SCIENCES SORBONNE UNIVERSITÉ

Optimization tools

Claire Boyer and Maxime Sangnier

Summary

Definition (Subdifferential)

Let $f : \mathbb{R}^d \to (-\infty, \infty]$ be a convex function. The subdifferential of f at $x \in \mathbb{R}^d$ is defined by

$$\partial f(x) = \{v \in \mathbb{R}^d : \forall y \in \mathbb{R}^d, f(y) \geq f(x) + v^\top (y - x)\}.$$

The elements of $\partial f(x)$ are called the subgradients of f at x.

Example

$f : x \in \mathbb{R} \mapsto |x|$ has a subdifferential for all x and

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

Proposition (Calculus of subgradients)

Let $f : \mathbb{R}^d \to (-\infty, \infty]$ be a convex function and $x \in \mathbb{R}^d$.

a) $\forall \alpha \geq 0, \partial(\alpha f)(x) = \alpha \partial f(x)$.

b) *If $f = \sum_{i=1}^{p} f_i$, with f_i convex, $\mathrm{dom}(f_i) = \mathbb{R}^d$, then $\partial f(x) = \sum_{i=1}^{p} \partial f_i(x)$ (Minkowski sum).*

c) *If $f : y \mapsto \max_{1 \leq i \leq p} f_i(y)$, with f_i convex, then*

$$\partial f(x) = \mathrm{conv}\left(\cup_{\substack{1 \leq i \leq p \\ f_i(x) = f(x)}} \partial f_i(x) \right).$$

Example

$f : x \in \mathbb{R}^d \mapsto \|x\|_1 = \sum_{i=1}^d \text{sign}(x_i)x_i = \max\{s^\top x : s \in \{\pm 1\}^d\}$.
The max is achieved for $s \in \{\pm 1\}^d$ such that $s_i = \text{sign}(x_i)$ if $x_i \neq 0$ and $s_i = \pm 1$ for $x_i = 0$, with $s^\top x = \|x\|_1$. As a consequence:

$$\begin{aligned}\partial f(x) &= \text{conv}\left(\left\{s \in \{\pm 1\}^d : s^\top x = \|x\|_1\right\}\right) \\ &= \left\{v \in \mathbb{R}^d : \|v\|_\infty \leq 1, v^\top x = \|x\|_1\right\}.\end{aligned}$$

Proposition (Subgradient of differentiable functions)

Let $f : \mathbb{R}^d \to \mathbb{R}$ be a convex and differentiable function at $x \in \mathbb{R}^d$. Then $\partial f(x) = \{\nabla f(x)\}$.

Proposition (Subgradient of differentiable functions)

Let $f : \mathbb{R}^d \to \mathbb{R}$ be a convex and differentiable function at $x \in \mathbb{R}^d$ and $g : \mathbb{R}^d \to \mathbb{R}$ be a convex function. Then

$$\partial(f + g)(x) = \{\nabla f(x)\} + \partial g(x).$$

Existence of subgradients

Definition (Relative interior)

Let $\mathcal{C} \subset \mathbb{R}^d$ be a convex set. The relative interior of $\mathcal{C}$ is

$$\text{relint}(\mathcal{C}) = \{x \in \mathcal{C} : \forall y \in \mathcal{C}, \exists \lambda > 1 : y + \lambda(x - y) \in \mathcal{C}\}.$$

In other words, in any direction from $x \in \text{relint}(\mathcal{C})$, there is always a point ahead of x which lies in $\text{relint}(\mathcal{C})$.

Remark

The relative interior of a convex set $\mathcal{C}$ is never empty. If $\mathcal{C}$ is a singleton, then $\text{relint}(\mathcal{C}) = \mathcal{C}$.

Proposition

Let $f : \mathbb{R}^d \to [-\infty, \infty]$ be a convex function and $x \in \mathrm{relint}(\mathrm{dom}(f))$ (which is well defined since $\mathrm{dom}(f)$ is convex). Then $\partial f(x)$ is non-empty.

Theorem (Fermat's rule)

Let $f : \mathbb{R}^d \to (-\infty, \infty]$ be a convex function. $x^\star \in \mathbb{R}^d$ is a global minimizer of f if and only if

$$0 \in \partial f(x^\star).$$

Exercise of last week

In the setting of compressed sensing, we observe m linear projections of an object $x \in \mathbb{R}^n$ where $m \ll n$. The data is contained in vector $y = Ax$ where A is the sensing matrix.

We want to solve an under-determined linear system. To do so, one can choose to pick the solution with the minimimal energy in ℓ^2-norm.

It can read as follows

$$\min_{z \in \mathbb{R}^n} \|z\|_2 \quad \text{s.t.} \quad y = Az.$$

Exercice 6

1. Show that solving the above problem is equivalent to solving the following problem
$$\min_{z \in \mathbb{R}^n} \|z\|_2^2 \quad \text{s.t.} \quad y = Az.$$
2. Is there a solution to the above problem ? Is it unique ?
3. Using the Lagrangian, write optimality conditions. Deduce an optimum.
4. Now, we want to find the same optimum but using only Fermat's rule.
   A. Show that the problem can be rewritten as follows
   $$\min_{z \in \mathbb{R}^n} \|z\|_2^2 + \chi_C(z)$$
   where χ_C is the characteristic function of the (convex) set $C := \{x : y = Ax\}$.
   B. Show that the subdifferential of a characteristic function of a convex set corresponds to the normal cone
   $$\mathcal{N}_C(x) = \{d, \langle d, y - x \rangle \leq 0, \quad \forall y \in C\}$$
   C. Show that the normal cone to the set $C := \{x : y = Ax\}$ is $\ker(A)^\perp = \text{ran}(A^*)$.
   D. Write the Fermat's rule for the last optimization problem and deduce an optimum.

A canonical convex optimization problem has the form:

$$\begin{aligned} \underset{x \in \mathcal{X}}{\text{minimize}} \; & f(x) \\ \text{s.t.} \quad & \begin{cases} \forall j \in [p] \colon g_j(x) \leq 0 \\ Ax = b, \end{cases} \end{aligned} \tag{P1}$$

where f and g_j are convex, $A \in \mathbb{R}^{m \times d}$ and $b \in \mathbb{R}^m$.

Proposition (Optimality criterion)

Assume that f is differentiable. Then $x^\star \in \mathcal{C}$ is a global minimizer of (P1) if and only if

$$\forall x \in \mathcal{C} \colon \nabla f(x^\star)^\top (x - x^\star) \geq 0.$$

Linear programs

A linear program (LP) is of the form:

$$\begin{aligned} \underset{x \in \mathbb{R}^d}{\text{minimize}} \; & c^\top x + d \\ \text{s.t.} \quad & \begin{cases} Gx \preccurlyeq h \\ Ax = b. \end{cases} \end{aligned}$$

Here, the feasible set is a polyhedron.

Quadratic programs

A quadratic program (QP) is

$$\begin{aligned} \underset{x\in\mathbb{R}^d}{\text{minimize}} \ & \frac{1}{2}x^\top Px + q^\top x + r \\ \text{s.t.} \ & \begin{cases} Gx \preccurlyeq h \\ Ax = b, \end{cases} \end{aligned}$$

where P is a positive semi-definite matrix.
Quadratic constraints $\implies$ quadratically constrained quadratic program (QCQP):

$$\begin{aligned} \underset{x\in\mathbb{R}^d}{\text{minimize}} \ & \frac{1}{2}x^\top Px + q^\top x + r \\ \text{s.t.} \ & \begin{cases} \forall j \in [p], \frac{1}{2}x^\top P_j x + q_j^\top x + r_j \leq 0 \\ Ax = b, \end{cases} \end{aligned}$$

where P_j are positive semi-definite.

Summary

Definition (Legendre-Fenchel transformation)

Let $f : \mathbb{R}^d \to [-\infty, \infty]$. The Legendre-Fenchel transformation (or convex conjugate) of f is:

$$f^* : y \in \mathbb{R}^d \mapsto \sup_{x \in \mathbb{R}^d} \left\{ y^\top x - f(x) \right\}.$$

Remark

If $f : \mathbb{R}^d \to \mathbb{R}$, f is differentiable and the supremum is attained in x^*, then x^* is such that

- $y = \nabla f(x^*)$;
- $f(x^*) = y^\top x^* - f^*(y)$.

Remark

Thus, $-f^*(y)$ is the intercept corresponding to the tightest affine minorant of f with "slope" y:

$$\forall x \in \mathbb{R}^d, \quad y^\top x^* - f(x^*) \geq y^\top x - f(x) \quad \text{so}$$
$$f(x) \geq f(x^*) + y^\top (x - x^*) = y^\top x - f^*(y).$$

Proposition (Some properties of the convex conjugate)

Let $f : \mathbb{R}^d \to [-\infty, \infty]$.

1. $f^*(0) = -\inf_{x \in \mathbb{R}^d} f(x)$.
2. *f^* is convex and lower semi-continuous.*
3. *If $\operatorname{dom}(f) \neq \emptyset$, then $\forall y \in \mathbb{R}^d : f^*(y) > -\infty$.*
4. *If f is convex and proper, then f^* is proper.*

Legendre-Fenchel transformation and duality

Proposition (Fenchel-Young inequality)

Let $f : \mathbb{R}^d \to [-\infty, \infty]$. *Then*

$$\forall (x, y) \in \mathbb{R}^d \times \mathbb{R}^d : \quad f(x) + f^*(y) \geq x^\top y.$$

Moreover, if f *is convex,* $\forall (x, y) \in \mathbb{R}^d \times \mathbb{R}^d$*:*

$$f(x) + f^*(y) = x^\top y \iff y \in \partial f(x).$$

Remark

This is an extension of the following inequality to non-quadratic functions f*:*

$$\forall x, y \in \mathbb{R}^d : \quad \frac{1}{2}\|x\|_2^2 + \frac{1}{2}\|y\|_2^2 \geq x^\top y.$$

Legendre-Fenchel transformation and duality

Example (Remarkable convex conjugates)

1. If $f = \frac{1}{2}\|\cdot\|_2^2$, then $f^* = \frac{1}{2}\|\cdot\|_2^2$.
2. If $f = \exp$, then $f^*(y) = y(\log(y) - 1)$ if $y > 0$, $f(y) = \infty$ if $y < 0$ and $f(0) = 0$.
3. Let $\mathcal{K} \subset \mathbb{R}^d$. If $f = \chi_{\mathcal{K}}$, then $f^* : y \in \mathbb{R}^d \mapsto \sup_{x \in \mathcal{K}} y^\top x$.

Definition (Dual norm)

Let $\|\cdot\|$ be a norm on $\mathbb{R}^d$. Its dual norm $\|\cdot\|_*$ is defined by:

$$\forall y \in \mathbb{R}^d: \quad \|y\|_* = \sup_{\|x\| \leq 1} y^\top x.$$

Proposition

Let $\|\cdot\|$ be a norm on $\mathbb{R}^d$.

1. $\|\cdot\|_*$ *is a norm.*
2. $\forall (x, y) \in \mathbb{R}^d \times \mathbb{R}^d: \quad y^\top x \leq \|x\| \|y\|_*.$
3. *The dual norm of a dual norm is the primal norm:* $(\|\cdot\|_*)_* = \|\cdot\|.$

Example (Dual norms)

1. Let $p > 1$ and $q > 1$ such that $\frac{1}{p} + \frac{1}{q} = 1$. Then $\|\cdot\|_p$ is dual to $\|\cdot\|_q$.
   We deduce Hölder's inequality:
   $\forall (x, y) \in \mathbb{R}^d \times \mathbb{R}^d : \quad y^\top x \leq \|x\|_p \|y\|_q$.
2. Particular cases are: ℓ_2 is self-dual, ℓ_1 and ℓ_∞ are dual.
3. For matrices, the Frobenius norm is self-dual, the spectral and the trace norms are dual.

Proposition (Convex conjugate of a norm)

Let $\|\cdot\|$ be a norm on $\mathbb{R}^d$. Then, $\|\cdot\|^* = \chi_{\{x \in \mathbb{R}^d : \|x\|_* \leq 1\}}$.

Proposition (Biconjugate, involution)

Let $f : \mathbb{R}^d \to [-\infty, \infty]$ *and* $f^{**} = (f^*)^*$ *the biconjugate of* f*. Then*

$$\forall x \in \mathbb{R}^d : f(x) \geq f^{**}(x).$$

In addition, if f *is convex, proper and lower semi-continuous, then*

$$f = f^{**}$$

and

$$y \in \partial f(x) \iff x \in \partial f^*(y).$$

Remark

The biconjugate f^{**} *is sometimes called the convex relaxation of* f*.*

Example (Link with convex conjugates)

Assume that we are interested in an optimization problem with linear constraints:

$$\begin{aligned} &\underset{x\in\mathbb{R}^d}{\text{minimize}}\ f(x) \\ &\text{s.t.} \quad \begin{cases} Ax \preccurlyeq b \\ Cx = d, \end{cases} \end{aligned}$$

where A, C, b and d are any matrices and vectors. Then the dual function is:

$$\forall(\lambda,\nu) \in \mathbb{R}^p \times \mathbb{R}^m:$$
$$G(\lambda,\nu) = -(\lambda^\top b + \nu^\top d) - f^*(-A^\top\lambda - C^\top\nu) - \chi_{\mathbb{R}^p_+}(\lambda).$$

Let us consider the optimization problem:

$$\underset{x\in\mathbb{R}^d}{\text{minimize}}\ f(Ax)+g(x), \tag{P2}$$

where $f\colon \mathbb{R}^p \to (-\infty,\infty]$, $g\colon \mathbb{R}^d \to (-\infty,\infty]$ are proper convex and $A\in\mathbb{R}^{p\times d}$.
This problem is equivalent to

$$\begin{aligned}\underset{x\in\mathbb{R}^d,y\in\mathbb{R}^p}{\text{minimize}}\ & f(y)+g(x)\\ \text{s.t.}\quad & Ax-y=0.\end{aligned}$$

The dual function to this last problem is:

$$\begin{aligned}\forall \nu \in \mathbb{R}^p \colon G(\nu) &= \inf_{x\in\mathbb{R}^d, y\in\mathbb{R}^p} \left\{ f(y) + g(x) + \nu^\top A x - \nu^\top y \right\} \\ &= -\sup_{y\in\mathbb{R}^p} \left\{ \nu^\top y - f(y) \right\} - \sup_{x\in\mathbb{R}^d} \left\{ -\nu^\top A x - g(x) \right\} \\ &= -f^*(\nu) - g^*(-A^\top \nu).\end{aligned}$$

Therefore, the dual problem of interest is:

$$\underset{\nu\in\mathbb{R}^p}{\text{maximize}} \; -f^*(\nu) - g^*(-A^\top \nu). \tag{P3}$$

Theorem

Let $f : \mathbb{R}^p \to (-\infty, \infty]$, $g : \mathbb{R}^d \to (-\infty, \infty]$ *be proper convex functions and* $A \in \mathbb{R}^{p \times d}$. *Assume that either* $\operatorname{dom}(f) = \mathbb{R}^p$ *or* $\operatorname{dom}(g) = \mathbb{R}^d$ *and that* $\exists x \in \mathbb{R}^d : Ax \in \operatorname{dom}(f)$. *If the optima are attained in (P2) and (P3), then strong duality holds:*

$$\min_{x \in \mathbb{R}^d} f(Ax) + g(x) = \max_{\nu \in \mathbb{R}^p} -f^*(\nu) - g^*(-A^\top \nu).$$

Moreover, a primal-dual optimum is a solution to the saddle-point problem:

$$\operatorname*{minimize}_{x \in \mathbb{R}^d} \operatorname*{maximize}_{\nu \in \mathbb{R}^p} g(x) + \nu^\top Ax - f^*(\nu).$$

Set constraint $f = \chi_{\mathcal{C}}$, where $\mathcal{C} \subset \mathbb{R}^d$ is a convex set.

$$\underset{x \in \mathbb{R}^d}{\text{minimize}}\ g(x) + \chi_{\mathcal{C}}(Ax - b),$$

whose dual is

$$\underset{\nu \in \mathbb{R}^p}{\text{maximize}}\ -b^\top \nu - \chi_{\mathcal{C}}^*(\nu) - g^*(-A^\top \nu).$$

	Constraint	Set $\mathcal{C}$	Conjugate
Equality	$Ax = b$	$\{0\}$	0
Ball	$\|Ax - b\| \leq 1$	unit $\|\cdot\|$-ball	$\|\cdot\|_*$
Conic inequality	$Ax \preccurlyeq_{\mathcal{K}} b$	$-\mathcal{K}$	$\chi_{\mathcal{K}^*}$

Table: Examples of constraints.

Norm regularization $f(y) = \|y - b\|$.

$$\underset{x \in \mathbb{R}^d}{\text{minimize}}\; g(x) + \|Ax - b\|.$$

Since, $f^*(\nu) = b^\top \nu + \chi_{\mathcal{B}}(\nu)$, where $\mathcal{B} = \{y \in \mathbb{R}^p : \|y\|_* \leq 1\}$, the dual reads:

$$\begin{aligned} \underset{\nu \in \mathbb{R}^p}{\text{maximize}}\; & -b^\top \nu - f^*(-A^\top \nu) \\ \text{s.t.} \quad & \|\nu\|_* \leq 1. \end{aligned}$$

Summary

Compressed sensing: finding $x \in \mathbb{R}^d$ such that

$$Ax = y,$$

where $A \in \mathbb{R}^{p \times d}$ is a sensing matrix and $y \in \mathbb{R}^p$ the vector of measurements.

Special features:

1. $p \ll d$, so the problem of finding x such that $y = Ax$ is under-determined;
2. the signal to recover is s-sparse ($s \in \mathbb{N}^*$).

Compressed sensing (optimization):

$$\begin{aligned} \underset{x \in \mathbb{R}^d}{\text{minimize}} \ & \|x\|_0 \\ \text{s.t.} \quad & Ax = y. \end{aligned}$$

A roughly equivalent formulation:

$$\begin{aligned} \underset{x \in \mathbb{R}^d}{\text{minimize}} \; & f(x) = \|Ax - y\|_2 \\ \text{s.t.} \quad & \|x\|_0 \leq s, \end{aligned}$$

where $s \in \mathbb{N}$ is a prescribed sparsity level.
If the signal to recover $x^\star \in \mathbb{R}^d$ is s-sparse, then it is a solution the previous optimization problem and $f(x^\star) = 0$.

Initial point $x_0 = 0 \in \mathbb{R}^d$ and $S_0 = \text{supp}(x_0) = \emptyset$ its support.

Algorithm Orthogonal matching pursuit (OMP)

$$
\begin{aligned}
(j_k, \alpha_k) &\in \underset{j \in [d], \alpha \in \mathbb{R}}{\arg\min} \|A(x_k + \alpha e_j) - y\|_2 \\
S_{k+1} &= S_k \cup \{j_k\} \\
x_{k+1} &\in \underset{x \in \mathbb{R}^d : \text{supp}(x) \subset S_{k+1}}{\arg\min} \|Ax - y\|_2,
\end{aligned}
$$

where e_j is the j^{th} canonical basis vector of $\mathbb{R}^d$.

- When the columns of A norm to 1, Step 1 boils down to finding $j \in [d]$, that maximizes $|(A_j)^\top (Ax_k - y)|$ (atom that is the most correlated to the residue $Ax_k - y$).
- Step 2 potentially increments the sparsity of the current iteration x_{k+1}: $\|x_{k+1}\|_0 \leq k + 1$, at each iteration k.
- Step 3 is an orthogonal projection, hence the name *orthogonal* matching pursuit.

Remark

Under some conditions, the orthogonal matching pursuit can recover any s-sparse signal $x^\star$ with at most s iterations. However, the weakness of orthogonal matching pursuit is that, once an incorrect index j has been selected, it remains in the support of the proposed solution. In this case, s iterations are not enough to recover an s-sparse signal.

Let $L_s \colon \mathbb{R}^d \to [d]$ be such that $L_s(x)$ is the index set of s largest absolute entries of x.
Let $H_s \colon \mathbb{R}^d \to \mathbb{R}^d$ be the hard-thresholding operator of order s: H_s is such that $H_s(x)$ has support $L_s(x)$ and equals x on its support (the other entries are 0).
Initial point $x_0 = 0 \in \mathbb{R}^d$.

Algorithm Compressive sampling matching pursuit (CoSaMP)

$$
\begin{aligned}
S_{k+1} &= \operatorname{supp}(x_k) \cup L_{2s}(A^\top (A x_k - y)) \\
u_{k+1} &\in \operatorname*{arg\,min}_{u \in \mathbb{R}^d : \operatorname{supp}(u) \subset S_{k+1}} \|Au - y\|_2 \\
x_{k+1} &= H_s(u_{k+1}).
\end{aligned}
$$

Remark

Orthogonal and compressive sampling matching pursuits require to estimate the sparsity of the signal $x^\star$ to recover. This is not an easy task.

Summary

Basis pursuit

Compressed sensing:

$$\begin{aligned} \underset{x \in \mathbb{R}^d}{\text{minimize}} \ & \|x\|_0 \\ \text{s.t.} \quad & Ax = y, \end{aligned}$$

where $A \in \mathbb{R}^{p \times d}$.

Since this problem is non-convex and even NP-hard in general, we would like to convexify it.

$\|\cdot\|_0$ is relatively well approximated by $\|\cdot\|_q^q$ when $q \to 0_+$ (not convex for $0 \leq q < 1$). The smallest value of q for which $\|\cdot\|_q^q$ is convex is $q = 1$.

Convexification:

$$\begin{aligned} \underset{x \in \mathbb{R}^d}{\text{minimize}} \ & \|x\|_1 \\ \text{s.t.} \quad & Ax = y. \end{aligned} \tag{P4}$$

This problem is often referred to as *Basis pursuit*.

Proposition (Sparsity of basis pursuit)

Assume that (P4) has a minimizer $x^\star \in \mathbb{R}^d$. Then $\|x^\star\|_0 \leq p$.

Basis pursuit as a linear program

Difficulty: the objective function of (P4) is not differentiable.

Proposition (Variational ℓ_1-norm)

$$\forall x \in \mathbb{R}^d : \quad \|x\|_1 = \min \left\{ \sum_{i=1}^{d} \xi_i^+ + \xi_i^- \ : x = \xi^+ - \xi^-, \ (\xi^+, \xi^-) \in (\mathbb{R}_+^d)^2 \right\}.$$

Proof.

Show that a minimizer is (ξ^+, ξ^-), with $\xi_i^+ = \max(0, x_i)$ and $\xi_i^- = \max(0, -x_i)$ $(\forall i \in [d])$. Then, remark that $\|x\|_1 = \sum_{i=1}^{d} \xi_i^+ + \xi_i^-$. □

Basis pursuit as a linear program

As a consequence of the previous proposition, (P4) can be reformulated in:

$$\begin{aligned} \underset{x\in\mathbb{R}^d}{\text{minimize}} \quad & \min_{(\xi^+,\xi^-)\in(\mathbb{R}^d)^2} \sum_{i=1}^{d} \xi_i^+ + \xi_i^- \\ & \text{s.t.} \quad \begin{cases} x = \xi^+ - \xi^- \\ \xi^+ \succcurlyeq 0 \\ \xi^- \succcurlyeq 0, \end{cases} \\ \text{s.t.} \quad & Ax = y, \end{aligned}$$

which becomes:

$$\begin{aligned} \underset{(\xi^+,\xi^-)\in(\mathbb{R}^d)}{\text{minimize}} \quad & \sum_{i=1}^{d} \xi_i^+ + \xi_i^- \\ \text{s.t.} \quad & \begin{cases} A(\xi^+ - \xi^-) = y \\ \xi^+ \succcurlyeq 0 \\ \xi^- \succcurlyeq 0. \end{cases} \end{aligned} \tag{P5}$$

We focus on an optimization problem of the form:

$$\begin{aligned} \underset{x \in \mathbb{R}^d}{\text{minimize}} \; & c^\top x \\ \text{s.t.} \quad & \begin{cases} Ax = b \\ x \succcurlyeq 0, \end{cases} \end{aligned} \tag{P6}$$

where $c \in \mathbb{R}^d$, $A \in \mathbb{R}^{p \times d}$ and $b \in \mathbb{R}^p$ are any matrices.

Remark

A linear program can always be written in the form of (P6).

The feasibility set of (P6) reads $\mathcal{C} = \{x \in \mathbb{R}^d : x \succcurlyeq 0, Ax = b\}$.
As a consequence, it is either:

1. empty, so (P6) is not feasible;
2. not compact;
3. or the convex hull of a finite number of points.

In Situation 3, $\mathcal{C}$ is called a polytope or a simplex.

Proposition (Solution of a linear program)

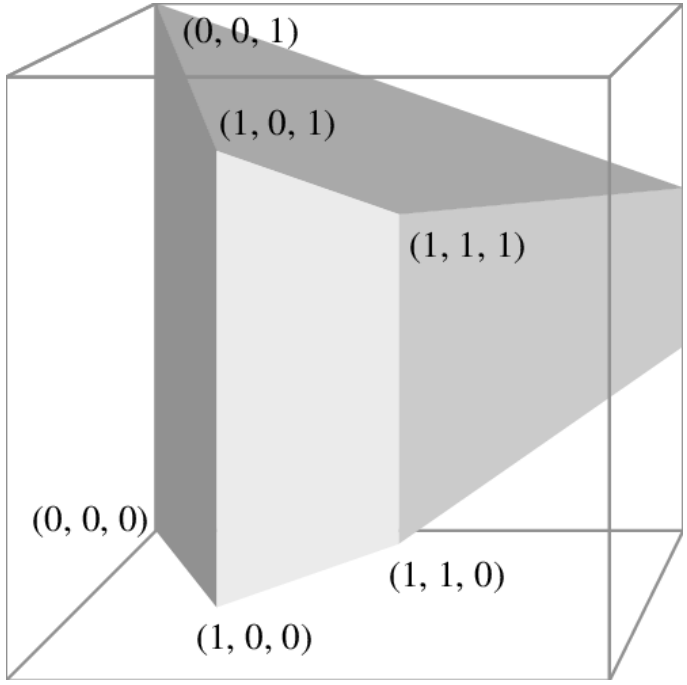
Let us assume that $\mathcal{C}$ is non-empty and compact. Then, (P6) has a solution, which is an extreme point of $\mathcal{C}$.

Proof.

A compact convex set is the convex hull of its extreme points. □

Linear programming

The simplex algorithm finds a path in the set of extreme points of $\mathcal{C}$ such that the objective function does not increase at each iteration. Generally, the simplex algorithms converges linearly in the number of constraints. However, the worst-case complexity is very bad. On the so-called Klee-Minty cube, the simplex algorithm exhibits poor performance (it visits all 2^p corners of the cube, where p is the number of constraints).



Linear programming

Here, we focus on the optimization problem:

$$\begin{aligned} \underset{x\in\mathbb{R}^d}{\text{minimize}}\ & f(x) \\ \text{s.t.}\quad & \begin{cases} \forall j \in [p]\colon g_j(x) \le 0 \\ Ax = b, \end{cases} \end{aligned}$$

where $A \in \mathbb{R}^{m\times d}$ is a rank m matrix, f and g_j are twice differentiable.
Barrier methods are also called interior point methods.
They are particularly useful when f and g_j are linear functions.

Reformulation:

$$\begin{aligned} \underset{x\in\mathbb{R}^d}{\text{minimize}}\ & f(x) + \sum_{j=1}^{m} \chi_{\mathbb{R}_-}(g_j(x)) \\ \text{s.t.}\quad & Ax = b, \end{aligned}$$

$\chi_{\mathbb{R}_-}$ can be approximated by a smooth barrier function.
For instance, the logarithmic barrier:

$$\phi\colon x \in \mathbb{R}^d \mapsto \begin{cases} -\sum_{j=1}^{m} \log(-g_j(x)) & \text{if } g_j(x) < 0, \forall j \in [m] \\ \infty & \text{otherwise.} \end{cases}$$

For $t > 0$, the function $x \in \mathbb{R}^d \mapsto \frac{1}{t}\phi(x)$ approximate $\chi_{\mathbb{R}^d_-}$ and the approximation improves as $t \to \infty$.

Proposition

The barrier ϕ is convex and twice differentiable.

For $t > 0$, the problem of interest becomes:

$$\begin{aligned} \underset{x \in \mathbb{R}^d}{\text{minimize}} \; & tf(x) + \phi(x) \\ \text{s.t.} \quad & Ax = b, \end{aligned} \tag{P7}$$

Proposition

Assume that strong duality holds for (P7) and let $x^\star(t)$ be a minimizer (P7) for $t > 0$. Then

$$0 \leq f(x^\star(t)) - p^* \leq \frac{m}{t},$$

where $p^* = \inf_{x \in \mathbb{R}^d} f(x) + \sum_{j=1}^m \chi_{\mathbb{R}_-}(g_j(x)) + \chi_b(Ax)$ is the infimum of the original problem.

Proof.

This comes from KKT conditions. □

Strictly feasible initial point $x_0 \in \mathbb{R}^d$, $t_0 > 0$ and $\mu > 1$.

Algorithm Barrier (or interior point) method.

$$x_{k+1} \in \underset{x \in \mathbb{R}^d : Ax = b}{\arg\min} \; t_k f(x) + \phi(x),$$

$$t_{k+1} = \mu t_k.$$

As a stopping criterion, one can use $\frac{m}{k} \leq \epsilon$ since this ratio bounds the difference $f(x_{k+1}) - p^*$.

Remark

1. *The first step of a barrier algorithm is generally performed thanks to Newton method.*
2. *x_k is used to initialize the algorithm for solving Step 1 (warm start). This makes the all story faster and explains why only several iterations are needed in barrier methods.*

Interior point methods are very reliable on small scale problems but are not workable for very large problems. First order methods seem to be the only option.

Summary

Primal methods

In this section, we consider a function $f : \mathbb{R}^d \to \mathbb{R}$, that is differentiable and convex, and we tackle the problem:

$$\underset{x \in \mathbb{R}^d}{\text{minimize}}\ f(x).$$

Initial point $x_0 \in \mathbb{R}^d$.

Algorithm Gradient descent.

$$x_{k+1} = x_k - \gamma_k \nabla f(x_k),$$

where $\gamma_k > 0$ is a step size to be tuned.

Minimizing a local quadratic approximation of f:

$$\forall x \in \mathbb{R}^d : \quad f(x) \approx f(x_k) + \nabla f(x_k)^\top (x - x_k) + \frac{1}{2\gamma_k} \|x - x_k\|_2^2,$$

where γ_k is unknown a priori.

Advantages and drawbacks

Advantages of gradient descent are:

1. every iteration is inexpensive;
2. it does not require second order information (Hessian of f).

However, gradient descent

1. is often slow (oscillation);
2. does not handle nondifferentiable functions.

Other first-order methods address one or both disadvantages.
Methods with improved convergence:

- quasi-Newton methods;
- accelerated gradient method.

Methods for nondifferentiable or constrained problems:

- subgradient method;
- proximal gradient method.

Newton method

This section deals with including second order information in gradient descent. We assume that f is twice differentiable.

Algorithm Newton method.

$$x_{k+1} = x_k - (\nabla^2 f(x_k))^{-1} \nabla f(x_k).$$

Minimizing a second-order approximation of f around x_k:

$$f(y) \approx f(x_k) + \nabla f(x_k)^\top (y - x_k) + \frac{1}{2}(y - x_k)^\top \nabla^2 f(x_k)(y - x_k).$$

Drawbacks: Fast convergence but:

- Convergence not always guaranteed.
- Expensive for large scale applications.

The Hessian can be approximated by a metric $H \in \mathbb{R}^{d \times d}$, that is symmetric positive definite.

Algorithm Quasi-Newton method.

Given an initial point $x_0 \in \mathbb{R}^d$ and an initial metric $H_0 \in \mathbb{R}^{d \times d}$, that is symmetric positive definite, iterate

$$x_{k+1} = x_k - \gamma_k H_k^{-1} \nabla f(x_k),$$
$$\text{set } H_{k+1} \text{ based on } H_k,$$

where $\gamma_k > 0$ are step sizes, which can be chosen by line search.

Newton method Setting $H_k = \nabla^2 f(x_k)$ makes the last algorithm boiling down to a Newton method with adaptive step size.
Broyden-Fletcher-Goldfarb-Shanno (BFGS) Setting $\Delta_x = x_{k+1} - x_k$ and $\Delta_y = \nabla f(x_{k+1}) - \nabla f(x_k)$, the BFGS update rule is:

$$H_{k+1} = H_k + \frac{1}{\Delta_x^\top \Delta_y} \Delta_y \Delta_y^\top - \frac{1}{\Delta_x^\top H_k \Delta_x} H_k \Delta_x \Delta_x^\top H_k.$$

Let us remark can the inverse can be computed efficiently:

$$H_{k+1}^{-1} = \left(I_d - \frac{1}{\Delta_x^\top \Delta_y} \Delta_x \Delta_y^\top \right) H_k^{-1} \left(I_d - \frac{1}{\Delta_x^\top \Delta_y} \Delta_y^\top \Delta_x \right) + \frac{1}{\Delta_x^\top \Delta_y} \Delta_x^\top \Delta_x,$$

where I_d is the identity matrix of size $d \times d$.
BFGS method converges for strongly convex functions (in that case $\Delta_x^\top \Delta_y > 0$).

Square root BFGS Same as previously but with $H_k = L_k L_k^\top$ (Cholesky decomposition). The updates rule is:

$$L_{k+1} = L_k \left(I_d + \frac{1}{\tilde{\Delta}_x^\top \tilde{\Delta}_x} (\alpha \tilde{\Delta}_y - \tilde{\Delta}_x) \tilde{\Delta}_x^\top \right),$$

where $\tilde{\Delta}_x = L_k \Delta_x$, $\tilde{\Delta}_y = L_k^{-1} \Delta_y$ and $\alpha = \frac{\tilde{\Delta}_x^\top \tilde{\Delta}_x}{\Delta_x^\top \Delta_y}$.

Limited-memory BFGS (L-BFGS) Leveraging the recursive formula of H_k^{-1}, we can compute a direction of descent $H_k^{-1} \nabla f(x_k)$ with only recursive updates of vectors.

L-BFGS goes beyond this remark by truncating the recursion to the last m (often $m \approx 30$) iterations.

This requires nevertheless to store the m last values of Δ_x and Δ_y.

We no longer require f to be differentiable (but f is still convex). Subgradient method is similar to gradient descent but replacing gradients by subgradients.

Algorithm Subradient descent.

$$x_{k+1} = x_k - \gamma_k v_k,$$

where $v_k \in \partial f(x_k)$ and $\gamma_k > 0$ is a step size.

Remark

Contrarily to a negative gradient $-\nabla f(x_k)$, a negative subgradient $-v$ ($v \in \partial f(x_k)$) is not a direction of descent in general. This means that the subgradient method is not a descent method.

Primal methods

Akin to gradient descent, several step size rules coexist:

- fixed step: γ_k is constant;
- fixed length: $\gamma_k \|v_k\|_2 = \|x_k - x_{k-1}\|_2$;
- diminishing step: $\gamma_k \to 0$, with $\sum_{k=1}^{\infty} \gamma_k = \infty$.

For fixed step sizes and fixed length, the subgradient method does not converge.
However, two cases are of interest: diminishing step sizes and fixed length for a given number of steps.

Diminishing step

Theorem (Convergence of subgradient method)

Assume that f has a minimizer $x^\star \in \mathbb{R}^d$ and that f is Lipschitz continuous with Lipschitz constant $L > 0$. For a diminishing step sizes $\gamma_k \to 0$, with $\sum_{k=1}^{\infty} \gamma_k = \infty$:

$$\min_{0 \le \ell \le k} f(x_\ell) - f(x^\star) \le \frac{\|x_0 - x^\star\|_2^2 + L^2 \sum_{\ell=1}^{k} \gamma_\ell^2}{2 \sum_{\ell=1}^{k} \gamma_\ell}.$$

Since $\frac{\sum_{\ell=1}^{k} \gamma_\ell^2}{2 \sum_{\ell=1}^{k} \gamma_\ell} \to 0$, $\min_{0 \le \ell \le k} f(x_\ell)$ converges to $f(x^\star)$.

Fixed length for a given number of iterations

Theorem (Convergence of subgradient method)

Assume that f has a minimizer $x^\star \in \mathbb{R}^d$ and that f is Lipschitz continuous with Lipschitz constant $L > 0$. Let $x_0 \in \mathbb{R}^d$ be an initial point close to a minimizer: $\|x_0 - x^\star\|_2 \leq R$, for $R > 0$. For a fixed step length: $\gamma_k \|v_{k-1}\|_2 = \frac{R}{\sqrt{k}}$:

$$\min_{0 \leq \ell \leq k} f(x_\ell) - f(x^\star) \leq \frac{LR}{\sqrt{k}}.$$

In addition, any other step length increases the bound.

Remark

The convergence rate of subgradient descent is optimal (we can construct an optimization problem for which convergence is in $O(1/\sqrt{k})$.

To sum up, subgradient descent:

1. handles nondifferentiable convex problems;
2. is an algorithm as simple as gradient descent;
3. has slow convergence;
4. does not provide easy stopping criterion.

Primal methods

We have seen at the beginning of this class that optimization problems in machine learning are often of the form:

$$\underset{x \in \mathbb{R}^d}{\text{minimize}}\ f(x) + g(x), \tag{P8}$$

where $f : \mathbb{R}^d \to \mathbb{R}$ is a differentiable and convex function and $g : \mathbb{R}^d \to (-\infty, \infty]$ is a convex function.
We leverage the special structure of this problem to introduce fast algorithms (compared to subgradient methods) even though the global function $f + g$ is not differentiable.

Definition (Proximal operator)

Let $h : \mathbb{R}^d \to [-\infty, \infty]$ be a proper, lower semi-continuous convex function. The proximal operator of h is defined by:

$$\forall x \in \mathbb{R}^d : \operatorname{prox}_h(x) = \underset{u \in \mathbb{R}^d}{\arg\min}\ h(u) + \frac{1}{2}\|u - x\|_2^2.$$

(By strong convexity prox_g is well defined.)

Example

- For $h = 0$, $\text{prox}_h(x) = x, \forall x \in \mathbb{R}^d$.
- Let $\mathcal{C} \subset \mathbb{R}^d$ be a closed convex set and $h = \chi_{\mathcal{C}}$. Then prox_h is the orthogonal projector on $\mathcal{C}$.
- For $h = \|\cdot\|_1$, prox_h is the *soft-thresholding* operator:

$$\forall x \in \mathbb{R}^d, \forall i \in [d]: \quad \text{prox}_h(x)_i = \begin{cases} x_i - 1 & \text{if } x_i \geq 1 \\ 0 & \text{if } |x_i| \leq 1 \\ x_i + 1 & \text{if } x_i \leq -1. \end{cases}$$

Theorem (Moreau decomposition)

Let $h\colon \mathbb{R}^d \to [-\infty, \infty]$ be a proper, lower semi-continuous convex function. Then

$$\forall x \in \mathbb{R}^d: \quad x = \text{prox}_h(x) + \text{prox}_{h^*}(x).$$

Primal methods

When:

1. $f : \mathbb{R}^d \to \mathbb{R}$ is convex and differentiable;
2. $g : \mathbb{R}^d \to (-\infty, \infty]$ is convex with an *easy-to-compute* proximal operator.

Algorithm Proximal gradient method.

$$x_{k+1} = \text{prox}_{\gamma_k g}\left(x_k - \gamma_k \nabla f(x_k)\right),$$

where $\gamma_k > 0$ is a step size.

Minimizing a local quadratic approximation of f:

$\forall x \in \mathbb{R}^d$:

$$\begin{aligned} f(x) + g(x) &\approx f(x_k) + \nabla f(x_k)^\top (x - x_k) + \frac{1}{2\gamma_k}\|x - x_k\|_2^2 + g(x) \\ &= g(x) + \frac{1}{2\gamma_k}\|x - (x_k - \gamma_k \nabla f(x_k))\|_2^2 - \frac{\gamma_k}{2}\|\nabla f(x_k)\|_2^2, \end{aligned}$$

where γ_k is unknown a priori.

Example (Soft-thresholding)

When $g = \|\cdot\|_1$, we obtain the soft-thresholding method, where we first perform a gradient step $x^+ = x_k - \gamma_k \nabla f(x_k)$, and then a soft-thresholding:

$$\forall i \in [d]: \quad (x_{k+1})_i = \begin{cases} x_i^+ - \gamma_k & \text{if } x_i \geq \gamma_k \\ 0 & \text{if } -\gamma_k \leq x_i \leq \gamma_k \\ x_i^+ + \gamma_k & \text{if } x_i \leq -\gamma_k. \end{cases}$$

Theorem (Convergence of the proximal method)

Consider (P8) with $f : \mathbb{R}^d \to \mathbb{R}$ being differentiable with L-Lipschitz gradient ($L > 0$), and $g : \mathbb{R}^d \to (-\infty, \infty]$ being proper, lower-semicontinuous and convex.
Let us assume that $F = f + g$ has a minimizer $x^\star \in \mathbb{R}^d$.
For a constant step size $\gamma_k = \frac{1}{L}$ ($\forall k \in \mathbb{N}$):

$$F(x_k) - F(x^\star) \leq \frac{L}{2k} \|x_0 - x^\star\|_2^2.$$

In addition, if f is μ-strongly convex ($\mu > 0$), then:

$$\|x_k - x^\star\|_2^2 \leq c^k \|x_0 - x^\star\|_2^2,$$

where $c = 1 - \frac{m}{L} \in (0, 1)$.

Proof.

Similar to gradient descent, but considering the direction $d_k = \frac{1}{\gamma_k}(x_k - \text{prox}_{\gamma_k g}(x_k - \gamma_k \nabla f(x_k)))$. □

Remark

d_k is a direction of descent. Thus the proximal gradient method is a descent method.

Special cases

We can derive three special cases of the proximal gradient method:

1. when $g = 0$, the proximal gradient method is a gradient descent;
2. when $g = \chi_{\mathcal{C}}$ for a set $\mathcal{C} \subset \mathbb{R}^d$, the proximal method is a projected gradient descent;
3. when $f = 0$, we get the proximal point method.

Nesterov's method: add a momentum term to accelerate the proximal gradient descent.
Initial $\lambda_0 = 0$ and initial points $x_0 = y_0 \in \mathbb{R}^d$.

Algorithm Accelerated proximal gradient method.

$$x_{k+1} = \text{prox}_{\gamma_k g}(y_k - \gamma_k \nabla f(y_k))$$

$$\lambda_{k+1} = \frac{1 + \sqrt{1 + 4\lambda_k^2}}{2}$$

$$y_{k+1} = x_{k+1} + \frac{\lambda_k - 1}{\lambda_{k+1}}(x_{k+1} - x_k),$$

where $\gamma_k > 0$ is a step size.

---

y_k is an extrapolated point where the proximal gradient step is performed.

Remark

In image processing and compressed sensing, this method is often called FISTA, for fast iterative shrinkage-thresholding algorithm.

As always, the step size may be set to $\frac{1}{L}$ or chosen by line search.

Theorem (Convergence of the accelerated proximal method)

Consider (P8) with $f : \mathbb{R}^d \to \mathbb{R}$ being differentiable with L-Lipschitz gradient ($L > 0$), and $g : \mathbb{R}^d \to (-\infty, \infty]$ being proper, lower-semicontinuous and convex. Let us assume that $F = f + g$ has a minimizer $x^\star \in \mathbb{R}^d$. For a constant step size $\gamma_k = \frac{1}{L}$ ($\forall k \in \mathbb{N}$):

$$F(x_k) - F(x^\star) \leq \frac{2L}{k^2} \|x_0 - x^\star\|_2^2.$$

Primal methods

Here, we focus on the optimization problem:

$$\underset{x\in\mathbb{R}^d}{\text{minimize}}\; f(x)+g(x), \tag{P9}$$

where $f:\mathbb{R}^d\to(-\infty,\infty]$ and $g:\mathbb{R}^d\to(-\infty,\infty]$ are two convex functions.
Contrarily to the proximal gradient method, the Douglas-Rachford does not require f to be differentiable.
Initial point $y_0\in\mathbb{R}^d$.

Algorithm Douglas-Rachford method.

$$\begin{aligned} x_k &= \text{prox}_{\gamma_k f}(y_k)\\ y_{k+1} &= y_k+\mu_k\left(\text{prox}_{\gamma_k g}\left(2x_k-y_k\right)-x_k\right), \end{aligned}$$

where $\gamma_k>0$ is a step size (without restriction) and $\mu_k\in(0,2)$.

Primal methods

Douglas-Rachford iteration can be written as fixed-point iteration:

$$y_{k+1} = y_k + \mu_k(F_{\gamma_k}(y_k) - y_k), \tag{1}$$

with $F_\gamma(y) = y + \operatorname{prox}_{\gamma g}\left(2\operatorname{prox}_{\gamma f}(y) - y\right) - \operatorname{prox}_{\gamma f}(y)$.

- Usual Douglas-Rachford algorithm: $\mu_k = 1$ $(\forall k \in \mathbb{N})$, $y_{k+1} = F_1(y_k)$.
- Over-relaxation: $1 < \mu_k < 2$.
- Under-relaxation: $0 < \mu_k < 1$.

Interpretation of γ_k: Iteration (1) corresponds to

$$y_{k+1} = y_k + \mu_k(F_1(y_k) - y_k),$$

for minimizing $x \mapsto \gamma_k f(x) + \gamma_k g(x)$.

Remark

In practice, we usually consider $\mu_k = \gamma_k = 1$ $(\forall k \in \mathbb{N})$.

Idea (with $\mu_k = \gamma_k = 1$): y is a fixed point of F_1 if and only if

1. $x = \text{prox}_f(y)$;
2. $0 = (y - x) + (x - y) \in \partial f(x) + \partial g(x)$.

It is enough that $y - x \in \partial f(x)$ and $x - y \in \partial g(x)$.

Theorem (Convergence of Douglas-Rachford method)

Consider (P9) with $f : \mathbb{R}^d \to (-\infty, \infty]$ and $g : \mathbb{R}^d \to (-\infty, \infty]$ being two proper, lower semi-continuous and convex functions. Let us assume that $f + g$ has a minimizer in $\mathbb{R}^d$. For a fixed step size $\gamma_k = \gamma > 0$ and relaxation parameter $\mu_k \in [\underline{\mu}, \overline{\mu}]$ $(\forall k \in \mathbb{N})$, where $0 < \underline{\mu} \leq \overline{\mu} < 2$, the sequence $(x_k)_{k \in \mathbb{N}}$ generated by the Douglas-Rachford method converges to a minimizer of $f + g$.

Summary

Primal-dual methods

We have seen previously that the proximal gradient method, used for minimizing a composite objective function, reduces to:

1. the gradient method when $g = 0$;
2. the proximal point method when $f = 0$.

In this section, we exploit these tow simple algorithms with a dual problem to devise primal-dual methods.

Primal-dual methods

Let us consider the optimization problem:

$$\underset{x\in\mathbb{R}^d}{\text{minimize}}\ f(x)+g(Ax), \tag{P10}$$

where $f\colon \mathbb{R}^d \to (-\infty,\infty]$, $g\colon \mathbb{R}^p \to (-\infty,\infty]$ are proper convex and $A\in\mathbb{R}^{p\times d}$.
This problem is equivalent to

$$\begin{aligned}\underset{x\in\mathbb{R}^d,y\in\mathbb{R}^p}{\text{minimize}}\ & f(x)+g(y)\\ \text{s.t.}\quad & Ax=y.\end{aligned}$$

The dual function to this last problem is:

$$\begin{aligned}\forall \nu \in \mathbb{R}^p \colon G(\nu) &= \inf_{x\in\mathbb{R}^d, y\in\mathbb{R}^p} \left\{ f(x) + g(y) + \nu^\top A x - \nu^\top y \right\} \\ &= -\sup_{y\in\mathbb{R}^p} \left\{ \nu^\top y - g(y) \right\} - \sup_{x\in\mathbb{R}^d} \left\{ -\nu^\top A x - f(x) \right\} \\ &= -g^*(\nu) - f^*(-A^\top \nu).\end{aligned}$$

Therefore, the dual problem of interest is:

$$\underset{\nu\in\mathbb{R}^p}{\text{maximize}} \; -g^*(\nu) - f^*(-A^\top \nu). \tag{P11}$$

Set constraint

Assume that $g = \chi_{\mathcal{C}}$, where $\mathcal{C} \subset \mathbb{R}^d$ is a convex set, and we aim at solving:

$$\underset{x \in \mathbb{R}^d}{\text{minimize}}\ f(x) + \chi_{\mathcal{C}}(Ax - b),$$

whose dual is

$$\underset{\nu \in \mathbb{R}^p}{\text{maximize}}\ -b^\top \nu - \chi_{\mathcal{C}}^*(\nu) - f^*(-A^\top \nu).$$

	Constraint	Set $\mathcal{C}$	Conjugate
Equality	$Ax = b$	$\{0\}$	0
Ball	$\|Ax - b\| \leq 1$	unit $\|\cdot\|$-ball	$\|\cdot\|_*$
Conic inequality	$Ax \preccurlyeq_{\mathcal{K}} b$	$-\mathcal{K}$	$\chi_{\mathcal{K}^*}$

Table: Examples of constraints.

With $\mathcal{K} \subset \mathbb{R}^p$ being a proper convex cone.

Norm regularization

$g(y) = \|y - b\|$ and we want to minimize $f(x) + \|Ax - b\|$.
$g^*(\nu) = b^\top \nu + \chi_{\mathcal{B}}(\nu)$, where $\mathcal{B} = \{y \in \mathbb{R}^p : \|y\|_* \leq 1\}$
The dual reads:

$$\begin{aligned} \underset{\nu \in \mathbb{R}^p}{\text{maximize}} \; & -b^\top \nu - f^*(-A^\top \nu) \\ \text{s.t.} \quad & \|\nu\|_* \leq 1. \end{aligned}$$

Primal-dual methods

We consider (P10) and its dual (P11) when $g = \chi_{\{b\}}$ ($b \in \mathbb{R}^p$), which boils down to the following primal:

$$\begin{aligned} \underset{x \in \mathbb{R}^d}{\text{minimize}} \; & f(x) \\ \text{s.t.} \quad & Ax = b, \end{aligned} \tag{P12}$$

and dual:

$$\underset{\nu \in \mathbb{R}^p}{\text{minimize}} \; b^\top \nu + f^*(-A^\top \nu). \tag{P13}$$

Starting with $(x_0, \nu_0) \in \mathbb{R}^d \times \mathbb{R}^p$, the method of interest is:

Algorithm Method of Lagrange multipliers.

$$x_{k+1} \in \underset{x \in \mathbb{R}^d}{\arg\min}\, L(x, \nu_k)$$

$$\nu_{k+1} = \nu_k + \gamma_k (Ax_{k+1} - b).$$

where $\gamma_k > 0$ is a step size (without restriction).

Proposition

Let $f : \mathbb{R}^d \to (-\infty, \infty]$ be a convex, proper and lower semi-continuous function and $\mu > 0$. Then f is μ-strongly convex if and only if f^* is differentiable and ∇f^* is μ^{-1}-Lipschitz continuous.

Theorem (Method of Lagrange multipliers)

If f is μ-strongly convex ($\mu > 0$), proper and lower semi-continuous, then the method of Lagrange multipliers for (P12) is the gradient method applied to (P13).
In addition, if the Lagrangian of (P12) has a saddle point and if $\gamma_k \leq \frac{\mu}{\sigma_A^2}$, where $\sigma_A > 0$ is the largest singular value of A, then $((x_k, \nu_k))_{k \in \mathbb{N}}$ converges to a saddle point of the Lagrangian.

Proof.

By the previous proposition, f^* is differentiable. The gradient method is:

$$
\begin{aligned}
& \nu_{k+1} = \nu_k + \gamma_k A \nabla f^*(-A^\top \nu_k) - \gamma_k b \\
\iff & \exists x_{k+1} \in \mathbb{R}^d : x_{k+1} = \nabla f^*(-A^\top \nu_k), \nu_{k+1} = \nu_k + \gamma_k (A x_{k+1} - b) \\
\iff & \exists x_{k+1} \in \mathbb{R}^d : -A^\top \nu_k \in \partial f(x_{k+1}), \nu_{k+1} = \nu_k + \gamma_k (A x_{k+1} - b) \\
\iff & \exists x_{k+1} \in \mathbb{R}^d : 0 \in \partial f(x_{k+1}) + A^\top \nu_k, \nu_{k+1} = \nu_k + \gamma_k (A x_{k+1} - b) \\
\iff & \exists x_{k+1} \in \operatorname*{arg\,min}_{x \in \mathbb{R}^d} f(x) + \nu_k^\top (Ax - b), \nu_{k+1} = \nu_k + \gamma_k (A x_{k+1} - b) \\
\iff & \exists x_{k+1} \in \operatorname*{arg\,min}_{x \in \mathbb{R}^d} L(x, \nu_k), \nu_{k+1} = \nu_k + \gamma_k (A x_{k+1} - b),
\end{aligned}
$$

where L is the Lagrangian of (P12).

□

Proximal point method

The proximal point method is obtained from the proximal gradient descent with $f = 0$.
In a general context, the proximal point method is defined for minimizing a proper convex lower semi-continuous function $h\colon \mathbb{R}^d \to (-\infty, \infty]$:

Algorithm Proximal point method.

$$x_{k+1} = \operatorname{prox}_{\gamma_k h}(x_k),$$

where $\gamma_k > 0$ is a step size (without restriction).

It is mainly a conceptual algorithm.
The step size γ_k affects both the number of iterations to reach an ϵ-solution and the cost of prox-evaluations.

Definition (Augmented Lagrangian function)

Let $\gamma > 0$ be a parameter. The augmented Lagrangian function associated to (P12) is:

$$L_\gamma : (x, \nu) \in \mathbb{R}^d \times \mathbb{R}^p \mapsto f(x) + \nu^\top (Ax - b) + \frac{\gamma}{2} \|Ax - b\|_2^2.$$

Starting with an initial primal-dual point $(x_0, \nu_0) \in \mathbb{R}^d \times \mathbb{R}^p$, the augmented Lagrangian method is:

Algorithm Augmented Lagrangian method.

$$x_{k+1} \in \operatorname*{arg\,min}_{x \in \mathbb{R}^d} L_{\gamma_k}(x, \nu_k)$$

$$\nu_{k+1} = \nu_k + \gamma_k (Ax_{k+1} - b),$$

where $\gamma_k > 0$ is a step size (without restriction).

Theorem (Augmented Lagrangian method)

If f is convex, proper and lower semi-continuous, then the augmented Lagrangian method for (P12) is the proximal point method applied to (P13).
As a consequence, $(\nu_k)_{k\in\mathbb{N}}$ converges to a dual solution.

Proof.

...

□

Proof.

Defining $h\colon \nu \in \mathbb{R}^p \mapsto b^\top \nu + f^*(-A^\top \nu)$, the proximal point method reduces to:

$$
\begin{aligned}
& \nu_{k+1} = \operatorname{prox}_{\gamma_k h}(\nu_k) \\
\iff & \nu_{k+1} = \operatorname*{arg\,min}_{\nu \in \mathbb{R}^p} \gamma_k h(\nu) + \frac{1}{2}\|\nu - \nu_k\|_2^2 \\
\iff & 0 \in \gamma_k \partial h(\nu_{k+1}) + \nu_{k+1} - \nu_k \\
\iff & 0 \in -\gamma_k A \partial f^*(-A^\top \nu_{k+1}) + \gamma_k b + \nu_{k+1} - \nu_k \\
\iff & \exists x_{k+1} \in \partial f^*(-A^\top \nu_{k+1}) : \nu_{k+1} = \nu_k + \gamma_k(Ax_{k+1} - b) \\
\iff & \exists x_{k+1} \in \mathbb{R}^d : -A^\top \nu_{k+1} \in \partial f(x_{k+1}), \nu_{k+1} = \nu_k + \gamma_k(Ax_{k+1} - b) \\
\iff & \exists x_{k+1} \in \mathbb{R}^d : 0 \in \partial f(x_{k+1}) + A^\top \nu_{k+1}, \nu_{k+1} = \nu_k + \gamma_k(Ax_{k+1} - b) \\
\iff & \exists x_{k+1} \in \mathbb{R}^d : 0 \in \partial f(x_{k+1}) + A^\top \nu_k + \gamma_k A^\top(Ax_{k+1} - b), \nu_{k+1} = \nu_k + \gamma_k(Ax_{k+1} - b) \\
\iff & \exists x_{k+1} \in \operatorname*{arg\,min}_{x \in \mathbb{R}^d} f(x) + \nu_k^\top(Ax - b) + \frac{\gamma_k}{2}\|Ax - b\|_2^2 : \nu_{k+1} = \nu_k + \gamma_k(Ax_{k+1} - b) \\
\iff & \exists x_{k+1} \in \operatorname*{arg\,min}_{x \in \mathbb{R}^d} L_{\gamma_k}(x, \nu_k) : \nu_{k+1} = \nu_k + \gamma_k(Ax_{k+1} - b).
\end{aligned}
$$

The assumption that f is convex, proper and lower semi-continuous is necessary to state that

$x_{k+1} \in \partial f^*(-A^\top \nu_{k+1}) \iff -A^\top \nu_{k+1} \in \partial f(x_{k+1})$.

Primal-dual methods

We consider (P10) and its dual (P11) when f and g are only proximable functions, which leads to primal:

$$\begin{aligned} \underset{x\in\mathbb{R}^d, y\in\mathbb{R}^p}{\text{minimize}} \; & f(x)+g(y) \\ \text{s.t.} \quad & Ax = y. \end{aligned} \tag{P14}$$

and dual:

$$\underset{\nu\in\mathbb{R}^p}{\text{maximize}} \; -g^*(\nu) - f^*(-A^\top \nu). \tag{P15}$$

As a reminder, the augmented Lagrangian for (P14) is:

$$L_\gamma(x,y,\nu) = f(x)+g(y)+\nu^\top(Ax-y)+\frac{\gamma}{2}\|Ax-y\|_2^2.$$

Primal-dual methods

Initial primal-dual point $(x_0, y_0, \nu_0) \in \mathbb{R}^d \times \mathbb{R}^p \times \mathbb{R}^p$.

Algorithm Alternating direction method of multipliers.

$$x_{k+1} \in \operatorname*{arg\,min}_{x \in \mathbb{R}^d} L_{\gamma_k}(x, y_k, \nu_k) = \operatorname*{arg\,min}_{x \in \mathbb{R}^d} \left(f(x) + \nu_k^\top Ax + \frac{\gamma}{2} \|Ax - y_k\|_2^2 \right)$$

$$y_{k+1} \in \operatorname*{arg\,min}_{y \in \mathbb{R}^p} L_{\gamma_k}(x_{k+1}, y, \nu_k) = \operatorname*{arg\,min}_{y \in \mathbb{R}^p} \left(g(y) - \nu_k^\top y + \frac{\gamma}{2} \|Ax_{k+1} - y\|_2^2 \right)$$

$$\nu_{k+1} = \nu_k + \gamma_k (Ax_{k+1} - y_{k+1}),$$

where $\gamma_k > 0$ is a step size (without restriction).

Theorem (Alternating direction method of multipliers)

If f and g are convex, proper and lower semi-continuous, then the alternating direction method of multipliers for (P14) is the Douglas-Rachford method applied to (P15).
As a consequence, $(\nu_k)_{k \in \mathbb{N}}$ converges to a dual solution.