Convex Analysis

Dirk Lorenz

Contents

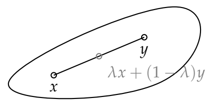
Please submit erros to d.lorenz@tu-braunschweig.de

1 Convex sets

We introduce the basic notion of convexity of sets. We will develop everything in the euclidean space $\mathbb{R}^d$, but most of what we will be doing, will also work in a general real and separable Hilbert space X, i.e. a real vector space that is equipped with an inner product and has a orthonormal basis.

Definition 1.1. A set $C \subset \mathbb{R}^d$ is *convex*, if for all $x, y \in C$ and $\lambda \in [0,1]$ it holds that$\lambda x + (1-\lambda)y \in C$. The term $\lambda x + (1-\lambda)y$ is called the *convex combination.*

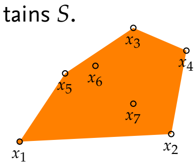
More general: For $x_1, \ldots, x_n \in \mathbb{R}^d$ and $\lambda_1, \ldots, \lambda_n \geq 0$ with $\sum_{i=1}^n \lambda_i = 1$ we call $x = \sum_{i=1}^n \lambda_i x_i$ a *convex combination.*



It holds: A set is convex if and only if it contains all convex combinations of its points.

Definition 1.2. The *convex hull* $\mathrm{conv}(S)$ of a subset $S \subset \mathbb{R}^d$ is the set of all convex combinations of points in S.

Another way to say this is: The convex hull is the smallest convex set that contains S.



$S = \{x_1, \ldots, x_7\}$

Example 1.3. For two convex sets $C_1, C_2 \subset \mathbb{R}^d$ it holds that the *Minkowski sum*

$$C_1 + C_2 := \{x = x_1 + x_2 \mid x_1 \in C_1,\ x_2 \in C_2\}$$

is again convex. $\triangle$

It's fairly straightforward to prove the following (and one should do so as an exercise):

Proposition 1.4. *The following sets are convex:*

1. $\alpha C = \{\alpha x \mid x \in C\}$, *for convex* $C \in \mathbb{R}^d$ *and* $\alpha \in \mathbb{R}$,
2. $AC \subset \mathbb{R}^m$, *for convex* $C \in \mathbb{R}^d$ *and a matrix* $A \in \mathbb{R}^{m \times d}$,
3. $C_1 \times C_2 \subset \mathbb{R}^{m+d}$ *for convex* $C_1 \in \mathbb{R}^m$ *and convex* $C_2 \in \mathbb{R}^d$,
4. $\bigcap_{i \in I} C_i$ *for convex* C_i *and any index set* I,
5. *the closure* $\overline{C}$ *and the interior* C° *for convex* C.

We will denote the closure of C also by $\mathrm{cl}(C)$ and the interior also by $\mathrm{int}(C)$.

Proof. Let us prove the second point: If x, y are in AC then there are u, v, such that $x = Au$ and $y = Av$. For any $\lambda \in [0,1]$ it holds that $\lambda u + (1-\lambda)v \in C$, and hence $\lambda x + (1-\lambda)y = \lambda Au + (1-\lambda)Av = A(\lambda u + (1-\lambda)v)$ is in AC. $\square$

The proofs of the remaining assertions are straightforward and left as exercise.

Definition 1.5. A set $S \subset \mathbb{R}^d$ is called *affine* if for all $x, y \in S$ and $\lambda \in \mathbb{R}$ it holds that$\lambda x + (1-\lambda)y \in S$.

Note that $\lambda x + (1-\lambda)y = y + \lambda(x-y)$, i.e. the point $\lambda x + (1-\lambda)y$ is reached by starting from y and going λ times the vector from y to x in the direction of x.

Of course, linear subspaces are also affine spaces and for every non-empty affine set, there is a unique subspace L and some vector a such that $S = a + L$. It's also clear that affine sets are convex.

A point x is an *affine combination* of $x_1, \ldots, x_n$ if there exists λ_i with $\sum_{i=1}^n \lambda_i = 1$ and $x = \sum_{i=1}^n \lambda_i x_i$.

Definition 1.6. The *affine hull* $\operatorname{aff}(S)$ of a set S is the set of all affine combinations of elements of S.

Alternatively, we could also define the affine hull of S as the smallest affine set which contains S.

An affine space enherits a topology from the surrounding space $\mathbb{R}^d$ and hence, we have closure and interior with respect this topology for subsets of affine spaces. This gives rise to the following notion:

Definition 1.7. The *relative interior* of some $S \subset \mathbb{R}^d$, denoted by $\operatorname{ri}(S)$, is the interior of S relative to the affine set $\operatorname{aff}(S)$, i.e.

$$\operatorname{ri}(C) = \{x \in C \mid \exists \epsilon > 0 : B_\epsilon(x) \cap \operatorname{aff}(C) \subset C\}.$$

The set $\overline{C} \setminus \operatorname{ri}(C)$ is called *relative boundary* of C.

You also see $\operatorname{relint}(S)$ for the relative interior. Note that there is no notion for relative closure as the closure is always within the affine hull.

Proposition 1.8.

1. *For convex $C \neq \emptyset$ it holds that $\operatorname{ri}(C)$ is also convex and it holds that $\operatorname{aff}(\operatorname{ri}(C)) = \operatorname{aff}(\overline{C})$.*
2. *For convex C, invertible $A \in \mathbb{R}^{d\times d}$ and any $b \in \mathbb{R}^d$ it holds that $A\operatorname{ri}(C) + b = \operatorname{ri}(AC + b)$ and for all $A \in \mathbb{R}^{m\times d}$ it holds that $A(\operatorname{ri}(C)) = \operatorname{ri}(AC)$.*
3. *For convex C we have $x \in \operatorname{ri}(C)$ if and only if for every $y \in \operatorname{aff}(C)$ there exists $\epsilon > 0$ such that $x \pm \epsilon(y - x) \in C$,*
4. *For convex C_1, C_2 it holds $\overline{C_1} = \overline{C_2}$ if and only if $\operatorname{ri}(C_1) = \operatorname{ri}(C_2)$,*
5. *For convex C_1, C_2 it holds $\operatorname{ri}(C_1 + C_2) = \operatorname{ri}(C_1) + \operatorname{ri}(C_2)$.*

We don't give a proof of this proposition, but note that point 2. is helpful to prove other statements about the relative interior: If $\operatorname{aff}(C)$ is an m-dimensional affine space, we can, without loss of generality, assume that $\operatorname{aff}(C)$ lies in the subspace $V = \{x \mid x_{m+1} = \cdots = x_n = 0\}$ and since this is a just a copy of $\mathbb{R}^m$, we can assume that C is fulldimensional, i.e. $\operatorname{aff}(C)$ is the full space.

Note that even if $C_1 \subset C_2$, it is generally not true, that $\operatorname{ri}(C_1)$ is contained in $\operatorname{ri}(C_2)$! This may be seen, for example, with C_2 being a (closed) square in $\mathbb{R}^2$ and C_1 being one of its sides.

Definition 1.9. A set of $n+1$ points $x_0, \dots, x_n \in \mathbb{R}^d$ are called *affinely independent* if the affine hull $\operatorname{aff}(\{x_0, \dots, x_n\})$ is an n-dimensional affine space.

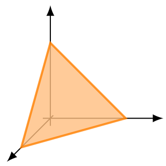
Since $\operatorname{aff}\{x_0, \dots, x_n\} = V + x_0$ where V is the subspace spanned by the vectors $x_1 - x_0, \dots, x_n - x_0$, we see that the vectors $x_0, \dots, x_n$ are affinely independent if and only if the vectors $x_1 - x_0, \dots, x_n - x_0$ are linearly independent.

Proposition 1.10. *If $x_0, \dots, x_n$ are affinely independent, each $x \in \operatorname{aff}\{x_0, \dots, x_n\}$ can be represented uniquely as an affine combination $x = \sum_{0=1}^n \lambda_i x_i$ and these λ_i are called barycentric coordinates of x with respect to the points $x_0, \dots, x_n$.*

Proof. If $M = \operatorname{aff}\{x_0, \dots, x_n\} = x_0 + V$ with$V = \operatorname{span}\{x_1 - x_0, \dots, x_n - x_0\}$, we can express each $y \in V$ uniquely as $y = \sum_{i=1}^n \lambda_i(x_i - x_0)$ and since each $x \in M$ is of the form $x_0 + y$ with $y \in V$, we express each $x \in M$ uniquely as $x = \sum_{i=1}^n \lambda_i(x_i - x_0) + x_0$, i.e. as an affine combination $x = \sum_{i=0}^n \lambda_i x_i$ with $\sum_{i=0}^n \lambda_i = 1$. □

If we have affinely independent points $x_0, \dots, x_n$ their convex hull is called a ***simplex***. Of special importance is the ***probability simplex*** (also called ***standard simplex***) which is the convex hull of the standard basis vectors e_i, i.e.

$$\Delta_d := \operatorname{conv}(e_1, \dots, e_d)$$



2 Hyperplanes and cones

For each $x_0, p \in \mathbb{R}^d$ with $p \neq 0$ the *hyperplane* through x_0 with normal vector p can be written with $\alpha = \langle x_0, p\rangle$ as

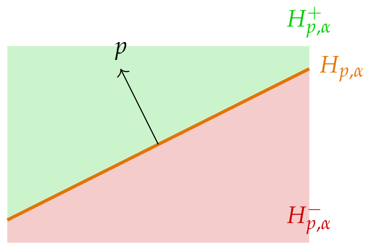
$$H_{p,\alpha} := \{x \in \mathbb{R}^d \mid \langle p, x\rangle = \alpha\} = \{x \mid \langle p, x - x_0\rangle = 0\}.$$

Hyperplanes are affine sets and the orthogonal projection of some x onto $H_{p,\alpha}$ is

$$\hat{x} = x - \tfrac{\langle p,x\rangle - \alpha}{\|p\|_2^2} p.$$

Moreover, there are the associated half-spaces

$$H^+_{p,\alpha} := \{x \in \mathbb{R}^d \mid \langle p, x\rangle \geq \alpha\}, \quad H^-_{p,\alpha} := \{x \in \mathbb{R}^d \mid \langle p, x\rangle \leq \alpha\}.$$



We will show (with the help of separation theorems) that a closed and convex set equals the intersection of all half-spaces that contain C.

A set C that is the intersection of finitely many half-spaces is called *polyhedral set*, in this case we can write

$$C = \{x \mid Ax \leq b, Bx = d\}$$

for some $A \in \mathbb{R}^{n\times d}$, $b \in \mathbb{R}^n$, $B \in \mathbb{R}^{m\times d}$, $c \in \mathbb{R}^m$.

Related to hyperplanes are *affine functionals* which are of the form

$$f(x) = \langle p, x\rangle + \alpha$$

for some $p \in \mathbb{R}^d$, $\alpha \in \mathbb{R}$. A hyperplane in $\mathbb{R}^{d+1}$ is of the form

$$H_{p,\alpha} = \{(x', x_{d+1}) \in \mathbb{R}^{d+1} \mid \langle p', x'\rangle + p_{d+1}x_{d+1} = \alpha\}.$$

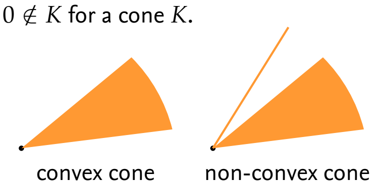
A "vertical" hyperplane is one with p_{d+1} but for the other (i.e. for $p_{d+1} \neq 0$ we have

$$x_{d+1} = \langle -\tfrac{p'}{p_{d+1}}, x'\rangle + \tfrac{\alpha}{p_{d+1}},$$

i.e., the hyperplane $H_{p,\alpha} \subset \mathbb{R}^{d+1}$ is the graph of the affine function $f : \mathbb{R}^d \to \mathbb{R}$, $f(x') = \langle -\tfrac{p'}{p_{d+1}}, x'\rangle + \tfrac{\alpha}{p_{d+1}}$.

Definition 2.1. A set $K \subset \mathbb{R}^d$ is called a *cone*, is $x \in K$ and $\lambda \geq 0$ implies $\lambda x \in K$.

Sometimes cones are defined with just $\lambda > 0$. In this case, one may have $0 \notin K$ for a cone K.



Proposition 2.2. *A cone K is convex if and only if $K + K \subset K$.*

Proof. "$\Rightarrow$": If K is convex and $x, y \in K$ we have $(x + y)/2 \in K$ and hence $x + y \in K$.

"$\Leftarrow$": Let $K + K \subset K$ hold and let $x, y \in K$, since K is a cone, we have $\lambda x, (1 - \lambda)x \in K$ for any $\lambda \geq 0$ and hence, $\lambda x + (1 - \lambda)y \in K$ by $K + K \subset K$. □

Example 2.3. 1. Every linear subspace is a convex cone, more precisely, a cone K is a linear subspace if and only if $K = -K$.

2. The halfspaces $H^+_{p,0} = \{x \mid \langle p, x\rangle \geq 0\}$ are convex cones.

3. An important convex cone is the non-negative orthant $K = \mathbb{R}^d_{\geq 0}\{x \mid x_1, \ldots, x_n \geq 0\}$.

We also write $x \geq 0$ is every component is non-negative.

4. The set $\{Ay \mid y \geq 0\}$ is a convex cone for $A \in \mathbb{R}^{d\times m}$ and cones of this form are called *finitely generated*.

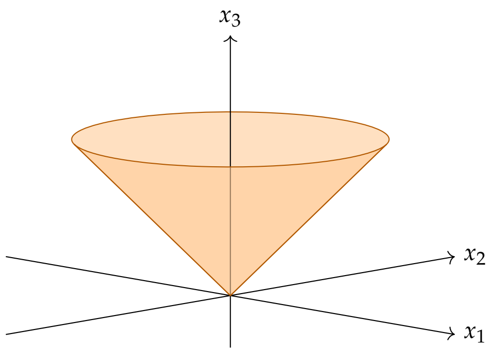
5. The set $\{x \mid A^T x \leq 0\}$ with $A \in \mathbb{R}^{d\times m}$ is a convex cone and cones of this type are called *polyhedral cones*.

6. The set

$$\mathbb{L}^{d+1} = \{(x', x_{d+1}) \in \mathbb{R}^{d+1} \mid \|x'\|_2 \leq x_{d+1}\}$$

is a convex cone called *second order cone* (or Lorentz-cone or, due to its shape, ice-cream cone).

The Lorentz cone $\mathbb{L}^3$:



7. The set of symmetric positive semi-definite matrices $\mathrm{Sym}^+_d = \{M \in \mathbb{R}^{d\times d} \mid M \succcurlyeq 0\}$ is a convex cone. For $d = 2$ symmetric matrices are of the form

$$M = \begin{bmatrix} x_1 & x_2 \\ x_2 & x_3 \end{bmatrix}$$

and such a matrix is positive semi-definite if $x_1 \geq 0$ and $x_1x_3 \geq x_2^2$. Using the change of variables $x = x_1 - x_3$, $y = 2x_2$, $z = x_1 + x_3$ one can see that M is positiv semi-definite if and only if $(x, y, z) \in \mathbb{L}^3$.

$\triangle$

Definition 2.4. For a non-empty set S one defines the *polar cone* by

$$S^* := \{p \in \mathbb{R}^d \mid \forall x \in S : \langle p, x\rangle \leq 0\} = \bigcap_{x\in S} H^-_{x,0}.$$

Consequently, S^* is always closed and convex. Since for all $x \in S$ and $p \in S^*$ we always have

$$\tfrac{\langle p,x\rangle}{\|p\|_2\|x\|_2} \leq 0,$$

and the left hand side defines the cosine of the angle between x and p, we see that this angle is always larger or equal to $\pi/2$.

One can see that $(K^*)^* = \overline{\mathrm{conv}\, K}$ for any set K and hence, for closed convex cones K it holds that $(K^*)^* = K$.

Example 2.5. 1. The polar cone of $K = \{0\}$ is $K^* = \mathbb{R}^d$.

2. If K is a linear subspace, one has $K^* = K^\perp$.

3. For the non-negative orthant $\mathbb{R}^d_{\geq 0}$ the polar cone is $(\mathbb{R}^d_{\geq 0})^* = \mathbb{R}^d_{\leq 0}$, i.e. the non-positive orthant.
4. For some $p \in \mathbb{R}^d$ and $K = H^-_{p,0} = \{x \mid \langle p, x\rangle \leq 0\}$ it holds that $K^* = \{\lambda p \mid \lambda \geq 0\}$.
5. The polar cone of the finitely generated cone $K = \{Ay \mid y \geq 0\}$ for some $A \in \mathbb{R}^{d\times m}$ is the polyhedral cone $K^* = \{p \mid A^T p \leq 0\}$.

△

Definition 2.6. For some set $S \subset \mathbb{R}^d$ and $x_0 \in S$ we define the *tangetial cone* of S in x_0 by

$$T_S(x_0) := \{v \in \mathbb{R}^d \mid \exists x_n \in S,\ x_n \to x_0,\ \lambda_n \searrow 0 : v = \lim_{n\to\infty}(x_n - x_0)/\lambda_n\}$$

and the *normalized directions* in $T_S(x_0)$ are the limits

$$\frac{x_n - x_0}{\|x_n - x_0\|_2} \to \frac{v}{\|v\|_2}.$$

If C is convex, one has

$$T_C(x_0) = \overline{\{t(x - x_0) \mid x \in C, t \geq 0\}}$$

which shows that $T_C(x_0)$ is closed.

Definition 2.7. For a convex C and $x_0 \in C$ we define the *normal cone* of C at x_0 as

$$N_C(x_0) := \{p \mid \forall x \in C : \langle p, x - x_0\rangle \leq 0\}.$$

By definition, the normal cone is always closed.

Theorem 2.8. *For a convex set C and $x_0 \in C$ it holds that*

$$(T_C(x_0))^* = N_C(x_0).$$

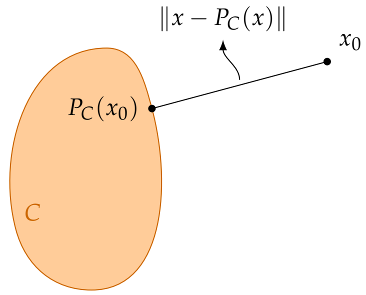
Proof. Let $p \in T_C(x_0)^*$. Then it holds that $\langle p, v\rangle \leq 0$ for all $v \in T_C(x_0)$, i.e. $\langle p, x - x_0\rangle \leq 0$.

Conversely, lez $p \in N_C(x_0)$. We have to show that $\langle p, v\rangle \leq 0$ for any $v \in T_C(x_0)$. Since we can write $v = \lim_{n\to\infty}(x_n - x_0)/\lambda_n$ with $x_n \to x_0$, $x_n \in C$ and $\lambda_n \searrow 0$, we have $\langle p, x_n - x_0\rangle \leq 0$ and hence $\langle p, v\rangle \leq 0$ as well. □

The following theorem would better fit in the next section, however, we still have some space here, so we state and prove it now:

Theorem 2.9 (Projection theorem)**.** *Let $C \subset \mathbb{R}^d$ be a nonempty, closed, convex set. Then, for any $x_0 \in \mathbb{R}^d$, there exists a unique $\hat{x}_0 \in C$, called orthogonal projection of x_0 onto C, and denoted by $P_C x_0 := \hat{x}_0$, such that*

$$\|x_0 - \hat{x}_0\|_2 = \inf_{x\in C}\|x_0 - x\|_2$$



and this element fulfills

$$\forall x \in C: \ \langle x_0 - \hat{x}_0, x - \hat{x}_0\rangle \leq 0. \tag{1}$$

Conversely, if $y \in C$ fulfills the variational inequality

$$\forall x \in C: \ \langle x_0 - y, x - y\rangle \leq 0, \tag{2}$$

then $y = P_C x_0$.

Proof. The function $f(x) = \|x - x_0\|_2$ is continuous and since C is closed, f takes its infimum in C (in fact, in the compact set $C \cap B_r(x_0)$ for some large r). By Weierstraß' theorem, the infimum is attained. This shows existence of a projection.

Now let $\hat{x}_0$ be an orthogonal projection of x_0 onto C. By convexity of C, we have $\hat{x}_0 + \lambda(x - \hat{x}_0) \in C$ for every $x \in C$ and $\lambda \in [0,1]$. Since f is minimal at $\hat{x}_0$, the same holds for $x \mapsto \|x - x_0\|_2^2$ which means

$$\begin{aligned} 0 &\leq \|\hat{x}_0 + \lambda(x - \hat{x}_0)\|_2^2 - \|\hat{x}_0 - x_0\|_2^2 \\ &= \lambda^2\|x - \hat{x}_0\|_2^2 - 2\lambda\langle\hat{x}_0 - x_0, \hat{x}_0 - x\rangle. \end{aligned}$$

For $\lambda > 0$ we can rearrange to

$$\langle\hat{x}_0 - x_0, \hat{x}_0 - x\rangle \leq \tfrac{\lambda}{2}\|x - \hat{x}_0\|_2^2$$

and with $\lambda \to 0$ we have shown that (1) holds.

Conversely, assume that the variational inequality (2) holds for some y. Then, by Cauchy-Schwarz, it holds for all $x \in C$ that

$$\begin{aligned} 0 &\geq \langle y - x_0, y - x\rangle = \langle y - x_0, y - x_0 + x_0 - x\rangle \\ &= \|y - x_0\|_2^2 + \langle y - x_0, x_0 - x\rangle \\ &\geq \|y - x_0\|_2^2 - \|y - x_0\|_2\|x - x_0\|_2. \end{aligned}$$

Dividing by $\|y - x_0\|_2$ we obtain $\|y - x_0\|_2 \leq \|x - x_0\|_2$ for all $x \in C$ and this say that y is the orthogonal projection of x_0 onto C.

Finally, we show uniqueness: Assume that $\hat{x}_0$ and $\tilde{x}_0$ are both orthogonal projection of x_0 onto C. Since $\hat{x}_0, \tilde{x}_0 \in C$ we can plug them into their variational inequalities and get

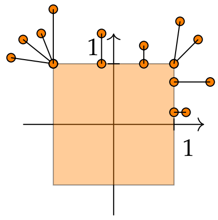
$$\langle x_0 - \hat{x}_0, \tilde{x}_0 - \hat{x}_0\rangle \leq 0, \qquad \langle x_0 - \tilde{x}_0, \hat{x}_0 - \tilde{x}_0\rangle \leq 0.$$

Adding both inequalities we get $\|\hat{x}_0 - \tilde{x}_0\|_2^2 \leq 0$ which means $\hat{x}_0 = \tilde{x}_0$. □

Example 2.10. We project onto balls in the p-norms for $p = 1, 2, \infty$:

1. First consider $p = \infty$ and the respective norm ball of radius $\lambda > 0$ around 0 is

$$B_\lambda^\infty(0) := \{x \mid \max(|x_1|, \ldots, |x_d|) \leq \lambda\}.$$

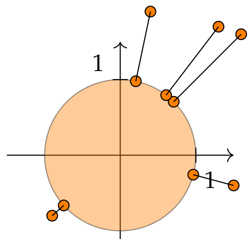


The projection of some x onto this ball is minimizing $f(x-y) = \|x-y\|_2$ over all $y \in B^\infty_\lambda(0)$, i.e. over all y with $|y_i| \leq \lambda$. This can be done componentwise and leads to

$$(P_{B^\infty_\lambda(0)}x)_i = \begin{cases} x_i, & \text{if } |x_i| \leq \lambda \\ \lambda \operatorname{sign}(x_i), & \text{if } |x_i| > \lambda \end{cases}$$

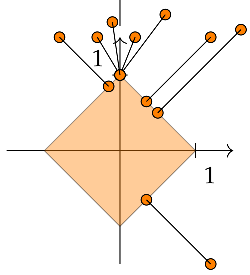
which can be written consisely by $P_{B^\infty_\lambda(0)}x = \min(\max(x, -\lambda), \lambda)$ where the minimum and maximum applied componentwise.

2. For $p = 2$ we simply need to shrink x if it is outside of the ball, i.e.



$$P_{B^2_\lambda(0)}x = \begin{cases} x, & \text{if } \|x\|_2 \leq \lambda \\ \lambda \frac{x}{\|x\|_2}, & \text{if } \|x\|_2 > \lambda. \end{cases} = \max(1, \tfrac{\lambda}{\|x\|_2})x$$

3. The case $p = 1$ is more complicated and there is no explicit formula. However, one can show the following: We define the *soft-shrinkage* (or soft-thresholding) function



$$S_\lambda(x) = \max(|x| - \lambda, 0) \operatorname{sign}(x)$$

and denote by π a permutation of $\{1, \dots, d\}$ which sorts the entries of x in decreasing order, i.e. $|x_{\pi(1)}| \geq |x_{\pi(2)}| \geq \dots \geq |x_{\pi(d)}| \geq 0$. Then, if m is the largest index such that $|x_{\pi(m)}| > 0$ and $\frac{|x_{\pi(1)}| + \dots + |x_{\pi(m)}| - \lambda}{m} \leq |x_{\pi(m)}|$, one has

$$P_{B^1_\lambda(0)}x = \begin{cases} x, & \text{if } \|x\|_1 \leq \lambda \\ S_\mu(x), & \text{if } \|x\|_1 > \lambda. \end{cases}$$

△

3 Projection and separation

Theorem 3.1. *Let $K \subset \mathbb{R}^d$ be a non-empty, closed and convex cone. Then every element in x_0 can be uniquely decomposed as*

$$x_0 = P_K x_0 + P_{K^*} x_0$$

and it holds that $P_K x_0 \perp P_{K^*} x_0$.

Proof. By the Projection Theorem (Theorem 2.9) one has for every $x \in K$

$$\langle x_0 - P_K x_0, x - P_K x_0 \rangle \leq 0. \tag{*}$$

For $x = 0$ we have $\langle x_0 - P_K x_0, P_K x_0 \rangle \geq 0$ and for $x = 2P_K x_0 \in K$ one has $\langle x_0 - P_K x_0, P_K x_0 \rangle \leq 0$ which implies that we have

$$\langle x_0 - P_K x_0, P_K x_0 \rangle = 0. \tag{**}$$

Thus, by (*) we have for all $x \in K$

$$\langle x_0 - P_K x_0, x \rangle \leq 0$$

and this means that $x_0 - P_K x_0 \in K^*$ by the definition of the polar cone. Moreover, since for all $x \in K^*$

$$\begin{aligned}\langle x_0 - (x_0 - P_K x_0), x - (x_0 - P_k x_0) \rangle &= \langle P_K x_0, x - x_0 + P_K x_0 \rangle \\ &= \langle P_K x_0, x \rangle + \langle P_K x_0, P_K x_0 - x_0 \rangle \leq 0\end{aligned}$$

we have (again by the Projection Theorem) that $x_0 - P_K x_0 = P_{K^*} x_0$. The orthogonality follows from (**). □

Since the polar cone of a subspace V is the orthogonal complement $V^\perp$ we obtain:

Corollary. *Let V be a non-empty subspace of $\mathbb{R}^d$. Then the orthogonal projection of x_0 onto V is characterized by*

$$\forall x \in V: \langle x_0 - P_V x_0, x \rangle = 0$$

and, moreover, $x_0 = P_V x_0 + P_{V^\perp} x_0$.

Now we come to the notion of separation:

Definition 3.2. Let C_1, C_2 be two sets. We say that a hyperplane $H_{p,\alpha}$

1. *separates* C_1 and C_2 if for all $x_1 \in C$ and $x_2 \in C_2$ it holds that

$$\langle p, x_1 \rangle \leq \alpha \leq \langle p, x_2 \rangle.$$

In terms of halfspaces: $C_1 \subset H^-_{p,\alpha}$ and $C_2 \in H^+_{p,\alpha}$.

2. *strictly separates* C_1 and C_2 if for all $x_1 \in C$ and $x_2 \in C_2$ it holds that

$$\langle p, x_1 \rangle < \alpha < \langle p, x_2 \rangle.$$

In terms of halfspaces: $C_1 \subset \operatorname{int}(H^-_{p,\alpha})$ and $C_2 \subset \operatorname{int}(H^+_{p,\alpha})$.

3. *properly separates* C_1 and C_2 if it separates the sets and there exist $x_i \in C_i$, $i = 1, 2$ such that

$$\langle p, x_1 \rangle < \langle p, x_2 \rangle.$$

Theorem 3.3. *If C is a non-empty, closed and convex set and $x_0 \notin C$, then we can strictly separate C from $\{x_0\}$, i.e. there exists p and $\epsilon > 0$ such that*

$$\sup_{x \in C} \langle p, x\rangle \leq \langle p, x_0\rangle - \epsilon.$$

Proof. By the Projection Theorem (Theorem 2.9) we have for all $x \in C$ that

$$\langle P_C x_0 - x_0, P_C x_0 - x\rangle \leq 0,$$

from which we deduce by adding and subtracting x_0 in the right argument that

$$\|P_C x_0 - x_0\|_2^2 \leq \langle x_0 - P_C x_0, x_0 - x\rangle$$

for all $x \in C$. But since $x_0 \notin C$, we have $\|P_C x_0 - x_0\|_2^2 \geq \epsilon > 0$ so we can take $p = x_0 - P_C x_0$. $\square$

Corollary 3.4. *For a closed and convex set C ist holds that C equals the intersection of all halfspaces that include C.*

The case of $C = \emptyset$ is clear. That C is a subset of said intersection is clear. And if $x \notin C$ we can find a hyperplane that separates C from x and hence, there is a corresponding halfspace that contains C, but not x and hence, x is not in said intersection.

Theorem 3.5 (Strong separation)**.** *Let C_1, C_2 be non-empty, convex and disjoint and let C_1 be closed and C_2 be compact. Then there exists $p \in \mathbb{R}^d$ and $\alpha \in \mathbb{R}$ such that $H_{p,\alpha}$ strictly separates C_1 and C_2, i.e. for all $x_1 \in C_1$, $x_2 \in C_2$ it holds that*

$$\langle p, x_1\rangle < \alpha < \langle p, x_2\rangle.$$

Proof. We consider the Minkowski sum $C := C_1 + (-C_2) = \{x_1 - x_2 \mid x_1 \in C_1, x_2 \in C_2\}$ which is also non-empty, closed and convex. Also $0 \notin C$ since C_1 and C_2 are disjoint. Hence, we can consider $\hat{x} = P_C 0 \in C$. Since this is a point in $C = C_1 - C_2$ we can write it as $\hat{x} = \hat{x}_1 - \hat{x}_2$ with $x_i \in C_i$, $i = 1, 2$. Now we set

$$x^* := \tfrac{\hat{x}_2 + \hat{x}_1}{2}, \quad p := \tfrac{\hat{x}_2 - \hat{x}_1}{2}, \quad \alpha := \langle p, x^*\rangle.$$

Note that $p \neq 0$. By the Projection Theorem, we get for all $x_1 \in C_1$, $x_2 \in C_2$ that

$$\langle 0 - (\hat{x}_1 - \hat{x}_2), x_1 - x_2 - (\hat{x}_1 - \hat{x}_2)\rangle \leq 0$$

from which we deduce (using $x^* - \hat{x}_1 = (\hat{x}_2 - \hat{x}_1)/2$ and $x^* - \hat{x}_2 = (\hat{x}_1 - \hat{x}_2)/2$)

$$\langle x^* - \hat{x}_1, x_1 - \hat{x}_1\rangle + \langle x^* - \hat{x}_2, x_2 - \hat{x}_2\rangle \leq 0.$$

Plugging in $x_1 = \hat{x}_1$ and $x_2 = \hat{x}_2$, respectively, we get for all $x_i \in C_i$ that

$$\langle x^* - \hat{x}_i, x_i - \hat{x}_i \rangle \leq 0.$$

This means that $\hat{x}_i$is the orthogonal projection of x^* onto C_i. Finally, since $x^* - \hat{x}_1 = p$ we get

$$\langle p, x_1 \rangle \leq \langle p, \hat{x}_1 \rangle = \langle p, x^* \rangle + \langle p, \hat{x}_1 - x^* \rangle = \alpha - \|p\|_2^2 < \alpha.$$

Similarly, (using $x^* - \hat{x}_2 = -p$) one shows that $\langle p, x_2 \rangle > \alpha$. □

We recall the following fact from analysis:

Proposition 3.6. *If $(S_x)_x$ is a familiy of compact subsets of a metric space such that the intersection of every finity subfamily of the S_x is non-empty, then $\bigcap_x S_x \neq \emptyset$.*

Proposition 3.7. *Let C be a non-empty convex set and $x_0 \notin C^\circ$. Then there exists a non-zero p such that for all $x \in C$ it holds that*

$$\langle p, x \rangle \leq \langle p, x_0 \rangle.$$

Proof. For every $x \in C$ we define

$$F_x := \{p \mid \|p\|_2 = 1,\ \langle p, x \rangle \leq \langle p, x_0 \rangle\}$$

and observe that each F_x is closed and since F_x is subset of the compact set $\{\|p\|_2 = 1\}$, it is compact as well. Now we show that every intersection of finitely many F_x is non-empty: For $x_1, \dots, x_n \in C$ we define

$$M := \operatorname{conv}(x_1, \dots, x_n).$$

Since C is convex, we have $M \subset C$ and hence, $x_0 \notin M$. Since M is non-empty, convex and closed, we can invoke the Projection Theorem to get that for all $x \in M$ it holds that

$$\langle x_0 - P_M x_0, x - P_M x_0 \rangle \leq 0$$

from which we deduce

$$\|x_0 - P_M x_0\|_2^2 \leq \langle x_0 - P_M x_0, x_0 - x \rangle.$$

Now we set $p = (x_0 - P_M x_0)/\|x_0 - P_M x_0\|$ and get $\langle p, x \rangle \leq \langle p, x_0 \rangle$ for all $x \in M$ and especially $\langle p, x_i \rangle \leq \langle p, x_0 \rangle$ for $i = 1, \dots, n$. This shows $p \in \bigcap_{i=1}^n F_{x_i}$. By the previous proposition, we conclude that $\bigcap_{x \in C} F_x$ is non-empty as well, which shows the assertion. □

Theorem 3.8 (Separation Theorem)**.** *Any two non-empty closed, convex and disjoint sets C_1 and C_2 can be separated by a hyperplane.*

Proof. We consider the Minkowski sum $C = C_1 + (-C_2)$ which is non-empty, closed and convex with $0 \notin C$. By Proposition 3.7 we can separate 0 from C and this hyperplane also separates C_1 and C_2. $\square$

We will also use the following slight generalization which even characterizes proper separation:

Theorem 3.9 (Proper separation theorem)**.** *Two non-empty, convex sets C_1 and C_2 can be properly separated if and only if $\operatorname{ri} C_1$ and $\operatorname{ri} C_2$ are disjoint.*

Proof. **"⇒":** Suppose that $H_{p,\alpha}$ separates C_1 and C_2 properly, i.e. for all $x_i \in C_i$ we have $\langle p, x_1\rangle \le \alpha \le \langle p, x_2\rangle$ and there exist $\bar{x}_i \in C_i$ such that $\langle p, \bar{x}_1\rangle < \langle p, \bar{x}_2\rangle$. We aim to show that for all $x_i \in \operatorname{ri} C$ we have that $\langle p, x_1\rangle < \langle p, x_2\rangle$ as well (which then implies that $\operatorname{ri} C_1$ and $\operatorname{ri} C_2$ are disjoint).

To do so, assume that there are $x_i \in \operatorname{ri} C_i$ such that $\langle p, x_1\rangle = \langle p, x_2\rangle$. By Proposition 1.8 we know that there exists $\epsilon > 0$ such that for $i = 1, 2$ we have

$$y_i := x_i - \epsilon(\bar{x}_i - x_i) = (1+\epsilon)x_i - \epsilon\bar{x}_i \in C_i.$$

Thus,

$$\begin{aligned}\langle p, y_1\rangle &= \langle p, (1+\epsilon)x_1 - \epsilon\bar{x}_1\rangle\\ &= \langle p, (1+\epsilon)x_2\rangle - \epsilon\langle p, \bar{x}_1\rangle\\ &> \langle p, (1+\epsilon)x_2\rangle - \epsilon\langle p, \bar{x}_2\rangle\\ &= \langle p, y_2\rangle.\end{aligned}$$

This contradicts the separation property of $H_{p,\alpha}$.

"⇐": Now let $\operatorname{ri}(C_1) \cap \operatorname{ri}(C_2) = \emptyset$, i.e. $0 \notin \operatorname{ri}(C_1) - \operatorname{ri}(C_2) = \operatorname{ri}(C_1 - C_2)$. We construct a properly separating hyperplane $H_{p,\alpha}$.

Remembering the remark after Proposition 1.8, we assume without loss of generality, that $\operatorname{aff}(C_1 - C_2) = \mathbb{R}^d$. Then $\operatorname{ri}(C_1 - C_2) = \operatorname{int}(C_1 - C_2)$ and by Proposition 3.7 there exists $p \neq 0$ such that for all $x_1 - x_2 \in \operatorname{int}(C_1 - C_2)$

$$\langle p, x_1 - x_2\rangle \ge 0.$$

We conclude that

$$\{x \in \mathbb{R}^d \mid \langle p, x\rangle \ge 0\} \supset \operatorname{int}(C_1 - C_2).$$

Since the set on the left is closed, we also get

$$\{x \in \mathbb{R}^d \mid \langle p, x\rangle \ge 0\} \supset \operatorname{cl}(C_1 - C_2).$$

Thus, we even have

$$\langle p, x_1 - x_2\rangle \ge 0$$

for all x_1, x_2 with $x_1 - x_2 \in C_1 - C_2$. Thus, we can separate C_1 and C_2 with this p and $\alpha := \sup_{x_2 \in C_2} \langle p, x_2 \rangle$. Finally, for $x_1 - x_2 \in \operatorname{int}(C_1 - C_2)$ we have the strict inequality $\langle p, x_1 - x_2 \rangle > 0$.

$\square$

For a non-empty convex set C one calls a halfspace *supporting*, if it contains C and has a point in the closure of C in its boundary. A *supporting hyperplane*, is one which is the boundary of a supporting halfspace. Our above results show that for every $x_0 \in \operatorname{cl}(C) \setminus \operatorname{int}(C)$, there exists a supporting hyperplane for C which contains x_0 (namely, one which separates x_0 from $\operatorname{int}(C)$).

4 Convex functions

When working with convex functions, it will be convenient to use the extended real numbers $\bar{\mathbb{R}} = [-\infty, \infty]$. We will use the following conventions:

$$
\begin{aligned}
&\text{For } -\infty < a \le \infty: && a + \infty = \infty + a = \infty,\\
&\text{For } -\infty \le a < \infty: && a - \infty = -\infty + a = -\infty,\\
&\text{For } 0 < a \le \infty: && a\infty = \infty a = \infty,\ a(-\infty) = (-\infty)a = -\infty,\\
&\text{For } -\infty \le a < 0: && a\infty = \infty a = -\infty,\ a(-\infty) = (-\infty)a = \infty,\\
& && 0\infty = \infty 0 = 0 = 0(-\infty) = (-\infty)0,\\
& && -(-\infty) = \infty,\\
& && \inf\emptyset = \infty,\\
& && \sup\emptyset = -\infty.
\end{aligned}
$$

The expressions $\infty - \infty$ and $-\infty + \infty$ will be left undefined intentionally.

We will use the notion of *indicator function* of a set S which is $i_S : \mathbb{R}^d \to \bar{\mathbb{R}}$ defined by

$$i_S(x) := \begin{cases} 0, & \text{if } x \in S\\ \infty, & \text{if } x \notin S. \end{cases}$$

Indicator functions are helpful to formulate constrained minimization problems $\min_{x\in S} f(x)$ as (formally) unconstrained minimization problems $\min_{x\in\mathbb{R}^d} f(x) + i_S(x)$.

Definition 4.1. A function $f : \mathbb{R}^d \to \bar{\mathbb{R}}$ is proper, if $f(x) > -\infty$ for all x and there exists x_0 such that $f(x_0) < \infty$. The *(effective) domain* of f is

$$\operatorname{dom} f := \{x \mid f(x) < \infty\}.$$

We say that f is *lower semicontinuous* (lsc) at x_0 if

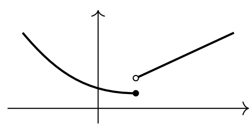
$$f(x_0) \le \liminf_{x\to x_0} f(x)$$

and f is called lsc if it is lsc at every point.

Equivalent formulation of lower semicontinuity are

- f is lsc at x_0 if for every λ with $\lambda < f(x_0)$ there exists $\epsilon > 0$ such that $f(x) \ge \lambda$ for all $x \in B_\epsilon(x_0)$.
- f is lsc at x_0 if $f(x_0) \le \lim_{k\to\infty} f(x_k)$ whenever $x_k \to x_0$ and $\lim f(x_k)$ exists.

Intuitively, a function is lsc, if it only may jump down in the limit.



It is a simple observation that an indicator function i_S is lower semicontinuous if and only if S is closed.

Example 4.2. Consider $f_i : \mathbb{R} \to \bar{\mathbb{R}}$ given by

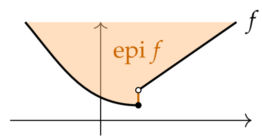
$$f_1 = i_{]1,1]}, \qquad f_2 = i_{[-1,1]}$$

$$f_3(x) = \begin{cases} 0, & \text{if } x = 0\\ 1, & \text{else.}\end{cases} \qquad f_4(x) = \begin{cases} \frac{1}{x}, & \text{if } x > 0\\ \infty, & \text{else.}\end{cases}$$

Then: f_1 is not lsc while f_2, f_3 and f_4 are. △

Definition 4.3. The *epigraph* of a function $f : \mathbb{R}^d \to \bar{\mathbb{R}}$ is

$$\operatorname{epi} f := \{(x,\alpha) \in \mathbb{R}^d \times \mathbb{R} \mid f(x) \le \alpha\}.$$



The *level sets* of f for level $\alpha \in \mathbb{R}$ are

$$\operatorname{lev}_\alpha f := \{x \in \mathbb{R}^d \mid f(x) \le \alpha\}$$

Theorem 4.4. *The following conditions for $f : \mathbb{R}^d \to \mathbb{R}$ are equivalent:*

i) f *is lsc,*

ii) $\operatorname{epi} f$ *is closed,*

iii) for all $\alpha \in \mathbb{R}$, $\operatorname{lev}_\alpha f$ *is closed.*

Proof. **i) $\Longrightarrow$ ii):** Let f be lsc and consider a sequence $(x_n, \alpha_n) \to (x, \alpha)$ with $(x_n, \alpha_n) \in \operatorname{epi} f$, i.e. $f(x_n) \leq \alpha_n$. Since f is lsc, we have $f(x) \leq \liminf f(x_n) \leq \lim \alpha_n = \alpha$ and this shows $(x, \alpha) \in \operatorname{epi} f$.

ii) $\Longrightarrow$ iii): Let $\operatorname{epi} f$ be closed and $\alpha \in \mathbb{R}$. If x_n converges to x and fulfilles $x_n \in \operatorname{lev}_\alpha f$, we also have $\lim(x_n, \alpha) = (x, \alpha)$ and $(x_n, \alpha) \in \operatorname{epi} f$. Since $\operatorname{epi} f$ is closed, the have $(x, \alpha) \in \operatorname{epi} f$ which shows $f(x) \leq \alpha$, and hence, $x \in \operatorname{lev}_\alpha f$ which shows the closedness.

iii) $\Longrightarrow$ i): Let $\operatorname{lev}_\alpha f$ be closed for all α and let $x_0 \in \mathbb{R}^d$. If $f(x_0) = -\infty$ holds, f is obviously lsc at x_0. Otherwise, let $\alpha < f(x_0)$ which means that $x_0 \notin \operatorname{lev}_\alpha f$. Since $\operatorname{lev}_\alpha f$ is closed, there is $\epsilon > 0$ such that $\operatorname{lev}_\alpha f \cap B_\epsilon(x_0) = \emptyset$. Thus, $f(x) > \alpha$ for all $x \in B_\epsilon(x_0)$ and hence, f is lsc at x_0. □

Proposition 4.5. *If* $f, g, f_i : \mathbb{R}^d \to \bar{\mathbb{R}}$ *are lsc,* $A \in \mathbb{R}^{d \times m}$ *and* $\alpha > 0$, *then the following functions are also lsc:*

i) $f + g$ *(if it's defined),* αf

ii) $f \circ A$

iii) $\inf(f, g)$

iv) $\sup_i f_i$ *for arbitrarily many* f_i.

It's not true that the infimum of infinitely many lsc f_i is again lsc. Can you think of an example?

Proof. The items i) and ii) follow directly from the definition. For iii) note that $(x, \alpha) \in \operatorname{epi}(\inf(f, g))$ if and only if $f(x) \leq \alpha$ or $g(x) \leq \alpha$. Hence, $\operatorname{epi}(\inf(f, g)) = \operatorname{epi} f \cup \operatorname{epi} g$. And since the union of two closed sets is closed, the assertion follows.

For item iv) note that $\operatorname{epi}(\sup_i f_i)$ is the intersection of all the sets $\operatorname{epi}(f_i)$ and since the intersection of closed sets is closed set, we are done. □

Since we characterized lower semi-continuity of a function by closedness of the epigraph, we can associate to every function $f : \mathbb{R}^d \to \bar{\mathbb{R}}$ which is not a lsc another lsc function $\bar{f}$ via closing the epigraph, i.e. $\bar{f}$ is characterized by

$$\operatorname{epi}(\bar{f}) = \overline{\operatorname{epi}(f)}.$$

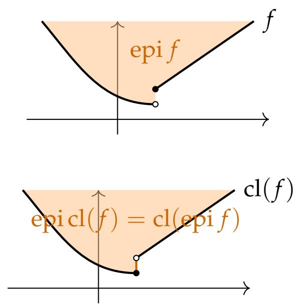
This function is called the *lower semi-continuous hull* of f.

Taking special care of the case when $f(x) = -\infty$ can occur, we define the *closure* of f by

$$\operatorname{cl}(f) := \begin{cases} \bar{f}, & \text{if } f(x) > -\infty \text{ for all } x \\ -\infty, & \text{if } f(x) = -\infty \text{ for some } x. \end{cases}$$

We call a function closed, if $\operatorname{cl} f = f$ and for functions which do not take the value $-\infty$, closed means the same as lsc and hence, we have

$$(\operatorname{cl} f)(x_0) = \liminf_{x \to x_0} (\operatorname{cl} f)(x) = \liminf_{x \to x_0} f(x) \le f(x_0).$$



Intuitively, the closure of a function moves the function values to the lowest possible values at points of discontinuity.

Example 4.6. For the function

$$f(x) = i_{]a,b[}(x) = \begin{cases} 0, & a < x < b \\ \infty, & \text{else,} \end{cases}$$

the closure (and also the lsc hull) is

$$\operatorname{cl} f = \bar{f} = i_{[a,b]}.$$

△

Definition 4.7. A function $f : \mathbb{R}^d \to \bar{\mathbb{R}}$ is called

- *convex*, if for all $x, y \in \operatorname{dom}(f)$, $\lambda \in [0,1]$ it holds that

  $$f(\lambda x + (1-\lambda)y) \le \lambda f(x) + (1-\lambda) f(y).$$

- *strictly convex*, if it is convex and for $x \ne y$ and $\lambda \in \,]0,1[$ it holds that

  $$f(\lambda x + (1-\lambda)y) < \lambda f(x) + (1-\lambda) f(y).$$

- *uniformly convex*, if there exists a strictly increasing function φ with $\varphi(0) = 0$ such that for all $x, y \in \operatorname{dom}(f)$, $\lambda \in [0,1]$ it holds that

  $$f(\lambda x + (1-\lambda)y) \le \lambda f(x) + (1-\lambda) f(y) - \lambda(1-\lambda)\varphi(\|x-y\|_2).$$

- *strongly convex* with constant $\sigma > 0$, if for all $x, y \in \operatorname{dom}(f)$, $\lambda \in [0,1]$ it holds that

  $$f(\lambda x + (1-\lambda)y) \le \lambda f(x) + (1-\lambda) f(y) - \tfrac{\sigma}{2}\lambda(1-\lambda)\|x-y\|_2^2.$$

One can show that a function f is strongly convex with constant σ if and only if the function $x \mapsto f(x) - \frac{\sigma}{2}\|x\|_2^2$ is convex.

It's clear that

$$\text{strongly convex} \implies \text{strictly convex} \implies \text{convex}$$

but the reverse implications do not hold: $f(x) = i_{[a,b]}$ is convex but not strictly so, $f(x) = x^4$ is stricly convex, but not strongly so. It's also clear that $\mathrm{dom}(f)$ is a convex set for any convex function f.

Put differently: A function is strongly convex (with modulus) σ, if it is uniformly convex with respect to $\varphi(t) = \frac{\sigma}{2}t^2$.
The constant of strong convexity is also called *modulus of strong convexity*.

Finally, we call a function f *concave* if $-f$ is convex.

By induction one can deduce from the defining in the inequality:

Lemma 4.8 (Jensen's inequality). *A function $f : \mathbb{R}^d \to \bar{\mathbb{R}}$ is convex if and only if for all $x_i \in \mathrm{dom}(f)$ and $\lambda_i \in [0,1]$ with $\sum_{i=1}^n \lambda_i = 1$ it holds that*

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

Lemma 4.9. *If a convex function f has a finite value at some point $x_0 \in \mathrm{ri}(\mathrm{dom}(f))$, then it is proper (i.e. it does not assume the value $-\infty$ anywhere).*

Proof. For a contradiction, assume that $x_1 \in \mathrm{dom}(f)$ exists with $f(x_1) = -\infty$. Since $x_0 \in \mathrm{ri}(\mathrm{dom}(f))$, there exists $x_2 \in \mathrm{ri}(\mathrm{dom}(f))$ and $\lambda \in [0,1]$ such that $x_0 = \lambda x_1 + (1-\lambda)x_2$. But, by convexity we would get

$$f(x_0) \leq \lambda f(x_1) + (1-\lambda)f(x_2) = -\infty$$

which contradicts $f(x_0)$ finite. □

Example 4.10.

1. Any affine function is convex as well as concave and only affine functions are both at the same time.
2. Any norm $f(x) = \|x\|$ is convex, but no norm is strictly convex since for $y = 0$ and $x \neq 0$ and $\lambda \in [0,1]$ we have

$$\|\lambda x + (1-\lambda)y\| = \lambda\|x\| + (1-\lambda)\|y\|.$$

3. Indicator functions i_C are convex exactly for convex sets C.

△

5 Characterization of convex functions

For differentiable functions, convexity can be described with derivatives:

Theorem 5.1. *If $f : \mathbb{R}^d \to \mathbb{R}$ is differentiable, the following conditions are equivalent:*

i) f is convex,

ii) the gradient $\nabla f : \mathbb{R}^d \to \mathbb{R}^d$ is monotone, i.e. for all $x, y \in \mathbb{R}^d$

$$\langle \nabla f(x) - \nabla f(y), x - y \rangle \geq 0,$$

iii) for all $x, y \in \mathbb{R}^d$ it holds that

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle.$$

If f is twice differentiable, then f is convex if and only if the Hessian $\nabla^2 f(x)$ is positive semi-definite for every x.

Strict convexity is characterized by strict inequalities for $x \neq y$ in i) and ii). Positive definiteness in the case of twice differentiability is sufficient but not necessary for strict convexity.

A function $g : \mathbb{R}^d \to \mathbb{R}^d$ is called *monotone* if for all x, y it holds that $\langle g(x) - g(y), x - y \rangle \geq 0$. Can you see, why this is called "monotonicity"? (Hint: consider $d = 1$.)

The equivalence of i) and iii) is quite remarkable: Besides the definition of convex function from above by "function lies below it's secants" we have the equivalent description from below by "function lies above it's tangents". This is comparable to the duality of the descriptions of convex sets from the inside (via convex combinations) and the outside (by separating hyperplanes).

The function $f(x) = x^4$ is strictly convex, but the second derivative vanishes (i.e., is not positive definite) at $x = 0$.

Proof. We first show i) $\iff$ iii) $\iff$ ii):

i) $\implies$ iii): For $\lambda \in]0, 1]$ we rearrange $f(\lambda y + (1-\lambda)x) \leq \lambda f(y) + (1-\lambda) f(x)$ to

$$\tfrac{f(x+\lambda(y-x)) - f(x)}{\lambda} \leq f(y) - f(x).$$

For $\lambda \to 0$ the left hand side converges to the directional derivative of f in x in the direction of $y - x$ which equals $\langle \nabla f(x), y - x \rangle$ since f is differentiable.

iii) $\implies$ i): Set $x_\lambda = \lambda x + (1-\lambda)y$ and note that $x - x_\lambda = (1-\lambda)(x - y)$ and $y - x_\lambda = -\lambda(x - y)$. We get the inequalities

$$f(x) \geq f(x_\lambda) + \langle \nabla f(x_\lambda), x - x_\lambda \rangle = f(x_\lambda) + (1-\lambda)\langle \nabla f(x_\lambda), x - y \rangle$$
$$f(y) \geq f(x_\lambda) + \langle \nabla f(x_\lambda), y - x_\lambda \rangle = f(x_\lambda) - \lambda \langle \nabla f(x_\lambda), x - y \rangle.$$

Multiplying the first inequality with λ and the second with $(1-\lambda)$ and adding the inequality shows the assertion.

iii) $\implies$ ii): We add $f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle$ and $f(x) \geq f(y) + \langle \nabla f(y), x - y \rangle$ and get $0 \geq \langle \nabla f(x) - \nabla f(y), y - x \rangle$ which shows the assertion.

ii) $\Longrightarrow$ iii): Since for $h(\lambda) = f(y + \lambda(x-y))$ we have $h'(\lambda) = \langle \nabla f(y+\lambda(x-y)), x-y\rangle$ and the fundamental theorem of calculus gives

$$\begin{aligned} f(x) - f(y) = h(1) - h(0) &= \int_0^1 h'(\lambda)\mathrm{d}\lambda \\ &= \int_0^1 \langle \nabla f(y+\lambda(x-y)), x-y\rangle \mathrm{d}\lambda. \end{aligned}$$

Subtracting $\langle \nabla f(x), x-y\rangle$ from both sides, we get

$$f(x) - f(y) - \langle \nabla f(x), x-y\rangle = \int_0^1 \langle \nabla f(y+\lambda(x-y)) - \nabla f(x), x-y\rangle \mathrm{d}\lambda.$$

The right hand side is non-negative (since $y + \lambda(x-y) - x = (1-\lambda)(y-x)$ and ∇f is monotone), by assumption and this shows i).

The claim on strict convexity follows by inspection of the above arguments in this case.

For twice differentiable functions we show the equivalence of positive semidefinite Hessian and ii): Set $x_\tau = x + \tau s$ for $\tau > 0$ and with ii) we get (using $\frac{\mathrm{d}}{\mathrm{d}\lambda}\langle \nabla f(x+\lambda s), s\rangle = \langle \nabla^2 f(x+\lambda s)s, s\rangle$)

$$\begin{aligned} 0 &\leq \tfrac{1}{\tau^2}\langle \nabla f(x_\tau) - f(x), x_\tau - x\rangle \\ &= \tfrac{1}{\tau}\langle \nabla f(x_\tau) - f(x), s\rangle = \tfrac{1}{\tau}\int_0^\tau \langle \nabla^2 f(x+\lambda s)s, s\rangle \mathrm{d}\lambda \end{aligned}$$

and $\tau \to 0$ shows that $\nabla^2 f(x) \succcurlyeq 0$. Conversely, if $\nabla^2 f(x) \succcurlyeq 0$ we can write (using the fundamental theorem of calculus twice)

$$\begin{aligned} f(y) &= f(x) + \langle \nabla f(x), y-x\rangle + \int_0^1\int_0^\tau \langle \nabla^2 f(x+\lambda(y-x))(y-x), y-x\rangle \mathrm{d}\lambda\mathrm{d}\tau \\ &\geq f(x) + \langle \nabla f(x), y-x\rangle \end{aligned}$$

which implies convexity of f.

$\square$

Example 5.2.
1. The function $f(x) = \frac{1}{2}\langle Ax, x\rangle + \langle a, x\rangle + b$ is convex as soon as the matrix A is positive semidefinite and it is strictly convex of A is positive definite (since $\nabla^2 f(x) = A$). Moreover, one can even show strong convexity of f for positive definite A (what is the modulus?).

2. The “log-sum-exp” function is

$$\operatorname{logsumexp}(x) = \log(\sum_{i=1}^d \exp(x_i)).$$

We abbreviate $h(x) = \sum_{i=1}^d \exp(x_i)$ and get

$$\partial_i \operatorname{logsumexp}(x) = \tfrac{\exp(x_i)}{h(x)}$$

and

$$\partial_j \partial_i \operatorname{logsumexp}(x) = \begin{cases} \frac{\exp(x_i)h(x) - \exp(2x_i)}{h(x)^2}, & i = j \\ -\frac{\exp(x_i + x_j)}{h(x)^2}, & i \neq j. \end{cases}$$

Thus, we can compute for $z \in \mathbb{R}^d$:

$$\begin{aligned}
\langle \nabla^2 \operatorname{logsumexp}(x) z, z\rangle &= \tfrac{1}{h(x)^2}\Big(h(x)\sum_{i=1}^d \exp(x_i) z_i^2 - \sum_{i,j=1}^d \exp(x_i + x_j) z_i z_j\Big)\\
&= \tfrac{1}{h(x)^2}\Big(\sum_{i,j=1}^d \exp(x_i + x_j) z_i^2 - \sum_{i,j=1}^d \exp(x_i + x_j) z_i z_j\Big)\\
&= \tfrac{1}{h(x)^2}\Big(\tfrac12\sum_{i,j=1}^d \exp(x_i + x_j) z_i^2 + \tfrac12\sum_{i,j=1}^d \exp(x_i + x_j) z_j^2\\
&\qquad - \sum_{i,j=1}^d \exp(x_i + x_j) z_i z_j\Big)\\
&= \tfrac{1}{h(x)^2}\Big(\sum_{i,j=1}^d \exp(x_i + x_j)\underbrace{\big(\tfrac12 z_i^2 + \tfrac12 z_j^2 - z_i z_j\big)}_{=\frac12 (z_i - z_j)^2 \geq 0}\Big) \geq 0
\end{aligned}$$

which shows convexity of logsumexp.

$\triangle$

If a function f is convex, its level set $\operatorname{lev}_\alpha f$ are all convex (if $f(x), f(y) \leq \alpha$, then $f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda) f(x) \leq \alpha$). The converse is not true (consider something like $f(x) = \log(1 + x^2)$ on the real line, for example). Convexity of the epigraph, though, does characterize convexity of the function:

Functions with convex level sets sometimes are called *quasi-convex*.

Proposition 5.3. *A function $f : \mathbb{R}^d \to \bar{\mathbb{R}}$ is convex if and only if* epi f *is a convex set.*

Proof. The case $f \equiv \infty$ is clear since in this case $\operatorname{epi} f = \emptyset$. So consider $\operatorname{dom}(f) \neq \emptyset$.

$\Rightarrow$: Let f be convex and $(x, a), (y, b) \in \operatorname{epi}(f)$, i.e. $f(x) \leq a$, $f(y) \leq b$. Hence, for $\lambda \in [0,1]$: $f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda) f(y) \leq \lambda a + (1-\lambda) b$. But this means that $\lambda \left[\begin{smallmatrix} x \\ a \end{smallmatrix}\right] + (1-\lambda)\left[\begin{smallmatrix} y \\ b \end{smallmatrix}\right] \in \operatorname{epi}(f)$.

$\Leftarrow$: Let $\operatorname{epi}(f)$ be convex and $x, y \in \operatorname{dom}(f)$ with $f(x), f(y) \neq -\infty$. Since $(x, f(x)), (y, f(y)) \in \operatorname{epi}(f)$, we have for $\lambda \in [0,1]$

Beware: Sometimes I write tuples as (x, a) and sometimes they will be $\left[\begin{smallmatrix} x \\ a \end{smallmatrix}\right]$ depending on the typographic circumstances.

$$\lambda(x, f(x)) + (1-\lambda)(y, f(y)) = (\lambda x + (1-\lambda) y, \lambda f(x) + (1-\lambda) f(y)) \in \operatorname{epi}(f)$$

and this means that $f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda) f(y)$. If $f(x) = -\infty$, then we use $(x, -N)$ instead of $(x, f(x))$ and let $N \to -\infty$. $\square$

We apply the separation of points from convex sets with hyperplanes to the epigraph and obtain:

Proposition 5.4. *If $f : \mathbb{R}^d \to \bar{\mathbb{R}}$ is a convex and proper, then there exists $p \in \mathbb{R}^d$ and $\alpha \in \mathbb{R}$ such that for all x it holds that*

$$f(x) \geq \langle p, x\rangle + \alpha.$$

Proof. Since $\operatorname{cl} f \leq f$ we only need to consider the case where f is closed.

If $x \notin \operatorname{dom}(f)$, the inequality is valid for any p and α. Now fix $x_0 \in \operatorname{dom}(f)$ and β such that $f(x_0) > \beta$, i.e. $(x_0, \beta) \notin \operatorname{epi}(f)$. Since $\operatorname{epi}(f)$ is closed and convex, we can use Theorem 3.5 to separate the compact singleton $\{(x_0, \beta)\}$ from $\operatorname{epi}(f)$. This means, that there exists $(\bar{p}, -b) \in \mathbb{R}^{d+1} \setminus \{0\}$ and $\epsilon > 0$ such that for all $x \in \operatorname{dom}(f)$ it holds that

$$\langle \bar{p}, x\rangle - bf(x) \leq \langle \bar{p}, x_0\rangle - b\beta - \epsilon. \qquad (*)$$

For $x = x_0$ we get

$$b(f(x_0) - \beta) \geq \epsilon > 0$$

and since $f(x_0) - \beta > 0$ we obtain $b > 0$ as well. Hence, with $p = \bar{p}/b$ we get, by dividing (*) by b,

$$f(x) \geq \langle p, x\rangle - \langle p, x_0\rangle + \beta$$

and we proved the claim with $\alpha = -\langle p, x_0\rangle + \beta$. □

The following corollary is a simple exercise.

Corollary 5.5. *If $f : \mathbb{R}^d \to \bar{\mathbb{R}}$ is convex, it holds for all $x_0 \in \mathbb{R}^d$*

$$\operatorname{cl} f(x_0) = \sup\{\langle p, x_0\rangle + \alpha \mid p \in \mathbb{R}^d, \alpha \in \mathbb{R} \text{ with } f(x) \geq \langle p, x\rangle + \alpha\ \forall x \in \mathbb{R}^d\}.$$

Convexity if preserved under several operations.

Proposition 5.6.

1. *$f_1 + f_2$ is convex if f_1 and f_2 are convex.*
2. *αf is convex if f is convex and $\alpha \geq 0$.*
3. *$f \circ A$ is convex if f is convex and A is linear.*
4. *$\varphi \circ f$ is convex if f is convex and $\varphi : \bar{\mathbb{R}} \to \bar{\mathbb{R}}$ is convex and increasing (with $\varphi(\infty) = \infty$ and $\varphi(-\infty) = -\infty$).*
5. *$\sup_{i \in I} f_i$ is convex if all the f_i are.*

All proofs are straightforward calculations.

Example 5.7. By Proposition 5.6 5. we see that $f(x) = \max(x_1, \ldots, x_d)$ is convex. △

6 Continuity of convex functions and minimizers

Surprisingly, the notion of convexity implies a certain continuity. We start with a lemma:

Lemma 6.1. *Let $f : \mathbb{R}^d \to \bar{\mathbb{R}}$ be proper and convex and suppose that it is locally bounded at x_0, i.e. there exists $\delta > 0$ and $m, M \in \mathbb{R}$ such that for $x \in B_{2\delta}(x_0)$ it holds that*

$$m \leq f(x) \leq M.$$

Then f is Lipschitz continuous on $B_\delta(x_0)$ with Lipschitz constant at most $(M-m)/\delta$.

Proof. The proof is a bit technical: Let $x_1, x_2 \in B_\delta(x_0)$, $x_1 \neq x_2$ and set

$$y := \left(1 + \tfrac{\delta}{\|x_1-x_2\|_2}\right)x_2 - \tfrac{\delta}{\|x_1-x_2\|_2}x_1 = x_2 + \delta\tfrac{x_1-x_2}{\|x_1-x_2\|_2}.$$

It holds that $\|y - x_2\|_2 \leq \delta$, i.e. $y \in B_{2\delta}(x_0)$. By rearranging to

$$x_2 = \tfrac{\|x_1-x_2\|_2}{\delta+\|x_1-x_2\|_2}y + \tfrac{\delta}{\delta+\|x_1-x_2\|_2}x_1$$

we see that x_2 is a convex combination of x_1 and y.

Since f is convex we get,

$$f(x_2) \leq \tfrac{\|x_1-x_2\|_2}{\delta+\|x_1-x_2\|_2}f(y) + \tfrac{\delta}{\delta+\|x_1-x_2\|_2}f(x_1)$$

which leads to

$$\begin{aligned} f(x_2) - f(x_1) &\leq \tfrac{\|x_1-x_2\|_2}{\delta+\|x_1-x_2\|_2}(f(y) - f(x_1)) \\ &\leq \tfrac{M-m}{\delta}\|x_1 - x_2\|_2. \end{aligned}$$

Since we can swap the roles of x_1 and x_2 this shows the desired Lipschitz continuity. ☐

Theorem 6.2. *Let $f : \mathbb{R}^d \to \bar{\mathbb{R}}$ be proper and convex. Then f is Lipschitz continuous on any compact subset C of $\operatorname{ri}(\operatorname{dom}(f))$.*

Proof.

1. (Restriction to fulldimensional case.) Let $C \subset \operatorname{ri}(\operatorname{dom}(f))$. Since C is a subset of the affine hull of the relative interior, we can restrict f to this affine set and hence, we may assume that $\operatorname{ri}(\operatorname{dom}(f)) = \operatorname{int}(\operatorname{dom}(f))$.

2. (Locally Lipschitz.) Let $x_0 \in C$ and since $x_0 \in \operatorname{int}(\operatorname{dom}(f))$, there exist $v_1, \dots, v_r \in \operatorname{dom}(f)$ and $\delta > 0$ such that

$$B_{2\delta}(x_0) \subset S := \operatorname{conv}(v_1, \dots, v_r) \subset \operatorname{dom}(f).$$

Hence, any $v \in B_{2\delta}(x_0)$ can be written as $v = \sum_{i=1}^r \alpha_i v_i$, $\sum_i \alpha_i = 1$, $\alpha_i \geq 0$ and convexity of f gives

$$f(v) \leq \sum_{i=1}^r \alpha_i f(v_i) \leq \max_i f(v_i) =: M.$$

This shows that f is bounded from above on $B_{2\delta}(x_0)$.

By Proposition 5.4 we know that there are $p \in \mathbb{R}^d$ and $\alpha \in \mathbb{R}$ such that

$$f(x) \geq \langle p, x\rangle + \alpha.$$

Hence, f is bounded from below on $B_{2\delta}(x_0)$ since the right hand side is so.

Thus, we can apply Lemma 6.1 and obtain that f is Lipschitz continuous on $B_\delta(x_0)$.

3. (Lipschitz on full C) The previous step shows that for any $x_0 \in C$ there exists $\delta(x_0) > 0$ and $L(x_0, \delta) \geq 0$ such that for all $x, y \in B_\delta(x_0)$ it holds that

$$|f(x) - f(y)| \leq L(x_0, \delta)\|x - y\|_2.$$

The balls $B_{\delta(x_0)}(x_0)$ cover the compact set C and hence, there are $x_1, \ldots, x_K$ such that C is covered by $B_{\delta(x_1)}(x_1), \ldots, B_{\delta(x_K)}(x_K)$. Hence f is Lipschitz continuous on C with constant $L = \max_k(L(x_k, \delta(x_k)))$.

□

Corollary 6.3. *If $f : \mathbb{R}^d \to \bar{\mathbb{R}}$ is proper and convex, then it is continuos on* $\operatorname{ri}(\operatorname{dom}(f))$ *relative to* $\operatorname{aff}(\operatorname{dom}(f))$. *If f is additionally finite everywhere, then it is continuous on $\mathbb{R}^d$.*

Now we start our treatment of minimization problems

$$\min_{x\in\mathbb{R}^d} f(x).$$

A ridiculously large class of pratically relevant problems can be written in this form (note that we can treat constraint problems just by setting $f = g + i_C$) and we will see some examples later in the lecture.

In addition to the usual definitions of local and global minima of functions, we will also need the set of minimizers which we will denote by

$$\operatorname{argmin} f := \{\hat{x} \in \operatorname{dom}(f) \mid f(\hat{x}) = \inf_x f(x)\}.$$

Similarly we have $\operatorname{argmax} f := \operatorname{argmin}(-f)$.

Now we are interested in conditions on f which ensure the follwing properties

i) f attains its minimum, i.e. $\operatorname{argmin} f \neq \emptyset$,

ii) every local minimizer of f is a global minimizer,

iii) the global minimizer is unique.

Mere convexity does not even ensure existence of minimizers, even in the proper case as $f(x) = \exp(x)$ shows.

Definition 6.4. A function $f : \mathbb{R}^d \to \bar{\mathbb{R}}$ is called *level-bounded* or *coercive* if all level sets $\mathrm{lev}_\alpha f$ are bounded.

Another way to put this: f is coercive if and only if $f(x) \to \infty$ whenever $\|x\|_2 \to \infty$.

$\Leftarrow$: Assume that $f(x) \to \infty$ whenever $\|x\|_2 \to \infty$. If $x \in \mathrm{lev}_\alpha f$, then $f(x) \leq \alpha$. Thus, there can't be an unbounded sequence in $\mathrm{lev}_\alpha f$ since this would contradict that $f(x) \leq \alpha$ for all $x \in \mathrm{lev}_\alpha f$.

$\Rightarrow$: We prove the contraposition: Assume that there exists $\|x_n\| \to \infty$ with $f(x_n) \leq C$. But then $\mathrm{lev}_C f$ is not bounded, since $x_n \in \mathrm{lev}_C f$.

Theorem 6.5. *Let $f : \mathbb{R}^d \to \bar{\mathbb{R}}$ be proper.*

i) If f is lsc and coercive, then $\operatorname{argmin} f$ is non-empty and compact.

ii) If f is convex, then every local minimizer is also global and the set $\operatorname{argmin} f$ is convex (but possibly empty).

iii) If f is strictly convex, then $\operatorname{argmin} f$ contains at most one point.

In other words: lsc and coercivity ensure existence of minimizers, convexity excluded non-global minimizers and, strict convexity ensures uniqueness of minimizers.

Proof. i) Since f is proper, we have that $\bar{\alpha} := \inf f < \infty$. Hence, there is a sequence x_n such that $f(x_n) \to \bar{\alpha}$. Since f is coercive, the sequence x_n is bounded, and hence, it has a convergent subsequence x_{n_k} with limit $\bar{x}$. Since f is lsc, this limit fulfills $f(\bar{x}) \leq \liminf_k f(x_{n_k}) = \bar{\alpha}$ and this shows $\bar{x} \in \operatorname{argmin} f$. Moreover $\operatorname{argmin} f = \mathrm{lev}_{\bar{\alpha}} f$ and since all level sets are closed (by lsc) and bounded (by coercivity), $\operatorname{argmin} f$ is compact.

ii) Let $\bar{x}$ be a local minimizer of f, i.e. there exists $\delta > 0$ such that $f(\bar{x}) \leq f(x)$ for all $x \in B_\delta(\bar{x})$. Now let $y \in \mathbb{R}^d$ and choose $\lambda \in \,]0, \delta/\|\bar{x} - y\|_2[$ with $\lambda < 1$. Then $\lambda y + (1-\lambda)\bar{x} \in B_\delta(\bar{x})$ and by convexity of f we get

$$f(\bar{x}) \leq f(\lambda y + (1-\lambda)\bar{x}) \leq \lambda f(y) + (1-\lambda) f(\bar{x}).$$

This implies $f(\bar{x}) \leq f(y)$ and hence, $\bar{x}$ is a global minimizer.

If $\bar{x}, x^*$ are global miminizers, i.e. $f(\bar{x}) = f(x^*) = \bar{\alpha}$, then for every $\lambda \in [0,1]$

$$f(\lambda\bar{x} + (1-\lambda)x^*) \leq \lambda f(\bar{x}) + (1-\lambda) f(x^*) = \bar{\alpha},$$

i.e. $\lambda\bar{x} + (1-\lambda)x^*$ is also a global minimizers and hence, the set of global minimizers ist convex.

iii) If f is strictly convex, assume that $\bar{x}$ and x^* are two different global minimizers. But then

$$f((\bar{x} + x^*)/2) < \tfrac{1}{2}(f(\bar{x}) + f(x^*))$$

and hence, $(\bar{x} + x^*)/2$ would be below the global minimum which is impossible.

$\square$

7 Inf-projection and inf-convolution

Definition 7.1. For a function $\varphi : \mathbb{R}^d \times \mathbb{R}^m \to \bar{\mathbb{R}}$ we define the *epigraphical projection* or *inf-projection* as

$$v(u) = \inf_{x \in \mathbb{R}^d} \varphi(x, u).$$

This function is called epigraphical projection, since the epigraph of v arises as the projection of the epigraph of φ.

Theorem 7.2. *If $\varphi : \mathbb{R}^d \times \mathbb{R}^m \to \bar{\mathbb{R}}$ is convex, the its epigraphical projection v is also convex. In particular, it holds for any convex set $C \subset \mathbb{R}^{m+1}$ that $v(u) := \inf\{\alpha \in \mathbb{R} \mid (\alpha, u) \in C\}$ is convex.*

Proof. For $(x_1, u_1), (x_2, u_2) \in \operatorname{dom} \varphi$ and $\lambda \in [0, 1]$ we have by convexity of φ and the definition of v that

$$\begin{aligned} \lambda \varphi(x_1, u_1) + (1 - \lambda)\varphi(x_2, u_2) &\geq \varphi(\lambda x_1 + (1 - \lambda)x_2, \lambda u_1 + (1 - \lambda)u_2) \\ &\geq v(\lambda u_1 + (1 - \lambda)u_2). \end{aligned}$$

Taking the infimum over x_1 and x_2 on the left hand side gives

$$\lambda v(u_1) + (1 - \lambda)v(u_2) \geq v(\lambda u_1 + (1 - \lambda)u_2)$$

which proves the first claim. For the further claim, apply the first part to $\varphi(\alpha, u) = \alpha + i_C(\alpha, u)$. □

Unfortunately, epigraphical projections are not always as nice as one would want them to be:

Example 7.3. Let $f : \mathbb{R} \to \mathbb{R}$ be convex and set

$$\varphi(x, u) = \begin{cases} f(x), & \exp(x) \leq u \\ \infty, & \text{else.} \end{cases}$$

This is a proper, convex function, and also lsc if f is lsc. But for $f(x) = x$ we get that

$$v(u) = \inf_{x \in \mathbb{R}} \varphi(x, u) = \begin{cases} \infty, & u \leq 0 \\ -\infty, & u > 0 \end{cases},$$

i.e. v is neither proper nor lsc. △

Definition 7.4. For two functions $f_1, f_2 : \mathbb{R}^d \to \bar{\mathbb{R}}$ we define the *inf-convolution* (also called *infimal convolution* or *epi-addition*) as

$$(f_1 \,\square\, f_2)(u) := \inf_{x+y=u} f_1(x) + f_2(y) = \inf_x f_1(x) + f_2(u - x)$$

If the infimum is attained whenever it is finite, we call the inf-convolution *exact*.

You may compare this definition to the standard convolution of two functions $f_1 * f_2(u) = \int f_1(x) f_2(u-x)dx$ and note that the integral ("generalized sum") has been replaces by an infimum ("generalized minimum") and the multiplication has been replaced by an addition.

Since different infima commute, we get for more than two functions

$$\begin{aligned} f_1 \,\square\, (f_2 \,\square\, f_3)(u) &= \inf_{x+y=u} f_1(x) + (f_2 \,\square\, f_3)(y) \\ &= \inf_{x+y=u} \Big[f_1(x) + \inf_{z+v=y} f_2(z) + f_3(v) \Big] \\ &= \inf_{\substack{x+y=u \\ z+v=y}} f_1(x) + f_2(z) + f_3(v) \\ &= \inf_{x+z+v=u} f_1(x) + f_2(z) + f_3(v) \end{aligned}$$

(and we see that inf-convolutions are associative.

Example 7.5. 1. For two sets S_1 and S_2 we get for the inf-convolution of their indicator functions

$$(i_{S_1} \,\square\, i_{S_2})(u) = \inf_{x+y=u} i_{S_1}(x) + i_{S_2}(y) = i_{S_1+S_2}(u).$$

2. If we take $f_1(x) = \|x\|$ (for some norm) and $f_2 = i_C$ the indicator function of a convex set C, we get as their inf-convolution

$$(f_1 \,\square\, f_2)(u) = \inf_{x \in C} \|u - x\| = d(u, C),$$

i.e., the distance function for the set C (with respect to the norm $\|\cdot\|$)

$\triangle$

Theorem 7.6. *Let $f_{1/2} : \mathbb{R}^d \to \bar{\mathbb{R}}$ be proper. Then it holds*

$$\operatorname{epi} f_1 + \operatorname{epi} f_2 \subset \operatorname{epi}(f_1 \,\square\, f_2)$$

and equality holds if and only if the inf-convolution is exact. Moreover, if f_1 and f_2 are convex, then so is $f_1 \,\square\, f_2$.

Proof. First, lets show the inclusion: Let $(x, \alpha) \in \operatorname{epi} f_1 + \operatorname{epi} f_2$, i.e. there exist $(\alpha_i, x_i) \in \operatorname{epi} f_i$ $(i = 1, 2)$ with $x = x_1 + x_2$, $\alpha = \alpha_1 + \alpha_2$ and $f_1(x_1) + f_2(x_2) \le \alpha_1 + \alpha_2 = \alpha$. Hence

$$(f_1 \,\square\, f_2)(x) = \inf_{y_1+y_2=x} f_1(y_1) + f_2(y_2) \le f_1(x_1) + f_2(x_2) \le \alpha$$

and this shows that $(x, \alpha) \in \operatorname{epi}(f_1 \,\square\, f_2)$.

Now let's show that there is equality precisely in the case of exactness: Let $(x, \alpha) \in \operatorname{epi}(f_1 \,\square\, f_2)$, i.e. $(f_1 \,\square\, f_2)(x) \le \alpha$. By exactness, there are x_1, x_2 with $x = x_1 + x_2$ and $(f_1 \,\square\, f_2)(x) = f_1(x_1) + f_2(x_2) \le \alpha$. Hence

$$(x, \alpha) = (x_1, f_1(x_1)) + (x_2, \underbrace{\alpha - f_1(x_1)}_{\ge f_2(x_2)}) \in \operatorname{epi} f_1 + \operatorname{epi} f_2.$$

For the converse implication, let $f = f_1 \Box f_2$ be finite at x, i.e. $(x, f(x)) \in \operatorname{epi}(f_1 \Box f_2) = \operatorname{epi} f_1 + \operatorname{epi} f_2$. Then, there exist $(x_i, \alpha_i) \in \operatorname{epi} f_i$ $(i = 1, 2)$ with $(x, f(x)) = (x_1, \alpha_1) + (x_2, \alpha_2)$ and thus, $f(x) = \alpha_1 + \alpha_2 \geq f_1(x_1) + f_2(x_2)$. However, since we also have $f(x) \leq f_1(x_1) + f_2(x_2)$ by the definition of the inf-convolution, we have equality and see that the infimum is attained at $x = x_1 + x_2$.

Finally, let f_1, f_2 be convex. By Proposition 5.3, we know that $\operatorname{epi} f_1 + \operatorname{epi} f_2$ is a convex set. It remains to note that

$$(f_1 \Box f_2)(x) = \inf\{\alpha \in \mathbb{R} \mid (x, \alpha) \in \operatorname{epi} f_1 + \operatorname{epi} f_2\}$$

and Theorem 7.2 on inf-projections shows that this is convex function. $\square$

Inf-convolutions are also not always as nice as we would like them to be:

Example 7.7.
1. The functions $f_1(x) = px$ $(p \in \mathbb{R})$ and $f_2(x) = \exp(x)$ are proper, convex, and continuous. Their inf-convolution is

   $$(f_1 \Box f_2)(u) = \begin{cases} p(u - \log(p))p + p, & \text{if } p > 0, \\ 0, & \text{if } p = 0, \\ -\infty, & \text{if } p < 0. \end{cases}$$

   Hence, for $p < 0$, the inf-convolution is not proper. Moreover, for $p = 0$

   $$\operatorname{epi} f_1 = \mathbb{R} \times [0, \infty[, \quad \operatorname{epi} f_2 = \{(x, \alpha) \mid \exp(x) \leq \alpha\}, \quad \operatorname{epi}(f_1 \Box f_2) = \mathbb{R} \times [0, \infty[$$

   and hence, $\operatorname{epi} f_1 + \operatorname{epi} f_2 \subset \operatorname{epi}(f_1 \Box f_2)$ with strict inclusion.

2. Consider

   $$C_1 := \{(x, y) \mid y \geq \exp(x)\}, \quad C_2 := \{(x, y) \mid y \geq \exp(-x)\}$$

   which are both non-empty, closed, convex sets in $\mathbb{R}^2$ and hence, their indicator functions i_{C_1} and i_{C_2} are proper, convex and lsc. The Minkowski sum of C_1 and C_2 is

   $$C_1 + C_2 = \mathbb{R} \times]0, \infty[$$

   which is not closed and hence, their inf-convolution $i_{C_1} \Box i_{C_2} = i_{C_1 + C_2}$ is not lsc.

$\triangle$

Interestingly, inf-convolutions are used in image processing. Here we do inf-convolution for arbitrary (bounded) functions and do not assume any convexity.

Example 7.8 (Morphological operations). A function $f : \mathbb{R}^d \to \mathbb{R}$ is considered as a gray scale image, i.e. $f(x)$ denotes the gray value of the image at location x. More precisely, one would deal with function $f : A \to [0,1]$ where A is the image domain (e.g. a two-dimensional domain such as a rectangle) and the value $f(x) = 0$ stands for the color black at position x while $f(x) = 1$ stands for white.

Let $B \subset \mathbb{R}^d$ be compact with $B = -B$. For a function $f : \mathbb{R}^d \to \mathbb{R}$ we define the *erosion* of f with *structure element* B as

$$\epsilon_B f(x) = \inf_{y \in B} f(x+y) = (f \,\square\, i_B)(x)$$

and there is corresponding *dilation* defined as $\delta_B f(x) = \sup_{y\in B} f(x+y)$. If f represents an image, the erosion shrinks the bright areas in the image, while the dilation expands them.

The operation

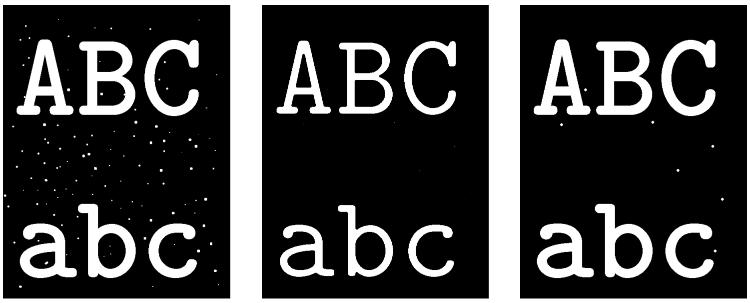
$$O_B f = \delta_{-B}(\epsilon_B f)(x)$$

is called *opening* of f and

$$C_B f = \epsilon_{-B}(\delta_B f)(x)$$

is called *closing* of f and $\triangle$

Here is an example of an image which is degraded by white spots on dark backgroud. The structure element is a circle (can you guess its radius?). The opening of the image eliminates the white spots but keeps the main structure (the letters) almost unchanged.



origrinal eroded opened

Another example is the elimination of an uneven background: To that end one calculates an image which does not contain any of the foreground structure (in this example, all dark structures should be eliminated) by a closing with a large structure element and then subtracts this background image from the original:

original background corrected

8 Proximal mappings and Moreau-Yoshida regularization

Definition 8.1. For a proper, convex and lsc function $f : \mathbb{R}^d \to \bar{\mathbb{R}}$ and $\lambda > 0$ we define the *proximal mapping* $\operatorname{prox}_{\lambda f} : \mathbb{R}^d \to \mathbb{R}^d$ by

$$\operatorname{prox}_{\lambda f}(x) := \operatorname*{argmin}_{y \in \mathbb{R}^d} \tfrac{1}{2}\|x - y\|_2^2 + \lambda f(y).$$

The *Moreau envelope* (also called *Moreau-Yoshida regularization*) ${}^\lambda f$ is

$$ {}^\lambda f(x) := \min_{y \in \mathbb{R}^d} \tfrac{1}{2\lambda}\|x - y\|_2^2 + f(y) = \left(\tfrac{1}{2\lambda}\|\cdot\|_2^2 \,\square\, f\right)(y).$$

Note that we could have defined the proximal mapping differently by taking the argmin of $\frac{1}{2\lambda}\|x - y\|_2^2 + f(y)$, i.e. the proximal mapping maps x to the point where the minimum in the Moreau envelope is assumed.

A direct consequence of the defintion is that

$$ {}^\lambda f(x) = \tfrac{1}{2\lambda}\|x - \operatorname{prox}_{\lambda f}(x)\|_2^2 + f(\operatorname{prox}_{\lambda f}(x)).$$

Since the inf convolution of convex functions is convex, the Moreau envelope is convex, but we still need to show that both the Moreau envelope and the proximal mapping are well defined (for the first we need to show that a minimizer always exist and this will, in turn, show that the minimium for the Moreau envelope is actually assumed).

Theorem 8.2. *For $f : \mathbb{R}^d \to \bar{\mathbb{R}}$ proper, convex and lsc and $\lambda > 0$ it holds that:*

1. *For every x there is exists a unique minimizer $\hat{x} = \operatorname{prox}_{\lambda f}(x)$ of $\frac{1}{2}\|x - y\|_2^2 + \lambda f(y)$.*
2. *This $\hat{x}$ is characterized by the variational inequality*
$$\forall y \in \mathbb{R}^d : \ \langle x - \hat{x}, y - \hat{x}\rangle + \lambda(f(\hat{x}) - f(y)) \leq 0.$$
3. *Some x^* is a minimizer of f if and only if x^* is a fixed point of $\operatorname{prox}_{\lambda f}$ for any $\lambda > 0$, i.e. $x^* = \operatorname{prox}_{\lambda f}(x^*)$.*
4. *The Moreau envelope ${}^\lambda f$ is differentiable with gradient*
$$\nabla({}^\lambda f)(x) = \tfrac{1}{\lambda}(x - \operatorname{prox}_{\lambda f}(x)).$$
5. *It holds that*
$$\operatorname*{argmin}_x f(x) = \operatorname*{argmin}_x {}^\lambda f(x).$$

Proof. 1. The objective in the definition of the proximal mapping is strictly convex and hence, has at most one minimizer. Let us now show existence of a minimizer: By Proposition 5.4 there exists an affine lower bound for f, i.e. we have $f(y) \geq \langle p, y\rangle + \alpha$ for some p and α. Hence, we have

$$\begin{aligned} g_x(y) := \tfrac{1}{2}\|y - x\|_2^2 + \lambda f(y) &\geq \tfrac{1}{2}\|y - x\|_2^2 + \lambda\langle p, y\rangle + \lambda\alpha \\ &= \tfrac{1}{2}\|y - x + \lambda p\|_2^2 + \lambda\langle p, x\rangle - \tfrac{\lambda^2}{2}\|p\|^2 + \lambda\alpha. \end{aligned}$$

Hence, the function on the left is coercive and since it also lsc, we see by Theorem 6.5, that minimizers exist.

We could also have observed that g_x is strongly convex and use an exercise from Sheet 4 to deduce coercivity.

2. If $\hat{x} = \operatorname{prox}_{\lambda f}(x)$, then for every $\mu \in [0,1]$ and y

$$g_x(\hat{x}) \leq g_x(\hat{x} + \mu(y - \hat{x}))$$

which we rearrange to

$$\begin{aligned}0 &\geq \lambda\big(f(\hat{x}) - f(\hat{x} + \mu(y-\hat{x}))\big) + \tfrac{1}{2}\big(\|\hat{x}-x\|_2^2 - \|\hat{x} - x + \mu(y-\hat{x})\|_2^2\big)\\ &\geq \lambda\big(f(\hat{x}) - f(\hat{x} + \mu(y-\hat{x}))\big) + \tfrac{1}{2}\big(-\mu^2\|y-\hat{x}\|_2^2 - 2\mu\langle \hat{x}-x, y-\hat{x}\rangle\big)\end{aligned}$$

By convexity of f we get

$$0 \geq \lambda\mu\big(f(\hat{x}) - f(y)\big) + \mu\langle \hat{x}-x, \hat{x}-y\rangle - \tfrac{\mu^2}{2}\|\hat{x}-y\|_2^2$$

Dividing by μ and letting $\mu \to 0$ shows the claim.

Conversely, let the variational inequality be fulfilled. This means that for all $y \in \mathbb{R}^d$

$$\begin{aligned}\lambda f(\hat{x}) + \tfrac{1}{2}\|\hat{x}-x\|_2^2 &\leq \lambda f(y) - \langle \hat{x}-x, \hat{x}-y\rangle + \tfrac{1}{2}\|\hat{x}-x\|_2^2\\ &\leq \lambda f(y) + \tfrac{1}{2}\|\hat{x}-y\|_2^2 - \langle \hat{x}-x, \hat{x}-y\rangle + \tfrac{1}{2}\|\hat{x}-x\|_2^2\\ &= \lambda f(y) + \tfrac{1}{2}\|y-x\|_2^2\end{aligned}$$

and this shows $\hat{x} = \operatorname{prox}_{\lambda f}(x)$.

3. If $\hat{x}$ is a minimizer of f, then $f(\hat{x}) \leq f(x)$ for all x and hence $\lambda f(\hat{x}) + \frac{1}{2}\|\hat{x}-\hat{x}\|_2^2 \leq \lambda f(x) + \frac{1}{2}\|x-\hat{x}\|_2^2$ which shows that $\hat{x} = \operatorname{prox}_{\lambda f}(\hat{x})$.

   Conversely, if $\hat{x} = \operatorname{prox}_{\lambda f}(\hat{x})$, we get, by 2., that for all y

$$\langle \hat{x}-\hat{x}, y-\hat{x}\rangle + \lambda\big(f(\hat{x}) - f(y)\big) \leq 0$$

   and this shows $f(\hat{x}) \leq f(y)$ for all y.

4. For $x_0 \in \mathbb{R}^d$ define $\hat{x}_0 := \operatorname{prox}_{\lambda f}(x_0)$ and $z := \frac{x_0 - \hat{x}_0}{\lambda}$. To show that ${}^\lambda f$ is differentiable at x_0 with $({}^\lambda f)'(x_0) = z$ we need to show that

$$r(u) := {}^\lambda f(x_0 + u) -{}^\lambda f(x_0) - \langle z, u\rangle = o(\|u\|_2).$$

   By the definition of the prox and the Moreau envelope, we have

$$\begin{aligned}{}^\lambda f(x_0) &= f(\hat{x}_0) + \tfrac{1}{2\lambda}\|\hat{x}_0 - x_0\|_2^2\\ {}^\lambda f(x_0+u) &= \min_x \Big[f(x) + \tfrac{1}{2\lambda}\|x - x_0 - u\|_2^2\Big]\\ &\leq f(\hat{x}_0) + \tfrac{1}{2\lambda}\|\hat{x}_0 - x_0 - u\|_2^2.\end{aligned}$$

   It follows (plugging in z)

$$r(u) \leq \tfrac{1}{2\lambda}\|\hat{x}_0 - x_0 - u\|_2^2 - \tfrac{1}{2\lambda}\|\hat{x}_0 - x_0\|_2^2 - \tfrac{1}{\lambda}\langle x_0 - \hat{x}_0, u\rangle = \tfrac{1}{2\lambda}\|u\|_2^2.$$

Using the convexity of r, we get

$$0 = r(0) = r(\tfrac{1}{2}u + \tfrac{1}{2}(-u)) \leq \tfrac{1}{2}(r(u) + r(-u))$$

and this shows

$$r(u) \geq -r(-u) \geq -\tfrac{1}{2\lambda}\|-u\|_2^2$$

and in total we get $|r(u)| \leq \frac{1}{2\lambda}\|u\|_2^2$ which proves the claim.

5. We know by 3. that $\hat{x}$ is a minimizer of f if and only if $\hat{x} = \operatorname{prox}_{\lambda f}(\hat{x})$ and by 4. this is equivalent to $\nabla({}^{\lambda}f)(\hat{x}) = 0$, i.e. exactly when $\hat{x}$ is a minimizer of ${}^{\lambda}f$.

□

Example 8.3 (Prox of indicators are projections)*.* Let $f(y) = i_C(y)$ for some non-empty, convex and closed C. Then (note that $\lambda i_C = i_C$)

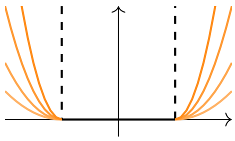
$$\operatorname{prox}_{i_C}(x) = \operatorname*{argmin}_{y\in\mathbb{R}^d} \tfrac{1}{2}\|x-y\|_2^2 + \lambda i_C(y) = \operatorname*{argmin}_{y\in C}\|x-y\| = P_C(x),$$

i.e. $\operatorname{prox}_{\lambda f}(x)$ is the orthogonal projection of x onto C (indepently of $\lambda > 0$).

The Moreau envelope is

$${}^{\lambda}i_C(x) = \tfrac{1}{2\lambda}\|x - P_C(x)\|_2^2 = \tfrac{1}{2\lambda}d(x,C)^2.$$

△



$\lambda = 1, \frac{1}{2}, \frac{1}{4}, \frac{1}{8}$

Example 8.4 (Prox of $|\cdot|$ is soft shrinkage)*.* Now consider $f(x) = |x|$. To calculate $\hat{x} := \operatorname{argmin}_y \frac{1}{2}(x-y)^2 + \lambda|y|$ we first observe that if $x \geq 0$, then $\hat{x} \geq 0$ and also $\operatorname{prox}_{\lambda f}(-x) = -\operatorname{prox}_{\lambda f}(x)$. Hence, we only consider $x > 0$ and have

$$\hat{x} = \operatorname*{argmin}_{y\geq 0} \tfrac{1}{2}(x-y)^2 + \lambda y.$$

We note that the unconstrained minimizer is $x - \lambda$, and if $0 \leq x \leq \lambda$, then the minimizer is 0. In total, we get that

$$\begin{aligned}\operatorname{prox}_{\lambda|\cdot|}(x) &= \begin{cases} x-\lambda, & \text{if } x > \lambda,\\ 0, & \text{if } |x| \leq \lambda,\\ x+\lambda, & \text{if } x < -\lambda,\end{cases}\\ &= \max(|x|-\lambda, 0)\operatorname{sign}(x) =: S_\lambda(x).\end{aligned}$$

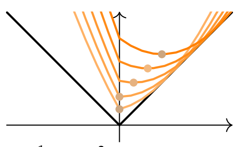
This function is known as the *soft thresholding* or *soft shrinkage* function.

To calculate the Moreau envelope of $|\cdot|$ we just plug in and get

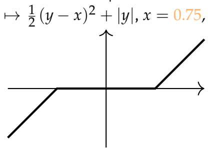
$$\begin{aligned}{}^{\lambda}f(x) &= \tfrac{1}{2\lambda}(x - S_\lambda(x))^2 + |S_\lambda(x)|\\ &= \begin{cases} |x| - \frac{\lambda}{2}, & \text{if } |x| > \lambda\\ \frac{1}{2\lambda}x^2, & \text{if } |x| \leq \lambda.\end{cases}\end{aligned}$$

This function is called *Huber function.*

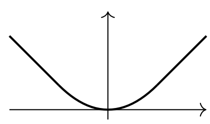
△



$y \mapsto \frac{1}{2}(y-x)^2 + |y|,\ x = 0.75, 1, 1.25, 1.5, 1.75$



Prox: $S_\lambda(x) = \operatorname{prox}_{|\cdot|}(x)$



Moreau envelope: ${}^{1}|\cdot|$

Lemma 8.5 (Calculus for proximal mappings). *For proper, convex and lsc functions f and $\lambda > 0$ it holds that:*

i) If $g(x) = f(x) + \alpha$ for $\alpha \in \mathbb{R}$, then $\operatorname{prox}_{\lambda g} = \operatorname{prox}_{\lambda f}$.

ii) If $g(x) = \tau f(\mu x)$ for $\tau > 0, \mu \in \mathbb{R}$, then $\operatorname{prox}_{\lambda g}(x) = \frac{1}{\mu}\operatorname{prox}_{\lambda \mu^2 \tau f}(\mu x)$.

iii) If $g(x) = f(x + x_0) + \langle p, x\rangle$, for $x_0, p \in \mathbb{R}^d$, then $\operatorname{prox}_{\lambda g}(x) = \operatorname{prox}_{\lambda f}(x + x_0 + \lambda p) - x_0$.

iv) If $g(x) = f(Qx)$ for orthonormal $Q \in \mathbb{R}^{d\times d}$, then $\operatorname{prox}_{\lambda g}(x) = Q^T \operatorname{prox}_{\lambda f}(Qx)$.

v) If $h(x, y) = f(x) + g(y)$ for proper, convex and lsc g, then $\operatorname{prox}_{\lambda h}(x, y) = (\operatorname{prox}_{\lambda f(x), \operatorname{prox}_{\lambda g}(y)}$.

Proof. The first three items are straightforward implications from the definition by appropriate substitutions. For item iv) we write

$$\operatorname{prox}_{\lambda g}(x) = \operatorname*{argmin}_y \tfrac{1}{2}\|x - y\|_2^2 + \lambda f(Qy)$$

and use that Q^T is onto by assumption. Hence, we can substutite $y = Q^T z$, minimize over z, but keep in mind that we are interested in the *minimizer*. Hence, we have, using $QQ^T = I$

$$\begin{aligned}\operatorname{prox}_{\lambda g}(x) &= Q^T \operatorname*{argmin}_z \tfrac{1}{2}\|x - Q^T z\|_2^2 + \lambda f(z)\\ &= Q^T \operatorname*{argmin}_z \tfrac{1}{2}\|Qx - z\|_2^2 + \lambda f(z)\end{aligned}$$

where we used that $\|Qu\|_2 = \|u\|_2$ since Q is orthonormal. For item v) just observe that the minimization can be carried out independently over x and y.

□

9 Subgradients

If you've found it surprising that convex functions automatically have some continuity property, you may find it even more surprising, that they also have some inbuild differentiability.

We recall the notion of directional derivative: For $f : \mathbb{R}^d \to \mathbb{R}$, $x_0, v \in \mathbb{R}^d$ we denote by

$$Df(x_0, v) := \lim_{h \searrow 0} \tfrac{f(x_0+hv)-f(x_0)}{h}$$

the *one-sided directional derivative* (whenever the limit exists).

Recall that in the case of differentiable f, it holds that $Df(x_0, v) = \langle \nabla f(x_0), v\rangle$, but that directional differentiability in all directions does not imply differentiability.

Definition 9.1. A function $f : \mathbb{R}^d \to \bar{\mathbb{R}}$ is *positively homogeneous* (of degree 1) if

$$\lambda > 0 \implies f(\lambda x) = \lambda f(x).$$

Positive homogeneity of degree p would mean $f(\lambda x) = \lambda^p f(x)$, but we will not need this notion.

Norms and semi-norms are obvious examples of positive homogeneous functions as well a linear functions.

Theorem 9.2. *Let $f : \mathbb{R}^d \to \bar{\mathbb{R}}$ be convex, $x_0 \in \operatorname{dom}(f)$, $v \in \mathbb{R}^d$. Then, the limit in the definition of the directional derivative exists in $\bar{\mathbb{R}}$ and it holds*

$$Df(x_0, v) = \inf_{h>0} \tfrac{f(x_0+hv)-f(x_0)}{h}.$$

Moreover, the function $Df(x_0, \cdot)$ is convex and positvely homogeneous and for any v it holds that

$$Df(x_0, v) \leq f(x_0 + v) - f(x_0)$$

and if $f(x_0 - v) \neq -\infty$ it also holds that

$$f(x_0) - f(x_0 - v) \leq Df(x_0, v).$$

In particular, $Df(x_0, v)$ is finite for all v if $x_0 \in \operatorname{int}\operatorname{dom}(f)$.

Proof. The proof relies on the following monotonicity property of the difference quotient: For $0 < h_1 < h_2$ it holds that

$$\tfrac{f(x_0+h_1 v)-f(x_0)}{h_1} \leq \tfrac{f(x_0+h_2 v)-f(x_0)}{h_2}.$$

To see this first note that if $x_0 + h_2 v \notin \operatorname{dom} f$, the inequality is always fulfilled. Hence, we assume that $x_0 + h_2 v \in \operatorname{dom} f$. By convexity of f we get

$$\begin{aligned} f(x_0 + h_1 v) - f(x_0) &= f\big(\tfrac{h_1}{h_2}(x_0 + h_2 v) + (1 - \tfrac{h_1}{h_2})x_0\big) - f(x_0) \\ &\leq \tfrac{h_1}{h_2} f(x_0 + h_2 v) + (1 - \tfrac{h_1}{h_2}) f(x_0) - f(x_0) \\ &= \tfrac{h_1}{h_2}\big(f(x_0 + h_2 v) - f(x_0)\big), \end{aligned}$$

and this shows the desired monotonicity.

The monotonicity implies that the limit in the definition of the directional derivative is actually and infimum and, moreover, $Df(x_0, v) \leq f(x_0 + v) - f(x_0)$ follows by taking $h = 1$.

To show the lower inequality, let v be such that $f(x_0 - v) > -\infty$. If $f(x_0 - v) = \infty$ or $Df(x_0, v) = \infty$, there is nothing to prove. Hence, assume that both quantities are finite. Since $\infty > Df(x_0, v) = \lim_{h\to 0}(f(x_0 + hv) - f(x_0))/h$, there is an $\bar{h} > 0$ such that $\frac{f(x_0+hv)-f(x_0)}{h}$ is finite for all $h \in \,]0, \bar{h}]$. Hence, $x_0 - v, x_0 + hv \in \operatorname{dom} f$ for these h and we get, again using convexity,

$$\begin{aligned} f(x_0) &= f\Big(\tfrac{1}{1+h}(x_0 + hv) + \tfrac{h}{1+h}(x_0 - v)\Big) \\ &\leq \tfrac{1}{1+h} f(x_0 + hv) + \tfrac{h}{1+h} f(x_0 - v). \end{aligned}$$

We rearrange to

$$f(x_0) - f(x_0 - v) \leq \tfrac{f(x_0+hv)-f(x_0)}{h}$$

and the claim follows for $h \searrow 0$.

Now we show that $Df(x_0, \cdot)$ is convex: For $v_1, v_2 \in \operatorname{dom} Df(x_0, \cdot)$ we have (for small enough h and $\lambda \in [0, 1]$)

$$\begin{aligned} f(x_0 + h(\lambda v_1 + (1-\lambda)v_2)) - f(x_0) &= f(\lambda(x_0 + hv_1) + (1-\lambda)(x_0 + hv_2)) - \lambda f(x_0) - (1-\lambda) f(x_0) \\ &\leq \lambda(f(x_0 + hv_1) - f(x_0)) + (1-\lambda)(f(x_0 + hv_2) - f(x_0)) \end{aligned}$$

Dividing by $h > 0$ and $h \searrow 0$ shows the desired convexity.

To show that $Df(x_0, \cdot)$ is positively homogeneous, we observe for $\lambda > 0$

$$\begin{aligned} Df(x_0, \lambda v) &= \lim_{h\searrow 0} \tfrac{f(x_0+\lambda hv)-f(x_0)}{h} \\ &= \lim_{\bar{h}\searrow 0} \lambda \tfrac{f(x_0+\bar{h}v)-f(x_0)}{\bar{h}} && (\bar{h} = \lambda h) \\ &= \lambda Df(x_0, v) \end{aligned}$$

as desired.

Finally, we show that $Df(x_0, v)$ is finite for $x_0 \in \operatorname{int}\operatorname{dom} f$ and all v: For every v there is $\lambda > 0$ such that $x_0 \pm \lambda v \in \operatorname{int}\operatorname{dom} f$ and thus

$$\tfrac{1}{\lambda}(f(x_0) - f(x_0 - \lambda v)) \leq \underbrace{\tfrac{1}{\lambda} Df(x_0, \lambda v)}_{=Df(x_0, v)} \leq \tfrac{1}{\lambda}(f(x_0 + \lambda v) - f(x_0))$$

and the claim follows since the leftmost and rightmost expressions are finite. □

Now recall that by Theorem 5.1 we know that a differentiable and convex f fulfills that for every x_0 it holds that

$$\forall x: \; f(x) \geq f(x_0) + \langle \nabla f(x_0), x - x_0 \rangle.$$

Definition 9.3. Let $f : \mathbb{R}^d \to \bar{\mathbb{R}}$ be convex and $x_0 \in \operatorname{dom} f$. Then $p \in \mathbb{R}^d$ is a *subgradient* of f at x_0 if

$$\forall x \in \mathbb{R}^d : \ f(x) \geq f(x_0) + \langle p, x - x_0 \rangle.$$

The set of all subgradients of f at x_0 is denoted by $\partial f(x_0)$ and called *subdifferential* of f at x_0.

If $x_0 \notin \operatorname{dom} f$ we set $\partial f(x_0) = \emptyset$. Furthermore we denote

$$\operatorname{dom} \partial f = \{x \mid \partial f(x) \neq \emptyset\}$$

and f is called *subdifferentiable* at x_0 if $x_0 \in \operatorname{dom} \partial f$.

A direct consequence of the definition is that $\partial f(x_0)$ is always closed and convex.

If $p_n \in \partial f(x)$ with $p_n \to p$, then $f(y) \geq f(x) + \langle p_n, y - x \rangle$ for all y and passing to the limit $p_n \to p$ shows that $p \in \partial f(x)$. If $p, q \in \partial f(x)$, then $f(y) = \lambda f(y) + (1-\lambda) f(y) \geq \lambda f(x) + \lambda \langle p, x - y \rangle + (1-\lambda) f(x) + (1-\lambda)\langle q, y - x \rangle = f(x) + \langle \lambda p + (1-\lambda) q, y - x \rangle$, i.e. $\lambda p + (1-\lambda) q \in \partial f(x)$.

Graphically, some p is a subgradient of f at x_0 if the function $x \mapsto f(x_0) + \langle p, x - x_0 \rangle$ is a supporting affine lower bound which is exact at x_0.

The nice thing about subgradient is, that they even exist where a convex function has a kink, and in this case there is more than one of them:

Example 9.4 (Subgradients of the absolute value). The function $f(x) = |x|$ on $\mathbb{R}$ fulfills

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

which can be seen by direct inspection. △

Proposition 9.5 (The subdifferential and directional derivatives). *Let $f : \mathbb{R}^d \to \bar{\mathbb{R}}$ be convex and $f(x_0) \in \mathbb{R}$. Then it holds that*

$$\partial f(x_0) = \{p \in \mathbb{R}^d \mid \forall v \in \mathbb{R}^d : \langle p, v \rangle \leq Df(x_0, v)\}. \qquad (*)$$

Proof. Denote the set on the right hand side of (*) by M.

$\partial f(x_0) \subset M$**:** Let $p \in \partial f(x_0)$ and set $x = x_0 + hv$ for $h > 0$. Then, by definition of the subdifferential:

$$f(x_0 + hv) - f(x_0) \geq \langle p, hv \rangle$$

and division by h and $h \searrow 0$ shows $p \in M$.

$\partial f(x_0) \supset M$**:** Now let $p \in M$. Theorem 9.2 shows that for all $v \in \mathbb{R}^d$.

$$\langle p, v \rangle \leq Df(x_0, v) \leq f(x_0 + v) - f(x_0)$$

and the claim follows if we set $x = x_0 + v$

□

Subgradients have some properties similar to gradients:

Proposition 9.6. *Let $f : \mathbb{R}^d \to \bar{\mathbb{R}}$ be convex. Then it holds:*

i) $\partial\varphi(x) = \partial f(x + x_0)$ *for* $\varphi(x) := f(x + x_0)$.

ii) $\partial\varphi(x) = \lambda\partial f(\lambda x)$ *for* $\varphi(x) := f(\lambda x)$.

iii) $\partial\varphi(x) = \lambda\partial f(x)$ *for* $\varphi(x) := \lambda f(x)$.

These are straightforward consequences of the definition and one should do the proofs as exercises.

Proposition 9.7. *A convex function $f : \mathbb{R}^d \to \bar{\mathbb{R}}$ is differentiable at x_0 if a only if f is continuous at x_0 and $\partial f(x_0)$ only has one element and in this case we have $\partial f(x_0) = \{\nabla f(x_0)\}$.*

Proof. Let f be differentiable at x_0. By the previous lemma we have that $p \in \partial f(x_0)$ fulfills $\langle p, v\rangle \leq \langle \nabla f(x_0), v\rangle$ for all v which implies $p = \nabla f(x_0)$.

To show the converse, we assume without loss of generality that $x_0 = 0$ and $\partial f(0) = \{w\}$. We define the convex function $F_v(t) = f(tv)$. Similarly to Proposition 9.6 ii) we see that $\partial F_v(0) = \{\langle w, v\rangle\}$ which implies for $t > 0$

$$0 \leq \tfrac{F_v(t) - F_v(0)}{t} - \langle w, v\rangle.$$

On the other hand, for $\epsilon > 0$ there exists t_ϵ such that $F_v(t_\epsilon) < F_v(0) + t_\epsilon\langle w, v\rangle + t_\epsilon\epsilon$ (note that $\langle w, v\rangle + \epsilon \notin \partial F_v(0)$). Convexity of F_v shows that for every $t \in [0, t_\epsilon]$ it holds that

$$\begin{aligned} F_v(t) &\leq \tfrac{t}{t_\epsilon} F_v(t_\epsilon) + \tfrac{t_\epsilon - t}{t_\epsilon} F_v(0) \\ &\leq F(0) + t\langle w, v\rangle + t\epsilon, \end{aligned}$$

and hence,

$$\tfrac{F_v(t) - F_v(0)}{t} - \langle w, v\rangle \leq \epsilon.$$

Since $\epsilon > 0$ was arbitrary, it follows that

$$\lim_{t \searrow 0} \tfrac{F_v(t) - F_v(0)}{t} = \langle w, v\rangle$$

which proves differentiablity of f and $\nabla f(0) = w$. □

Example 9.8. 1. We begin with the very simple example $f(x) = |x|$ on the real line. In one dimension, we can charactarize the subdifferential completely by left- and right-derivatives, since these are the directional derivatives $Df(x, -1)$ and $Df(x, 1)$, respectively. For the absolute value we get

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

2. Now consider the function

$$f(x) = \begin{cases} \infty, & \text{if } x < 0 \text{ or } x > 1 \\ -\sqrt{x}, & \text{if } 0 \leq x \leq 1. \end{cases}$$

We get

$$\partial f(x) = \begin{cases} \emptyset, & \text{if } x \leq 0, \\ \{-\frac{1}{2\sqrt{x}}\}, & \text{if } 0 < x < 1, \\ [-\frac{1}{2}, \infty[, & \text{if } x = 1, \\ \emptyset, & \text{if } x > 1. \end{cases}$$

Note that the subdifferential $\partial f(x_0)$ can be empty or unbounded if $x_0 \notin \operatorname{int} \operatorname{dom} f$. The next theorem shows that none of this can happen in the interior of the domain.

$\triangle$

10 Subdifferential calculus

Proper and convex functions always a subgradients in the interior of their domain.

Theorem 10.1. *If $f:\mathbb{R}^d \to \bar{\mathbb{R}}$ be proper and convex. Then it holds that $\partial f(x_0)$ is non-empty and bounded if $x_0 \in \operatorname{int}\operatorname{dom} f$.*

There is sharper result, namely that the subdifferential $\partial f(x_0)$ is non-empty and bounded if and only if $x_0 \in \operatorname{ri}\operatorname{dom} f$, but we will not prove this here.

Proof. If $x_0 \in \operatorname{int}\operatorname{dom} f$, then f is locally Lipschitz-continuous at x_0 (Lemma 6.1), hence, there is $\delta > 0$ such that $|f(x) - f(x_0)| < 1$ for $|x - x_0| < \delta$. In particular, for some $p \in \partial f(x_0)$ and all x with $\|x\|_2 < 1$ we have

$$1 > f(x_0 + \delta x) - f(x_0) \geq \langle p, \delta x\rangle.$$

Hence, $\langle p, x\rangle \leq 1/\delta$, and taking the supremum over all x with $\|x\|_2 < 1$ shows $\|p\|_2 \leq 1/\delta$.

To show that $\partial f(x_0)$ is non-empty, note that $\operatorname{int}(\operatorname{epi} f)$ is not empty (since it contains an open set of the form $B_\delta(x_0) \times]f(x_0), \infty[$). Moreover $(x_0, f(x_0))$ is not in $\operatorname{int}(\operatorname{epi} f)$ and hence we can, by Proposition 3.7, find $(p_0, t) \in \mathbb{R}^d \times \mathbb{R}$ for fulfills $(p_0, t) \neq (0, 0)$ and

$$\langle p_0, x\rangle + ts \leq \lambda \quad \text{if } x \in \operatorname{dom} f, f(x) \leq s \quad \text{and} \quad \langle p_0, x_0\rangle + tf(x_0) \geq \lambda.$$

With $x = x_0$ and $s = f(x_0)$ we see that $\lambda = \langle p_0, x_0\rangle + tf(x_0)$. Moreover, we see that $t < 0$, since $t > 0$ would lead to a contradiction by sending $t \to \infty$ and $t = 0$ would imply $\langle p_0, x - x_0\rangle \leq 0$ for all $x \in B_\delta(x_0)$ and this would imply $p_0 = 0$ which would contradict $(p_0, t) \neq 0$. With $p = -p_0/t$ we get

$$f(x_0) + \langle p, x - x_0\rangle \leq f(x_0)$$

for all $x \in \operatorname{dom} f$ and this shows $p \in \partial f(x_0)$. $\square$

Derivatives can be used to find local minima of functions. For convex functions, this is accomplished by subgradients. However, due to convexity, the first order condition is necessary and sufficient.

Theorem 10.2 (Fermat's rule)**.** *Let $f:\mathbb{R}^d \to \bar{\mathbb{R}}$ be proper and convex. Then it holds that $\hat{x}$ is a global minimizer of f if and only if $0 \in \partial f(\hat{x})$.*

Proof. Let $\hat{x}$ be a global minimizer. Hence, we have for all x

$$\begin{aligned} f(x) &\geq f(\hat{x}) \\ &= f(\hat{x}) + \langle 0, x - \hat{x}\rangle \end{aligned}$$

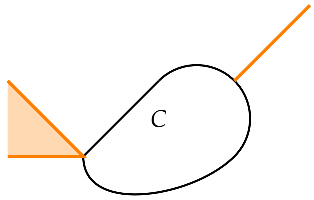
and this shows $0 \in \partial f(\hat{x})$. Conversely, if $0 \in \partial f(\hat{x})$, then the above subgradient inequality holds for all x and hence, $\hat{x}$ is a global minimizer. $\square$

We close this secton with another important example:

Example 10.3. Let C be convex and non-empty. Then the subdifferential of the indicator function of C is given by

$$\partial i_C(x_0) = \begin{cases} \{p \in \mathbb{R}^d \mid \forall x \in C : \langle p, x - x_0\rangle \leq 0\}, & \text{if } x_0 \in C \\ \emptyset, & \text{if } x_0 \notin C \end{cases}$$

First, we observe, that for $x_0 \in \operatorname{int} C$, it holds that $\partial i_C(x_0) = \{0\}$ (as the function is locally constant there). If $x \in \partial C$, then there is geometric meaning of the subgradients: A vector p is a subgradient, if the angle between p and the line from x_0 to any point $x \in C$ is larger than 90°. We have seen this property already and indeed it holds that the subgradient of the indicator is in fact the normal cone, i.e.



$$\partial i_C(x) = N_C(x).$$

△

However, the situation is a bit more subtle for the sum-rule and the concatenation with general linear operators:

Theorem 10.4. *i) Let $f, g : \mathbb{R}^d \to \bar{\mathbb{R}}$ be proper and convex. Then it holds for every x that*

$$\partial f(x) + \partial g(x) \subset \partial(f + g)(x).$$

Equality holds if there exists some $\bar{x}$ such that $\bar{x} \in \operatorname{dom}(f) \cap \operatorname{dom}(g)$ and f is continuous at $\bar{x}$.

Note that the point $\bar{x}$ is totally unrelated to the x in the assertion and that only one of the two functions need to be continuous at $\bar{x}$.

ii) Let $f : \mathbb{R}^d \to \bar{\mathbb{R}}$ be proper and convex and $A \in \mathbb{R}^{d\times n}$. Then $\varphi(x) := f(Ax)$ satisfies

$$A^T \partial f(Ax) \subset \partial\varphi(x)$$

for all x and equality holds if there exists $\bar{x}$ such that f is continuous and finite at $A\bar{x}$.

Proof. i) We start with the inclusion: Let $p \in \partial f(x) + \partial g(x)$, i.e. we have $p = q + r$ with $p \in \partial f(x)$, $r \in \partial g(x)$. Hence, we have for all y that

$$\begin{aligned} f(y) &\geq f(x) + \langle q, y - x\rangle \\ g(y) &\geq g(x) + \langle r, y - x\rangle \end{aligned}$$

and adding these inequalities shows that $p = q + r \in \partial(f + g)(x)$.

Now assume that f is continuous at some $\bar{x}$ with $\bar{x} \in \operatorname{dom}(f) \cap \operatorname{dom}(g)$ and it remains to show that $\partial(f + g)(x) \subset \partial f(x) + \partial g(x)$. If $f(x) = \infty$ or $g(x) = \infty$, then the inclusion is clear, since then $\partial(f + g)(x) = \emptyset$.

Hence, let $x \in \operatorname{dom}(f) \cap \operatorname{dom}(g)$. We have to prove that every $p \in \partial(f+g)(x)$ can be decomposed into $p = q + r$ with $q \in \partial f(x)$ and $r \in \partial g(x)$. The subgradient inequality for $p \in \partial(f+g)(x)$ is

$$f(y) + g(y) \geq f(x) + g(x) + \langle p, y - x\rangle$$

for all y. We rewrite this as

$$F(y) := f(y) - f(x) - \langle p, y - x\rangle \geq g(x) - g(y) =: G(y) \tag{*}$$

and consider the sets

$$C_1 := \{(y,\alpha) \in \mathbb{R}^d \times \mathbb{R} \mid \alpha \geq F(y)\} = \{(y,\alpha) \in \operatorname{dom}(f) \times \mathbb{R} \mid \alpha \geq F(y)\}$$
$$C_2 := \{(y,\alpha) \in \mathbb{R}^d \times \mathbb{R} \mid G(y) \geq \alpha\} = \{(y,\alpha) \in \operatorname{dom}(g) \times \mathbb{R} \mid G(y) \geq \alpha\}.$$

Since $F(x) = G(x) = 0$, both sets are non-empty and since f and g are convex (hence F is convex and G is concave), both sets are convex. By (*), the sets C_1 and C_2 only share boundary points. Hence, by the proper separation theorem (Theorem 3.9), there exists a nonzero (q,a) such that

$$\left\langle \begin{bmatrix} q \\ a \end{bmatrix}, \begin{bmatrix} y_2 \\ \alpha_2 \end{bmatrix} \right\rangle \leq \left\langle \begin{bmatrix} q \\ a \end{bmatrix}, \begin{bmatrix} y_1 \\ \alpha_1 \end{bmatrix} \right\rangle$$

for $(y_i, \alpha_i) \in C_i$, $i = 1, 2$.

We aim to show $a > 0$: Since $(x, \alpha_1) \in C_1$ for $\alpha_1 \geq 0$ and $(x, \alpha_2) \in C_2$ for $\alpha_2 \leq 0$, we have that $a\alpha_2 \leq a\alpha_1$. Hence, $a \geq 0$.

Now we show that $a > 0$, i.e. we only need to show that $a \neq 0$. If $a = 0$ would hold, than we would have

$$\langle q, y_2\rangle \leq \langle q, y_1\rangle, \quad \text{for } y_1 \in \operatorname{dom} F = \operatorname{dom} f,\ y_2 \in \operatorname{dom}(-G) = \operatorname{dom} g$$
$$\langle q, \bar{x}\rangle \leq \langle q, y_1\rangle \quad \text{for } y_1 \in \operatorname{dom} F = \operatorname{dom} f.$$

In particular, the continuity of f would imply that for $y_1 = \bar{x} \pm \Delta x \in \operatorname{dom}(f)$ (which holds for all Δx small enough), that $0 \leq \langle q, \pm\Delta x\rangle$ for all these Δx and this would imply $q = 0$ which contradicts $(q,a) \neq 0$.

Thus $a > 0$ and we can assume $a = 1$ without loss of generality. We have for $(y, F(y)) \in C_1$ and $(y, G(y)) \in C_2$ that whenever y is in $\operatorname{dom}(f)$ or $\operatorname{dom}(g)$, respectively, that

$$\langle q, y\rangle + F(y) \geq b \qquad\qquad \langle q, y\rangle + G(y) \leq b.$$

With $y = x \in \operatorname{dom}(f) \cap \operatorname{dom}(g)$, we get the equality $\langle q, x\rangle = b$ and we get

$$\forall y \in \operatorname{dom}(f):\ f(y) - f(x) - \langle p, y - x\rangle + \langle q, y\rangle \geq \langle q, x\rangle$$

which means

$$\forall y \in \operatorname{dom}(f):\ f(y) \geq f(x) + \langle p - q, y - x\rangle.$$

We conclude $p - q \in \partial f(x)$ and similarly, we get $q \in \partial g(x)$ which proves the claim.

ii) For the inclusion let $q = A^T p$ with $p \in \partial f(Ax)$ we have for all y that

$$\begin{aligned}\varphi(x) + \langle q, y - x\rangle &= f(Ax) + \langle A^T p, y - x\rangle \\ &= f(Ax) + \langle p, Ay - Ax\rangle \le f(Ay) = \varphi(y)\end{aligned}$$

and this shows that $q \in \partial\varphi(x)$.

For the equality, assume that $q \in \partial\varphi(x)$, i.e. for all y

$$f(Ax) + \langle q, y - x\rangle \le f(Ay).$$

We aim to squeeze a separating hyperplane into this inequality. We define

$$C_1 = \operatorname{epi} f, \quad C_2 = \{(Ay, f(Ax) + \langle q, y - x\rangle) \in \mathbb{R}^d \times \mathbb{R} \mid y \in \mathbb{R}^d\}.$$

We note that $\operatorname{int}(C_1)$ is not empty and that $\operatorname{int}(C_1) \cap C_2 = \emptyset$. Hence, by Theorem 3.9 we can find some $0 \ne (q^0, \alpha_0) \in \mathbb{R}^d \times \mathbb{R}$ such that

$$\begin{aligned}\langle q^0, \bar{y}\rangle + \alpha_0\alpha &\le \lambda \quad \forall \bar{y} \in \operatorname{dom} f, \alpha \ge f(\bar{y}) \\ \langle q^0, Ay\rangle + \alpha_0(f(Ax) + \langle q, y - x\rangle) &\ge \lambda, \quad \forall y \in \mathbb{R}^d.\end{aligned}$$

Again $\alpha_0 > 0$ can not occur ($\alpha \to \infty$ would lead to a contradiction) and $\alpha_0 \ne 0$ follows from the continuity of f at $A\bar{x}$.

If $\alpha_0 = 0$, then $\langle q^0, \bar{y}\rangle \le \langle q^0, Ay\rangle$ for all $\bar{y} \in \operatorname{dom} f$ and $y \in \mathbb{R}^d$. Hence we can choose $y = x \pm \Delta x$ and $\bar{y} = A(x \pm \Delta x)$ and would get $q^0 = 0$ as well, contradicting $(q^0, \alpha_0) \ne 0$.

If we set $\bar{y} = Ax$, $\alpha = f(\bar{y})$ and $y = x$ we conclude $\lambda = \langle q^0, Ax\rangle + \alpha_0 f(Ax)$. By the second inequality above we get

$$\langle q^0, Ay - Ax\rangle + \alpha_0\langle q, y - x\rangle \ge 0 \quad \forall y \in \mathbb{R}^d$$

and hence $q = -\frac{1}{\alpha_0} A^T q^0$. Setting $p = -\frac{1}{\alpha_0} q^0$ we get from the first inequality above, by rearranging and dividing by $\alpha_0 < 0$, that

$$\langle p, z - Ax\rangle + f(Ax) \le f(z), \quad \forall z \in \operatorname{dom} f$$

and this means that $p \in \partial f(Ax)$. Hence, $\partial\varphi(x) \subset A^T \partial f(Ax)$. ☐

11 Applications of subgradients

Example 11.1. Let's consider counterexamples to show that the sum-rule rule and the chain rule for linear maps are indeed not always true:

1. Let $f = i_{[0,\infty[}$ and

$$g(x) = \begin{cases} -\sqrt{-x}, & x \leq 0 \\ \infty, & x > 0. \end{cases}$$

The subdifferentials are

$$\partial f(x) = \begin{cases} \emptyset, & x < 0 \\ ]-\infty, 0], & x = 0 \\ \{0\}, & x > 0 \end{cases}, \quad \partial g(x) = \begin{cases} \{\frac{1}{2\sqrt{-x}}\}, & x < 0 \\ \emptyset, & x \geq 0. \end{cases}$$

Thus, the sum of the subdifferentials is always empty: $\partial f(x) + \partial g(x) = \emptyset$ for all x.

The sum of f and g, however, is

$$(f+g)(x) = i_{\{0\}}(x)$$

and hence, the subdifferential of the sum is

$$\partial(f+g)(x) = \begin{cases} \emptyset, & x \neq 0, \\ \mathbb{R}, & x = 0, \end{cases}$$

which is much larger.

2. For the chain rule for linear maps there is a very simple counterexample: Consider

$$f(x) = \begin{cases} \infty, & \text{if } x < 0, \\ -\sqrt{x}, & \text{if } x \geq 0, \end{cases}$$

with $\partial f(0) = \emptyset$ and $A = 0$ (the 1×1 zero-matrix). Then $\varphi(x) = f(Ax) \equiv f(0) = 0$, i.e. $\partial\varphi(0) = \{0\}$. Hence

$$\partial\varphi(0) = \{0\} \supsetneq \emptyset = A\partial f(A0).$$

△

Here is a positive result:

Corollary 11.2. *Let $f : \mathbb{R}^d \to \bar{\mathbb{R}}$ be proper, lsc, and convex and let $C \subset \mathbb{R}^d$ be non-empty, closed, and convex such that* $\operatorname{dom} f \cap \operatorname{int} C \neq \emptyset$. *Then it holds*

$$\hat{x} \in \operatorname*{argmin}_{x \in C} f(x) \iff \begin{cases} 1. & \hat{x} \in C \\ 2. & 0 \in \partial f(\hat{x}) + N_C(\hat{x}) \end{cases}$$

The condition that $\operatorname{dom} f \cap \operatorname{int} C \neq \emptyset$ ensures that the subgradient sum-rule is fulfilled for $f + i_C$ and $N_C(\hat{x}) = \partial i_C(\hat{x})$ proves the claim.

Here are two reformulations of the optimality conditions: $\hat{x}$ solves $\min_{x\in C} f(x)$ if and only if $\hat{x} \in C$ and one of the following conditions holds

A. $\exists p \in \partial f(\hat{x}) \forall x \in C\colon \langle p, x - \hat{x}\rangle \geq 0$,

B. $\exists p \in \partial f(\hat{x}) \forall \gamma > 0\colon \hat{x} = P_C(\hat{x} - \gamma p)$.

Condition A.: It holds that $0 \in \partial f(\hat{x}) + N_C(\hat{x})$ if and only if there ist $p \in \partial f(\hat{x})$ and $-p \in N_C(\hat{x})$. Writing out the latter condition gives the assertion.

Condition B: The Projection Theorem (Theorem 2.9) we can characterize $\hat{x} = P_C(\hat{x} - \gamma p)$ by

$$\forall x \in C\colon \ \langle \hat{x} - \gamma p - \hat{x}, x - \hat{x}\rangle \leq 0$$

and this is equivalent to

$$\forall x \in C\colon \langle p, x - \hat{x}\rangle \geq 0.$$

The above results allow us to obtain very simple algorithms in several situations.

Example 11.3 (Non-negative least squares). We consider a least squares problem $\min \frac{1}{2}\|Ax - b\|_2^2$ and want to find only non-negative solutions. i.e. we add the constraint $x \geq 0$. We can do this by adding an indicator function of the non-negative orthant $C = \mathbb{R}^d_{\geq 0}$, i.e. we consider

$$\min_x \tfrac{1}{2}\|Ax - b\|_2^2 + i_{\mathbb{R}^d_{\geq 0}}(x).$$

We use the projection characterization (item B. above): Projecting onto the non-negative orthant is just clipping away the negative entries, i.e we take the positive part of the vector:

$$P_{\mathbb{R}^d_{\geq 0}}(x) = \max(x, 0) =: x_+.$$

Moreover, the function $f(x) = \frac{1}{2}\|Ax - b\|_2^2$ is convex and differentiable, hence $\partial f(x) = \{A^T(Ax - b)\}$ and thus, solutions are characterized by

$$\hat{x} = (\hat{x} - \gamma A^T(A\hat{x} - b))_+$$

for any $\gamma > 0$. It turns out that for suitable γ (namely $0 < \gamma < 2/\|A^TA\|$) the corresponding fixed point iteration

$$x^{k+1} = (x^k - \gamma A^T(Ax^k - b))_+$$

converges to a solution (we will prove this later). $\triangle$

A little bit more general:

Example 11.4 (Projected subgradient method). Let $f : \mathbb{R}^d \to \bar{\mathbb{R}}$ be proper, convex and lsc and let C be non-empty, closed and convex and consider the general convexly constrained convex optimization problem $\min_{x\in C} f(x)$

Condition B. now reads as: For some $\hat{p} \in \partial f(\hat{x})$ it holds that

$$\hat{x} = P_C(\hat{x} - \gamma\hat{p})$$

which we can turn into a fixed-point iteration

$$\begin{aligned} \text{Choose } & p^k \in \partial f(x^k), \\ \text{Set } & x^{k+1} = P_C(x^k - \gamma p^k). \end{aligned}$$

Let us analyze this method a little bit: If x^* denotes any solution of the problem, then $P_C(x^*) = x^*$ and since P_C is Lipschitz continuous with constant 1 (we will show this later) we get

$$\begin{aligned} \|x^{k+1} - x^*\|_2^2 &= \|P_C(x^k - \gamma p^k) - P_C(x^*)\|_2^2 \\ &\le \|x^k - \gamma p^k - x^*\|_2^2 \\ &= \|x^k - x^*\|_2^2 - 2\gamma\langle p^k, x^k - x^*\rangle + \gamma^2\|p^k\|_2^2 \\ &\le \|x^k - x^*\|_2^2 - 2\gamma(f(x^k) - f(x^*)) + \gamma^2\|p^k\|_2^2 \end{aligned}$$

where the last step uses the subgradient inequality $f(x^*) \ge f(x^k) + \langle p^k, x^* - x^k\rangle$. Since $f(x^k) - f(x^*) \ge 0$ (x^* is a minimizer) we see that a step with a small enough stepsize γ should reduce the distance to any minimizer. Of course, we can also use a stepsize γ_k that changes with each iteration, leading to an estimate

$$\|x^{k+1} - x^*\|_2^2 \le \|x^k - x^*\|_2^2 - 2\gamma_k(f(x^k) - f(x^*)) + \gamma_k^2\|p^k\|_2^2$$

We can rearrange this to

$$\gamma_k(f(x^k) - f(x^*)) \le \tfrac{1}{2}\gamma_k^2\|p^k\|_2^2 + \tfrac{1}{2}\|x^k - x^*\|_2^2 - \tfrac{1}{2}\|x^{k+1} - x^*\|_2^2.$$

Now we sum up these inequalities for $k = 0, \dots, N$ and get

$$\begin{aligned} \sum_{k=0}^N \gamma_k(f(x^k) - f(x^*)) &\le \tfrac{1}{2}\sum_{k=0}^N \gamma_k^2\|p^k\|_2^2 + \tfrac{1}{2}\|x^0 - x^*\|_2^2 - \tfrac{1}{2}\|x^{N+1} - x^*\|_2^2 \\ &\le \tfrac{1}{2}\sum_{k=0}^N \gamma_k^2\|p^k\|_2^2 + \|x^0 - x^*\|_2^2. \end{aligned}$$

To get a convergent method, we assume that the norms of the subgradients are bounded, i.e. for all k we have $\|p^k\|_2 \le L$ for some $L > 0$. Furthermore we denote $f^* := f(x^*)$, $f_{\text{best}}^N := \min_{k=0,\dots,N} f(x^k)$ and $D^2 = \frac{1}{2}\|x^0 - x^*\|_2^2$. Then we get

$$\Big(\sum_{k=0}^N \gamma_k\Big)(f_{\text{best}}^N - f^*) \le \tfrac{L^2}{2}\sum_{k=0}^N \gamma_k^2 + D^2.$$

Finally, this leads to

$$(f_{\text{best}}^N - f^*) \le \frac{D^2 + \frac{L^2}{2}\sum_{k=0}^N \gamma_k^2}{\sum_{k=0}^N \gamma_k}.$$

This shows: The best function value converges towards the minimal one, if we assume that the stepsizes γ_k fulfill

$$\frac{\sum_{k=0}^N \gamma_k^2}{\sum_{k=0}^N \gamma_k} \overset{N\to\infty}{\longrightarrow} 0.$$

This can be accomplished, for example, if $\sum_{k=0}^\infty \gamma_k^2$ converges, while $\sum_{k=1}^\infty \gamma_k$ diverges which is the case, for example, for $\gamma_k = 1/(k+1)$. However, other stepsizes also make sense, e.g. one can consider $\gamma_k = 1/\sqrt{k+1}$. More precisely, we get the estimates (by comparison with the respective integrals)

We use that for a function f that decreases on an interval $[K-1, N+1]$ it holds that $\int_K^{N+1} f(x)dx \le \sum_{k=K}^N f(k) \le \int_{K-1}^N f(x)dx$.

$$\begin{aligned}
\log(N+2) = \int_0^{N+1} \frac{1}{x+1}dx \le \sum_{k=0}^N \frac{1}{k+1} &\le 1 + \int_0^N \frac{1}{x+1}dx = 1 + \log(N+1)\\
2\sqrt{N+2} - 2 = \int_0^{N+1} \frac{1}{\sqrt{x+1}}dx \le \sum_{k=0}^N \frac{1}{\sqrt{k+1}} &\le 1 + \int_0^N \frac{1}{\sqrt{x+1}}dx = 2\sqrt{N+1} - 1\\
\sum_{k=0}^N \frac{1}{(k+1)^2} &\le 1 + \int_0^N \frac{1}{(x+1)^2}dx = 2 - \frac{1}{N+1}.
\end{aligned}$$

Hence, we get for the stepsize $\gamma_k = 1/(k+1)$ that

$$(f_{\text{best}}^N - f^*) \le \frac{D^2 + \frac{L^2}{2}(2 - \frac{1}{N+1})}{\log(N+2)}.$$

and for the stepsize $\gamma_k = 1/\sqrt{k+1}$ that

$$(f_{\text{best}}^N - f^*) \le \frac{D^2 + \frac{L^2}{2}(1 + \log(N+1))}{2\sqrt{N+2} - 2}.$$

If one wants to achieve the best result with a fixed number N of iterations, one can proceed differently: Here one would like to minimize the ratio

$$R(\gamma) = \frac{D^2 + \frac{L^2}{2}\sum_{k=0}^N \gamma_k^2}{\sum_{k=0}^N \gamma_k}$$

over all variables γ_k. We can take the gradient with respect to the vector $\gamma = (\gamma_1, \gamma_2, \dots, \gamma_N)$ and get

$$\nabla R(\gamma) = \frac{L^2\gamma\sum_{k=0}^N \gamma_k - (D^2 + \frac{L^2}{2}\sum_{k=0}^N \gamma_k^2)\mathbb{1}}{\left(\sum_{k=0}^N \gamma_k\right)^2}$$

where $\mathbb{1}$ denotes the vector of all ones. Setting this to zero we observe that γ should be a constant vector, i.e. $\gamma = h\mathbb{1}$ for some h. Plugging this in, we need to solve

$$L^2(N+1)h^2 = \tfrac{L^2}{2}(N+1)h^2 + D^2$$

leading to the constant stepsize

$$\gamma_k \equiv \frac{\sqrt{2}D}{L\sqrt{N+1}}.$$

This gives us the guarantee that

$$f_{\text{best}}^N - f^* \leq 2D^2 \tfrac{1}{\sqrt{N+1}}.$$

In conclusion we get the following:

stepsize γ_k	$\frac{1}{k+1}$	$\frac{1}{\sqrt{k+1}}$	$\frac{\sqrt{2}D}{L\sqrt{N+1}}$
estimate $f_{\text{best}}^N - f^*$	$\frac{L^2}{2}\frac{D^2+2-\frac{1}{N+1}}{\log(N+2)}$	$\frac{L^2}{2}\frac{D^2+1+\log(N+1)}{2\sqrt{N+2}-2}$	$\frac{2D^2}{\sqrt{N+1}}$

So, theoretically, the option with fixed step-size has the best worst-case guarantee, but note that further iterations will not improve this any more, and moreover, an estimate on $\|x^0 - x^*\|_2$ is needed. In practice, the stepsize $1/(k+1)$ often leads to best results, but the choice of step-sizes for subgradient methods is a delicate issue.

Here is an example for the problem of *least absolute deviations* (LAD) which is

$$\min_{x\in\mathbb{R}^n} \|Ax - b\|_1,$$

i.e. one minimizes the sum of the absolute deviations and not the sum of the squares. This approach is useful if b containes additive noise which follows a Laplace distribution. Since the 1-norm is convex, lsc and everywhere continous, we can apply the chain rule from Theorem 10.4. Since the subgradient of the 1-norm is obtain from the known subgradient of the absolute value by applying this component-wise, we obtain

$$\partial\|x\|_1 = \operatorname{Sign}(x)$$

with the so-called *multivalued sign* function which acts componentwise as

$$\operatorname{Sign}(x_i) = \begin{cases} \{-1\}, & x_i < 0 \\ [-1,1], & x_i = 0 \\ \{1\}, & x_i > 0. \end{cases}$$

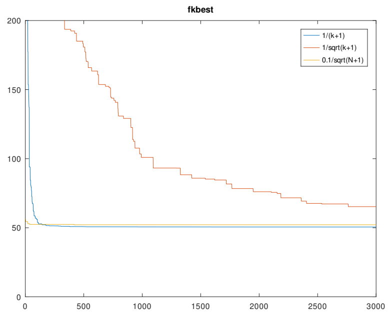
By the chain rule and Lemma 9.6 we get as subgradient of $f(x) = \|Ax - b\|_1$

$$\partial f(x) = A^T \operatorname{Sign}(Ax - b).$$

Hence, the subgradient method for the LAD problem can be implemented by choosing the ordinary sign-function

$$x^{k+1} = x^k - \gamma_k A^T \operatorname{sign}(Ax^k - b).$$

△



12 Proximal algorithms

Lemma 12.1. *The proximal mapping* $\operatorname{prox}_{\lambda f}$ *for a convex and lsc function and any* $\lambda > 0$ *is Lipschitz continuous with constant* 1*, i.e. it holds*

$$\|\operatorname{prox}_{\lambda f}(x) - \operatorname{prox}_{\lambda f}(y)\|_2 \leq \|x - y\|_2.$$

Proof. We use the variational inequality from Theorem 8.2 for $\hat{x} = \operatorname{prox}_{\lambda f}(x)$ and $\hat{y} = \operatorname{prox}_{\lambda f}(y)$ and plug in $\hat{y}$ and $\hat{x}$ is the other inequality to get

$$\begin{aligned}\langle x - \hat{x}, \hat{y} - \hat{x}\rangle + \lambda(f(\hat{x}) - f(\hat{y})) &\leq 0\\ \langle y - \hat{y}, \hat{x} - \hat{y}\rangle + \lambda(f(\hat{y}) - f(\hat{x})) &\leq 0.\end{aligned}$$

Adding these inequalities gives

$$\begin{aligned}0 &\geq \langle x - \hat{x}, \hat{y} - \hat{x}\rangle + \langle y - \hat{y}, \hat{x} - \hat{y}\rangle\\ &= \langle x - y - (\hat{x} - \hat{y}), \hat{y} - \hat{x}\rangle\\ &= \langle x - y, \hat{y} - \hat{x}\rangle + \|\hat{y} - \hat{x}\|_2^2.\end{aligned}$$

By Cauchy-Schwarz, we get

$$\|\hat{y} - \hat{x}\|_2^2 \leq \|x - y\|_2 \|\hat{y} - \hat{x}\|_2$$

which shows the claim. □

We start with a simple algorithm to solve an unconstrained minimization problem $\min_x f(x)$ (the algorithm is, in this very simple form, not practically useful, but will be good to know, since it can be used as a building block for further methods). The method is called the *proximal point method* and simply iterates

$$x^{k+1} = \operatorname{prox}_{t_k f}(x^k)$$

for some sequence $t_k > 0$ of stepsizes.

Lemma 12.2. *The sequence* (x^k) *from the proximal point method fulfills*

$$f(x^{k+1}) \leq f(x^k) - \tfrac{1}{t_k}\|x^{k+1} - x^k\|_2^2.$$

Proof. From the variational inequality (Theorem 8.2) we get with $x = y = x^k$ and $\hat{x} = x^{k+1}$ that

$$t_{k+1}(f(x^{k+1}) - f(x^k)) + \langle x^k - x^{k+1}, x^k - x^{k+1}\rangle \leq 0$$

from which the claim follows. □

This can be used to show convergence of the method:

Theorem 12.3. *Let* f *be proper, convex and lsc and let* $f^* = f(x^*) = \min_x f(x)$*. Then it holds that the sequence* (x^k) *generated by the proximal point method fulfills*

$$\begin{aligned}\|x^{k+1} - x^*\|_2 &\leq \|x^k - x^*\|_2 \quad \textit{and}\\ f(x^{N+1}) - f^* &\leq \tfrac{\|x^0 - x^*\|_2^2}{2\sum_{k=0}^N t_k}.\end{aligned}$$

Proof. Again by the variational inequality from Theorem 8.2 (now with $x = x^k$, $\hat{x} = x^{k+1}$ and $y = x^*$

$$\begin{aligned} f(x^{k+1}) &\leq f^* - \tfrac{1}{t_k}\langle x^k - x^{k+1}, x^* - x^{k+1}\rangle \\ &= f^* + \tfrac{1}{t_k}\Big(\langle x^k - x^{k+1}, x^k - x^*\rangle - \|x^k - x^{k+1}\|_2^2\Big) \end{aligned}$$

where we inserted $-x^k + x^k$ to get to the second equality. Now we use $\langle a, b\rangle = \frac{1}{2}(\|a\|_2^2 + \|b\|_2^2 - \|a - b\|_2^2)$ with $a = x^k - x^{k+1}$ and $b = x^k - x^*$ to get

$$\begin{aligned} f(x^{k+1}) &\leq f^* + \tfrac{1}{2t_k}\Big(\|x^k - x^{k+1}\|_2^2 + \|x^k - x^*\|_2^2 - \|x^{k+1} - x^*\|_2^2 - 2\|x^k - x^{k+1}\|_2^2\Big) \\ &\leq f^* + \tfrac{1}{2t_k}\Big(\|x^k - x^*\|_2^2 - \|x^{k+1} - x^*\|_2^2\Big). \end{aligned}$$

Since $f^* \leq f(x^{k+1})$ we get, on the one hand, $\|x^{k+1} - x^*\|_2 \leq \|x^k - x^*\|_2$ and, on the other hand, summing up the estimates from $k = 0, \dots, N$ we get

$$2\sum_{k=0}^{N} t_k(f(x_{k+1}) - f^*)) \leq \|x^0 - x^*\|_2^2.$$

Since we already know that $f(x^{k+1}) - f^* \geq f(x^{N+1}) - f^*$ we get

$$2\Big(\sum_{k=0}^{N} t_k\Big)(f(x^{N+1}) - f^*) \leq \|x^0 - x^*\|_2^2$$

as claimed. □

Hence, we see that larger stepsizes are better, but one should emphasize that the proximal point method is merely a theoretical algorithm, as each step needs the evaluation of the proximal map for the objective and this may be no simpler than the original problem.

In fact in can be simpler, since the objective in the minimization problem for the prox is strongly convex even if f is just convex.

Now we come to a more practical algorithm and this relies on a very fruitful idea: If the objective function in our optimization problem is the sum of two convex function, i.e.

$$\min_{x\in\mathbb{R}^d} f(x) + g(x)$$

we may try to treat both terms differently, depending on their properties. Methods that are derived from splitting the objective additively into different parts go under the name *splitting methods*. Two different properties that will be useful are the following:

- *L-smoothness*: $g : \mathbb{R}^d \to \mathbb{R}$ is convex and differentiable and, moreover, that the gradient ∇g is Lipschitz-continuous with constant L, i.e. $\|\nabla g(y) - \nabla g(x)\|_2 \leq L\|y - x\|_2$.

- *proximability*: f is of the form that $\operatorname{prox}_{\lambda f}$ is simple to evaluate. This is the case, for example for the 1-norm $\|x\|_1$ or indicator functions i_C of convex sets if the projection onto these sets is simple (e.g. positivity constraints, hyperplanes, balls).

We have just seen in Lemma 12.2 that the proximal map reduces the objective (for this part) and the next lemma shows that one can also get a guaranteed descent of the objective by doing gradient steps for L-smooth functions:

Lemma 12.4. *If $g : \mathbb{R}^d \to \mathbb{R}$ is L-smooth, then it holds that*

$$g(y) - g(x) - \langle \nabla g(x), y - x\rangle \leq \tfrac{L}{2}\|x - y\|_2^2.$$

Proof. By the fundamental theorem of calculus we get

$$\begin{aligned} g(y) - g(x) - \langle \nabla g(x), y - x\rangle &= \int_0^1 \langle \nabla g(x + \tau(y-x)) - \nabla g(x), y - x\rangle \mathrm{d}\tau \\ &\leq \int_0^1 \|\nabla g(x + \tau(y-x)) - \nabla g(x)\|_2 \|y - x\|_2 \mathrm{d}\tau \\ &\leq \int_0^1 L\|\tau(y-x)\|_2\|y-x\|_2\mathrm{d}\tau \\ &\leq \tfrac{L}{2}\|x-y\|_2^2. \end{aligned}$$

$\square$

This allows to guarantee a reduction the value of $g(x)$ by making a gradient step $x^+ = x - \lambda\nabla g(x)$: We use the above lemma with $y = x^+$ and the fact that $x^+ - x = -\lambda\nabla g(x)$ to get

$$\begin{aligned} g(x^+) &\leq g(x) + \langle \nabla g(x), x^+ - x\rangle + \tfrac{L}{2}\|x^+ - x\|_2^2 \\ &= g(x) - \lambda\|\nabla g(x)\|_2^2 + \tfrac{L\lambda^2}{2}\|\nabla g(x)\|_2^2 \\ &= g(x) - \lambda(1 - \tfrac{L}{2}\lambda)\|\nabla g(x)\|_2^2. \end{aligned}$$

Hence, we get a guaranteed descent if both $\lambda > 0$ and $1 - \frac{L}{2}\lambda > 0$, i.e. if $0 < \lambda < 2/L$.

How can we use these two ingredients to come up with a method that can mininize the objective $f + g$? One idea is, to replace the differentiable part of the objective by a simpler upper bound at some current iterate x^k: Inspired by Lemma 12.4 we define x^{k+1} by

$$x^{k+1} = \operatorname*{argmin}_x \Big[f(x) + g(x^k) + \langle \nabla g(x^k), x - x^k\rangle + \tfrac{1}{2\lambda}\|x - x^k\|_2^2\Big].$$

We can drop the terms in the objective which do not depend on x and multiply by λ to get

$$x^{k+1} = \operatorname*{argmin}_x \Big[\lambda f(x) + \langle \lambda\nabla g(x^k), x - x^k\rangle + \tfrac{1}{2\lambda}\|x - x^k\|_2^2\Big].$$

We complete the square to see that

$$\begin{aligned}
&\langle \lambda\nabla g(x^k), x - x^k\rangle + \tfrac{1}{2\lambda}\|x - x^k\|_2^2 \\
&= \tfrac{1}{2}\|\lambda\nabla g(x^k)\|^2 + \langle \lambda\nabla g(x^k), x - x^k\rangle + \tfrac{1}{2\lambda}\|x - x^k\|_2^2 - \tfrac{1}{2}\|\lambda\nabla g(x^k)\|^2 \\
&= \tfrac{1}{2}\|x - (x^k - \lambda\nabla g(x^k))\|_2^2 - \tfrac{1}{2}\|\lambda\nabla g(x^k)\|_2^2.
\end{aligned}$$

Again dropping terms that do not affect the minimizer, we finally get that

$$\begin{aligned}
x^{k+1} &= \operatorname*{argmin}_x \lambda f(x) + \tfrac{1}{2}\|x - (x^k - \lambda\nabla g(x^k))\|_2^2 \\
&= \operatorname{prox}_{\lambda f}(x^k - \lambda\nabla g(x^k)).
\end{aligned}$$

This method does a gradient step for the smooth part and a proximal step for the proximable part.

Example 12.5 (Regularized least squares for inverse problems). One example for which the proximal gradient method gained popularity is the case of regularized least squares problems: For some $A \in \mathbb{R}^{m\times n}$ and $b \in \mathbb{R}^m$ one can consider the least squares problem $\min_x \frac{1}{2}\|Ax - b\|_2^2$. This is used in many contexts, e.g. in statistics for regression but also in the context of signal processing or inverse problems where some quantity of interest $x^\dagger \in \mathbb{R}^n$ can only be measured indirectly, namely one can only observe $b^\delta = Ax^\dagger + \eta$ where A is a (known) linear map which models the measurement process, and η is an unknow error (e.g. due to measurement noise or modelling errors). Since b^δ is also affected by noise, it is pointless to solve $Ax = b^\delta$ exactly and a least squares approach seems more reasonable. In addition it may also happen than $n > m$, i.e. we do not have enough measurements to reconstruct $x^\dagger$ even from a noise-free b^δ. More precisely, the minimizers of the least squares problem are characterized by the equation $A^TAx = A^Tb^\delta$, but since A^TA does not have full rank, there are still multiple solutions, but due to noise, none of these seems to be reasonable. In this case one uses prior knowledge on the unknown solution, and this is done by specifying a *regularization functional* $R : \mathbb{R}^n \to \bar{\mathbb{R}}$ which gives small values $R(x)$ to "reasonable" x and large values $R(x)$ for "undesired" x. The regularization functional is also called *penalty function* since is penalizes undesired vectors. Together, we end with the *regularized least squares problem*

$$\min_x \tfrac{1}{2}\|Ax - b^\delta\|_2^2 + \alpha R(x)$$

where α is a positive *regularization parameter* that can emphasize the regularization (large α) or tone it down (small α).

If R is convex, lsc and proximable, one can use the proximal gradient method to solve this minimization problem: We take $g(x) = \frac{1}{2}\|Ax - b^\delta\|_2^2$ and $f(x) = \alpha R(x)$ and since $\nabla g(x) = A^T(Ax - b^\delta)$ is Lipschitz continuous with constant $L = \|A^TA\|$ one iterates

$$x^{k+1} = \operatorname{prox}_{\lambda\alpha R}(x^k - \lambda A^T(Ax^k - b^\delta))$$

with some $\lambda \in \,]0, 2/\|A^T A\|[$.

$\triangle$

13 Convex conjugation

There is an important notion of duality for convex functions, namely, the one of convex conjugation (and this duality is related to the characterization of closed, convex set as intersection of half-spaces). In Corollary 5.5 we have already seen that a closed convex function is equal to the supremum of all affine functions that are below said function. This leads to the following definition:

Definition 13.1. For a function $f : \mathbb{R}^d \to \bar{\mathbb{R}}$ we define the *(Fenchel) conjugate* $f^* : \mathbb{R}^d \to \bar{\mathbb{R}}$ as

$$f^*(p) = \sup_x \Big[\langle p, x\rangle - f(x)\Big]$$

Moreover, the *biconjugate* is

$$f^{**}(x) = \sup_p \Big[\langle p, x\rangle - f^*(p)\Big].$$

If f is proper, then f^* is the pointwise supremum of affine functions and hence, it is always convex and lower semicontinuous (even when f has none of these properties).

We consider a few one-dimensional examples:

Example 13.2.

1. Let $f(x) = \alpha x^2$ for $\alpha > 0$. Then the conjugate is

$$f^*(p) = \sup_x [px - \alpha x^2].$$

To calculate the supremum we take the derivative with respect to x, set it to zero and plug it in:

$$\begin{aligned} p - 2\alpha x = 0 &\implies x = \tfrac{p}{2\alpha} \\ &\implies \sup_x [px - \alpha x^2] = p\tfrac{p}{2\alpha} - \alpha\tfrac{p^2}{4\alpha^2} = \tfrac{1}{4\alpha}p^2, \end{aligned}$$

hence,

$$f^*(p) = \tfrac{1}{4\alpha}p^2.$$

Similarly, one observes that $f^{**}(x) = f(x)$. (Note that $f^* = f$ for $\alpha = \frac{1}{2}$, i.e. for $f(x) = x^2/2$.)

2. Let $f(x) = \exp(x)$. We consider different cases:

   - If $p < 0$, then $px - \exp(x)$, is unbounded from above and we get $f^*(p) = \infty$.
   - If $p = 0$, then $-\exp$ has the supremum 0 (which is not attained) and we have $f^*(0) = 0$.

- If $p > 0$, then $px - \exp(x)$ is bounded from above and we calculate as in the previous example

$$\begin{aligned} p - \exp(x) = 0 &\implies x = \log(p) \\ &\implies \sup_x [px - \exp(x)] = p\log(p) - p. \end{aligned}$$

Hence, the conjugate is

$$f^*(p) = \begin{cases} \infty, & \text{if } p < 0 \\ 0, & \text{if } p = 0 \\ p\log(p) - p, & \text{if } p > 0. \end{cases}$$

One may verify in a similar way that $f^{**}(x) = \exp(x) = f(x)$.

3. For $f(x) = |x|$ we have $f^*(x) = \sup_x px - |x|$ and we see by direct inspection that

$$f^*(x) = \begin{cases} \infty, & \text{if } p > -1 \\ 0, & \text{if } -1 \leq p \leq 1 \\ \infty, & \text{if } p > 1 \end{cases}$$

i.e. $f^*(x) = i_{[-1,1]}(x)$. Again, we get $f^{**}(x) = |x| = f(x)$.

4. For a non-convex example, consider

$$f(x) = \begin{cases} \infty, & \text{if } |x| > 1 \\ 1 - x^2, & \text{if } |x| \leq 1. \end{cases}$$

In this case one has $f^*(x) = |x|$ and by the previous example, $f^{**}(x) = i_{[-1,1]}(x) \neq f(x)$. However, $f^{**} = \operatorname{cl} f$.

$\triangle$

We will see later that the observed behavior $f^{**} = f$ for convex and lsc functions and $f^{**} = \operatorname{cl} f$ qre true in general.

Lemma 13.3. *Let $f, g : \mathbb{R}^d \to \bar{\mathbb{R}}$. Then:*

i) If $f \geq g$, then $f^* \leq g^*$.

ii) For all p, x where $f(x)$ or $f^*(p)$ are finite we have Fenchel's inequality

$$f(x) + f^*(p) \geq \langle x, p \rangle.$$

iii) It holds Fenchel's equality

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

The equivalence $p \in \partial f(x) \iff x \in \partial f^*(p)$ is called *subgradient inversion theorem*.

iv) It holds $f^{**} \leq f$.

Proof. i) If $f \geq g$, then

$$\begin{aligned} g^*(p) &= \sup_x [\langle p, x\rangle - g(x)] \\ &\geq \sup_x [\langle p, x\rangle - f(x)] = f^*(p). \end{aligned}$$

ii) Follows directly from the definition of the conjugate by replacing the supremum by any value of the argument.

iii) By rearranging the subgradient inequality, we have $p \in \partial f(x)$ if and only if for all y it holds $\langle x, p\rangle - f(x) \geq \langle p, y\rangle - f(y)$. Taking the supremum over all y shows $\langle x, p\rangle - f(x) \geq f^*(p)$ and by Fenchel's inequality, we get equality. Since this argument works both ways, the claim is proven.

iv) We consider

$$f^{**}(x) = \sup_p [\langle p, x\rangle - f^*(p)]$$

and use ii) to estimate the term in the supremum by $f(x)$ (which is then idependent from p).

$\square$

Lemma 13.4. *Let $f : \mathbb{R}^d \to \bar{\mathbb{R}}$. Then it holds:*

i) If f is not proper, then $f^* \equiv \infty$ or $\equiv -\infty$.

ii) If f is proper, then $f^* > -\infty$.

iii) If f is proper and convex, then f^* is proper and moreover we have $(\operatorname{cl} f)^* = f^*$ and $f^{**} = \operatorname{cl} f$.

Proof. i) For non-proper f we have two cases. In the first case there is x_0 such that $f(x_0) = -\infty$. But then there is no affine function below f and hence $f^* \equiv \infty$. In the second case $f \equiv \infty$, end then $f^* \equiv -\infty$.

ii) Now let f be proper. If we had $f^*(p_0) = -\infty$ for some p_0, then we would have for every x that $-\infty \geq \langle p_0, x\rangle - f(x)$, but this implies $f(x) \geq \infty$ for all x.

iii)
- We show that f^* is proper: By Proposition 5.4, we know that there exist $p_0 \in \mathbb{R}^d$ and $\alpha \in \mathbb{R}$ such that $f(x) \geq \langle p_0, x\rangle + \alpha$ for all x, and hence $f^*(p_0) \leq -\alpha$.
- Since f is proper, we have to show $(\bar{f})^* = f^*$ and by Lemma 13.3 i) we deduce from $\bar{f} \leq f$ the inequality $(\bar{f})^* \geq f^*$. For the lsc envelope we have by the Fenchel inequality that

  $$\bar{f}(x_0) = \liminf_{x\to x_0} f(x) \geq \liminf_{x\to x_0} [\langle p, x\rangle - f^*(p)] \quad = \langle p, x_0\rangle - f^*(p).$$

  This leads to $f^*(p) \geq \langle p, x_0\rangle - \bar{f}(x_0)$ and taking the supremum over all x_0 shows $f^* \geq (\bar{f})^*$.

The statement $f^{**} = \operatorname{cl} f$ follows from the next theorem.

The theorem says that $f = f^{**}$ if and only if f is convex and lsc, and hence $f^{**}(f^*)^* = ((\mathrm{cl}\, f)^*)^* = \mathrm{cl}\, f$ since $\mathrm{cl}\, f$ is convex and lsc.

□

Theorem 13.5 (Fenchel-Moreau)**.** *A proper function $f : \mathbb{R}^d \to \mathbb{R}^d$ is convex and lsc if and only if $f = f^{**}$*

Proof. Since any conjugate is always convex and lsc, the reverse implication is clear.

Now let f be proper, convex and lsc. We already know that f^* is proper, convex and lsc as well and that $f^{**} \leq f$. Now we show the following claim:

If $f(x) > \alpha$ for some x and α, then $f^{**}(x) \geq \alpha$.

Once, the claim is proven, the theorem follows: If $x \notin \mathrm{dom}(f)$, then we can choose α arbitrarily large and see that $f^{**}(x) = \infty$ as well. If $x \in \mathrm{dom}(f)$, and $f^{**}(x) = f(x) - \epsilon$ for some $\epsilon > 0$, then we could chose $\alpha = f(x) - \epsilon/2$ and would get $f^{**}(x) < \alpha$ which would contradict $f^{**}(x) \geq \alpha$.

Now we prove the claim: If $(x, \alpha) \notin \mathrm{epi}(f)$, we can, since $\mathrm{epi}(f)$ is closed and convex, strictly separate (x, α) from $\mathrm{epi}(f)$, i.e. there is $(p, a) \in \mathbb{R}^{d+1}$ and $\epsilon > 0$ such that

$$\sup_{(y,\beta)\in\mathrm{epi}(f)} \langle \begin{bmatrix} p \\ a \end{bmatrix}, \begin{bmatrix} y \\ \beta \end{bmatrix} \rangle \leq \langle \begin{bmatrix} p \\ a \end{bmatrix}, \begin{bmatrix} x \\ \alpha \end{bmatrix} \rangle - \epsilon.$$

which says that for all $(y, \beta) \in \mathrm{epi}(f)$ we have

$$\langle p, y \rangle + a\beta \leq \langle p, x \rangle + a\alpha - \epsilon \tag{*}$$

If we had $a > 0$ we would get a contradiction with $\beta \to \infty$.

Hence we have $a \leq 0$. If $a < 0$, we divide (*) by $-a > 0$ and set $\bar{p} = -p/\alpha$ to get

$$\langle \bar{p}, y \rangle - \beta \leq \langle \bar{p}, x \rangle - \alpha + \tfrac{\epsilon}{a}$$

We set $\beta = f(y)$ and take the supremum over all y on the left-hand side to get

$$f^*(\bar{p}) \leq \langle \bar{p}, x \rangle - \alpha + \tfrac{\epsilon}{a} < \langle \bar{p}, x \rangle - \alpha.$$

We rearrange and use the Fenchel inequality (Lemma 13.3) to get

$$\alpha < \langle \bar{p}, x \rangle - f^*(\bar{p}) \leq f^{**}(x).$$

In the remaining case $a = 0$ the inequality (*) turns into

$$\langle p, y \rangle \leq \langle p, x \rangle - \epsilon$$

and still holds for all $y \in \mathrm{dom}(f)$. If we had $x \in \mathrm{dom}(f)$ we would get a contradiction by choosing $y = x$, so we have $x \notin \mathrm{dom}(f)$ if

$a = 0$. So we have $f(x) = \infty$ in this case and for all q we have by Fenchel's inequality again

$$\langle q, y\rangle - f(y) \leq f^*(q).$$

Multiplying the second to last inequality by $\mu > 0$ and adding the last one gives

$$\langle q + \mu p, y\rangle - f(y) \leq f^*(q) + \mu\langle p, x\rangle - \mu\epsilon.$$

Taking the supremum over y gives

$$f^*(q + \mu p) \leq f^*(q) + \mu\langle p, x\rangle - \mu\epsilon,$$

which leads to

$$\langle q, x\rangle - f^*(q) + \mu\epsilon \leq \langle q + \mu p, x\rangle - f^*(q + \mu p) \leq f^{**}(x).$$

Since the left hand side goes to ∞ for $\mu \to \infty$, this shows $f^{**}(x) = \infty$, as desired. $\square$

14 Conjugation calculus

Proposition 14.1. *If $f : \mathbb{R}^d \to \bar{\mathbb{R}}$ is proper, then $f^* = f$ it and only if $f(x) = \frac{1}{2}\|x\|_2^2$.*

Proof. We have seen in the previous lecture that αx^2 has the conjugate $\frac{1}{4\alpha}p^2$ so with $a = \frac{1}{2}$ they are equal. Applying this componentwise we see that the conjugate of $\frac{1}{2}\|x\|_2^2$ is the function itself.

For the converse, assume $f = f^*$. By Fenchel's inequality we have $f(x) + f^*(p) \geq \langle x, p\rangle$ and with $p = x$ we get $f(x) \geq \frac{1}{2}\|x\|_2^2 =: h(x)$. By Lemma 13.3 i) we know that $f(x) = f^*(x) \leq h^*(x) = \frac{1}{2}\|x\|_2^2$ which proves the claim. □

Lemma 14.2. *Let $f : \mathbb{R}^d \to \bar{\mathbb{R}}$, $A \in \mathbb{R}^{d\times d}$ be invertible, $a \in \mathbb{R}$, $b \in \mathbb{R}^d$ and $\lambda \in \mathbb{R}\setminus\{0\}$. Then it holds that*

- *i) If $\varphi(x) = f(x) + a$, then $\varphi^*(p) = f^*(p) - a$.*
- *ii) If $\varphi(x) = f(\lambda x)$, then $\varphi^*(p) = f^*(\lambda^{-1}p)$.*
- *iii) If $\varphi(x) = \lambda f(x)$, $\lambda > 0$, then $\varphi^*(p) = \lambda f^*(\lambda^{-1}p)$.*
- *iv) If $\varphi(x) = f(x) - \langle b, x\rangle$, then $\varphi^*(p) = f^*(p + b)$.*
- *v) If $\varphi(x) = f(Ax + b)$, then $\varphi^*(p) = f^*(A^{-T}p) - \langle A^{-T}p, b\rangle$.*

Proof. These are straightforward calculations:

- i) $\varphi^*(p) = \sup_x \big[\langle p, x\rangle - f(x) - a\big] = f^*(p) - a$.
- ii) This is a special case of v).
- iii) $\varphi^*(p) = \sup_x \big[\langle p, x\rangle - \lambda f(x)\big] = \lambda \sup_x \big[\langle \lambda^{-1}p, x\rangle - f(x)\big] = \lambda f^*(\lambda^{-1}p)$.
- iv) $\varphi^*(p) = \sup_x \big[\langle p, x\rangle - (f(x) - \langle b, x\rangle)\big] = \sup_x \big[\langle p + b, x\rangle - f(x)\big] = f^*(p + b)$.
- v) $\varphi^*(p) = \sup_x \big[\langle p, x\rangle - f(Ax + b)\big] = \sup_y \big[\langle p, A^{-1}(y - b)\rangle - f(y)\big] = \sup_y \big[\langle A^{-T}p, y - b\rangle - f(y)\big] = f^*(A^{-T}p) - \langle A^{-T}p, b\rangle$.

□

By these rules one immediately sees that for $f(x) = \frac{1}{2}\|x - b\|_2^2$ one has $f^*(p) = \frac{1}{2}\|p\|_2^2 + \langle p, b\rangle$, for example.

The conjugation of the infimal convolution is readily calculated:

Lemma 14.3. *If $f_1, f_2 : \mathbb{R}^d \to \bar{\mathbb{R}}$ are proper, then it holds that*

$$(f_1 \,\square\, f_2)^* = f_1^* + f_2^*.$$

Proof. This can be seen as follows:

$$\begin{aligned}(f_1 \,\square\, f_2)^*(p) &= \sup_x \left[\langle p, x\rangle - \inf_{x_1+x_2=x}(f_1(x_1) + f_2(x_2))\right]\\ &= \sup_x \sup_{x_1+x_2=x} \left[\langle p, x\rangle - f_1(x_1) - f_2(x_2)\right]\\ &= \sup_{x_1,x_2} \left[\langle p, x_1\rangle + \langle p, x_2\rangle - f_1(x_1) - f_2(x_2)\right]\\ &= f_1^*(p) + f_2^*(p).\end{aligned}$$

□

One would be tempted to assume that one also has that $(f_1 + f_2)^* = f_1^* \,\square\, f_2^*$ but this does not holds without further assumptions.

Here is a counterexample:

Example 14.4. Let $f, g : \mathbb{R}^2 \to \bar{\mathbb{R}}$ defined by

$$f(x) = \begin{cases} -\sqrt{x_1 x_2}, & x_1, x_2 \geq 0 \\ \infty, & \text{else} \end{cases}, \qquad g(x) = I_{\{x_1=0\}}(x).$$

The sum is $(f+g)(x) = I_{\{x_1=0,\ x_2\geq 0\}}(x)$ and has the conjugate

$$(f+g)^*(p) = \sup_{x_1=0,\ x_2\geq 0} \left[p_1x_1 + p_2x_2\right] = I_{\{p_2\leq 0\}}(p).$$

The individual conjugates are

$$\begin{aligned}g^*(p) &= \sup_{x_1=0} \left[p_1x_1 + p_2x_2\right] = I_{\{p_2=0\}}(p)\\ f^*(p) &= \sup_{x_1,x_2\geq 0} \left[p_1x_2 + p_2x_2 + \sqrt{x_1x_2}\right]\end{aligned}$$

We see that

$$f^*(p) = I_{\{p_1p_2\leq -1\}}(p).$$

Hence, the infimal convoluiton is

$$(f^* \,\square\, g^*)(p) = \inf_{q+r=p} I_{\{p_2=0\}}(q) + I_{\{p_1p_2\leq -1\}}(r) = I_{\{p_2=0\}+\{p_1p_2\leq -1\}}(p) = I_{\{p_2<0\}}(p).$$

Note that $(f+g)^* \neq (f^* \,\square\, g^*)(p)$ and that the latter is not even lsc. △

We show a very general result which is due to Attouch and Brezis from 1986: Let $f, g : \mathbb{R}^d \to \bar{\mathbb{R}}$ be two proper and convex functions and assume that

$$\bigcup_{\lambda\geq 0} \lambda(\operatorname{dom}(f) - \operatorname{dom}(g)) \quad \text{is a subspace of } \mathbb{R}^d. \tag{Q}$$

Note that in Example 14.4 we have $\operatorname{dom}(f) = \{x_1, x_2 \geq 0\}$ and $\operatorname{dom}(g) = \{x_2 = 0\}$ and hence $\operatorname{dom}(f) - \operatorname{dom}(g) = \{x_2 \geq 0\}$. Thus, (Q) is not fulfilled

Theorem 14.5. *If $f, g : \mathbb{R}^d \to \bar{\mathbb{R}}$ are proper, convex, and lsc and satisfy condition (Q), then*

$$(f+g)^* = f^* \,\Box\, g^*$$

and the inf-convolution on the right is exact.

Proof. Step 0: First we note that for any decomposition $p = p_1 + p_2$ we get

$$\begin{aligned}(f+g)^*(p) &= \sup_x \left[\langle x, p\rangle - f(x) - g(x)\right] \\ &= \sup_x \left[\langle x, p_1\rangle + \langle x, p_2\rangle - f(x) - g(x)\right] \\ &\leq \sup_x \left[\langle x, p_1\rangle - f(x)\right] + \sup_x \left[\langle x, p_2\rangle - g(x)\right] \\ &= f^*(p_1) + g^*(p_2).\end{aligned}$$

Taking the infimum over all such decomposition on the right hand side shows

$$(f+g)^* \leq f^* \,\Box\, g^*.$$

Step 1: Now we claim that if the more restrictive condition

$$\bigcup_{\lambda \geq 0} \lambda(\operatorname{dom}(f) - \operatorname{dom}(g)) = \mathbb{R}^d \tag{Q'}$$

holds, then $f^* \,\Box\, g^*$ is lsc on $\mathbb{R}^d$. Let $\mu \in \mathbb{R}$ and

$$C := \operatorname{lev}_\mu(f^* \,\Box\, g^*) = \{p \in \mathbb{R}^d \mid (f^* \,\Box\, g^*)(p) \leq \mu\}$$

and aim to show that C is closed. For $\epsilon > 0$ consider

$$C_\epsilon := \{q + r \in \mathbb{R}^d \mid f^*(q) + g^*(r) \leq \mu + \epsilon\}.$$

By definition of the infimal convolution we have $(f^* \,\Box\, g^*)(p) \leq \mu$ if and only if $p = q + r$ such that for all $\epsilon > 0$ it holds that $f^*(q) + g^*(r) \leq \mu + \epsilon$. In other words: It holds that

$$C = \bigcap_{\epsilon > 0} C_\epsilon$$

and hence, it is enough, to prove that all the C_ϵ are closed. To that end, we consider the sets

$$\begin{aligned}K &= C_\epsilon \cap \overline{B_t(0)} \\ &= \{q + r \in \mathbb{R}^d \mid f^*(q) + g^*(r) \leq \mu + \epsilon,\ \|q + r\| \leq t\}.\end{aligned}$$

If all these K are closed, then C_ϵ is closed. Let

$$H = \{(q, r) \in \mathbb{R}^d \times \mathbb{R}^d \mid f^*(q) + g^*(r) \leq \mu + \epsilon,\ \|q + r\| \leq t\}.$$

Since the map $(q, r) \mapsto f^*(q) + g^*(r)$ is closed (both f^* and g^* are lsc), we see that H is closed.

We show that H is bounded: To show this, we show that there is a constant $C(x,y)$ such that for all $(q,r) \in H$ it holds that $\langle x,q\rangle + \langle y,r\rangle \leq C(x,y)$. By assumption (Q') we can write every (x,y) as

$$x - y = \lambda(u - v)$$

with some $u \in \mathrm{dom}(f)$, $v \in \mathrm{dom}(g)$ and $\lambda \geq 0$. Then (using the inequality by Fenchel and Cauchy-Schwarz)

$$\begin{aligned}\langle x,q\rangle + \langle y,r\rangle &= \lambda\langle u,q\rangle + \lambda\langle v,r\rangle + \langle y - \lambda v, q + r\rangle\\ &\leq \lambda(f^*(q) + f(u) + g^*(r) + g(v)) + \|q + r\|\|y - \lambda v\|\\ &\leq \lambda(\mu + \epsilon + f(u) + g(v)) + t\|y - \lambda v\| = C(x,y).\end{aligned}$$

This shows that H is bounded, and hence, compact. It remains to note that K is the image of H under the linear map $(q,r) \mapsto q + r$ and hence, K is also compact, hence closed.

Step 2: Now we prove that if condition (Q') is fulfilled, we have $(f+g)^* = f^* \,\square\, g^*$ and that the inf-convolution is exact. By Lemma 14.3 we have $(f^* \,\square\, g^*)^* = f^{**} + g^{**} = f + g$ and another conjugation shows

$$(f+g)^* = (f^* \,\square\, g^*)^{**},$$

But our previous step showed that $f^* \,\square\, g^*$ is lsc and hence $(f^* \,\square\, g^*)^{**} = f^* \,\square\, g^*$.

To see that the inf-convolution is exact, we note that we can see (similar to he first step) that for each μ the set $\{q \mid f^*(q) + g^*(p - q) \leq \mu\}$ is closed and hence, the infimum in the definition of the inf-convolution is attained.

Step 3: In the last step we get rid of the restrictive assumption (Q'). We use the following fact: If $A \subset \mathbb{R}^d$ is convex and $\bigcup_{\lambda\geq 0} \lambda A$ is a subspace, then $0 \in A$ and hence $\bigcup_{\lambda>0} \lambda A = \bigcup_{\lambda\geq 0} \lambda A$.

Let $a \in A$. Since $a \in \bigcup_{\lambda\geq 0} \lambda A$ and the latter set is a vector space, there is $b \in A$ such that $-a = \lambda b$. But then we have that

$$\tfrac{1}{1+\lambda}a + \tfrac{\lambda}{1+\lambda}b \in A$$

but the convex combination on the right hand side is 0.

From Assumption (Q) and the previous observation it follows that $\mathrm{dom}(f) \cap \mathrm{dom}(g) \neq \emptyset$ and we may assume without loss of generality that $0 \in \mathrm{dom}(f) \cap \mathrm{dom}(g)$. We define the subspace $V = \bigcup_{\lambda\geq 0} \lambda(\mathrm{dom}(f) - \mathrm{dom}(g))$ and note $\mathrm{dom}(f) \subset V$ and $\mathrm{dom}(g) \subset V$. Hence, we could have worked in V from the start and since V is isomorophic to $\mathbb{R}^n$, the proof is complete.

□

As a consequence, the subgradient sum-rule also holds if (Q) is fulfilled:

Corollary 14.6. *Let f and g be proper, convex and lsc and fulfull condition (Q). Then it holds that*

$$\partial(f+g) = \partial f + \partial g.$$

As we have seen, the inclusion $\partial f + \partial g \subset \partial(f+g)$ always holds. For the reverse inclusion let $p \in \partial(f+g)(x)$. By Fenchel's equality (Lemma 13.3) we get

$$(f+g)(x) + (f+g)^*(p) = \langle p, x\rangle.$$

Theorem 14.5 shows $(f+g)^* = f^* \,\square\, g^*$ and that the inf-convolution is exact, i.e. we have $(f+g)^*(p) = f^*(p-q) + g^*(q)$ for some q. We get

$$f(x) + g(x) + f^*(p-q) + g^*(q) = \langle p-q, x\rangle + \langle q, x\rangle.$$

We conclude $p - q \in \partial f(x)$ and $q \in \partial g(x)$ (since $q \notin \partial g(x)$ would imply $g(x) + g^*(q) > \langle x, q\rangle$ from which we would get $f(x) + f^*(p-q) < \langle p-q, x\rangle$ which contradict Fenchel's inequality). Hence, we have $p \in \partial f(x) + \partial g(x)$.

Finally, let us note that condition (Q) is more general than the assumption in Theorem 10.4 namely that

$$\exists x \in \operatorname{dom}(f) \cap \operatorname{dom}(g),\ f \text{ continuous at } x \implies \text{(Q) fulfilled.}$$

If there is $x \in \operatorname{dom}(f) \cap \operatorname{dom}(g)$ and f is continuous at f, then $x \in \operatorname{int}\operatorname{dom}(f)$, i.e. $B_\epsilon(x) \subset \operatorname{dom}(f)$ for some ϵ. Hence $\operatorname{dom}(f) - \operatorname{dom}(g) \supset B_\epsilon(x) - \{x\} = B_\epsilon(0)$ and we see that $\bigcup_{\lambda \geq 0} \lambda(\operatorname{dom}(f) - \operatorname{dom}(g)) = \mathbb{R}^d$ while for (Q) we only need that the union is a subspace.

15 Fenchel-Rockafellar duality

The Fenchel equality states that

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

Hence, we see that if x^* is a minimizer of f, then we know that $x \in \partial f^*(0)$. In other words: The subgradient of the conjugate at zero shows us, where minimizers of f are. Hence, knowing conjugate functions is quite helpful to treat minimization problems.

In this section, we use conjugate functions to derive a quite general notion of duality between optimization problem (which includes the notion of duality of linear programs, for example).

The problems which we will treat in this section are of the form

$$\min_{x\in\mathbb{R}^n} f(x) + g(Ax)$$

where $A \in \mathbb{R}^{m\times n}$ and $f : \mathbb{R}^n \to \bar{\mathbb{R}}$ and $g : \mathbb{R}^m \to \bar{\mathbb{R}}$ are two proper, convex and lsc functions. We have seen examples for this, e.g. in Example 12.5 where we minimized $\frac{1}{2}\|Ax - b\|_2^2 + \alpha R(x)$, i.e. we could take $f = R$ and $g(y) = \frac{1}{2}\|y - b\|_2^2$.

To motivate the duality, we express g via its conjugate and get

$$\begin{aligned}\min_{x\in\mathbb{R}^n} f(x) + g(Ax) &= \min_{x\in\mathbb{R}^n} f(x) + \sup_{y\in\mathbb{R}^m} \langle y, Ax\rangle - g^*(y)\\ &= \min_{x\in\mathbb{R}^n}\sup_{y\in\mathbb{R}^m} f(x) + \langle y, Ax\rangle - g^*(y).\end{aligned}$$

If we would know that the supremum was a maximum and that we could swap minimum and maximum, this would be equal to

$$\begin{aligned}\max_{y\in\mathbb{R}^n}\min_{x\in\mathbb{R}^n} f(x) + \langle y, Ax\rangle - g^*(y) &= \max_{y\in\mathbb{R}^m}\min_{x\in\mathbb{R}^n}\big[f(x) + \langle A^T y, x\rangle\big] - g^*(y)\\ &= \max_{y\in\mathbb{R}^m} -\big[\max_{x\in\mathbb{R}^n}\langle -A^T y, x\rangle - f(x)\big] - g^*(y)\\ &= \max_{y\in\mathbb{R}^m} -f^*(-A^T y) - g^*(y).\end{aligned}$$

The problem in the last line is called the *(Fenchel-Rockafellar) dual problem*. Note that the problem is a concave maximization problem, but since it is equivalent to

$$\min_{y\in\mathbb{R}^m} f^*(-A^T y) + g^*(y)$$

it is of exactly the same type than the problem we started with (and this problem is called *primal problem* in this context).

Let us explore the relation of the primal and dual problem in general. We start from middle ground, namely with a so-called *saddle point problem*, i.e. we have a function $L : \mathbb{R}^n \times \mathbb{R}^m \to \bar{\mathbb{R}}$ and want to find a pair (x^*, y^*) such that

$$x^* \in \operatorname*{argmin}_{x\in\mathbb{R}^n} L(x,y^*)$$
$$y^* \in \operatorname*{argmax}_{y\in\mathbb{R}^m} L(x^*,y).$$

Any such pair will be called *saddle point* of L. Put differently, the solution (x^*, y^*) should satisfy

$$\forall x,y:\ L(x^*,y) \leq L(x,y) \leq L(x,y^*). \tag{*}$$

Proposition 15.1. *For any L it holds the min-max inequality*

$$\inf_x \sup_y L(x,y) \geq \sup_y \inf_x L(x,y).$$

Moreover, (x^*, y^*) is a saddle point of L if and only if

$$\min_x \max_y L(x,y) = \max_y \min_x L(x,y) = L(x^*,y^*).$$

The statement should be read as "(x^*, y^*) is a saddle point, if and only if the minima and maxima exist and the equality holds".

Proof. For all $\bar{x}, \bar{y}$ it holds that

$$\sup_y L(\bar{x},y) \geq L(\bar{x},\bar{y}) \geq \inf_x L(x,\bar{y}).$$

Taking the infimum over all $\bar{x}$ on the left and the supremum over all $\bar{y}$ on the right shows the inequality.

Now, let (x^*, y^*) be a saddle point. From the formulation (*) above and the min-max inequality we get that

$$\begin{aligned} L(x^*,y^*) &\geq \max_y L(x^*,y) \geq \inf_x \max_y L(x,y) \\ &\geq \sup_y \inf_x L(x,y) = \sup_y \min_x L(x,y) \geq \min_x L(x,y^*) = L(x^*,y^*). \end{aligned}$$

Check, that the change from sup to max and inf to min is justified in all places.

Conversely, assume that the interchange of minimum and maximum gives the same result. Then

$$\min_x L(x,y^*) \leq L(x^*,y^*) \leq \max_y L(x^*,y)$$

which shows that (x^*, y^*) is a saddle point. □

Example 15.2. In many cases, one only has the min-max inequality, but no saddle points exist, even though maxima and minima exist. The simplest example may be

$$L(x,y) = \sin(x+y).$$

It holds that

$$\inf_y \sup_x \sin(x+y) = 1 > -1 = \sup_x \inf_y \sin(x+y)$$

even though the infima and suprema are attained. △

Definition 15.3. For a saddle-point problem with function L define $F(x) = \sup_y L(x,y)$ and $G(y) = \inf_x L(x,y)$. Then the corresponding *primal* and *dual problem* are

$$\min_x F(x) \quad \text{and} \quad \max_y G(y),$$

respectively.

One sees that if (x^*, y^*) is a saddle point of L, then x^* solves the primal problem and y^* solves the dual problem and one has the that $F(x^*) = G(y^*)$, i.e. the primal and dual optimal values coincide.

A little more terminology: If we have

$$\inf_x \sup_y L(x,y) = \sup_y \inf_x L(x,y)$$

we say that *strong duality* holds for the saddle point problem, while the min-max inequality shows that one always has *weak duality*. In terms of primal and dual problems: The primal and dual problems of a saddle point problem always obey *weak duality* i.e. it always holds that $\inf_x F(x) \geq \sup_y G(y)$ and if $\inf_x F(x) = \sup_y G(y)$, even *strong duality* holds. Note that strong duality does not imply that the infimum or supremum are attained.

If a saddle point problem does not obey strong duality, we say that there is a *duality gap* and the difference $\inf_x \sup_y L(x,y) - \sup_y \inf_x L(x,y)$ is called value of the *duality gap*.

Coming back to the problem

$$\min_x f(x) + g(Ax)$$

from the beginning of the section we see that this is the primal problem of the saddle point problem for

$$L(x,y) = f(x) + \langle Ax, y\rangle - g^*(y)$$

and the respective dual problem is

$$\max_y -f^*(-A^T y) - g^*(y).$$

In this context, the function L is also called *Lagrangian* of the problem. Weak duality

$$\inf_x f(x) + g(Ax) \geq \sup_y -f^*(-A^T y) - g^*(y)$$

always holds. Moreover, we will denote the *primal objective* by $F(x) = f(x) + g(Ax)$ and the *dual objective* by $G(y) = -f^*(-A^T y) - g^*(y)$.

The knowledge of the dual problem is useful, to get cheap estimates on the distance to optimality:

Proposition 15.4. *For convex, proper and lsc function f and g and matrix A define the gap function*

$$\operatorname{gap}(x,y) := f(x) + g(Ax) + f^*(-A^T y) + g^*(y) = F(x) - G(y).$$

Moreover, denote the primal and dual objective by F and G as above. Then it holds for any pair $(\bar{x},\bar{y})$ where $G(\bar{y}) > -\infty$ that

$$\operatorname{gap}(\bar{x},\bar{y}) \geq F(\bar{x}) - \inf_x F(x).$$

Proof. By weak duality one has $F(\bar{x}) \geq \inf_x F(x) \geq \sup_y G(y) \geq G(\bar{y})$ and especially $\inf_x F(x) \geq G(\bar{y})$. Hence, we have

$$F(\bar{x}) - \inf_x F(x) \leq F(\bar{x}) - G(\bar{y}) = \operatorname{gap}(\bar{x},\bar{y}).$$

□

What is the corresponding statement for the distance to dual optimality?

Theorem 15.5 (Fenchel-Rockafellar duality)**.** *Let $f : \mathbb{R}^n \to \bar{\mathbb{R}}$, $g : \mathbb{R}^m \to \bar{\mathbb{R}}$ be proper, convex and lsc and let $A \in \mathbb{R}^{n\times m}$. If*

$$\bigcup_{\lambda\geq 0} \lambda(\operatorname{dom}(g) - A\operatorname{dom}(f)) = \mathbb{R}^m$$

then strong duality holds, i.e.

$$\inf_{x\in\mathbb{R}^n} f(x) + g(Ax) = \max_{y\in\mathbb{R}^m} -f^*(-A^T y) - g^*(y)$$

(and especially the max on the right hand side is attained).

Proof. We define $\Phi : \mathbb{R}^n \times \mathbb{R}^m \to \bar{\mathbb{R}}$ by $\Phi(x,y) = f(x) + g(y)$ and let $M = \{(x,Ax) \mid x \in \mathbb{R}^n\}$ (i.e. M is the graph of A).

We aim to show that

$$\bigcup_{\lambda\geq 0} \lambda(\operatorname{dom}(\Phi) - M) = \mathbb{R}^n \times \mathbb{R}^m.$$

To that end, let $(x,y) \in \mathbb{R}^n \times \mathbb{R}^m$. By assumption, there exists $\lambda \geq 0$, $u \in \operatorname{dom}(f)$ and $v \in \operatorname{dom}(g)$ such that

$$y - Tx = \lambda(v - Tu)$$

and one can show (as in the beginning of Step 3 in Theorem 14.5) that one can even choose $\lambda > 0$. If we set $a = u - x/\lambda$ we get

$$x = \lambda(u-a), \quad y = \lambda(v - Ta)$$

and this shows that indeed $(x,y) \in \bigcup_{\lambda\geq 0} \lambda(\operatorname{dom}(\Phi) - M)$.

Now we define $\Psi(x,y) = I_M(x,y)$ and note that we have just shown that condition (Q) is fulfilled for Φ and Ψ. Thus we have by Theorem 14.5

$$(\Phi + \Psi)^* = \Phi^* \,\square\, \Psi^*$$

and the infimal convolution is exact. Actually, we only need this equality at 0 since

$$\begin{aligned}(\Phi+\Psi)^*(0) &= \sup_{(x,y)} -\Phi(x,y)-\Psi(x,y) = \sup_{y=Ax} -f(x)-g(y)\\ &= -\inf_x f(x)+g(Ax)\end{aligned}$$

and

$$\begin{aligned}(\Phi^* \,\square\, \Psi^*)(0) &= \min_{(x,y)} \Phi^*(x,y)+\Psi^*(-(x,y)) = \min_{x=-A^Ty} f^*(x)+g^*(y)\\ &= \min_y f^*(-A^Ty)+g^*(y).\end{aligned}$$

$\square$

16 Examples of duality and optimality systems

Example 16.1 (LP duality). A *linear program* is an optimization problem where the objective function is linear and where there are linear equality and inequality constraints. The standard form of a linear program is: Given $c \in \mathbb{R}^n$, $A \in \mathbb{R}^{m\times n}$ and $b \in \mathbb{R}^m$ solve

$$\min_{x\in\mathbb{R}^n} \langle c, x\rangle \quad \text{subject to} \quad Ax \leq b$$

where the inequality is understood componentwise. In other words

$$\min_{x\in\mathbb{R}^n} \{\langle c, x\rangle \mid Ax \leq b\}.$$

We rewrite this in the context of Fenchel-Rockafellar duality as

$$\min_{x\in\mathbb{R}^n} f(x) + g(Ax)$$

with $f(x) = \langle c, x\rangle$, $g(v) = I_{\mathbb{R}^m_{\leq 0}}(v - b)$. The conjugates are

$$f^*(p) = \sup_x \langle p - c, x\rangle = \begin{cases} 0, & \text{if } p - c = 0 \\ \infty, & \text{else,} \end{cases} = I_{\{c\}}(p),$$

$$g^*(y) = \sup_{v-b\leq 0} \langle v, y\rangle = \begin{cases} \langle b, y\rangle, & \text{if } y \geq 0 \\ \infty, & \text{else.} \end{cases}$$

Hence, the dual problem is

$$\max_{y\in\mathbb{R}^m} -f^*(-A^T y) - g^*(y) = \max_y \{-\langle b, y\rangle \mid -A^T y = c,\ y \geq 0\}$$

I.e. the (Fenchel-Rockafellar) dual of a linear problem is another linear program, namely

$$\max_{y\in\mathbb{R}^m} -\langle b, y\rangle \quad \text{subject to} \quad \begin{aligned} A^T y + c &= 0 \\ y &\geq 0. \end{aligned}$$

If $m < n$, then the dual problem has fewer variables (but more constraints). △

Example 16.2 (Equality constrained norm minimization). We consider the primal problem

$$\min_{x\in\mathbb{R}^n} \|x\| \quad \text{subject to} \quad Ax = b$$

with $A \in \mathbb{R}^{m\times n}$, $b \in \mathbb{R}^m$ and $\|\cdot\|$ denoting any norm on $\mathbb{R}^n$. With $f(x) = \|x\|$ and $g(v) = I_{\{b\}}(v)$ this is of the form $\min_x f(x) + g(Ax)$. The conjugate of g is simply

$$g^*(y) = \langle b, y\rangle$$

and the conjugate of f is

$$f^*(p) = \sup_x \langle p, x\rangle - \|x\|.$$

With the notion of *dual norm*, defined by

$$\|p\|_* = \sup_{\|x\|\leq 1} \langle p, x\rangle$$

which fulfills $\langle p, x\rangle \leq \|x\|\|p\|_*$ we can express the conjugate of f as

$$f^*(p) = I_{\{\|\cdot\|_*\leq 1\}}(p).$$

Hence, the dual problem is

$$\max_{y\in\mathbb{R}^m} -\langle b, y\rangle \quad \text{subject to} \quad \|A^T y\|_* \leq 1.$$

In the case of the 1-norm (whose dual is the ∞-norm), the primal

$$\min_{Ax=b} \|x\|_1$$

has the dual

$$\max_{\|A^T y\|_\infty \leq 1} -\langle b, y\rangle$$

which can be written as a linear program

By the previous example, we know that the primal should also be a linear program, can you see how to write it a such?

$$\begin{aligned}\max_{y\in\mathbb{R}^m} -\langle b, y\rangle \quad \text{subject to} \quad A^T y &\leq \mathbb{1}\\ -A^T y &\leq \mathbb{1}.\end{aligned}$$

$\triangle$

If both the subgradient sum-rule and the subgradient chain-rule hold for the objective

$$\min_x f(x) + g(Ax)$$

an optimal x^* is characterized by the inclusion

$$0 \in \partial f(x^*) + A^T \partial g(Ax^*).$$

Fenchel-Rockafellar duality allows for an alternative optimality system that uses the dual variable:

Proposition 16.3. *Let f, g be proper, convex and lower semicontinuous and assume strong duality is fulfilled and that the primal problem has a solution, i.e. we have*

$$\min_x f(x) + g(Ax) = \max_y -f^*(-A^T y) - g^*(y)$$

then a pair (x^*, y^*) is a saddle point of $f(x) + \langle Ax, y\rangle - g^*(y)$ if and only if

$$\begin{aligned}-A^T y^* &\in \partial f(x^*)\\ y^* &\in \partial g(Ax^*).\end{aligned}$$

By the subgradient inversion theorem (Lemma 13.3), the primal dual optimality system is also equivalent to

$$\begin{aligned}-A^T y^* &\in \partial f(x^*)\\ Ax^* &\in \partial g^*(y^*).\end{aligned}$$

This primal-dual optimality system (or Fenchel-Rockafellar duality system) is also equivalent to x^* begin a solution to the primal problem and y^* being a solution to the dual problem.

Proof. A pair (x^*, y^*) is optimal if and only if

$$-f^*(-A^T y^*) - g^*(y^*) = f(x^*) + g(Ax^*)$$

to which we add and subtract $\langle y^*, Ax^*\rangle$ to get

$$\langle -A^T y^*, x^*\rangle + \langle y^*, Ax^*\rangle - f^*(-A^T y^*) - g^*(y^*) = f(x^*) + g(Ax^*).$$

By Fenchel's inequality we have $\langle -A^T y^*, x^*\rangle \leq f(x^*) + f^*(-A^T y^*)$ and $\langle y^*, Ax^*\rangle \leq g(Ax^*) + g^*(y^*)$, and see that the previous equality if equivalent to

$$\langle -A^T y^*, x^*\rangle = f(x^*) + f(-A^T y^*) \quad \text{and} \quad \langle y^*, Ax^*\rangle = g(Ax^*) + g^*(y^*)$$

which, by Fenchel's equality, is equivalent to the primal-dual optimality system. □

Example 16.4 (Primal-dual optimality for LPs). Let us work out the primal-dual optimality system for the LP from Example 16.1. The subgradient of f is just $\partial f(x) = c$ (independent of x). The subgradient of g fulfills

$$\partial g(v) = \begin{cases} 0, & \text{if } v < b, \\ \emptyset, & \text{if } v \not\leq b \end{cases}$$

In the remaining (interesting) cases where some of the inqualities $v_i \leq b_i$ are tight, we have for any $w \in \partial g(v)$ that $w_i \in \,]-\infty, 0]$. Hence, the primal-dual optimality system is

The last condition is a so-called *complimentarity condition* and it states that at least one of the quantities y_i or $(Ax - b)_i$ has to be zero for every i.

$$\begin{aligned} -A^T y^* \in \partial f(x^*) &\implies \quad -A^T y^* = c \\ y^* \in \partial g(Ax^*) &\implies \quad Ax^* \leq b, \text{ and } \begin{cases} y_i^* = 0, & \text{if } (Ax^*)_i < b_i \\ y_i^* \leq 0, & \text{if } (Ax^*)_i = b_i. \end{cases} \end{aligned}$$

△

Example 16.5 (Primal-dual optimality system for constrained norm minimization). In the norm minimization example (Example 16.2), we have $\partial g^*(v)$ is empty if $v \neq b$, but equal to $\mathbb{R}$ for $v = b$. It total we get as optimality system

$$\begin{aligned} -A^T y^* \in \partial f(x^*) &\implies & -A^T y^* &= \partial\|x^*\| \\ y^* \in \partial g(Ax^*) &\implies & Ax^* &= b. \end{aligned}$$

Let's consider special cases:

- Let $\|x\| = \|x\|_1$. Then the subgradient fulfills

$$p \in \partial\|x\|_1 \iff \begin{cases} p_i = \operatorname{sign}(x_i), & \text{if } x_i \neq 0 \\ |p_i| \leq 1, & \text{if } x_i = 0. \end{cases}$$

  Hence, the primal-dual optimality system is

$$\begin{aligned} Ax^* &= b, \\ |A^T y|_i &\leq 1, \\ (A^T y)_i &= \operatorname{sign}(x_i) \text{ if } x_i \neq 0. \end{aligned}$$

- In the case of the 2-norm one would rather take $f(x) = \frac{1}{2}\|x\|_2^2$ (with subgradient $\partial f(x) = \{x\}$) and get the primal-dual optimality system

$$\begin{aligned} -A^T y^* &= x^* \\ Ax^* &= b. \end{aligned}$$

△

In some cases the inclusion $-A^T y \in \partial f(x)$ can help to recover a primal solution from a dual solution: By subgradient inversion (Lemma 13.3) the inclusion is equivalent to $x \in \partial f^*(-A^T y)$. If ∂f^* is single valued, this even leads to a single primal solution correspoding the any dual solution. The next propositions shows that this is the case, for example, when f is strongly convex.

Proposition 16.6. *Let $f : \mathbb{R}^d \to \bar{\mathbb{R}}$ be proper, convex and lsc. Then the subgradient is monotone, in the sense that for $p_i \in \partial f(x_i)$, $i = 1, 2$, then*

$$\langle p_1 - p_2, x_1 - x_2 \rangle \geq 0.$$

If f is strongly convex with constant μ, then it even holds that

$$\langle p_1 - p_2, x_1 - x_2 \rangle \geq \mu \|x_1 - x_2\|_2^2.$$

Proof. The first claim follows by adding the two subgradient inequalities $f(x_2) \geq f(x_1) + \langle p_1, x_2 - x_1 \rangle$ and $f(x_1) \geq f(x_2) + \langle p_2, x_1 - x_2 \rangle$. For the second claim, apply the first one to the convex function $g(x) = f(x) - \frac{\mu}{2}\|x\|_2^2$ with $p_i - \mu x_i \in \partial g(x_i)$. □

Proposition 16.7. *If $f : \mathbb{R}^d \to \bar{\mathbb{R}}$ is proper, strongly convex with constant μ and lsc, then:*

i) $\operatorname{dom}(f^*) = \mathbb{R}^d$,

ii) *f^* is $1/\mu$-smooth and moreover* $\nabla f^*(p) = \operatorname{argmax}_x \langle p, x \rangle - f(x)$,

Recall that f^* is L-smooth if it is differentiable with ∇f^* being Lipschitz continuous with constant L.

Proof. i) Since $f^*(p) = \sup_x \langle x, p \rangle - f(x)$, we see that strong convexity of f ensures existence of maximizers for every p, and this gives a finite value for the supremum for every p.

ii) By Fermat's principle, some x maximizes $\langle p, x \rangle - f(x)$ exactly if $p \in \partial f(x)$ which is, by subgradient inversion, equivalent to $x \in \partial f^*(p)$. By strong convexity of f, the maximizer x is unique for every p, which shows that $\partial f^*(p)$ is a singleton and Proposition 9.7 implies differentiablity of f^* and

$$\partial f^*(p) = \{\nabla f^*(p)\} = \{x\}.$$

This also implies that $\nabla f^*(p) = \operatorname{argmax}_x \langle p, x \rangle - f(x)$.

Proposition 16.6 shows for $p \in \partial f(x)$ and $p' \in \partial f(x')$ that

$$\langle p - p', x - x' \rangle \geq \mu \|x - x'\|^2.$$

Subgradient inversion gives $x \in \nabla f^*(p)$ and $x' \in \nabla f^*(p')$ and this gives

$$\langle p - p', \nabla f^*(p) - \nabla f^*(p')\rangle \geq \mu \|\nabla f^*(p) - \nabla f^*(p')\|_2^2.$$

Applying Cauchy-Schwarz's inequality shows $\|\nabla f^*(p) - \nabla f^*(p')\|_2 \leq \frac{1}{\mu}\|p - p'\|_2$ as claimed.

$\square$

17 Classes of optimization problems

We are going to develop a theory that will allow us to say how hard a certain class of optimization problems is. To that end we will have to pin down some ingredients:

- The problem class: How do we describe a problem from a class? What properties do we assume for a problem?
- The algorithm: What information of the problem can be used by the algorithm?
- A notion of approximate solution: How do we measure how good some approximate solution is?

We will analyze iterative algorithms and our goal is, to quantify how many steps are needed to get an answer with a given accuracy.

We will start with a very simple case (that is actually not related to convex analysis at all):

Global optimization of Lipschitz functions on bounded domains.

Problem class: Let $Q = [0,1]^n$ be the unit cube and $f : Q \to \mathbb{R}$ be Lipschitz continuous with constant L with respect to the ∞-norm. The problem we consider is

$$\min_{x \in Q} f(x).$$

We consider Lipschitz continuous function for some reason: Since we want a ceertain accuracy for some given iterate x^k, we would like to able to estimate $f(x^k) - \min_{x \in Q} f(x)$ in some way. If f is merely continuous, the function may vary as much as it like in a small neighborhood, and with Lipschitz continuity we are able to bound such variations.

Oracle: In each step we are able to query the function value $f(x)$ for some point $x \in Q$.

This is called a *zeroth order oracle* and methods that only use functions evaluations are called *zeroth order methods.*

Here is a very simple algorithm:

Algorithm 1: Grid search

Input: $p \in \mathbb{N}$

For all $(i_1, \dots, i_n) \in \{0, \dots, p\}^n$ assemble points

$$x_{(i_1,\dots,i_n)} = \left[\tfrac{i_1}{p}, \dots, \tfrac{i_n}{p}\right]^T$$

In every point $x_{(i_1,\dots,i_n)}$ evaluate the functional value;

Find $\bar{x}$ among the $x_{(i_1,\dots,i_n)}$, which has the smallest objection value;

Result: Pair $(\bar{x}, f(\bar{x}))$

Let us analyze this method:

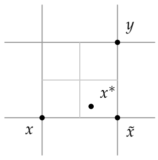
Theorem 17.1 (Upper complexity bound for grid search). *Denote $f^* := \min_{x \in Q} f(x)$ and let $\bar{x}$ be the output of Algorithm 1. Then it holds that*

$$f(\bar{x}) - f^* \leq \tfrac{L}{2p}.$$

Proof. Let $f^* = f(x^*)$. Then, x^* lies in one grid cell, i.e. we have

$$x := x_{(i_1,\dots,i_n)} \leq x^* \leq y_{(i_1+1,\dots,i_n+1)} =: y.$$

From these two surrounding points we define the best point on the grid nearby as



$$(\tilde{x})_i = \begin{cases} y_i, & x_i^* \geq \frac{x_i+y_i}{2} \\ x_i, & \text{else.} \end{cases}$$

Then we have $|\tilde{x}_i - x_i^*| \leq 1/(2p)$ and since $\tilde{x}$ is among the grid points, we have

$$f(\bar{x}) - f^* \leq f(\tilde{x}) - f(x^*) \leq L\|\tilde{x} - x^*\|_\infty \leq \tfrac{L}{2p}. \qquad \square$$

In other words: To reach an accuracy $f(\bar{x}) - f^* \leq \epsilon$ we need $p \geq L/(2\epsilon)$, and since the number of points in the grid is $N = (p+1)^n$, the number of function evaluations (or calls to the oracle) is at least

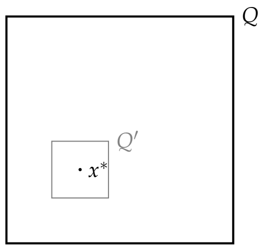
$$N \geq (\tfrac{L}{2\epsilon} + 1)^n.$$

While this is a lot, the next theorem shows, that one can not do much better than that. We can prove such result, by constructing a problem that tries to annoy a given problem as much as possible.

Theorem 17.2. *If $\epsilon < L/2$, there is no method that solves the problem with less than $(\lfloor L/(2\epsilon)\rfloor)^n$ function evaluations.*

Proof. Set $p = \lfloor L/(2\epsilon)\rfloor$ (note $p \geq 1$) and assume that there is method that needs $N < p^n$ function evaluations to solve any problem in our class. We construct an "annoying" f: We build f such that $f(x) = 0$ for all points that have been tried, but $f^* = \min f$ being as small as possible.

Since $p \geq 1$, and since we have $N < p^n$ evaluation points, there has to be a small cube $Q' \subset Q$ of sidelength $1/p$ that does not contain any of these evaluation points. Let x^* be the midpoint of Q' and set



$$f(x) = \min(0, L\|x - x^*\|_\infty - \epsilon')$$

with $\epsilon < \epsilon' < L/(2p)$. This f is L-Lipschitz by definition and equals 0 outside of Q' (since $\epsilon'/L < 1/(2p)$). Its minimal values is $f^* = f(x^*) = -\epsilon'$. Hence, the accuracy of this method is $\epsilon' > \epsilon$. $\square$

We have seen that grid seach need roughly $(\frac{1}{\epsilon})^n$ iterations and the lower bound we got in the theorem is of the same order of magnitude. Hence, grid search is actually not a bad method for this class, althoug the numer of iterations can be insanely large. This indicates, that general problems from this class can be very hard.

Here are further problems classes that we will deal with:

Convex and L-Lipschitz.

Problem class: $f : \mathbb{R}^n \to \bar{\mathbb{R}}$, proper, convex and L-Lipschitz with respect to the 2-norm and the problem is

$$\min_{x \in \mathbb{R}^n} f(x).$$

Oracle: At a given point x^k we are able to query one subgradient $p^k \in \partial f(x^k)$.

Method: The iterates will only move in the set

$$x^k \in x^0 + \operatorname{span}\{p^0, \dots, p^{k-1}\}$$

i.e. in each step we can only move into directions of subgradients which we have already encoutered.

Convex and L-smooth.

Problem class: $f : \mathbb{R}^n \to \mathbb{R}$ convex and differentiable with L-Lipschitz gradient and the problem is

$$\min_{x \in \mathbb{R}^n} f(x).$$

Oracle: At a given point x^k we are able to query the gradient $g^k = \nabla f(x^k)$.

Method: The iterates will only move in the set

$$x^k \in x^0 + \operatorname{span}\{g^0, \dots, g^{k-1}\}$$

i.e. in each step we can only move into directions of gradients which we have already encoutered.

Convex, L-smooth and μ-strongly convex. Same as before, and we additionally assume that f is strongly convex with constant $\mu > 0$.

Now we collecd a few more facts about smooth and strongly convex functions:

Theorem 17.3. *The following conditions are equivalent to f being convex and L-smooth (each holding for all x, y and $\lambda \in [0, 1]$):*

The property in iii) is called *cocoercivity* of the gradient ∇f. It is not to be confused with strong convexity where one has a lower bound of the form $\mu/2\|x - y\|_2^2$.

i) $0 \leq f(y) - f(x) - \langle \nabla f(x), y - x \rangle \leq \frac{L}{2}\|x - y\|^2$

ii) $f(x) + \langle \nabla f(x), y - x \rangle + \frac{1}{2L}\|\nabla f(x) - \nabla f(y)\|^2 \leq f(y)$

iii) $\frac{1}{L}\|\nabla f(x) - \nabla f(y)\|^2 \leq \langle \nabla f(x) - \nabla f(y), x - y \rangle$

iv) $\langle \nabla f(x) - \nabla f(y), x - y \rangle \leq L\|x - y\|^2$

v) $\lambda f(x) + (1 - \lambda) f(y) \geq f(\lambda x + (1 - \lambda) y) + \frac{\lambda(1-\lambda)}{2L}\|\nabla f(x) - \nabla f(y)\|^2$

vi) $\lambda f(x) + (1 - \lambda) f(y) \leq f(\lambda x + (1 - \lambda) y) + \frac{\lambda(1-\lambda)L}{2}\|x - y\|^2$

Proof. Claim i) is just convexity combined with Lemma 12.4.

Now fix x_0 and consider $\varphi(y) = f(y) - \langle \nabla f(x_0), y\rangle$ which has its minimum at $y^* = x_0$. By i) we have

$$\begin{aligned}\varphi(y^*) &\leq \varphi(y - \tfrac{1}{L}\nabla\varphi(y)) \leq \varphi(y) + \tfrac{L}{2}\|\tfrac{1}{L}\nabla\varphi(y)\|^2 + \langle\nabla\varphi(y), -\tfrac{1}{L}\nabla\varphi(y)\rangle\\ &= \varphi(y) - \tfrac{1}{2L}\|\nabla\varphi(y)\|^2\end{aligned}$$

which shows ii) since$\nabla\varphi(y) = \nabla f(y) - \nabla f(x_0)$.

Inequality iii) follows from ii) by adding two copies of ii) with x and y swapped and by Cauchy-Schwarz we get L-smoothness from iii).

Inequality iv) follows from i) by adding two copies and vice versa iv) implies i) by

$$\begin{aligned}f(y) - f(x) - \langle\nabla f(x), y - x\rangle &= \int_0^1 \langle\nabla f(x + \tau(y-x)) - \nabla f(x), y - x\rangle \mathrm{d}\tau\\ &\leq \tfrac{L}{2}\|y - x\|^2.\end{aligned}$$

To get v) from ii) set $x_\alpha = \alpha x + (1-\alpha)y$ and note

$$\begin{aligned}f(x) &\geq f(x_\alpha) + \langle\nabla f(x_\alpha), (1-\alpha)(x-y)\rangle + \tfrac{1}{2L}\|\nabla f(x) - \nabla f(x_\alpha)\|^2\\ f(y) &\geq f(x_\alpha) - \langle\nabla f(x_\alpha), \alpha(y-x)\rangle + \tfrac{1}{2L}\|\nabla f(y) - \nabla f(x_\alpha)\|^2.\end{aligned}$$

Multiplying by α and $(1-\alpha)$, respectively, and adding we get v). Conversely, we get ii) from v) by dividing by $1-\alpha$ and $\alpha \to 1$.

Similarly, one shows the equivalence of i) and vi). □

Lemma 17.4. *A differentiable function $f : \mathbb{R}^n \to \bar{\mathbb{R}}$ is μ-strongly convex if and only if for all x, y we have*

$$f(y) \geq f(x) + \langle\nabla f(x), y - x\rangle + \tfrac{\mu}{2}\|y - x\|_2^2.$$

Proof. Just apply Theorem 5.1 iii) to the convex function $g(x) = f(x) - \frac{\mu}{2}\|x\|_2^2$ with gradient $\nabla g(x) = \nabla f(x) - \mu x$. □

Theorem 17.5. *Let f be μ-strongly convex. Then it holds for all x, y that*

$$\begin{aligned}f(y) &\leq f(x) + \langle\nabla f(x), y - x\rangle + \tfrac{1}{2\mu}\|\nabla f(x) - \nabla f(y)\|_2^2\\ \langle\nabla f(x) - \nabla f(y), x - y\rangle &\leq \tfrac{1}{\mu}\|\nabla f(x) - \nabla f(y)\|_2^2.\end{aligned}$$

Proof. For the first inequality we fix x and tilt f by defining $\varphi(y) = f(y) + \langle\nabla f(x), y\rangle$. Note that $\nabla\varphi(x) = 0$ and thus φ is minimal at x. Since φ is still μ-strongly convex, we have $\varphi(x) \geq \min_z \varphi(z) \geq \min_z \left[\varphi(y) + \langle\nabla\varphi(y), z - y\rangle + \frac{\mu}{2}\|z - y\|_2^2\right]$. We calculate that the minimum on the right as attained at $z = y - \frac{1}{\mu}\nabla\varphi(y)$ and has the value $\varphi(y) - \frac{1}{2\mu}\|\nabla\varphi(y)\|_2^2$ which is exactly the first inequality.

For the second inequality just swap the roles of x and y in the first and add both of them. □

Theorem 17.6. *Let f be μ-strongly convex and L-smooth with $L \geq \mu$. Then it holds for all x, y that*

$$\langle \nabla f(x) - \nabla f(y), x - y\rangle \geq \tfrac{\mu L}{\mu+L}\|x-y\|_2^2 + \tfrac{1}{\mu+L}\|\nabla f(x) - \nabla f(y)\|_2^2.$$

Proof. We define the function $\varphi(x) = f(x) - \frac{\mu}{2}\|x\|_2^2$ which is still convex and since the gradient is $\nabla\varphi(x) = \nabla f(x) - \mu x$, is it $L-\mu$-smooth.

In the case $L = \mu$ we see from Theorem 17.3, iv) and Proposition 16.6 that $\langle \nabla f(x) - \nabla f(y), x - y\rangle = \mu\|x-y\|_2^2$ from which we conclude $f(x) = \frac{\mu}{2}\|x\|_2^2$ and the theorem holds.

For $\mu < L$ we get from Theorem 17.3 iii) that

$$\langle \nabla\varphi(x) - \nabla\varphi(y), x - y\rangle \geq \tfrac{1}{L-\mu}\|\nabla\varphi(x) - \nabla\varphi(y)\|_2^2.$$

The left hand side evaluates to

$$\langle \nabla f(x) - \nabla f(y), x - y\rangle - \mu\|x-y\|_2^2$$

while the right hand side is

$$\tfrac{1}{L-\mu}\big(\|\nabla f(x) - \nabla f(y)\|_2^2 - 2\mu\langle \nabla f(x) - \nabla f(y), x-y\rangle + \mu^2\|x-y\|_2^2\big).$$

Plugging this in and cleaning up proves the result. □

18 Convergence rates and worst case analysis

Now we will apply the concept of "annoying problems" to more classes of optimization problems. We start with the class of convex and L-Lipschitz functions, i.e. we do not assume differentiability or strong convexity and can only use subgradients in each iteration. Here is the result on worst case analysis:

Theorem 18.1. *For every $k \in \{0,\dots,n\}$ there exists some convex function $f_k : \mathbb{R}^n \to \bar{\mathbb{R}}$ which is Lipschitz contiuous with constant L, and has a minimum $f_k^* = f_k(x^*)$ at some x^* with $\|x^*\|_2 \leq R$ and an oracle that gives a subgradient $p \in \partial f(x)$ such that any sequence x^k that fulfills $x^k \in x^0 + \operatorname{span}\{p^0,\dots,p^{k-1}\}$ fulfills*

$$f_k(x^k) - f_k^* \geq \tfrac{LR}{2(1+\sqrt{k+1})}.$$

Proof. For some constant $\mu, \gamma > 0$ (to be defined later) we set

$$f_k(x) = \gamma \max_{1\leq i\leq k} x_i + \tfrac{\mu}{2}\|x\|_2^2.$$

The subdifferential is

$$\partial f_k(x) = \gamma \operatorname{conv}\{e_i \mid i \in I(x)\} + \mu x$$

where $I(x) := \{j \in \{1,\dots,k\} \mid x_j = \max_{1\leq i\leq k} x_i\}$. This function is Lipschitz continuous on any ball $B_\rho(0)$, since by the subgradient inequality we have for all $p_k(x) \in \partial f_k(x)$ that

$$f_k(y) - f_k(x) \leq \langle p_k(y), y - x\rangle \leq \|p_k(y)\|_2\|x-y\|_2 \leq (\mu\rho + \gamma)\|x-y\|_2$$

and hence $L = \mu\rho + \gamma$ is a Lipschitz constant.

Solving the inclusion $0 \in \partial f_k(x)$ we see that a minimizer is at x^{k*} given by

$$(x^{k*})_i = \begin{cases} -\frac{\gamma}{\mu k}, & \text{if } 1 \leq i \leq k\\ 0, & \text{else.}\end{cases}$$

The norm of the minimizer and the optimal value are

$$R_k := \|x^{k*}\|_2 = \sqrt{k\left(\tfrac{\gamma}{\mu k}\right)^2} = \tfrac{\gamma}{\mu\sqrt{k}}, \qquad f_k^* := f_k(x^{k*}) = -\tfrac{\gamma^2}{\mu k} + \tfrac{\mu}{2}R_k^2 = -\tfrac{\gamma^2}{2\mu k}.$$

Let us initialize the method with $x^0 = 0$ and see what we can get. We aim to show that the j-th iterate ($j \leq k$) x^j has all entries with indices $i = j+1,\dots,n$ equal to zero. In the first step, our oracle gives us the subgradient $p_0 = \gamma e_1$ (others would be possible, but this is the worst choice) and hence, x^1 has all entries x_i^1 equal to zero with the only possible exception of $i = 1$ and this proves the case $j = 1$.

For an induction assume that x^j fulfills the assumption. Our oracle gives the subgradient $p^j = \mu x^j + \gamma e_{i^*}$ with $i^* \leq j+1$ (note that the first j entries may all be negative so that $i^* = j+1$ is possible), and this shows the claim.

Hence, in the first $k-1$ steps, the objective value fulfills

$$f_k(x^i) \geq \max_{1\leq j\leq k} x^i_j \geq 0.$$

Now we choose our constants as

$$\gamma = \tfrac{\sqrt{k}L}{1+\sqrt{k}}, \quad \mu = \tfrac{L}{(1+\sqrt{k})R}$$

and observe that

$$f_k^* = -\tfrac{\gamma^2}{2\mu k} = -\tfrac{LR}{2(1+\sqrt{k})}$$

and $\|x^0 - x^*\|_2 = \frac{\gamma}{\mu\sqrt{k}} = R$. Finally, we compute that the function is indeed Lipschitz continuous on the ball $B_R(0)$ with constant $\mu R + \gamma = L$ as desired. ☐

Hence, we conclude that no algorithm that only uses subgradient steps is able to solve *all* optimization problems with Lipschitz-continuous convex objective with less than $\mathcal{O}(1/\sqrt{k})$ operations. One says: The iterations complexity of convex optinization is $\mathcal{O}(1/\sqrt{k})$.

Now let us analyze the next class of problems: Convex and L-smooth objectives. At each iteration x^k we can query the gradient $g^k = \nabla f(x^k)$ and the k-th iterate is assumed to be in the set

$$x^0 + \operatorname{span}\{g^0, \ldots, g^{k-1}\}.$$

We will call any method that fulfills this assumption a *first order method*.

Here is an annoying objective that is difficult for all such methods: For given $L > 0$ and $0 \leq k \leq n$ let $f_k : \mathbb{R}^n \to \mathbb{R}$ be defined by

$$f_k(x) = \tfrac{L}{4}\Big[\tfrac{1}{2}\big((x_1)^2 + \sum_{i=1}^{k}(x_i - x_{i+1})^2 + (x_k)^2\big) - x_1\Big].$$

We rewrite this objective with the matrix

$$D_k = \left[\begin{array}{ccccc|c} -1 & & & & & \\ 1 & -1 & & & & \\ & \ddots & \ddots & & & 0_{k+1,n-k} \\ & & 1 & -1 & & \\ & & & 1 & & \\ \hline & & 0_{n-k-1,k} & & & 0_{n-k-1,n-k} \end{array}\right].$$

as

$$f_k(x) = \tfrac{L}{4}\big(\tfrac{1}{2}\|D_k x\|_2^2 + \langle e_1, x\rangle\big)$$

and we have

$$\nabla f_k(x) = \tfrac{L}{4}(D_k^T D_k x - e_1) \quad \text{and} \quad \nabla^2 f(x) = \tfrac{L}{4} D_k^T D_k.$$

One sees

$$A_k = D_k^T D_k = \left[\begin{array}{cccc|c} 2 & -1 & & & \\ -1 & 2 & \ddots & & \\ & \ddots & & -1 & 0_{k,n-k} \\ & & -1 & 2 & \\ \hline & & 0_{n-k,k} & & 0_{n-k,n-k} \end{array}\right]$$

and hence, we get

$$\begin{aligned} 0 \leq \tfrac{L}{4}\|D_k s\|^2 &= \langle \nabla^2 f_k(x)s, s\rangle \\ &= \tfrac{L}{4}\Big((s^{(1)})^2 + \sum_{i=1}^{k-1}(s^{(i)} - s^{(i+1)})^2 + (s^{(k)})^2\Big) \\ &\leq \tfrac{L}{4}\Big((s^{(1)})^2 + \sum_{i=1}^{k-1} 2\big((s^{(i)})^2 + (s^{(i+1)})^2\big) + (s^{(k)})^2\Big) \\ &\leq L\sum_{i=1}^{n}(s^{(i)})^2 = L\|s\|^2. \end{aligned}$$

This shows $0 \preccurlyeq \nabla^2 f_k(x) \preccurlyeq LI$ and we have proven the f_k is convex and L-smooth.

Proposition 18.2. *The function f_k has a minimizer*

$$x_i^* = \begin{cases} 1 - \frac{i}{k+1}, & \text{if } i = 1, \dots, k \\ 0, & \text{if } i = k+1, \dots, n. \end{cases}$$

and optimal value and norm

$$f_k^* = f_k(x^*) = \tfrac{L}{8}(-1 + \tfrac{1}{k+1}) \quad \text{and} \quad \|x^*\|_2^2 \leq \tfrac{k+1}{3}, \text{ respectively.}$$

Note that the entries of the minimizer (and hence also f_k^* and its norm as well) depend on the value k. We will make use of this in the following.

Proof. The optimality condition is

$$0 = \nabla f_k(x) = \tfrac{L}{4}(A_k x - e_1)$$

which is

$$\left[\begin{array}{cccc|c} 2 & -1 & & & \\ -1 & 2 & \ddots & & \\ & \ddots & & -1 & 0_{k,n-k} \\ & & -1 & 2 & \\ \hline & & 0_{n-k,k} & & 0_{n-k,n-k} \end{array}\right] \begin{bmatrix} x_1 \\ \vdots \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

We see that $x_i = 0$ for $i = k+1, \dots, n$. The first equation is $2x_1 - x_2 = 1$, which gives

$$x_2 = 2x_1 - 1.$$

Plugging this is the second equation $-x_1 + 2x_2 - x_3 = 0$ gives

$$x_3 = 3x_1 - 2.$$

Proceeding in this way, we get for $i = 2, \dots, k$ from the $i-1$th equation that

$$x_i = ix_1 - (i-1). \tag{*}$$

The k-th equation $-x_{k-1} + 2x_k = 0$ is then

$$0 = -(k-1)x_1 + (k-2) + 2(kx_1 - (k-1)) = (k+1)x_1 - k.$$

which shows $x_1 = 1 - 1/(k+1)$. Plugging this in (*) shows the formula for the minimizer.

For the minimal value we just plug in and get

$$\begin{aligned} f_k^* = f_k(x^*) &= \tfrac{L}{4}\big(\tfrac{1}{2}\|Dx^*\|^2 - \langle x^*, e_1\rangle\big) \\ &= \tfrac{L}{4}\big(\tfrac{1}{2}\langle \underbrace{A_k x^*}_{e_1}, x^*\rangle - \langle x^*, e_1\rangle\big) = -\tfrac{L}{8}\langle x^*, e_1\rangle \\ &= \tfrac{L}{8}(-1 + \tfrac{1}{k+1}). \end{aligned}$$

Finally, we esimate

$$\begin{aligned} \|x^*\|^2 = \sum_{i=1}^n (x_i^*)^2 &= \sum_{i=1}^k \Big(1 - \tfrac{i}{k+1}\Big)^2 = \sum_{i=1}^k \Big(1 - \tfrac{2i}{k+1} + \tfrac{i^2}{(k+1)^2}\Big) \\ &= k - \tfrac{2}{k+1} \underbrace{\sum_{i=1}^k i}_{=\frac{k(k+1)}{2}} + \tfrac{1}{(k+1)^2} \underbrace{\sum_{i=1}^k i^2}_{\leq \frac{(k+1)^3}{3}} \leq \tfrac{k+1}{3}. \end{aligned}$$

□

We used the estimate $(k+1)^3 = \sum_{i=0}^k[(i+1)^3 - i^3] = \sum_{i=0}^k[3i^2 + 3i + 1] \geq 3\sum_{i=1}^k i^2$. The exact sum would be $\sum_{i=1}^k i^2 = \frac{k(k+1)(2k+1)}{6}$ (which we will use in the next section).

Now let us analyze how methods according to our definition perform for this particular function.

Lemma 18.3. *Let $1 \leq p \leq n$ and $x_0 = 0$. Then it holds for every sequence x^k with*

$$x^k \in L_k := \operatorname{span}\{\nabla f_p(x_0), \dots, \nabla f_p(x_{k-1})\}$$

and $k \leq p$ that $x^k = (*, \dots, *, 0 \dots, 0)$, i.e. only the first k entries can be different from zero.

Proof. Recall that $\nabla f_p(x) = \frac{L}{4}(A_p x - e_1)$, and hence we see that $\nabla f_p(x^0) = -\frac{L}{4}e_1$, and hence, x^1 fulfills the claim.

No proceed inductively: if x^k only has the first k entries non-zero, then, since A_p is tridiagonal, $g^{k+1} = \nabla f_p(x^k)$ has only the first $k+1$ entries different zero and the same holds for x^{k+1}. □

Corollary 18.4. *For every sequence x^k, $k = 0, \dots, p$ with $x^0 = 0$ and $x^k \in L_k$ it holds that $f_p(x^k) \geq f_k^*$.*

We only observe that since $x^k \in L_k$, it only has the first k components non-zero and thus, that $f_p(x^k) = f_k(x^k) \geq f_k^*$.

19 More worst case analysis

Now comes the main theorem on the complexity of first order method for convex and L-smooth functions.

Theorem 19.1. *Let k be such that $1 \leq k \leq \frac{1}{2}(n-1)$ and $x_0 \in \mathbb{R}^n$. Then there exists a convex and L-smooth $f : \mathbb{R}^n \to \mathbb{R}$ such that any first order method produces iterates such that*

$$\begin{aligned} f(x^k) - f^* &\geq \tfrac{3L\|x^0-x^*\|_2^2}{32(k+1)^2} \\ \|x^k - x^*\|_2^2 &\geq \tfrac{1}{8}\|x^0 - x^*\|_2^2. \end{aligned}$$

Proof. Without loss of generality, we assume $x^0 = 0$ (otherwise use $\tilde{f}(x) = f(x + x^0)$). For the first inequality, fix k and consider $f = f_{2k+1}$ (the one which we defined on page 81). We know from Corollary 18.4 and Proposition 18.2

$$f(x^k) = f_{2k+1}(x^k) = f_k(x^k) \geq f_k^* = \tfrac{L}{8}(-1 + \tfrac{1}{k+1})$$

Note that $f^* = f^*_{2k+1} = \frac{L}{8}(-1 + \frac{1}{2k+2})$ (again by Proposition 18.2).

and get

$$\frac{f(x^k) - f^*}{\|x^0 - x^*\|_2^2} \geq \frac{\frac{L}{8}\big(-1 + \frac{1}{k+1} - (-1 + \frac{1}{2k+2})\big)}{\frac{1}{3}(2k+2)} = \frac{3L}{8}\frac{\frac{1}{k+1} - \frac{1}{2(k+1)}}{2(k+1)} = \frac{3L}{32(k+1)^2}.$$

For the second inequality we use that x^k is zero in the first components and that we know the extries of x^* by Proposition 18.2 to estimate

$$\begin{aligned} \|x^k - x^*\|^2 &\geq \sum_{i=k+1}^{2k+1} (x_i^*)^2 = \sum_{i=k+1}^{2k+1}\Big(1 - \tfrac{i}{2k+2}\Big)^2 \\ &= \sum_{i=k+1}^{2k+1}\Big(1 - \tfrac{i}{k+1} + \tfrac{i^2}{4(k+1)^2}\Big). \end{aligned}$$

We calculate

$$\begin{aligned} \sum_{i=k+1}^{2k+1} i &= \sum_{i=1}^{2k+1} i - \sum_{i=1}^{k} i = \tfrac{1}{2}\Big((2k+1)(2k+2) - k(k+1)\Big) \\ &= \tfrac{1}{2}(3k+2)(k+1) \end{aligned}$$

Recall the famous $\sum_{i=1}^n i = \frac{n(n+1)}{2}$.

and

$$\begin{aligned} \sum_{i=k+1}^{2k+1} i^2 &= \sum_{i=1}^{2k+1} i^2 - \sum_{i=1}^{k} i^2 \\ &= \tfrac{1}{6}\Big((2k+1)(2k+2)(4k+3) - k(k+1)(2k+1)\Big) \\ &= \tfrac{1}{6}(k+1)(2k+1)(7k+6). \end{aligned}$$

Here we use $\sum_{i=1}^k i^2 = \frac{k(k+1)(2k+1)}{6}$.

Combining this, we get (again using Proposition 18.2)

$$\begin{aligned}
\|x^k - x^*\|^2 &\geq k + 1 - \tfrac{1}{k+1}\tfrac{(3k+2)(k+1)}{2} + \tfrac{1}{4(k+1)^2}\tfrac{(k+1)(2k+1)(7k+6)}{6}\\
&= -\tfrac{k}{2} + \tfrac{(2k+1)(7k+6)}{24(k+1)}\\
&= \tfrac{(2k+1)(7k+6)-12k(k+1)}{24(k+1)}\\
&= \tfrac{2k^2+7k+6}{24(k+1)}\\
&\geq \tfrac{2k^2+7k+6}{24(k+1)}\tfrac{3}{2(k+1)}\|x^0 - x^*\|^2\\
&= \tfrac{2k^2+7k+6}{16(k+1)^2}\|x^0 - x^*\|^2\\
&\geq \tfrac{2(k^2+2k+1)}{16(k+1)^2}\|x^0 - x^*\|^2\\
&= \tfrac{1}{8}\|x^0 - x^*\|^2. \qquad \square
\end{aligned}$$

Some interpretation of the above theorem:

- The result only holds for roughly the first $k \leq \frac{1}{2}(n-1)$ iterates. However, if the problem size n get large, $\frac{1}{2}(n-1)$ may already be more iterations than one wants to perform.
- Although, the result is somehow dependent on the dimension, it still says, that we can't get better results without using anything that is specific for the case of dimension n.
- Roughly, we see that the distance to the optimal function value goes down like $1/k^2$. Put differently: To guarantee $f(x_k) - f^* \leq \epsilon$ we need at least

  $$k \geq \sqrt{\tfrac{3L}{32}}\tfrac{\|x_0 - x^*\|}{\sqrt{\epsilon}} - 1$$

  iterations. Another way to see it: If a method would be this fast, we would need to multiply the number of iterations by $\sqrt{2}$ to cut the distance to optimality in half.
- The iterates themself may converge arbitrarily slow.

Now, let's move on to a more restrictive family of problems: L-smooth and μ-strongly convex functions. In this case we still get to evaluate the gradient of the objective at the current iterate and move in the span of the previous gradients but we will be a bit more ambitios and aim to get some $\bar{x}$, such that

$$f(\bar{x}) - f^* \leq \epsilon, \quad \text{and} \quad \|\bar{x} - x^*\|_2^2 \leq \epsilon.$$

Note that this aim makes sense, since for strongly convex objectives, there is only one minimizer x^*.

We will state the annoying obective in infinite dimensions. Since we aim to obtain a dimensionless result, the dimension should not enter anyway and we could work with infinite dimensions as well. The simplest infinite dimensional Hilbert space is $\ell^2 = \{x = (x_i) \in \mathbb{R}^{\mathbb{N}} \mid \sum_{i=1}^{\infty} x_i^2 < \infty\}$ with inner product $\langle x, y\rangle = \sum_{i=1}^{\infty} x_i y_i$ and norm $\|x\| = \big(\sum_{i=1}^{\infty} x_i^2\big)^{1/2}$.

Here is the annoying objective for this class of functions: For constants $\mu > 0$, $Q_f > 1$ define

$$f_{\mu,Q_f}(x) = \tfrac{\mu(Q_f-1)}{8}\Big[(x_1)^2 + \sum_{i=1}^{\infty}(x_i - x_{i+1})^2 - 2x_1\Big] + \tfrac{\mu}{2}\|x\|^2.$$

With

$$D = \begin{bmatrix} -1 & & & \\ 1 & -1 & & \\ & 1 & \ddots & \\ & & & \ddots \end{bmatrix}$$

this is

$$f_{\mu,Q_f}(x) = \tfrac{\mu(Q_f-1)}{8}\Big[\|Dx\|^2 - 2\langle x, e_1\rangle\Big] + \tfrac{\mu}{2}\|x\|^2$$

and we have with

$$A = D^*D = \begin{bmatrix} 2 & -1 & \\ -1 & 2 & \ddots \\ & \ddots & \ddots \end{bmatrix}$$

that $\nabla^2 f_{\mu,Q}(x) = \frac{\mu(Q_f-1)}{4}A + \mu I$ (where I denotes the identity operator on ℓ^2). Similarly to the annoying function in the class of convex and L-smooth functions one gets $0 \preccurlyeq A \preccurlyeq 4I$ and hence

$$\mu I \preccurlyeq \nabla^2 f_{\mu,Q}(x) \preccurlyeq (\mu(Q_f-1)+\mu)I = \mu Q_f I.$$

This shows that $f_{\mu,Q}$ is indeed μ-strongly convex and L-smooth with $L = \mu Q_f$. The number

$$Q_f = L/\mu$$

is also called *condition number* of the objective.

Lemma 19.2. *The function $f = f_{\mu,Q_f}$ has the unique minimizer x^* with entries $x_k^* = q^k$ with $q = \frac{\sqrt{Q_f}-1}{\sqrt{Q_f}+1}$.*

Proof. The optimality condition is

$$0 = \nabla f_{\mu,Q_f}(x) = \Big(\tfrac{\mu(Q_f-1)}{4}A + \mu I\Big)x - \tfrac{\mu(Q_f-1)}{4}e_1 = 0,$$

i.e., we have to solve

$$\Big(A + \tfrac{4}{Q_f-1}I\Big)x = e_1.$$

The first and the kth equation ($k \geq 2$) are

$$\begin{aligned} \Big(2 + \tfrac{4}{Q_f-1}\Big)x_1 - x_2 &= 1 \\ -x_{k-1} + \Big(2 + \tfrac{4}{Q_f-1}\Big)x_k - x_{k+1} &= 0 \end{aligned}$$

and after reformulation we get

$$2\tfrac{Q_f+1}{Q_f-1}x_1 - x_2 = 1$$
$$x_{k-1} - 2\tfrac{Q_f+1}{Q_f-1}x_k + x_{k+1} = 0.$$

These equations have a solution of the form $x_k^* = q^k$, where q is the smallest solution of

$$q^2 - 2\tfrac{Q_f+1}{Q_f-1}q + 1 = 0$$

and this is

$$\begin{aligned} q &= \tfrac{Q_f+1}{Q_f-1} - \sqrt{\left(\tfrac{Q_f+1}{Q_f-1}\right)^2 - 1} \\ &= \tfrac{Q_f+1}{Q_f-1} - \sqrt{\tfrac{Q_f^2+2Q_f+1-Q_f^2+2Q_f-1}{(Q_f-1)^2}} \\ &= \tfrac{Q_f+1-2\sqrt{Q_f}}{Q_f-1} = \tfrac{(\sqrt{Q_f}-1)^2}{(\sqrt{Q_f}+1)(\sqrt{Q_f}-1)} \\ &= \tfrac{\sqrt{Q_f}-1}{\sqrt{Q_f}+1} < 1 \end{aligned}$$

Just check that the first equation

$$2\tfrac{Q_f+1}{Q_f-1}q - q^2 = 1$$

and is fulfilled by construction. And the second equation is

$$\begin{aligned} 0 &= q^{k-1} - 2\tfrac{Q_f+1}{Q_f-1}q^k + q^{k+1} \\ &= q^{k-1}\left(1 - 2\tfrac{Q_f+1}{Q_f-1}q + q^2\right) \end{aligned}$$

and hence, is fulfilled as well.

□

Theorem 19.3. *For every $x^0 \in \ell^2$, $L \geq \mu > 0$ there exists a μ-strongly convex and L-smooth function f (with condition number $Q = L/\mu$) with minimum $f^* = f(x^*)$ such that every first order method for f fulfills*

$$\|x^k - x^*\|^2 \geq \left(\tfrac{\sqrt{Q}-1}{\sqrt{Q}+1}\right)^{2k}\|x_0 - x^*\|^2$$
$$f(x^k) - f^* \geq \tfrac{\mu}{2}\left(\tfrac{\sqrt{Q}-1}{\sqrt{Q}+1}\right)^{2k}\|x_0 - x^*\|^2$$

Proof. Without loss of generality we assume $x^0 = 0$ and choose $f = f_{\mu,Q}$. Then

$$\|x^0 - x^*\|^2 = \sum_{i=1}^{\infty}|x_i^*|^2 = \sum_{i=1}^{\infty}q^{2i} = \tfrac{q^2}{1-q^2}.$$

Since $\nabla^2 f_{\mu,Q}(x)$ is tridiagonal and we have $f'_{\mu,Q}(0) = e_1$, we conclude (as we did previously) that x^k has non-zero entries only in the first k entries. This shows the first estimate:

$$\begin{aligned} \|x_k - x^*\|^2 &\geq \sum_{i=k+1}^{\infty}|x_i^*|^2 = \sum_{i=k+1}^{\infty}q^{2i} \\ &= \sum_{i=0}^{\infty}q^{2i} - \sum_{i=0}^{k}q^{2i} = \tfrac{1}{1-q^2} - \tfrac{1-q^{2(k+1)}}{1-q^2} \\ &= \tfrac{q^{2(k+1)}}{1-q^2} = q^{2k}\|x_0 - x^*\|^2. \end{aligned}$$

Lemma 17.4 with $x = x^*$ and $y = x^k$ gives

$$f(x^k) - f(x^*) \geq \tfrac{\mu}{2}\|x^k - x^*\|^2 \geq \tfrac{\mu}{2}q^{2k}\|x^k - x^*\|^2$$

which shows the second estimate. □

20 Subgradient method and gradient descent

In the case of convex and Lipschitz continuous functions on bounded domains, we have already analyzed the subgradient method in Example 11.4. Here we just recall the facts: The method is as follows. Initialize x^0 and iterate for some stepsize $\gamma_k > 0$

$$\begin{aligned} p^k &\in \partial f(x^k), \\ x^{k+1} &= P_C(x^k - \gamma_k x^k). \end{aligned}$$

The fundamental estimate we got was: if we denote by f^k_{best} the smallest objective value among the first k iterates and denote by $D^2 = \frac{1}{2}\|x^0 - x^*\|_2^2$ for any solution x^*, then it holds

$$f^k_{\text{best}} - f^* \le \frac{D^2 + \frac{L^2}{2}\sum_{i=1}^k \gamma_i^2}{\sum_{i=0}^k \gamma_i}.$$

From this one deduces (using further estimates from Example 11.4):

Theorem 20.1. *If we choose the stepsize $\gamma_i = C/\sqrt{k+1}$ for some $C > 0$ and $i = 0,\dots,k$, then we get*

$$f^k_{\text{best}} - f^* = \mathcal{O}_{k\to\infty}(1/\sqrt{k+1}).$$

If we choose $\gamma_i = C/\sqrt{i+1}$ for all i, then it holds for all k that

$$f^k_{\text{best}} - f^* = \mathcal{O}_{k\to\infty}(\log(k+1)/\sqrt{k+1}).$$

Comparing this to our worst case result from Theorem 18.1 we see that the very simple subgradient method can already be made optimal up to constants. One the one hand, this is a lucky case, but on the other hand, the optimal rate is bad and the simplest idea is good enough, so we may also say that this class is just too broad to allow for a general and practical method.

The notion "optimal up to constants" is sometime called "order optimal".

Let us test our luck with the simplest method for convex and smooth problems: gradient descent. We start with the L-smooth case:

Theorem 20.2. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be convex and L-smooth with minimum $f^* = f(x^*)$ and let x^k be defined by*

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

for $0 < h < 2/L$. Then it holds that

$$f(x^k) - f^* \le \frac{2(f(x^0)-f^*)\|x^0-x^*\|_2^2}{2\|x^0-x^*\|_2^2 + kh(2-Lh)(f(x^0)-f^*)}$$

Proof. We denote $r_k = \|x^k - x^*\|$ and estimate

$$\begin{aligned}
r_{k+1}^2 &= \|x^k - x^* - h\nabla f(x^k)\|_2^2\\
&= r_k^2 - 2h \underbrace{\langle \nabla f(x^k), x^k - x^*\rangle}_{=\langle \nabla f(x^k) - \nabla f(x^*), x^k - x^*\rangle} + h^2\|\nabla f(x^k)\|_2^2\\
&\leq r_k^2 - \tfrac{2h}{L}\|\nabla f(x^k) - \nabla f(x^*)\|_2^2 + h^2\|\nabla f(x^k)\|_2^2\\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\text{(Thm. 17.3 iii))}\\
&= r_k^2 - h(\tfrac{2}{L} - h)\|\nabla f(x^k)\|_2^2.
\end{aligned}$$

We see that $r_k \leq r_0$ and by Theorem 17.3 i) we get (denoting $\omega = h(1 - \frac{L}{2}h)$)

$$\begin{aligned}
f(x^{k+1}) &\leq f(x^k) + \langle \nabla f(x^k), x^{k+1} - x^k\rangle + \tfrac{L}{2}\|x^{k+1} - x^k\|_2^2\\
&= f(x^k) - \omega\|\nabla f(x^k)\|_2^2
\end{aligned}$$

We further abbreviate $\Delta_k = f(x^k) - f^*$ and get by convexity of f and Cauchy-Schwarz

$$\Delta_k \leq \langle \nabla f(x^k), x^k - x^*\rangle \leq r_k\|\nabla f(x^k)\| \leq r_0\|\nabla f(x^k)\|.$$

Together with the above we obtain

$$f(x^{k+1}) \leq f(x^k) - \omega\tfrac{\Delta_k^2}{r_0^2}.$$

Subtracting f^* on both sides we get the recurrence $\Delta_{k+1} \leq \Delta_k - \frac{\omega}{r_0^2}\Delta_k^2 = \Delta_k(1 - \frac{\omega}{r_0^2}\Delta_k)$ which implies $\Delta_{k+1} \leq \Delta_k$ and can be rearranged to

$$\frac{1}{\Delta_{k+1}} \geq \frac{1}{\Delta_k} + \frac{\omega}{r_0^2}\frac{\Delta_k}{\Delta_{k+1}} \geq \frac{1}{\Delta_k} + \frac{\omega}{r_0^2} \geq \cdots \geq \frac{1}{\Delta_0} + \frac{\omega}{r_0^2}(k+1).$$

This finally gives

$$\Delta_k \leq \frac{1}{\frac{1}{\Delta_0} + \frac{\omega}{r_0^2}k} = \frac{\Delta_0 r_0^2}{r_0^2 + \Delta_0\omega k}.$$

□

To get a cleaner bound, we optimize the right hand side over the step-size h and get:

Corollary 20.3. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be convex and L-smooth with minimum $f^* = f(x^*)$ and let x^k be defined by*

$$x^{k+1} = x^k - \tfrac{1}{L}\nabla f(x^k).$$

Then it holds that

$$f(x^k) - f^* \leq \tfrac{2L\|x_0 - x^*\|_2^2}{k+4}.$$

We want to make $h(2-Lh)$ as large as possible, and this is the case for $h^* = 1/L$. Then $h^*(2-Lh^*) = 1/L$. Furthermore, use L-smoothness and $\nabla f(x^*) = 0$ to estimate $f(x^0) - f^* \leq \frac{L}{2}\|x^0 - x^*\|_2^2$. This simplifies the upper bound from Theorem 20.2 to $\frac{2L\|x_0 - x^*\|_2^2}{k+4}$.

Reading the result in a different way, we see that we need about $\mathcal{O}(1/\epsilon)$ iterations to reach a guaranteed accuracy of $f(x^k) - f^* \leq \epsilon$. Thus, to cut the distance to optimality in half, we need twice as many iterations.

For the strongly convex case, we can do much better:

Theorem 20.4. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be μ-strongly convex and L-smooth (i.e. with condition number $Q = L/\mu$) with minimum $f^* = f(x^*)$ and let x^k be defined by*

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

fulfills

$$\|x^k - x^*\|_2^2 \leq (1 - \tfrac{2h\mu L}{\mu+L})^k \|x^0 - x^*\|_2^2$$

for $0 < h \leq 2/(L+\mu)$. The right hand side is minimal for the step-size $h = \frac{2}{L+\mu}$ and in thsi case we get

$$\|x^k - x^*\|_2 \leq \Big(\frac{Q-1}{Q+1}\Big)^k \|x^0 - x^*\|_2,$$
$$f(x^k) - f^* \leq \frac{L}{2}\Big(\frac{Q-1}{Q+1}\Big)^{2k} \|x^0 - x^*\|_2^2.$$

Proof. Similar to the proof in the L-smooth case we arrive at

$$\begin{aligned} r_{k+1}^2 &= \|x^k - x^* - h\nabla f(x^k)\|_2^2 \\ &= r_k^2 - 2h\langle \nabla f(x^k) - \nabla f(x^*), x^k - x^*\rangle + h^2\|\nabla f(x^k)\|_2^2, \end{aligned}$$

but now we use Theorem 17.6 to bound

$$\begin{aligned} r_{k+1}^2 &\leq r_k^2 - \tfrac{2h\mu L}{\mu+L} r_k^2 - \tfrac{2h}{\mu+L}\|\nabla f(x^k)\|_2^2 + h^2\|\nabla f(x^k)\|_2^2 \\ &= (1 - \tfrac{2h\mu L}{\mu+L}) r_k^2 + h(h - \tfrac{2}{\mu+L})\|\nabla f(x^k)\|_2^2. \end{aligned}$$

Since $0 \leq h \leq 2/(\mu+L)$ we get $r_{k+1}^2 \leq (1 - \frac{2h\mu L}{\mu+L}) r_k^2$. To minimize the factor on the right hand, we simply need to choose h as large as possible, and this is $h = 2/(L+\mu)$. In this case we get

$$\begin{aligned} 1 - \frac{2h\mu L}{\mu+L} &= 1 - \frac{4\mu L}{(L+\mu)^2} = \frac{(L+\mu)^2 - 4\mu L}{(L+\mu)^2} \\ &= \frac{(L-\mu)^2}{(L+\mu)^2} = \Big(\frac{Q-1}{Q+1}\Big)^2. \end{aligned}$$

□

Reading this result in a different, we see that we need about $\mathcal{O}(\log(\epsilon))$ iterations to reach a guaranteed accuracy of $f(x^k) - f^* \leq \epsilon$. Thus, to cut the distance to optimality in half, we just need to add a constant number of iterations (and this number depends on the size of the contraction factor $((Q-1)/(Q+1))^2$). A similar claim is true for the distance to optimum $\|x^k - x^*\|_2$

If we compare our performance bounds to the wort case analysis from previous sections, we see:

- For the convex and L-smooth case, the worst case bound and bound for the gradient method combine to the sandwich inequality

  $$\tfrac{3L\|x^0-x^*\|_2^2}{32(k+1)^2} \leq f(x^k) - f^* \leq \tfrac{2L\|x_0-x^*\|_2^2}{k+4}.$$

  The most notable difference is, that the lower bound is $\mathcal{O}(k^{-2})$ while the upper bound is only $\mathcal{O}(k^{-1})$, and hence, of worse order. There could be different reasons for this. For example, our annoying function was not the worst possible or our analysis of the gradient method could be improved. But in fact, something else is true: The gradient method is not optimal for this class, but a slight adaptation of the method is!

- For the μ-strongly convex and L-smooth case, the worst case bound and the bound for the gradient method for the distance to optimum give

  $$\Big(\frac{\sqrt{Q}-1}{\sqrt{Q}+1}\Big)^{2k}\|x_0 - x^*\|_2^2 \leq \|x^k - x^*\|_2^2 \leq \Big(\frac{Q-1}{Q+1}\Big)^{2k}\|x^0 - x^*\|_2^2$$

  Both bounds are geometrically convergent, but since $Q > 1$ and the function $t \mapsto (t-1)/(t+1)$ is stricly increasing, the constant if the lower bound is always better than the one in the upper bound. For large values of Q, the difference get notably large. Here we ask ourselves as well: What is the reason for this discrepancy. Again it will turn out that there is simple adaption of the gradient method that will be optimal order.

For $Q = 100$, we get $(Q-1)/(Q+1) \approx 0.98$, i.e. a 2% improvement of $\|x^k - x^*\|$ in every iteration, while $(\sqrt{Q}-1)(\sqrt{Q}+1) \approx 0.82$ which is an 18% improvement. For $Q = 10000$ the improvement is 0.02% vs. 2%.

We will see the accelerated gradient methods in both cases in the next section.

21 Accelerated gradient descent

As we have seen, the gradient method does not match our lower bound both in the L-smooth and the μ-strongly convex and L-smooth case. Surprisingly, there are two very simple adaptions of the method that do at least match the optimal order. The methods are due to Nesterov and are of the following form: Initialize with $x^{-1}, x^0 \in \mathbb{R}^n$ and iterate

$$\begin{aligned} y^k &= x^k + \alpha_k(x^k - x^{k-1}) \\ x^{k+1} &= y^k - h_k \nabla f(y^k) \end{aligned} \tag{*}$$

for some extrapolation parameters α_k and step-sizes h_k.

Note that the first step is not a convex combination of x^k and x^{k+1}, but an *extrapolation*. From the place x^k of the kth iterate, we move some more in the direction of where we came from.

This method will not be a descent method (i.e. $f(x^k)$ is not decreasing in every step), but the next lemma will help us to show convergence nonetheless.

Lemma 21.1. *If f is convex and L-smooth, $0 < h \leq 1/L$ and $x^+ = x - h\nabla f(x)$ is a gradient step, then it holds for all y that*

$$f(x^+) + \tfrac{\|x^+ - y\|_2^2}{2h} \leq f(y) + \tfrac{\|x-y\|_2^2}{2h}.$$

Proof. Since x^+ is a gradient step, we know from previous proofs that it holds for L-smooth functions that $f(x^+) \leq f(x) - h(1 - \frac{Lh}{2})\|\nabla f(x)\|_2^2 \leq f(x) - \frac{h}{2}\|\nabla f(x)\|_2^2$. Using this we get

$$\begin{aligned} f(x^+) + \tfrac{\|x^+ - y\|_2^2}{2h} &= f(x^+) + \frac{\|x-y\|_2^2}{2h} - \langle \nabla f(x), x - y\rangle + \tfrac{h}{2}\|\nabla f(x)\|_2^2 \\ &\leq f(x) + \langle \nabla f(x), y - x\rangle + \frac{\|x-y\|_2^2}{2h} \\ &\leq f(y) + \frac{\|x-y\|_2^2}{2h}. \end{aligned}$$

□

Moreover, we write the extrapolation step as

$$y^k = x^k + \tfrac{t_k - 1}{t_{k+1}}(x^k - x^{k-1}) \tag{**}$$

which will simplify the proofs.

Proposition 21.2. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be convex and L-smooth with minimum $f^* = f(x^*)$, $x_0, x_{-1} \in \mathbb{R}^n$, $0 < h \leq 1/L$ and t_k such that $t_1 = 1$, $t_k \geq 1$ and $t_k^2 - t_{k+1}^2 + t_{k+1} \geq 0$. Then it holds that the sequence x^k generated by the accelerated gradient method (**), (*) fulfills*

$$f(x^k) - f^* \leq \frac{\|x^0 - x^*\|_2^2}{2ht_k^2}.$$

Proof. We use Lemma 21.1 with the special points $x = y^k$, $x^+ = x^{k+1}$ and $y = (1 - \frac{1}{t_{k+1}})x^k + \frac{1}{t_{k+1}}x^*$ and get

$$\begin{aligned} &f(x^{k+1}) + \tfrac{\|x^{k+1} - (1 - \frac{1}{t_{k+1}})x^k - \frac{1}{t_{k+1}}x^*\|^2}{2h} \\ &\qquad \leq f\Big((1 - \tfrac{1}{t_{k+1}})x^k + \tfrac{1}{t_{k+1}}x^*\Big) + \tfrac{\|y^k - (1 - \frac{1}{t_{k+1}})x^k - \frac{1}{t_{k+1}}x^*\|^2}{2h}. \end{aligned}$$

We define the auxiliary variable

$$u^{k+1} = x^k + t_{k+1}(x^{k+1} - x^k)$$

and get

$$x^{k+1} - (1 - \tfrac{1}{t_{k+1}})x^k - \tfrac{1}{t_{k+1}}x^* = \tfrac{1}{t_{k+1}}(u^{k+1} - x^*)$$

and

$$\begin{aligned} y^k - (1 - \tfrac{1}{t_{k+1}})x^k - \tfrac{1}{t_{k+1}}x^* &= x^k + \tfrac{t_k - 1}{t_{k+1}}(x^k - x^{k-1}) - (1 - \tfrac{1}{t_{k+1}})x^k - \tfrac{1}{t_{k+1}}x^* \\ &= \tfrac{1}{t_{k+1}}\Big((x^k + (t_k - 1)(x^k - x^{k-1}) - x^*)\Big) \\ &= \tfrac{1}{t_{k+1}}(u^k - x^*). \end{aligned}$$

We plug this in the first equality, use convexity of f (recall that $t_k \geq 1$, i.e. $1/t_k \leq 1$) and get

$$f(x^{k+1}) + \tfrac{\|u^{k+1} - x^*\|^2}{2ht_{k+1}^2} \leq (1 - \tfrac{1}{t_{k+1}})f(x^k) + \tfrac{1}{t_{k+1}}f(x^*) + \tfrac{\|u^k - x^*\|^2}{2ht_{k+1}^2}.$$

We rearrange to

$$f(x^{k+1}) - f^* - (1 - \tfrac{1}{t_{k+1}})(f(x^k) - f^*) \leq \tfrac{\|u^k - x^*\|^2}{2ht_{k+1}^2} - \tfrac{\|u^{k+1} - x^*\|^2}{2ht_{k+1}^2}.$$

Using the abbreviations

$$f_k = f(x^k) - f^*, \quad v_k = \|u^k - x^*\|^2$$

we get, multiplying by t_{k+1}^2

$$t_{k+1}^2 f_{k+1} - (t_{k+1}^2 - t_{k+1})f_k \leq \frac{v_k - v_{k+1}}{2h}.$$

We sum these inequalities and obtain

$$t_{K+1}^2 f_{K+1} + \sum_{k=0}^{K}(t_k^2 - t_{k+1}^2 + t_{k+1})f_k \leq \frac{v_0 - v_K}{2h}.$$

Since the coefficients in the sum are, by assumption, non-negative and $v_K \geq 0$, we have (using $t_1 = 1$)

$$f(x^{K+1}) - f^* \leq \frac{v_0}{2ht_{K+1}^2} = \frac{\|u^0 - x^*\|^2}{2ht_{K+1}^2} = \frac{\|x^0 - x^*\|^2}{2ht_{K+1}^2},$$

which proves the claim. □

We notice that the result needs the inequality $t_k^2 - t_{k+1}^2 + t_{k+1} \geq 0$ and that the upper bound decays faster, the quicker the sequence t_k grows. Hence, we would like the t_k to grow as fast as possible and hence we choose it such that the above inequality is always strict. This gives

$$t_{k+1} = \sqrt{t_k^2 + \tfrac{1}{4}} + \tfrac{1}{2}.$$

It's simple to implement this strategy in practice and one can show that $t_k \geq \frac{k+1}{2}$. If one wants a simpler value for the t_k, one could try to do a little worse in the inequality and omit the $+\frac{1}{4}$, leading to $t_{k+1} = t_k + \frac{1}{2}$.

Here is a result which gives a simple choice of t_k which directly translates to a choice of α_k:

Corollary 21.3. *The iterates of the accelerated gradient method* (*) *with* $\alpha_k = \frac{k-1}{k+a}$ *and* $a \geq 2$ *and stepsize* $0 < h \leq 1/L$ *fulfill*

$$f(x^k) - f^* \leq \frac{a^2\|x^0 - x^*\|_2^2}{2h(k+a-1)^2}.$$

Proof. The choice $t_k = \frac{k+a-1}{a}$ gives $\alpha_k = \frac{k-1}{k+a}$ and also fulfills

$$\begin{aligned} t_k^2 - t_{k+1}^2 + t_{k+1} &= \Big(\frac{k+a-1}{a}\Big)^2 - \Big(\frac{k+a}{a}\Big)^2 + \frac{k+a}{a} \\ &= \tfrac{1}{a^2}\Big((k+a-1)^2 - (k+a)^2 + a(k+a)\Big) \\ &= \tfrac{1}{a^2}\Big((k+a)^2 - 2(k+a) + 1 - (k+a)^2 + a(k+a)\Big) \\ &= \tfrac{1}{a^2}\Big((a-2)(k+a) + 1\Big) \end{aligned}$$

which is non-negative for $a \geq 2$. Plugging things into the result of Proposition 21.2 gives the result. □

The optimal values for h and α in the upper bound in this results are the largest possible h and the smallest possible a, i.e. $h = 1/L$ and $a = 2$, i.e. $\alpha_k = \frac{k-1}{k+2}$ leading the upper bound

$$f(x^k) - f^* \leq \frac{2L\|x^0 - x^*\|_2^2}{(k+1)^2}.$$

This upper bound is actually pretty close to our lower bound $\frac{3L\|x^0-x^*\|_2^2}{32(k+1)^2}$.

Now we turn to the case of L-smooth and μ-strongly convex functions. In this case a constant value for α_k works well:

Theorem 21.4. *Let* $f : \mathbb{R}^n \to \mathbb{R}$ *be convex,* L*-smooth and* μ*-strongly convex (i.e. the condition number is* $Q = L/\mu$*) with minimum* $f^* = f(x^*)$*. Then it holds that for any initialization* $x^0, y^0 \in \mathbb{R}^n$ *the iterates of the accelerated gradient method*

$$\begin{aligned} x^{k+1} &= y^k - \tfrac{1}{L}\nabla f(y^k), \\ y^{k+1} &= x^{k+1} + \tfrac{\sqrt{Q}-1}{\sqrt{Q}+1}(x^{k+1} - x^k), \end{aligned}$$

fulfill

$$\begin{aligned} f(x^k) - f^* &\leq \big(1 - \tfrac{1}{\sqrt{Q}}\big)^k \tfrac{\mu+L}{2}\|x^0 - x^*\|_2^2, \\ \|x^k - x^*\|_2 &\leq \big(1 - \tfrac{1}{\sqrt{Q}}\big)^{k/2}\sqrt{Q+1}\|x^0 - x^*\|_2. \end{aligned}$$

Proof. The proof is quite technical and consists of several steps.

1) We define
$$\Phi_0(x) = f(y^0) + \tfrac{\mu}{2}\|x - y^0\|^2$$
and proceed recursively as
$$\Phi_{k+1}(x) = \Big(1 - \tfrac{1}{\sqrt{Q}}\Big)\Phi_k(x) + \tfrac{1}{\sqrt{Q}}\big(f(y_k) + \langle \nabla f(y^k), x - y^k\rangle + \tfrac{\mu}{2}\|x - y^k\|^2\big)$$
Since f is μ-strongly convex we get
$$\Phi_{k+1}(x) \leq \Big(1 - \tfrac{1}{\sqrt{Q}}\Big)\Phi_k(x) + \tfrac{1}{\sqrt{Q}} f(x).$$

2) We claim that
$$\Phi_k(x) \leq f(x) + \Big(1 - \tfrac{1}{\sqrt{Q}}\Big)^k (\Phi_0(x) - f(x)).$$
The case $k = 0$ is clear and to show the claim by induction we calculate
$$\begin{aligned}\Phi_{k+1}(x) &\leq \Big(1 - \tfrac{1}{\sqrt{Q}}\Big)\Phi_k(x) + \tfrac{1}{\sqrt{Q}} f(x)\\ &\leq \Big(1 - \tfrac{1}{\sqrt{Q}}\Big)\Big[f(x) + \Big(1 - \tfrac{1}{\sqrt{Q}}\Big)^k (\Phi_0(x) - f(x))\Big] + \tfrac{1}{\sqrt{Q}} f(x)\\ &= f(x) + \Big(1 - \tfrac{1}{\sqrt{Q}}\Big)^{k+1} (\Phi_0(x) - f(x)).\end{aligned}$$

3) Now we claim that $f(x^k) \leq \min_x \Phi_k(x)$ and show this by induction again:
For $k = 0$ it holds $\Phi_0(x) = f(y^0) + \frac{\mu}{2}\|x - y^0\|^2$ which is minimal for $x = y^0 = x^0$.
Now denote $\Phi_k^* = \min_x \Phi_k(x)$ and since x^{k+1} is a gradient step from y^k with stepsize $1/L$ we get
$$\begin{aligned}f(x^{k+1}) &\leq f(y^k) - \tfrac{1}{2L}\|\nabla f(y^k)\|^2\\ &= \Big(1 - \tfrac{1}{\sqrt{Q}}\Big)\underbrace{f(x^k)}_{\leq \Phi_k^*} + \Big(1 - \tfrac{1}{\sqrt{Q}}\Big)\underbrace{(f(y^k) - f(x^k))}_{\leq \langle \nabla f(y^k), y^k - x^k\rangle} + \tfrac{f(y^k)}{\sqrt{Q}} - \tfrac{1}{2L}\|\nabla f(y^k)\|^2\end{aligned}$$
Hence, we need to show that
$$\Phi_{k+1}^* \geq \Big(1 - \tfrac{1}{\sqrt{Q}}\Big)\Phi_k^* + \Big(1 - \tfrac{1}{\sqrt{Q}}\Big)\langle \nabla f(y^k), y^k - x^k\rangle + \tfrac{f(y^k)}{\sqrt{Q}} - \tfrac{1}{2L}\|\nabla f(y^k)\|^2 \tag{3}$$
We show this is several steps:

a) For all x and k we have $\nabla^2\Phi_k(x) = \mu I_n$: For $k = 0$ this is clear by definition and a close look on the recursive definition shows that this property stays true for Φ_k as well.

b) Hence, Φ_k can be written as $\Phi_k(x) = \Phi_k^* + \frac{\mu}{2}\|x - v^k\|^2$ with some v^k. One can see that
$$v^{k+1} = \Big(1 - \tfrac{1}{\sqrt{Q}}\Big)v^k + \tfrac{1}{\sqrt{Q}} y^k - \tfrac{1}{\mu\sqrt{Q}}\nabla f(y^k).$$

We take the derivative of Φ_{k+1} to get
$$\begin{aligned}\nabla\Phi_{k+1}(x) &= \Big(1 - \tfrac{1}{\sqrt{Q}}\Big)\nabla\Phi_k(x) + \tfrac{1}{\sqrt{Q}}\nabla f(y^k)\\ &\quad + \tfrac{\mu}{\sqrt{Q}}(x - y^k)\\ &= \Big(1 - \tfrac{1}{\sqrt{Q}}\Big)\mu(x - v^k) + \tfrac{1}{\sqrt{Q}}\nabla f(y^k)\\ &\quad + \tfrac{\mu}{\sqrt{Q}}(x - y^k)\end{aligned}$$
The condition $0 = \nabla\Phi_{k+1}(v^{k+1})$ implies the recursion for v_k.

By the recursive definition of Φ_{k+1} we get

$$\begin{aligned}\Phi^*_{k+1} + \tfrac{\mu}{2}\|y^k - v^{k+1}\|^2 &= \Phi_{k+1}(y^k)\\ &= \Big(1-\tfrac{1}{\sqrt{Q}}\Big)\Big(\Phi^*_k + \tfrac{\mu}{2}\|y^k - v^k\|^2\Big) + \tfrac{1}{\sqrt{Q}}f(y^k)\\ &= \Big(1-\tfrac{1}{\sqrt{Q}}\Big)\Phi^*_k + \tfrac{\mu}{2}\Big(1-\tfrac{1}{\sqrt{Q}}\Big)\|y^k - v^k\|^2 + \tfrac{1}{\sqrt{Q}}f(y^k) \qquad (4)\end{aligned}$$

and the recursion for v^{k+1} gives

$$\begin{aligned}\|y^k - v^{k+1}\|^2 &= \|\Big(1-\tfrac{1}{\sqrt{Q}}\Big)(y^k - v^k) - \tfrac{1}{\mu\sqrt{Q}}\nabla f(y^k)\|^2\\ &= \Big(1-\tfrac{1}{\sqrt{Q}}\Big)^2\|y^k - v^k\|^2 + \tfrac{1}{\mu^2 Q}\|\nabla f(y^k)\|^2\\ &\quad - \tfrac{2}{\mu\sqrt{Q}}\Big(1-\tfrac{1}{\sqrt{Q}}\Big)\langle\nabla f(y^k), v^k - y^k\rangle.\end{aligned}$$

We plug this into (4) and get

$$\begin{aligned}\Phi^*_{k+1} &= \Big(1-\tfrac{1}{\sqrt{Q}}\Big)\Phi^*_k + \tfrac{1}{\sqrt{Q}}f(y^k) + \tfrac{\mu}{2}\Big(1-\tfrac{1}{\sqrt{Q}}\Big)\|y^k - v^k\|^2\\ &\quad - \tfrac{\mu}{2}\Big(1-\tfrac{1}{\sqrt{Q}}\Big)^2\|y^k - v^k\|^2 - \tfrac{1}{2\mu Q}\|\nabla f(y^k)\|^2\\ &\quad + \tfrac{1}{\sqrt{Q}}\Big(1-\tfrac{1}{\sqrt{Q}}\Big)\langle\nabla f(y^k), v^k - y^k\rangle\\ &= \Big(1-\tfrac{1}{\sqrt{Q}}\Big)\Phi^*_k + \tfrac{1}{\sqrt{Q}}f(y^k) + \tfrac{\mu}{2\sqrt{Q}}\Big(1-\tfrac{1}{\sqrt{Q}}\Big)\|y^k - v^k\|^2\\ &\quad - \tfrac{1}{2L}\|\nabla f(y^k)\|^2 + \tfrac{1}{\sqrt{Q}}\Big(1-\tfrac{1}{\sqrt{Q}}\Big)\langle\nabla f(y^k), v^k - y^k\rangle.\end{aligned}$$

Here we used $(1-\frac{1}{\sqrt{Q}}) - (1-\frac{1}{\sqrt{Q}})^2 = \frac{1}{\sqrt{Q}}(1-\frac{1}{\sqrt{Q}})$.

c) We claim that $v^k - y^k = \sqrt{Q}(y^k - x^k)$. For $k = 0$ both sides are zero and recursively we get

$$\begin{aligned}v^{k+1} - y^{k+1} &= \Big(1-\tfrac{1}{\sqrt{Q}}\Big)v^k + \tfrac{1}{\sqrt{Q}}y^k - \tfrac{1}{\mu\sqrt{Q}}\nabla f(y^k) - y^{k+1}\\ &= \Big(1-\tfrac{1}{\sqrt{Q}}\Big)(y^k + \sqrt{Q}(y^k - x^k)) + \tfrac{1}{\sqrt{Q}}y^k - \tfrac{\sqrt{Q}}{L}\nabla f(y^k) - y^{k+1}\\ &= \sqrt{Q}(y^{k+1} - x^{k+1}).\end{aligned}$$

We used $x^{k+1} = y^k - \frac{1}{L}\nabla f(y^k)$ and $(\sqrt{Q}-1)x^k = 2\sqrt{Q}x^{k+1} - (\sqrt{Q}-1)y^{k+1}$ which we get from the extrapolation step.

This shows

$$\begin{aligned}\Phi^*_{k+1} &= \Big(1-\tfrac{1}{\sqrt{Q}}\Big)\Phi^*_k + \tfrac{1}{\sqrt{Q}}f(y^k) + \tfrac{\mu\sqrt{Q}}{2}\Big(1-\tfrac{1}{\sqrt{Q}}\Big)\|y^k - x^k\|^2\\ &\quad - \tfrac{1}{2L}\|\nabla f(y^k)\|^2 + \Big(1-\tfrac{1}{\sqrt{Q}}\Big)\langle\nabla f(y^k), y^k - x^k\rangle.\end{aligned}$$

This finally shows the inequality (3) and hence

$$f(x^k) \leq \Phi^*_k.$$

4) Using step 2), the definition of Φ_0 and that ∇f is L-Lipschitz

(more concretely, we use $f(y^0) - f^* \leq \frac{L}{2}\|y^0 - x^*\|^2$)

$$\begin{aligned}
f(x^k) - f^* &\leq \Phi_k^* - f^* \leq \Phi_k(x^*) - f^* \\
&\leq f(x^*) + \Big(1 - \tfrac{1}{\sqrt{Q}}\Big)^k (\Phi_0(x^*) - f(x^*)) - f^* \\
&= \Big(1 - \tfrac{1}{\sqrt{Q}}\Big)^k \Big(f(y^0) + \tfrac{\mu}{2}\|x^* - y^0\|^2 - f^*\Big) \\
&\leq \Big(1 - \tfrac{1}{\sqrt{Q}}\Big)^k \tfrac{\mu+L}{2}\|y^0 - x^*\|^2.
\end{aligned}$$

5) The second inequality simply follows since f is μ-strongly convex and thus

$$\|x^k - x^*\|^2 \leq \tfrac{2}{\mu}(f(x^k) - f^*). \qquad \square$$

We note the slightly different upper and lower bound here: We have

$$\Big(\frac{\sqrt{Q}-1}{\sqrt{Q}+1}\Big)^{2k}\|x_0 - x^*\|_2^2 \leq \|x^k - x^*\|_2^2 \leq (1 - \tfrac{1}{\sqrt{Q}})^k \sqrt{Q+1}\|x^0 - x^*\|_2^2$$

However, since

$$\begin{aligned}
\Big(1 - \tfrac{1}{\sqrt{Q}}\Big)^k &\leq \exp(-\tfrac{k}{\sqrt{Q}}), \\
\Big(\frac{\sqrt{Q}-1}{\sqrt{Q}+1}\Big)^{2k} &\geq \exp(-\tfrac{4k}{\sqrt{Q}-1}),
\end{aligned}$$

both bounds are quite close.

One the one hand, we can raise $(1+t) \leq \exp(t)$ to the k-th power and set $t = -1/r$ to get $(1-1/r)^k \leq \exp(-k/r)$, and on the other hand, we rearrange to $\exp(-t) \leq 1/(1+t)$ and set $t = 2/(\sqrt{Q}-1)$ to get $\exp(-2/(\sqrt{Q}-1) \leq 1/(1+2/(\sqrt{Q}-1)) = (\sqrt{Q}-1)/(\sqrt{Q}+1)$.

We have seen that a simple extrapolation step turns the gradient method into a method that is of optimal order in both the convex and L-smooth case and the L-smooth and μ-strongly convex case. However, one needs to choose the extrapolation step differently in both cases.

22 Analysis of the proximal gradient method and its acceleration

In this section we analyze the convergence of the proximal gradient method. We recall the setup for this method: For some convex and lsc $f : \mathbb{R}^d \to \mathbb{R}$ and a convex and L-smooth g we consider the problem

$$\min_x f(x) + g(x).$$

We assume that we can evaluate the proximal operator prox_{hf} of f for any $h > 0$ and the gradient ∇g at some point in every iteration. The proximal gradient method alternates a gradient step for g and a proximal step for f:

$$x^{k+1} = \operatorname{prox}_{hf}(x^k - h\nabla g(x^k)).$$

To analyze the method we introduce the map

$$G_h(x) = \tfrac{1}{h}\Big(x - \operatorname{prox}_{hf}(x - h\nabla g(x))\Big)$$

and call it the *gradient map* of the proximal gradient method. By definition we have

$$\operatorname{prox}_{hf}(x - h\nabla g(x)) = x - hG_h(x)$$

and hence, the gradient map is the "(additive) step of the proximal gradient method".

Lemma 22.1. *For all $h > 0$ it holds that*

$$G_h(x) - \nabla g(x) \in \partial f(x - hG_h(x))$$

and if g is L-smooth, then for $0 < h \leq 1/L$ it holds that

$$g(x - hG_h(x)) \leq g(x) - h\langle \nabla g(x), G_h(x)\rangle + \tfrac{h}{2}\|G_h(x)\|^2.$$

Proof. We recall that $u = \operatorname{prox}_{hf}(x)$ iff and only if $(x-u)/h \in \partial f(u)$. Using this with $x - hG_h(x)$ for u and $x - h\nabla g(x)$ for x shows the first claim. For the second claim we use Lemma 12.4 with $y = x - hG_h(x)$ to get

$$g(x - hG_h(x)) \leq g(x) - h\langle \nabla g(x), G_h(x)\rangle + \tfrac{h^2L}{2}\|G_h(x)\|_2^2$$

from which the inequality follows. □

Lemma 22.2. *Let f be convex and lsc and g be convex and L-smooth. Then it holds for $F = f + g$, $0 < h \leq 1/L$ and all z that*

$$F(x - hG_h(x)) \leq F(z) + \langle G_h(x), x - z\rangle - \tfrac{h}{2}\|G_h(x)\|_2^2.$$

Proof. By the second point in Lemma 22.1, we get the estimate

$$\begin{aligned}F(x-hG_h(x)) &= f(x-hG_h(x)) + g(x-hG_h(x))\\ &\leq f(x-hG_h(x)) + g(x) - h\langle\nabla g(x), G_h(x)\rangle + \tfrac{h}{2}\|G_h(x)\|_2^2\end{aligned}$$

and by the first point of the same lemma we know (by the subgradient inequality) that

$$\begin{aligned}f(z) &\geq f(x-hG_h(x)) + \langle G_h(x)-\nabla g(x), z-x+hG_h(x)\rangle\\ &= f(x-hG_h(x)) + h\|G_h(x)\|_2^2 + \langle G_h(x)-\nabla g(x), z-x\rangle - h\langle\nabla g(x), G_h(x)\rangle.\end{aligned}$$

Combining these inequalities we get (using convexity of g in the second inequality)

$$\begin{aligned}F(x-hG_h(x)) &\leq f(z) + g(x) + \langle G_h(x)-\nabla g(x), x-z\rangle - \tfrac{h}{2}\|G_h(x)\|_2^2\\ &\leq f(z) + g(z) + \langle G_h(x), x-z\rangle - \tfrac{h}{2}\|G_h(x)\|_2^2\end{aligned}$$

as claimed. □

Proposition 22.3. *Let f be convex and lsc and g be convex and L-smooth. Then it holds for $F = f+g$, $x^* = \operatorname{argmin} F$, $F^* = F(x^*)$, and $x^+ = \operatorname{prox}_{hf}(x - h\nabla g(x))$ and $0 < h \leq 1/L$ that*

i) For all y we have $F(x^+) + \frac{1}{2h}\|x^+ - y\|_2^2 \leq F(y) + \frac{1}{2h}\|x-y\|_2^2$.

ii) $F(x^+) \leq F(x) - \frac{h}{2}\|G_h(x)\|_2^2$.

iii) $F(x^+) - F^* \leq \frac{1}{2h}(\|x-x^*\|_2^2 - \|x^+ - x^*\|_2^2)$.

iv) $\|x^+ - x^*\|_2 \leq \|x-x^*\|_2$.

Note that this is exactly the same inequality than the one for the gradient step in Lemma 21.1

Proof. We have $x^+ = x - hG_h(x)$ and by Lemma 22.2 we have by completing the square

$$\begin{aligned}F(x^+) &\leq F(y) + \langle G_h(x), x-y\rangle - \tfrac{h}{2}\|G_h(x)\|_2^2\\ &= F(y) - \tfrac{1}{2h}\|x-hG_h(x)-y\|_2^2 + \tfrac{1}{2h}\|x-y\|_2^2\end{aligned}$$

and point i) follows. Point ii) follows with $y = x$, point iii) with $y = x^*$ and point iv) follows from iii) since the left hand side is non-negative. □

Theorem 22.4. *Let f be convex and lsc and g be convex and L-smooth. Then it holds for $F = f+g$, $x^* = \operatorname{argmin} F$, $F^* = F(x^*)$, that the iterates $x^{k+1} = \operatorname{prox}_{hf}(x^k - h\nabla g(x^k)$ with $h = 1/L$ fulfill*

$$F(x^k) - F^* \leq \tfrac{L\|x^0 - x^*\|_2^2}{2k}.$$

Proof. Using point iii) in Proposition 22.3 with $x^+ = x^{k+1}$ and $x = x^k$ we get

$$F(x^{k+1}) - F^* \leq \tfrac{L}{2}(\|x^k - x^*\|_2^2 - \|x^{k+1} - x^*\|_2^2)$$

and summing this inequality we get

$$\sum_{i=1}^{k+1}\left(F(x^i) - F^*\right) \leq \tfrac{L}{2}\left(\|x^0 - x^*\|_2^2 - \|x^{k+1} - x^*\|_2^2\right)$$

Since $F(x^i)$ is decreasing (Proposition) 22.3 ii)) we have

$$F(x^{k+1}) - F^* \leq \tfrac{1}{k+1}\sum_{i=1}^{k+1}\left(F(x^i) - F^*\right) \leq \tfrac{L}{2(k+1)}\left(\|x^0 - x^*\|_2^2 - \|x^{k+1} - x^*\|_2^2\right)$$

□

Interestingly, the same acceleration that works for the gradient method also works for the proximal gradient. This generalization is due to Amir Beck and Marc Teboulle and it goes as follows: Initialize with $x^{-1} = x^0 \in \mathbb{R}^d$ and iterate

$$\begin{aligned} y^k &= x^k + \alpha_k(x^k - x^{k-1})\\ x^{k+1} &= \operatorname{prox}_{h_k f}(y^k - h_k\nabla g(y^k)). \end{aligned}$$

for some sequence α_k of positive extrapolation parameters and some stepsizes $h_k > 0$. We will use the following reformulation (which we already used in the proof of Proposition 21.2) of the method: We initialize with $y^0 = x^0 \in \mathbb{R}^d$ and iterate

$$\begin{aligned} x^k &= \operatorname{prox}_{h_k f}(y^{k-1} - h_k\nabla g(y^{k-1}))\\ u^k &= t_k x^k + (1 - t_k)x^{k-1} \\ y^k &= (1 - \tfrac{1}{t_{k+1}})x^k + \tfrac{1}{t_{k+1}}u^k. \end{aligned} \tag{*}$$

Lemma 22.5. *If $\alpha_k = \frac{t_k - 1}{t_{k+1}}$, then the iterates of the sequence x^k generated by (*) are equal to the iterated of the accelerated proximal gradient method.*

Proof. We take the iteration (*) and plug u^k in the definition of y^k and obtain

$$\begin{aligned} y^k &= (1 - \tfrac{1}{t_{k+1}})x^k + \tfrac{1}{t_{k+1}}\left(t_k x^k + (1 - t_k)x^{k-1}\right)\\ &= x^k + \tfrac{t_k - 1}{t_{k+1}}(x^k - x^{k-1}) \end{aligned}$$

as desired. □

Proposition 22.6. *Let f convex and lsc, g be convex and L-smooth, $F = f + g$ and let x^* be a minimizer of F and $F^* = F(x^*)$. For the sequences x^k, u^k, y^k generated by (*) with $h_k = h \in]0, 1/L]$ it holds that*

$$t_{k+1}^2(F(x^{k+1}) - F^*) + \sum_{i=1}^{k}(t_i^2 - t_{i+1}^2 + t_{i+1})(F(x^i) - F^*) \leq \tfrac{1}{2h}(\|u^0 - x^*\|_2^2 - \|u^{k+1} - x^*\|_2^2).$$

Proof. We start with Proposition 22.3 i) with $x = y^k$, $x^+ = x^{k+1}$ and the special point $z = (1 - \frac{1}{t_{k+1}})x^k + \frac{1}{t_{k+1}}x^*$ to get

$$\begin{aligned} F(x^{k+1}) &+ \tfrac{1}{2h}\|x^{k+1} - (1 - \tfrac{1}{t_{k+1}})x^k - \tfrac{1}{t_{k+1}}x^*\|_2^2 \\ &\leq F\Big((1 - \tfrac{1}{t_{k+1}})x^k + \tfrac{1}{t_{k+1}}x^*\Big) + \tfrac{1}{2h}\|y^k - (1 - \tfrac{1}{t_{k+1}})x^k + \tfrac{1}{t_{k+1}}x^*\|_2^2 \\ &\qquad = F\Big((1 - \tfrac{1}{t_{k+1}})x^k + \tfrac{1}{t_{k+1}}x^*\Big) + \tfrac{1}{2h}\|\tfrac{1}{t_{k+1}}(u^k - x^*)\|_2^2. \end{aligned}$$

Convexity of F gives

$$F\Big((1 - \tfrac{1}{t_{k+1}})x^k + \tfrac{1}{t_{k+1}}x^*\Big) \leq (1 - \tfrac{1}{t_{k+1}})F(x^k) + \tfrac{1}{t_{k+1}}F(x^*)$$

and since $\frac{1}{t_{k+1}}u^k = x^{k+1} - (1 - \frac{1}{t_{k+1}})x^k$ we get

$$\big(F(x^{k+1}) - F^*\big) - (1 - \tfrac{1}{t_{k+1}})\big(F(x^k) - F^*\big) \leq \tfrac{1}{2ht_{k+1}^2}\big(\|u^k - x^*\|_2^2 - \|u^{k+1} - x^*\|_2^2\big).$$

Multiplying by t_{k+1}^2 and summing up this inequality gives the assertion.

□

We again see that the quantity $t_i^2 - t_{i+1}^2 + t_{i+1}$ plays an important role here. Similarly to the accelerated gradient method from the previous section we can deduce (since $u^0 = x^0$)

$$F(x^{k+1}) - F^* \leq \tfrac{\|x^0 - x^*\|_2^2}{2ht_{k+1}^2}.$$

as soon as $t_i^2 - t_{i+1}^2 + t_{i+1} \geq 0$. We've seen that, for example $t_k = (k+1)/2$ is a valid choice which leads to:

Theorem 22.7. *If we choose $h = 1/L$, and $t_k = (k+1)/2$ in the previous proposition, we have*

$$F(x^{k+1}) - F^* \leq \tfrac{2L\|x^0 - x^*\|_2^2}{(k+2)^2}.$$

Note that this estimate is significantly better than the one for the standard proximal gradient method, which only guaranteed that $F(x^{k+1}) - F^* \leq \frac{L\|x^0 - x^*\|_2^2}{2(k+1)}$. Another way of phrasing this, is to ask, how many iterations one needs to guarantee that $F(x^k) - F^* \leq \epsilon$ holds. For the proximal gradient method, this holds if

$$k \geq C\epsilon^{-1}$$

while for Nesterov's accelerated proximal gradient method, we get this for

$$k \geq C\epsilon^{-1/2}$$

(with a different constant) and $\epsilon^{-1/2}$ grows notably slower than ϵ^{-1} for $\epsilon \to 0$.

23 Monotone operators

From this point on, we will formulate the theory in the context of general real and finite dimensional Hilbert space X. That means that X is a real vector space, equipped with an inner product $\langle x, y\rangle$ and which is complete with respect to the induced norm $\|x\| = \sqrt{\langle x, x\rangle}$. Everything which we did so far still holds true and we will consider the finite dimensional case throughout. Note that finite dimensional Hilbert spaces are all isomorphic to $\mathbb{R}^d$ equipped with an inner product of the form $\langle x, y\rangle = x^T My$ for some symmetric positive definite $M \in \mathbb{R}^{d\times d}$.

In the infinite dimensional case, several things need to be adapted due to the reason that closed and bounded sets will not be compact anymore. Moreover, one has to use the notion of weak convergence and some arguments will get more involved.

One of the main differences between the usual gradient of a differentiable function and the subgradient of a convex, lsc function is, that the subgradient is, in general, a set and not a singleton. Hence, it seems natural, to consider set-valued maps and we will introduce such maps now.

A set valued map on X would be

$$A : X \to \mathfrak{P}(X),$$

i.e. $A(x)$ is a subset of X. However, we will have a different view on set valued maps: A set valued map is fully described, by its graph

$$\operatorname{gr} A := \{(x, y) \in X \times X \mid y \in A(x)\}$$

and, vice versa, every subset $\mathcal{A} \subset X \times X$ gives rise to a set valued map $f(x) = \{y \in X \mid (x, y) \in \mathcal{A}\}$.

Definition 23.1. A set valued map $A : X \rightrightarrows X$ is characterized by its graph

$$\operatorname{gr} A \subset X \times X,$$

and we write $y \in A(x)$ if $(x, y) \in \operatorname{gr} A$.

The *domain* of A is

$$\operatorname{dom} A := \{x \mid A(x) \neq \emptyset\}.$$

The inverse of a set value map always exists is defined by

$$\operatorname{gr} A^{-1} := \{(y, x) \in X \times X \mid y \in A(x)\},$$

in other words

$$x \in A^{-1}(y) \iff y \in A(x).$$

Algebraic operators of set valued maps are defined in the usual way: If $A, B : X \rightrightarrows X$ and $\alpha \in \mathbb{R}$ we define

We will often omit the parentheses and write Ax instead of $A(x)$, even though A in set-valued and non-linear.

$$\begin{aligned}
(A + B)(x) &:= A(x) + B(x) = \{y + y' \mid y \in A(x), y' \in B(x)\}\\
(\alpha A)(x) &:= \alpha A(x) = \{\alpha y \mid y \in A(x)\}\\
(B \circ A)(x) &:= \bigcup_{y\in A(x)} B(y)
\end{aligned}$$

One further notion that can be extended to set valued map, is the one of *monotonicity*: Recall that a function $f : \mathbb{R} \to \mathbb{R}$ is monotone, if $(x-y)(f(x)-f(y))$ always has the same sign (if the expression is non-negative, we have an increasing function, if it is non-positive, we have a decreasing function). For set valued operators one always chooses the increasing case and defines:

Definition 23.2. A set valued map $A : X \rightrightarrows X$ is *monotone*, if for all $y \in A(x)$ and $y' \in A(x')$ it holds that

$$\langle y - y', x - x' \rangle \geq 0.$$

The map A is called *strongly monotone* with constant $\mu > 0$ if $A - \mu I$ is monotone, or, put differently, for all $y \in A(x)$ and $y' \in A(x')$ it holds that

By I we denote the identity mapping, the $Ix = x$, e.g, $(A - \mu I)(x)$ is the set $A(x) - \mu x$.

$$\langle y - y', x - x' \rangle \geq \mu \|x - x'\|^2.$$

Finally, A is called *maximally monotone*, if there is no monotone map with a larger graph, i.e. for all $(x, y) \notin \operatorname{gr} A$ there exists $(x', y') \in \operatorname{gr} A$ such that

$$\langle x - x', y - y' \rangle < 0.$$

Lemma 23.3. *A mulivalued monotone map A is maximally monotone, if and only if it holds that whenever $\langle x - y, u - v \rangle \geq 0$ holds for all $(y, v) \in \operatorname{gr} A$, then $(x, u) \in \operatorname{gr} A$ (i.e. $u \in A(x)$).*

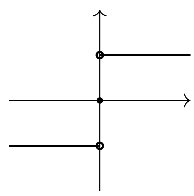
Proof. Let A be maximally monotone and assume that whenever $v \in Ay$ holds, we have $\langle x - y, u - v \rangle \geq 0$. But then $u \notin A(x)$ would contradict the maximality, since we could enlarge the graph $\operatorname{gr} A$ by adding the element (x, u) without destroying monotonicity.

For the reverse implication assume that A is monotone, but not maximally so. Then there is another monotone operator $\tilde{A}$ with a larger graph, i.e. there exist $(x, u) \in \operatorname{gr} \tilde{A}$ but $(x, u) \notin \operatorname{gr} A$. But then (since $u \in \tilde{A}x$ and $\tilde{A}$ is monotone) we still have $\langle x - y, u - v \rangle \geq 0$ for all $(y, v) \in \operatorname{gr} A \subset \operatorname{gr} \tilde{A}$, but not $(x, u) \in \operatorname{gr} A$. □

Example 23.4. Examples of a multivalued monotone maps arise as sign-functions. The ordinary sign-function

$$\operatorname{sign}(x) = \begin{cases} -1, & x < 0 \\ 0, & x = 0 \\ 1, & x > 0 \end{cases}$$

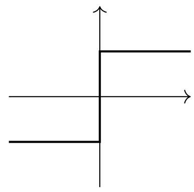
(when considered as a multivalued function with value $\{\operatorname{sign}(x)\}$) is monotone. However, it is not strongly monotone and not maximally monotone. The latter is true, since the *multivalued sign*



Sign : $\mathbb{R} \rightrightarrows \mathbb{R}$ defined by

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

has a graph that contains the graph of sign but is still monotone. The multivalued sign is maximally monotone, though. △



Note that $\mathrm{Sign}(x) = \partial|x|$.

Arguing that a monotone map is maximal using the definition of maximal monotonicity is not straightforward. The following proposition is sometimes easier to use.

Proposition 23.5. *If A is monotone and $I + A$ is onto, then A is maximally monotone.*

The reverse implication is also true and known as Minty's theorem, and we will prove it in the next section.

Proof. Assume that A is monotone and $I + A$ is onto. We aim to prove maximality of A by using the characterization in Lemma 23.3. Fix (x, u) such that for all $(y, v) \in \mathrm{gr}\, A$ it holds that

$$\langle x - y, u - v\rangle \geq 0.$$

If we can show that $u \in A(x)$ follows, the assertion follows from Lemma 23.3. Since $I + A$ is onto, there is a solution y of the inclusion $x + u \in (I + A)(y) = y + A(y)$. In other words, there exists a $v \in A(y)$ such that $x + u = y + v$. Since (u, x) fulfills our assumption, we know that

$$0 \leq \langle x - y, \underbrace{u - v}_{=y-x}\rangle = -\|x - y\|^2$$

and hence, $x = y$ and also $u = v$. Thus $u \in A(x)$ as claimed. □

Maximal monotonicity implies a certain closedness of the graph:

Lemma 23.6. *Let $A : X \rightrightarrows X$ be maximally monotone. Then it holds that $\mathrm{gr}\, A$ is closed, i.e. if $u_n \in Ax_n$ and $x_n \to x$ and $u_n \to u$, then $u \in Ax$.*

Proof. For any $(y, v) \in \mathrm{gr}\, A$ we have that $\langle x_n - y, u_n - v\rangle \geq 0$. Hence, this also holds in the limit, i.e. we have for all $(y, v) \in \mathrm{gr}\, A$ that $\langle x - y, u - v\rangle \geq 0$ and Lemma 23.3 shows that $u \in Ax$. □

Proposition 23.7. *Let $f : X \to \bar{\mathbb{R}}$ be proper and convex. Then ∂f is a monotone map. If f is also lsc., then ∂f is maximally monotone.*

Proof. If f is proper and convex and we have x, x' with subgradients $p \in \partial f(x)$, $p' \in \partial f(x')$, we can evaluate the subgradient inequalities at the other points and get

$$\begin{aligned} f(x') &\geq f(x) + \langle p, x' - x\rangle \\ f(x) &\geq f(x') + \langle p', x - x'\rangle. \end{aligned}$$

Adding these inequalities gives

$$0 \geq \langle p - p', x' - x \rangle$$

which is equivalent to $0 \leq \langle p - p', x - x' \rangle$.

If f is lsc, then by Theorem 8.2 $\frac{1}{2}\|y - x\|^2 + f(y)$ has a minimizer $z = \operatorname{prox}_f(x)$. Since $\|y - x\|^2$ is continuous and defined everywhere, the subgradient sum rule gives

$$0 \in z - x + \partial f(z)$$

which we rearrange to

$$x \in z + \partial f(z) = (I + \partial f)(z).$$

This shows that $(I + \partial f)$ is onto, and hence, by Proposition 23.5 ∂f is maximally monotone. □

The fact that subgradients of proper, convex and lsc functions are maximally monotone shows: A *convex optimization problem*

$$\min_{x \in X} f(x)$$

for a proper, convex and lsc function f is equivalent to the *monotone inclusion*

$$0 \in Ax$$

with $A = \partial f$. Hence, the study of monotone inclusions is worthwhile, since any algorithm which can be used to solve monotone inclusions can (in priciple) be turned into an algorithm for convex optimization problems.

Note that, however, not every maximally monotone operator is a subgradient. (Try to find an example. Hint: Consider linear maps on $\mathbb{R}^2$.) Hence, the class of monotone inclusions is in fact larger than the class of convex optimization problems.

To see that the Proposition 23.7 need not to hold for functions that aren't lsc, consider the function $f(x) = i_{]0,\infty[}(x)$. The subgradient is

$$\partial f(x) = \begin{cases} \emptyset, & x \leq 0 \\ \{0\}, & x > 0 \end{cases}$$

which is monotone. But it is not maximally monotone, since one can enlarge it's graph without destroying monotonicity. In fact, this can be done in different ways. Two extreme ways would be to consider the single-valued function

$$x \mapsto \{0\}$$

which is the subgradient of the zero function, of the multi-valued function

$$x \mapsto \begin{cases} \emptyset, & x < 0 \\ ]-\infty, 0], & x = 0 \\ \{0\}, & x > 0 \end{cases}$$

which is the subgradient of $i_{[0,\infty]}$.

Recall from Section 8, that the proximal map of a proper, convex and lsc function $f : X \to \bar{\mathbb{R}}$ is defined as

$$\operatorname{prox}_f(x) = \operatorname*{argmin}_y \tfrac{1}{2}\|x - y\|^2 + f(x).$$

Using Fermat's characterization of minimizers, the sum-rule for subgradient, and our notion of inverse for multivalued function we can characterize $z = \operatorname{prox}_f(x)$ as

$$\begin{aligned} & z = \operatorname*{argmin}_y \tfrac{1}{2}\|y - x\|^2 + f(y) \\ \iff & 0 \in z - x + \partial f(z) \\ \iff & x \in z + \partial f(z) = (I + \partial f)(z) \\ \iff & z = (I + \partial f)^{-1}(x), \end{aligned}$$

i.e. we have

$$\operatorname{prox}_f(x) = (I + \partial f)^{-1}(x).$$

This is handy way to figure out proximal mappings of one-dimensional functions graphically.

24 Resolvents and non-expansive operators

Here are few more useful facts about maximally monotone operators:

Proposition 24.1. *For every monotone operator $A : X \rightrightarrows X$, there esists a maximally monotone extesion $\bar{A} : X \rightrightarrows X$.*

Proof. We set

$$\mathcal{A} = \{A' : X \rightrightarrows X \mid A' \text{ monotone, } \operatorname{gr} A \subset \operatorname{gr} A'\}.$$

This set is partially ordered by graph inclusion and by Zorn's lemma there is a maximal ordered subset $\mathcal{A}_0$. We now define $\bar{A} = \bigcup_{A' \in \mathcal{A}_0} A'$ which is monotone and also maximal by definition. □

Proposition 24.2. *If a continuous map $A : X \to X$ is monotone, then it is maximally monotone.*

Proof. Assume that for some $(x', v') \in Ax$ it holds that $\langle v' - Ax, x' - x\rangle \geq 0$ for all x. We have to show that $v' = Ax'$. Then we take $x = x' - \epsilon u$ for some $\epsilon > 0$ and some $u \in X$. Then we have $\langle v' - A(x' - \epsilon u), u\rangle \geq 0$ since $A(x' + \epsilon u) \to Ax'$ for $\epsilon \to 0$ by continuity, we get that $\langle v' - Ax', u\rangle \geq 0$. Since we can do this for all u, we conclude that $v' - Ax' = 0$ as desired. □

We have just seen in the last section that $\operatorname{prox}_f = (I + \partial f)^{-1}$ and we already know that the proximal map is quite helpful when it comes to minimization problems. Hence, we define for monotone operators:

Definition 24.3. Let $A : X \rightrightarrows X$ be monotone. Then the *resolvent* of A is

$$J_A = (I + A)^{-1}.$$

Hence we can write $J_{\partial f} = \operatorname{prox}_f$.

Proposition 24.4. *Let $A : X \rightrightarrows X$ be a monotone map with* $\operatorname{dom} A \neq \emptyset$. *Then J_A is (at most) single-valued.*

Proof. Assume that J_A is has more than one value, i.e. there exist x_1, x_2 suchthat $x_i \in (I + A)^{-1}(y)$, i.e. $y \in (I + A)(x_i)$ for $i = 1, 2$. This means that

$$y - x_1 \in A(x_1), \qquad y - x_2 \in A(x_2).$$

But since A is monotone, we obtain

$$0 \leq \langle x_1 - x_2, y - x_1 - (y - x_2)\rangle = -\|x_1 - x_2\|^2$$

which shows $x_1 = x_2$. □

Definition 24.5. A map $T : X \to X$ is called *non-expansive* if for all x_1, x_2 it holds that

$$\|T(x_1) - T(x_2)\| \leq \|x_1 - x_2\|.$$

The map T is called a *contraction*, if

$$\|T(x_1) - T(x_2)\| \leq q\|x_1 - x_2\|$$

for some $q < 1$.

Non-expansiveness is the same as being Lipschitz continuous with constant one.

Similarly, a set valued mapping S is non-expansive, if for $y_1 \in S(x_1)$, $y_2 \in S(x_2)$ it holds that $\|y_1 - y_2\| \leq \|x_1 - x_2\|$.

We will see shortly, that the non-expansive set-valued mappings are actually always single valued.

(Maximally) monotone maps are related to non-expansive maps in several ways.

Proposition 24.6 (Minty parametrization)**.** *Let $J : X \times X \to X \times X$ be defined by*

$$J(x, v) = \begin{bmatrix} x + v \\ -x + v \end{bmatrix} = \begin{bmatrix} I & I \\ -I & I \end{bmatrix} \begin{bmatrix} x \\ v \end{bmatrix}, \quad J^{-1}(z, w) = \tfrac{1}{2} \begin{bmatrix} z - w \\ z + w \end{bmatrix}.$$

For two set-value mappings $A, B : X \rightrightarrows X$ assume that their graphs are related by

$$\operatorname{gr} B = J(\operatorname{gr} A), \qquad \operatorname{gr} A = J^{-1}(\operatorname{gr} B).$$

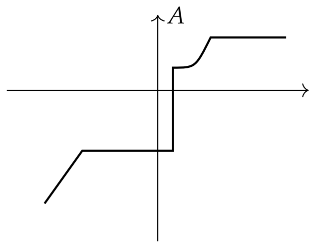
Then it holds

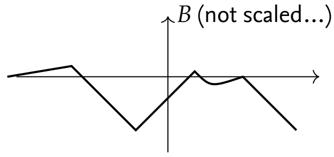
i) A is monotone if and only if B is non-expansive.

ii) We have

$$B = I - 2I \circ (I + A)^{-1}, \qquad A = (I - B)^{-1} \circ 2I - I.$$

For $n = 1$, the map J is a clockwise rotation by $\pi/4$ (and a scaling by $\sqrt{2}$)of the plane $\mathbb{R} \times \mathbb{R}$ in which the graphs live.





Proof. For $(z_j, w_j) = J(x_j, v_j) = (x_j + v_j, v_j - x_j)$, $j = 1, 2$ it holds that $z_j + w_j = 2v_j$ and $z_j - w_j = 2x_j$ and hence,

$$\begin{aligned} &\|z_2 - z_1\|^2 - \|w_2 - w_1\|^2 \\ &\quad = \langle (z_2 - z_1) + (w_2 - w_1), (z_2 - z_1) - (w_2 - w_1) \rangle \\ &\quad = \langle (z_2 + w_2) - (z_1 + w_1), (z_2 - w_2) - (z_1 - w_1) \rangle \\ &\quad = \langle 2v_2 - 2v_1, 2x_2 - 2x_1 \rangle = 4\langle v_2 - v_1, x_2 - x_1 \rangle. \end{aligned}$$

We conclude

$$\|w_2 - w_1\| \leq \|z_2 - z_1\| \iff \langle v_2 - v_1, x_2 - x_1 \rangle \geq 0.$$

This shows that B is non-expansive, if and only if A is monotone, i.e. i) is proven.

To prove ii), note that the condition $\operatorname{gr} B = J(\operatorname{gr} A)$ means that if $(z, w) \in J(\operatorname{gr} A)$, then for some $v \in Ax$ one has $z = v + x$ and $w = v - x$. This is equivalent to $z \in (I + A)x$ and $w = (z - x) - x = z - 2x$, Hence, we have $w \in Bz$ if and only if $w = z - 2x$ for some $x \in (I + A)^{-1}z$ which says that $B = I - 2I \circ (I + A)^{-1}$. For the second equality, we rearrange to $(I - B) = 2I \circ (I + A)^{-1}$ which leads to $(I + A) = (I - B)^{-1} \circ 2I$. $\square$

Theorem 24.7 (Minty). *If A is maximally monotone, then $I + \lambda A$ is onto for every $\lambda > 0$. Moreover, $(I + \lambda A)^{-1}$ is also maximally monotone and single valued.*

Proof. Of course, we can restrict to $\lambda = 1$. Since A is maximally monotone, the map B from Proposition 24.6 is non-expansive and also maximal in the sense that we can't enlarge its graph without destroying the non-expansiveness. Now we use the following result:

> Let $X \subset \mathbb{R}^n$ and $F : X \to \mathbb{R}^m$ be Lipschitz continuous. Then F has an extension $\bar{F} : \mathbb{R}^n \to \mathbb{R}^m$ which has the same Lipschitz constant.

The proof is fairly technical and uses Zorn's Lemma (hence, it is not constructive) and can be found in "Variational Analysis" by Rockafellar and Wets as Theorem 9.58.

By this result, we know that B is a maximal non-expansive mapping if and only if $\operatorname{dom} B = X$ which is the case if any only if $\operatorname{dom}(I + A)^{-1} = X$. Since $(I + A)^{-1}$ is single valued, continuous and monotone, it has to be maximally monotone. □

There is another graphical interpretation of the resolvant: For $\lambda > 0$ consider the map

$$M_\lambda : X \times X \to X \times X, \quad M_\lambda(x, v) = \begin{bmatrix} x + \lambda v \\ x \end{bmatrix}.$$

Since we have $x = J_{\lambda A} z$ (i.e. $(z, x) \in \operatorname{gr} J_{\lambda A}$) if and only if $\frac{z-x}{\lambda} \in Ax$ (i.e $(x, (z - x)/\lambda) \in \operatorname{gr} A$). And since $M_\lambda(x, (z - x)/\lambda) = (z, x)$ we see that

$$\operatorname{gr} J_{\lambda A} = M_\lambda \operatorname{gr} A.$$

Resolvents have special properties which make them useful to build algorithms, one of which is their non-expansiveness:

Definition 24.8. A map $A : X \to X$ is called *firmly non-expansive* if

$$\|A(x_1) - A(x_2)\|^2 + \|(I - A)(x_1) - (I - A)(x_2)\|^2 \le \|x_1 - x_2\|^2.$$

Obviously, a firmly non-expansive operator is also non-expansive, but there is more:

Proposition 24.9. *Let $D \subset X$ be non-empty and $A : D \to X$. Then the following are equivalent:*

i) *A is firmly non-expansive.*

ii) *$I - A$ is firmly non-expansive.*

iii) *$2A - I$ is non-expansive.*

iv) For all $x, y \in D$ it holds that

$$\|Ax - Ay\|^2 \leq \langle x - y, Ax - Ay\rangle.$$

v) For all $x, y \in D$ it holds that

$$0 \leq \langle Ax - Ay, (I - A)x - (I - A)y\rangle.$$

Proof. The equivalence of i) and ii) is clear, since the defining inequality from Definition 24.8 stays the same if we replace A by $I - A$.

For the equivalence of ii) and iii) we first observe the auxiliary identity

$$\|2u - v\|^2 = 2\|u - v\|^2 + 2\|u\|^2 - \|v\|^2.$$

(which can be verified, for example, by expanding both sides). With $u = x - y$ and $v = Ax - Ay$ we arrive at

$$\|(2A - I)x - (2A - I)y\|^2 = 2\|(A - I)x - (A - I)y\|^2 + 2\|Ax - Ay\|^2 - \|x - y\|^2$$

which shows the equivalence of ii) and iii).

For the equivalence of ii) and iv) use $\|(I - A)x - (I - A)y\|^2 = \|x - y\|^2 - 2\langle x - y, Ax - Ay\rangle + \|Ax - Ay\|$ and the equivalence of iv) and v) is clear. □

Proposition 24.10. *The resolvent J_A of a maximally monotone operator A is firmly non-expansive.*

Proof. Let $y_i = J_A(x_i)$, $i = 1, 2$, i.e

$$x_1 - y_1 \in A(y_1), \qquad x_2 - y_2 \in A(y_2).$$

By monotonicity we get

$$\langle x_1 - y_1 - x_2 + y_2, y_1 - y_2\rangle \geq 0$$

which leads to

$$0 \leq \langle y_1 - y_2, x_1 - x_2\rangle - \|y_1 - y_2\|^2.$$

Hence, we can conclude

$$\|J_A(x_1) - J_A(x_2)\|^2 \leq \langle J_A(x_1) - J_A(x_2), x_1 - x_2\rangle$$

and the previous proposition show s the result. □

Using Cauchy-Schwarz here would show non-expansiveness already.

The following lemma will be helpful later on.

Lemma 24.11. *If A is maximally monotone, it holds that*

$$AJ_A = I - J_A,$$

especially, the map AJ_A is single valued.

Proof. Let $u \in AJ_A y$. We add $J_A y$ on both sides and get $u + J_A y \in AJ_A y + J_A y = (I + A)J_A y$. Since $J_A = (I + A)^{-1}$, the right hand side is single valued and equal to y and we get $u = y - J_A y$ as claimed. □

25 Relaxed Mann iterations

We come back to the development of optimization algorithms and their analysis. First, we have an important notion that will help to ease the analysis.

Definition 25.1. Let $C \subset X$ be non-empty. A sequence (x_n) is called *Fejér monotone* with respect to C if

$$\forall x \in C : \|x_{n+1} - x\| \leq \|x_n - x\|.$$

Proposition 25.2. *Let $C \subset X$ be non-empty and (x_n) be Fejér monotone with respect to C. Then it holds:*

i) (x_n) is bounded.

ii) For every $x \in C$ it holds that sequence $\|x_n - x\|$ converges.

iii) The distance $d(x_n, C)$ is decreasing and converges.

iv) For m, n it holds hat $\|x_{n+m} - x_n\| \leq 2d(x_n, C)$.

Proof. Assertion i) follows, since $\|x_n - x\| \leq \|x_0 - x\|$ for some (even every) $x \in C$, and since $\|x_n - x\|$ is decreasing and bounded from below it converges which is ii). For iii) note that for every $x \in C$

$$d(x_{n+1}, C) = \inf_{y \in C} \|x_{n+1} - y\| \leq \|x_{n+1} - x\| \leq \|x_n - x\|.$$

Taking the infimum over x shows the claim. Finally, by the triangle inequality

$$\|x_{n+m} - x_n\| \leq \|x_{n+m} - x\| + \|x_n - x\| \leq 2\|x_n - x\|$$

and taking the infimum over x shows the claim. □

Definition 25.3. For a set X and a map $T : X \to X$ we denote by $\operatorname{Fix} T$ the set of fixed points of T, i.e. $\operatorname{Fix} T = \{x \mid Tx = x\}$.

Given a map $T : X \to X$, Banach's fixed point theorem states that the iteration $x_{n+1} = Tx_n$ converges to a fixed point if T is a *contraction*, i.e., when it's Lipschitz continuous with constant strictly smaller than one. In the case of a merely non-expansive map T this may fail (as can be seen by $T = -I$, for example). However, the following fundamental result on fixed point iterations state that we still get convergence under mild additional assumptions:

Theorem 25.4 (Krasnosel'skiĭ-Mann fixed point theorem). *Let $D \subset X$ be non-empty, closed and convex and $T : D \to D$ be non-expansive such that $\operatorname{Fix} T$ is not empty. For $x_0 \in D$ define the Mann iteration*

$$x_{n+1} = Tx_n, \; n = 0, 1, \ldots$$

and assume that $x_n - Tx_n \xrightarrow{n\to\infty} 0$. Then it (x_n) converges to a fixed point of T.

A sequence (x_n) of iterates that fullfills $x_n - Tx_n \to 0$ is called *asymtotically regular*.

Proof. For each $x \in \operatorname{Fix} T$ it holds that

$$\|x_{n+1} - x\| = \|Tx_n - Tx\| \leq \|x_n - x\|$$

and hence, (x_n) is Fejér monotone with respect to $\operatorname{Fix} T$ and hence, it's bounded and has cluster points. Now assume that x^* is any cluster point of (x_n), i.e. we have $x_{n_k} \xrightarrow{k\to\infty} x^*$.

Now we work along this subsequence and get

$$\begin{aligned}\|x^* - Tx^*\|^2 &= \|x_{n_k} - Tx^*\|^2 - \|x_{n_k} - x^*\|^2 - 2\langle x_{n_k} - x^*, x^* - Tx^*\rangle \\ &= \|x_{n_k} - Tx_{n_k}\|^2 + 2\langle x_{n_k} - Tx_{n_k}, Tx_{n_k} - Tx^*\rangle \\ &\quad + \|Tx_{n_k} - Tx^*\|^2 - \|x_{n_k} - x^*\|^2 - 2\langle x_{n_k} - x^*, x^* - Tx^*\rangle \\ &\leq \|x_{n_k} - Tx_{n_k}\|^2 + 2\langle x_{n_k} - Tx_{n_k}, Tx_{n_k} - Tx^*\rangle - 2\langle x_{n_k} - x^*, x^* - Tx^*\rangle.\end{aligned}$$

Using the assumption $x_{n_k} - Tx_{n_k} \to 0$, we get that all terms on the right hand side vanish in the limit $k \to \infty$ and this shows that $x^* = Tx^*$, i.e. $x^* \in \operatorname{Fix} T$.

By Proposition 25.2 iii) we know that $d(x_n, \operatorname{Fix} T)$ converges, but since $x_{n_k} \to x^* \in \operatorname{Fix} x$, $d(x_n, \operatorname{Fix} T)$ has to converge to 0. $\square$

The following theorem shows that even for merely non-expansive operators T we can get a method that converges towards a fixed point of T by *relaxation*:

Theorem 25.5 (Krasnosel'skiĭ-Mann iteration). *Let $D \subset X$ be non-empty, closed and convex and $T : D \to D$ be non-expansive with $\operatorname{Fix} T \neq \emptyset$. Let (λ_n) be a sequence in $[0,1]$ such that $\sum_{n=0}^{\infty} \lambda_n(1-\lambda_n)$ diverges. For some $x_0 \in D$ define the*

$$x_{n+1} = x_n + \lambda_n(Tx_n - x_n) = \lambda_n Tx_n + (1-\lambda_n)x_n.$$

Then it holds that

i) *(x_n) is Fejér monotone with respect to $\operatorname{Fix} T$,*

ii) $Tx_n - x_n \xrightarrow{n\to\infty} 0$,

iii) *(x_n) converges to some fixed point of T.*

Proof. Since D is convex, the sequence of iterates is alway well defined.

For any $y \in \operatorname{Fix} T$ we use the equation $\|\lambda u + (1-\lambda)v\|^2 + \lambda(1-\lambda)\|u - v\|^2 = \lambda\|u\|^2 + (1-\lambda)\|v\|^2$ to get

$$\begin{aligned}\|x_{n+1} - y\|^2 &= \|(1-\lambda_n)(x_n - y) + \lambda_n(Tx_n - y)\|^2 \\ &= (1-\lambda_n)\|x_n - y\|^2 + \lambda_n\|Tx_n - Ty\|^2 \\ &\quad - \lambda_n(1-\lambda_n)\|Tx_n - x_n\|^2 \\ &\leq \|x_n - y\|^2 - \lambda_n(1-\lambda_n)\|Tx_n - x_n\|^2\end{aligned}$$

and hence, (x_n) is Fejér monotone with respect to Fix T. Applying this inequality recursively, we get

$$\sum_{n=0}^{\infty} \lambda_n(1-\lambda_n)\|Tx_n - x_n\|^2 \leq \|x_0 - y\|^2.$$

Since the series over $\lambda_n(1-\lambda_n)$ diverges, we get that $\liminf_{n\to\infty}\|Tx_n - x_n\|^2 = 0$. However, since $x_{n+1} - x_n = \lambda_n(Tx_n - x_n)$ we get from the triangle inequality

$$\begin{aligned}\|Tx_{n+1} - x_{n+1}\| &= \|Tx_{n+1} - Tx_n + (1-\lambda_n)(Tx_n - x_n)\|\\ &\leq \underbrace{\|x_{n+1} - x_n\|}_{=\|\lambda_n(Tx_n - x_n)\|} + (1-\lambda_n)\|Tx_n - x_n\|\\ &= \|Tx_n - x_n\|\end{aligned}$$

and thus, $\|Tx_n - x_n\| \to 0$. The convergence of x_n to a fixed point now follows from Theorem 25.4. □

Definition 25.6 (Averaged operators)**.** Let $D \subset X$ and $\alpha \in \,]0,1[$. A non-expansive map $T : D \to X$ is called *α-averaged* if there exists another non-expansive operator $R : D \to X$ such that$T = (1-\alpha)I + \alpha R$. If T is α-averaged for some $\alpha \in \,]0,1[$ we call T *averaged.*

Note that every averaged operatoris non-expansive but not every non-expansive operator is averaged (consider $T = -I$).

Proposition 25.7. *A firmly non-expansive operator is 1/2-averaged.*

Proof. If T is firmly non-expansive, then, by Proposition 24.9, $R := 2T - I$ is non-expansive and hence $T = \frac{1}{2}I + \frac{1}{2}R$. □

For averaged operators, we get a slightly stronger convergence result for relaxed Mann iterations:

Proposition 25.8. *Let $\alpha \in \,]0,1[$ and $T : X \to X$ be α-averaged with* Fix $T \neq \emptyset$. *Furthermore, let $\lambda_n \in [0, 1/\alpha]$ such that $\sum_{n=0}^{\infty} \lambda_n(1-\alpha\lambda_n)$ diverges. For some $x_0 \in X$ define the iteration*

$$x_{n+1} = x_n + \lambda_n(Tx_n - x_n) = \lambda_n Tx_n + (1-\lambda_n)x_n.$$

Then it holds that

i) (x_n) *is Fejér monotone with respect to* Fix T*,*

ii) $Tx_n - x_n \xrightarrow{n\to\infty} 0$*,*

iii) (x_n) *converges to some fixed point of* T*.*

Proof. As T is α-averaged, we have $T = (1-\alpha)I + \alpha R$ with some non-expansive R, i.e. $R = (1 - \frac{1}{\alpha})I + \frac{1}{\alpha}T$. Moreover, $Tx = x$ holds if and only if $Rx = x$, i.e. R has the same fixed points as T. With $\mu_n = \alpha\lambda_n$ we can rewrite the iteration as

$$\begin{aligned} x_{n+1} &= x_n + \lambda_n((1-\alpha)I + \alpha R)x_n - x_n) \\ &= x_n + \alpha\lambda_n(Rx_n - x_n) = x_n - \mu_n(Rx_n - x_n). \end{aligned}$$

Now $\mu_n \in [0,1]$ and $\sum_{n=0}^{\infty} \mu_n(1-\mu_n) = \infty$ by assumption and the result follows from Theorem 25.5. □

Hence, we get the following convergence result for the convergence of relaxed iterations for firmly non-expansive maps:

Corollary 25.9. *Let $T : X \to X$ be firmly non-expansive with* $\operatorname{Fix} T \neq \emptyset$. *Then it holds that the iterates $x_{n+1} = x_n + \lambda_n(Tx_n - x_n)$ converge to a fixed point of T if $\lambda_n \in [0,2]$ such that $\sum_{n=0}^{\infty} \lambda_n(1-\lambda_n)$ diverges.*

In particular, the iterates $x_{n+1} = Tx_n$ converge to a fixed point of T for a firmly non-expansive mapping T.

Follows from the previous proposition and the result that firmly non-expansive operators and $\frac{1}{2}$ averaged.

Here are a few more results about averaged operators:

Proposition 25.10. *Let $D \subset X$ be nonempty, $T : D \to X$ be non-expansive and $\alpha \in]0,1[$. Then the following are equivalent:*

i) T is α-averaged.

ii) $(1-\frac{1}{\alpha})I + \frac{1}{\alpha}T$ is non-expansive.

iii) For all $x, y \in D$ is holds that $\|Tx - Ty\|^2 \leq \|x-y\|^2 + \frac{1-\alpha}{\alpha}\|(I-T)x - (I-T)y\|^2$.

Proof. **i) $\iff$ ii):** T is α-averaged if and only if $T = (1-\alpha)I + \alpha R$ for some non-expansive R and this is exactly if $R = \frac{1}{\alpha}T + (1-\frac{1}{\alpha})I$.

ii) $\iff$ iii): We use the elementary equality $\|(1-\frac{1}{\alpha})u + \frac{1}{\alpha}v\|^2 = (1-\frac{1}{\alpha})\|u\|^2 + \frac{1}{\alpha}\|v\|^2 + \frac{1-\alpha}{\alpha^2}\|u-v\|^2$ to write

$$\begin{aligned} \|Rx - Ry\|^2 &= \|(1-\tfrac{1}{\alpha})(x-y) + \tfrac{1}{\alpha}(Tx-Ty)\|^2 \\ &= (1-\tfrac{1}{\alpha})\|x-y\|^2 + \tfrac{1}{\alpha}\|Tx-Ty\|^2 \\ &\quad + \tfrac{1-\alpha}{\alpha^2}\|(I-T)x - (I-Tx)y\|^2 \end{aligned}$$

from which the equivalence follows.

□

Proposition 25.11. *Let $D \subset X$ be nonempty, $T : D \to X$ be L-Lipschitz continuous with $0 < L < 1$. Then T is $\frac{L+1}{2}$-averaged.*

Proof. Set $\alpha = \frac{L+1}{2}$ and note

$$R = (1-\tfrac{1}{\alpha})I + \tfrac{1}{\alpha}T = \tfrac{L-1}{L+1}I + \tfrac{2}{L+1}T$$

is Lipschitz continuous with constant $\frac{L-1}{L+1} + \frac{2}{L+1} = 1$. Hence, R is non-expansive and by Proposition 25.10 we see that T is α-averaged. □