences*, vol. 8, no. 3, pp. 1752–1797, 2015.

[45] G. Aubert and L. Vese, "A variational method in image recovery", *SIAM Journal on Numerical Analysis*, vol. 34, no. 5, pp. 1948–1979, 1997.

[46] D. Luenberger and Y. Ye, *Linear and Nonlinear Programming*, Springer, New York, 3 edition, 2008.

[47] A. Lewis and M. Overton, "Nonsmooth optimization via quasi-newton methods", *Math. Programming, online 2012*.

[48] P. G. Ciarlet, *Introduction to Numerical Linear Algebra and Optimization*, Cambridge University Press, 1989.

[49] M. Minoux, *Programmation mathématique. Théorie et algorithmes, T.1*, Dunod, Paris, 1983.

[50] P. G. Ciarlet, *Introduction à l'analyse numérique matricielle et à l'optimisation*, Collection mathématiques appliquées pour la maîtrise. Masson, Paris, 1990.

[51] C. Vogel, *Computational Methods for Inverse Problems*, Frontiers in Applied Mathematics Series, Number 23. SIAM, 2002.

[52] S. Wright, *Primal-Dual Interior-Point Methods*, SIAM Publications, Philadelphia, 1997.

[53] Y. Nesterov, *Introductory Lectures on Convex Optimization: a Basic Course*, Kluwer Academic, Dordrecht, 2004.

[54] Y. Nesterov, "Smooth minimization of non-smooth functions", *Math. Program. (A)*, vol. 1, no. 103, pp. 127–152, 2005.

[55] J.-F. Aujol, "Some first-order algorithms for total variation based image restoration", *Journal of Mathematical Imaging and Vision*, vol. 34, no. 3, pp. 307–327, July 2009.

[56] P. Weiss, L. Blanc-Féraud, and G. Aubert, "Efficient schemes for total variation minimization under constraints in image processing", *SIAM Journal on Scientifigc Computing*, vol. 31, no. 3, pp. 2047–2080, 2009.

[57] Y. Nesterov, "Gradient methods for minimizing composite objective function", Tech. Rep., Université catholique de Louvain, Center for Operations Research and Econometrics (CORE), CORE Discussion Papers 2007076, Sep. 2007.

[58] C. Zalinescu, *Convex analyzis in general vector spaces*, World Scientific, River Edge, N.J., 2002.

[59] L. Rudin, S. Osher, and C. Fatemi, "Nonlinear total variation based noise removal algorithm", *Physica*, vol. 60 D, pp. 259–268, 1992.

[60] J. E. Besag, "Digital image processing : Towards Bayesian image analysis", *Journal of Applied Statistics*, vol. 16, no. 3, pp. 395–407, 1989.

[61] D. Donoho, I. Johnstone, J. Hoch, and A. Stern, "Maximum entropy and the nearly black object", *Journal of the Royal Statistical Society B*, vol. 54, no. 1, pp. 41–81, 1992.

[62] P. Moulin and J. Liu, "Analysis of multiresolution image denoising schemes using generalized Gaussian and complexity priors", *IEEE Transactions on Image Processing*, vol. 45, no. 3, pp. 909–919, Apr. 1999.

[63] M. Belge, M. Kilmer, and E. Miller, "Wavelet domain image restoration with adaptive edge-preserving regularization", *IEEE Transactions on Image Processing*, vol. 9, no. 4, pp. 597–608, Apr. 2000.

[64] A. Antoniadis and J. Fan, "Regularization of wavelet approximations", *Journal of Acoustical Society America*, vol. 96, no. 455, pp. 939–967, Sep. 2001.

[65] M. Nikolova, "Minimizers of cost-functions involving nonsmooth data-fidelity terms. Application to the processing of outliers", *SIAM Journal on Numerical Analysis*, vol. 40, no. 3, pp. 965–994, 2002.

[66] M. Nikolova, "A variational approach to remove outliers and impulse noise", *Journal of Mathematical Imaging and Vision*, vol. 20, no. 1, Jan. -Mar. 2004.

[67] T. Chan and S. Esedoglu, "Aspects of total variation regularized l^1 function approximation", *SIAM Journal on Applied Mathematics*, vol. 65, pp. 1817â–1837, 2005.

[68] J. Yang, Y. Zhang, and W. Yin, "An efficient TVL1 algorithm for deblurring multichannel images corrupted by impulsive noise", *SIAM Journal on Scientific Computing*, vol. 31, no. 4, pp. 2842–2865, June 2009.

[69] V. Duval, J.-F. Aujol, and Y. Gousseau, "The TVL1 model: a geometric point of view", *SIAM Journal on Multiscale Modeling and Simulation*, vol. 8, no. 1, pp. 154–189, 2009.

[70] R. T. Rockafellar and J. B. Wets, *Variational analysis*, Springer-Verlag, New York, 1998.

[71] L. C. Evans and R. F. Gariepy, *Measure theory and fine properties of functions*, Studies in Advanced Mathematics. CRC Press, Roca Baton, FL, 1992.

[72] J.-J. Moreau, "Proximité et dualité dans un espace hilbertien", *Bulletin de la S. M. F.*

[73] Y. Wang, J. Yang, W. Yin, and Y. Zhang, "A new alternating minimization algorithm for total variation image reconstruction", *SIAM Journal on Imaging Sciences*, vol. 1, no. 3, pp. 248â–272, 2008.

[74] P. L. Combettes and V. R. Wajs, "Signal recovery by proximal forward-backward splitting", *SIAM Multiscale Model. Simul.*, vol. 4, no. 4, pp. 1168–1200, 2005.

[75] N. Z. Shor, *Minimization Methods for Non-Differentiable Functions*, vol. 3, Springer-Verlag, 1985.

[76] S. Alliney and S. A. Ruzinsky, "An algorithm for the minimization of mixed l_1 and l_2 norms with application to Bayesian estimation", *IEEE Transactions on Signal Processing*, vol. SP-42, no. 3, pp. 618–627, Mar. 1994.

[77] H. Fu, M. Ng, M. Nikolova, and J. L. Barlow, "Efficient minimization methods of mixed $\ell_1 - \ell_l$ and $\ell_2 - \ell_1$ norms for image restoration", *SIAM Journal on Scientific Computing*, vol. 27, no. 6, pp. 1881–1902, 2006.

[78] H. Bauschke, S. M. Moffat, and X. Wang, "Firmly nonexpansive mappings and maximally monotone operators: Correspondence and duality", *Set-Valued Anal.*, vol. 20, no. 1, pp. 131–153, 2012.

[79] R. T. Rockafellar, "Monotone operators and the proximal point algorithm", *SIAM Journal on Control and Optimization*, vol. 14, no. 5, pp. 877–898, Aug. 1976.

[80] J. Eckstein and D. P. Bertsekas, "On the Douglas-Rachford splitting method and the proximal point algorithm for maximal monotone operators", *Math. Programming: Series A and B*, vol. 55, no. 3, pp. 293–318, July . 1992.

[81] J.-J. Moreau, "Fonctions convexes duales et points proximaux dans un espace hilbertien", *CRAS Sér. A Math.*, vol. 255, pp. 2897–2899, 1962.

[82] P. L. Combettes and J.-C. Pesquet, *Proximal Splitting Methods in Signal Processing*, p. 185â212, Bauschke, H.H.; Burachik, R.S.; Combettes, P.L.; Elser, V.; Luke, D.R.; Wolkowicz, H. (Eds.), Springer-Verlag, 2011.

[83] J. Yang, W. Yin, Y. Zhang, and Y. Wang, "A fast algorithm for edge-preserving variational multichannel image restoration", *SIAM Journal on Imaging Sciences*, vol. 2, no. 2, pp. 569–592, 2009.

[84] D. Gabay, *Applications of the method of multipliers to variational inequalities*, M. Fortin and R. Glowinski, editors, North-Holland, Amsterdam, 1983.

[85] P. Tseng, "Applications of a splitting algorithm to decomposition in convex programming and variational inequalities", *SIAM Journal on Control and Optimization*, vol. 29, no. 1, pp. 119–138, 1991.

[86] P. Tseng, "A modified forward-backward splitting method for maximal monotone mappings", *SIAM Journal on Control and Optimization*, vol. 38, no. 1, pp. 431–446, 2000.

[87] P.-L. Lions and B. Mercier, "Splitting algorithms for the sum of two nonlinear operators", *SIAM Journal on Numerical Analysis*, vol. 16, no. 6, pp. 964–979, Dec. . 1979.

[88] P.-L. Lions, "Une méthode itérative de resolution d'une inéquation variationnelle", *Israel Journal of Mathematics*, vol. 31, no. 2, pp. 204–208, June . 1978.

[89] G. B. Passty, "Ergodic convergence to a zero of the sum of monotone operators in hilbert space", *Journal of Mathematical Analysis and Applications*, vol. 72, 1979.

[90] P. L. Combettes, "Solving monotone inclusions via compositions of nonexpansive averaged operators", *Optimization*, vol. 53, no. 5, Dec. 2004.

[91] J. Eckstein and B. F. Svaiter, "A family of projective splitting methods for the sum of two maximal monotone operators", *Math. Program., Ser. B*, vol. 111, no. 1-2, pp. 173–199, Jan. . 2008.

[92] P. L. Combettes and J.-C. Pesquet, "A Douglas-Rachford splittting approach to nonsmooth convex variational signal recovery", *IEEE Journal of Selected Topics in Signal Processing*, vol. 1, no. 4, pp. 564–574, Dec. 2007.

[93] A. Chambolle and T. Pock, "A first-order primal-dual algorithm for convex problems with applications to imaging", *J. Math. Imaging Vision*, vol. 40, no. 1, pp. 120–145, 2011.

[94] K. J. Arrow, L. Hurwicz, and H. Uzawa, *Studies in linear and nonlinear programming*, Chenery, H.B., Johnson, S.M., Karlin, S., Marschak, T., Solow, R.M. (eds.). Stanford University Press, Stanford, 1958.

[95] Y. Drori, S. Sabach, and M. Teboulle, "A simple algorithm for a class of nonsmooth convex-concave saddle-point problems", *Operations Research Letters*, vol. 43, Feb. 2015.

[96] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers", *Foundations and Trends in Machine Learning,*, vol. 3, no. 1, pp. 1–122, 2011.

[97] J. Eckstein, "Augmented Lagrangian and Alternating Direction Methods for convex optimization: A tutorial and some illustrative computational results", Tech. Rep., Rutgers University, NJ, RRR 32-2012, 2012.

[98] T. Goldstein, B. O'Donoghue, S. Setzer, and R. Baraniuk, "Fast alternating direction optimization methods", *SIAM Journal on Imaging Sciences*, vol. 7, no. 3, pp. 1588–1623, 2014.

[99] D. Han and X. Yuan, "A note on the alternating direction method of multipliers", *Journal of Optimization Theory and Applications*, vol. 155, no. 1, pp. 227–238, Oct. 2012.