Convex Optimization with Applications to Image Processing

J. Lellmann

Lecture Notes

Mathematical Tripos, Part III
University of Cambridge, Michaelmas 2013

Please send questions and corrections to:
jl707@damtp.cam.ac.uk

Last updated 13/02/2014

Table of contents

Chapter 1
Introduction

1.1 Motivation

Why optimization in image processing? It is a convenient strategy to design algorithms

1. by defining what the *result* should *look like*, not *how* to find it,
2. that can be analyzed,
3. that are *modular* and can be easily modified as requirements change.

A classical example is image denoising, because the problem is very clear: we are given an input image $I\colon \Omega \subseteq \mathbb{R}^d \to [0,1]$ that is corrupted by noise – such as sensor noise or JPEG artifacts – and would like to reconstruct the uncorrupted image u as accurately as possible.

Simply convolving the image with a Gaussian kernel or computing the average in a small window around each point results in an image that is much too smooth, i.e., much of the structure in the original image is lost.

A common first approach is to use "ad hoc" methods. For example, we might notice that natural images tend to have a high amount of self-similarity, so it seems reasonable to try to find similar patches in the image, and for each point compute the average over the corresponding pixels in similar patches.

Such methods may work, and are often simple to implement and fast. The disadvantage is that if they do *not* work, it may be very hard to precisely point out why. Similarly it is often difficult to make specific statements about the quality of their output – which features does it remove or preserve? Does it remove features below a certain size? How strong can the noise be if we want it effectively removed? Does it keep edges?

A very useful way out of this dilemma is to follow a *variational* approach. We postulate that the output of our method is the minimizer of a function

$$\min_u f(u).$$

This abstracts from the actual *implementation* of the algorithm, and allows to focus on *modelling* the problem, i.e., finding a function f such that the minimizer has the desired properties.

A prototypical variational model is the following:

$$\min_{u:\Omega\to\mathbb{R}} f(u) := \frac{1}{2}\int_\Omega \|u - I\|^2\,dx + \lambda \int_\Omega \|\nabla u\|^2\,dx.$$

The left term is commonly called the *data term*, and serves to keep u close to the observed input image I. The right term, which we refer to as the *regularization term*, is weighted by a scalar $\lambda > 0$ and advocates solutions that are smooth. Finding suitable regularizers is often the more difficult part, as they encode what we know about the characteristics of the desired solution – our *prior knowledge* that does not depend on the actual data.

Apart from this intuitive explanation, there is also a *statistical* interpretation: minimizing f is the same as maximizing $e^{-f(x)}$. In the above case, this is the product of a Gaussian density with mean I and another density that encodes the regularizer. The minimizer of the variational model is therefore the image u that is most likely with respect to a certain density that depends on the input image.

A drawback of the above model is that it tends to over-smooth edges. The reason can be easily seen by inspecting the regularizer: it clearly prefers continuous transitions between two values c_1 and c_2 to a "jump" of the same height (going from 0 to 1 via 1/2 is penalized by $(1/2)^2 + (1/2)^2 = 1/2$, but a sharp transition from 0 to 1 is penalized by $1^2 = 1$). We have to make sure that they cost the same or we will get smoothing!

This leas to the Rudin-Osher-Fatemi (ROF) model:

$$\min_{u:\Omega\to\mathbb{R}} f(u) := \frac{1}{2}\int_\Omega \|u - I\|^2\,dx + \lambda \int_\Omega \|\nabla u\|\,dx \tag{1.1}$$

(the notation is not very precise, since the ROF model requires a generalization of the gradient to discontinuous functions, but it is enough to illustrate the point). The idea is that monotone transitions between two values have exactly the same regularization cost, regardless of how sharp the transition is.

For the ROF model it is possible to show that if the input image I assumes only the values 0 and 1 then jumps will be preserved. This highlights an important point: it is often relatively easy to explain unexpected results from basic properties of the objective, and then to slightly *modify* the objective to get rid of these properties.

Taking the idea one step further, we arrive at the $\mathrm{TV} - L^1$ model:

$$\min_{u:\Omega\to\mathbb{R}} f(u) := \int_\Omega \|u - I\|\,dx + \lambda \int_\Omega \|\nabla u\|\,dx.$$

Compared to ROF, this has the advantage that it does not reduce the contrast of the image. In fact, it is possible to show that characteristic function of discs with radius of at least $\lambda/2$ will be preserved, and discs with radius strictly less than $\lambda/2$ will be completely removed – we have a good intuition of what happens to features of a certain size.

The drawback of variational methods is that the model description in terms of f does not include any hint on how to implement the numerical minimization. In fact, there can be many obstacles such as large-scale problems, f being non-differentiable, or having multiple local minima.

This raises an important question: let us assume that the solver hits the stopping criterion and returns a potential minimizer u of f, but we find that u is not what we would like the output to look like. It could be that the model f is not appropriate, but it could also be that the model is fine but the solver just did not find a good minimizer.

This is where the concept of *convexity* comes into play:

$$\textbf{convexity} \Rightarrow \textbf{local minimizer} = \textbf{global minimizer}.$$

If f if *convex* (and in fact the ROF model (1.1) is), then every local minimizer is also a global minimizer. Therefore, if the output is not as expected (and the solver returns a local minimum), we know that the only correct way to improve the situation is to modify the model. Therefore convexity *decouples* the modelling and implementation/optimization stage.

A difficulty when implementing convex solvers for image processing tasks is that we often have to deal with problem that are

$$\textbf{non-smooth} \quad \text{and} \quad \textbf{large-scale}.$$

This causes several problems: we cannot evaluate the gradient everywhere, and much less the Hessian. Even if we could, the gradient is not Lipschitz-continuous and does not carry any information about how close we are to the minimum. If we decide to replace the non-smooth terms by a smoothed version (such as $\|\nabla u\| \approx (u_x^2 + u_y^2 + \varepsilon^2)^{1/2}$, close to the non-smoothness the gradient can be a very bad approximation of the behaviour of the function.

But the good news is:

$$\textbf{convexity} \Rightarrow \textbf{(very often) efficiently solvable}.$$

In the last decades very efficient general-purpose and dedicated convex solvers have been developed, so that even problems with 10 million variables can often be solved in reasonable time. This is not the case for general non-convex problems: minimizing

$$\min_{x \in \mathbb{R}^n} f(x) \quad \text{s.t.} \quad x_i\,(1 - x_i) = 0,$$

is already challenging for $n = 50$ and practically impossible in general for $n = 1000$, because at minimum f needs to be evaluated at 2^n points.

1.2 Course Layout

1. *Introduction:* variational approach, data term and regularizer, separation of modelling and implementation
2. *Existence:* extended real-valued functions, lower semi-continuity, level-boundedness, existence of minimizers of extended real-valued functions.
3. *Convexity:* convex sets and functions, epigraphs and effective domain, Jensen inequality, characterization of convex functions and convex calculus, local minimizers are global minimizers, derivative tests, projections, convex hulls.
4. *Cones and Generalized Inequalities:* cones, generalized inequalities, conic programs, standard-, second-order-, and semidefinite cones.
5. *Subgradients:* set-valued mappings, subdifferential, generalized Fermat condition, normal cones, subdifferentials as normal cones of the epigraph, relative interior, subdifferential calculus.
6. *Conjugate Functions:* convex hull and closure of a function, envelope representation of convex sets and functions, Legendre-Fenchel transform and properties, inversion rule for the subdifferential, conjugate calculus, support functions.
7. *Duality in Optimization:* primal and dual problems, perturbation formulation, weak and strong duality, necessary and sufficient primal-dual optimality conditions, dual for conic problems, Lagrangians, saddle-point formulation of the optimality conditions.
8. *Numerical Optimality:* difficulties in non-smooth optimization, numerical primal-dual gap, infeasibilities.
9. *First-Order Methods:* forward and backward steps, proximal step formulation, forward and backward stepping, splitting principle, gradient-projection, primal-dual methods, Augmented Lagrangian.
10. *Interior-Point Methods:* canonical barriers, primal-dual central path, duality gap on the central path, tracing the central path.

11. *Using Solvers and Discretization Issues:* transforming problems into normal forms, representing multi-dimensional data in vector form, implementing linear operators, discretizing variational problems, adjoint differential operators
12. *Support Vector Machines:* supervised and unsupervised machine learning, linear maximum-margin classifiers, reformulation as convex problem, dual problem and optimality conditions, computing primal from dual solutions, support vectors, kernel trick
13. *Total Variation:* functions of bounded variation, dual formulation of the total variation, coarea formula, higher-order total variation, infimal convolution, conjugate of infinal convolution, total generalized variation, Meyer's G-norm, non-local regularizers.
14. *Relaxation:* non-convexity in image processing, Chan-Vese, Mumford-Shah, convex relaxation, genralized coarea condition, thresholding theorem, discretized energy, anisotropy.

1.3 Suggested Reading

- Rockafellar, Wets: Variational Analysis, 2009: The notation and structure of the variational analysis part of this course follows Rockafellar's excellent book. The book can be a little abstract at times as it develops a much more general theory that also covers nonconvex functions. The predecessor classic (Convex Analysis) from the same author contains several simpler proofs for the convex case.
- Boyd, Vandenberghe: Convex Optimization, 2004: A good overview with many applications, examples and exercises. Focuses on the classical KKT representation of optimality conditions. An excellent read, and available online for free.
- Ben-Tal, Nemirovski: Lectures on Modern Convex Optimization, 2001: A good intuitive introduction to interior point methods including some complexity analysis.
- Paragios, Chen, Faugeras: Handbook of Mathematical Models in Computer Vision, 2006: This is a good complement to the course as it covers the modelling aspect and is a good reference and includes many of the standard methods in modern image processing.

Chapter 2
Existence

2.1 Extended real-valued functions

In the literature, optimization problems are commonly formulated using an *objective function* $f_0\colon \mathbb{R}^n \to \mathbb{R}$ and *constraint functions* $f_1, ..., f_m\colon \mathbb{R}^n \to \mathbb{R}$, e.g.,

$$\begin{aligned} &\min_x f_0(x) \quad s.t. \quad x \in C, \\ &C = \{x \in \mathbb{R}^n \,|\, f_i(x) \leqslant 0, i = 1, ..., m\}. \end{aligned}$$

By allowing $+\infty$ as the value of the objective function we can rewrite this in a very compact form:

$$\min_x f(x),$$

where $f\colon \mathbb{R}^n \to \mathbb{R} \cup \{\pm\infty\}$, with the definition $x \notin C \Leftrightarrow f(x) = +\infty$.

Definition 2.1. *(extended real line) We define* $\bar{\mathbb{R}} := \mathbb{R} \cup \{+\infty, -\infty\}$ *with the rules:*

1. $\infty + c = \infty, \quad -\infty + c = -\infty$ *for all* $c \in \mathbb{R}$,
2. $0 \cdot \infty = 0, \quad 0 \cdot (-\infty) = 0$,
3. $\inf \mathbb{R} = \sup \emptyset = -\infty, \quad \inf \emptyset = \sup \mathbb{R} = +\infty$.
4. $+\infty - \infty = -\infty + \infty = +\infty$ *(sometimes; careful:* $-\infty = \lambda\,(\infty - \infty) \neq \lambda\,\infty - \lambda\,\infty = \infty$ *if* $\lambda < 0$*)*

Remark 2.2. (rules for extended real values)

$$\begin{aligned} \inf_x \{f(x) + g(x)\} &\geqslant \inf_x f(x) + \inf_x g(x), \ \inf \lambda f = \lambda \inf f, \lambda \geqslant 0, \ \inf f = -\sup(-f), \\ \sup_x \{f(x) + g(x)\} &\leqslant \sup_x f(x) + \sup_x g(x), \ \sup \lambda f = \lambda \sup f, \lambda \geqslant 0, \ \sup f = -\inf(-f), \\ \inf_{x,y} \{f(x) + g(y)\} &= \inf_x f(x) + \inf_y g(y). \end{aligned}$$

The last rule does *not* hold for sup! Example: $f(x) = x$, $g(y) = -\infty$, then

$$\sup_{x,y} \{f(x) + g(y)\} = \sup_{x,y} (x - \infty) = \sup_{x,y} (-\infty) = -\infty \neq +\infty = +\infty - \infty = \sup_x f + \sup_y g.$$

Another special case to watch out for is that $\inf C \leqslant \sup C$ does *not* hold if $C = \emptyset$. Also $a \geqslant b \Leftrightarrow a - b \geqslant 0$ always holds, but $a \geqslant b \Leftrightarrow b - a \leqslant 0$ does not!

Definition 2.3. *(indicator function) For* $C \subseteq \mathbb{R}^n$*, denote*

$$\delta_C\colon \mathbb{R}^n \to \bar{\mathbb{R}}, \quad \delta_C(x) := \begin{cases} 0, & x \in C, \\ +\infty, & x \notin C. \end{cases}$$

Example 2.4. (constrained minimization via addition of indicator function): Assume $f\colon \mathbb{R}^n \to \mathbb{R}$, $C \subseteq \mathbb{R}^n$, $C \neq \emptyset$. Then

$$x' \text{ minimizes } f \text{ over } C \;\Leftrightarrow\; x' \text{ minimizes } f + \delta_C \text{ over } \mathbb{R}^n.$$

Definition 2.5. *(argmin, effective domain, proper) For $f\colon \mathbb{R}^n \to \bar{\mathbb{R}}$, denote*

1. $\operatorname{dom} f := \{x \in \mathbb{R}^n \,|\, f(x) < +\infty\}$ *(effective domain, set of feasible solutions)*
2. $\arg\min f := \begin{cases} \emptyset, & f \equiv +\infty, \\ \{x \in \mathbb{R}^n \,|\, f(x) = \inf f\}, & f < +\infty. \end{cases}$ *(set of minimizers/optimal solutions)*
3. *f is "proper"* $:\Leftrightarrow \operatorname{dom} f \neq \emptyset$ *and* $f(x) > -\infty \forall x \in \mathbb{R}^n$ *(i.e., $f \not\equiv +\infty$ and $f > -\infty$).*

By the definition of the $\arg\min$, if $x \in \arg\min f$ then $f(x) < +\infty$, however $f(x) = -\infty$ is possible. Proper functions are the "interesting" function for minimization: if $f = +\infty$ then the problem does not have any solution, and if there are x such that $f(x) = -\infty$ then $\arg\min f$ consists of exactly these points.

Definition 2.6. *(epigraphs, level sets) For $f\colon \mathbb{R}^n \to \bar{\mathbb{R}}$, denote*

$$\operatorname{epi} f \;:=\; \{(x, \alpha) \in \mathbb{R}^n \times \mathbb{R} \,|\, f(x) \leqslant \alpha\}.$$

Epigraphs are an alternative way to define functions and are often a convenient way to derive properties functions using theorem about sets. While every epigraph is a set, not every set is the epigraph of a function – in fact a set C is an epigraph iff for every x there is an $\alpha \in \bar{\mathbb{R}}$ such that $C \cap (x \times \mathbb{R}) = [\alpha, +\infty]$, i.e., all vertical one-dimensional sections must be closed upper half-lines. If f is proper then $\operatorname{epi} f$ is not empty and does not include a complete vertical line.

2.2 Existence of minimizers

Definition 2.7. *(lower-semicontinuity) For $f\colon \mathbb{R}^n \to \bar{\mathbb{R}}$, define*

$$\liminf_{x \to x'} f(x) \;:=\; \lim_{\delta \searrow 0} \inf_{\|x - x'\|_2 \leqslant \delta} f(x) \;=\; \lim_{k \to \infty} \inf_{\|x - x'\|_2 \leqslant 1/k} f(x).$$

f is "lower semi-continuous (lsc) at x'" $:\Leftrightarrow$

$$f(x') \;\leqslant\; \liminf_{x \to x'} f(x). \tag{2.1}$$

f is "lsc on $\mathbb{R}^n$" (or just "lsc") $:\Leftrightarrow$ *f is lsc at every $x' \in \mathbb{R}^n$.*

Proposition 2.8. *(sequence characterization of lower semi-continuity)*

$$\liminf_{x \to x'} f(x) \;=\; \min\{\alpha \in \bar{\mathbb{R}} \,|\, \exists (x^k) \to x' \colon f(x^k) \to \alpha\}. \tag{2.2}$$

In particular, f is lsc at x' iff

$$f(x') \;\leqslant\; \liminf_{k \to \infty} f(x^k) \quad \text{for all convergent sequences } x^k \to x'. \tag{2.3}$$

Proof. First part:

"$\leqslant$": We show that $\liminf_{x\to x'} f(x) \leqslant \alpha$ for all $\alpha \in S := \{\alpha \in \bar{\mathbb{R}} | \exists (x^k) \to x' \colon f(x^k) \to \alpha\}$. For such an α take $(x^k) \to x'$ from the definition of S. Then $f(x^k) \to \alpha$, so for all j exists k_j such that $x^{k_j} \in \mathcal{B}_{1/j}(x')$. Therefore

$$\begin{aligned} \inf_{x \in B_{1/j}(x')} f(x) &\leqslant f\big(x^{k_j}\big) \\ \overset{j\to\infty}{\Rightarrow} \underbrace{\lim_{j\to\infty} \inf_{x \in B_{1/j}(x')} f(x)}_{\liminf_{x\to x'} f(x)} &\leqslant \lim_{j\to\infty} f\big(x^{k_j}\big) \overset{x^k \to x}{=} \lim_{k\to\infty} f(x^k) = \alpha. \end{aligned}$$

"$\geqslant$": We show that there exists a sequence $x^k \to x'$ such that $f(x^k) \to \liminf_{x\to x'} f(x)$: if this is true, then "$\geqslant$" holds and we also have shown that the "min" notation is justified, i.e., S contains a minimal element.

For every k we can find $x^k \in \mathcal{B}_{1/k}(x')$ such that

$$f(x^k) \leqslant \inf_{x \in \mathcal{B}_{1/k}(x')} f(x) + \frac{1}{k}$$

(this requires that the lim inf is finite, but we can make a similar argument by adding $-\infty$ if it is $-\infty$). Since $f(x^k) \geqslant \inf_{x \in \mathcal{B}_{1/k}(x')} f(x)$ this implies

$$\inf_{x \in \mathcal{B}_{1/k}(x')} f(x) \leqslant f(x^k) \leqslant \inf_{x \in \mathcal{B}_{1/k}(x')} f(x) + \frac{1}{k}.$$

Since this holds for all k we can take $k \to \infty$ and obtain that

$$\lim_{k\to\infty} f(x^k) = \lim_{k\to\infty} \inf_{x \in \mathcal{B}_{1/k}(x')} f(x)$$

By definition $x^k \to x'$, so we have found the desired sequence.

Second part: This is based on the fact that we can identify the limits of converging sequences with the lim inf s of arbitrary sequences, because we can always find a subsequence converging to the lim inf.

"$\Leftarrow$": By 1., take any (x^k) from the definition of S such that $f(x^k) \to \alpha' := \min S$. Then $x^k \to x'$, therefore from $f(x') \leqslant \liminf_{k\to\infty} f(x^k) = \alpha' = \liminf_{x\to x'} f(x)$.

"$\Rightarrow$": Take any convergent sequence $x^k \to x'$ in (2.3). Choose a subsequence such that $f\big(x^{k_j}\big) \overset{j\to\infty}{\to} \liminf_{k\to\infty} f(x^k)$. Then $\big(x^{k_j}\big)_{j=1}^{\infty}$ is in S, therefore

$$\liminf_{k\to\infty} f(x^k) = \lim_{j\to\infty} f\big(x^{k_j}\big) \overset{(2.2)}{\geqslant} \alpha' \overset{1.}{=} \liminf_{x\to x'} f(x) \overset{(2.1)}{\geqslant} f(x').$$

□

Example 2.9.

1. $f(x) = \begin{cases} 1, & x > 0 \\ 0, & x \leqslant 0 \end{cases}$ is lsc,
2. $f(x) = \begin{cases} 1, & x \geqslant 0, \\ 0, & x < 0 \end{cases}$ is *not* lsc (but f is lsc in all $x \neq 0$),
3. $f(x) = \delta_C$ is lsc in the open sets $\operatorname{int} C$ and $\operatorname{ext} C$. It is lsc on $\mathbb{R}^n$ iff C is closed: $\liminf_{x\to x'} \delta_C(x)$ is always 0 for $x' \in \operatorname{bnd} C$, therefore δ_C is lsc iff $\delta_C(x) = 0$ for all $x \in \operatorname{bnd} C$, which is the case iff C is closed.

Theorem 2.10. *(semicontinuity and the epigraph) Let $f\colon \mathbb{R}^n \to \bar{\mathbb{R}}$. Then the following properties are equivalent:*

1. *f is lsc on $\mathbb{R}^n$,*
2. *$\operatorname{epi} f$ is closed in $\mathbb{R}^n \times \mathbb{R}$,*

3. the sublevelsets $\operatorname{lev}_{\leqslant\alpha} f := \{x \in \mathbb{R}^n \,|\, f(x) \leqslant \alpha\}$ *are closed in* $\mathbb{R}^n$ *for all* $\alpha \in \bar{\mathbb{R}}$.

Proof. *Idea* $(1 \Leftrightarrow 2)$: epi f can only be not closed along vertical lines.

1.⇒2.: Assume $(x^k, \alpha^k) \in \operatorname{epi} f$, $(x^k, \alpha^k) \to (x, \alpha)$, $\alpha \in \mathbb{R}$. Then $f(x^k) \leqslant \alpha^k$ (because $(x^k, \alpha^k) \in \operatorname{epi} f$), and $\liminf_{k\to\infty} f(x^k) \leqslant \liminf_{k\to\infty} \alpha^k = \alpha$. The second part of Lemma gives $f(x) \leqslant \liminf_{k\to\infty} f(x^k)$; therefore $f(x) \leqslant .... \leqslant \alpha$ and $(x, \alpha) \in \operatorname{epi} f$.

2.⇒3.: We have

$$\begin{aligned} & \operatorname{epi} f \text{ closed} \\ \Rightarrow\ & \operatorname{epi} f \cap (\mathbb{R}^n \times \{\alpha'\}) \text{ closed for all } \alpha' \in \mathbb{R} \\ \Leftrightarrow\ & \{(x, \alpha) \in \mathbb{R}^n \times \mathbb{R} | \alpha = \alpha', f(x) \leqslant \alpha\} \text{ closed for all } \alpha' \in \mathbb{R} \\ \Leftrightarrow\ & \{x \in \mathbb{R}^n \,|\, f(x) \leqslant \alpha'\} = \operatorname{lev}_{\leqslant\alpha'} f \text{ closed in } \mathbb{R}^n \text{ for all } \alpha' \in \mathbb{R}. \end{aligned}$$

This shows 3. for all $\alpha' \in \mathbb{R}$. For $\alpha' = +\infty$ we have $\operatorname{lev}_{\leqslant+\infty} f = \mathbb{R}^n$ which is always closed. For the case $\alpha' = -\infty$ we note that

$$\operatorname{lev}_{\leqslant-\infty} f \;=\; \bigcap_{k\in\mathbb{N}} \operatorname{lev}_{\leqslant k} f,$$

is also closed as the countable intersection of closed sets.

3.⇒1.: For any x', from Lemma 2.8 we get a sequence $x^k \to x'$ with $f(x^k) \to \liminf_{x\to x'} f(x) =: C$. If $C = +\infty$ we are done, since then $f(x') \leqslant \liminf_{x\to x'} f(x) = +\infty$ always holds.

Assume now $C \in \mathbb{R}$. Then for every $\varepsilon > 0$ we can find $K(\varepsilon)$ such that $f(x^k) \leqslant C + \varepsilon$ for all $k \geqslant K(\varepsilon)$. This implies

$$x^k \;\in\; \operatorname{lev}_{\leqslant C+\varepsilon} f,$$

for all $\varepsilon > 0$ and $k \geqslant K(\varepsilon)$. Since all level sets are closed by assumption and the subsequence converges to x', we get

$$\begin{aligned} x' \;&\in\; \operatorname{lev}_{\leqslant C+\varepsilon} f \quad \forall \varepsilon > 0 \\ \Leftrightarrow f(x') \;&\leqslant\; C + \varepsilon \quad \forall \varepsilon > 0, \end{aligned}$$

therefore

$$f(x') \;\leqslant\; C = \lim \inf_{x\to x'} f(x),$$

which shows that f is lsc in x'.

If $C = -\infty$ then we use a similar argument, replacing $\operatorname{lev}_{\leqslant C+\varepsilon}$ by $\operatorname{lev}_{\leqslant -1/\varepsilon}$. □

Definition 2.11. $f: \mathbb{R}^n \to \bar{\mathbb{R}}$ *is "level-bounded"* $:\Leftrightarrow \operatorname{lev}_{\leqslant\alpha} f$ *is bounded for all* $\alpha \in \mathbb{R}$.

Example 2.12. $f_1(x) := x^2$ is level-bounded. $f_2(x) := x$ is not bounded below and not level-bounded. $f_3(x) := 1/|x|$ (with the $+\infty$ extension at 0) is bounded below and not level-bounded. $f_4(x) = \min\{1, |x|\}$ is bounded from below and *not* level-bounded, since $\operatorname{lev}_{\leqslant\alpha} f = \mathbb{R}^n$ for $\alpha \geqslant 1$.

Proposition 2.13. $f: \mathbb{R}^n \to \bar{\mathbb{R}}$ *is level-bounded if and only if* $f(x^k) \to +\infty$ *for all sequences* (x^k) *satisfying* $\|x^k\|_2 \to +\infty$.

Proof. "⇒": Assume we have $\|x^k\|_2 \to +\infty$. Then for any $\alpha \in \mathbb{R}$ there is $K(\alpha)$ such that $x^k \notin \operatorname{lev}_{\leqslant\alpha} f$ for $k \geqslant K(\alpha)$, because all these sets are bounded. Therefore $f(x^k) > \alpha$ for all $k \geqslant K(\alpha)$. This holds for all α, therefore $f(x^k) \to +\infty$.

"⇐": if f is *not* level-bounded then there is an α such that $\operatorname{lev}_{\leqslant\alpha} f$ is unbounded. Thus we can find a sequence x^k in this set with $\|x^k\|_2 \to \infty$ and $f(x^k) \leqslant \alpha$. Therefore $f(x^k) \nrightarrow +\infty$. □

The property in Prop. 2.13 is also often referred to as *coercivity* in the literature.

Theorem 2.14. *(existence of minimizers): Assume $f\colon \mathbb{R}^n \to \bar{\mathbb{R}}$ is lsc, level-bounded, and proper. Then*

$$\inf_x f(x) \;\in\; (-\infty, +\infty),$$

and $\arg\min f$ *is nonempty and compact.*

Proof. f is proper $\Rightarrow f < +\infty \Rightarrow \inf f < +\infty$. We have

$$\begin{aligned} \arg\min f &= \{x \in \mathbb{R}^n \,|\, f(x) \leqslant \inf f\} \\ &= \{x \in \mathbb{R}^n \,|\, f(x) \leqslant \alpha \;\forall \alpha \in \mathbb{R}\colon \alpha > \inf f\} \\ &= \bigcap_{\alpha\in\mathbb{R}, \alpha > \inf f} \operatorname{lev}_{\leqslant\alpha} f. \end{aligned}$$

For any α in the intersection, $\operatorname{lev}_{\leqslant\alpha} f$ is closed since f is lsc (Thm. 2.10, 3.). But f is also level-bounded, therefore $\operatorname{lev}_{\leqslant\alpha} f$ is bounded. Together $\operatorname{lev}_{\leqslant\alpha} f$ is compact.

We can also make the intersection countable (use $\alpha = \inf f + 1/k$ if $\inf f \in \mathbb{R}$, and $\alpha = -k$ if $\inf f = -\infty$). All sets in the intersection are nonempty and compact. Therefore Cantor's intersection theorem states that their intersection is also nonempty (it is also compact).

(Why? Take any sequence with $x^k \in \operatorname{lev}_{\leqslant\alpha_k} f$, we can do this explicitly in $\mathbb{R}^n$ through inf. (x^k) has a converging subsequence because it lies completely in $\operatorname{lev}_{\leqslant\alpha_1} f$ which is compact. Denote the limit by x'. For every k we can find a "tail" of (x^k) that lies completely in $\operatorname{lev}_{\leqslant\alpha_k}$ (and still converges to x'), and because all these sets are closed we have $x' \in \operatorname{lev}_{\leqslant\alpha_k}$. Therefore x' is also contained in the intersection.)

The only thing that remains to be shown is that $\inf f > -\infty$. Assume that $\inf f = -\infty$. By the previous part, $\arg\min f \neq \emptyset$. Therefore there exists $x \in \arg\min f$, but then $f(x) = \inf f = -\infty$, which contradicts the properness of f. □

Remark 2.15. The proof of the theorem does not require full level-boundedness, it suffices to have $\operatorname{lev}_{\leqslant\alpha} f$ bounded and nonempty for *at least one* $\alpha \in \mathbb{R}$: closedness of the sublevelsets follows from f being lsc, and boundedness is only required for all $\operatorname{lev}_{\leqslant\alpha'}$ with $\alpha' \leqslant \alpha$, for which it automatically holds because $\operatorname{lev}_{\leqslant\alpha'} \subseteq \operatorname{lev}_{\leqslant\alpha}$.

An example for such a function is $f(x) = 1 - e^{-|x|}$, which is bounded from above by 1 and therefore not level-bounded, but it is lsc, proper, and attains its minimum in $x = 0$. All the sets $\operatorname{lev} \leqslant \alpha$ are bounded for $\alpha < 1$, and, as one would expect $\arg\min f = \{0\}$ is nonempty, closed and convex.

Proposition 2.16. *(lower semi-continuity of sums and scalar multiples):*

1. *f, g lsc and proper $\Rightarrow$ $f + g$ lsc*
2. *f lsc, $\lambda \geqslant 0 \Rightarrow \lambda f$ lsc,*
3. *$f\colon \mathbb{R}^n \to \bar{\mathbb{R}}$ lsc and $g\colon \mathbb{R}^m \to \mathbb{R}^n$ continuous $\Rightarrow$ $f \circ g$ lsc.*

Proof. Exercise. □

Chapter 3
Convexity

Definition 3.1. *(convex sets and functions)*

1. $f\colon \mathbb{R}^n \to \bar{\mathbb{R}}$ *is "convex"* $:\Leftrightarrow$

$$f((1-\tau)\,x+\tau y) \leqslant (1-\tau)\,f(x)+\tau f(y) \quad \forall x, y \in \mathbb{R}^n, \tau \in (0,1). \tag{3.1}$$

2. $C \subseteq \mathbb{R}^n$ *is "convex"* $:\Leftrightarrow$ δ_C *is convex* $\Leftrightarrow (1-\tau)\,x+\tau y \in C \quad \forall x, y \in C, \tau \in (0,1)$.
3. $f\colon \mathbb{R}^n \to \bar{\mathbb{R}}$ *is "strictly convex"* $:\Leftrightarrow$ f *convex and (3.1) holds strictly for all* $x \neq y$ *with* $f(x), f(y) \in \mathbb{R}$ *and for all* $\tau \in (0,1)$.

Remark 3.2. $C \subseteq \mathbb{R}^n$ is convex iff for every two points $x, y \in C$ the whole line segment between them is contained in C:

$$\{(1-\tau)\,x+\tau y | t \in (0,1)\} \;\subseteq\; C \quad \forall x, y \in C.$$

If $-f$ is convex then we say that f is *concave*. Not that this does not correspond to reversing the inequality sign in (3.1) alone due to the $+\infty-\infty=+\infty$ convention (Def. 2.1), but also requires to change the convention to $+\infty-\infty=-\infty$ in addition to the reversed sign. In order to avoid this problem we prefer to say that $-f$ is convex to saying that f is concave.

Example 3.3.

1. $\mathbb{R}^n$ is convex,
2. $\{x \in \mathbb{R}^n | x > 0\}$ is convex,
3. $\{x \in \mathbb{R}^n |\, \|x\|_2 \leqslant 1\}$ is convex,
4. $\{x \in \mathbb{R}^n |\, \|x\|_2 \leqslant 1, x \neq 0\}$ is *not* convex,
5. the half-spaces $\{x | a^\top x + b \geqslant 0\}$ are convex,
6. $f(x) = a^\top x + b$ is convex (inequality holds as an equality) but *not* strictly convex,
7. $f(x) = \|x\|_2^2$ is strictly convex,
8. $f(x) = \|x\|_2$ is convex but *not* strictly convex.

Definition 3.4. *(convex combination) Assume* $x_0, \ldots, x_m \in \mathbb{R}^n$ *and* $\lambda_0, \ldots, \lambda_m \geqslant 0$, $\sum_{i=0}^m \lambda_i = 1$. *We call the linear combination*

$$\sum_{i=0}^{m} \lambda_i\, x_i$$

a "convex combination" of the points $x_0, \ldots, x_m$.

Theorem 3.5. *(convex combinations and Jensen's inequality)*

1. $f\colon \mathbb{R}^n \to \bar{\mathbb{R}}$ *convex* $\Leftrightarrow$

$$f\left(\sum_{i=0}^{m} \lambda_i x_i\right) \leqslant \sum_{i=0}^{m} \lambda_i f(x_i) \text{ for all } m \geqslant 0, x_i \in \mathbb{R}^n, \lambda_i \geqslant 0, \sum_{i=0}^{m} \lambda_i = 1. \tag{3.2}$$

2. $C \subseteq \mathbb{R}^n$ *convex* $\Leftrightarrow$ C *contains all convex combinations of its elements.*

Proof.

1. "$\Leftarrow$": For $m=1$, (3.2) is the convexity condition.

   "$\Rightarrow$": Induction: $m=0$ is trivial, and the case $m=1$ follows from convexity. Assume that (3.2) holds for all $m' < m$ for some $m \geqslant 2$, and consider an arbitrary convex combination $x = \lambda_0 x_0 + \ldots + \lambda_m x_m$.

   W.l.o.g. assume that all $\lambda_i > 0$ (if not we can remove it from the sum together with the corresponding x_i) and all $\lambda_i < 1$ (if not then all other $\lambda_i = 0$ and we have the trivial case $m=0$). Then

$$\begin{aligned} x &= \lambda_0 x_0 + \ldots + \lambda_m x_m \\ &= (1-\lambda_m) \sum_{i=0}^{m-1} \frac{\lambda_i}{1-\lambda_m} x_i + \lambda_m x_m. \end{aligned}$$

   Thu

$$\begin{aligned} f(x) &\overset{m'=1}{\leqslant} (1-\lambda_m) f\left(\sum_{i=0}^{m-1} \frac{\lambda_i}{1-\lambda_m} x_i\right) + \lambda_m f(x_m) \\ &\overset{m'=m-1}{\leqslant} (1-\lambda_m) \sum_{i=0}^{m-1} \frac{\lambda_i}{1-\lambda_m} f(x_i) + \lambda_m f(x_m) \\ &= \sum_{i=0}^{m} \lambda_i f(x_i). \end{aligned}$$

2. C is convex iff δ_C convex, and

$$\delta_C\left(\sum_{i=0}^{m} \lambda_i x_i\right) \leqslant \sum_{i=0}^{m} \lambda_i \delta_C(x_i)$$

   iff there exists $x_i \notin C$ or $\sum_{i=0}^{m} \lambda_i x_i \in C$. Therefore (3.2) holds for δ_C iff C contains all convex combinations of its elements, and 2. follows from 1.

□

Proposition 3.6. *(effective domain of convex functions)*

$$f\colon \mathbb{R}^n \to \bar{\mathbb{R}} \text{ convex } \Rightarrow \text{ dom } f \text{ convex}.$$

Proof. Let $x, y \in \text{dom } f$, $\tau \in (0,1)$. Then $f(x), f(y) < +\infty$ and by convexity $f((1-\tau)x + \tau y) \leqslant (1-\tau) f(x) + \tau f(y) < \infty$, therefore $(1-\tau)x + \tau y \in \text{dom } f$. □

Proposition 3.7. *(convexity of the epigraph)*

1. $f\colon \mathbb{R}^n \to \bar{\mathbb{R}}$ *convex* $\Leftrightarrow$ epi f *is convex in* $\mathbb{R}^n \times \mathbb{R}$.
2. $f\colon \mathbb{R}^n \to \bar{\mathbb{R}}$ *convex* $\Leftrightarrow$ $\{(x,\alpha) \in \mathbb{R}^n \times \mathbb{R} | f(x) < \alpha\}$ *(strict epigraph set) is convex in* $\mathbb{R}^n \times \mathbb{R}$.

Proof.

1.

$$\begin{aligned}
\text{epi } f \text{ convex} &\Leftrightarrow (1-\tau)(x_0,\alpha_0)+\tau(x_1,\alpha_1)\in \text{epi } f \ \forall (x_0,\alpha_0),(x_1,\alpha_1)\in \text{epi } f,\\
&\quad \tau\in(0,1)\\
&\Leftrightarrow f((1-\tau)x_0+\tau x_1)\leqslant (1-\tau)\alpha_0+\tau\alpha_1 \,\forall x_0,x_1,\forall \alpha_0\geqslant f(x_0),\\
&\quad \alpha_1\geqslant f(x_1),\tau\in(0,1)\\
&\Leftrightarrow f((1-\tau)x_0+\tau x_1)\leqslant (1-\tau)f(x_0)+\tau f(x_1)\,\forall x_0,x_1\forall\tau\in(0,1)
\end{aligned}$$

(last equivalence: "⇒" follows immediately with $\alpha_i = f(x_i)$, "⇐" follows because $\tau\in(0,1)$ and therefore $(1-\tau)\geqslant 0$ and $\tau\geqslant 0$).

2. similar with strict inequalities.

□

Proposition 3.8. *(convexity of sublevelsets) $f\colon \mathbb{R}^n\to\bar{\mathbb{R}}$ convex $\Rightarrow$ $\text{lev}_{\leqslant\alpha}f$ convex for all $\alpha\in\bar{\mathbb{R}}$.*

Proof. $\alpha=+\infty$ is trivial ($\text{lev}_{\leqslant+\infty}f=\mathbb{R}^n$). Let $\alpha<+\infty$, $x,y\in\text{lev}_{\leqslant\alpha}f$ and $\tau\in(0,1)$, then

$$\begin{aligned}
f((1-\tau)x+\tau y) &\overset{f\text{ convex}}{\leqslant} (1-\tau)f(x)+\tau f(y)\\
&\overset{\text{lev}_{\leqslant\alpha}}{\leqslant} (1-\tau)\alpha+\tau\alpha\\
&= \alpha.
\end{aligned}$$

Therefore $(1-\tau)x+\tau y\in\text{lev}_{\leqslant\alpha}f$. □

Theorem 3.9. *(global optimality) Assume $f\colon\mathbb{R}^n\to\bar{\mathbb{R}}$ is convex. Then*

1. *$\arg\min f$ is convex.*
2. *x is a local minimizer of f $\Rightarrow$ x is a global minimizer of f.*
3. *f strictly convex and proper $\Rightarrow$ f has at most one global minimizer.*

Proof.

1. If $f\not\equiv+\infty$ then $\arg\min f=\text{lev}_{\leqslant\inf f}f$ and we can use Prop. 3.8. If $f=+\infty$ then $\arg\min f=\emptyset$, which is convex.

2. Assume x is a local minimizer and $y\in\mathbb{R}^n$, and assume $f(y)<f(x)$, i.e., x is *not* a global minimizer. By convexity:

$$\begin{aligned}
f((1-\tau)x+\tau y) &\leqslant (1-\tau)f(x)+\tau f(y)\,\forall\tau\in(0,1)\\
\Rightarrow f(x+\tau(y-x)) &< (1-\tau)f(x)+\tau f(x)\,\forall\tau\in(0,1)\\
\Rightarrow f(x+\tau(y-x)) &< f(x)\,\forall\tau\in(0,1).
\end{aligned}$$

$\Rightarrow x$ cannot be a local minimizer.

3. Assume x, y are global minimizers. Then $f(x)=f(y)=\inf f\in\mathbb{R}$ (f is proper, Thm. 2.14). If $x\neq y$ then by strict convexity:

$$\begin{aligned}
f((1-\tau)x+\tau y) &< (1-\tau)f(x)+\tau f(y) \quad \forall\tau\in(0,1)\\
\Rightarrow f((1-\tau)x+\tau y) &< \inf f \quad \forall\tau\in(0,1).
\end{aligned}$$

This is impossible, therefore $x = y$. □

Proposition 3.10. *(operations that preserve convexity) Let I be an arbitrary index set. Then*

1. *$f_i, i \in I$ convex $\Rightarrow$ $f(x) := \sup_{i\in I} f_i(x)$ is convex,*
2. *$f_i, i \in I$ strictly convex, I finite $\Rightarrow$ $f(x) := \sup_{i\in I} f_i(x)$ is strictly convex,*
3. *$C_i, i \in I$ convex $\Rightarrow$ $\bigcap_{i\in I} C_i$ convex,*
4. *$f_k, k \in \mathbb{N}$ convex $\Rightarrow$ $f(x) := \limsup_{k\to\infty} f_k(x)$ is convex.*

Proof. Exercise.

1. from the definition of convexity and $a_i \leqslant b_i \Rightarrow \sup_{i\in I} a_i \leqslant \sup_{i\in I} b_i$.
2. same with strict inequalities for finite sup.
3. $\delta_{\cap_{i\in I} C_i} = \sup_{i\in I} \delta_{C_i}$ and 1.
4. similar to 1. □

Example 3.11.

1. The union of convex sets is generally *not* convex (but can be): $C = [0,1]$, $D = [2,3]$.
2. $f\colon \mathbb{R}^n \to \mathbb{R}$ convex and C convex $\Rightarrow$ $f + \delta_C$ is convex $\Rightarrow$ set of minimizers of f on C is convex (Thm. 3.9).
3. $f(x) = |x| = \max\{x, -x\}$ is convex (Prop. 3.10).
4. $f(x) = \|x\|_2 = \sup_{\|y\|_2\leqslant 1} y^\top x$ is convex (Prop. 3.10) (similarly: $f(x) = \|x\|_p$ is convex, look at dual norm $\|\cdot\|_q$ with $1/p + 1/q = 1$). It is *not* strictly convex: set $x = 0, y \neq 0$, then $f(x/2 + y/2) = f(y/2) = f(x)/2 + f(y)/2$.

Theorem 3.12. *(derivative tests) Assume $C \subseteq \mathbb{R}^n$ is open and convex, and $f\colon C \to \mathbb{R}$ (i.e., real-valued!) is differentiable. Then the following conditions are equivalent:*

1. *f is [strictly] convex,*
2. *$\langle y - x, \nabla f(y) - \nabla f(x)\rangle \geqslant 0$ for all $x, y \in C$ [and > 0 if $x \neq y$],*
3. *$f(x) + \langle y - x, \nabla f(x)\rangle \leqslant f(y)$ for all $x, y \in C$ [and $< f(y)$ if $x \neq y$],*
4. *if f is additionally twice differentiable: $\nabla^2 f(x)$ is positive semidefinite for all $x \in C$.*

If f is twice differentiable and $\nabla^2 f$ is positive definite, then f is strictly convex. The opposite does not *hold.*

Proof. Exercise, reduce to the one-dimensional case. □

Remark 3.13. The second condition is a *monotonicity* condition on the gradient: in one dimension it becomes

$$(y - x)\,(f'(y) - f'(x)) \geqslant 0,$$

which is equivalent to $y > x \Leftrightarrow f'(y) \geqslant f'(x)$. This means that the derivative must increase when going towards larger values of x. The third condition says that f is never below any of its local linear approximations. The fourth condition in one dimension means $f'' \geqslant 0$, the graph is "curved upwards".

Proposition 3.14.

1. *(nonnegative linear combination) Assume* $f_1, ..., f_m \colon \mathbb{R}^n \to \bar{\mathbb{R}}$ *are convex,* $\lambda_1, ..., \lambda_m \geqslant 0$*. Then*

$$f := \sum_{i=1}^{m} \lambda_i f_i$$

is convex. If at least one of the f_i *with* $\lambda_i > 0$ *is strictly convex, then* f *is strictly convex.*

2. *(separable sum) Assume* $f_i \colon \mathbb{R}^{n_i} \to \bar{\mathbb{R}}, i = 1, ..., m$ *are convex. Then*

$$f \colon \mathbb{R}^{n_1} \times ... \times \mathbb{R}^{n_m} \to \bar{\mathbb{R}},$$
$$f(x_1, ..., x_m) := \sum_{i=1}^{m} f_i(x_i)$$

is convex. If all f_i *are strictly convex, then* f *is strictly convex.*

3. *(linear composition) Assume* $f \colon \mathbb{R}^m \to \bar{\mathbb{R}}$ *is convex,* $A \in \mathbb{R}^{m \times n}$*, and* $b \in \mathbb{R}^m$*. Then*

$$g(x) \;:=\; f(A\,x + b)$$

is convex.

Proof.

1. From Def. 3.1:

$$\begin{aligned} f((1-\tau)\,x + \tau y) &= \sum_{i=1}^{m} \lambda_i\, f_i((1-\tau)\,x + \tau y) \\ &\leqslant[<] \sum_{i=1}^{m} \lambda_i\,((1-\tau)\,f_i(x) + \tau\, f_i(y)) \\ &= (1-\tau)\,f(x) + \tau f(y). \end{aligned}$$

2. From Def. 3.1:

$$\begin{aligned} f((1-\tau)\,x + \tau y) &= \sum_{i=1}^{m} f_i((1-\tau)\,x_i + \tau y_i) \\ &\leqslant[<] \sum_{i=1}^{m} ((1-\tau)\,f_i(x_i) + \tau\, f_i(y_i)) \\ &= (1-\tau)\,f(x_i) + \tau f(y_i). \end{aligned}$$

3. From Def. 3.1:

$$\begin{aligned} g((1-\tau)\,x + \tau y) &= f(A\,((1-\tau)\,x + \tau y) + b) \\ &= f((1-\tau)(A\,x + b) + \tau\,(A\,y + b)) \\ &\underset{f \text{ convex}}{\leqslant} (1-\tau)\,f(A\,x + b) + \tau f(A\,y + b) \\ &= (1-\tau)\,g(x) + \tau g(y). \end{aligned}$$

□

Proposition 3.15. *(convexity properties of sets)*

1. $C_1, ..., C_m$ *convex* $\Rightarrow C_1 \times ... \times C_m$ *convex.*
2. $C \subseteq \mathbb{R}^n$ *convex,* $A \in \mathbb{R}^{m \times n}$*,* $b \in \mathbb{R}^m$*,* $L(x) := A\,x + b \Rightarrow L(C)$ *convex.*
3. $C \subseteq \mathbb{R}^m$ *convex,* $A \in \mathbb{R}^{m \times n}$*,* $b \in \mathbb{R}^m$*,* $L(x) := A\,x + b \Rightarrow L^{-1}(C)$ *convex.*

4. *C_1, C_2 convex $\Rightarrow C_1 + C_2$ convex.*
5. *C convex, $\lambda \in \mathbb{R} \Rightarrow \lambda C$ convex.*

Proof. Exercise. □

Definition 3.16. *For any set $S \subseteq \mathbb{R}^n$ and any point $x \in \mathbb{R}^n$, we define the "projection of x onto S" as*

$$\Pi_S(y) := \arg\min_{x \in S} \|x - y\|_2.$$

Proposition 3.17. *Assume that $C \subseteq \mathbb{R}^n$ is convex, closed, and $C \neq \emptyset$. Then Π_C is single-valued, i.e., the projection of x onto C is unique for every $x \in \mathbb{R}^n$.*

Proof. We can rewrite the problem using a more convenient (differentiable) objective:

$$\begin{aligned} \Pi_C(y) &= \arg\min_{x \in C} \frac{1}{2} \|x - y\|_2^2 \\ &= \arg\min_{x} \frac{1}{2} \|x - y\|_2^2 + \delta_C(x) \end{aligned}$$

Quick proof:

- $x \mapsto \frac{1}{2}\|x-y\|_2^2$ is lsc (it is continuous) and level-bounded ($f(x) \to +\infty$ as $\|x\|_2 \to +\infty$) and proper (never $-\infty$, not always $+\infty$)
  $\delta_C(x)$ is lsc because C is closed (Ex. 2.9, alternatively Thm. 2.10: epi δ_C is closed $\Rightarrow$ δ_C lsc).
  $\Rightarrow f(x) := \frac{1}{2}\|x-y\|_2 + \delta_C$ is lsc, level-bounded and proper
  $\Rightarrow \arg\min f \neq \emptyset$ by Thm. 2.14.
- $x \mapsto \frac{1}{2}\|x-y\|_2^2$ is strictly convex (derivative is $x-y$, thus $\langle y-x, \nabla f(y) - \nabla f(x)\rangle = \|y-x\|^2 > 0$ if $x \neq y$, this yields strict convexity using Thm. 3.12). We could also use the second-order criterion in Thm. 3.12 and verify that $\nabla^2\left(\frac{1}{2}\|\cdot - y\|_2^2\right) = I$ is positive definite.
  Alternative proof without using differentiability: $s \mapsto (s-t)^2$ is strictly convex (prove this directly) $\Rightarrow x \mapsto \|x-y\|_2^2$ is strictly convex (sum of strictly convex functions) $\Rightarrow x \mapsto \frac{1}{2}\|x-y\|_2^2$ is strictly convex (positive multiple of strictly convex function)
  δ_C is convex because C is convex (Rem. 3.2) and proper because $C \neq \emptyset$ (otherwise δ_C would be $+\infty$)
  $\Rightarrow f$ is strictly convex (sum of strictly convex and convex function)
  $\Rightarrow f$ has at most one minimizer by Thm. 3.9.

□

Definition 3.18. *(convex hull) For an arbitrary set $S \subseteq \mathbb{R}^n$,*

$$\operatorname{con} S := \bigcap_{C \text{ convex}, S \subseteq C} C$$

is the "convex hull" of S.

Remark 3.19. The convex hull $\operatorname{con} S$ is the smallest convex set that contains S: $\operatorname{con} S$ is convex by Prop. 3.10 (intersection of convex sets) and every set C that is convex and contains S also contains $\operatorname{con} S$ by definition.

Theorem 3.20. *(convex hulls from convex combinations) Assume $S \subseteq \mathbb{R}^n$, then*

$$\operatorname{con} S \;=\; \left\{ \sum_{i=0}^{p} \lambda_i x_i \,\middle|\, x_i \in S, \lambda_i \geqslant 0, \sum_{i=0}^{p} \lambda_i = 1, p \geqslant 0 \right\}.$$

Proof. We denote the right-hand side by D and need to show that $\operatorname{con} S = D$.

"$\supseteq$": $S \subseteq \operatorname{con} S$. $\operatorname{con} S$ is convex $\Rightarrow$ (Thm 3.5): $\operatorname{con} S$ contains all convex combinations of points in $\operatorname{con} S \Rightarrow \operatorname{con} S$ contains all convex combinations of points in $S \Leftrightarrow D \subseteq \operatorname{con} S$.

"$\subseteq$": if $x, y \in D$ then for some $x_i, y_i, \lambda_i, \mu_i$:

$$\begin{aligned}(1-\tau)\,x + \tau\, y \;&=\; (1-\tau)\sum_{i=0}^{m_x} \lambda_i x_i + \tau \sum_{i=0}^{m_y} \mu_i y_i \\ &=\; \sum_{i=0}^{m_x} (1-\tau)\lambda_i x_i + \sum_{i=0}^{m_y} \tau \mu_i y_i.\end{aligned}$$

with

$$\begin{aligned}\sum_{i=0}^{m_x} (1-\tau)\,\lambda_i + \sum_{i=0}^{m_y} \tau \lambda_i \;&=\; (1-\tau)\sum_{i=0}^{m_x} \lambda_i + \tau \sum_{i=0}^{m_y} \mu_i \\ &=\; (1-\tau)\cdot 1 + \tau \cdot 1 \\ &=\; 1.\end{aligned}$$

$\Rightarrow$ convex combinations of elements in D are in D.

$\Rightarrow$(Thm 3.5) D is convex.

$\Rightarrow$(Def. 3.18, $S \subseteq D$) $\operatorname{con} S \subseteq D$. □

Definition 3.21. *(closure, interior) For a set $C \subseteq \mathbb{R}^n$, denote*

$$\begin{aligned}\operatorname{cl} C \;&:=\; \{x \in \mathbb{R}^n \,|\, \textit{for all (open) neighborhoods } N \textit{ of } x \textit{ we have } N \cap C \neq \emptyset\}, \\ \operatorname{int} C \;&:=\; \{x \in \mathbb{R}^n \,|\, \textit{there exists an (open) neighborhood } N \textit{ of } x \textit{ such that } N \subseteq C\}, \\ \operatorname{bnd} C \;&:=\; \operatorname{cl} C \setminus \operatorname{int} C.\end{aligned}$$

Remark 3.22. The closure is to closedness what the convex hull is to convexity:

$$\operatorname{cl} C \;=\; \bigcap_{S \text{ closed}, C \subseteq S} S.$$

Chapter 4
Cones and Generalized Inequalities

Definition 4.1. *(cones)* $K \subseteq \mathbb{R}^n$ *"cone"* $:\Leftrightarrow$

$$0 \in K, \quad \lambda x \in K \quad \forall x \in K, \lambda \geqslant 0.$$

A cone K *is "pointed"* $:\Leftrightarrow$

$$x_1 + \ldots + x_m = 0, \quad x_i \in K \; \forall i \in \{1, \ldots, m\} \;\Rightarrow\; x_1 = \ldots = x_m = 0.$$

Note that cones can also be *nonconvex*, such as the cone $K = (\mathbb{R}_{\geqslant 0} \times \{0\}) \cup (\{0\} \times \mathbb{R}_{\geqslant 0})$.

Proposition 4.2. *(convex cones) Assume* $K \subseteq \mathbb{R}^n$ *is an arbitrary set. Then the following conditions are equivalent:*

1. K *is a convex cone,*
2. K *is a cone and* $K + K \subseteq K$*,*
3. $K \neq \emptyset$ *and* $\sum_{i=0}^m \alpha_i x_i \in K$ *for all* $x_i \in K$ *and* $\alpha_i \geqslant 0$ *(not necessarily summing to* 1*).*

Proof. Exercise, idea: $x = \sum_{i=0}^m \alpha_i x_i \in K$, $\alpha_i \geqslant 0 \Leftrightarrow \sum_{i=0}^m \frac{\alpha_i}{\sum_j \alpha_j} x \in K$, and the latter is a convex combination with coefficients summing to 1 (if there is at least one $\alpha_i \neq 0$). □

Proposition 4.3. *(pointed cones) Assume* K *is a convex cone. Then*

$$K \text{ pointed} \;\Leftrightarrow\; K \cap -K = \{0\}.$$

Proof. "⇒": $x \in K \cap -K \Rightarrow x, -x \in K$. Then $0 = x + (-x) \overset{\text{Def. 4.1}}{\Rightarrow} x = -x = 0$.

"⇐": K not pointed ⇒ there exist $x_1 + \ldots + x_m = 0$, $x_i \neq 0$, $x_i \in K$ (w.l.o.g. with $x_1 \neq 0$) ⇒ $x_1 + (x_2 + \ldots + x_m) = 0$. Therefore $x_2 + \ldots + x_m = -x_1$, and $x_2 + \ldots + x_m \in K$ (Prop. 4.2) ⇒ $x_1 \in K \cap -K \Rightarrow K \cap -K \neq \{0\}$. □

Proposition 4.4. *(generalized inequalities) For a closed convex cone* $K \subseteq \mathbb{R}^n$ *we define the "generalized inequality"*

$$x \geqslant_K y \;:\Leftrightarrow\; x - y \in K.$$

Then

1. $x \geqslant_K x$ *(reflexivity),*
2. $x \geqslant_K y$, $y \geqslant_K z \Rightarrow x \geqslant_K z$ *(transitivity),*
3. $x \geqslant_K y \Rightarrow -y \geqslant_K -x$*,*
4. $x \geqslant_K y, \lambda \geqslant 0 \Rightarrow \lambda x \geqslant_K \lambda y$*,*
5. $x \geqslant_K y, x' \geqslant_K y' \Rightarrow x + x' \geqslant_K y + y'$*,*

6. *If $x^k \to x$ and $y^k \to y$ with $x^k \geqslant_K y^k$ for all $k \in \mathbb{N}$, then $x \geqslant_K y$.*

7. *$x \geqslant_K y, y \geqslant_K x \Rightarrow x = y$ for all $x, y \in \mathbb{R}^n$ (antisymmetry) holds iff K is pointed.*

If "$\geqslant$" is a relation on $\mathbb{R}^n$ satisfying 1.-6., then it can be represented as $\geqslant_K$ for a closed convex cone.

Proof. 1.-6. from the definition of a cone. Converse: recover $K = \{x \in \mathbb{R}^n | x \geqslant 0\}$, then $x \geqslant y \overset{5.}{\Leftrightarrow} x - y \geqslant 0 \Leftrightarrow x - y \in K \Leftrightarrow x \geqslant_K y$. Then show that K is a closed convex cone. Antisymmetry: $x \geqslant_K y, y \geqslant_K x \Leftrightarrow x - y \in K, y - x \in K \Leftrightarrow x - y \in K \cap -K$. This holds for all x, y iff $K \cap -K = \{0\}$, which by Prop. 4.3 is equivalent to K being pointed. □

Definition 4.5. *(conic program) For any pointed, closed, convex cone $K \subseteq \mathbb{R}^m$, a matrix $A \in \mathbb{R}^{m \times n}$ and vectors $c \in \mathbb{R}^n$, $b \in \mathbb{R}^m$, we define the "conic program" or "conic problem" (CP)*

$$\begin{aligned} &\inf_x c^\top x \\ &s.t.\ \ A\,x \geqslant_K b. \end{aligned}$$

Example 4.6. (standard cone): The "standard cone"

$$K_n^{\mathrm{LP}} \ := \ \{x \in \mathbb{R}^n | x_1, ..., x_n \geqslant 0\}$$

is a pointed, closed, convex cone. The associated conic program is the "linear program" (LP)

$$\begin{aligned} &\inf_x c^\top x \\ &\text{s.t. } A\,x \geqslant b. \end{aligned}$$

This is surprisingly powerful: for example, the problem

$$\min_x |x_1 - x_2| \quad \text{s.t.} \quad x_1 = -1, x_2 \geqslant 0$$

could be written as

$$\min_{x,y} y \quad \text{s.t.} \quad y \geqslant |x_1 - x_2|, x_1 \geqslant -1, x_1 \leqslant -1, x_2 \geqslant 0,$$

then

$$\min_{x,y} \ y \quad \text{s.t.} \quad y \geqslant x_1 - x_2, y \geqslant x_2 - x_1, x_1 \geqslant -1, -x_1 \geqslant 1, x_2 \geqslant 0,$$

and finally

$$\begin{aligned} \min_{(x_1,x_2,y) \in \mathbb{R}^3} \quad & (0,0,1) \begin{pmatrix} x_1 \\ x_2 \\ y \end{pmatrix} \\ \text{s.t.} \quad & \begin{pmatrix} -1 & 1 & 1 \\ 1 & -1 & 1 \\ 1 & 0 & 0 \\ -1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ y \end{pmatrix} \geqslant \begin{pmatrix} 0 \\ 0 \\ -1 \\ 1 \\ 0 \end{pmatrix}. \end{aligned}$$

Example 4.7. (second-order cone): The "second-order cone" (also called "Lorentz cone", "ice-cream cone")

$$K_n^{\mathrm{SOCP}} \ := \ \left\{ x \in \mathbb{R}^n \middle| x_n \geqslant \sqrt{x_1^2 + ... + x_{n-1}^2} \right\}$$

is a pointed, closed, convex cone. Conic programs with $K = K_{n_1}^{\text{SOCP}} \times ... \times K_{n_l}^{\text{SOCP}}$ are called "second-order conic programs" (SOCP).

Example:

$$\min_x \ \|x\|_2 \quad \text{s.t.} \quad x_1 + x_2 \geqslant 1.$$

This can be rewritten as a second-order cone program:

$$\begin{array}{ll} \min\limits_{x,y} & y \\ \text{s.t.} & y \geqslant \|x\|_2, x_1 + x_2 - 1 \geqslant 0. \\ & \Leftrightarrow I_3 \begin{pmatrix} x \\ y \end{pmatrix} \geqslant_{K_3^{\text{SOCP}}} 0, \quad \begin{pmatrix} 1 & 1 & 0 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} \geqslant_{K_1^{\text{SOCP}}} \begin{pmatrix} 1 \end{pmatrix}. \end{array}$$

Example 4.8. (positive semidefinite cone): The "positive semidefinite cone"

$$K_n^{\text{SDP}} := \{X \in \mathbb{R}^{n \times n} | X \text{ symmetric positive semidefinite}\}$$

is a pointed, closed, convex cone. Conic programs with $K = K_{n_1}^{\text{SDP}} \times ... \times K_{n_l}^{\text{SDP}}$ are called "semidefinite programs" (SDP):

$$\begin{array}{ll} \inf\limits_{x \in \mathbb{R}^n} & c^\top x \\ \text{s.t.} & A\,x - b \text{ positive semidefinite.} \end{array}$$

Here A is a linear operator $A: \mathbb{R}^n \to \mathbb{R}^{m \times m}$, and $b \in \mathbb{R}^{m \times m}$. Often x and c are also written as matrices $X, C \in \mathbb{R}^{n \times n}$ with the inner product $\langle C, X \rangle := \sum_{i,j} C_{ij} X_{ij}$ replacing $c^\top x$.

Proof. The set of symmetric matrices $\mathbb{R}_{\text{sym}}^{n \times n}$ is a closed convex cone: every set

$$K_{ij} := \{A \in \mathbb{R}^{n \times n} | A_{ij} = A_{ji}\}$$

is a closed convex cone. Finite intersections of closed convex cones are closed convex cones, so $\mathbb{R}_{\text{sym}}^{n \times n} = \bigcap_{i \neq j} K_{ij}$ is a closed convex cone.

Closedness: $x^\top A^k x \geqslant 0$ for all x and k implies $x^\top A\,x \geqslant 0$ for $A^k \to A$.

Cone: follows immediately, $0 \in K$ and A psd $\Rightarrow \lambda A$ psd if $\lambda \geqslant 0$ (nonneg. eigenvalues)

Convex cone: A, $B \in K$. Then A, B are symmetric. For every $x \in \mathbb{R}^n$: $x^\top A\,x \geqslant 0$, $x^\top B\,x \geqslant 0$ (A, B pos. semidef.) $\Rightarrow x^\top (A + B)\,x \geqslant 0 \Rightarrow$ $A + B$ pos. semidef. $\Rightarrow A + B \in K$. Therefore (Prop. 4.2) K is a convex cone.

Pointed: $A \in K, A \in -K \Rightarrow$ A symm., all eigenvalues $\geqslant 0$, all eigenvalues $\leqslant 0 \Rightarrow K = 0$. $\square$

Chapter 5
Subgradients

Definition 5.1. *(set-valued mappings) X, U sets. Then*

$$S\colon X \to 2^U$$

is a "set-valued mapping $S\colon X \rightrightarrows U$".

Remark 5.2. The set-valued mappings are the relations on $X \times U$:

$$S \rightsquigarrow \operatorname{gph} S = \{(x,u) | u \in S(x)\} =: R \subseteq X \times U \;\rightsquigarrow\; S(x) = \{u | (x,u) \in R = \operatorname{gph} S\}.$$

Definition 5.3. *(domain, range, inverse) $S\colon X \rightrightarrows U$. Then*

$$\begin{aligned} S^{-1}(u) &:= \{x \in X | u \in S(x)\} \\ \operatorname{dom} S &:= \{x \in X | S(x) \neq \emptyset\}, \\ \operatorname{rge} S &:= \{u \in U | \exists x \in X\colon u \in S(x)\}. \end{aligned}$$

Remark 5.4. $\operatorname{gph} S^{-1}$ is the "transpose" of the graph of S:

$$u \in S(x) \Leftrightarrow (x,u) \in \operatorname{gph} S \Leftrightarrow (u,x) \in \operatorname{gph} S^{-1} \Leftrightarrow x \in S^{-1}(u).$$

This also shows that $(S^{-1})^{-1} = S$ always holds ($(S^{-1})^{-1}(x) = \{u | x \in S^{-1}(u)\} = \{u | u \in S(x)\} = S(x)$).

Definition 5.5. *(subgradients) For any $f\colon \mathbb{R}^n \to \bar{\mathbb{R}}$ and $x \in \mathbb{R}^n$,*

$$\partial f(x) := \{v \in \mathbb{R}^n | f(x) + \langle v, y - x\rangle \leqslant f(y)\ \forall y \in \mathbb{R}^n\}$$

is the set of "subgradients of f at x". The set-valued mapping $\partial f\colon \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ is the "subdifferential mapping of f".

Note that for constant functions $f \equiv c \in \mathbb{R}$ we get $\partial f(x) = \{0\}$ for all $x \in \mathbb{R}^n$. However, for the constant function $f \equiv +\infty$ and $f \equiv -\infty$, from $+\infty \geqslant +\infty$ and $-\infty \geqslant -\infty$ (in fact equality holds), we conclude that

$$\partial(+\infty)(x) = \mathbb{R}^n, \quad \partial(-\infty)(x) = \mathbb{R}^n, \quad \forall x \in \mathbb{R}^n.$$

Proposition 5.6. *(subgradients of differentiable functions) Assume that $f, g\colon \mathbb{R}^n \to \bar{\mathbb{R}}$ are convex. Then:*

1. *if f is differentiable at x, then $\partial f(x) = \{\nabla f(x)\}$,*
2. *if f is differentiable at x and $g(x) \in \mathbb{R}$, then $\partial(f+g)(x) = \partial g(x) + \nabla f(x)$.*

Proof. 1. follows from 2. using $g \equiv 0$, since then $\partial g(x) = \{0\}$ for all x.

To show 2., we first show $\partial(f+g)(x) \supseteq \partial g(x) + \nabla f(x)$. If $v \in \partial g(x)$, then

$$\begin{aligned} f(y) &\geqslant f(x) + \langle \nabla f(x), y - x\rangle, \\ g(y) &\geqslant g(x) + \langle v, y - x\rangle \end{aligned}$$

The first inequality uses the same argument as in Thm. 3.12 (we cannot apply the theorem directly because we do not know if f is differentiable in a neighborhood of x, but the proof is the same; specifically for all $t>0$, we have

$$\begin{aligned}
& \frac{f(y)-f(x)-\langle \nabla f(x), y-x\rangle}{\|y-x\|_2} \\
= \quad & \frac{t\,f(y)-t\,f(x)-\langle \nabla f(x), t\,(y-x)\rangle}{t\,\|y-x\|_2} \\
= \quad & \frac{(1-t)\,f(x)+t\,f(y)-f(x)-\langle \nabla f(x), t\,(y-x)\rangle}{t\,\|y-x\|_2} \\
\overset{f \text{ convex}}{\geqslant} \quad & \frac{f(x+t\,(y-x))-f(x)-\langle \nabla f(x), t\,(y-x)\rangle}{t\,\|y-x\|_2}.
\end{aligned}$$

This holds for all $t>0$ and the last term converges to 0 from the definition of differentiability. Thus $f(y)-f(x)-\langle \nabla f(x), y-x\rangle \geqslant 0$.)

Adding the inequalities for f and g we obtain

$$f(y)+g(y) \;\geqslant\; f(x)+g(x)+\langle v+\nabla f(x), y-x\rangle,$$

which shows $v+\nabla f(x)\in\partial(f+g)(x)$.

To show $\partial(f+g)(x)\subseteq \partial g(x)+\nabla f(x)$, assume that $v\in\partial(f+g)(x)$. Then $(f+g)(y)-(f+g)(x)\geqslant\langle v, y-x\rangle$ for any $y\in\mathbb{R}^n$ per definition. Thus in particular

$$\liminf_{z\to x} \frac{f(z)+g(z)-f(x)-g(x)-\langle v, z-x\rangle}{\|z-x\|_2}\geqslant 0.$$

Because f is differentiable in x we find that the lim inf is exactly the same as the lim inf of

$$\frac{-f(z)+f(x)+\langle \nabla f(x), z-x\rangle+f(z)-f(x)+g(z)-g(x)-\langle v, z-x\rangle}{\|z-x\|_2}$$

This is possible because $f(z)$ and $f(x)$ must be finite close to x from the definition of differentiability, and we can simplify:

$$\liminf_{z\to x} \frac{g(z)-g(x)-\langle v-\nabla f(x), z-x\rangle}{\|z-x\|_2}\geqslant 0.$$

Now consider $z(t):=(1-t)\,x+t\,y$. Then the above equation states

$$\liminf_{t\searrow 0} \frac{g(z(t))-g(x)-\langle v-\nabla f(x), z(t)-x\rangle}{\|z(t)-x\|_2}\geqslant 0.$$

From the definition of convexity we get

$$g(z(t)) \;\leqslant\; (1-t)\,g(x)+t\,g(y).$$

Thus

$$\liminf_{t\searrow 0} \frac{(1-t)\,g(x)+t\,g(y)-g(x)-\langle v-\nabla f(x), z(t)-x\rangle}{\|z(t)-x\|_2} \;\geqslant\; 0.$$

We assumed that $g(x)$ is finite, so $(1-t)\,g(x)-g(x)=-t\,g(x)$ (if $g(x)=+\infty$ this gives $\infty-\infty=-\infty$, which is wrong). Also $z(t)-x=t\,(y-x)$, and we get

$$\begin{aligned}
& \liminf_{t\searrow 0} \frac{-t\,g(x)+t\,g(y)-\langle v-\nabla f(x), t\,(y-x)\rangle}{t\,\|y-x\|_2}\geqslant 0 \\
\Leftrightarrow \;\; & \liminf_{t\searrow 0} \{g(y)-g(x)-\langle v-\nabla f(x), y-x\rangle\}\geqslant 0 \\
\Leftrightarrow \;\; & g(y)\geqslant g(x)+\langle v-\nabla f(x), y-x\rangle
\end{aligned}$$

Since y was arbitrary this shows $v - \nabla f(x) \in \partial g(x)$, so $v \in \partial g(x) + \nabla f(x)$, and finally $\partial(f+g) \subseteq \partial g + \nabla f$.

□

Theorem 5.7. *(Generalized Fermat) Assume $f: \mathbb{R}^n \to \bar{\mathbb{R}}$ is proper. Then*

$$x \in \arg\min f \;\Leftrightarrow\; 0 \in \partial f(x).$$

Proof.

$$\begin{array}{rl} & 0 \in \partial f(x) \\ \Leftrightarrow & f(x) + \langle 0, y - x\rangle \leqslant f(y) \quad \forall y \in \mathbb{R}^n \\ \Leftrightarrow & f(x) \leqslant f(y) \quad \forall y \in \mathbb{R}^n \\ \overset{f \text{ proper}}{\Leftrightarrow} & x \in \arg\min f. \end{array}$$

In the last equivalence the properness is required since $\arg\min +\infty = \emptyset$ by definition, but $\partial(+\infty)(x) = \mathbb{R}^n \ni 0$ for all x. □

Definition 5.8. *For a convex set $C \subseteq \mathbb{R}^n$ and $x \in C$, the "normal cone" $N_C(x)$ at x is defined as*

$$N_C(x) \;:=\; \{v \in \mathbb{R}^n \,|\, \langle v, y - x\rangle \leqslant 0 \;\forall y \in C\}.$$

By convention, $N_C(x) := \emptyset$ for $x \notin C$.

It can be easily seen that $N_C(x)$ is in fact a cone if $x \in C$.

Proposition 5.9. *(subdifferential of indicator functions) Assume $C \subseteq \mathbb{R}^n$ is convex with $C \neq \emptyset$, then*

$$\partial \delta_C(x) \;=\; N_C(x).$$

Proof.

For $x \in C$:

$$\begin{array}{rcl} \partial \delta_C(x) & = & \{v \in \mathbb{R}^n \,|\, \delta_C(x) + \langle v, y - x\rangle \leqslant \delta_C(y) \;\forall y \in \mathbb{R}^n\} \\ & = & \{v \in \mathbb{R}^n \,|\, 0 + \langle v, y - x\rangle \leqslant 0 \;\forall y \in C\} = N_C(x). \end{array}$$

For $x \notin C$: Since $C \neq \emptyset$ we can find a $y \in C$, i.e., $\delta_C(y) = 0$. But then for any $v \in \partial \delta_C(x)$ we have

$$\delta_C(x) + \langle v, y - x\rangle \;=\; +\infty > 0 = \delta_C(y).$$

Thus $v \notin \partial \delta_C(x)$ for any v, and we conclude $\partial \delta_C(x) = \emptyset$. □

Proposition 5.10. *Assume $C \subseteq \mathbb{R}^n$ is closed and convex with $C \neq \emptyset$, and $x \in \mathbb{R}^n$. Then*

$$y = \Pi_C(x) \;\Leftrightarrow\; x - y \in N_C(y),$$

Proof. From Prop. 3.17 we know that $y = \Pi_C(x)$ is the unique minimizer of

$$f(y') \;:=\; \frac{1}{2}\|y' - x\|_2^2 + \delta_C(y').$$

This is the case iff $0 \in \partial f(y)$ (Thm. 5.7). From Prop. 5.6 and Prop. 5.9 we know that $\partial f(y) = y - x + N_C(y)$, thus

$$\begin{aligned} y = \Pi_C(x) &\Leftrightarrow y \in \arg\min f \\ &\Leftrightarrow 0 \in \partial f(y) \\ &\Leftrightarrow 0 \in y - x + N_C(y) \\ &\Leftrightarrow x - y \in N_C(y). \end{aligned}$$

□

A consequence of Prop. 5.10 is that by looking at the normal cone for a fixed $x \in C$, we can find *all* points y that get projected into x.

Proposition 5.11. *(subdifferentials as normal cones) Assume $f: \mathbb{R}^n \to \bar{\mathbb{R}}$ is proper and convex. Then*

$$\partial f(x) = \begin{cases} \emptyset, & x \notin \operatorname{dom} f, \\ \{v \in \mathbb{R}^n | (v, -1) \in N_{\operatorname{epi} f}(x, f(x))\}, & x \in \operatorname{dom} f. \end{cases}$$

Also, if $x \in \operatorname{dom} f$ then

$$N_{\operatorname{dom} f}(x) = \{v \in \mathbb{R}^n | (v, 0) \in N_{\operatorname{epi} f}(x, f(x))\}.$$

Proof. $x \notin \operatorname{dom} f$: f proper $\Rightarrow$ ex. y such that $f(y) \in \mathbb{R}$. If $v \in \partial f(x)$ then by definition of the subdifferential

$$\begin{aligned} v^\top (y - x) + f(x) &\leqslant f(y) \\ \Rightarrow +\infty &\leqslant f(y) \in \mathbb{R}. \end{aligned}$$

This is impossible, thus $\partial f(x) = \emptyset$.

$x \in \operatorname{dom} f$:

$$\begin{aligned} & v \in \partial f(x) \\ \Leftrightarrow\ & v^\top (y - x) + f(x) \leqslant f(y) \quad \forall y \in \mathbb{R}^n \\ \Leftrightarrow\ & v^\top (y - x) + f(x) \leqslant \alpha \quad \forall y \in \operatorname{dom} f, \forall \alpha \geqslant f(y), \alpha \in \mathbb{R} \\ \Leftrightarrow\ & v^\top (y - x) + (-1)(\alpha - f(x)) \leqslant 0 \quad \forall (y, \alpha) \in \operatorname{epi} f \\ \Leftrightarrow\ & (v, -1) \in N_{\operatorname{epi} f}(x, f(x)). \end{aligned}$$

Second part:

$$\begin{aligned} & v \in N_{\operatorname{dom} f}(x) \\ \Leftrightarrow\ & v^\top (y - x) \leqslant 0 \quad \forall y \in \operatorname{dom} f \\ \Leftrightarrow\ & v^\top (y - x) + 0 \cdot (\alpha - f(x)) \leqslant 0 \quad \forall y \in \operatorname{dom} f, \forall \alpha \geqslant f(y) \\ \Leftrightarrow\ & v^\top (y - x) + 0 \cdot (\alpha - f(x)) \leqslant 0 \quad \forall (y, \alpha) \in \operatorname{epi} f \\ \Leftrightarrow\ & (v, 0) \in N_{\operatorname{epi} f}(x, f(x)). \end{aligned}$$

□

Example 5.12. The subdifferential of $f: \mathbb{R} \to \bar{\mathbb{R}}$, $f(x) = |x|$ $(x \in \mathbb{R})$ is

$$\partial f(x) = \begin{cases} \{1\}, & x > 0, \\ \{-1\}, & x < 0, \\ [-1, 1], & x = 0. \end{cases}$$

Computing subdifferentials is generally not an easy task. Fortunately in many cases they can be found by combining the subdifferentials of simpler functions, but this requires a little bit more caution than for differentiable functions.

Definition 5.13. *(relative interior) For any set $C \subseteq \mathbb{R}^n$, we define the "affine hull" and the "relative interior"*

$$\begin{aligned} \operatorname{aff} C &:= \bigcap_{A \text{ affine}, C \subseteq A} A, \\ \operatorname{rint} C &:= \{x \in \mathbb{R}^n | \text{ex. (open) neighborhood } N \text{ of } x \text{ with } N \cap \operatorname{aff} C \subseteq C\}. \end{aligned}$$

Proposition 5.14. *(rules for subdifferentials)*

1. *Assume $f: \mathbb{R}^n \to \bar{\mathbb{R}}$ is convex. Then*

$$\begin{aligned} g(x) = f(x+y) &\Rightarrow \partial g(x) = \partial f(x+y) \quad y \in \mathbb{R}^n, \\ g(x) = f(\lambda x) &\Rightarrow \partial g(x) = \lambda \partial f(\lambda x), \quad \lambda \neq 0, \\ g(x) = \lambda f(x) &\Rightarrow \partial g(x) = \lambda \partial f(x), \quad \lambda > 0. \end{aligned}$$

   [example where 2. does not hold: take $f(0) = 1$, $f(1) = -\infty$, then $\partial f(0)$ is empty but $\partial(f(0\cdot)) = \partial(0) = \{0\}$]

2. *Assume $f: \mathbb{R}^n \to \bar{\mathbb{R}}$ is proper and convex, and $A \in \mathbb{R}^{n \times m}$ such that*

$$\{A y | y \in \mathbb{R}^m\} \cap \operatorname{rint} \operatorname{dom} f \neq \emptyset.$$

   If $x \in \operatorname{dom}(f \circ A) = \{y \in \mathbb{R}^m | A y \in \operatorname{dom} f\}$, then

$$\partial(f \circ A)(x) = A^\top \partial f(A x).$$

3. *Assume $f_1, ..., f_m: \mathbb{R}^n \to \bar{\mathbb{R}}$ are proper and convex, and*

$$\operatorname{rint} \operatorname{dom} f_1 \cap ... \cap \operatorname{rint} \operatorname{dom} f_m \neq \emptyset.$$

   If $x \in \operatorname{dom} f$, then

$$\partial(f_1 + ... + f_m)(x) = \partial f_1(x) + ... + \partial f_m(x).$$

Proof. 1. can be shown directly. The proof for 2. is surprisingly difficult and we refer to [Roc70, 23.8,23.9]. 3. then follows from 2. with $A = (I|...|I)^\top$ and $f(x_1, ..., x_m) = f_1(x_1) + ... + f_m(x_m)$. □

Remark 5.15. The conditions on the relative interiors Prop. 5.14 are important: in 3., it is easy to see that the direction "$\supseteq$" always holds. But the direction "$\subseteq$" requires more. Set

$$\begin{aligned} f_1(x) &:= \delta_{\mathcal{B}_1(0)}, \\ f_2(x) &:= \delta_C, \; C := \{1\} \times \mathbb{R}. \end{aligned}$$

Then $\partial f_1(1,0) = \mathbb{R}_{\geqslant 0} \times \{0\}$ and $\partial f_2(1,0) = N_C(1,0) = \mathbb{R} \times \{0\}$. Thus

$$(\partial f_1 + \partial f_2)(1,0) = \mathbb{R} \times \{0\}.$$

But $f_1 + f_2 = \delta_{\{(1,0)\}}$, thus $\partial(f_1 + f_2)(1,0) = \mathbb{R}^2$. This is only possible since f_1, f_2 do not satisfy the condition on the relative interiors in Prop. 5.14.

Chapter 6
Conjugate Functions

6.1 The Legendre-Fenchel Transform

Definition 6.1. *(convex hull of a function) For* $f\colon \mathbb{R}^n \to \bar{\mathbb{R}}$*,*

$$\operatorname{con} f(x) \;=\; \sup_{g \leqslant f,\, g \text{ convex}} g(x)$$

is the "convex hull" of f*.*

Remark 6.2. The supremum on the right-hand side is convex (Prop. 3.10), majorized by f and clearly majorizes every convex function majorized by f. Therefore $\operatorname{con} f$ is the greatest convex function majorized by f.

Definition 6.3. *(closure of a function) For* $f\colon \mathbb{R}^n \to \bar{\mathbb{R}}$*, the "(lower) closure"* $\operatorname{cl} f$ *is defined as*

$$\operatorname{cl} f(x) \;:=\; \liminf_{y \to x} f(y).$$

Proposition 6.4. *(closure and the epigraph) For* $f\colon \mathbb{R}^n \to \bar{\mathbb{R}}$ *we have*

$$\operatorname{epi}(\operatorname{cl} f) \;=\; \operatorname{cl}(\operatorname{epi} f).$$

Moreover, if f *is convex then* $\operatorname{cl} f$ *is convex.*

Proof.

$$\begin{aligned}(x,\alpha) \in \operatorname{cl}(\operatorname{epi} f) \;&\Leftrightarrow\; \exists x^k \to x, \alpha^k \to \alpha,\, f(x^k) \leqslant \alpha^k \\ &\Leftrightarrow\; \underbrace{\liminf_{y \to x} f(y)}_{\operatorname{cl} f(x)} \leqslant \alpha \Leftrightarrow (x,\alpha) \in \operatorname{epi}(\operatorname{cl} f).\end{aligned}$$

This holds because if there exists such a sequence, then the $\liminf$ is $\leqslant\alpha$, and if the $\liminf$ is $\leqslant\alpha$ then take a sequence x^k with $f(x^k) \to \liminf$. Then $(x, \liminf_{y\to x} f(x)) \in \operatorname{cl}(\operatorname{epi} f)$, and therefore in particular $(x,\alpha) \in \operatorname{cl}(\operatorname{epi} f)$ since $\alpha \geqslant \liminf_{y\to x} f(x)$.

Second part: if f is convex then $\operatorname{epi} f$ is convex. For $x, y \in \operatorname{epi}(\operatorname{cl} f) = \operatorname{cl}(\operatorname{epi} f)$ we have $x^k \to x$, $y^k \to y$ with $x^k, y^k \in \operatorname{epi} f$. For every $\tau \in (0,1)$, $z := (1-\tau)\, x + \tau\, y$ is the limit of $(1-\tau)\, x^k + \tau\, y^k$, and all these points are in $\operatorname{epi} f$ because $\operatorname{epi} f$ is convex (f is convex). Thus z is in $\operatorname{cl}\operatorname{epi} f$ and therefore in $\operatorname{epi}(\operatorname{cl} f)$. This shows that $\operatorname{epi}(\operatorname{cl} f)$ is convex and therefore $\operatorname{cl} f$ is convex. □

Note that the corresponding statement for the convex closure does *not* hold, i.e.,

$$\operatorname{epi}(\operatorname{con} f) \;\neq\; \operatorname{con}(\operatorname{epi} f)$$

in general. One example is the function

$$f(x) \;=\; \begin{cases} 1, & x \geqslant 0, \\ 0, & x < 0. \end{cases}$$

Proposition 6.5. *(closure, alternative definition) For $f\colon \mathbb{R}^n \to \bar{\mathbb{R}}$, we have*

$$(\operatorname{cl} f)(x) \;=\; \sup_{g \leqslant f,\, g \text{ lsc}} g(x).$$

Proof. "$\leqslant$": $\operatorname{epi}(\operatorname{cl} f)$ is closed by Prop. 6.4 and therefore $\operatorname{cl} f$ is lsc by Thm. 2.10. Also $\operatorname{cl} f \leqslant f$ (i.e., $\liminf_{y \to x} f(y) \leqslant f(x)$), because of the definition of the lim inf (or take the constant sequence $y^k = x$). Together this shows $\leqslant$.

"$\geqslant$": If $g \leqslant f$ and g is lsc, then

$$g(x) \overset{\text{lsc}}{\leqslant} \liminf_{y \to x} g(y) \overset{g \leqslant f}{\leqslant} \liminf_{y \to x} f(y) = (\operatorname{cl} f)(x).$$

□

Theorem 6.6. *(envelope representation of sets) Assume that $C \subseteq \mathbb{R}^n$ is closed and convex. Then*

$$C \;=\; \bigcap_{(b,\beta),\, C \subseteq H_{b,\beta}} H_{b,\beta}, \quad \text{where} \quad H_{b,\beta} := \{x \in \mathbb{R}^n \mid \langle x, b\rangle - \beta \leqslant 0\}.$$

C is thus the intersection of all closed half-spaces containing it.

Proof. If $C = \mathbb{R}^n$ or $C = \emptyset$ we are done (there are no such half-spaces; the intersection of all half-spaces is empty).

If $x \in C$ then x is also contained in the intersection.

If $x \notin C$ then set $y := \Pi_C(x)$. By Prop. 5.10 we know that $v := x - y \in N_C(y)$. This means

$$\begin{aligned} & \langle v, z - y\rangle \leqslant 0 \;\forall z \in C \\ \Leftrightarrow\; & \left\langle \underbrace{v}_{b}, z \right\rangle - \underbrace{\langle v, y\rangle}_{\beta} \leqslant 0 \;\forall z \in C \\ \Leftrightarrow\; & C \subseteq H_{b,\beta}. \end{aligned}$$

But $\langle v, x - y\rangle = \langle x - y, x - y\rangle = \|x - y\|^2 > 0$ because $y \in C$ and $x \notin C$ and therefore $x \neq y$. Thus

$$x \;\notin\; H_{b,\beta},$$

which shows that x is not contained in the intersection. □

In fact not all the half-spaces are needed in Thm. 6.6. If we assume $C \neq \emptyset$ then it is enough to intersect all *supporting* half-spaces, i.e., the half-spaces whose associated hyperplane touches C. The following theorem shows that if C is the epigraph of a proper lsc convex function, we do not need the vertical half-spaces.

Theorem 6.7. *(envelope representation of convex functions) Assume $f\colon \mathbb{R}^n \to \bar{\mathbb{R}}$ is proper, lsc and convex. Then*

$$f(x) \;=\; \sup_{g \text{ affine}, g \leqslant f} g(x).$$

Proof. [RW04, 12.1] f is lsc $\Rightarrow$ epi f is closed (Thm. 2.10); f is convex $\Rightarrow$ epi f is convex (Prop. 3.7). Therefore epi f is the intersection of all half-spaces containing it (in $\mathbb{R}^n \times \mathbb{R}$) by Thm. 6.6:

$$\operatorname{epi} f \;=\; \bigcap_{(b,c),\beta \in S} H_{(b,c),\beta} =: I_1,$$

where $H_{(b,c),\beta} = \{(x,\alpha) \in \mathbb{R}^n \times \mathbb{R} | \langle (x,\alpha),(b,c)\rangle - \beta \leqslant 0\}$ and S defines the set of half-spaces containing epi f.

Claim:

$$g \text{ affine } \;\Leftrightarrow\; \exists (b,c),\beta \text{ with } c<0, \operatorname{epi} g = H_{(b,c),\beta}.$$

"$\Rightarrow$": $g(x) = \langle b,x\rangle - \beta$. Then $(x,\alpha) \in \operatorname{epi} g \Leftrightarrow \langle b,x\rangle - \beta \leqslant \alpha \Leftrightarrow \langle (b,-1),(x,\alpha)\rangle - \beta \leqslant 0 \Leftrightarrow (x,\alpha) \in H_{(b,-1),\beta}$.

"$\Leftarrow$": $(x,\alpha) \in H_{(b,c),\beta}$ with $c<0 \Leftrightarrow \langle b,x\rangle + \alpha\, c - \beta \leqslant 0 \Leftrightarrow \langle b,x\rangle - \beta \leqslant (-c)\,\alpha$. Since $c<0$ this is equivalent to $\langle -b/c, x\rangle - (-\beta/c) \leqslant \alpha \Leftrightarrow (x,\alpha) \in \operatorname{epi} g$ with $g(x) = \langle -b/c, x\rangle + \beta/c$.

Using the claim, since

$$\operatorname{epi}\left(\sup_{g \text{ affine}, g \leqslant f} g(x) \right) \;=\; \bigcap_{g \text{ affine}, g \leqslant f} \operatorname{epi} g$$

and $g \leqslant f \Leftrightarrow \operatorname{epi} f \subseteq \operatorname{epi} g$, we only need to show that

$$I_1 := \underbrace{\bigcap_{(b,c),\beta \in S} H_{(b,c),\beta}}_{\operatorname{epi} f} \;=\; \underbrace{\bigcap_{(b,c),\beta \in S, c<0} H_{(b,c),\beta}}_{\operatorname{epi}(\sup \ldots)} =: I_2.$$

The direction "$\subseteq$" is clear. For "$\supseteq$" we need to show that if $(\bar{x},\bar{\alpha}) \notin I_1$ then $(\bar{x},\bar{\alpha}) \notin I_2$, i.e., there exist $(b,c),\beta \in S$ with $c<0$ such that $(\bar{x},\bar{\alpha}) \notin H_{(b,c),\beta}$.

Assume that $(\bar{x},\bar{\alpha}) \notin I_1$. Then there exist $(b_1,c_1),\beta_1 \in S$ such that $(\bar{x},\bar{\alpha}) \notin H_{(b_1,c_1),\beta}$. If $c_1 > 0$ then

$$\begin{aligned} \operatorname{epi} f \subseteq H_{(b_1,c_1),\beta} \;&=\; \{(x,\alpha) | \langle b_1,x\rangle + \alpha\, c_1 - \beta \leqslant 0\} \\ &\overset{c_1>0}{=}\; \left\{ (x,\alpha) \middle| \alpha \leqslant \frac{1}{c_1}(\beta - \langle b_1,x\rangle) \right\}. \end{aligned}$$

If this were true then epi f would be contained in a lower half space, which cannot hold since epi $f \neq \emptyset$ (f is proper). This shows that $c_1 \leqslant 0$.

If $c_1 < 0$ then $(\bar{x},\bar{\alpha}) \notin I_2$ by definition. Therefore the only difficult case is $c_1 = 0$.

Assume $c_1 = 0$. Then we define

$$g_1(x) \;:=\; \langle b_1,x\rangle - \beta_1.$$

If $x \in \operatorname{dom} f$ then $(x,\, f(x)) \in \operatorname{epi} f$, therefore $(x,\, f(x)) \in H_{(b_1,c_1),\beta_1}$ and $g_1(x) = \langle x,\, b_1\rangle - \beta_1 \leqslant 0$ on $\operatorname{dom} f$.

Now take any (b',c'), $\beta' \in S$ with $c' < 0$. Such a triple exists: if not, $c = 0$ holds for all $(b,c),\beta \in S$, which means that epi f is only bounded by vertical hyperplanes, and therefore f is $-\infty$ on all of $\operatorname{dom} f$, which contradicts the properness of f.

We then define the associated affine function

$$g_2(x) \;:=\; \langle -b'/c', x\rangle - (-\beta'/c').$$

Since epi $f \subseteq H_{(b',c'),\beta'} = \operatorname{epi} g_2$ we get $(x,\, f(x)) \in \operatorname{epi} g_2$ and therefore $g_2(x) \leqslant f(x)$ $\forall x \in \operatorname{dom} f$.

For $\lambda \geqslant 0$ we construct the function

$$g^\lambda \;:=\; \lambda\, g_1 + g_2.$$

The crucial point is that $g^\lambda(x) \leqslant f(x)$ for $x \in \operatorname{dom} f$ by the above considerations. On the other hand, because $(\bar{x}, \bar{\alpha}) \notin H_{(b_1,c_1),\beta_1}$ we know that $g_1(\bar{x}) > 0$, and, since $\operatorname{dom} g_2 = \mathbb{R}^n$, we can choose λ large enough such that $g^\lambda(\bar{x}) > \bar{\alpha}$, i.e., $(\bar{x}, \bar{\alpha}) \notin \operatorname{epi} g^\lambda$. Then $g^\lambda \leqslant f$ and therefore

$$\operatorname{epi} f \;\subseteq\; \operatorname{epi} g^\lambda = H_{(\lambda b_1 + b_2, -1), \lambda\beta_1 + \beta_2}.$$

The half-space on the right is included in the intersection in I_2 because it has $c = -1 < 0$. But $(\bar{x}, \bar{\alpha}) \notin \operatorname{epi} g^\lambda$, therefore $(\bar{x}, \bar{\alpha})$ cannot be contained in I_2. □

Definition 6.8. *(Legendre-Fenchel transform) Let $f\colon \mathbb{R}^n \to \bar{\mathbb{R}}$, then*

$$\begin{aligned} &f^*\colon \mathbb{R}^n \to \bar{\mathbb{R}}, \\ &f^*(v) := \sup_{x \in \mathbb{R}^n} \{\langle v, x\rangle - f(x)\} \end{aligned}$$

is the "conjugate to f". The mapping $f \mapsto f^*$ is the "Legendre-Fenchel transform".

Remark 6.9. The intuition here is that for some b, β, we have

$$\begin{aligned} \langle b, x\rangle - \beta \leqslant f(x) \quad \forall x \;&\Leftrightarrow\; \langle b, x\rangle - f(x) \leqslant \beta \quad \forall x \\ &\Leftrightarrow\; \beta \geqslant \sup_x \{\langle b, x\rangle - f(x)\} \\ &\Leftrightarrow\; \beta \geqslant f^*(b) \\ &\Leftrightarrow\; (b, \beta) \in \operatorname{epi} f^*. \end{aligned}$$

The conjugate thus characterizes all the affine functions majorized by f by providing the offset β for any given linear part b. Also, for any given slope b, all affine functions $\langle b, x\rangle - \beta$ with $\beta \geqslant f^*(b)$ are majorized by $\langle b, x\rangle - f^*(b)$ and therefore can be left out when taking the intersection in Thm. 6.7. This means that the Legendre-Fenchel transform describes a reduced set of affine functions that are necessary to reconstruct f by returning the offset β for a given slope b.

Theorem 6.10. *(Legendre-Fenchel transform) Assume $f\colon \mathbb{R}^n \to \bar{\mathbb{R}}$. Then*

$$f^* = (\operatorname{con} f)^* = (\operatorname{cl} f)^* = (\operatorname{cl}\operatorname{con} f)^*$$

and

$$f^{**} := (f^*)^* \;\leqslant\; f.$$

If $\operatorname{con} f$ is proper, then f^* and f^{**} are proper, lsc and convex, and

$$f^{**} \;=\; \operatorname{cl}\operatorname{con} f.$$

If $f\colon \mathbb{R}^n \to \bar{\mathbb{R}}$ is proper, lsc and convex, then

$$f^{**} \;=\; f.$$

Proof. First statement: we have

$$\begin{aligned} (v, \beta) \in \operatorname{epi} f^* \;&\Leftrightarrow\; \beta \geqslant f^*(v) \\ &\Leftrightarrow\; \beta \geqslant \langle v, x\rangle - f(x) \quad \forall x \in \mathbb{R}^n \\ &\Leftrightarrow\; \langle v, x\rangle - \beta \leqslant f(x) \quad \forall x \in \mathbb{R}^n. \end{aligned} \tag{6.1}$$

The $(v, \beta) \in \operatorname{epi} f^*$ define all the affine functions majorized by f.

We claim that for every affine function h we have

$$h \leqslant f \Leftrightarrow h \leqslant \operatorname{cl} f \Leftrightarrow h \leqslant \operatorname{con} f \Leftrightarrow h \leqslant \operatorname{cl} \operatorname{con} f.$$

Indeed, if $h \leqslant \operatorname{cl} \operatorname{con} f$ or $h \leqslant \operatorname{cl} f$ or $h \leqslant \operatorname{con} f$, then $h \leqslant f$ (by definition). If $h \leqslant f$ then $h \leqslant \operatorname{cl} f$, $h \leqslant \operatorname{con} f$ and $h \leqslant \operatorname{cl} \operatorname{con} f$, because h is lsc and convex and $\operatorname{cl} f$, $\operatorname{con} f$ and $\operatorname{cl} \operatorname{con} f$ are the *largest* functions in their class $\leqslant f$ (for $\operatorname{cl} \operatorname{con} f$: $h \leqslant f \Rightarrow h \leqslant \operatorname{con} f \Rightarrow h \leqslant \operatorname{cl} \operatorname{con} f$ by the arguments for $\operatorname{con} f$, $\operatorname{cl} f$ applied to f, $\operatorname{con} f$).

Using the claim, in the last line of (6.1) we can replace f by $\operatorname{cl} f$, $\operatorname{con} f$, or $\operatorname{cl} \operatorname{con} f$ to get

$$\begin{aligned} \operatorname{epi} f^* &= \operatorname{epi} (\operatorname{cl} f)^* = \operatorname{epi} (\operatorname{con} f)^* = \operatorname{epi} (\operatorname{cl} \operatorname{con} f)^*. \\ \Rightarrow f^* &= (\operatorname{cl} f)^* = (\operatorname{con} f)^* = (\operatorname{cl} \operatorname{con} f)^*. \end{aligned}$$

For the inequality: we have

$$\begin{aligned} f^{**}(y) &= \sup_v \{\langle v, y\rangle - f^*(v)\} = \sup_v \left\{ \langle v, y\rangle - \sup_x \{\langle v, x\rangle - f(x)\} \right\} \\ &= \sup_v \left\{ \langle v, y\rangle + \inf_x \{f(x) - \langle v, x\rangle\} \right\} \\ &\underset{\text{set } x=y}{\leqslant} \sup_v \{\langle v, y\rangle + f(y) - \langle v, y\rangle\} \\ &= f(y). \end{aligned}$$

Second statement: Assume $\operatorname{con} f$ is proper. We claim that $\operatorname{cl} \operatorname{con} f$ is proper, lsc and convex. Lower semi-continuity is clear, convexity from Prop. 6.4. For the properness, observe that $\operatorname{cl} \operatorname{con} f \leqslant \operatorname{con} f$ (because the epigraph increases). Because $\operatorname{con} f < +\infty$ this shows $\operatorname{cl} \operatorname{con} f < +\infty$. It remains to show that $\operatorname{cl} \operatorname{con} f(x) > -\infty$ always (example sheets).

From the claim we know that $\operatorname{cl} \operatorname{con} f$ is proper, lsc, convex and can apply Thm. 6.7 :

$$\begin{aligned} \operatorname{cl} \operatorname{con} f(x) &= \sup_{g \text{ affine}, g \leqslant \operatorname{cl} \operatorname{con} f} g(x) \\ &= \sup_{(v,\beta) \in \operatorname{epi} (\operatorname{cl} \operatorname{con} f)^*} \{\langle v, x\rangle - \beta\} \\ &= \sup_{(v,\beta) \in \operatorname{epi} f^*} \{\langle v, x\rangle - \beta\} \\ &= \sup_{(v,\beta) \in \operatorname{epi} f^*} \{\langle v, x\rangle - f^*(v)\} \\ &= \sup_{v \in \mathbb{R}^n} \{\langle v, x\rangle - f^*(v)\} \\ &= f^{**}(v). \end{aligned}$$

To show that f^* is proper, lsc and convex: we know that $f^* = (\operatorname{cl} \operatorname{con} f)^*$ is the conjugate of a proper lsc convex function. $\operatorname{epi} f^*$ is closed and convex as the intersection of closed convex sets (see (6.1), sets run over x) $\Rightarrow$ f^* is lsc convex. Properness: $\operatorname{con} f$ is proper $\Rightarrow$ there is at least one x' such that $\operatorname{con} f(x') < +\infty$, thus $f^*(v) = \sup\ldots \geqslant \langle v, x'\rangle - \operatorname{con} f(x')$, i.e., f^* is lower-bounded by an affine function (or $+\infty$) and can therefore never take the value $-\infty$. Therefore the only way for f^* *not* to be proper is $f^* = +\infty$. But then by the previous arguments

$$\begin{aligned} \operatorname{cl} \operatorname{con} f &\underset{\text{assumption}}{=} f^{**} \\ &= (+\infty)^* \\ &= -\infty. \end{aligned}$$

This is not possible because we showed that $\operatorname{cl} \operatorname{con} f$ is proper (see above, [RW04, 2.32]).

Last statement: f is convex, therefore $\operatorname{con} f = f$. Thus $\operatorname{con} f$ is proper. Also $\operatorname{con} f = f$ is lsc, therefore $f^{**} = \operatorname{cl}\operatorname{con} f = \operatorname{cl} f = f$. □

The condition that $\operatorname{con} f$ if proper does not appear to be easy to validate at first. It turns out that we can just compute f^*, and if we obtain neither $+\infty$ nor $-\infty$ then $\operatorname{con} f$ must be proper, as the following proposition shows:

Proposition 6.11. *Assume $f: \mathbb{R}^n \to \bar{\mathbb{R}}$. Then*

1. *$\operatorname{con} f$ is* not *proper $\Rightarrow$ $f^* \equiv +\infty$ or $f^* \equiv -\infty$.*
2. *in particular: f^* proper $\Rightarrow$ $\operatorname{con} f$ proper.*

Proof. The first statement follows in the same way as in the proof of the second statement in Thm. 6.10: if $\operatorname{con} f$ is not proper then $\operatorname{con} f = +\infty$ or there is x' s.t. $\operatorname{con} f(x') = -\infty$. If $\operatorname{con} f = +\infty$ then

$$f^*(v) \;=\; (\operatorname{con} f)^*(v) = \sup_x \langle v, x\rangle - \operatorname{con} f(x) = \sup_x -\infty = -\infty.$$

If there is x' s.t. $\operatorname{con} f(x') = -\infty$ then

$$f^*(v) \;=\; (\operatorname{con} f)^*(v) = \sup_x \langle v, x\rangle - (\operatorname{con} f)(x) \geqslant \langle v, x'\rangle - (\operatorname{con} f)(x') = +\infty.$$

The second statement is a special case (f proper $\Rightarrow$ $f \notin \{-\infty, +\infty\}$). □

6.2 Duality Correspondences

The following beautifully symmetric theorem is at the core of many of the later proofs.

Theorem 6.12. *(inversion rule for subdifferentials) Assume $f: \mathbb{R}^n \to \bar{\mathbb{R}}$ proper, lsc, convex. Then*

$$\partial f^* \;=\; (\partial f)^{-1},$$

specifically

$$v \in \partial f(x) \Leftrightarrow f(x) + f^*(v) = \langle v, x\rangle \Leftrightarrow x \in \partial f^*(v).$$

Moreover,

$$\begin{aligned} \partial f(x) &= \arg\max_{v'} \{\langle v', x\rangle - f^*(v')\}, \\ \partial f^*(v) &= \arg\max_{x'} \{\langle v, x'\rangle - f(x)\}. \end{aligned}$$

Proof. [RW04, 11.3] We have

$$\begin{aligned} & f(x) + f^*(v) = \langle v, x\rangle \\ \Leftrightarrow\quad & f^*(v) = \langle v, x\rangle - f(x) \\ \Leftrightarrow\quad & x \in \arg\max_{x'} \{\langle v, x'\rangle - f(x')\} \qquad (6.2) \\ \overset{\text{Thm. 6.10}}{\Leftrightarrow}\quad & 0 \in \partial(-\langle v, \cdot\rangle + f)(x) \\ \overset{\text{Prop. 5.6}}{\Leftrightarrow}\quad & v \in \partial f(x). \end{aligned}$$

This shows the second statement (apply to f^* and use $f^{**} = f$ from Thm. 6.10, f is proper, lsc, convex). The first statement is just a reformulation. The third statement then follows from the center line of (6.2). □

Prop. 6.12 provides a way to compute the subdifferentials of f and f^* by solving an optimization problem. This is useful in theory, but can be hard in practice. In fact,

$$\partial f^*(0) \;=\; \arg\max_x -f(x) = \arg\min_x f(x),$$

i.e., computing the subdifferential of f^* at a *single point* is generally as hard as finding the whole set of minimizers of f.

Proposition 6.13. *(duality correspondences) For proper, lsc, convex $f\colon \mathbb{R}^n \to \bar{\mathbb{R}}$ we have*

$$\begin{aligned}
(f(\cdot) - \langle a, \cdot\rangle)^* &= f^*(\cdot + a),\\
(f(\cdot + b))^* &= f^*(\cdot) - \langle \cdot, b\rangle,\\
(f(\cdot) + c)^* &= f^*(\cdot) - c,\\
(\lambda f(\cdot))^* &= \lambda f^*(\cdot/\lambda) \quad (\lambda > 0)\\
(\lambda f(\cdot/\lambda))^* &= \lambda f^*(\cdot) \quad (\lambda > 0).
\end{aligned}$$

Proof. Follows directly from the definition. □

Proposition 6.14. *(conjugation in product spaces) Let $f_i\colon \mathbb{R}^{n_i} \to \bar{\mathbb{R}}, i = 1, \ldots, m$ be proper and*

$$f(x_1, \ldots, x_m) \;:=\; f_1(x_1) + \ldots + f_m(x_m).$$

Then

$$f^*(v_1, \ldots, v_m) \;=\; f_1^*(v_1) + \ldots + f_m^*(v_m).$$

Proof. Follows directly from the definition. □

Definition 6.15. *(support functions) For any set $S \subseteq \mathbb{R}^n$ we define the "support function"*

$$\sigma_S(v) \;:=\; \sup_{x \in S} \langle v, x\rangle = (\delta_S^*)(v).$$

Definition 6.16. *(positively homogeneous functions) A function $f\colon \mathbb{R}^n \to \bar{\mathbb{R}}$ is said to be "positively homogeneous" iff $0 \in \operatorname{dom} f$ and*

$$f(\lambda x) \;=\; \lambda f(x) \quad \forall x \in \mathbb{R}^n, \forall \lambda > 0.$$

Proposition 6.17. *(support functions, polar cones) The set of positively homogeneous proper lsc convex functions and the set of closed convex nonempty sets are in one-to-one-correspondence through the Legendre-Fenchel transform:*

$$\delta_C \underset{*}{\longleftrightarrow} \sigma_C, \qquad x \in \partial\sigma_C(v) \Leftrightarrow x \in C, \sigma_C(v) = \langle v, x\rangle \Leftrightarrow v \in N_C(x).$$

In particular, the set of closed convex cones is in one-to-one correspondence with itself: for any cone K define the "polar cone" (also sometimes referred to as the "dual cone") as

$$K^* \;:=\; \{v \in \mathbb{R}^d \,|\, \langle v, x\rangle \leqslant 0 \;\forall x \in K\}.$$

Then

$$\delta_K \underset{*}{\longleftrightarrow} \delta_{K^*}, \qquad x \in N_{K^*}(v) \Leftrightarrow v \in N_K(x) \Leftrightarrow x \in K,\; v \in K^*,\; \langle x,\, v\rangle = 0 \Leftrightarrow 0 \leqslant_K x \perp v \geqslant_{K^*} 0.$$

Proof. (example sheet) Support functions: If $f(x) = \delta_C(x)$ then

$$f^*(x) \;=\; \sigma_C(x).$$

C is nonempty closed convex $\Rightarrow$ δ_C is proper lsc convex $\Rightarrow$ by Thm. 6.10, f^* is proper lsc convex. Also $f^*(0) = \sup_{x\in C} \langle 0, x\rangle = 0$ and

$$f^*(\lambda v) \;=\; \sup_{x\in C} \langle \lambda v, x\rangle = \lambda \sup_{x\in C} \langle v, x\rangle = \lambda f^*(v), \quad \lambda > 0,$$

therefore $f^* = \sigma_C$ is positively homogeneous.

On the other hand, assume that g is a pos. hom. lsc convex function. We know that $g(0) = 0$, therefore $g^*(x) \geqslant 0$. Then for any $\lambda > 0$ we have

$$g^*(x) \;=\; \sup_{v\in\mathbb{R}^n} \{\langle x, \lambda v\rangle - g(\lambda v)\} \overset{\text{pos. hom}}{=} \sup_{v\in\mathbb{R}^n} \lambda\,(\langle x, v\rangle - g(v)) = \lambda\, g^*(x).$$

Thus $g^*(x) \in \{0, +\infty\}$, and g^* is an indicator function of some set C, which must be closed convex and nonempty because g^* is proper lsc convex.

One-to-one correspondence: from $(\delta_C)^{**} = \delta_C$.

From Thm. 6.10 we get

$$\begin{aligned} &x \in \partial\sigma_C(v) \Leftrightarrow v \in \partial(\sigma_C^*)(x) \Leftrightarrow v \in \partial\delta_C(x) \overset{\text{Prop. 5.9}}{\Leftrightarrow} v \in N_C(x), \\ &\Leftrightarrow \delta_C(x) + \sigma_C(v) = \langle v, x\rangle \Leftrightarrow x \in C, \sigma_C(v) = \langle v, x\rangle. \end{aligned}$$

Cones: We claim: g is a positively homogeneous lsc convex proper indicator function $\Leftrightarrow$ $g = \delta_K$ with K closed convex cone: for $\lambda > 0$ we have $x \in K \Leftrightarrow \lambda x \in K$, therefore

$$\delta_K(\lambda x) \;=\; \begin{cases} 0, & \lambda x \in K, \\ +\infty, & \text{otherwise} \end{cases} = \begin{cases} 0, & x \in K, \\ +\infty, & \text{otherwise} \end{cases} = \delta_K(x) \overset{\delta_K(x)\in\{0,+\infty\}}{=} \lambda\,\delta_K(x).$$

Thus δ_K is positively homogeneous. Other direction: $g = \delta_C$ positively homogeneous lsc convex indicator function, then $0 \in \operatorname{dom} g$, thus $0 \in C$, and $x \in C \Rightarrow g(x) = 0 \Rightarrow g(\lambda x) \overset{\text{pos. hom}}{=} \lambda\, g(x) = 0 \Rightarrow \lambda x \in C$, thus C is a cone.

Take any such pos. hom. lsc convex indicator function δ_K, then by the first part (forward direction) δ_K^* is positively homogeneous lsc convex. But δ_K is also positively homogeneous lsc convex, therefore by the first part (backward direction) δ_K^* is an indicator function.

$\Rightarrow \delta_K^* = \delta_{K'}$ for some closed convex cone K'. We have

$$\begin{aligned} \delta_K^*(v) < +\infty \quad &\Leftrightarrow \quad \sup_{x\in K} \langle v, x\rangle < +\infty \\ &\overset{K \text{ cone}}{\Leftrightarrow} \quad \sup_{x\in K} \langle v, x\rangle \leqslant 0 \\ &\Leftrightarrow \quad v \in \{v' | \langle v', x\rangle \leqslant 0 \;\forall x \in K\} = K^*, \end{aligned}$$

thus $K' = K^*$ and $\delta_K^* = \delta_{K^*}$. □

Chapter 7
Duality in Optimization

Definition 7.1. *(primal and dual optimization problems, perturbation formulation) Assume $f\colon \mathbb{R}^n \times \mathbb{R}^m \to \bar{\mathbb{R}}$ is proper, lsc, and convex. We define the "primal" and "dual" problems*

$$\inf_{x\in\mathbb{R}^n} \varphi(x),\ \varphi(x) := f(x,0), \qquad \sup_{y\in\mathbb{R}^n} \psi(y),\ \psi(y) := -f^*(0,y).$$

and the "inf-projections"

$$p(u) := \inf_x f(x,u), \quad q(v) := \inf_y f^*(v,y) = -\sup_y \{-f^*(v,y)\}.$$

f is sometimes called a "perturbation function" for φ, and p the associated "marginal function".

A typical example is $f(x,u) = \frac{1}{2}\|x - I\|_2^2 + \delta_{\geqslant 0}(Ax - b + u)$. The extra variables u are used to *perturb* the constraints $Ax \geqslant b$ to $Ax \geqslant b - u$.

Proposition 7.2. *Assume f satisfies the assumptions in Def. 7.1. Then*

1. *φ and $-\psi$ are lsc and convex.*
2. *p, q are convex.*
3. *$p(0)$ and $p^{**}(0)$ are the optimal values of the primal and dual problems:*

$$p(0) = \inf_x \varphi(x), \quad p^{**}(0) = \sup_y \psi(y).$$

4. *The primal and dual problems are feasible iff the domain of their associated marginal function contains 0:*

$$\begin{aligned} \inf_x \varphi(x) < +\infty &\ \Leftrightarrow\ 0 \in \operatorname{dom} p, \\ \sup_y \psi(y) > -\infty &\ \Leftrightarrow\ 0 \in \operatorname{dom} q. \end{aligned}$$

Proof.

1. f proper lsc convex $\Rightarrow$ f^* is proper lsc convex (Thm. 6.10, con f is proper) $\Rightarrow$ φ, ψ lsc convex (not necessarily proper!).
2. Convexity of p: We consider the *strict* epigraph set of p:

$$\begin{aligned} E &:= \left\{(u,\alpha) \in \mathbb{R}^m \times \mathbb{R} \,\middle|\, p(u) = \inf_{x\in\mathbb{R}^n} f(x,u) < \alpha\right\} \\ &= \{(u,\alpha) \in \mathbb{R}^m \times \mathbb{R} | \exists x\colon f(x,u) < \alpha\} \\ &= A(\underbrace{\{(x,u,\alpha) \in \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R} | f(x,u) < \alpha\}}_{=:E'}) \\ &= A(E'), \end{aligned}$$

where A is the linear (coordinate projection) mapping $A(x, u, \alpha) := (u, \alpha)$. The first equality requires the strict inequality, i.e., we cannot just use epi p for the argument. E' is the strict epigraph of f and thus convex (Prop. 3.7) $\Rightarrow A(E')$ is convex (Prop. 3.15)). Thus p is convex (Prop. 3.7).

The same argument applied to q and f^* shows that q is convex.

3. First part by definition:

$$p(0) \;=\; \inf_x f(x,0) = \inf_x \varphi(x).$$

Second part: we know

$$\begin{aligned} p^*(y) &= \sup_u \{\langle y,u\rangle - p(u)\} \\ &= \sup_u \left\{\langle y,u\rangle - \inf_x f(x,u)\right\} \\ &= \sup_{u,x} \{\langle y,u\rangle - f(x,u)\} \\ &= \sup_{u,x} \{\langle (0,y),(x,u)\rangle - f(x,u)\} \\ &= f^*(0,y) \;=\; -\psi(y). \end{aligned}$$

Therefore

$$\begin{aligned} p^{**}(0) &= \sup_y \langle 0,y\rangle - p^*(y) \\ &= \sup_y \psi(y). \end{aligned}$$

4. From the definitions: $0 \in \operatorname{dom} p \Leftrightarrow p(0) < +\infty \Leftrightarrow \inf_x f(x,0) < +\infty \Leftrightarrow \inf \varphi < +\infty$, similar for q.

□

Theorem 7.3. *(weak and strong duality) Assume f satisfies the assumptions in Def. 7.1. Then "weak duality" always holds:*

$$\inf_x \varphi(x) \;\geqslant\; \sup_y \psi(y), \tag{7.1}$$

and under certain conditions the infimum and supremum are equal and finite ("strong duality"):

$$p(0) \in \mathbb{R} \text{ and } p \operatorname{lsc} \text{in } 0 \;\Leftrightarrow\; \inf_x \varphi(x) = \sup_y \psi(y) \in \mathbb{R}.$$

The difference $\inf \varphi - \sup \psi$ *is the "duality gap".*

Proof. From [ET99, Prop. 2.1]: For the inequality:

$$\inf_x \varphi(x) = p(0) \overset{\text{Thm. 6.10}}{\geqslant} p^{**}(0) \overset{\text{Prop. 7.2}}{=} \sup_y \psi(y).$$

Equality holds if and only if $p(0) = p^{**}(0)$. We thus have to show

$$p(0) \in \mathbb{R} \text{ and } p \text{ lsc in } 0 \;\Leftrightarrow\; p(0) = p^{**}(0) \in \mathbb{R}.$$

"⇐": Since $p^{**}(0) \leqslant \operatorname{cl} p(0) \leqslant p(0)$ holds for arbitrary p the right-hand side implies $\liminf_{y\to 0} p(y) = \operatorname{cl} p(0) = p(0) \in \mathbb{R}$, thus p is lsc in 0.

"⇒": We claim that if the left-hand side holds then $\operatorname{cl} p$ is proper lsc convex. Convexity and lower semi-continuity is clear from Prop. 7.2 and the definition of the closure. $\operatorname{cl} p$ must then also be proper: $\operatorname{cl} p$ is not constant $+\infty$ because $\operatorname{cl} p(0) \leqslant p(0) < +\infty$. If there were y s.t. $\operatorname{cl} p(y) = -\infty$ then ($\operatorname{cl} p$ convex) $\operatorname{cl} p((1-t)\,0 + t\,y) \leqslant (1-t)\operatorname{cl} p(0) + t \operatorname{cl} p(y) = -\infty$ for all $t \in (0,\,1)$. Since by assumption $\operatorname{cl}\, p(0) = p(0) \in \mathbb{R}$, this means $\operatorname{cl}\, p(t\;y) = -\infty$ for all $t \in (0,\,1)$. Moreover, $t\;y \to 0$ for $t \to 0$ and $\operatorname{cl}\, p$ is lsc (in particular in 0), thus $\operatorname{cl} p(0) \leqslant \liminf_{t\to 0} \operatorname{cl} p(t\,y) = -\infty$. But this would mean $p(0) = -\infty$, because p is lsc in 0, which implies $p(0) = \operatorname{cl} p(0)$. Thus $\operatorname{cl} p$ must be proper, lsc, and convex.

An alternative way of proving that $\operatorname{cl} p$ is proper is to use the fact that any improper, convex, lsc function is constant $+\infty$ or $-\infty$ (example sheets), which contradicts $p(0) \in \mathbb{R}$.

Because $p^* = (\operatorname{cl} p)^*$ always holds (Thm. 6.10),

$$(p^*)^*(0) \overset{\text{Thm. 6.10}}{=} ((\operatorname{cl} p)^*)^*(0) = (\operatorname{cl} p)^{**}(0) \overset{\text{Thm. 6.10, cl } p \text{ proper lsc conv.}}{=} \operatorname{cl} p(0) \overset{p \text{ lsc in } 0}{=} p(0).$$

Together with the first part this shows $\inf \varphi = \sup \psi = p(0)$, which is finite by assumption. □

Proposition 7.4. *(primal-dual optimality conditions) Assume f satisfies the assumptions in Def. 7.1. Then we have the "primal-dual optimality conditions"*

$$(0, y') \in \partial f(x', 0) \Leftrightarrow \left\{ \begin{array}{l} x' \in \arg\min\limits_x \varphi(x), \\ y' \in \arg\max\limits_y \psi(y), \\ \inf\limits_x \varphi(x) = \sup\limits_y \psi(y) \end{array} \right\} \Leftrightarrow (x', 0) \in \partial f^*(0, y'). \tag{7.2}$$

The set of "primal-dual optimal points" (x', y') satisfying (7.2) is either empty or equal to $(\arg\min \varphi) \times (\arg\max \psi)$.

Proof. We know from Prop. 6.12 that

$$\begin{aligned} (0, y') \in \partial f(x', 0) &\Leftrightarrow (x', 0) \in \partial f^*(0, y') \\ &\Leftrightarrow f(x', 0) + f^*(0, y') = \langle x', 0\rangle + \langle 0, y'\rangle \\ &\Leftrightarrow f(x', 0) = -f^*(0, y') \in \mathbb{R} \\ &\Leftrightarrow \varphi(x') = \psi(y') \in \mathbb{R}. \end{aligned}$$

Because $\inf \varphi \geqslant \sup \psi$ always and $\varphi(x') = \psi(y')$ shows $\inf \varphi \leqslant \sup \psi$, this is equivalent to

$$\inf_x \varphi(x) = \sup_y \psi(y) \in \mathbb{R}, \quad x' \in \arg\min \varphi,\, y' \in \operatorname{argmax} \psi.$$

This is again equivalent to

$$\inf_x \varphi(x) = \sup_y \psi(y), \quad x' \in \arg\min \varphi,\, y' \in \operatorname{argmax} \psi,$$

since equality with an infinite value would imply either $\varphi(x') = +\infty$ or $\psi(y') = -\infty$, both of which are explicitly excluded through the definition of the $\arg\min$.

If the set of x', y' that satisfy the conditions is non-empty, then $\inf \varphi = \sup \psi$ must hold with a finite value as seen above. Thus x', y' satisfy the conditions iff $x' \in \arg\min \varphi$ and $y' \in \arg\max \psi$, which proves the last statement. □

Proposition 7.5. *(sufficient conditions for strong duality) Assume f satisfies the assumptions in Def. 7.1. Then*

$$\begin{aligned} 0 \in \operatorname{int} \operatorname{dom} p \quad \textit{or} \quad 0 \in \operatorname{int} \operatorname{dom} q \;&\Rightarrow\; \inf_x \varphi(x) = \sup_y \psi(y) \quad (S') \\ 0 \in \operatorname{int} \operatorname{dom} p \quad \textit{and} \quad 0 \in \operatorname{int} \operatorname{dom} q \;&\Rightarrow\; \inf_x \varphi(x) = \sup_y \psi(y) \in \mathbb{R} \quad (S) \end{aligned}$$

(note that in the first case equality may hold with the value $+\infty$ or $-\infty$) and

$$0 \in \operatorname{int}\operatorname{dom} p \text{ and } \inf_x \varphi(x) \in \mathbb{R} \;\Leftrightarrow\; \arg\max_y \psi(y) \text{ nonempty and bounded} \quad (P),$$
$$0 \in \operatorname{int}\operatorname{dom} q \text{ and } \sup_y \psi(y) \in \mathbb{R} \;\Leftrightarrow\; \arg\min_x \varphi(x) \text{ nonempty and bounded} \quad (D).$$

In particular, if any of the conditions (S), (P), (D) holds, then strong duality holds, i.e., $\inf \varphi = \sup \psi \in \mathbb{R}$. *Moreover, if (S) holds, or (P) and (D) both hold, then there exist x', y' satisfying the primal-dual optimality conditions (7.2).*

Also,

$$(P) \;\Rightarrow\; \partial p(0) = \arg\max_y \psi(y),$$
$$(D) \;\Rightarrow\; \partial q(0) = \arg\min_x \varphi(x).$$

Proof. Assume $0 \in \operatorname{int}\operatorname{dom} p$. If $p(0) = -\infty$ then $p^{**}(0) \leqslant p(0) = -\infty$, thus $\sup \psi = p^{**}(0) = p(0) = \inf \varphi = -\infty$. The only other possibility is $p(0) \in \mathbb{R}$ since $0 \in \operatorname{int}\operatorname{dom} p$. Since $\operatorname{cl} p = p$ on $\operatorname{int}\operatorname{dom} p$ (example sheets) we know that if $0 \in \operatorname{int}\operatorname{dom} p$ then p is lsc in 0 and we can apply Thm. 7.3 to get $p^{**}(0) = p(0) \in \mathbb{R}$, which shows $\inf \varphi = \sup \psi$ (now with a finite value).

We can apply the same argument to $f'(x, y) := f^*(y, x)$, the approach is completely symmetric:

$$\begin{aligned}
\varphi'(x) &= f'(x,0) = f^*(0,x) = -\psi(x),\\
\psi'(y) &= -f'^*(0,y) \overset{f \text{ proper, lsc, convex}}{=} -f(y,0) = \varphi(y),\\
p'(u) &= \inf_x f'(x,u) = \inf_x f^*(u,x) = q(u),\\
q'(v) &= \inf_y f'^*(v,y) = \inf_y f(y,v) = p(v),\\
\inf \varphi' &= -\sup \psi,\\
\sup \psi' &= -\inf \varphi.
\end{aligned}$$

Then $0 \in \operatorname{int}\operatorname{dom} q \Rightarrow 0 \in \operatorname{int}\operatorname{dom} p'$. From the first part we know that then $\inf \varphi' = \sup \psi'$, but this means $-\sup\psi = -\inf \varphi$, thus $\inf \varphi = \sup \psi$.

If both $0 \in \operatorname{int}\operatorname{dom} p, 0 \in \operatorname{int}\operatorname{dom} q$ hold, then additionally $+\infty > p(0) \geqslant p^{**}(0) = \sup \psi = -q(0) > -\infty$, thus the value is finite.

Non-emptyness and boundedness: See [RW04, Thm. 11.39 proof]; the idea is to show that $0 \in \operatorname{int}\operatorname{dom} p$ if and only if ψ is proper (lsc convex) and level-bounded.

Subdifferential: If (P) holds then $0 \in \operatorname{int}\operatorname{dom} p$ and $p(0) \in \mathbb{R}$. Then we know that $\operatorname{cl} p(0) = p(0) \in \mathbb{R}$. $\operatorname{cl} p$ is then proper (and lsc convex) by the proof of Thm. 7.3, thus by Thm. 6.12

$$\begin{aligned}
\partial(\operatorname{cl} p)(0) &= \arg\max_y \{\langle 0, y\rangle - (\operatorname{cl} p)^*(y)\} = =a = \arg\max \psi.\\
&= \arg\max \{-(\operatorname{cl} p)^*\}\\
&\overset{\text{Thm. 6.10}}{=} \arg\max \{-p^*\}\\
&= \arg\max \psi.
\end{aligned}$$

But

$$\begin{aligned}
\partial p(0) &= \{v | p(0) + \langle v, x\rangle \leqslant p(x) \quad \forall x\}\\
&\overset{\operatorname{cl} p(0) = p(0)}{=} \{v | \operatorname{cl} p(0) + \langle v, x\rangle \leqslant p(x) \quad \forall x\}\\
&= \{v | \operatorname{cl} p(0) + \langle v, x\rangle \leqslant \operatorname{cl} p(x) \quad \forall x\}\\
&= \partial(\operatorname{cl} p)(0).
\end{aligned}$$

The second-to-last inequality follows because the affine functions majorized by p are exactly the affine functions majorized by $\operatorname{cl} p$. Together we get $\partial p(0) = \arg\max_y \psi(y)$.

Using duality, a similar argument shows the corresponding statement for q. □

Proposition 7.6. *Assume $k\colon \mathbb{R}^n \to \bar{\mathbb{R}}$ and $h\colon \mathbb{R}^m \to \bar{\mathbb{R}}$ are both proper, lsc, convex, and $A \in \mathbb{R}^{m\times n}$, $b \in \mathbb{R}^m$, $c \in \mathbb{R}^n$. For*

$$f(x,u) \;:=\; \langle c, x\rangle + k(x) + h(Ax - b + u)$$

the primal and dual problems are of the form

$$\inf_x \varphi(x), \quad \varphi(x) := \langle c, x\rangle + k(x) + h(Ax - b),$$
$$\sup_y \psi(y), \quad \psi(y) := -\langle b, y\rangle - h^*(y) - k^*(-A^\top y - c)$$

with

$$\begin{aligned} \operatorname{int}\operatorname{dom} p &= \operatorname{int}(\operatorname{dom} h - A \operatorname{dom} k) + b, \\ \operatorname{int}\operatorname{dom} q &= \operatorname{int}(\operatorname{dom} k^* - (-A^\top)\operatorname{dom} h^*) + c, \end{aligned}$$

and the optimality conditions

$$\left\{\begin{matrix} -A^\top y' - c \in \partial k(x') \\ y' \in \partial h(Ax' - b) \end{matrix}\right\} \Leftrightarrow \left\{\begin{array}{l} x' \in \arg\min_x \varphi(x), \\ y' \in \arg\max_y \psi(y), \\ \inf_x \varphi(x) = \sup_y \psi(y) \end{array}\right\} \Leftrightarrow \left\{\begin{matrix} Ax' - b \in \partial h^*(y') \\ x' \in \partial k^*(-A^\top y' - c) \end{matrix}\right\}.$$

Proof. f is convex (Prop. 3.14 sum is convex, Prop. 3.14 $f(Ax) + b$ is convex) and lsc (Prop. 2.16, sums of proper lsc functions qqare lsc, the composition of an lsc function with a continuous function is lsc). f is also proper because $f(x,u) = -\infty$ implies k or h not proper, and for some $x \in \operatorname{dom} k \neq \emptyset$ there must exist u s.t. $Ax - b + u \in \operatorname{dom} h$ because $\operatorname{dom} h \neq \emptyset$ and there are no other constraints on h.

With this choice we have $\varphi(x) = f(x,0)$ and (this is an important trick!)

$$\begin{aligned} f^*(v,y) &= \sup_{x,u} \langle x, v\rangle + \langle u, y\rangle - \langle c, x\rangle - k(x) - h(Ax - b + u) \\ &\overset{w = Ax-b+u}{=} \sup_{x,w} \langle x, v\rangle + \langle w - Ax + b, y\rangle - \langle c, x\rangle - k(x) - h(w) \\ &= \sup_{x,w} \langle x, -A^\top y + v - c\rangle - k(x) + \langle w, y\rangle - h(w) + \langle b, y\rangle \\ &= \langle b, y\rangle + \sup_x \{\langle x, -A^\top y + v - c\rangle - k(x)\} + \sup_w \{\langle w, y\rangle - h(w)\} \\ &= \langle b, y\rangle + k^*(-A^\top y + v - c) + h^*(y). \end{aligned}$$

Thus $\psi(y) = -f^*(0, y)$ (the case $v = 0$). Because f is proper, lsc, convex we can apply Prop. 7.4 and Prop. 7.5. We have

$$\begin{aligned} u \in \operatorname{dom} p \Leftrightarrow \inf_x f(x,u) < +\infty \;&\Leftrightarrow\; \inf_x \langle c, x\rangle + k(x) + h(Ax - b + u) < +\infty \\ &\Leftrightarrow\; \exists x\colon \langle c, x\rangle + k(x) + h(Ax - b + u) < +\infty \\ &\Leftrightarrow\; \exists x \in \operatorname{dom} k\colon h(Ax - b + u) < +\infty \\ &\Leftrightarrow\; \exists x \in \operatorname{dom} k\colon Ax - b + u \in \operatorname{dom} h \\ &\Leftrightarrow\; \exists x \in \operatorname{dom} k\colon u \in \operatorname{dom} h - Ax + b \\ &\Leftrightarrow\; u \in \operatorname{dom} h - A \operatorname{dom} k + b. \end{aligned}$$

Thus

$$0 \in \operatorname{int}\operatorname{dom} p \;\Leftrightarrow\; 0 \in \operatorname{int}(\operatorname{dom} h - A \operatorname{dom} k + b) \Leftrightarrow b \in \operatorname{int}(A \operatorname{dom} k - \operatorname{dom} h).$$

Similarly for $\operatorname{int}\operatorname{dom} q$.

Subdifferential of f: we have

$$f(x,u) = g(x,(u + A\,x)) \quad \text{with} \quad g(x,w) := \langle c, x\rangle + k(x) + h(w-b).$$

Then by the definition of the subdifferential (separable sum $\Rightarrow$ product of subdifferentials)

$$\partial g(x,w) \;=\; (c + \partial k(x)) \times (\partial h(w-b)).$$

Also, by Prop. 5.14 (chain rule) with $F \cdot (x,u) := \begin{pmatrix} I & 0 \\ A & I \end{pmatrix} \begin{pmatrix} x \\ u \end{pmatrix}$ and $f = g \circ F$ we get

$$\begin{aligned} \partial f(x,u) &= F^\top (\partial g(F(x,u))) \\ &= \{(c + v + A^\top y, y) | v \in \partial k(x), y \in \partial h(A\,x + u - b)\}. \end{aligned}$$

Therefore

$$(0,y') \;\in\; \partial f(x',0) \Leftrightarrow \begin{cases} 0 = c + v + A^\top y, \\ y' = y, \\ v \in \partial k(x'), \\ y \in \partial h(A\,x' - b + 0) \end{cases} \Leftrightarrow \begin{cases} -A^\top y' - c \in \partial k(x'), \\ y' \in \partial h(A\,x' - b). \end{cases}$$

Similarly for f^*. □

Example 7.7. (conic problems) Assume that K, L are pointed closed convex cones with polar cones K^*, L^* and consider the problem

$$\inf_x \langle c, x\rangle \quad \text{s.t.} \quad A\,x - b \geqslant_{L^*} 0, x \geqslant_K 0,$$

in alternative notation

$$\inf_x \langle c, x\rangle + \delta_K(x) + \delta_{L^*}(A\,x - b)$$

By Prop. 7.6 the dual problem is

$$\sup_y \psi(y), \quad \psi(y) := -\langle b, y\rangle - h^*(y) - k^*(-A^\top y - c),$$

which can be rewritten as

$$\begin{aligned} &\sup_y - \langle b, y\rangle - \delta^*_{L^*}(y) - \delta^*_K(-A^\top y - c) \\ =\; &\sup_y - \langle b, y\rangle - \delta_L(y) - \delta_{K^*}(-A^\top y - c). \end{aligned}$$

Thus the dual problem is

$$\sup - \langle b, y\rangle \quad \text{s.t.} \quad -A^\top y - c \geqslant_{K^*} 0, y \geqslant_L 0.$$

For *self-dual* cones with $K^* = -K$, such as $K = \mathbb{R}^n_{\geqslant 0}$ (and the same for L) we obtain

$$\begin{aligned} \inf_x \langle c, x\rangle \quad &\text{s.t.} \quad A\,x \leqslant_L b, x \geqslant_K 0, \\ \sup_y - \langle b, y\rangle \quad &\text{s.t.} \quad -A^\top y \leqslant_K c, y \geqslant_L 0. \end{aligned}$$

On important special case is the *Linear Programming duality*, where $K = \mathbb{R}^n_{\geqslant 0}$ and $L = \delta_0$:

$$\begin{aligned} \inf \langle c, x\rangle \quad &\text{s.t.} \quad A\,x = b, x \geqslant 0, \\ \sup - \langle b, y\rangle \quad &\text{s.t.} \quad -A^\top y \leqslant c. \end{aligned}$$

Note that $K = -K^*$ but the dual constraint on y disappears because $L^* = \mathbb{R}^n$.

Proposition 7.8. *(Lagrangian) Assume $f: \mathbb{R}^n \times \mathbb{R}^m \to \bar{\mathbb{R}}$ is proper, lsc, convex. We define the associated* Lagrangian *as*

$$l(x, y) := -f(x, \cdot)^*(y),$$

i.e.,

$$l(x, y) \;:=\; \inf_u \{ f(x, u) - \langle y, u \rangle \}.$$

Then $l(\cdot, y)$ is convex for every y, $-l(x, \cdot)$ is lsc and convex for every x, y, and

$$\begin{aligned} f(x, \cdot) \;&=\; (-l(x, \cdot))^*, \\ (v, y) \in \partial f(x, u) \;&\Leftrightarrow\; v \in \partial_x l(x, y) \text{ and } u \in \partial_y(-l)(x, y). \end{aligned}$$

Proof. Denote $g(x, y, u) := f(x, u) - \langle y, u \rangle$. f is proper lsc convex $\Rightarrow$ g is proper lsc convex. Thus

$$l(\cdot, y) \;=\; \inf_u g(\cdot, y, u)$$

is convex (but does not have to be proper for a specific y), see the proof of Prop. 7.2. Define $f_x(y) := f(x, y)$, then $-l(x, \cdot) = f_x^*(\cdot)$ and f_x is either $+\infty$ or proper lsc convex, therefore $-l(x, \cdot)$ is either $-\infty$ or proper lsc convex by Thm. 6.10, but always lsc convex as claimed.

We have, by definition of the subgradient,

$$\begin{aligned} (v, y) \in \partial f(x, u) \;&\Leftrightarrow\; f(x', u') \geqslant f(x, u) + \langle v, x' - x \rangle + \langle y, u' - u \rangle \quad \forall x', u' \\ &\Leftrightarrow\; f(x', u') - \langle y, u' \rangle \geqslant f(x, u) + \langle v, x' - x \rangle - \langle y, u \rangle \quad \forall x', u' \\ &\Leftrightarrow\; \inf_{u'} \{ f(x', u') - \langle y, u' \rangle \} \geqslant f(x, u) - \langle y, u \rangle + \langle v, x' - x \rangle \, \forall x'. \end{aligned} \tag{7.3}$$

Setting $x' = x$ the last line implies

$$\inf_{u'} \{ f(x, u') - \langle y, u' \rangle \} \;\geqslant\; f(x, u) - \langle y, u \rangle,$$

which is again equivalent to

$$\inf_{u'} \{ f(x, u') - \langle y, u' \rangle \} \;=\; f(x, u) - \langle y, u \rangle,$$

because $\inf_u \ldots \leqslant \ldots$ always holds (set $u = u'$). Thus we can continue (7.3) via

$$\begin{aligned} (7.3) \;&\Leftrightarrow\; \begin{cases} \inf_{u'} \{ f(x', u') - \langle y, u' \rangle \} \geqslant f(x, u) - \langle y, u \rangle + \langle v, x' - x \rangle \quad \forall x' \\ \inf_{u'} \{ f(x, u') - \langle y, u' \rangle \} = f(x, u) - \langle y, u \rangle \end{cases} \\ &\Leftrightarrow\; \begin{cases} \inf_{u'} \{ f(x', u') - \langle y, u' \rangle \} \geqslant \inf_{u'} \{ f(x, u') - \langle y, u' \rangle \} + \langle v, x' - x \rangle \quad \forall x' \\ \inf_{u'} \{ f(x, u') - \langle y, u' \rangle \} = f(x, u) - \langle y, u \rangle \end{cases} \\ &\Leftrightarrow\; \begin{cases} l(x', y) \geqslant l(x, y) + \langle v, x' - x \rangle \quad \forall x' \\ f(x, u') \geqslant f(x, u) + \langle y, u' - u \rangle \, \forall u' \end{cases} \\ &\Leftrightarrow\; \begin{cases} y \in \partial_u f(x, u), \\ v \in \partial_x l(x, y). \end{cases} \end{aligned}$$

Because f_x is $+\infty$ or proper lsc convex, by Prop. 6.12 (inversion of subdifferentials under Legendre-Fenchel transform) the first condition is equivalent to $u \in \partial f_x^*(y)$, which is the same as

$$u \;\in\; \partial(-l(x, \cdot))^{**}(y) = \partial_y(-l)(x, y).$$

Therefore we get

$$(7.3) \;\Leftrightarrow\; \begin{cases} u \in \partial_y(-l)(x,y), \\ v \in \partial_x l(x,y) \end{cases}$$

which shows the assertion. □

The Lagrangian opens a particularly nice way to formulate the optimality conditions:

Definition 7.9. *(saddle points) For any function $l\colon \mathbb{R}^n \times \mathbb{R}^m \to \bar{\mathbb{R}}$ we say that (x', y') is a* saddle point *of l iff*

$$l(x, y') \geqslant l(x', y') \geqslant l(x', y) \quad \forall x, y.$$

The set of all saddle-points is denoted by $\operatorname{sp} l$.

Remark 7.10. The condition $(x', y') \in \operatorname{sp} l$ is equivalent to

$$\inf_x l(x, y') = l(x', y') = \sup_y l(x', y),$$

i.e., x' minimizes $l(\cdot, y')$ and y' maximizes $l(x', \cdot)$.

The direction "⇐" is clear, "⇒" follows since

$$\inf_x l(x, y') \;\leqslant\; l(x', y')$$

always holds (set $x = x'$); together with the saddle-point condition we get equality (similarly for the supremum).

Proposition 7.11. *Assume f is proper, lsc, convex with associated Lagrangian l. Then*

$$\begin{aligned} \varphi(x) &= \sup_y l(x, y), \\ \psi(y) &= \inf_x l(x, y), \end{aligned}$$

and the primal and dual problems can be written as

$$\begin{aligned} \inf_x \varphi(x) &= \inf_x \sup_y l(x, y), \\ \sup_y \psi(y) &= \sup_y \inf_x l(x, y). \end{aligned}$$

Moreover, we have the optimality condition

$$\left\{ \begin{array}{l} x' \in \arg\min_x \varphi(x), \\ y' \in \arg\max_y \psi(y), \\ \inf_x \varphi(x) = \sup_y \psi(y) \end{array} \right\} \Leftrightarrow (x', y') \in \operatorname{sp} l \Leftrightarrow \left\{ \begin{array}{l} 0 \in \partial_x l(x', y'), \\ 0 \in \partial_y(-l)(x', y') \end{array} \right\}.$$

Proof. For fixed y (without any assumptions),

$$\begin{aligned} \inf_x l(x, y) &= \inf_x -\sup_u \{\langle y, u\rangle - f(x, u)\} \\ &= -\sup_{x,u} \{\langle (x, u), (0, y)\rangle - f(x, u)\} \\ &= -f^*(0, y) \\ &= \psi(y). \end{aligned}$$

For fixed x,

$$\begin{aligned}\sup_y l(x,y) &= \sup_y -f_x^*(y)\\ &= \sup_y \langle 0,y\rangle - f_x^*(y)\\ &= f_x^{**}(0)\\ &= f_x(0)\\ &= f(x,0)\\ &= \varphi(x).\end{aligned}$$

The equality $f_x^{**}=f$ holds because f_x is either proper lsc convex or $+\infty$, in which case $f_x^*=-\infty$ and $f_x^{**}=+\infty$.

By Prop. 7.4 the optimality condition is equivalent to $(0,y')\in\partial f(x',0)$, which by Prop. 7.8 is equivalent to having $0\in\partial_x l(x',y')$ and $0\in\partial_y(-l)(x',y')$. Since l and $-l$ are convex, this is exactly the saddle-point condition $(x',y')\in\operatorname{sp} l$ (Def. 7.9) and by Rem. 7.10 and the first part of the proposition equivalent to to $\varphi(x')=l(x',y')=\psi(y')$. □

Note that Prop. 7.11 only gives sufficient conditions for optimality. The following theorem answers the question of how to construct f from a given Lagrangian such that f is proper lsc convex and for every minimizer x' we can find a dual feasible point y' such that (x',y') is a saddle point.

Proposition 7.12. *Assume $X\subseteq\mathbb{R}^n$ and $Y\subseteq\mathbb{R}^m$ are nonempty, closed, convex, and*

$$L\colon X\times Y\to\mathbb{R}$$

is a continuous function with $L(\cdot,y)$ convex for every y and $-L(x,\cdot)$ convex for every x. Then

$$l(x,y) := L(x,y)+\delta_X(x)-\delta_Y(y),$$

with the convention $+\infty-\infty=+\infty$ on the right, is the Lagrangian to

$$f(x,u) := \sup_y\{l(x,y)+\langle u,y\rangle\} = (-l(x,\cdot))^*(u).$$

f is proper, lsc, and convex, i.e., Prop. 7.11 applies with primal and dual problems

$$\begin{aligned}&\inf_{x\in X}\sup_{y\in Y}L(x,y)=\inf_x\varphi(x),\quad \varphi(x):=\delta_X(x)+\sup_{y\in Y}L(x,y),\\ &\sup_{y\in Y}\inf_{x\in X}L(x,y)=\sup_y\psi(y),\quad \psi(y):=-\delta_Y(y)+\inf_{x\in X}L(x,y).\end{aligned}\tag{7.4}$$

Moreover, if X and Y are bounded, then $\operatorname{sp} l$ is nonempty and bounded.

Proof. For every x, $-l(x,\cdot)$ is either $-\infty$ (if $x\notin X$) or $-L(x,\cdot)+\delta_Y(\cdot)$ which is proper (L finite, $Y\neq\emptyset$), lsc (L continuous, Y closed) and convex (L convex, Y convex). Either way we have

$$-l(x,\cdot) = (-l(x,\cdot))^{**}=f(x,\cdot)^*$$

as required for l to be the Lagrangian to f (Prop. 7.8). We have $f(x,u)$ the supremum of convex functions (ranging over y) and therefore convex.

For the lower-semicontinuity, consider the mapping $g_y\colon(x,u)\mapsto L(x,u)+\langle u,y\rangle$. For every $y\in Y$, g_y is lower semi-continuous. The pointwise supremum of an arbitrary family of lsc functions is again lsc; this follows because their epigraphs are closed and therefore their intersection is also closed. This shows that the mapping

$$(x,u)\mapsto \sup_{y\in Y}\{L(x,y)+\langle u,y\rangle\}=\sup_y\{L(x,y)+\langle u,y\rangle-\delta_Y(y)\}$$

is lsc, and therefore $f(x,u) = \sup_y \{L(x,y) + \langle u, y\rangle - \delta_Y(y)\} + \delta_X(x)$ is lsc as well (the last step requires X to be closed and makes use of the convention $+\infty - \infty = 0$).

Properness: if $f(x,u) \equiv +\infty$ then $l(x,\cdot) = -(f(x,\cdot))^* \equiv +\infty$ for all x, which is not possible because it would mean $X = \emptyset$. For any fixed x, $f(x,\cdot)$ is either $+\infty$ or $x \in X$, in which case $-l(x,\cdot) = -L(x,\cdot) + \delta_Y$. This is proper lsc convex ($Y \neq \emptyset$). Thus $f(x,\cdot)$ must be proper lsc convex (Thm. 6.10), which means it cannot assume the value $-\infty$. Thus $f(x,u) \neq -\infty$ always; together f is jointly proper. All in all, the conditions in Prop. 7.11 are fulfilled.

The relations in (7.4) follow directly from the definition of l if one takes some caution to respect the convention $+\infty - \infty = +\infty$:

$$\begin{aligned} \psi(y) &= \inf_x \{L(x,y) + \delta_X(x) - \delta_Y(y)\} \\ &= \Big(\inf_{x\in X} L(x,y)\Big) - \delta_Y(y). \end{aligned}$$

The inf on the left side is always finite because $X \neq \emptyset$ and L is finite. Also

$$\begin{aligned} \varphi(x) &= \sup_y \{L(x,y) + \delta_X(x) - \delta_Y(y)\} \\ &= \begin{cases} +\infty, & x \notin X, \\ \sup_y L(x,y) - \delta_Y(y), & x \in X \end{cases} \\ &= \Big(\sup_{y\in Y} L(x,y)\Big) + \delta_X(x). \end{aligned}$$

To show that the set of saddle points is nonempty, consider

$$p(u) = \inf_{x\in X} \sup_{y\in Y} \{L(x,y) + \langle u, y\rangle\} \leqslant \sup_{x\in X} \sup_{y\in Y} \{L(x,y) + \langle u, y\rangle\}$$

(note that the inequality requires $X, Y \neq \emptyset$). X and Y are compact, therefore the right side is bounded and we get $\operatorname{dom} p = \mathbb{R}^m$; similarly $\operatorname{dom} q = \mathbb{R}^n$. In particular $0 \in \operatorname{int}\operatorname{dom} p$ and $0 \in \operatorname{int}\operatorname{dom} q$, and from Prop. 7.5 we obtain that the optimality conditions have a solution with a finite value. By Prop. 7.11 this implies

$$\operatorname{sp} l = (\operatorname{argmin} \varphi) \times (\arg\max \psi) \neq \emptyset.$$

Both sets are bounded because X and Y are bounded, therefore $\operatorname{sp} l$ is bounded as well. □

Example 7.13. For

$$l(x,y) := \langle c, x\rangle + k(x) - \langle b, y\rangle - h^*(y) + \langle A\,x, y\rangle$$

with k, h proper lsc convex we get

$$\begin{aligned} \varphi(x) &= \sup_y l(x,y) \\ &= \langle c, x\rangle + k(x) + h(A\,x - b) \\ \psi(y) &= \inf_x l(x,y) = -\langle b, y\rangle - h^*(y) - \sup_x \{\langle x, -A^\top y - c\rangle - k(x)\} \\ &= -\langle b, y\rangle - h^*(y) - k^*(-A^\top y - c). \end{aligned}$$

and

$$\begin{aligned} f(x,u) &= \sup_y \{l(x,y) + \langle u, y\rangle\} \\ &= \sup_y \{\langle c, x\rangle + k(x) - \langle b, y\rangle - h^*(y) + \langle A\,x, y\rangle + \langle u, y\rangle\} \\ &= \langle c, x\rangle + k(x) - h(A\,x - b + u). \end{aligned}$$

Example 7.14. (shrinkage) We consider the problem

$$\inf_x \frac{1}{2}\|x-a\|_2^2 + \lambda\|x\|_1. \tag{7.5}$$

We can reformulate the problem in "inf sup" form by rewriting the 1-norm using its dual norm: with the definition $Y := \{y \,|\, \|y\|_\infty \leqslant 1\}$,

$$\inf_x \sup_{y\in Y} \frac{1}{2}\|x-a\|_2^2 + \langle x, y\rangle.$$

We directly get the dual

$$\sup_{y\in Y} \inf_x \frac{1}{2}\|x-a\|_2^2 + \langle x, y\rangle.$$

The advantage of this formulation is that the Lagrangian is differentiable, so the (unconstrained!) inner problem can be solved explicitly by setting its gradient to zero, and we find that it solved by $x = a - y$. Substituting this we obtain

$$\sup_{y\in Y} -\frac{1}{2}\|y\|_2^2 + \langle a, y\rangle = \sup_{y\in Y} -\frac{1}{2}\|y-a\|_2^2.$$

This means that the dual problem is solved by computing the *projection* of a onto Y,

$$y' \;=\; \Pi_Y(a),$$

which can be computed explicitly and separately for each component. We can even obtain a primal solution from y': we know that x' minimizes $l(\cdot, y')$, so

$$x' \;=\; \arg\min \frac{1}{2}\|x-a\|_2^2 + \langle x, y'\rangle = \arg\min \frac{1}{2}\|x-a\|_2^2 + \langle x, \Pi_Y(a)\rangle.$$

The solution is unique, therefore we obtain the primal solution $x' = a - \Pi_Y(a)$. This operation is known as *shrinkage*, because it shrinks the value of a towards zero.

Note that recovering the primal solution from the dual in this way is only possible because of the uniqueness of the primal solution: in the general case, not every minimizer x'' of $l(\cdot, y')$ leads to a saddle point (x'', y'), as it may still violated the second half of the saddle-point condition.

The fact that problem (7.5) can be solved explicitly – and therefore exactly – and includes both the smooth 2-norm as well as the non-smooth 1-norm has made it a very popular problem to be solved as a sub-step for solving more complicated problems, see Chapter 9.

Example 7.15. (complementarity conditions) We consider Lagrangians of the form

$$l(x, y) \;:=\; \langle c, x\rangle + k(x) + \langle A\,x - b, y\rangle - \delta_L(y),$$

with a closed convex cone L and proper lsc convex function k, with the associated primal and dual problems

$$\begin{aligned} \inf_x \langle c, x\rangle + k(x) \quad &\text{s.t.} \quad A\,x \geqslant_{L^*} b, \\ \sup_y -\langle b, y\rangle - k^*(-A^\top y - c) \quad &\text{s.t.} \quad y \geqslant_L 0. \end{aligned}$$

Looking at the optimality conditions in terms of the Lagrangian as in Prop. 7.11,

$$\begin{aligned} 0 \in \partial_x l(x, y) \;&\Leftrightarrow\; 0 \;\in\; c + \partial k(x) + A^\top y, \\ 0 \in \partial_y(-l)(x, y) \;&\Leftrightarrow\; 0 \;\in\; -(A\,x - b) + N_L(y), \end{aligned}$$

we can apply the second part of Prop. 6.17 to rewrite the last condition through $v \in N_L(y) \Leftrightarrow v \in L^*, y \in L, \langle x, y \rangle = 0$, and get

$$0 \in c + A^\top y + \partial k(x),$$
$$0 \leqslant_L y \perp (A\,x - b) \geqslant_{L^*} 0.$$

The orthogonality constraint in the second equation is also known as a *complementarity condition.* To see why, set $A = I$, $b = 0$, and $L = \mathbb{R}_{\geqslant 0}$. The equation is then

$$0 \leqslant y \perp x \leqslant 0.$$

Because of the inequalities, no term the inner product $\langle y, x \rangle = 0$ can be positive, therefore $\langle y, x \rangle = 0$ is equivalent to $y_i\,x_i = 0$ for all i – the variables x_i and y_i are *complementary* in the sense that if one of them is nonzero, the other one must be zero (they could still both be zero, however).

Chapter 8
Numerical Optimality

When designing optimization methods an important question is what stopping criterion to choose. A very common pitfall is the following:

1. Fix a $\delta > 0$, x^0.
2. iterate: $k = 1, 2, \ldots$
   a. compute x^{k+1} from x^k,
   b. stop if $|\varphi(x^{k+1}) - \varphi(x^k)| < \delta$, or if $\|x^{k+1} - x^k\| < \delta$,
   c. $k \leftarrow k + 1$.

This approach suffers from all kinds of problems:

- It is also very dependent on the scaling of the problem or the date by constant factors.
- It stops when the solver is *slow*, which – without further knowledge about the algorithm – does not imply that the iterate is close to the solution. In fact the trivial update $x^{k+1} \leftarrow x^k$ "converges" after one iteration but the solution is useless.
- We do not get any information about how close x^k is to the minimizer x' or how close $f(x^k)$ is to $f(x')$.

Ultimately we would like our method to find a solution withing a guaranteed distance to the optimal solution:

Definition 8.1. *For $\varphi: \mathbb{R}^n \to \bar{\mathbb{R}}$, a point x is an ε-optimal solution if*

$$\varphi(x) - \inf \varphi \leqslant \varepsilon.$$

This is the "ideal" stopping criterion – it guarantees that energy-wise x is not much worse than a true minimizer x'. Unfortunately $\varphi(x')$ is generally unknown. What can we do?

8.1 The smooth and the non-smooth case

We first consider the smooth case. Assume that f is convex with a unique minimizer, and $f \in C^2$. The usual (local) convergence analysis for iterative methods hinges on two important assumptions.

1. f is *strongly convex*: $\nabla^2 f(x) \geqslant \sigma_- I$ for some $\sigma_- > 0$ and all x, i.e., the eigenvalues of the Hessian are uniformly bounded away from zero.

We get (Taylor)

$$f(y) \;=\; f(x) + \langle \nabla f(x), y-x\rangle + \frac{1}{2}(y-x)^\top \nabla^2 f(z)(y-x)$$

for some $z = (1-t)\,x + t\,y$, $t \in [0,1]$. Thus

$$f(y) \;\geqslant\; f(x) + \langle \nabla f(x), y-x\rangle + \frac{\sigma_-}{2}\,\|y-x\|_2^2.$$

This is a better bound than convexity gives; convexity alone corresponds to $\sigma_- = 0$.

a. The right-hand side is maximized by $\bar{y} = x - \frac{1}{\sigma_-}\nabla f(x)$, substituting this gives

$$f(y) \;\geqslant\; f(x) - \frac{1}{2\,\sigma_-}\,\|\nabla f(x)\|_2^2.$$

In particular for $x = x^k$ and $y = x'$ (a minimizer) we get

$$f(x^k) - f(x') \;\leqslant\; \frac{1}{2\,\sigma_-}\,\|\nabla f(x^k)\|_2^2.$$

This means that $f(x^k)$ approaches $f(x')$as $\|\nabla f(x^k)\|_2 \to 0$.

b. Another consequence is (x' is optimal!)

$$\begin{aligned} 0 \geqslant f(x') - f(x^k) \;&\geqslant\; \langle \nabla f(x^k), x^k - x'\rangle + \frac{\sigma_-}{2}\,\|x^k - x'\|_2^2 \\ &\geqslant\; -\|\nabla f(x^k)\|_2\,\|x^k - x'\|_2 + \frac{\sigma_-}{2}\,\|x^k - x'\|_2^2, \end{aligned}$$

thus

$$\|x^k - x'\|_2 \;\leqslant\; \frac{2}{\sigma_-}\,\|\nabla f(x^k)\|_2,$$

meaning that also x^k approaches x' as $\|\nabla f(x^k)\|_2 \to 0$.

2. The Hessian of f is uniformly bounded from above: $\nabla^2 f \leqslant \sigma_+ I$ for some $\sigma_+ > 0$ and all x. Then (again using Taylor and minimizing both sides over x)

$$f(x^k) - f(x') \;\geqslant\; \frac{1}{2\,\sigma_+}\,\|\nabla f(x^k)\|_2^2.$$

This gives the reverse statement: $\|\nabla f(x^k)\|_2$ approaches 0 as $f(x^k) \to f(x')$.

Together we can upper- and lower-bound $f(x')$ using quadratic upper and lower bounds on f. The convergence speeds of gradient-based methods depends heavily on the *condition*

$$\kappa := \frac{\sigma_+}{\sigma_-}.$$

In general convex optimizations, problems are usually

- *non-differentiable* ($\Rightarrow \sigma_+ = +\infty$) and
- have *affine* regions ($\Rightarrow \sigma_- = 0$)

(this is most easily visualized using a simple function such as $f(x) = |x|$). Therefore they do not have a good "classical" condition (not even locally, again consider $f(x) = |x|$).

One problem is that the subgradients usually carry no information about how close x^k is to x' (or even just how close $f(x^k)$ is to $f(x')$). Another viewpoint is the following:

Subgradients in convex optimization only provide a lower bound for the function, and may be a very bad linear approximation!

Some issues that arise from this:

1. Since f is not differentiable, it can have many subgradients at any given point. Evaluating whole subdifferentials is generally very hard. Which and how do we pick one of the subgradients? (The smallest one? That again amounts to solving a minimization problem over the set of subgradients. Or a random one? Then we have to deal with non-deterministic methods.)
2. Even if $x^k \to x'$ and $f(x^k) \to f(x)$, we can have $\partial f(x^k) \equiv \{v\}$ until $x' = x^k$ holds *exactly* (smooth case: 2. shows that $\|\nabla f(x^k)\|_2 \to 0$). In practice we only have finite precision, so this condition cannot be checked exactly.
3. The minimizer can be non-unique, e.g., $f(x) = \max\{x - 1, -x - 1\}$. Even if $f(x^k) \to f(x')$ this does not say anything about convergence of the sequence x^k. We could possibly pass onto a converging subsequence if f is level-bounded, but that is only useful in theory rather than in an actual implementation.
4. Even if all conditions hold and σ_+ and σ_- exist, they are not much value in practice, because they are often unknown or can only be estimated very roughly, and therefore provide conservative estimates that are useless in practice. This can lead to slow convergence or bad estimates if we would like to know how close *one specific* iterate x^k is to the minimizer rather than get an asymptotic convergence rate.

We need a more reliable *practical* criterion to check optimality.

8.2 The numerical primal-dual gap

Proposition 8.2. *(numerical primal-dual gap) Assume (x^k, y^k) is a primal-dual feasible pair (or "primal-dual feasible"), i.e., $x^k \in \operatorname{dom}\varphi$ and $y^k \in \operatorname{dom}\psi$. Then*

$$\varphi(x^k) \;\geqslant\; \psi(y^k)$$

and

$$0 \leqslant \varphi(x^k) - \inf\varphi \;\leqslant\; \varphi(x^k) - \psi(y^k) \quad =:\; \gamma(x^k, y^k) =:\; \gamma.$$

The quantity γ is the "numerical primal-dual gap". If $\gamma < \varepsilon$ then x^k is an ε-optimal solution with "optimality certificate" y^k.

Proof. The first inequality follows directly from weak duality. The second inequality follows from $\varphi(x') \geqslant \psi(y^k)$ which again holds due to weak duality. □

This means that if for a given x^k we can find y^k such that $\varphi(x^k) - \psi(y^k) \leqslant \varepsilon$ we have *proof* that x^k is an ε-optimal solution, with y^k acting as a *certificate* of optimality! This is an important concept – solvers can *prove* that their solution is ε-optimal.

If strong duality holds (i.e., $\inf\varphi = \sup\psi \in \mathbb{R}$), then such (x^k, y^k) always exist for arbitrarily small $\varepsilon > 0$, as we can take any minimizing/maximizing pair of sequences for the primal/dual problems. Existence of a certificate for $\varepsilon = 0$ requires existence of a primal-dual optimal pair.

The numerical primal-dual gap is translation invariant, i.e., if $f'(x, u) := f(x, u) + c$ for some $c \in \mathbb{R}$ (thus $\varphi'(x) = \varphi(x) + c$, $\psi'(y) = \psi'(y) + c$; verify this using $\psi(y) = -f^*(0, y)$) then

$$\begin{aligned} \varphi(x^k) - \psi(y^k) \;&\leqslant\; \varepsilon \\ \Leftrightarrow \varphi'(x^k) - \psi'(y^k) \;&\leqslant\; \varepsilon. \end{aligned}$$

It is however *not* scale invariant, i.e., if $f'(x,u) = c\,f(x,u)$ for $c>0$ (then $\varphi'(x) = c\,\varphi(x)$, $\psi'(y) = -f^*(0,\ y) = -\sup_{x,u}\ \{\langle u,\ y\rangle\ -\ c\ f(x,\ u)\} = -\sup_{x,u}\ \left\{c\ \left\langle u,\ \frac{1}{c}\ y\right\rangle\ -\ c\ f(x,\ u)\right\} = -c\,f^*\left(0, \frac{y}{c}\right) = c\,\psi(y/c)$) and even if we get the iterates $(x^k, c\,y^k)$, then

$$\varphi'(x^k) - \psi'(y^k)\ \leqslant\ c\,\varepsilon,$$

i.e., if we stop on the numerical primal-dual gap we get different solutions although the problem is effectively the same!

Therefore in practice the normalized gap is often used instead:

Definition 8.3. *(normalized gap) We define the* normalized numerical primal-dual gap *as*

$$\bar\gamma := \bar\gamma(x^k, y^k)\ :=\ \frac{\varphi(x^k) - \psi(y^k)}{\psi(y^k)}.$$

We get

$$\begin{aligned}\bar\gamma &\overset{y' \text{ dual optimal}}{\geqslant} \frac{\varphi(x^k) - \psi(y^k)}{\psi(y')}\\ &\overset{\text{weak duality}}{\geqslant} \frac{\varphi(x^k) - \psi(y^k)}{\varphi(x')}\\ &\overset{\text{weak duality}}{\geqslant} \frac{\varphi(x^k) - \varphi(x')}{\varphi(x')}.\end{aligned}$$

Therefore a small normalized gap $\bar\gamma \leqslant \varepsilon$ guarantees that x^k is $\varepsilon\,\varphi(x')$-optimal. $\bar\gamma$ is scale-invariant due to the normalization. It is however *not* translation-invariant – by adding a large constant c to φ, ψ we can make it arbitrarily small:

$$\gamma'\ =\ \frac{\varphi(x^k) + c - \psi(y^k) - c}{\psi(y^k) + c} = \frac{\varphi(x^k) - \psi(y^k)}{\psi(y^k) + c}.$$

The normalized gap also requires $\varphi(x') > 0$. Nevertheless, it is an excellent stopping criterion and widely used in practice.

Common values are in the range of $\delta = 10^{-4}$ for inaccurate solutions – which are often almost identical to high-accuracy solutions in image processing – to $\delta = 10^{-8}$ if precise solutions are required.

8.3 Infeasibilities

When computing the numerical primal-dual gap there is an important detail: We need to work with the *full extended real-valued* φ and ψ. This is very easy to lose track of, as the following example shows.

Example 8.4. Consider

$$\begin{aligned}\inf_{x\in[1,2]} |x| &= \inf_{x\in[1,2]} \sup_{y\in[-1,1]} x\,y\\ &= \sup_{y\in[-1,1]} \inf_{x\in[1,2]} x\,y\\ &= \sup_{y\in[-1,1]} \min\{y, 2\,y\}.\end{aligned}$$

An optimal primal-dual pair is $(x', y') = (1,1)$, then $\varphi(x') = |x'| = 1 = \min\{1, 2\cdot 1\} = \psi y'$, i.e., the numerical duality gap is zero.

Assume we have a point $x^k = 1 + \delta_k$, $y^k = 1 + \delta_k$ with $\delta_k \searrow 0$, then $(x^k, y^k) \to (x', y')$, and

$$|x^k| - \min\{y^k, 2\,y^k\} \;=\; 1 + \delta_k - (1 + \delta_k) = 0.$$

This means the numerical primal-dual gap is computed as *zero*, although none of the iterates x^k are primal optimal. In fact, we could produce a *negative* gap by setting $y^k = 1 + 2\,\delta_k$, which would clearly violate weak duality.

The reason for this is that we did not use the complete primal and dual objectives to compute the gap. In fact

$$\varphi(x) = |x| + \delta_{[1,2]}(x), \quad \psi(y) = \min\{y, 2\,y\} - \delta_{[-1,1]}(y).$$

In our example we get $\varphi(x^k) = 1$, but $\psi(y^k) = -\infty$ because the y^k are *not* dual feasible. The numerical primal-dual gap is therefore not zero but $+\infty$!

Unfortunately we cannot accurately deal with this issue – the primal or dual constraints could be non-trivial equality constraints, and in the general case cannot be fulfilled exactly using finite precision arithmetic. Also, we would like to have an indicator of how the primal-dual sequence converges even if it is infeasible.

One way to deal with this issue is to split the primal and dual objectives:

Definition 8.5. *Assume that*

$$\begin{aligned}\varphi(x) &= \varphi_0(x) + \sum_{i=1}^{n_p} \delta_{g_i(x)\leqslant 0},\\ \psi(y) &= \psi_0(y) - \sum_{i=1}^{n_d} \delta_{h_i(x)\leqslant 0},\end{aligned}$$

where $\operatorname{dom}\varphi_0 = \operatorname{dom}\psi_0 = \mathbb{R}^n$ *and* $g_i\colon \mathbb{R}^n \to \mathbb{R}$, $h_i\colon \mathbb{R}^m \to \mathbb{R}$ *are suitable continuous real-valued convex functions, i.e., the primal and dual constraints are of the form*

$$\begin{aligned} g_i(x) &\leqslant 0, \quad i \in \{1, \ldots, n_p\},\\ h_i(y) &\leqslant 0, \quad i \in \{1, \ldots, n_d\}.\end{aligned}$$

Then the primal and dual infeasibilities are defined as

$$\begin{aligned}\eta_p &:= \max\{0, g_1(x^k), \ldots, g_{n_p}(x^k)\},\\ \eta_d &:= \max\{0, h_1(y^k), \ldots, h_{n_d}(y^k)\}.\end{aligned}$$

Writing the objectives in such a way is often natural. We can then use the stopping criterion

$$\max\{\bar\gamma_0, \eta_p, \eta_d\} \;\leqslant\; \varepsilon, \quad \bar\gamma_0 := \frac{\varphi_0(x^k) - \psi_0(y^k)}{\psi_0(y^k)}.$$

This enforces not only a small primal-dual gap but also that the primal and dual iterates are close to being feasible.

For Ex. 8.4 we get

$$\begin{aligned}\bar\gamma_0 &= (1 + \delta_k - (1 + \delta_k))/(1 + \delta_k) = 0,\\ \eta_p &= \max\{0, 1 - x^k, x^k - 2\} = 0,\\ \eta_d &= \max\{0, y^k - 1, -1 - y^k\} = \delta_k.\end{aligned}$$

This makes it obvious that although the apparent numerical duality gap is zero, it cannot be fully trusted because the dual infeasibility is not zero.

Chapter 9
First-Order Methods

9.1 Forward and backward steps

Traditional gradient descent for differentiable functions:

$$x^{k+1} \;=\; x^k - \tau_k \nabla f(x^k)$$

for some step size sequence t^k.

Idea: The basic gradient descent discretizes the PDE

$$x_t \;=\; -\nabla f(x)$$

using explicit (forward) Euler steps with step length τ_k. For non-smooth f this is generally not possible, but what if we change it to

$$x_t \;\in\; \partial f(x)?$$

Indeed it is possible to show that if $f\colon \mathbb{R}^n \to \bar{\mathbb{R}}$ is proper lsc convex with $\arg\min f \neq \emptyset$, and $x_0 \in \operatorname{cl}\operatorname{dom} f$, then there exists a *unique* path $x(t), t \in [0,\infty)$ with $x(0) = x_0$, $\partial f(x(t)) \neq \emptyset$ for all t and

$$\frac{d}{dt}x(t) \;\in\; -\partial f(x(t)) \quad \text{for a.e. } t,$$

where $t \mapsto x(t)$ is absolutely continuous on all intervals $[\delta, \infty)$ and $x(t)$ converges to a minimizer of f (Bruck 1975).

In order to discretize this, there are two straightforward choices:

Definition 9.1. *For $f\colon \mathbb{R}^n \to \bar{\mathbb{R}}$ we define*

1. *the "forward step":*

$$F_{\tau_k f}(x^k) := (I - \tau_k \partial f)\, x^k,$$

2. *the "backward step":*

$$B_{\tau_k f}(x^k) := (I + \tau_k \partial f)^{-1} x^k.$$

These are called "forward" and "backward" steps, since they correspond to forward and backward Euler discretizations of the gradient descent PDE. The optimization problem can be seen as finding a *zero* of the set-valued mapping $\partial f\colon \mathbb{R}^n \rightrightarrows \mathbb{R}^n$, i.e.,

$$\text{find } x \in \mathbb{R}^n \text{ s.t. } 0 \;\in\; \partial f(x).$$

The forward and backward steps then amount to

$$x^{k+1} \in x^k - \tau_k \partial f(x^k), \qquad x^{k+1} \in x^k - \tau_k \partial f(x^{k+1}).$$

We see that the only difference is that in the backward step, the *new* iterate x^{k+1} is used to compute the subdifferential.

The backward step is also known as *proximal step*, because it can be reformulated as minimizing f together with an additional quadratic *prox term* that keeps the new iterate close to the previous iterate:

Proposition 9.2. *(proximal steps) If $f\colon \mathbb{R}^n \to \bar{\mathbb{R}}$ is proper lsc convex with $\tau > 0$, then the backward step is*

$$B_{\tau f}(x) \;=\; \arg\min_y \left\{ \frac{1}{2}\|y-x\|_2^2 + \tau f(y) \right\}$$

and therefore unique.

Proof.

$$\begin{aligned} y &\in B_{\tau f}(x) \\ \Leftrightarrow y &\in (I+\tau\partial f)^{-1}(x) \\ \Leftrightarrow x &\in (I+\tau\partial f)(y) \\ \Leftrightarrow 0 &\in y - x + \tau\partial f(y) \\ \Leftrightarrow y &\in \arg\min \left\{ \frac{1}{2}\|y-x\|_2^2 + \tau f(y) \right\}. \end{aligned}$$

□

If we are given a problem

$$\inf_x f(x),$$

without any additional knowledge about the structure of f the most straightforward approaches are to apply forward or backward steps to f:

Example 9.3. (forward and backward stepping) Assume f is proper lsc convex and $\arg\min f \neq \emptyset$.

1. *Forward stepping:*

$$x^{k+1} \;\in\; F_{\tau_k f}(x^k).$$

Convergence:

- Only restricted convergence results available (line search for smooth functions; but basic result for non-smooth functions requires knowledge of $\inf f$).
- Also: convergence of x^k to minimizer if there is $\alpha > 0$ such that $0 < \inf \tau_k \leqslant \sup \tau_k < 2\,\alpha$ and [Eck89, 3.3]

$$\langle y-x, h-g\rangle \geqslant \alpha\,\|h-g\|_2^2 \;\;\forall\;\; g\in\partial f(x), h\in\partial f(y).$$

In the smooth case this implies the Lipschitz condition:

$$\begin{aligned} \|y-x\|_2\,\|\nabla f(y)-\nabla f(x)\|_2 &\geqslant \langle y-x, \nabla f(y)-\nabla f(x)\rangle \\ &\geqslant \alpha\,\|\nabla f(x)-\nabla f(x)\|_2^2 \\ \Rightarrow \|\nabla f(x)-\nabla f(x)\|_2 &\leqslant \frac{1}{\alpha}\|y-x\|_2. \end{aligned} \tag{9.1}$$

Advantages:

- requires very little information (just a single subgradient)

- if $\inf \tau_k > 0$ then fixed points are minimizers of f [Eck89, 3.6].

Disadvantages:

- sequence not unique, depends on selection of subgradient
- can get stuck if x^k becomes infeasible

2. *Backward stepping:*

$$x^{k+1} \;=\; B_{\tau_k f}(x^k).$$

Convergence:

- $\sum_k \tau_k^2 = +\infty$ is sufficient (i.e., no upper bound on step size!)

Advantages:

- *unique* sequence
- cannot get stuck, infeasible starting point possible

Disadvantages:

- substeps are as hard as the original problem (but strictly convex)

The last point is generally the dealbreaker for backward steps. But we often encounter problems where the objective is a *sum* of several terms, each of which is easy to minimize:

$$f(x) \;=\; g(x) + h(x).$$

The question is, can we find a minimum of f by combining suitable alternating minimization steps of g and h?

Example 9.4. (splitting principle) Assume $f = g + h$ such that $\partial f = \partial g + \partial h$ with f, g, h proper lsc convex and $\arg\min f \neq \emptyset$.

1. *Backward-Backward:*

$$x^{k+1} \;=\; B_{\tau_k h}\, B_{\tau_k g}(x^k).$$

Convergence:

- *in the mean,* i.e.,

$$y^k \;:=\; \left(\sum_{k'=0}^{k} \tau_{k'}\, x^{k'+1}\right) \Big/ \left(\sum_{k'=0}^{k} \tau_{k'}\right)$$

converges to a minimizer of f if $\tau_k > 0$, $\sum_k \tau_k = +\infty$, $\sum_k \tau_k^2 < +\infty$.

Advantages:

- few restrictions on step size

Disadvantages:

- step size needs to approach 0
- fixed points (i.e., x with $x = B_{\tau h}\, B_{\tau g}(x)$ for some *fixed* τ) are generally *not* minimizers of f!
- convergence in the mean raises numerical difficulties for large number of steps

2. *Forward-Backward:*

$$x^{k+1} \;\in\; B_{\tau_k h}\, F_{\tau_k g}(x^k).$$

A very popular special case is *gradient-projection*, where $f(x) = g(x) + \delta_C(x)$, g is differentiable, $C \neq \emptyset$ closed and convex. Then

$$\begin{aligned} x^{k+1} &\in \arg\min_x \left\{ \frac{1}{2} \|y - (x^k - \tau_k \nabla g(x^k))\|_2^2 + \delta_C(x) \right\} \\ &= \underbrace{\Pi_C\Big(x^k - \underbrace{\tau_k \nabla g(x^k)}_{\text{gradient}} \Big)}_{\text{projection}}. \end{aligned}$$

Convergence:

- gradient-projection: if (9.1) holds and $0 < \inf \tau_k \leqslant \sup \tau_k < 2\,\alpha$.
- general case: in the mean if $\sum_k \tau_k = +\infty$ and $\sum_k \tau_k^2 < +\infty$, *and* the chosen sequence of subgradients $y^k \in F_{\tau_k g}(x^k)$ is bounded (this is not clear and needs to be shown separately)

Advantages:

- fixed points are the minimizers of f
- requires backward steps on h only, and can therefore deal with relatively complicated functions g (think $g(x) = \|A\ x\ -\ b\|_2^2$ where backward steps involve solving large systems)

Disadvantages:

- sequence not unique, may get stuck
- can escape from minimizer! (if $\partial g(x') \neq \{0\}$, i.e., contains non-zero subgradients)

Example 9.5. (ISTA) Assume we would like to reconstruct an image $y \in \mathbb{R}^n$ from the observation $b = R\,y$, where R is a linear operator. This could be a blurring operator, a super-resolution operator, or just the identity for denoising of a given image.

As prior knowledge about the image to be recovered we make the first assumption that

$$y \;=\; W x,$$

where $W \in \mathbb{R}^{n \times m}$ is a set of basis vectors (possible overcomplete). One choice is to use a wavelet basis consisting of scaled and translates copies of a "mother wavelet", for which the product $x \mapsto W\,x$ can be evaluated quickly although W is large and non-sparse. The second and key assumption is that x is *sparse*, i.e., it has only a few non-zeros: we assume that the image can be represented using only a few of the basis functions.

An approach to recover y is to set $A = R\,W$ and find

$$x \;\in\; \arg\min_x \left\{ \frac{1}{2} \|A\,x - b\|_2^2 + \lambda\, \|x\|_1 \right\}.$$

The left term ensures that the resulting image is close to the observation, the right term promotes sparsity of x. If we apply forward-backward splitting with $f = g + h$, $g(x) = \frac{1}{2}\|A\,x - b\|_2$, $h(x) = \lambda\,\|x\|_1$, we get

$$x^{k+1} \;=\; \arg\min_x \left\{ \frac{1}{2} \|x - (x^k - \tau_k\, A^\top (A\,x^k - b))\|_2^2 + \tau_k\,\lambda\, \|x\|_1 \right\}.$$

This is know as ISTA (iterative shrinkage thresholding) and has two strong points:

- It only requires us to do matrix multiplications involving A and $A^\top$, as opposed to solving linear equation systems involving A.

- The proximal step is *separable*, and can be solved explicitly using *shrinkage* (see example sheets).
- Convergence: $O\Big(\frac{1}{k}\Big)$, where k is the number of steps. A variant (FISTA) employs over-relaxation and an adaptive choice of the τ_k to achieve $O\Big(\frac{1}{k^2}\Big)$. This is – although far from a linear convergence rate – often enough in practice and widely used.

9.2 Primal-dual methods

A slightly different approach is to not try to minimize the function, but to solve the primal-dual *optimality conditions* instead. This is particularly fruitful in convex optimization, as a primal-dual solution not only yields a global minimizer of f, but also an *optimality certificate* in form of the dual feasible (and optimal, in case of a solution) point.

Consider the primal and dual problems with associated Lagrangian

$$\begin{aligned} \inf_x \varphi(x), \qquad & \varphi(x) = g(x) + h\,(A\,x), \\ \sup_y \psi(y), \qquad & \psi(y) = -g^*(-A^\top\,y) - h^*(y), \\ l(x,y) \;&=\; g(x) - h^*(y) + \langle A\,x, y\rangle. \end{aligned}$$

The optimality (saddle-point) conditions are

$$\begin{aligned} x' \;\in\; \arg\min l(x,y') \quad &=\quad \arg\min_x \{g(x) + \langle A^\top\,y', x\rangle\}, \\ y' \;\in\; \arg\max l(x',y) \quad &=\quad \arg\min_y \{h^*(y) + \langle -A\,x, y\rangle\}. \end{aligned}$$

An obvious strategy is to do alternate between improving the primal and dual objectives, keeping the other variable fixed. In the following we generally assume that strong duality holds.

Example 9.6. (primal-dual methods)

1. *PFBS (proximal forward-backward splitting):* We apply a full minimization step with respect to y, followed by a backward step on x. The minimization with respect to y can be rewritten as follows:

$$\begin{aligned} y^{k+1} \;&\in\; \arg\min_y \{h^*(y) + \langle -A\,x^k, y\rangle\} \\ \Leftrightarrow 0 \;&\in\; \partial h^*(y^{k+1}) - A\,x^k \\ \overset{\partial h^* = (\partial h)^{-1}}{\Leftrightarrow}\; y^{k+1} \;&\in\; \partial h(A\,x). \end{aligned}$$

   The backward step is then

$$\begin{aligned} x^{k+1} \;&\in\; \arg\min_x \left\{ g(x) + \langle A^\top\,y^{k+1}, x\rangle + \frac{1}{2\,\tau_k}\,\|x - x^k\|_2^2 \right\} \\ &=\; \arg\min_x \left\{ g(x) + \frac{1}{2\,\tau_k}\,\|x - (x^k - \tau_k\,A^\top\,y^{k+1})\|_2^2 \right\}. \end{aligned}$$

   By the chain rule (under sufficient regularity assumptions) $A^\top\,y^{k+1} \in \partial(h \circ A)(x^k)$, so PFBS actually amounts to a forward step on $h \circ A$ followed by a backward step on g – it is in fact the same as forward-backward splitting in the operator splitting approach!

Convergence:

- if $\nabla(h \circ A)$ is Lipschitz continuous (in particular differentiable!) with constant L, and $0 < \inf \tau_k \leqslant \sup \tau_k < 2/L$.

Advantages:

- requires proximal steps on g only
- compared to forward-backward splitting, provides an alternative way to compute the subgradient by minimizing h^* and generates a sequence of dual iterates y^k that can be used to compute the numerical primal-dual gap.

Disadvantages:

- step size restriction, need to estimate L

2. *Modified PDHG (primal-dual hybrid gradient):* We alternate between backward steps with respect to x and y with an additional over-relaxation:

$$\begin{aligned} y^{k+1} &= B_{\sigma_k(-l(\bar{x}^k,\cdot))}(y^k), \\ x^{k+1} &= B_{\tau_k(l(\cdot,y^{k+1}))}(x^k), \\ \bar{x}^{k+1} &= x^{k+1} + \theta_k\,(x^{k+1} - x^k). \end{aligned}$$

Convergence:

- in the mean if $\sigma_k \equiv \sigma, \tau_k \equiv \tau$, $\theta_k \equiv \theta = 1$, $\sigma\,\tau < 1/\|A\|^2$, then with $O(1/k)$.
- in the mean if g or h^* strongly convex with constant α: $O(1/k^2)$ for $\|x^k - x'\|_2$ (needs adaptive θ_k)
- in the mean if g and h^* strongly convex with constants α, β: $O(\omega^k)$ with $\omega = \frac{1+\theta}{2\left(1+\frac{\sqrt{\sigma\beta}}{\|A\|^2}\right)}$, i.e., linear convergence rate.

Advantages:

- few requirements for convergence, flexible
- efficiently removes linearity A

Disadvantages:

- step size bound
- convergence in the mean (although convergence in (x^k, y^k) is observed)
- even if $\|A\|$ is known, the ratio σ/τ may still need to be tuned for good convergence

Example 9.7. (Augmented Lagrangian): We shift h into the primal objective by adding an artificial variable z:

$$\begin{aligned} \inf_x \varphi(x), \qquad & \varphi(x) = g(x) + h\,(z) + \delta_{Ax=z}, \\ l((x,z),y) &= g(x) + h(z) + \langle A\,x - z, y\rangle. \end{aligned}$$

The dual variables y only occur as *multipliers*. Note the optimality conditions:

$$\begin{cases} 0 \in \partial_{(x,z)} l((x,z),y) \\ 0 \in \partial_y(-l)((x,z),y) \end{cases} \Leftrightarrow \begin{cases} 0 \in \partial g(x) + A^\top y, 0 \in \partial h(z) - y \\ A\,x = z \end{cases}$$

$$\Leftrightarrow \begin{cases} 0 \in \partial g(x) + A^\top y \\ y \in \partial h(A\,x) \end{cases} \overset{\text{subgrad. inversion}}{\Leftrightarrow} \begin{cases} 0 \in \partial g(x) + A^\top y \\ 0 \in \partial h^*(y) - A\,x \end{cases}.$$

This shows that solutions $((x,z),y)$ of the new problem are exactly the triples $((x,A\,x),y)$ where $(x,\,y)$ is a primal-dual solution for l! In particular, y is always a dual solution for the original problem, although we did not explicitly introduce it as such.

We *augment* the Lagrangian by adding a quadratic penalty:

$$\bar{l}\,((x,z),y) \;:=\; g(x)+h(z)+\langle A\,x-z,y\rangle+\frac{1}{2\,\sigma}\,\|A\,x-z\|_2^2.$$

This does not change the set of minimizers, because $A\,x=z$ holds for every minimizer. We then alternate between minimization with respect to x and z and gradient ascent for y with step size σ:

$$\begin{aligned}
x^{k+1} &= \arg\min_x\left\{g(x)+\frac{1}{2\,\sigma}\,\|A\,x-z^k+\sigma\,y^k\|_2^2\right\},\\
z^{k+1} &= \arg\min_z\left\{h(z)+\frac{1}{2\,\sigma}\,\|z-A\,x^{k+1}-\sigma\,y^k\|_2^2\right\},\\
y^{k+1} &= y^k+\frac{1}{\sigma}\,(A\,x^{k+1}-z^{k+1}).
\end{aligned}$$

This is known as the *alternating direction method of multipliers* (ADMM).

Convergence:

- for *any* $\sigma>0$ we have $x^k\to x'$ and $y^k\to y'$ where (x',y') primal-dual optimal pair; additionally $z^k\to A\,x'$

Advantages:

- very few assumptions required for convergence

Disadvantages:

- We need to solve problems involving the term $\|A\,x-\cdot\|_2^2$, which can be difficult due to the coupling of the variables in x.

Chapter 10
Interior-Point Methods

Setting: We would like to solve

$$\begin{aligned} &\inf \langle c, x\rangle \quad \text{s.t.} \quad A\,x - b \geqslant_K 0, \\ \triangleq &\inf \langle c, x\rangle \quad + \quad \delta_K(A\,x - b), \end{aligned}$$

where K is a proper closed convex cone.

Idea: We would like to use Newton's method

$$x^{k+1} \;=\; x^k - (\nabla^2 f(x))^{-1} \nabla f(x)$$

since it is usually very fast (locally quadratic convergence rate $\|x^{k+1} - x'\| \leqslant c\,\|x^k - x'\|^2$). Problem: the constraints!

Idea: We replace the indicator function $\delta_K(z)$ by an approximation $F(z)$ such that $F(z) \to +\infty$ as $z \to \operatorname{bd} K$ and solve

$$\inf t\,\langle c, x\rangle + F(A\,x - b).$$

As $t \to +\infty$ the minimizers should approach the minimizers of the original problem. Nevertheless, they all lie in the *interior* of the constraint set K, hence then name "interior-point methods" as opposed to methods that travel on the boundary of the constraint set. Two ideas:

- Can we *guarantee* that we get a solution with a certain accuracy, i.e., how do we need to choose t?
- As $t \to +\infty$ the problems become potentially more difficult, as the optimal point moves toward the boundary, and the condition of F become worse the closer one allows points to be from the boundary. Can we do this iteratively, i.e., if we know a solution for a certain t^k, can we quickly find a solution for a larger t^{k+1}?

Definition 10.1. *(canonical barrier) For a cone K we define the "canonical barriers" $F = F_K$ and associated parameters θ_F:*

- $K = K_n^{\mathrm{LP}} = \{x \in \mathbb{R}^n | x_1, ..., x_n \geqslant 0\}$*, we have the canonical barrier*

$$F(x) \;=\; \sum_{i=1}^{n} -\log x_i, \quad \theta_F = n,$$

- $K = K_n^{\mathrm{SOCP}} = \left\{ x \in \mathbb{R}^n \middle| x_n \geqslant \sqrt{x_1^2 + ... + x_{n-1}^2} \right\}$,

$$F(x) \;=\; -\log\left(x_n^2 - x_1^2 - ... - x_{n-1}^2\right), \quad \theta_F = 2,$$

- $K = K_n^{\mathrm{SDP}} := \{X \in \mathbb{R}^{n\times n} | X \text{ symmetric positive semidefinite}\}$,

$$F(X) \;=\; -\log\det X, \quad \theta_F = n.$$

- $K = K^1 \times K^2$, *then* $F_K(x^1, x^2) = F_{K^1}(x^1) + F_{K^2}(x^2)$ *with* $\theta_F = \theta_{F^1} + \theta_{F^2}$.

Proposition 10.2. *[Nem 6.3.1, 6.3.2, Boyd 11.6.1] If F is a canonical barrier for K, then F is smooth on* $\operatorname{dom} F = \operatorname{int} K$ *and strictly convex,*

$$F(t\,x) = F(x) - \theta_F \log t \qquad \forall x \in \operatorname{dom} F,$$

and for $x \in \operatorname{dom} F$ we have

1. $-\nabla F(x) \in \operatorname{dom} F$,
2. $\langle \nabla F(x), x \rangle = -\theta_F$,
3. $-\nabla F(-\nabla F(x)) = x$,
4. $-\nabla F(t\,x) = -\frac{1}{t} \nabla F(x)$.

Proof. [Boyd 11.6.1] differentiate with respect to t at $t = 1$. □

Again we consider (assume that A has full column rank (linearly independent columns), Nem p.53 Ass. A).

$$\inf \langle c, x \rangle \quad \text{s.t.} \quad A\,x - b \geqslant_K 0.$$

The *dual* problem is

$$\sup \langle -b, y \rangle \quad \text{s.t.} \quad -A^\top y = c,\ y \geqslant_{K^*} 0.$$

We replace y by $-y$:

$$\sup \langle b, y \rangle \quad \text{s.t.} \quad A^\top y = c,\ y \leqslant_{K^*} 0.$$

We then assume that K is self-dual ($K^* = K$) and get

$$\sup \langle b, y \rangle \quad \text{s.t.} \quad A^\top y = c,\ y \geqslant_K 0.$$

In the remainder of this chapter we generally assume that the primal and dual problems are strictly feasible, i.e., there exist x, y so that $A\,x - b \in \operatorname{int} K$, $A^\top y = c$, and $y \in \operatorname{int} K$. This guarantees that we can actually find interior point, and that strong duality holds.

Proposition 10.3. *(primal-dual central path) The* primal central path *is the mapping*

$$t \;\mapsto\; x(t) = \arg\min \{ t\,\langle c, x \rangle + F(A\,x - b) \}.$$

The dual central path *is the mapping*

$$t \;\mapsto\; y(t) = \arg\min \{ -t \langle b, y \rangle + F(y) + \delta_{A^\top y = c} \}.$$

The primal-dual central path *is the mapping $t \mapsto z(t) := (x(t), y(t))$. These central paths exist and are unique for all $t \geqslant 0$. Also,*

$$y(t) \;=\; -\frac{1}{t} \nabla F(A\,x(t) - b),$$

and (x, y) is on the central path, i.e., $(x, y) = (x(t), y(t))$ for some $t > 0$, if and only if

$$\begin{aligned} &A\,x - b \in \operatorname{dom} F, \\ &A^\top y = c, \\ &t\,y + \nabla F(A\,x - b) = 0. \end{aligned}$$

Proof. Existence, uniqueness: see Nem. 6.4.2 (strict convexity etc.).

Second part: From Prop. 10.2 we get that $y := -\frac{1}{t}\,\nabla F(A\,x - b) \in \operatorname{dom} F$ because $A\,x - b \in \operatorname{dom} F$ (x is strictly feasible) and $\operatorname{dom} F$ is a cone. Also, if x is on the primal central path, then

$$\begin{aligned} 0 &= t\,c + A^\top \nabla F(A\,x - b) \\ \Leftrightarrow c &= -\frac{1}{t}\,A^\top \nabla F(A\,x - b). \end{aligned}$$

Therefore

$$A^\top y = -\frac{1}{t}\,A^\top \nabla F(A\,x - b) = c.$$

Together this shows that y is feasible.

The dual optimality conditions are therefore

$$0 \in -t\,b + \nabla F(y) + \underbrace{N_{A^\top y = c}}_{\operatorname{rge} A = A\mathbb{R}^n}.$$

We check:

$$\begin{aligned} -t\,b + \nabla F(y) &= -t\,b + \nabla F\left(-\frac{1}{t}\,\nabla F(A\,x - b)\right) \\ &\overset{10.2}{=} -t\,b + t\,\nabla F(-\nabla F(A\,x - b)) \\ &\overset{10.2}{=} -t\,b - t\,(A\,x - b) \\ &= -t\,A\,x \in \operatorname{rge} A. \end{aligned}$$

Therefore y is a dual solution and therefore *the* dual solution (because it is unique).

Last part: Multiplying the last line by $A^\top$, we get

$$\begin{aligned} 0 &= t\,A^\top y + A^\top \nabla F(A\,x - b) \\ &= t\,c + A^\top \nabla F(A\,x - b), \end{aligned}$$

which is the optimality condition for the primal central point. □

Proposition 10.4. *(duality gap along the path) For feasible x, y (i.e., $A\,x - b \in K$, $A^\top y = c$, $y \in K$), the duality gap is*

$$\varphi(x) - \psi(y) = \langle y, A\,x - b\rangle.$$

Moreover, for points $(x(t), y(t))$ on the central path, the duality gap is

$$\varphi(x(t)) - \psi(y(t)) = \frac{\theta_F}{t}.$$

Proof.

$$\begin{aligned} \varphi(x) - \psi(y) &= \langle c, x\rangle - \langle b, y\rangle \\ &= \langle A^\top y, x\rangle - \langle b, y\rangle \\ &= \langle y, A\,x - b\rangle. \end{aligned}$$

$$\begin{aligned} \varphi(x(t)) - \psi(y(t)) &= \langle y(t), A\,x(t) - b\rangle \\ &\overset{\text{Prop. 10.3}}{=} \langle -t^{-1}\,\nabla F(A\,x(t) - b), A\,x(t) - b\rangle \\ &\overset{\text{Prop. 10.2}}{=} \frac{\theta_F}{t}. \end{aligned}$$

□

Remark: This has an intriguing consequence: If we are given x, y on the central path, we can immediately find the unique t to which they belong by computing the numerical gap!

We do not have to compute *exact* solutions, it is essentially enough to stay *close* to the central path:

Proposition 10.5. *(near the central path) We define*

$$\|v\|_x^* \;:=\; (v^\top \nabla^2 F(A\,x-b)^{-1}\,v)^{1/2},$$

$z := (x,y)$*, so* $z(t)$ *is the primal-dual central path, and*

$$\operatorname{dist}(z,z(t)) \;=\; \|t\,y+\nabla F(A\,x-b)\|_x^*.$$

Then for $A\,x-b\in\operatorname{dom}F$*,* $y\in\operatorname{dom}F$*,* $A^\top y=c$*, we have*

$$\operatorname{dist}(z,z(t))\leqslant 1 \;\Rightarrow\; \varphi(x)-\psi(y)\leqslant 2\,(\varphi(x(t))-\psi(y(t)))=\frac{2\,\theta_F}{t}.$$

Proof. See Nemirovski. □

(path tracing) Idea: From 10.3 we know that a primal-dual optimal pair for a *fixed* t can be found by solving

$$A^\top y=c,\quad t\,y+\nabla F(A\,x-b)=0.$$

Idea: Newton for the nonlinear equation system; we linearize:

$$\nabla F(A\,x^{k+1}-b)=\nabla F(A\,x^k-b+A\,\Delta x)\approx\nabla F(A\,x^k-b)+\nabla^2 F(A\,x^k-b)\,A\,\Delta x$$

and get the following linear equation system for the steps Δx and Δy:

$$\begin{aligned} A^\top(y^k+\Delta y) &= c,\\ t^{k+1}(y^k+\Delta y)+\nabla F(A\,x^k-b)+\nabla^2 F(A\,x^k-b)\,A\,\Delta x &= 0. \end{aligned}$$

Multiply the last line by $A^\top$ from the left (to eliminate y) and substitute; we get

$$\begin{aligned} \Delta x &= H^{-1}(-t^{k+1}c-A^\top\nabla F(A\,x^k-b)),\quad H:=A^\top\nabla^2F(A\,x^k-b)\,A,\\ \Delta y &= -(t^{k+1})^{-1}(\nabla F(A\,x^k-b)+\nabla^2F(A\,x^k-b)\,A\,\Delta x)-y^k. \end{aligned}$$

Here we need the assumption that A has full column rank in order for H to be regular. We apply a step in the direction of $(\Delta x,\Delta y)$:

$$(x^{k+1},y^{k+1}) \;=\; (x^k,y^k)+\tau_k\,(\Delta x,\Delta y) \tag{10.1}$$

with the step size τ_k found using line search or a full Newton step ($\tau_k=1$):

Theorem 10.6. *Assume* $0<\rho\leqslant\kappa<\frac{1}{10}$*,* $t^k>0$ *fixed and* $z^k=(x^k,y^k)$ *strictly feasible, i.e.,* $A\,x^k-b\in\operatorname{dom}F$*,* $y^k\in\operatorname{dom}f$*, such that*

$$\operatorname{dist}(z^k,z(t^k)) \;<\; \kappa.$$

If we apply a full *Newton step in (10.1) with* $\tau_k=1$ *and* $t^{k+1}:=\left(1+\frac{\rho}{\sqrt{\theta_F}}\right)t^k$ *to generate* z^{k+1}*, then* x^{k+1},y^{k+1} *are strictly primal and dual feasible, and*

$$\operatorname{dist}(z^{k+1},z(t^{k+1})) \;<\; \kappa$$

as well.

(The idea is to bound the step for t^k so as to keep the Newton method within its region of quadratic convergence. The factor $\frac{1}{10}$ comes from a nasty expression involving ρ and κ

Advantages:

- We can update the penalty after a *single* Newton step!
- In particular, this means that

$$\varphi(x^k) - \psi(y^k) \;\leqslant\; 2\,\frac{\theta_F}{t^k} = \frac{2\,\theta_F}{t_0}\left(1 + \frac{\rho}{\sqrt{\theta_F}}\right)^{-k}.$$

  We have guaranteed global *linear* convergence with *explicit* bounds!
- In practice, we can often do much larger steps for t (we could imagine computing t from the duality gap!)
- Rapid convergence if the Newton steps can be solved in reasonable time.

Disadvantages:

- Relatively complicated.
- Many detail problems: finding a feasible point, computing the Newton step (or Quasi-Newton approximation)
- Requires problem-specific code for medium-scale problems to be efficient.

Chapter 11
Using Solvers and Discretization Issues

See `cvx_demo.zip`, available online.

Chapter 12
Support Vector Machines

12.1 Introduction to machine learning

General problem: Given a sample set of *feature vectors* $\mathcal{X} = \{x^1, ..., x^n\}$, taken from some *feature space* $\mathcal{F} \subseteq \mathbb{R}^m$, find a *classifier* function h_θ: $\mathbb{R}^m \to \mathcal{C}$ from a class of functions $\{h_\theta | \theta \in \Theta\}$ that "best" maps each feature vector into a corresponding class from the set $\mathcal{C}$, where usually $\mathcal{C}$ is a finite subset of $\mathbb{Z}$.

- In *unsupervised learning*, only the feature vectors $\mathcal{X}$ are known, together with some prior knowledge about the structure of the feature space and the distribution of the data. The task is to automatically find classes and *cluster* the samples into sets that are simlar with respect to some criterion.

  A typical application is *clustering:* an online shop might be interested to automatically classify their customers into groups to better target the individual group interests, or they might want to improve categorization of their products on the website. In image processing, one might want to separate foreground and background of an image, only knowing that they look "different", but not what they look like – it might be a white bird against a blue sky, but it also might be a squirrel on a tree. Other applications include the detection of "unusual" events, e.g., in a video stream, or identifying and mapping the data points to few low-dimensional subspaces as in Principal Component Analysis (PCA).

  Generally, there is no *a priori* knowledge on what the individual classes should be, and often the number of classes isn't even known. This makes unsupervised learning a very hard problem.

- In *supervised learning*, the number of classes is generally known, and there is a set of *training data* $\mathcal{T} = \{(x^1, y^1), ..., (x^n, y^n)\}$, consisting of pairs of a feature vector $x^i \in \mathcal{F}$ and a class label $y^i \in \mathcal{C}$.

  Usually such data is obtained by hand-labeling data or using historical data. For example, we might want to predict whether it is adivsable to buy a particular stock based on the recent development of the stock price. In image processing, we might be interested in finding a way to detect cancer cells from regular cells based on a library of cell images that have been classified beforehand by an expert, classify classify brain regions or scans, detect vertebrae, faces or body parts, or generally identify objects in images for search indexing or finding related images.

In this chapter we will only consider *supervised* learning. Most supervised learning techniques can be summarized in the following energy-based framework:

For a given training data set $\mathcal{T} = \{(x^1, y^1), ..., (x^n, y^n)\}$, find the classifier function h_θ from a family $\{h_\theta | \theta \in \Theta\}$ such that the total *loss*

$$f(\theta, \mathcal{T}) \;:=\; \sum_{i=1}^{n} g(y^i, h_\theta(x^i)) + R(\theta) \tag{12.1}$$

is minimized:

- The function g penalizes a deviation of the *predicted* class label $h_\theta(x^i)$ from the *actual* (known) class label y^i.
- The *regularizer* R can make the problem well-defined if there is not enough input data to uniquely identify θ, and can counter-act the problem of *overfitting:* If the size of the training data set is small compared to the parameter space Θ, there is a danger of fitting the classifier to the specific structure of the training data set, instead of the distribution underlying the training data – in effect, the classifier will work very well for the training data, but will perform poorly on feature vectors that are not in the training data set.

In this chapter we will mostly deal with the two-class case, more precisely, $h_\theta: \mathbb{R}^m \to \{-1, 1\}$; however, most concepts can be generalized to multiple classes. For the two-class case, the set $(h_\theta)^{-1}(\{0\})$ defines the *decision boundary*, i.e., the set that separates the two *class regions* $(h_\theta)^{-1}(\{1\})$ and $(h_\theta)^{-1}(\{-1\})$.

12.2 Linear Classifiers

Primal problem.

A simple but very powerful idea is to look for *linear* classifiers of the form

$$\begin{aligned} &h_\theta: \mathbb{R}^m \to \{-1, 1\}, \\ &h_\theta(x) := \operatorname{sgn}(\langle w, x\rangle + b), \quad \text{where } \theta = (w, b). \end{aligned} \tag{12.2}$$

This corresponds to finding a separating hyperplane $H := \{x \,|\, \langle w, x\rangle + b = 0\}$ that separates the two sets of points. Even if this is possible, there will be ambiguity as to how to exactly choose the hyperplane, so regularization is needed.

We will first look at *maximum margin classifiers*. Here the *margin* – the minimal (signed, in the direction of the desired class label y^i) distance between points x^i in $\mathcal{T}$ and the hyperplane H is *maximized:*

$$\sup_{w,b} \min_{i\in\{1,\dots,n\}} y^i \cdot \left\{ \left\langle \frac{w}{\|w\|_2}, x^i \right\rangle + \frac{b}{\|w\|_2} \right\}$$

We rewrite

$$\begin{aligned} &\sup_{w,b,c} c \quad \text{s.t. } c \leqslant y^i \cdot \left\{ \left\langle \frac{w}{\|w\|_2}, x^i \right\rangle + \frac{b}{\|w\|_2} \right\},\ i \in \{1, \dots, n\} \\ = &\sup_{w,b,c} c \quad \text{s.t. } \|w\|_2\, c \leqslant y^i \cdot \{\langle w, x^i\rangle + b\},\ i \in \{1, \dots, n\}. \end{aligned}$$

We now substitute $c' = c\,\|w\|_2$ and obtain

$$= \sup_{w,b,c'} c'/\|w\|_2 \quad \text{s.t. } c' \leqslant y^i \cdot \{\langle w, x^i\rangle + b\},\ i \in \{1, \dots, n\}.$$

For any solution (w, b, c'), any scalar multiple $(\lambda\, w, \lambda\, b, \lambda\, c')$ with $\lambda > 0$ will also be a solution. Assuming that there exists at least one solution with $c' \geqslant 0$ (i.e., the data sets can in fact be exactly separated), we can thus pick one of the solutions by forcing $c' = 1$, and obtain

$$\sup_{w,b} \frac{1}{\|w\|_2} \quad \text{s.t. } 1 \leqslant y^i \cdot \{\langle w, x^i\rangle + b\},\ i \in \{1, \dots, n\}. \tag{12.3}$$

Finally, maximizing $\|w\|^{-1}$ is equivalent to minimizing $\frac{1}{2}\|w\|^2$, so we obtain

$$\inf_{w,b} \frac{1}{2}\|w\|_2^2 \quad \text{s.t. } 1 \leqslant y^i \cdot \{\langle w, x^i\rangle + b\},\ i \in \{1,\ldots,n\}. \tag{12.4}$$

This is convex problem with quadratic objective, and can be rewritten in SDP form (an SOCP formulation is equally possible but leads to a different dual problem). In terms of the framework in (12.1), the quadratic part $\frac{1}{2}\|w\|_2^2$ constitutes the regularizer R, while the *indicator function* of the constraints is the loss function g – in effect, the loss for a correct classification is always 0, while incorrect classifications have infinite loss and are therefore prohibited.

Dual problem and optimality conditions.

Rewriting the primal problem as

$$\inf_{w,b}\left\{k(w,b) + h\left(M\begin{pmatrix} w \\ b\end{pmatrix} - 1\right)\right\}, \quad k(w,b) = \frac{1}{2}\|w\|_2^2,\ h(z) = \delta_{\geqslant 0}(z), \tag{12.5}$$

$$M = \begin{pmatrix} y^1 \cdot (x^1)^\top & y^1 \\ \vdots & \vdots \\ y^n \cdot (x^n)^\top & y^n \end{pmatrix}.$$

and using (note the conjugate of k with respect to b),

$$k^*(u,c) \;=\; \frac{1}{2}\|u\|_2^2 + \delta_0(c), \quad h^*(v) = \delta_{\leqslant 0}(v),$$

from Ex. 7.13 and Prop. 7.6 we obtain the saddle-point form

$$\begin{aligned} &\inf_{w,b}\sup_z k(w,b) + \left\langle M\begin{pmatrix} w \\ b\end{pmatrix} - 1, z\right\rangle - h^*(z) \\ =\;& \inf_{w,b}\sup_z \frac{1}{2}\|w\|^2 + \left\langle M\begin{pmatrix} w \\ b\end{pmatrix} - 1, z\right\rangle - \delta_{\leqslant 0}(z), \end{aligned} \tag{12.6}$$

and dual problem

$$\sup_z -\sum_{i=1}^n z_i - \delta_{\leqslant 0}(z) - \frac{1}{2}\left\|-\sum_{i=1}^n y^i\, x^i\, z_i\right\|_2^2 - \delta_0\left(\sum_{i=1}^n y^i\, z_i\right).$$

We get the *dual formulation of the linear SVM,*

$$\begin{aligned} \inf_{z\in\mathbb{R}^n}\ & \frac{1}{2}\left\|\sum_{i=1}^n y^i\, x^i\, z_i\right\|_2^2 + e^\top z \\ \text{s.t.}\ & z \leqslant 0, \\ & \sum_{i=1}^n y^i\, z_i = 0. \end{aligned} \tag{12.7}$$

Support vector machines are generally solved in their dual form (we will later see why exactly). As in Ex. 7.15, the primal-dual optimality conditions can be stated in terms of the Lagrangian in (12.6) in complementarity form,

$$0 \in \begin{pmatrix} w \\ 0\end{pmatrix} + M^\top z, \tag{12.8}$$

$$0 \leqslant \left(M\begin{pmatrix} w \\ b\end{pmatrix} - 1\right) \perp z \leqslant 0. \tag{12.9}$$

Again as in Ex. 7.15, as all terms in the last equality cannot be positive due to the preceding two conditions, the complementarity condition reduces to n scalar equalities of the form

$$(y^i \cdot \{\langle w, x^i\rangle + b\} - 1) \cdot z_i \;=\; 0. \tag{12.10}$$

Therefore, any point x^i with $z_i \neq 0$ must satisfy $y^i\,(\langle x^i, w\rangle + b) = 1$, i.e.,

$$z_i \neq 0 \;\Rightarrow\; \left|\left\langle x^i, \frac{w}{\|w\|_2}\right\rangle - \frac{b}{\|w\|_2}\right| = \frac{1}{\|w\|_2},$$

which by (12.3) is the minimum distance between any point x^j and the separating hyperplane. Therefore, the x^i with $z^i \neq 0$ have minimum distance.

Evaluating the linear function.

In order to classify a new incoming sample x using h_θ in (12.2), we need to compute the linear term $\langle w, x\rangle + b$. Fortunately, given a dual solution z, we can easily recover w from the optimality conditions (12.8). In fact,

$$\langle w, x\rangle \;=\; \left\langle -\sum_{i=1}^{n} z^i\, y^i\, x^i, x\right\rangle = -\sum_i z^i\, y^i\, \langle x^i, x\rangle. \tag{12.11}$$

To recover b, we can use (12.10) and plug in any support vector (having $z_i \neq 0$), or take the mean over all support vectors to increase numerical stability.

12.3 The Kernel Trick

The trouble with linear classifiers is that in practice, data can very rarely be separated by a hyperplane, i.e., the decision boundary must be of a more complicated shape. Extending the maximum-margin approach as in the previous section to non-linear decision boundaries is entirely non-trivial, and generally impossible except for special cases. But fortunately it turns out that we actually do not need to – instead, we transform the samples x^i *before* applying a linear classifier:

We transform the original training samples $\mathcal{T} = \{(x^1, y^1), \ldots, (x^n, y^n)\}$ using a nonlinear embedding $\eta\colon \mathcal{F} \to \bar{\mathcal{F}}$, where $\mathcal{F} \subseteq \mathbb{R}^m$ is the original feature space and $\bar{\mathcal{F}} \subseteq \mathbb{R}^{\bar{m}}$ is some higher-dimensional feature space, usually with $\bar{m} \gg m$:

$$\bar{\mathcal{T}} \;:=\; \{(\bar{x}^1, y^1), \ldots, (\bar{x}^n, y^n)\}, \quad \bar{x}^i := \eta(x^i).$$

The SVM approach then gives us a linear classifier function $\bar{h_\theta}(\bar{x}) = \mathrm{sgn}(\langle w, x\rangle - b)$ *in the space* $\bar{\mathcal{F}}$, which induces the *nonlinear* decision function

$$h_\theta(x) \;=\; \mathrm{sgn}(\langle w, \eta(x)\rangle - b)$$

in the original space. For example, for two-dimensional data we could define

$$\eta(x) \;:=\; \big(x_1^2, x_2^2, \sqrt{2}\,x_1\,x_2, \sqrt{2}\,x_1, \sqrt{2}\,x_2, 1\big). \tag{12.12}$$

The decision boundary can then be *any polynomial of degree* 2 *or less:*

$$h_\theta(x) \;=\; \mathrm{sgn}\big(w_1\,x_1^2 + w_2\,x_2^2 + w_3\,\sqrt{2}\,x_1\,x_2 + w_4\,\sqrt{2}\,x_1 + w_5\,\sqrt{2}\,x_2 + w_6 - b\big).$$

Unfortunately, such spaces can become huge very quickly for moderately large m – for polynomials of degree 2 or less, we already need

$$\binom{m}{2} = m\,(m-1)$$

terms. As this defines the size of the optimization problem (12.5) and of the coefficient vector w, it becomes too large very quickly.

The first key insight is the following: *The size of the* z *in dual problem (12.7) does not depend on the size* m *of the feature space* – the size of z is defined by the *number of samples* n instead!

Unfortunately, the objective

$$\inf_{z\in\mathbb{R}^n} \frac{1}{2}\left\|\sum_{i=1}^n y^i\,\eta(x^i)\,z_i\right\|_2^2 + e^\top z$$

still requires to map x^i into the high-dimensional space $\bar{\mathcal{F}}$. We start by rewriting the quadratic part as

$$\begin{aligned}\frac{1}{2}\left\|\sum_{i=1}^n y^i\,\eta(x^i)\,z_i\right\|_2^2 &= \sum_{i,j=1}^n y^i\,y^j\,\langle\eta(x^i),\eta(x^j)\rangle\,z_i\,z_j \\ &= \sum_{i,j=1}^n y^i\,y^j\,\kappa(x^i,x^j)\,z_i\,z_j, \quad \kappa(x,x') := \langle\eta(x),\eta(y)\rangle. \end{aligned} \tag{12.13}$$

This is where the second key insight comes into play: By carefully choosing η, the *kernel* κ can be evaluated *without computing* η: For η as in (12.12),

$$\kappa(x,x') = x_1^2 x_1'^2 + \ldots + 2\,x_2\,x_2' + 1 = (\langle x,x'\rangle + 1)^2$$

does the trick. Once we have solved the dual problem, the inner product $\langle w,\,\eta(x)\rangle$ is, according to (12.11),

$$\begin{aligned}\langle w,\eta(x)\rangle &= -\sum_{i=1}^n z^i\,y^i\,\langle\bar{x}^i,\eta(x)\rangle \\ &= -\sum_{i=1}^n z^i\,y^i\,\kappa(x^i,x). \end{aligned} \tag{12.14}$$

From a solution $z\in\mathbb{R}^n$ of the dual problem we can therefore compute b and evaluate the decision function $h_\theta(x) = \langle w,\eta(x)\rangle - b$ without ever evaluating η! This concept is known as the *kernel trick*.

In order to evaluate h_θ we still need to store the training vectors x^i in order to compute the terms $\kappa(x^i,x)$. But from (12.14) we see that we can discard any vectors with $z^i = 0$, and only need to store vectors x^i where $z^i \neq 0$. Any such vector must be a support vector and thus minimizes the distance to the separating hyperplane, which in practice is a very rare occurence – in order to evaluate (12.14), we will usually have to store only very few of the x^i, rather than the whole training data set.

We have shown that by solving the dual problem, we can find and evaluate an optimal non-linear classifier in the potentially very high-dimensional space $\bar{\mathcal{F}}$ without ever having to evaluate η. In fact, looking at (12.13), we find that we do not even have to *know* η explicitly: We can substitute *any* kernel κ, as long as the dual problem stays convex. A sufficient condition for this is that the matrix $M\in\mathbb{R}^{n\times n}$, $M_{ij} = \kappa(x^i,x^j)$ is symmetric and positive semidefinite for all choices of n and $\{x^1,\ldots,x^n\}$ (the y^i correspond to multiplication with diagonal matrices and do not affect positive semidefiniteness).

Additionally, the extended feature space $\bar{\mathcal{F}}$ can be infinite-dimensional. It can be shown that for any positive definite κ, an embedding η into a so-called *reproducing kernel Hilbert space* can be found that defines κ through η and its inner product (see for example [?]).

Within these bounds, the kernel can be chosen freely to suit the application. Popular choices are:

$$\begin{aligned}\kappa(x,x') &= \langle x,x'\rangle^q && \text{(monomials of degree } q\text{)},\\ \kappa(x,x') &= (\langle x,x'\rangle + 1)^q && \text{(polynomials of degree } q \text{ or less)},\\ \kappa(x,x') &= \exp\left(-\frac{1}{2}\|x-x'\|_2^2/\sigma^2\right) && \text{(Gaussian kernel)},\end{aligned}$$

and many more, see [Bis06, Chapt. 6.3] for an overview.

Chapter 13
Total Variation and Applications

13.1 Functions of Bounded Variation

In this section we will have a more detailed look at one of the most-often used non-smooth regularizers, the total variation (TV) of a function. Details can be found in [AFP00].

As a motivation, for C^1 functions we can formulate the regularizer

$$f(u) \;=\; \int_\Omega \|\nabla u(x)\|_2 \, dx. \tag{13.1}$$

We have already seen in practice that this regularizer can be very useful. To motivate the choice of the term "total variation" for f, consider the one-dimensional case, and assume that u monotononeously non-decreasing between two points a and b, i.e., $u'(x) \geqslant 0$ for all $x \in [a,b]$. Then

$$\int_a^b \|\nabla u(x)\|_2 \, dx \;=\; \int_a^b |u'(x)| \, dx = \int_a^b u'(x) \, dx = u(b) - u(a).$$

The same argument can be made if u is non-increasing, in which case we get $u(a) - u(b)$. This means that if u is monotonous between two points a and b, then $f(u) = |u(a) - u(b)|$, and the behaviour of u between the points is completely irrelevant, as long as it is monotonous. Hence f counts the the differences between extreme points of u, which gives rise to the term "variation" of u.

For general functions that are not necessarily differentiable (and may even be discontinuous), we use the following definition. We generally assume Ω to be an open Lipschitz domain in $\mathbb{R}^n$.

Definition 13.1. *For $u \in L^1(\Omega, \mathbb{R}^m)$, the* total variation *of u is defined as*

$$\mathrm{TV}(u) \;:=\; \sup_{v \in C_c^1(\Omega, \mathbb{R}^{m\times n}), \|v\|_\infty \leqslant 1} \int_\Omega \langle u, \operatorname{Div} v \rangle \, dx, \tag{13.2}$$

where $v(x) = (v^1(x), ..., v^m(x))^\top$, $\operatorname{Div} v = (\operatorname{div} v^1, ..., \operatorname{div} v^m)$, and for $v \in C_c^1(\Omega, \mathbb{R}^{m\times n})$,

$$\|v\|_\infty \;=\; \sup_{x\in\Omega} \|v(x)\|_2.$$

The space of functions of bounded variation *is defined as*

$$\mathrm{BV}(\Omega, \mathbb{R}^m) \;:=\; \{u \in L^1(\Omega, \mathbb{R}^m) | \mathrm{TV}(u) < +\infty\}.$$

This can be understood as follows: If $u \in C^1$, i.e., ∇u exists as a function, then

$$\int_\Omega \|\nabla u(x)\|_2 \, dx \;=\; \int_\Omega \sup_{v(x)\in\mathbb{R}^{m\times n}, \|v\|_2 \leqslant 1} \langle \nabla u, v(x) \rangle \, dx,$$

which is just rewriting the norm using its dual norm. It is possible to show that the supremum and the integral can be swapped and v can be restricted to C_c^1 functions:

$$\int_\Omega \|\nabla u(x)\|_2 \, dx \;=\; \sup_{v \in C_c^1(\Omega, \mathbb{R}^{m\times n}), \|v\|_\infty \leqslant 1} \int_\Omega \langle \nabla u, v \rangle \, dx.$$

As all v have compact support, we can use the divergence theorem applied to $u_i \, v^i$ (which is just partial integration):

$$0 \;=\; \int_\Omega \operatorname{div}(u_i \, v^i) \, dx \Rightarrow \int_\Omega \langle \nabla u_i, v^i \rangle \, dx = -\int_\Omega u_i \operatorname{div} v^i \, dx,$$

and get

$$\int_\Omega \|\nabla u(x)\|_2 \, dx \;=\; \sup_{v \in C_c^1(\Omega, \mathbb{R}^{m\times n}), \|v\|_\infty \leqslant 1} \int_\Omega \langle u, \operatorname{Div} v \rangle \, dx = \operatorname{TV}(u).$$

The is also sometimes referred to as the dual formulation of the total variation. The space BV can also be defined as the space of all $u \in L^1$ so that the gradient of u exists as a finite radon measure in Ω, denoted by Du, but we will not go into these details. The importance in formulation (13.2) as opposed to (13.1) is that it does not require u to be differentiable, or even continuous.

An important property of the total variation is that for characteristic functions of sets, it reduces to boundary length/area of the set:

Proposition 13.2. *Assume $A \subset \Omega$ is a set so that its boundary is C^1 and satisfies $\mathcal{H}^{n-1}(\Omega \cap \partial A) < \infty$. Define*

$$1_A(x) \;:=\; \begin{cases} 1, & x \in A, \\ 0, & x \notin A. \end{cases}$$

Then

$$\operatorname{TV}(1_A) \;=\; \mathcal{H}^{n-1}(\Omega \cap \partial A).$$

Proof. The idea is that

$$\begin{aligned} \operatorname{TV}(1_A) &= \sup_{v \in C_c^1(\Omega, \mathbb{R}^n), \|v\|_\infty \leqslant 1} \int_\Omega 1_A \operatorname{div} v dx \\ &= \sup_{v \in C_c^1(\Omega, \mathbb{R}^n), \|v\|_\infty \leqslant 1} \int_A \operatorname{div} v dx \\ &= \sup_{v \in C_c^1(\Omega, \mathbb{R}^n), \|v\|_\infty \leqslant 1} \int_{\partial A} \langle v, n \rangle \, ds \end{aligned}$$

by Gauss' theorem. This immediately show the lower bound $\operatorname{TV}(1_A) \leqslant \mathcal{H}^{n-1}(\partial A)$. The upper bound is slightly more difficult, intuitively we need $v(x) = n$ on the boundary, but have to show that v can be made sufficiently smooth, for example by constructing as a suitable L^1 function and approximating it using C_c^1 functions, see [AFP00] for general remarks. □

Penalizing boundary length is a very natural way of favouring smooth contours, which makes the total variation a good candidate for the problem of finding an optimal set, such as in segmentation problems.

Theorem 13.3. *(Coarea Formula): If $u \in \mathrm{BV}(\Omega)$, then*

$$\mathrm{TV}(1_{\{x|u(x)>t\}}) < +\infty \text{ for } \mathcal{L}^1\text{-a.e. } t \in \mathbb{R},$$

and

$$\mathrm{TV}(u) \;=\; \int_{-\infty}^{+\infty} \mathrm{TV}(1_{\{u>t\}})\, dt.$$

Proof. See [AFP00, Thm. 3.4]. □

This can often be used to reduce the study of functions of bounded variation to the study of their (super-)levelsets $\{u > t\}$.

Analogous to the definition of BV, we can define higher-order BV spaces by applying this idea to the gradients of $W^{k,1}$ functions as follows. For simplicity we assume the scalar valued case, i.e., $u\colon \Omega \to \mathbb{R}$.

Definition 13.4. *For $\Omega \subseteq \mathbb{R}^d$ and $k \geqslant 1$, we define the space $\mathrm{BV}^k(\Omega)$ as*

$$\mathrm{BV}^k \;:=\; \left\{ u \in W^{k-1,1} \middle| \nabla^{k-1} u \in \mathrm{BV}\Big(\Omega, \mathbb{R}^{d^{k-1}}\Big) \right\}$$

and the higher-order total variation as

$$\begin{aligned} \mathrm{TV}^k(u) \;&=\; \sup_{v \in C_c^k(\Omega, \mathrm{Sym}^k(\mathbb{R}^d)), \|v\|_\infty \leqslant 1} \int_\Omega u \,\mathrm{div}^k\, v\, dx, \\ &=\; \mathrm{TV}(\nabla^{k-1} u) \end{aligned} \tag{13.3}$$

where $\mathrm{Sym}^k(\mathbb{R}^d)$ is the space of symmetric tensors v of order k with arguments in $\mathbb{R}^d$:

$$v \;=\; (v_{i_1,\dots,i_k})_{i_1,\dots,i_k=1,\dots,d},$$

and $v(z^1,\dots,z^k) = v\big(z^{\pi(1)},\dots,z^{\pi(k)}\big)$ for all permutations π of $\{1,\dots,k\}$. The k-divergence is defined as

$$\mathrm{div}^k\, v \;=\; \sum_{\bar{i}=(i_1,\dots,i_k)\in\{1,\dots,d\}^k} \partial_{x_{i_1}} \cdots \partial_{x_{i_k}} v_{i_1,\dots,i_k}.$$

For $k = 1$ this reduces to the usual total variation, as then $v \in C_c^\infty(\Omega, \mathbb{R}^d)$, the symmetry condition disappears, and

$$\mathrm{div}^1\, v \;=\; \sum_{i_1 \in \{1,\dots,d\}} \partial_{x_{i_1}} v_{i_1} = \mathrm{div}\, v.$$

Note that all functions in BV^k must have a $(k-1)$-th derivatve in L^1, which means a certain regularity (see, e.g., the Sobolev embedding theorem).

In practice, TV^k can again be implemented by discretizing the k-th order derivatives using a matrix $G^k \in \mathbb{R}^{nd^k \times n}$ and using a special norm defined as the sum of 2-norms over the rows of $G^k u$:

$$\mathrm{TV}^k(u) \;=\; \|G^k u\|_* := \sum_{i=1}^{n} \|(G^k u)_i\|_2,$$

or the equivalent dual formulation

$$\mathrm{TV}^k(u) \;=\; \sup \left\{ \langle u, (G^k)^\top v \rangle \middle| v \in \mathbb{R}^{d^k \times n}, v_{\cdot j} \in \mathrm{Sym}^k(\mathbb{R}^d), \|v_{\cdot i}\|_2 \leqslant 1 \right\}.$$

Usually finite-difference discretizations of G^k will lead to a symmetric tensor (e.g., for $d = 2$ we would expect the discretization to only produce symmetric Hessians). If this is the case, then $(G^k u)_i \in \mathrm{Sym}^k(\mathbb{R}^d)$, which is a linear subspace of $\mathbb{R}$, and for an arbitrary, possibly non-symmetric tensor $v \in \mathbb{R}^{d^k}$, we can replace the inner product $\langle G^k u, v\rangle$ by $\left\langle G^k u, \Pi_{\mathrm{Sym}^k(\mathbb{R}^d)}(v)\right\rangle$. Since $\left\|\Pi_{\mathrm{Sym}^k(\mathbb{R}^d)}(v)\right\|_2 \leqslant \|v\|_2$ (the projection onto linear subspaces can only decrease the norm), this shows that we can drop the symmetry constraint on v, and get

$$\mathrm{TV}^k(u) \;=\; \sup\left\{\langle u, (G^k)^\top v\rangle \middle| v \in \mathbb{R}^{d^k \times n}, \|v_{i\cdot}\|_2 \leqslant 1\right\}, \tag{13.4}$$

which is very similar to the formulation for the "plain" TV regularizer, and can in most cases be easily substituted.

13.2 Infimal Convolution and TGV

We consider the following reformulation of the basic ROF problem:

$$\inf_u \left\{\frac{1}{2}\|u-g\|_2^2 + \lambda\,\mathrm{TV}(u)\right\} \;=\; \inf_{u,w,u+w=g}\left\{\frac{1}{2}\|w\|_2^2 + \lambda\,\mathrm{TV}(u)\right\}. \tag{13.5}$$

Solving (13.5) can be seen as *decomposing* the given data g into two components u and w, where w (the "noise") is small with respect to the L^2 norm – essentially assuming that the noise follows a Gaussian distribution – and u has a small total variation, i.e., it "looks like" a natural image as measured by the regularizer.

This decomposition can be easily extended to more terms – in particular, as the total variation favours piecewise constant images in practice, small affine changes in the image may end up in the noise variable w instead, which is not ideal. On the other hand, TV^2 favours affine regions, but cannot handle discontinuities. Adding TV^2 to the decomposition allows discontinuities as well as affine structures:

$$\inf_{u,v,w,u+v+w=g}\left\{\frac{1}{2}\,\|w\|_2^2 + \lambda\,\mathrm{TV}(u) + \mu\,\mathrm{TV}^2(v)\right\}.$$

We can then separately examine the piecewise constant ("cartoon") part u, the piecewise affine part v, and the noise w, or use $u+v$ to get a denoised version of g.

Definition 13.5. *(infimal convolution) For functions $f_1, \ldots, f_n \colon X \to \bar{\mathbb{R}}$ for any set X, we define the inf-convolution $(f_1 \square \cdots \square f_n) \colon X \to \bar{\mathbb{R}}$ as*

$$(f_1 \square \cdots \square f_k)(x) = \inf_{z^1,\ldots,z^k, z^1+\ldots+z^k=x} (f_1(z^1) + \ldots + f_k(z^k)).$$

In this primal formulation, infimal convolutions are difficult to deal with, as even evaluating them involves solving an optimization problem. Fortunately, they have a very concise dual representation:

Proposition 13.6. *(conjugates of infimal convolutions) Assume $f_1, \ldots, f_k \colon \mathbb{R}^n \to \bar{\mathbb{R}}$ are proper, lsc, convex. Then*

$$(f_1 \square \cdots \square f_k) \;=\; (f_1^* + \ldots + f_k^*)^*.$$

Proof. See example sheets. □

Example 13.7. If $f, g: \mathbb{R}^n \to \mathbb{R}$ are proper, lsc, convex, and positively homogeneous – as is the case if f and g are norms or seminorms – then by Prop. 6.17 we know that $f = (\delta_C)^*$ and $g = (\delta_D)^*$ for some closed convex non-empty sets $C, D \subseteq \mathbb{R}^n$. Then

$$(f \square g)^* \;=\; \delta_C + \delta_D = \delta_{C \cap D},$$

and

$$(f \square g) \;=\; (\delta_{C \cap D})^*.$$

For example, for discretizations $G \in \mathbb{R}^{2n \times n}$ and $H \in \mathbb{R}^{4n \times n}$ of the gradient and Hessian, we can write

$$\begin{aligned} \mathrm{TV}(u) &= \sup_{v \in \mathbb{R}^{2\times n} \| v^i \|_2 \leqslant 1 \forall i} \langle u, -G^\top v \rangle = \sup \left\{ \langle u, v' \rangle \middle| v' = G^\top v, \left\| v'^i \right\|_2 \leqslant 1 \forall i \right\} \\ &= \delta_C^*, \quad C := \{ v' \in \mathbb{R}^n | \exists v \in \mathbb{R}^{2\times n} : v' = G^\top v, \| v^i \|_2 \leqslant 1 \forall i \}, \\ \mathrm{TV}^2(u) &= \delta_D^*, \quad D := \{ w' \in \mathbb{R}^n | \exists w \in \mathbb{R}^{4\times n} : w' = H^\top w, \| w^i \|_2 \leqslant 1 \forall i \}. \end{aligned} \tag{13.6}$$

Then the *combined* regularizer

$$h = (\mathrm{TV} \square \mathrm{TV}^2)$$

has the set representation

$$h = \delta_E^*, \quad E := \{ z \in \mathbb{R}^n | \exists v \in \mathbb{R}^{2\times n}, w \in \mathbb{R}^{4\times n} : z = G^\top v = H^\top w, \| v^i \|_2 \leqslant 1, \| w^i \|_2 \leqslant 1 \forall i \}.$$

Solving a quadratic optimization problem with h as a regularizer amounts to a backward step on h:

$$\arg\min_u \left(\frac{1}{2} \| u - g \|_2^2 + \lambda h(u) \right) \;=\; B_{\lambda h}(g).$$

From the example sheets we know that $B_{\lambda f}(x) = x - \lambda B_{\lambda^{-1} f^*}(x/\lambda)$, thus

$$\begin{aligned} B_{\lambda h}(g) &= g - \lambda B_{\lambda^{-1} \delta_E}(g/\lambda) \\ &= g - \lambda \Pi_E(g/\lambda). \end{aligned}$$

This can be interpreted as splitting

$$\begin{aligned} &g = u + y, \\ &u = \mathrm{argmin} \ldots = g - \lambda \Pi_E(g/\lambda), \\ &y = g - u = \lambda \Pi_E(g/\lambda). \end{aligned}$$

This has a very natural interpretation: For a given image g, we compute the *noise* by (nonlinearly) projecting onto the set E – effectively, E defines the noise that we would like to allow – and the original image by subtracting the noise.

The infimal convolution of two regularizers corresponds to *intersecting their sets of "allowed" noise*.

In practice it was found that the combined $(\mathrm{TV} \square \mathrm{TV}^2)$ regularizer can often be improved upon by using the *Total Generalized Variation* (TGV) [SST11, BKP10] instead:

Definition 13.8. *(Total Generalized Variation) For $u \in L^1$, $k \geqslant 1$, and $\alpha = (\alpha_0, \ldots, \alpha_{k-1}) > 0$, the* Total Generalized Variation *is defined as*

$$\mathrm{TGV}_\alpha^k(u) \;=\; \sup \left\{ \int_\Omega u \, \mathrm{div}^k v \middle| v \in C_c^k(\Omega, \mathrm{Sym}^k(\mathbb{R}^d)), \| \mathrm{div}^j v \|_\infty \leqslant \alpha_j, 0 \leqslant j \leqslant k-1 \right\}$$

with the convention that $\operatorname{div}^0 v = v$.

Comparing this to the definition of TV^k in (13.3), the difference is in the additional constraints on the lower-order divergences on v.

Consider the case $n=2$ and $\alpha=(1,1)$, and assume that the discretization of the gradient can be decomposed as $G^2 = G_2 G_1$, where $G_1 \in \mathbb{R}^{2n\times n}$ discretizes the first-order derivatives of u, and $G_2 \in \mathbb{R}^{4n\times 2n}$ discretizes the first-order derivatives of a vector field (in particular ∇u) with suitable boundary conditions. The suitable discretization for div^1 is then $-G_2^\top$, which leads to (again assuming symmetric discretization of the gradient)

$$\begin{aligned}\mathrm{TGV}^2(u) &= \sup\left\{\langle u, (G^2)^\top v\rangle \middle| v\in C_1, \left(-G_2^\top\right) v \in C_2\right\}\\ C_1 &= \{v\in\mathbb{R}^{4n\times n} |\, \|v_{\cdot i}\|_2 \leqslant 1\ \forall i\},\\ C_2 &= \{z\in\mathbb{R}^{2n\times n} |\, \|z_{\cdot i}\|_2 \leqslant 1\ \forall i\}.\end{aligned}$$

We can rewrite this as follows:

$$\begin{aligned}\mathrm{TGV}^2(u) &= \sup_v \left\{\langle u, (G_2 G_1)^\top v\rangle - \delta_{C_1}(v) - \delta_{C_2}\left(-G_2^\top v\right)\right\}\\ &= \sup_v \left\{\left\langle -G_1 u, -G_2^\top v\right\rangle - \delta_{C_1}(v) - \delta_{C_2}\left(-G_2^\top v\right)\right\}\\ &= \sup_{v,w} \left\{-\langle G_1 u, w\rangle - \delta_{C_1}(v) - \delta_{C_2}(w) - \delta_{\{0\}}\left(G_2^\top v + w\right)\right\}\\ &= \sup_{v,w} \left\{-\langle G_1 u, w\rangle - \delta_{C_1}(v) - \delta_{C_2\times\{0\}}\left(\begin{pmatrix} 0 & I\\ G_2^\top & I\end{pmatrix}\begin{pmatrix} v\\ w\end{pmatrix}\right)\right\}.\end{aligned}$$

This has standard duality form, and we obtain

$$\begin{aligned}\mathrm{TGV}^2(u) &= \inf_{y^1,y^2} \delta^*_{C_1}(-G_2 y^2) + \delta_{\{0\}}(y^1 + y^2 - G_1 u) + \delta^*_{C_2}(y^1)\\ &= \inf_{y^1,y^2,y^1+y^2=G_1 u} \|y^1\|_{C_2} + \|G_2 y^2\|_{C_1}. \end{aligned} \tag{13.7}$$

This is also known as the "differentiation cascade" formulation of TGV^2, and can be extended to TGV^k. Comparing this to the standard infimal convolution,

$$\begin{aligned}(\mathrm{TV}\square\mathrm{TV}^2)(u) &= \inf_{z^1,z^2,z^1+z^2=u} (\mathrm{TV}(y) + \mathrm{TV}_2(u))\\ &= \inf_{z^1,z^2,z^1+z^2=u} (\|G_1 u\|_{C_2} + \|G_2 G_1 u\|_{C_1}),\end{aligned} \tag{13.8}$$

we see that TGV^2 is again a certain kind of infimal convolution, but instead of splitting u into multiple components, the *gradient* of u is split. The cascading formulation is also a convenient way of converting the *dual* energy in (13.8) into a form that can be handled by most solvers.

13.3 Meyers G-Norm

Meyer's G-norm [Mey01, AC04] was introduced as a regularizer adapted to *textured* regions, and is defined as

$$\|u\|_G = \inf\{\|v\|_\infty | \operatorname{div} v = u, v\in L^\infty(\mathbb{R}^d)\}.$$

After discretization, we again get

$$\|u\|_G = \inf_v \{\delta^*_C(v) + \delta_{-G^\top v = u}\}.$$

In saddle-point form:

$$\begin{aligned}
\|u\|_G &= \inf_v \{\delta_C^*(v) + \delta_{-G^\top v = u}\} \\
&= \inf_v \sup_w \{\delta_C^*(v) + \langle w, G^\top v + u\rangle\} \\
&= \sup_w \inf_v \{\delta_C^*(v) + \langle w, G^\top v + u\rangle\} \\
&= \sup_w -\sup_v \{\langle v, G\,w\rangle - \delta_C^*(v) - \langle w, u\rangle\} \\
&= \sup_w \{\langle w, u\rangle - \delta_C(G\,w)\} \\
&= \sup_w \{\langle w, u\rangle |\, \|G\,w\|_\infty \leqslant 1\} \\
&= \sup_w \{\langle w, u\rangle |\, \mathrm{TV}(w) \leqslant 1\}.
\end{aligned}$$

The G-norm is the *dual* norm to the total variation, i.e., the norm associated with the *unit ball* with respect to TV. In particular,

$$\|\cdot\|_G = \delta^*_{\{u|\mathrm{TV}(u)\leqslant 1\}} = \delta^*_{\mathcal{B}_{\mathrm{TV}}}(u), \quad \mathcal{B}_{\mathrm{TV}} := \{u|\mathrm{TV}(u) \leqslant 1\}.$$

In the same way, we get (compare (13.6))

$\sup\{\langle u, -G^\top v\rangle |\, \|v\|_\infty \leqslant 1\}$

$$\begin{aligned}
\mathrm{TV}(u) &= \sup\{\langle u, -G^\top v\rangle |\, \|v\|_\infty \leqslant 1\}, \quad \mathcal{B}_G := \{u|\, \|u\|_G \leqslant 1\}. \\
&= \sup\{\langle u, w\rangle | \exists v\colon \|v\|_\infty \leqslant 1, w = -G^\top v\} \\
&= \sup\{\langle u, w\rangle | \|w\|_G \leqslant 1\} \\
&= \delta^*_{\mathcal{B}_G}(u), \quad \mathcal{B}_G := \{u|\, \|u\|_G \leqslant 1\}.
\end{aligned}$$

Consider the problem of finding

$$\arg\min_u \frac{1}{2}\|u - f\|_2^2 \text{ s.t. } u \in \lambda\,\mathcal{B}_G,$$

i.e., separating the "texture" component of f with the assumption that the texture "level" as measured in the G-norm is at most λ. The solution is just the projection $\Pi_{\lambda G}(f)$, which we can already relate to the ROF problem: From the example sheets we know that $B_f = I - B_{f^*}$, and obviously if $\mathrm{TV}(u) = \delta^*_{\mathcal{B}_G}$ then $\lambda\,\mathrm{TV}(u) = \delta_{\lambda\mathcal{B}_G}$, so

$$\Pi_{\lambda\mathcal{B}_G} = f - B_{\lambda\,\mathrm{TV}(u)} = f - \arg\min_u \left\{\frac{1}{2}\,\|u - f\|_2^2 + \lambda\,\mathrm{TV}(u)\right\}.$$

This explains why the G-norm is a good candidate for regularizing data containing texture: Solving L^2-$\|\cdot\|_G$ *removes* the part of an image that can be explained by a "structure" component with low total variation.

In the same way, we can rewrite *TV-constrained* problems as

$$\arg\min_u \frac{1}{2}\,\|u - f\|_2^2 \text{ s.t. } \mathrm{TV}(u) \leqslant \lambda \;=\; f - \arg\min_u \left\{\frac{1}{2}\,\|u - f\|_2^2 + \lambda\,\|u\|_G\right\}.$$

13.4 Non-local regularization

Total variation regularization is inherently local, in the sense that for each point a small neighbourhood is considered to contain enough information to determine the regularization cost at that point. While this is very convenient for analysis, on real-world images this assumption may not always be justified: Many images contain textured regions that are highly oscillatory, but in a very regular sense – such as striped or checkedboard patterns. Such regions would be highly penalized by a total variation regularizer, even if they contain no noise.

In order to better cope with such situations, *nonlocal* regularizers have been proposed [BCM05, GO08]. The idea is to measure the regularity of the image at a given point not by the dissimilary of its gray value to its immediate neighbours, but to other points in the image that are in a similar location of the repeating pattern. These points can be potentially far away, which gives rise to the name, non-local regularization.

In a discrete setting, this can be achieved as follows.

Definition 13.9. *Denote* $\Omega = \{1, ..., n\}$. *For* $u \in \mathbb{R}^n$ *and* $x, y \in \mathbb{R}^n$ *and a nonnegative weighting function* $w: \Omega^2 \to \mathbb{R}_{\geqslant 0}$, *we define the* non-local partial derivative $\partial_y u(x)$ *as*

$$\partial_y u(x) \;=\; (u(y) - u(x))w(x, y).$$

The non-local gradient *of* u *at* x *for the weighting function* w *is the* n*-vector*

$$\nabla_w \colon \mathbb{R}^n \to \mathbb{R}^{n\times n},$$
$$\nabla_w u(x) = (\partial_y u(x))_{x\in\Omega}.$$

The difference between ∇_w and a usual discrete gradient is that $\nabla_w u(x)$ is an n-vector of all "partial derivatives" to all other points in the image, instead of a (usually) 2-vector of the partial derivatives in x- and y-direction.

We can equally define a corresponding non-local divergence, which sums up all partial derivatives:

$$\operatorname{div}_w v(x) \;=\; \sum_{y\in\Omega} (v(x, y) - v(y, x))w(x, y).$$

With the usual Euclidean inner products in $\mathbb{R}^n$ and $\mathbb{R}^{n\times n}$, it can be seen that we have a discrete "divergence theorem",

$$\langle -\operatorname{div}_w v, u\rangle \;=\; \langle v, \nabla_w u\rangle.$$

We can now define regularizers based on the non-local gradient.

$$J(u) \;=\; \sum_{x\in\Omega} g(\nabla_w u). \tag{13.9}$$

Most gradient-based regularizers that involve norms can be immediately extended to this setting. There is some freedom, for example in the TV case we could define the *gradient-* or *difference-based* versions

$$\mathrm{TV}^g_{\mathrm{NL}}(u) = \sum_{x\in\Omega} \|\nabla_w u\|_2 \quad \text{or} \quad \mathrm{TV}^d_{\mathrm{NL}} = \sum_{x\in\Omega} \|\nabla_w u\|_1.$$

For the ordinary total variation regularization, the 2-norm approach gives better results as it better respects the isotropy of the total variation, but non-local regularization is inherently anisotropic in any case due to the weighting function w, so there is no obvious "better" candidate.

A relevant difference in practice is that while $\mathrm{TV}^d_{\mathrm{NL}}$ only contains terms of the form $|(u(y) - u(x))\ w(x,\ y)|$ and is separable otherwise, $\mathrm{TV}^g_{\mathrm{NL}}$ is a sum of 2-norms of n-dimensional vectors and is therefore much less separable, which makes it less attractive from an optimization viewpoint. Nevertheless, both regularizers are convex and can be reformulated in SOCP ($\mathrm{TV}^g_{\mathrm{NL}}$) or LP ($\mathrm{TV}^d_{\mathrm{NL}}$) form.

In the same way, it is possible to generalize many other gradien-based regularizers such as the p-norm, the G-norm to nonlocal gradients.

The defining feature is the choice of the weighting function w. First we note that (13.9) includes most usual discretizations of the total variation – setting $w(x, y) = \frac{1}{h}$ (where h is the grid spacing) if x and y are neighbours, and $w(x, y) = 0$ otherwise, leaves the nonlocal gradient essentially as the local gradient, extended with zeros.

The intuition is that $w(x, y)$ should be large if the neighbourhoods of x and y are similar, as measured by a *patch distance*. A classical choice is to consider the patch distance

$$d_u(x, y) \;=\; \int_\Omega K_\sigma(p)\,(u(y+t) - u(x+t))^2\,dt,$$

which is just the ℓ^2 distance weighted by a Gaussian K_σ centered at zero with variance σ^2. While it would be possible to set, e.g., $w(x, y) = 1/(\varepsilon + d_u(x, y))$, this leaves us with a very large optimization problem: For $\mathrm{TV}^d_{\mathrm{NL}}$, the objective would contain at least n^2 non-smooth terms. Even for moderately-sized images with $n \approx 100000$ this is clearly not feasible. Therefore the weights are usually pruned before. The original choice is to define the set

$$A(x) \;:=\; \arg\min_A \left\{ \sum_{y\in\Omega} d_u(x, y) \,\middle|\, A \subseteq S(x), |A| = k \right\}$$

for a given *search neighborhood* $S(x)$, which consists of the k points around x with smallest distance. The weights are then simply set as

$$w(x, y) \;=\; \begin{cases} 1, & y \in A(x) \text{ or } x \in A(y), \\ 0, & \text{otherwise.} \end{cases}$$

The reason for introducing the search neighbourhood $S(x)$ is that computing $d_u(x, y)$ for all pairs of points x, y can already be too expensive.

Generally non-local regularizers tend to work very well in practice, but they also require more parameters to be chosen. In particular, the search window weights K_σ should be large enough for the noise error to average out when comparing patches, but small enough to get a good resolution. The main obstacle when implementing such methods is computational, i.e., how to quickly compute the patch distances and how to prune the weights in a way that keeps the optimization problem tractable.

Chapter 14
Relaxation

Many real-world problems are inherently non-convex. A typical example is the problem of segmenting an image into two regions: For given image data g, find a *set* $C \subseteq \Omega$ that best describes the foreground in the sense that it fits to the given data, but also adheres to some prior knowledge about the typical *shape* of the foreground.

A typical energy is the *Chan-Vese* model [CV01],

$$f_{\mathrm{CV}}(C, c_1, c_2) := \left\{ \int_C (g - c_1)^2 \, dx + \int_{\Omega \setminus C} (g - c_2)^2 \, dx + \lambda \, \mathcal{H}^{d-1}(C) \right\},$$

where $\mathcal{H}^{d-1}(C)$ is the perimeter (i.e., length or area) of the boundary ∂C, and minimization is performed over (C, c_1, c_2). The constants c_1 and c_2 describe the typical value of g inside (c_1) and outside (c_2) of C, i.e., the problem consists in identifying the foreground and background region together with *model parameters* (c_1, c_2) for each region.

The Chan-Vese model is in fact a special case of the *Mumford-Shah* model [MS89], one of the best-studied – but still not fully understood – models in image processing:

$$f_{\mathrm{MS}}(K, u) \;=\; \int_\Omega (g - u)^2 \, dx + \mu \int_{\Omega \setminus K} \|\nabla u\|_2^2 dx + \nu \, \mathcal{H}^{d-1}(K),$$

where $K \subseteq \Omega$ is closed and u is differentiable outside of K, i.e. $u \in C^1(\Omega \setminus K)$. Essentially, this corresponds to the $L^2 - L^2$ denoising approach (?), with the exception that u is allowed to be discontinuous on a "boundary" set K that should be "small" as measured by the Hausdorff term.

If one requires that u is piecewise constant, and furthermore assumes at most two values c_1 and c_2, the term involving $\|\nabla u\|^2$ vanishes, and Mumford-Shah energy simplifies to the Chan-Vese energy.

An efficient way of reformulating f_{CV} in a way that can be handled numerically is to represent the set C using its indicator function 1_C, and require $1_C \in \mathrm{BV}(\Omega)$. As the total variation of an indicator function is just the perimeter of the set, we obtain

$$f_{\mathrm{CV}}(C, c_1, c_2) \;=\; \int_\Omega 1_C \, (g - c_1)^2 + (1 - 1_C) \, (g - c_2)^2 \, dx + \lambda \, \mathrm{TV}(1_C).$$

We can now introduce a function $u \colon \Omega \to \{0, 1\}, u \in \mathrm{BV}(\Omega)$, and minimize

$$\begin{aligned} & \inf_{u:\Omega \to \{0,1\}, u \in \mathrm{BV}(\Omega), c_1, c_2} \int_\Omega u \, (g - c_1)^2 + (1 - u) \, (g - c_2)^2 \, dx + \lambda \, \mathrm{TV}(u) \\ = \; & \inf_{u:\Omega \to \{0,1\}, u \in \mathrm{BV}(\Omega), c_1, c_2} \int_\Omega u \, ((g - c_1)^2 - (g - c_2)^2) + (g - c_2)^2 \, dx + \lambda \, \mathrm{TV}(u). \end{aligned}$$

There are two difficulties to overcome:

1. The optimization problem is of *combinatorial* nature, i.e., the constraint set consists of a discrete set of points. This makes it difficult to apply optimization methods that rely on taking small steps towards a minimizer based on derivative information.

2. Even if the constraint set was convex, the objective is not jointly convex in (u, c_1, c_2).

The second problem cannot be easily overcome. However, if we fix either u or c_1, c_2, then the problem is convex in the other variable. In the remainder of this chapter we will therefore assume that c_1 and c_2 are known and fixed. A possible way to obtain a (at least local) solution for unknown c_1, c_2 is to alternate between minimization with respect to u and c_1, c_2.

The first point can be addressed by a *relaxation* approach: Instead of our original, nonconvex constraint set

$$\{u \in \mathrm{BV}(\Omega) | u(x) \in \{0,1\} \text{ a.e.}\}$$

we pass on to the *convex hull* of the constraint set:

$$\{u \in \mathrm{BV}(\Omega) | u(x) \in [0,1] \text{ a.e.}\}.$$

We obtain the energy

$$\inf_{u:\Omega\to[0,1], u\in\mathrm{BV}(\Omega)} \int_\Omega u\left((g-c_1)^2 - (g-c_2)^2\right) dx + (g-c_2)^2 + \lambda\,\mathrm{TV}(u).$$

This is in fact a convex problem. Moreover, we can set $s(x) := (g-c_1)^2 - (g-c_2)^2$ and ignore the constant term $(g-c_2)^2$ as it does not change the set of minimizers, and obtain

$$\inf_{u:\Omega\to[0,1], u\in\mathrm{BV}(\Omega)} f(u), \quad f(u) := \langle u, s\rangle_{L^1} + \lambda\,\mathrm{TV}(u). \tag{14.1}$$

This is even more appealing once one realizes that this problem is still convex if we replace the terms $(g-c_i)^2$ by *arbitrary* (integrable) functions h_1 and h_2, and set $s := h_1 - h_2$. We have found a variation of the problem that is always convex, no matter how complicated the data term is.

There is a slightly ambiguous understanding of what is meant by a *convex relaxation* approach – while it is frequently used to loosely describe the process of constructing *some* convex problem from a non-convex problem, in other publications it is understood in the strict sense of replacing a non-convex function f by its convex hull $\mathrm{con}\, f$.

The cost of removing the non-convexity is that minimizers of f are not necessarily indicator functions anymore. We can say the following:

Proposition 14.1. *Assume c_1, c_2 are fixed. If u is a minimizer of f, and $u(x) \in \{0,1\}$ a.e. (in particular, $u = 1_C$ for some set $C \subseteq \Omega$), then C is a minimizer of $f_{\mathrm{CV}}(\cdot, c_1, c_2)$.*

Proof. Assume that $u = 1_C$ for some set C,

$$f(1_C) \;=\; f_{\mathrm{CV}}(C, c_1, c_2).$$

Since C is a minimizer of f, for every set C' we have

$$f_{\mathrm{CV}}(C', c_1, c_2) = f(1_{C'}) \;\geqslant\; f(1_C) = f_{\mathrm{CV}}(C, c_1, c_2),$$

i.e., C must be a minimizer of f_{CV}. □

Unfortunately, the minimizer u of f may not be a characteristic function.

Definition 14.2. *Denote $\mathcal{C} := \mathrm{BV}(\Omega, [0,1]) := \{u \in \mathrm{BV}(\Omega) | u(x) \in [0,1] \text{ a.e.}\}$. For $u \in \mathcal{C}$, and $\alpha \in [0,1]$, define*

$$\bar{u}_\alpha \;:=\; 1_{\{u>a\}}, \quad 1_{\{u>\alpha\}}(x) = \begin{cases} 1, & u(x) > \alpha, \\ 0, & u(x) \leqslant \alpha. \end{cases}$$

We say that a functional $f: \mathcal{C} \to \mathbb{R}$ *satisfies the* generalized coarea condition *iff*

$$f(u) \;=\; \int_0^1 f(\bar{u}_\alpha)\, d\alpha \quad \forall u \in \mathcal{C}.$$

Proposition 14.3. *Assume that* $s \in L^\infty(\Omega)$ *and* Ω *is bounded. Then the function* f *in (14.1) satisfies the generalized coarea condition.*

Proof. For $f(u) = \mathrm{TV}(u)$, the condition is exactly the coarea formula (Thm. 13.3). As the condition is additive, we only have to show it for the linear part:

$$\begin{aligned}\int_\Omega s(x)\, u(x)\, dx \;&=\; \int_\Omega s(x) \int_0^1 1_{\{u(x)>\alpha\}} d\alpha \\ &=\; \int_0^1 \int_\Omega s(x)\, 1_{\{u(x)>\alpha\}} d\alpha.\end{aligned}$$

Swapping the integral using Fubini's theorem requires to show that $|s(x)\ 1_{\{u(x)>\alpha\}}|$ is bounded, which holds since $\int_\Omega |s(x)\, 1_{\{u(x)>\alpha\}}| \leqslant \|s\|_\infty |\Omega| < \infty$ due to $s \in L^\infty(\Omega)$. □

The most important result in this section is the following, generalized from [CEN06]:

Theorem 14.4. *(Thresholding Theorem) Assume that* $f: \mathrm{BV}(\Omega, [0,1]) \to \mathbb{R}$ *satisfies the generalized coarea condition, and* u^* *satisfies*

$$u^* \;\in\; \arg \min_{u \in \mathrm{BV}(\Omega,[0,1])} f(u).$$

Then for almost every $\alpha \in [0,1]$*, the thresholded function* $\bar{u}^*_\alpha = 1_{\{u^*>\alpha\}}$ *satisfies*

$$\bar{u}^*_\alpha \;\in\; \arg \min_{u \in \mathrm{BV}(\Omega,\{0,1\})} f(u).$$

Proof. We define the set of α violating the assertion, $\mathcal{S} := \{\alpha \in [0,1] | f(\bar{u}^*_\alpha) \neq f(u^*)\}$. Since $\bar{u}^*_\alpha \in \mathcal{C}_{\{0,1\}}$ and $\mathcal{C}_{\{0,1\}} \subseteq \mathcal{C}$, we have for any minimizer $u^*_{\{0,1\}}$ of f over $\mathcal{C}_{\{0,1\}}$,

$$f(u^*) \leqslant f(u^*_{\{0,1\}}) \leqslant f(\bar{u}^*_\alpha), \tag{14.2}$$

thus $\mathcal{S} = \{\alpha \in [0,1] | f(u^*) < f(\bar{u}^*_\alpha)\}$. Moreover, if $\alpha \notin \mathcal{S}$, then $f(u^*) = f(u^*_{\{0,1\}}) = f(\bar{u}^*_\alpha)$ by (14.2). Therefore, in order to show the theorem it suffices to show that $\mathcal{S}$ is an $\mathcal{L}^1$-zero set.

Assume the contrary holds, i.e. $\mathcal{L}^1(\mathcal{S}) > 0$. Then there must be $\varepsilon > 0$ such that

$$\mathcal{S}_\varepsilon \;:=\; \{\alpha \in [0,1] | f(u^*) \leqslant f(u^*_\alpha) - \varepsilon\} \tag{14.3}$$

has also nonzero measure, since otherwise $\mathcal{S}$ would be the countable union of zero measure sets, $\mathcal{S} = \bigcup_{i \in \mathbb{N}} \mathcal{S}_{1/n}$, and would consequently have zero measure as well. Then

$$\begin{aligned}f(u^*) \quad &= \quad \int_{[0,1]\setminus\mathcal{S}_\varepsilon} f(u^*) d\alpha + \int_{\mathcal{S}_\varepsilon} f(u^*) d\alpha && (14.4)\\ &\overset{u^* \text{ optimal}}{\leqslant} \quad \int_{[0,1]\setminus\mathcal{S}_\varepsilon} f(\bar{u}^*_\alpha) d\alpha + \int_{\mathcal{S}_\varepsilon} f(u^*) d\alpha && (14.5)\\ &\overset{\text{definition of } \mathcal{S}_\varepsilon}{\leqslant} \quad \int_{[0,1]\setminus\mathcal{S}_\varepsilon} f(\bar{u}^*_\alpha) d\alpha + \int_{\mathcal{S}_\varepsilon} (f(\bar{u}^*_\alpha) - \varepsilon) d\alpha && (14.6)\\ &\overset{\text{linearity}}{=} \quad \int_0^1 f(\bar{u}^*_\alpha) d\alpha - \varepsilon\, \mathcal{L}^1(\mathcal{S}_\varepsilon). && (14.7)\end{aligned}$$

But we assumed $\mathcal{L}^1(\mathcal{S}_\varepsilon) > 0$, therefore

$$f(u^*) \;<\; \int_0^1 f(\bar{u}^*_\alpha) d\alpha. \tag{14.8}$$

This is a contradiction to (?), therefore $\mathcal{L}^1(\mathcal{S}) = 0$ and the assertion follows. □

At the heart of the proof is the generalized coarea condition (?). It has the following intuitive interpretation:

1. The function u may be written in the form of a "generalized convex combination" of (an infinite number of) extreme points $E_u := \{\bar{u}_\alpha | \alpha \in [0, 1]\}$ of the constraint set, i.e. the unit ball in $\mathrm{BV}(\Omega)$. As shown in [Fle57] based on a result by Choquet [Cho56], and noted in [Str83, p.127], extreme points of this constraint set are (multiples of, but in this case equal to) indicator functions.
2. The extreme points $(\bar{u}_\alpha)$ and coefficients in this convex combination can be explicitly found. In fact, the coefficients are all equal to $1/|[0,1]| = 1$, i.e. u is the barycenter of the points in E_u.
3. For any convex f, the inequality
$$\int_0^1 f(\bar{u}_\alpha)\, d\alpha \;\geqslant\; f(u) \tag{14.9}$$
always holds. The coarea formula (?) is therefore equivalent to the reverse inequality.

In fact, the original proof of the coarea formula [FR60] relies on showing (14.9) and using the fact that (?) holds for piecewise linear u [FR60, (1.5c)]. Approximating an arbitrary $u \in \mathrm{BV}(\Omega)$ by a sequence of piecewise linear functions, this result is then transported to the general case.

Remark 14.5. It is important to understand that Thm. 14.4 in its current form only holds *before* discretization: Assume the problem is discretized according to

$$\min_{u \in \mathbb{R}^n} f(u),\; f(u) := \langle u, s \rangle + \lambda \, \|Gu\|_*, \quad \text{s.t. } u_i \in [0,1], \tag{14.10}$$

where $\|Gu\|_*$ is a suitable norm implementing the total variation. From Thm. 14.4 one might naturally assume that if u^* solves (?), then for almost every $\alpha \in [0,1]$, the thresholded minimizer $\bar{u}^*_\alpha$ solves the combinatorial problem

$$\min_{u \in \mathbb{R}^n} f(u) \quad \text{s.t. } u_i \in \{0,1\}. \tag{14.11}$$

This is generally not the case! The reason is that in order to transfer Thm. 14.4 to the discretized problem, the *discretized* objective functions needs to satisfy the generalized coarea condition. While the linear term does not pose a problem, the usual "isotropic" discretizations of the total variation, such as $\sum_i \|G_i u\|_2$, with G_i implementing forward, central, or finite differences on a staggered grid, do *not* have this property.

In finite dimensions, the integral

$$\int_0^1 f(\bar{u}_\alpha)\, d\alpha$$

is also known as the *Lovasz extension* of an energy $f\colon \{0,1\}^n \to \mathbb{R}$. Energies with a *convex* Lovasz extension are called "submodular", and play a central role in combinatorial optimization, as a large problem class that can be solved efficiently.

Submodular energies that consist of a sum of terms involving at most two variables u_i can be solved by computing a minimal cut through a graph, which can be achieved in polynomial time by solving the dual problem with specialized "maximum-flow" methods [BVZ01, Ber98].

The energy (14.10) can be made to satisfy a generalized coarea property, for example by choosing a forward difference discretization for G, and setting

$$\|Gu\|_* \;=\; \sum_i \|G_i\, u\|_1 = \sum_{i,j} \left(|u_{(i+1),j} - u_{i,j}| + |u_{i,(j+1)} - u_{i,j}|\right)$$

(try to verify this as an interesting exercise). Unfortunately this discretization is not isotropic anymore – i.e., it does not converge to the actual total variation as the grid spacing goes to zero –, and shows a bias towards horizontal and vertical edges.

This appears to limit the usefulness of the thresholding theorem for practical purposes. However, in practice it turns out that very often thresholded minimizers of (14.10) for *non-submodular* (but isotropic) discretization generate better – in terms of visual appearance, not in terms of the energy – segmentations than solving the combinatorial problem. Also, in many cases it can be advantageous to drop the requirement that the discretized solution should only assume the values $\{0, 1\}$ and allow intermediate values in $[0, 1]$ – later processing steps such as computing statistics on the geometry (center, area, boundary length, variance, higher-order moments, etc.) and extraction of a boundary contour can actually benefit from the increased accuracy that comes from allowing non-integer labels.

Bibliography

[AC04] Jean-Fran⟨c|o⟩is Aujol and Antonin Chambolle. Dual norms and image decomposition models. Tech. Rep. 5130, INRIA, 2004.

[AFP00] L. Ambrosio, N. Fusco and D. Pallara. *Functions of Bounded Variation and Free Discontinuity Problems*. Clarendon Press, 2000.

[BCM05] A. Buades, B. Coll and J.-M. Morel. A non-local algorithm for image denoising. In *Comp. Vis. Patt. Recogn.* 2005.

[Ber98] D. P. Bertsekas. *Network Optimization: Continuous and Discrete Models*. Athena Scientific, 1998.

[Bis06] C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[BKP10] K. Bredies, K. Kunisch and T. Pock. Total generalized variation. *J. Imaging Sci.*, 3(3):294–526, 2010.

[BVZ01] Y. Boykov, O. Veksler and R. Zabih. Fast approximate energy minimization via graph cuts. *Patt. Anal. Mach. Intell.*, 23(11):1222–1239, 2001.

[CEN06] T. F. Chan, S. Esedolu and M. Nikolova. Algorithms for finding global minimizers of image segmentation and denoising models. *J. Appl. Math.*, 66(5):1632–1648, 2006.

[Cho56] G. Choquet. Existence des représentations intégrales au moyen des points extrémaux dans les cônes convexes. *C. R. Acad. Sci.*, 243:669–702, 1956.

[CV01] T. F. Chan and L. A. Vese. Active contours without edges. *IEEE Trans. Image Proc.*, 10(2):266–277, 2001.

[Eck89] J. Eckstein. *Splitting Methods for Monotone Operators with Application to Parallel Optimization*. PhD thesis, MIT, 1989.

[ET99] I. Ekeland and R. Témam. *Convex analysis and variational problems*. SIAM, 1999.

[Fle57] W. H. Fleming. Functions with generalized gradient and generalized surfaces. *Annali di Mathematica Pura ed Applicata*, 44(1), 1957.

[FR60] W. H. Fleming and R. Rishel. An integral formula for total gradient variation. *Archiv der Mathematik*, 11(1):218–222, 1960.

[GO08] G. Gilboa and S. Osher. Nonlocal operators with applications to image processing. *Multiscale Mod. Simul.*, 7:1005–1028, 2008.

[Mey01] Y. Meyer. *Oscillating Patterns in Image Processing and Nonlinear Evolution Equations*, volume 22 of *Univ. Lect. Series*. AMS, 2001.

[MS89] D. Mumford and J. Shah. Optimal approximations by piecewise smooth functions and associated variational problems. *Comm. Pure Appl. Math.*, 42:577–685, 1989.

[Roc70] R. T. Rockafellar. *Convex Analysis*. Princeton Univ. Press, 1970.

[RW04] R. T. Rockafellar and R. J.-B. Wets. *Variational Analysis*. Springer, 2nd edition, 2004.

[SST11] S. Setzer, G. Steidl and T. Teuber. Infimal convolution regularizations with discrete l1-type functionals. *Comm. Math. Sci.*, 9(3):797–872, 2011.

[Str83] G. Strang. Maximal flow through a domain. *Math. Prog.*, 26:123–143, 1983.