Optimization

Marc Castella
marc.castella@telecom-sudparis.eu

Télécom SudParis

September 7, 2023

Part I

Introduction

Topics in the course

- Introduction and generalities about optimization
- Notions of convexity
  - Convex sets and functions
  - Separation theorem
- Optimization problems
  - (Convex optimizations problems: LP, QP, SOCP, SDP)
  - Optimality conditions
- Duality
  - Lagrange duality
  - Conjugate function and Fenchel duality
  - Karush-Kuhn-Tucker optimality conditions
- Algorithms
  - Notions on unconstrained optimization (gradient, Newton)
  - Notions on constrained optimization (interior points)
  - Basic introduction to proximal methods

Optimization softwares

Many free and commercial softwares exist for optimization:

- optimization solvers: SeDuMi, SDPT3, CPLEX, Gurobi, Mosek, ...
- high level modelling languages and parsers: CVX, YALMIP, ...

but many algorithms are not that complicated and can be programmed (e.g. with Matlab/Scientific Python)!

Useful references

Convex optimization:

- Boyd and Vandenberghe, *Convex Optimization* (Cambridge University Press)
- http://stanford.edu/~boyd/
- Borwein and Lewis, *Convex Analysis and Nonlinear Optimization, Theory and Examples* (Canadian Mathematical Society)

Proximal algorithms:

- N. Parikh and S. Boyd, *Proximal Algorithms* (Foundations and Trends in Optimization, 1(3):123-231, 2014)
- S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, *Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers* (Foundations and Trends in Machine Learning, 3(1):1–122, 2011.)

Some notations

$\mathbb{R}$ $\mathbb{R}_+$ $\mathbb{R}_{++}$	real numbers nonnegative (≥ 0) numbers positive (> 0) numbers
$\mathbb{S}^n$ $\mathbb{S}^n_+/\mathbb{S}^n_{++}$	$n \times n$ real-valued symmetric matrices $n \times n$ sym. positive semidefinite /definite matrices
$A^\top$ $\operatorname{tr} A$	transpose of the matrix A trace of the matrix A
$\mathbf{1}$ $\|.\|_2$ $\|.\|_1$ / $\|.\|_\infty$ $\sup$ / $\inf$	all ones (column) vector Euclidian norm ℓ_1 / ℓ_∞ norm supremum / infimum
$\preceq_K$ / $\succeq_K$ / $\prec_K$ / $\succ_K$ $\preceq$ / $\succeq$ ($\prec$ / $\succ$)	inequalities wrt to cone K. If not specified, K is positive orthant or $\mathbb{S}^n_+$
$[.]_+$	positive part $[x]_+ = \max(0, x)$

Optimization problems

Unconstrained optimization problem

Given a function $f_0 : \mathbb{R}^n \to \mathbb{R}$, find $x^\star \in \mathbb{R}^n$ such that:

$$\forall x \in \mathbb{R}^n : \quad f_0(x^\star) \leq f_0(x)$$

Constrained optimization problem

Given functions $f_0 : \mathbb{R}^n \to \mathbb{R}$ and $f_i : \mathbb{R}^n \to \mathbb{R}$ for $i = 1, \ldots, m$, find $x^\star$ such that:

$$\begin{aligned} & f_i(x^\star) \leq 0, i = 1, \ldots, m \\ & f_0(x^\star) \leq f_0(x), \quad \forall x \in \mathbb{R}^n \text{ such that } f_i(x) \leq 0, i = 1, \ldots, m \end{aligned}$$

Discrete optimization (not covered in this course):

f_0 and f_i are functions $\mathcal{D} \to \mathbb{R}$ with:

- $\mathcal{D}$ finite : combinatorial optimization problem
- $\mathcal{D} = \mathbb{Z}$: integer programming

Optimization problem

$$\begin{cases} \min.\ f_0(x) \\ \quad \text{s.t. } f_i(x) \leq 0, \quad i = 1, \ldots, m \end{cases}$$

- $x = (x_1, \ldots, x_n)^\top$: optimization variables
- $f_0 : \mathbb{R}^n \to \mathbb{R}$: objective function
- $f_i : \mathbb{R}^n \to \mathbb{R}, i = 1, \ldots, m$: constraint functions

Optimal value: $p^\star := \inf\{f_0(x) \,|\, f_i(x) \leq 0,\ \text{for } i = 1, \ldots, m\}$

Optimal solution: $x^\star$ satisfies $f_i(x^\star) \leq 0,\ i = 1, \ldots, m$ and:

$$f_0(x^\star) \leq f_0(x) \text{ for all } x \text{ that satisfy } f_i(x) \leq 0,\ i = 1, \ldots, m\,.$$

Examples

Portfolio optimization

- variables: amounts invested in different assets
- constraints: budget, max./min. investment per asset, minimum return
- objective: overall risk or return variance

Data fitting

- variables: model parameters
- constraints: prior information, parameter limits
- objective: measure of misfit or prediction error

Signal restoration

- variables: signal values
- constraints: prior informations, value limits
- objective: data fit + regularization

Example: (linear) classification

- Training data $(f_i, c_i)_{i=1,\dots,m}$ where for any $i = 1, \dots, m$:
  - $f_i \in \mathbb{R}^n$: features,
  - $c_i \in \{+1, -1\}$: category.
- Classify new data $f \in \mathbb{R}^n$ in the two classes.
  Linear classifier $\hat{c} = \mathrm{sign}(x^\top f)$: find weight vector x
- Associated optimization problem with ℓ_2 regularization:

$$\min_x . \sum_{i=1}^{m} \varphi\Big(- c_i(x^\top f_i)\Big) + \gamma \|x\|_2 \qquad (\gamma = \text{const. } > 0)$$

where cost function $\varphi(z)$ can be:

- $\varphi(z) = \mathbb{1}(z \geq 0)$
- $\varphi(z) = \log(1 + e^{-z})$ (logistic regression)
- $\varphi(z) = [1 - z]_+$ (support vector machine)
- $\varphi(z) = e^z$

General optimization problem:

- very difficult to solve (if nonconvex)
- methods involve some compromise, *e.g.*:
  - **local** optimization method (*nonlinear programming*): not always finding the solution
  - **global** optimization: very long computation time, worst case complexity grows exponentially with problem size

$\rightsquigarrow$ These algorithms are often based on solving convex subproblems

Convex optimization problems can be solved efficiently and reliably:

- least-squares problems (analytical solution even exist in this case)
- linear programming problems
- many other convex programming problems

Convex optimization problem

$$\begin{cases} \min.\ f_0(x) \\ \text{s.t. } f_i(x) \leq b_i, \quad i = 1, \dots, m \end{cases}$$

- Objective and constraint functions are convex
- Includes as special cases: least squares, linear programming
- Convex optimization is "almost a technology":
  - reliable and efficient algorithms (but generally no analytical solutions)
  - computation time (roughly) proportional to $\max\{n^3, n^2m, F\}$ where F is cost of evaluating f_i's and their first+second derivatives
- Many problems can be solved via convex optimization:
  - often difficult to recognize
  - many tricks for transforming problems

Euclidian space

Euclidian space $\mathbf{E}$ (finite dimension) with inner-product $\langle ., . \rangle$

- Often $\mathbf{E} = \mathbb{R}^n$ and $\langle x, y \rangle = x^\top y = \sum_{i=1}^n x_i y_i$
- (Euclidian) norm $\|x\|_2 = \sqrt{\langle x, x \rangle}$
- Cauchy-Schwarz inequality: $|\langle x, y \rangle| \leq \|x\|_2 \|y\|_2$
- Orthogonal complement:

$$G^\perp = \{y \in \mathbf{E} \,|\, \langle x, y \rangle = 0 \text{ for all } x \in G\}$$

- Ball of center x_0 radius $r \geq 0$:

$$B(x_0, r] = \{x \in \mathbf{E} \,|\, \|x - x_0\| \leq r\} \qquad \text{(closed ball)}$$
$$B(x_0, r[= \{x \in \mathbf{E} \,|\, \|x - x_0\| < r\} \qquad \text{(open ball)}$$

Dual norm

Let $\|.\|$ be a norm on $\mathbf{E}$.
Associated **dual norm** $\|.\|_*$:

$$\|z\|_* := \sup_{\|x\|\leq 1} \langle z, x\rangle$$

- $\langle z, x\rangle \leq \|z\|_* \|x\|$
- Dual norm of $\|.\|_2$ is itself.
- $\|.\|_\infty$ and $\|.\|_1$ are dual norms of each other.
- Dual of ℓ_p-norm is ℓ_q norm with $\frac{1}{p} + \frac{1}{q} = 1$.
- $\|.\|_{**} = \|.\|$ (need not hold in infinite dimensional spaces)

Open and closed sets

Interior, closure, boundary

interior of a set C:

$$\operatorname{int} C = \{x \in C \mid B(x, \varepsilon[\subset C \text{ for sufficiently small } \varepsilon\}$$

A set C is **open** if $C = \operatorname{int} C$ and **closed** if its complement is open.

closure of a set C:

$$\operatorname{cl} C = \{x \in \mathbf{E} \mid \text{for any (small) } \varepsilon, B(x, \varepsilon[\cap C \neq \emptyset\}$$

boundary of a set C: $\operatorname{bd} C = \operatorname{cl} C \setminus \operatorname{int} C$

core of a set C = set of points $x \in C$ such that for any direction $d \in \mathbf{E}$, $x + td \in C$ for all small t. Note that $\operatorname{int} C \subseteq \operatorname{core} C$ (but $\operatorname{core} C$ may be larger than $\operatorname{int} C$).

Linear maps, adjoint, null space

$\mathbf{E}$ and $\mathbf{F}$ two Euclidian spaces.

- $A : \mathbf{E} \to \mathbf{F}$ is linear if $A(\lambda x + \mu y) = \lambda Ax + \mu Ay$ for any $x, y \in \mathbf{E}$ and $\lambda, \mu \in \mathbb{R}$.
- Linear functions $\mathbf{E} \to \mathbb{R}$ have the form $\langle a, . \rangle$ for some $a \in \mathbf{E}$
- Affine functions = linear + constant
- **Adjoint** of A is the linear map $A^* : \mathbf{F} \to \mathbf{E}$ such that:

$$\langle A^* y, x \rangle = \langle y, Ax \rangle \text{ for any } x \in \mathbf{E}, y \in \mathbf{F}$$

  - If $\mathbf{E} = \mathbb{R}^n$, $\mathbf{F} = \mathbb{R}^p$, adjoint of $A : \mathbb{R}^n \to \mathbb{R}^p$ is given by $A^\top$
- Null space (kernel): $\operatorname{Ker} A = \{x \in \mathbf{E} \,|\, Ax = 0\}$

Symmetric matrices

- Set of symmetric matrices: $\mathbb{S}^n = \{M \in \mathbb{R}^{n\times n} \,|\, M^\top = M\}$
- Positive semidefinite matrices: $\mathbb{S}^n_+ = \{M \in \mathbb{S}^n \,|\, x^\top M x \geq 0 \text{ for all } x\}$
- Positive definite matrices: $\mathbb{S}^n_{++} = \{M \in \mathbb{S}^n \,|\, x^\top M x > 0 \text{ for all } x \neq 0\}$
- Inner product:

$$\langle A, B\rangle = \operatorname{tr} AB \text{ for } A, B \in \mathbb{S}^n$$

- $M \in \mathbb{S}^n_+$ (resp. $\mathbb{S}^n_{++}$) will be written $M \succeq 0$ (resp. $M \succ 0$). Similarly (see later):

$$A - B \in \mathbb{S}^n_+ \Leftrightarrow A \succeq B \qquad\qquad A - B \in \mathbb{S}^n_{++} \Leftrightarrow A \succ B$$

Domain and extended-value function

Let f be a function $\mathbf{E} \to \mathbb{R}$ (often, $\mathbf{E} = \mathbb{R}^n$).

Domain: $\operatorname{dom} f = \{x \in \mathbf{E} \,|\, f(x) \text{ exists}\}$ $\qquad (\operatorname{dom} f \subset \mathbf{E})$

If $f : \operatorname{dom} f \to \mathbb{R}$, we use the extended-value extension of f:

$$\begin{aligned} f :& \mathbf{E} \to \mathbb{R} \cup \{+\infty\} \\ & x \mapsto \begin{cases} f(x) & \text{if } x \in \operatorname{dom} f \\ +\infty & \text{if } x \notin \operatorname{dom} f \end{cases} \end{aligned}$$

- Often simplifies the notation and provides a unifying view.
- $\operatorname{dom} f = \{x \in \mathbf{E} \,|\, f(x) < \infty\}$
  - If $\operatorname{dom} f \neq \emptyset$, the function is said **proper**

Extended-value functions

Examples

- **Log-barrier** $f : \mathbb{R} \to \mathbb{R}$ defined by:

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

$\operatorname{dom} f = \mathbb{R}_{--}$

- **Indicator** function of a given set $C \subset \mathbf{E}$:

$$\imath_C(x) = \begin{cases} 0 & \text{if } x \in C, \\ +\infty & \text{otherwise.} \end{cases}$$

$\operatorname{dom} \imath_C = C$

Gradient vector

Let $f : \mathbb{R}^n \to \mathbb{R}$.

- **Gradient** (column) vector $\nabla f(x)$:

$$[\nabla f(x)]_i = \frac{\partial f(x)}{\partial x_i}$$

  First-order approximation of f near $\overline{x}$:

$$\hat{f}_1(x) = f(\overline{x}) + \nabla f(\overline{x})^\top (x - \overline{x})$$

- Ex:

$$\begin{aligned} f(x) &= a^\top x & \nabla f(x) &= a \\ g(x) &= x^\top M x & \nabla g(x) &= (M + M^\top)x \\ & & &= 2Mx \text{ if } M \text{ symmetric.} \end{aligned}$$

Hessian matrix

Let $f : \mathbb{R}^n \to \mathbb{R}$.

- **Hessian** matrix $\nabla^2 f(x)$ (often denoted by $H(x)$ in this course):

$$[\nabla^2 f(x)]_{ij} = \frac{\partial^2 f(x)}{\partial x_i \partial x_j}$$

  Second-order approximation of f near $\overline{x}$:

$$\hat{f}_2(x) = f(\overline{x}) + \nabla f(\overline{x})^\top (x - \overline{x}) + \frac{1}{2}(x - \overline{x})^\top \nabla^2 f(\overline{x})(x - \overline{x})$$

- Ex:

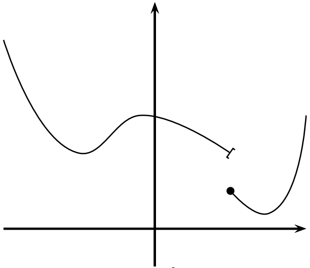
$$\begin{aligned} f(x) &= a^\top x & \nabla^2 f(x) &= 0 \\ g(x) &= x^\top M x & \nabla^2 g(x) &= M + M^\top \\ & & &= 2M \text{ if } M \text{ symmetric.} \end{aligned}$$

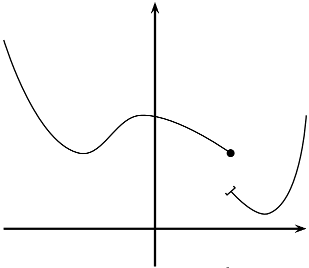
Lower semi-continuous function (l.s.c.)

f is l.s.c. if and only if at any point x:

$$x_n \xrightarrow[n\to\infty]{} x \quad \Rightarrow f(x) \leq \lim_{n\to\infty} f(x_n)$$

f is l.s.c. $\Leftrightarrow$ epigraph $\{(x,t) \in \mathbb{R}^n \times \mathbb{R} \,|\, f(x) \leq t\}$ is a closed set


l.s.c.


non l.s.c.

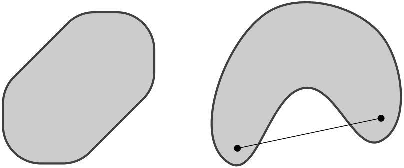
Part II

Convexity, convex optimization

Convex set

Convex set: contains line segment between any two points in the set

$$x, y \in C,\ 0 \leq \theta \leq 1 \Rightarrow \theta x + (1-\theta) y \in C$$

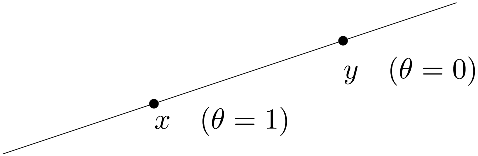


- Points of the form $\theta x + (1-\theta)y$ with $0 \leq \theta \leq 1$ corresponds to the line segment between x and y.

Affine set

Affine set: the line through any two points in the set is contained in the set

$$x, y \in C,\ \theta \in \mathbb{R} \Rightarrow \theta x + (1-\theta) y \in C$$



- Points of the form $\theta x + (1-\theta)y$ with $\theta \in \mathbb{R}$ corresponds to the line through x and y.

Affine and convex hull

Affine hull of set C = all affine combinations of points in C

$$\operatorname{aff} C = \{\theta_1 x_1 + \cdots + \theta_k x_k \,|\, x_i \in C, \theta_1 + \cdots + \theta_k = 1\}$$

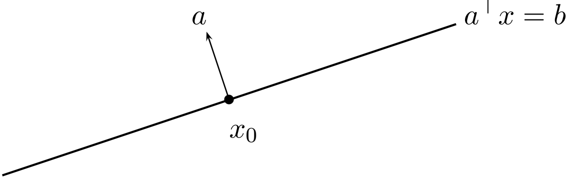
Convex hull of set C = all convex combinations of points in C

$$\operatorname{conv} C = \{\theta_1 x_1 + \cdots + \theta_k x_k \,|\, x_i \in C, \theta_i \geq 0, \theta_1 + \cdots + \theta_k = 1\}$$

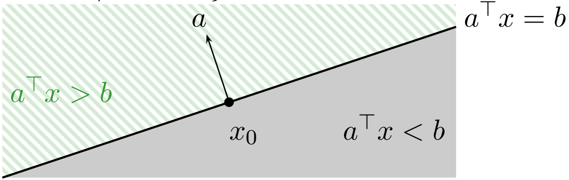
Hyperplanes and halfspaces in $\mathbb{R}^n$

Let $a \in \mathbb{R}^n, a \neq 0$ and $b \in \mathbb{R}$:

- Hyperplane: $\{x \in \mathbb{R}^n \,|\, a^\top x = b\}$: convex and affine



- Halfspace: $\{x \in \mathbb{R}^n \,|\, a^\top x \leq b\}$: convex but not affine



- a is the normal vector
- The hyperplane separates the whole space $\mathbb{R}^n$ in two halfspaces

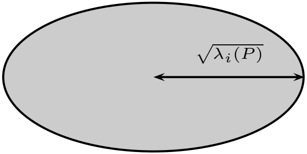
Balls and ellipsoids

Euclidian ball:

$$\begin{aligned} B(\overline{x}, r] &= \{x \mid \|x - \overline{x}\|_2 \leq r\} = \{x \mid (x - \overline{x})^\top (x - \overline{x}) \leq r^2\} \\ &= \{\overline{x} + ru \mid \|u\|_2 \leq 1\} \end{aligned}$$

Ellipsoid:

$$\mathcal{E} = \{x \mid (x - \overline{x})^\top P^{-1} (x - \overline{x}) \leq 1\} \quad \text{where } P \in \mathbb{S}^n_{++}$$



With $A = P^{1/2}$, other representation: $\mathcal{E} = \{\overline{x} + Au \mid \|u\|_2 \leq 1\}$

Operations that preserve convexity (1/3)

Intersection : the intersection of any number of convex sets is convex

Ex:

- **Polyhedra**: intersection of a finite number of hyperplanes/halfspaces
  - $\mathcal{P} = \{x \,|\, {a_j}^\top x \leq b_j, j = 1, \ldots, m, {c_i}^\top x = d_i, i = 1, \ldots, p\}$
  - Simplex $\{\theta_0 v_0 + \cdots + \theta_k v_k \,|\, \theta \succeq 0, \mathbf{1}^\top \theta = 1\}$
    ($v_0, \ldots, v_k$ *affinely independent*)
- Intersection of halfspaces:
  $\{x \in \mathbb{R}^m \,/\, |\sum_{k=1}^m x_k \cos kt| \leq 1, \forall t \in [-\pi/3, \pi/3]\}$
- Positive semidefinite matrices: $\mathbb{S}_+^n = \bigcap_{x \neq 0} \{M \in \mathbb{S}^n \,|\, x^\top M x \geq 0\}$

Convex hull of a set S: intersection of all convex sets containing S.

Operations that preserve convexity (2/3)

Affine transformation: the image and inverse image of a convex set under an affine function is convex.

Ex:

- Scaling, translation, projection.
- Sum $S_1 + S_2 = \{x + y \; x \in S_1, \; y \in S_2\}$
- Partial sum $\{(x, y_1 + y_2) \; ; \; (x, y_1) \in S_1, \; (x, y_2) \in S_2\}$

- Polyhedron (inverse image of nonnegative orthant)
- Ellipsoid (image/inverse image of the unit Euclidian ball)
- Solution set of a Linear Matrix Inequality (LMI):
  $\{x \in \mathbb{R}^n \,|\, x_1 A_1 + \cdots + x_n A_n \preceq B\}$
  where $B, A_1, \ldots, A_n$ are given in $\mathbb{S}^p$

Operations that preserve convexity (3/3)

Perspective function

$$P(x,t) = \frac{x}{t} \text{ where } P : \mathbb{R}^n \times \mathbb{R}_{++} \to \mathbb{R}^n$$

$\to$ image and inverse image through perspective remains convex.

Linear-fractional $f(x) = \frac{Ax+b}{c^\top x+d}$ with $\operatorname{dom} f = \{x \,|\, c^\top x + d\} > 0 \to$ preserve convexity (as a composition of affine and perspective functions).

Relative interior

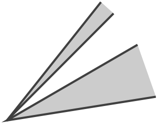
interior of a set C:

$$\operatorname{int} C = \{x \in C \mid B(x, \varepsilon[\subset C \text{ for sufficiently small } \varepsilon\}$$

relative interior of a set C = interior of C relative to its affine hull:

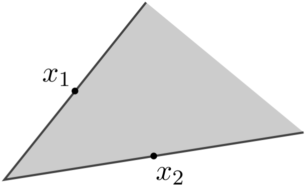
$$\operatorname{relint} C = \{x \in C \mid B(x, \varepsilon[\cap \operatorname{aff} C \subseteq C \text{ for sufficiently small } \varepsilon\}$$

Cones



Cone C: for every $x \in C$ and $\theta \geq 0$, we have $\theta x \in C$

Convex cone C: for every $x_1, x_2 \in C$ and $\theta_1, \theta_2 \geq 0$, we have $\theta_1 x_1 + \theta_2 x_2 \in C$



- Conic hull of a set C:
  $\{\theta_1 x_1 + \cdots + \theta_k x_k \,|\, x_i \in C, \theta_i \geq 0, i = 1, \ldots, k\}$

Examples of cones

- **Nonnegative orthant** $\mathbb{R}^n_+ = \{x \in \mathbb{R}^n \,|\, x_i \geq 0, i = 1, \dots, n\}$
- **Positive semidefinite matrices**

$$\mathbb{S}^n_+ = \{M \in \mathbb{S}^n \,|\, M \succeq 0\} = \{M \in \mathbb{S}^n \,|\, x^\top M x \geq 0, \forall x \in \mathbb{R}^n\}$$

where $\mathbb{S}^n$ is the set of symmetric matrices.

- **Norm cone**

$$\{(x, t) \in \mathbb{R}^{n+1} \,|\, \|x\| \leq t\}$$

When $\|.\| = \|.\|_2$, also called quadratic / second-order / Lorentz cone

- **Cone of positive polynomials**

$$K = \{p \in \mathbb{R}^n \,|\, p_1 + p_2 t + \cdots + p_n t^{n-1} \geq 0, \forall t \in [0, 1]\}$$

Normal cone

Normal cone to a convex set C at $\overline{x} \in C$:

$$\mathcal{N}_C(\overline{x}) = \{d \in \mathbf{E} \mid \langle d, x - \overline{x} \rangle \leq 0, \forall x \in C\}$$

when $\mathbf{E} = \mathbb{R}^n$, simplifies to:

$$\mathcal{N}_C(\overline{x}) = \{d \in \mathbb{R}^n \mid d^\top (x - \overline{x}) \leq 0, \forall x \in C\}$$

Proper cones, generalized inequalities

Proper cone:

- convex
- closed
- solid (*i.e.* nonempty interior)
- pointed (*i.e.* contains no line: $x \in K, -x \in K \Rightarrow x = 0$)

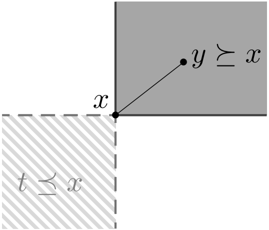
Generalized inequalities w.r.t. proper cone K:

$$x \preceq_K y \Leftrightarrow y - x \in K$$

$$x \prec_K y \Leftrightarrow y - x \in \operatorname{int} K \qquad \text{(interior of } K\text{)}$$

Examples of generalized inequalities

- $K = \mathbb{R}^n_+$ gives usual partial ordering on $\mathbb{R}^n$ (componentwise)



$$x \preceq_K y \iff x_i \leq y_i, \forall i$$

- $K = \mathbb{S}^n_+ =$ set of **symmetric positive semidefinite matrices**

$$A \preceq B \iff B - A \in \mathbb{S}^n_+$$

- $K =$ **cone of positive polynomials**

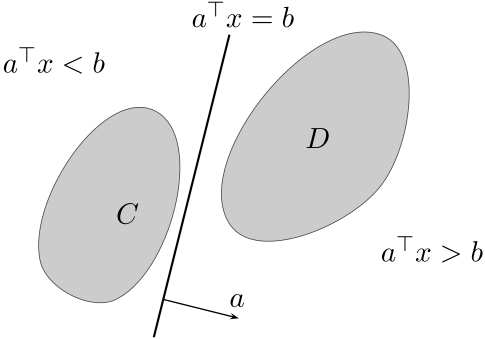
$$p \preceq_K q \iff 0 \leq (q_1 - p_1) + (q_2 - p_2)t + \cdots + (q_n - p_n)t^{n-1}, \forall t$$

Separating hyperplane

Separating hyperplane theorem If C and D are disjoint convex sets ($C \cap D = \emptyset$), there exist $a \neq 0, a \in \mathbb{R}^n$ and $b \in \mathbb{R}$ such that:

$$\forall x \in C, a^\top x \leq b \qquad \text{and } \forall x \in D, a^\top x \geq b$$

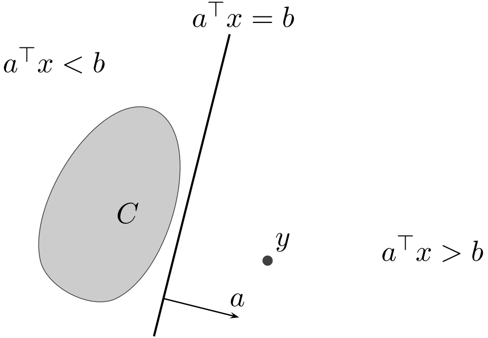
The hyperplane $\{x \,|\, a^\top x = b\}$ separates C and D.



Strict separation

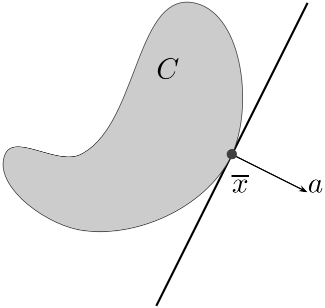
Basic separation If C closed and convex and $y \notin C$, there exist $a \neq 0, a \in \mathbb{R}^n$ and $b \in \mathbb{R}$ such that:

$$\forall x \in C, \quad a^\top x \leq b < a^\top y$$



Supporting hyperplanes

Supporting hyperplane $C \subset \mathbb{R}^n$, $\overline{x} \in \operatorname{bd} C$ If $a \neq 0$ and $\forall x \in C, a^\top x \leq a^\top \overline{x}$, then $\{x \in \mathbb{R}^n \,|\, a^\top x = a^\top \overline{x}\}$ is a supporting hyperplane of C.



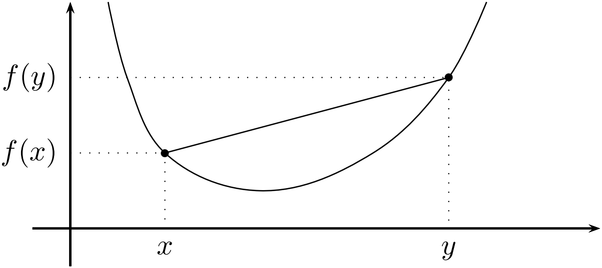
If C is convex, then there exist a supporting hyperplane at every boundary point of C.

Convex function

$f : \mathbb{R}^n \to \mathbb{R}$ is **convex** if $\mathrm{dom}\, f$ is a convex set and

$$f(\theta x + (1-\theta)y) \leq \theta f(x) + (1-\theta) f(y)$$

for all $x, y \in \mathrm{dom}\, f$, $0 \leq \theta \leq 1$.

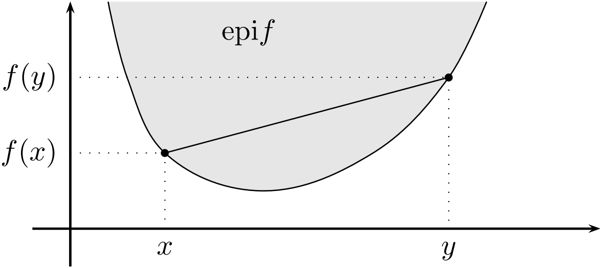


- **strictly convex** when: $f(\theta x + (1-\theta)y) < \theta f(x) + (1-\theta) f(y)$
- f is **concave** if $(-f)$ is convex.

Epigraph

The **epigraph** of a function $f : \mathbb{R}^n \to \mathbb{R}$ is:

$$\text{epi} f := \{(x, t) \in \mathbb{R}^{n+1} | x \in \text{dom}\, f, f(x) \leq t\}$$



- f is convex if and only if its epigraph is convex.
- sublevel set: $C_\alpha := \{x \,|\, f(x) \leq \alpha\}$

▷ C_α is a convex set if f convex

Jensen's inequality

For a convex function f:

- $f(\theta x + (1-\theta)y) \leq \theta f(x) + (1-\theta) f(y)$: called **Jensen's** inequality
- extends to
  - sums (finite or not): for $\theta_1, \ldots, \theta_p \geq 0,\ \theta_1 + \cdots + \theta_p = 1$:

$$f(\theta_1 x_1 + \cdots + \theta_p x_p) \leq \theta_1 f(x_1) + \cdots + \theta_p f(x_p)$$

  - integrals and expected values: if $p(x)$ is a pdf with support $S \subset \operatorname{dom} f$:

$$f(\int_S x p(x)\, dx) \leq \int_S f(x) p(x)\, dx \qquad f(\mathbb{E}\{X\}) \leq \mathbb{E}\{f(X)\}$$

Examples of convex/concave functions

convex

- $\|x\|$
- $\max(x_1, \ldots, x_n)$
- $f(x, y) = \frac{x^2}{y}$ with $\operatorname{dom} f = \mathbb{R} \times \mathbb{R}_{++}$
- $\log(e^{x_1} + \cdots + e^{x_n})$

concave

- $f(x) = \left(\prod_{i=1}^n x_i\right)^{1/n}$
- $f(X) = \log\det X$ with $\operatorname{dom} f = \mathbb{S}^n_{++}$.

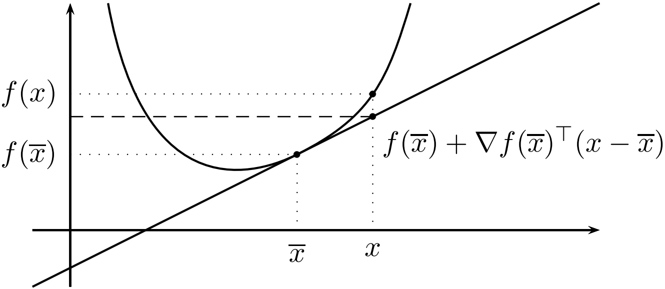
convex and **concave**

- affine functions: $f(x) = a^\top x + b$

First order conditions

Differentiable f with convex domain is convex if and only if:

$$f(x) \geq f(\overline{x}) + \nabla f(\overline{x})^\top (x - \overline{x}) \quad \forall x, \overline{x} \in \operatorname{dom} f$$



The linear approximation of f is a global underestimator.

Second order conditions

Twice differentiable f with convex domain:

$$f \text{ convex } \Leftrightarrow \nabla^2 f(x) \succeq 0 \quad \forall x \in \mathrm{dom}\, f$$

If $\nabla^2 f(x) \succ 0 \quad \forall x \in \mathrm{dom}\, f$, then f strictly convex.

- Ex: $f(x) = \frac{1}{2}x^\top Px + q^\top x + r$ defined on $\mathbb{R}^n$ is:
  - convex iff $P \succeq 0$ (concave iff $P \preceq 0$),
  - strictly convex iff $P \succ 0$ (strictly concave iff $P \prec 0$).

Operations that preserve convexity (1/3)

Nonnegative weighted sums: $f = w_1 f_1 + \cdots + w_m f_m$ is convex if $f_1, \ldots, f_m$ convex and $w_1, \ldots, w_m \geq 0$.

Composition with an affine mapping: $x \mapsto f(Ax + b)$ is convex (resp. concave) if f convex (resp. concave)

Pointwise maximum: $x \mapsto \max\{f_1(x), \ldots, f_m(x)\}$ is convex if $f_1, \ldots, f_m$ convex (extends to supremum).

Operations that preserve convexity (2/3)

Composition: let $g : \mathbb{R}^n \to \mathbb{R}^k$ and $h : \mathbb{R}^k \to \mathbb{R}$ and $f = h \circ g : \mathbb{R}^n \to \mathbb{R}$ defined by $f(x) = h(g(x))$.

- f is convex if h is convex nondecreasing and g is convex,
- f is convex if h is convex nonincreasing and g is concave,
- f is concave if h is concave nondecreasing and g is concave,
- f is concave if h is concave nonincreasing and g is convex.

(Easy proof in simple real valued differentiable case.)

Operations that preserve convexity (3/3)

Minimization: if $f(x,y)$ convex in (x,y), $C \neq \emptyset$, $g(x) = \inf_{y \in C} f(x,y)$ is convex in x provided $g(x) > -\infty$ for some x.

Perspective of a function: perspective function of $f : \mathbb{R}^n \to \mathbb{R}$ is $g : \mathbb{R}^{n+1} \to \mathbb{R}$ defined by

$$g(x,t) = tf(x/t)$$

The perspective preserves convexity.

How to prove convexity?

1. verify definition, often simplified by restricting to a line:
   - ▷ f is convex if and only if it is convex when restricted to any line that intersects $\operatorname{dom} f$

   Ex: prove concavity of $f(X) = \log\det X$ with $\operatorname{dom} f = \mathbb{S}^n_{++}$.
2. for twice differentiable functions, second-order condition
3. show that f is obtained from simple convex functions by operations that preserve convexity.

Optimization problem in standard form

General form, **non convex** (but can be):

$$\begin{cases} \min.\ f_0(x) \quad (x \in \mathcal{D} \subset \mathbb{R}^n) \\ \quad \text{s.t. } f_i(x) \leq 0, \quad i = 1, \dots, m \\ \qquad\quad h_j(x) = 0, \quad j = 1, \dots, p \end{cases}$$

- $x = (x_1, \dots, x_n)^\top$: optimization variables
- $f_0 : \mathcal{D} \to \mathbb{R}$: objective or cost function
- $f_i : \mathcal{D} \to \mathbb{R}, i = 1, \dots, m$: inequality constraint functions
- $h_j : \mathcal{D} \to \mathbb{R}, j = 1, \dots, p$: equality constraint functions

optimal value: $p^\star := \inf\{f_0(x) | f_i(x) \leq 0, h_j(x) = 0, x \in \mathcal{D}\}$

- $p^\star = +\infty$: problem unfeasible (no x satisfies the constraints)
- $p^\star = -\infty$: problem unbounded below

Vocabulary, remarks

- Constraints:
  - implicit: $x \in \mathcal{D}$ intersection of all functions domain: $\mathcal{D} \subset \operatorname{dom} f_i$ and $\mathcal{D} \subset \operatorname{dom} h_j$
  - explicit: $f_i(x) \leq 0$, $h_j(x) = 0$
  - unconstrained problem: only implicit constraints
- Feasible point: any x that satisfies the constraint.
  - feasibility problem = find a feasible point = special case of general problem with $f_0(x) = 0$
- optimal point $x^\star$:
  - $x^\star$ global optimal if feasible and $p^\star = f_0(x^\star) \leq f_0(x)$ for any feasible x
  - $x^\star_{\mathrm{loc}}$ local optimum if feasible and $f_0(x^\star_{\mathrm{loc}}) \leq f_0(x)$ for any x such that $\|x - x^\star_{\mathrm{loc}}\| \leq \alpha$ and x feasible.

Convex optimization problem (standard form)

$$\begin{cases} \text{min. } f_0(x) \quad (x \in \mathcal{D} = \cap_{i=0}^m \operatorname{dom} f_i) \\ \text{s.t. } f_i(x) \leq 0, \quad i = 1, \dots, m \\ \qquad {a_i}^\top x = b_i, \quad i = 1, \dots, p \end{cases}$$

- **objective** f_0 and **constraint** functions $f_1, \dots, f_m$ are convex
- **equality constraints** are **affine.**

often written as:

$$\begin{cases} \text{min. } f_0(x) \quad (x \in \mathcal{D}) \\ \text{s.t. } f_i(x) \leq 0, \quad i = 1, \dots, m \\ \qquad Ax = b \end{cases}$$

Remark: can be written **with inequalities only**.
Indeed, for $i = 1, \dots, p$, replace the equalities by the two inequalities ${a_i}^\top x - b_i \leq 0$ and $-{a_i}^\top x + b_i \leq 0$

Feasible set of a convex optimization problem

- General convex problem with inequalities only:

$$\begin{cases} \min.\ f_0(x) \quad (x \in \mathcal{D}) \\ \quad \text{s.t. } f_i(x) \leq 0, \quad i = 1, \dots, m \end{cases}$$

- for all i, the sublevel set $C_i = \{x \in \mathbb{R}^n \mid f_i(x) \leq 0\}$ is convex (follows from convexity of f_i)
- **feasible set** $X := \mathcal{D} \cap \bigcap_{i=1}^m C_i$ is **convex**
- A convex optimization problem minimizes a convex function over a convex set (take care: some convex sets may be nasty and intractable)

Global / local optimality for a convex optimization problem

Any locally optimal point of a **convex** problem is **globally** optimal.

Proof: Let $x^\star_{\text{loc}}$ be a local optimum. For an $R > 0$,

$$\forall x \text{ feasible, } \|x - x^\star_{\text{loc}}\| < R \Rightarrow f_0(x^\star_{\text{loc}}) \leq f_0(x)\,.$$

$x^\star_{\text{loc}}$ not global $\Rightarrow f_0(\overline{x}) < f_0(x^\star_{\text{loc}})$ for a feasible $\overline{x}$.
Let $z = (1-\theta)x^\star_{\text{loc}} + \theta\overline{x}$ with $\theta = \frac{R}{2\|\overline{x} - x^\star_{\text{loc}}\|} < 1$ and use convexity to get a contradiction:

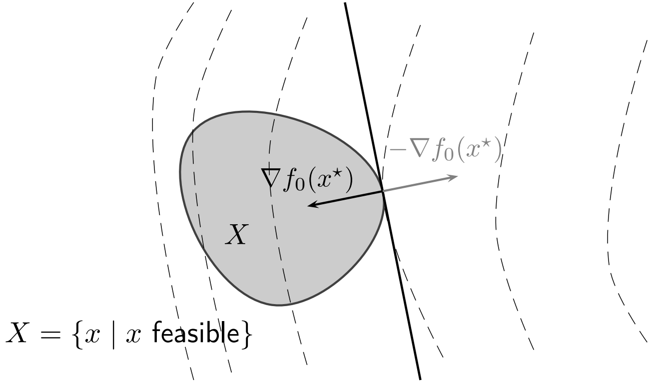
$$f_0(x^\star_{\text{loc}}) \leq f_0(z) \leq (1-\theta)f_0(x^\star_{\text{loc}}) + \theta f_0(\overline{x}) < f_0(x^\star_{\text{loc}})$$

Optimality criterion

For convex and differentiable f_0 ($\operatorname{dom} f_0$ open).

$x^\star$ is optimal if and only if:

- $x^\star$ feasible and: $\nabla f_0(x^\star)^\top (x - x^\star) \geq 0$ for all feasible x.



- Equivalent condition: $-\nabla f_0(x^\star) \in \mathcal{N}_X(x^\star)$ (normal cone)

Optimality criterion

(examples, see the exercises)

Particular cases, with differentiable f_0 ($\mathrm{dom}\, f_0$ open):

- unconstrained problem: $\min.\ f_0(x)$

$$x^\star \text{ optimal} \Leftrightarrow \quad \nabla f_0(x^\star) = 0, \quad x^\star \in \mathrm{dom}\, f_0$$

- equality constrained problem: $\begin{cases} \min.\ f_0(x) \\ \text{s.t. } Ax = b \end{cases}$

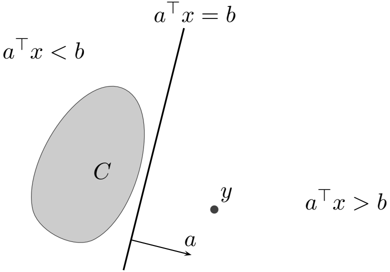
$$x^\star \text{ optimal} \Leftrightarrow \quad \nabla f_0(x^\star) + A^\top \nu^\star = 0, \quad Ax^\star = b, \quad x^\star \in \mathrm{dom}\, f_0$$

- minimization over nonnegative orthant: $\begin{cases} \min.\ f_0(x) \\ \text{s.t. } x \succeq 0 \end{cases}$

$$\begin{aligned} x^\star \text{ optimal} \Leftrightarrow \quad & x^\star \succeq 0, \quad \nabla f_0(x^\star) \succeq 0, \\ & x_i^\star [\nabla f_0(x^\star)]_i = 0, i = 1, \ldots, n \end{aligned}$$

Strict separation

Basic separation If C closed and convex and $y \notin C$, there exist $a \neq 0, a \in \mathbb{R}^n$ and $b \in \mathbb{R}$ such that $\forall x \in C, \quad a^\top x \leq b < a^\top y$.



Proof: Let $\overline{x}$ be a minimizer of $f(x) = \frac{\|x-y\|^2}{2}$ on C (which exists).
Optimality condition $-\nabla f(\overline{x}) \in \mathcal{N}_C(\overline{x})$, yields for all $x \in C$

$$(y - \overline{x})^\top (x - \overline{x}) \leq 0 \qquad \text{that is:}$$

$$\underbrace{(y - \overline{x})}_{=:a}{}^\top x \leq \underbrace{(y - \overline{x})^\top \overline{x}}_{=:b} < (y - \overline{x})^\top y\,.$$

Part III

Duality and optimality conditions

Lagrangian (inequality constraints only)

$$\begin{cases} \min.\ f_0(x) & x \in \mathcal{D} \subset \mathbb{R}^n \\ \quad \text{s.t. } f_i(x) \leq 0, & i = 1, \dots, m \end{cases}$$

with $\mathcal{D} := \cap_{i=1}^m \operatorname{dom} f_i$.

Lagrangian $\mathcal{L} : \mathcal{D} \times \mathbb{R}^m \to \mathbb{R}$

$$\mathcal{L}(x, \lambda) := f_0(x) + \sum_{i=1}^m \lambda_i f_i(x)$$

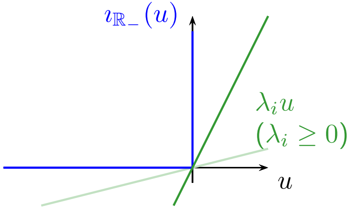
- λ_i are Lagrange multipliers, $\lambda = (\lambda_1, \dots, \lambda_m)^\top$.

Lagrangian: linear approximation interpretation

Equivalent unconstrained form:

$$\min.\ f(x) := f_0(x) + \sum_{i=1}^{m} \imath_{\mathbb{R}_-}(f_i(x))$$

Replace indicator functions by "soft" constraint/underestimator:



For $\lambda \succeq 0$:

$$\mathcal{L}(x,\lambda) := f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) \leq f(x)$$

Lagrange dual function

Dual function

$$\mathcal{L}_D(\lambda) := \inf_{x\in\mathcal{D}} \mathcal{L}(x,\lambda) = \inf_{x\in\mathcal{D}} \left(f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) \right)$$

- $\mathcal{L}_D$ is **concave** (even if non convex problem), can be $-\infty$
- **Lower bound property**: if $\lambda \succeq 0$, then $\mathcal{L}_D(\lambda) \leq p^\star$

Proof: for $\lambda \succeq 0$ and x feasible:

$$\mathcal{L}(x,\lambda) := f_0(x) + \sum_{i=1}^{m} \underbrace{\lambda_i}_{\geq 0} \underbrace{f_i(x)}_{\leq 0} \leq f_0(x)$$

Taking the infimum on the l.h.s yields $\mathcal{L}_D(\lambda) \leq f_0(x)$ for any feasible x and hence the result.

The dual problem

Lagrange dual problem

$$\begin{cases} \max. \ \mathcal{L}_D(\lambda) \\ \text{s.t. } \lambda \succeq 0 \end{cases}$$

$$d^\star := \sup_{\lambda \succeq 0} \mathcal{L}_D(\lambda)$$

- It is a convex problem
- λ dual feasible if $\lambda \succeq 0, \lambda \in \operatorname{dom} \mathcal{L}_D$

Weak duality: $d^\star \leq p^\star$ always holds (also for nonconvex problems)
$p^\star - d^\star$ is called **duality gap**.

Weak and strong duality

Weak duality (always holds): $d^\star \leq p^\star$

Strong duality: $d^\star = p^\star$

- does not hold in general
- holds for convex problems under **constraint qualifications** (see later).

Duality and max-min inequality

Primal with optimal value $p^\star$: $\begin{cases} \min.\ f_0(x) \quad (x \in \mathcal{D}) \\ \ \text{s.t. } f_i(x) \leq 0, \quad i = 1, \ldots, m \end{cases}$

- Lagrangian: $\mathcal{L}(x, \lambda) := f_0(x) + \sum_{i=1}^m \lambda_i f_i(x)$
- Primal reads also:

$$p^\star = \inf_{x \in \mathcal{D}} \sup_{\lambda \succeq 0} \mathcal{L}(x, \lambda)$$

- Dual problem:

$$d^\star = \sup_{\lambda \succeq 0} \inf_{x \in \mathcal{D}} \mathcal{L}(x, \lambda)$$

- We have (max-min inequality):

$$\sup_{\lambda \succeq 0} \inf_{x \in \mathcal{D}} \mathcal{L}(x, \lambda) \leq \inf_{x \in \mathcal{D}} \sup_{\lambda \succeq 0} \mathcal{L}(x, \lambda)$$

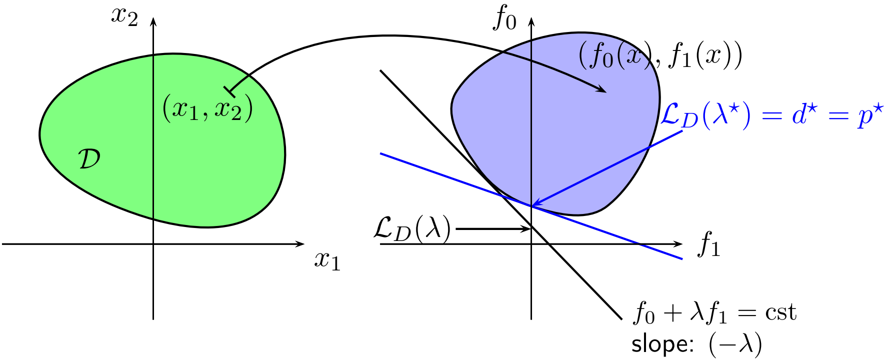
Strong duality when strong max-min/saddle-point property satisfied

Geometric interpretation of duality

Convex case

$$\begin{cases} \min\limits_{x\in\mathcal{D}} f_0(x) \\ \text{ s.t. } f_1(x) \le 0 \end{cases}$$

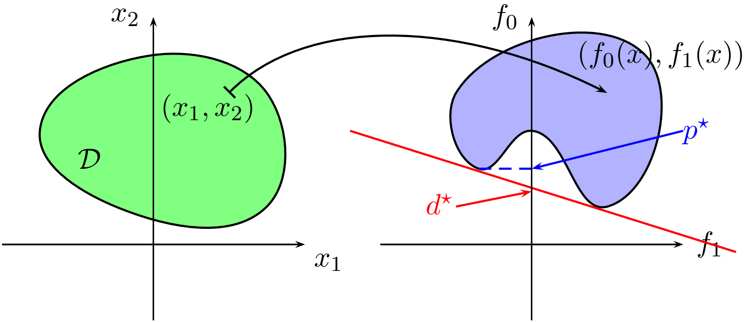
$$\mathcal{L}_D(\lambda) = \inf_{x\in\mathcal{D}} f_0(x) + \lambda f_1(x)$$



Geometric interpretation of duality

Non-convex case

$$\begin{cases} \min_{x \in \mathcal{D}} f_0(x) \\ \text{ s.t. } f_1(x) \leq 0 \end{cases} \qquad \mathcal{L}_D(\lambda) = \inf_{x \in \mathcal{D}} f_0(x) + \lambda f_1(x)$$



Lagrangian (inequality constraints only)

$$\begin{cases} \min.\ f_0(x) & x \in \mathcal{D} \subset \mathbb{R}^n \\ \quad \text{s.t. } f_i(x) \leq 0, & i = 1, \ldots, m \end{cases}$$

with $\mathcal{D} := \cap_{i=1}^m \operatorname{dom} f_i$.

Lagrangian $\mathcal{L} : \mathcal{D} \times \mathbb{R}^m \to \mathbb{R}$

$$\mathcal{L}(x, \lambda) := f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x)$$

- λ_i are Lagrange multipliers, $\lambda = (\lambda_1, \ldots, \lambda_m)^\top$.

Lagrangian sufficient conditions

Assume $(x^\star, \lambda^\star) \in \mathcal{D} \times \mathbb{R}^m$ satisfies:

$$\forall i = 1, \ldots, m, \qquad f_i(x^\star) \leq 0 \qquad \text{(primal feasability)}$$

$$\forall i = 1, \ldots, m, \qquad \lambda_i^\star \geq 0 \qquad \text{(dual feasability)}$$

$$\forall i = 1, \ldots, m, \qquad \lambda_i^\star f_i(x^\star) = 0 \qquad \text{(complementary slackness)}$$

$$\forall x \text{ feasible}, \quad \mathcal{L}(x^\star, \lambda^\star) \leq \mathcal{L}(x, \lambda^\star) \qquad (x^\star \text{ minimizes } \mathcal{L}(., \lambda^\star))$$

then, $x^\star$ is optimal (global minimum).

Proof: For any feasible x:
$f_0(x^\star) = \mathcal{L}(x^\star, \lambda^\star) \leq \mathcal{L}(x, \lambda^\star) = f_0(x) + \sum_{i=1}^m \underbrace{\lambda_i^\star f_i(x)}_{\leq 0} \leq f_0(x)$

- $\lambda^\star$: Lagrange multiplier vector
- Remark: no convexity!

KKT conditions (Karush-Kuhn-Tucker)

Convex case: sufficient conditions

Assume $(x^\star, \lambda^\star) \in \operatorname{int} \mathcal{D} \times \mathbb{R}^m$ satisfies:

$\forall i = 1, \dots, m, \quad f_i(x^\star) \leq 0$ (primal feasability)

$\forall i = 1, \dots, m, \quad \lambda_i^\star \geq 0$ (dual feasabiilty)

$\forall i = 1, \dots, m, \quad \lambda_i^\star f_i(x^\star) = 0$ (complementary slackness)

$\nabla f_0(x^\star) + \sum_{i=1}^m \lambda_i^\star \nabla f_i(x^\star) = 0$ ($x^\star$ critical point of the Lagrangian)

then, if the problem is **convex**, $x^\star$ is optimal.

- $\lambda^\star$: Lagrange multiplier vector
- Remark: for convex functions $f_0, f_1, \dots, f_m$, last condition implies $\mathcal{L}(x^\star, \lambda^\star) \leq \mathcal{L}(x, \lambda^\star)$

Necessary optimality conditions (Fritz-John)

$$\begin{cases} \min.\ f_0(x) & x \in \mathcal{D} \subset \mathbb{R}^n \\ \ \text{s.t. } f_i(x) \leq 0, & i = 1, \dots, m \end{cases}$$

- **Active set** at point x: $I(x) = \{i \in \{1, \dots, m\} \mid f_i(x) = 0\}$
- Fritz-John optimality conditions:
  If $x^\star_{\text{loc}} \in \operatorname{int} \mathcal{D}$ is a **local minimizer**, there exist $\lambda_0, \lambda_1, \dots, \lambda_m \geq 0$ such that:

$$\lambda_0 \nabla f_0(x^\star_{\text{loc}}) + \sum_{i \in I(x^\star_{\text{loc}})} \lambda_i \nabla f_i(x^\star_{\text{loc}}) = 0$$

- For $i \notin I(x^\star_{\text{loc}})$, complementary slackness yields $\lambda_i = 0 \rightsquigarrow$ terms don't appear above.
- To rule out the case $\lambda_0 = 0$, **constraint qualification** at $x^\star_{\text{loc}}$ (required for KKT to be necessary conditions)

Local constraint qualifications

Constraint qualifications **at a point** x:

- MFCQ (Mangasarian-Fromovitz constraint qualification):
  there is a direction d satisfying $\nabla f_i(x)^\top d < 0$ for all $i \in I(x)$
- LICQ (linear independence constraint qualification):
  $\{\nabla f_i(x)\}_{i \in I(x)}$ are linearly independent

Obviously: LICQ $\Rightarrow$ MFCQ

Global constraint qualification (Slater)

- **Slater** constraint qualification for convex problem with constraints $f_i(x) \leq 0, \quad i = 1, \ldots, m$

➤ there exists $\hat{x} \in \operatorname{relint} \mathcal{D}$ with $f_i(\hat{x}) < 0, \quad i = 1, \ldots, m$

- Refinement: **affine inequalities need not be strict**. For constraints $\begin{cases} f_i(x) \leq 0, \quad i = 1, \ldots, m \\ Ax \leq b, \quad Cx = d \end{cases}$

➤ there exists $\hat{x} \in \operatorname{relint} \mathcal{D}$ with $f_i(\hat{x}) < 0, \quad i = 1, \ldots, m$ and $Ax \leq b, Cx = d$

+ For a convex problem: Slater $\Rightarrow$ MFCQ at any feasible point.
+ Slater $\approx$ there exist a strictly feasible point
+ Slater $\Rightarrow$ strong duality and dual value attained when $d^\star > -\infty$

KKT necessary optimality conditions

Suppose $x^\star_{\text{loc}}$ is a **local minimizer** of

$$\inf\{f_0(x) \mid x \in \mathcal{D}, f_i(x) \leq 0, i = 1, \dots, m\}$$

If MFCQ holds at $x^\star_{\text{loc}}$, there is a Lagrange multiplier vector $\lambda^\star$ for $x^\star_{\text{loc}}$:

$$\forall i = 1, \dots, m, \qquad f_i(x^\star_{\text{loc}}) \leq 0 \qquad \text{(primal feasibility)}$$

$$\forall i = 1, \dots, m, \qquad \lambda^\star_i \geq 0 \qquad \text{(dual feasibility)}$$

$$\forall i = 1, \dots, m, \qquad \lambda^\star_i f_i(x^\star_{\text{loc}}) = 0 \qquad \text{(complementary slackness)}$$

$$\nabla f_0(x^\star_{\text{loc}}) + \sum_{i=1}^m \lambda^\star_i \nabla f_i(x^\star_{\text{loc}}) = 0 \qquad \text{(}x^\star_{\text{loc}}\text{ critical point of the Lagrangian)}$$

Remarks:

- No convexity here, but local minimizer considered.
- For convex problems, above conditions are necessary and sufficient for global optimality.

Lagrangian (inequality constraints only)

$$\begin{cases} \text{min. } f_0(x) & x \in \mathcal{D} \subset \mathbb{R}^n \\ \quad \text{s.t. } f_i(x) \leq 0, & i = 1, \ldots, m \end{cases}$$

with $\mathcal{D} := \cap_{i=1}^m \operatorname{dom} f_i$.

Lagrangian $\mathcal{L} : \mathcal{D} \times \mathbb{R}^m \to \mathbb{R}$

$$\mathcal{L}(x, \lambda) := f_0(x) + \sum_{i=1}^m \lambda_i f_i(x)$$

- λ_i are Lagrange multipliers, $\lambda = (\lambda_1, \ldots, \lambda_m)^\top$.

Necessary optimality conditions (through strong duality)

If **strong duality** holds, $x^\star$,$\lambda^\star$ are primal, dual optimal. Then:

- $x^\star$ minimizes $x \mapsto \mathcal{L}(x, \lambda^\star)$

$\rightsquigarrow$ $\nabla_x \mathcal{L}(x, \lambda^\star)|_{x^\star} = 0$ (see next slide)

- $\lambda_i^\star f_i(x^\star) = 0, \ i = 1, \dots, m$ (complementary slackness)

$$\lambda_i^\star > 0 \Rightarrow f_i(x^\star) = 0 \qquad f_i(x^\star) < 0 \Rightarrow \lambda_i^\star = 0$$

Proof: (write all inequalities, which become equalities)

$$d^\star = \mathcal{L}_D(\lambda^\star) = \inf_{x \in \mathcal{D}} \mathcal{L}(x, \lambda^\star) \leq \mathcal{L}(x^\star, \lambda^\star) \leq f_0(x^\star) = p^\star$$

where $\mathcal{L}(x^\star, \lambda^\star) = f_0(x^\star) + \sum_{i=1}^m \lambda_i^\star f_i(x^\star)$

Remark: no convexity assumption

Necessary KKT conditions (through strong duality)

If **strong duality** holds, $x^\star$,$\lambda^\star$ are primal, dual optimal, then the following conditions (called KKT) hold:

1. Primal constraints: $f_i(x^\star) \leq 0,$ for $i = 1, \ldots, m$
2. Dual constraints: $\lambda_i^\star \geq 0,$ for $i = 1, \ldots, m$
3. Complementary slackness: $\lambda_i^\star f_i(x^\star) = 0$ for $i = 1, \ldots, m$
4. Gradient of Lagrangian w.r.t. x vanishes at $x^\star$:

$$\nabla f_0(x^\star) + \sum_{i=1}^{m} \lambda_i^\star \nabla f_i(x^\star)$$

Remark: no convexity assumption

KKT sufficient conditions for convex problem

If $\overline{x}, \overline{\lambda}$ satisfy KKT for a convex problem, then they are primal/dual optimal.

1. Primal constraints: $f_i(\overline{x}) \leq 0$, for $i = 1, \ldots, m$
2. Dual constraints: $\overline{\lambda}_i \geq 0$, for $i = 1, \ldots, m$
3. Complementary slackness: $\overline{\lambda}_i f_i(\overline{x}) = 0$ for $i = 1, \ldots, m$
4. $\nabla_x \mathcal{L}(\overline{x}, \overline{\lambda}, \overline{\nu}) = \nabla f_0(\overline{x}) + \sum_{i=1}^m \overline{\lambda}_i \nabla f_i(\overline{x}) = 0$

Indeed:

$$\begin{aligned} f_0(\overline{x}) &= \mathcal{L}(\overline{x}, \overline{\lambda}) \quad \text{from compl. slackness and primal feas.} \\ &= \mathcal{L}_D(\overline{\lambda}) \quad \text{from vanishing of } \nabla_x \mathcal{L}(\overline{x}, \overline{\lambda}) \text{ and convexity.} \end{aligned}$$

KKT necessary and sufficient conditions for convex problem

For a convex problem, if **Slater's condition** is satisfied:

- Strong duality holds,
- Dual optimal value is attained when $d^\star > -\infty$ (*i.e.* there exists $\lambda^\star$ such that $\mathcal{L}_D(\lambda^\star) = d^\star = p^\star$),
- KKT conditions are sufficient and necessary for global optimality.

Remark: This generalizes $\nabla f_0(x^\star) = 0$ for unconstrained problem.

Perturbation and sensitivity analysis (1/2)

- Unperturbed optimization problem and dual

$$p^\star : \begin{cases} \min.\ f_0(x) \\ \quad \text{s.t. } f_i(x) \leq 0, \quad 1 \leq i \leq m \end{cases} \qquad \begin{cases} \max.\ \mathcal{L}_D(\lambda) \\ \quad \text{s.t. } \lambda \succeq 0 \end{cases}$$

- Perturbed problem and dual

$$p^\star(u) : \begin{cases} \min.\ f_0(x) \\ \quad \text{s.t. } f_i(x) \leq u_i, \quad 1 \leq i \leq m \end{cases} \qquad \begin{cases} \max.\ \mathcal{L}_D(\lambda) - u^\top \lambda \\ \quad \text{s.t. } \lambda \succeq 0 \end{cases}$$

Optimal value $p^\star(u)$ as a function of parameters u
(for the original problem $p^\star = p^\star(0)$)

Perturbation and sensitivity analysis (2/2)

Assume for problem, strong duality and $\lambda^\star$ dual optimal.

- Global sensitivity:

$$\begin{aligned} p^\star(u) &\geq \mathcal{L}_D(\lambda^\star) - u^\top \lambda^\star \text{ (weak duality pert. prob.)} \\ &\geq p^\star(0) - u^\top \lambda^\star \text{ (strong duality)} \end{aligned}$$

- Local sensitivity: if $p^\star(u)$ differentiable at 0:

$$\lambda_i^\star = -\frac{\partial p^\star(0)}{\partial u_i}$$

Proof: take $u = te_i$ where e_i is i^{th} canonical basis vector and get $\frac{p^\star(te_i) - p^\star(0)}{t} \geq -\lambda_i^\star$ for $t > 0$ or $\leq -\lambda_i^\star$ for $t < 0$.

- Interpretation: ...

Lagrangian and dual function

$$\begin{cases} \text{min. } f_0(x) \quad (x \in \mathcal{D} \subset \mathbb{R}^n) \\ \quad \text{s.t. } f_i(x) \le 0, \quad i = 1, \dots, m \\ \qquad\quad h_j(x) = 0, \quad j = 1, \dots, p \end{cases}$$

Lagrangian $\mathcal{L} : \mathcal{D} \times \mathbb{R}^m \times \mathbb{R}^p \to \mathbb{R}$
(λ_i, ν_j are Lagrange multipliers)

$$\mathcal{L}(x, \lambda, \nu) := f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) + \sum_{j=1}^{p} \nu_j h_j(x)$$

Dual function $\mathcal{L}_D(\lambda, \nu) := \inf_{x \in \mathcal{D}} \mathcal{L}(x, \lambda, \nu)$

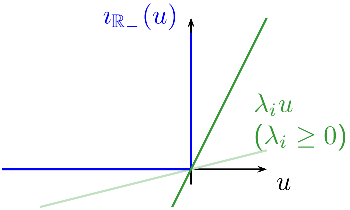
- $\mathcal{L}_D$ is **concave** (even if non convex problem), can be $-\infty$
- **Lower bound property**: if $\lambda \succeq 0$, $\nu \in \mathbb{R}^p$, then $\mathcal{L}_D(\lambda, \nu) \le p^\star$

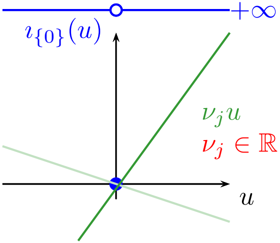
Lagrangian: linear approximation interpretation

Equivalent unconstrained form:

$$\min.\ f(x) := f_0(x) + \sum_{i=1}^{m} \imath_{\mathbb{R}_-}(f_i(x)) + \sum_{j=1}^{p} \imath_{\{0\}}(h_j(x))$$

Replace indicator functions by "soft" constraint/underestimator:





For $\lambda \succeq 0$ and $\nu \in \mathbb{R}^p$, $\mathcal{L}(x, \lambda, \nu) \leq f(x)$.

Lagrange dual function

Lagrange dual problem

$$d^\star := \sup_{\lambda \succeq 0, \nu \in \mathbb{R}^p} \mathcal{L}_D(\lambda, \nu) = \begin{cases} \max. \ \mathcal{L}_D(\lambda, \nu) \\ \quad \text{s.t. } \lambda \succeq 0 \end{cases}$$

- It is a convex problem.
- λ, ν are dual feasible if $\lambda \succeq 0, \nu \in \mathbb{R}^p, (\lambda, \nu) \in \operatorname{dom} \mathcal{L}_D$

Weak duality (always holds): $d^\star \leq p^\star$

$p^\star - d^\star$ is called **duality gap**.

Strong duality: $d^\star = p^\star$

- does not hold in general.
- holds for convex problems under **constraint qualifications**.

Duality and max-min inequality

Primal with optimal value $p^\star$: $\begin{cases} \min.\ f_0(x) \quad (x \in \mathcal{D}) \\ \text{s.t. } f_i(x) \leq 0, \quad i = 1, \dots, m \\ \qquad h_j(x) = 0, \quad j = 1, \dots, p \end{cases}$

- Lagrangian: $\mathcal{L}(x, \lambda, \nu) := f_0(x) + \sum_{i=1}^m \lambda_i f_i(x) + \sum_{j=1}^p \nu_j h_j(x)$
- Primal reads also:

$$p^\star = \inf_{x \in \mathcal{D}} \sup_{\nu \in \mathbb{R}^p, \lambda \succeq 0} \mathcal{L}(x, \lambda, \nu)$$

- Dual problem:

$$d^\star = \sup_{\nu \in \mathbb{R}^p, \lambda \succeq 0} \inf_{x \in \mathcal{D}} \mathcal{L}(x, \lambda, \nu)$$

- We have (max-min inequality):

$$d^\star = \sup_{\nu \in \mathbb{R}^p, \lambda \succeq 0} \inf_{x \in \mathcal{D}} \mathcal{L}(x, \lambda, \nu) \leq \inf_{x \in \mathcal{D}} \sup_{\nu \in \mathbb{R}^p, \lambda \succeq 0} \mathcal{L}(x, \lambda, \nu) = p^\star$$

Strong duality when strong max-min/saddle-point property satisfied.

KKT optimality conditions

$$
\begin{aligned}
&f_i(x^\star) \leq 0, \quad i = 1, \dots, m && \text{(primal feasability)}\\
&h_j(x^\star) = 0, \quad j = 1, \dots, p && \\
&\lambda_i^\star \geq 0, \quad i = 1, \dots, m && \text{(dual feasability)}\\
&(\nu_j^\star \in \mathbb{R}, \quad j = 1, \dots, p) && \\
&\lambda_i^\star f_i(x^\star) = 0, \quad i = 1, \dots, m && \text{(complementary slackness)}\\
&\mathcal{L}(x^\star, \lambda^\star, \nu^\star) \leq \mathcal{L}(x, \lambda^\star, \nu^\star), \quad \forall x \text{ feasible} && (x^\star \text{ minimizes } \mathcal{L}(., \lambda^\star, \nu^\star))
\end{aligned}
$$

KKT optimality conditions

$$f_i(x^\star) \leq 0, \quad i = 1, \dots, m \qquad \text{(primal feasability)}$$

$$h_j(x^\star) = 0, \quad j = 1, \dots, p$$

$$\lambda_i^\star \geq 0, \quad i = 1, \dots, m \qquad \text{(dual feasability)}$$

$$(\nu_j^\star \in \mathbb{R}, \quad j = 1, \dots, p)$$

$$\lambda_i^\star f_i(x^\star) = 0, \quad i = 1, \dots, m \qquad \text{(complementary slackness)}$$

$$\nabla f_0(x^\star) + \sum_{i=1}^{m} \lambda_i^\star \nabla f_i(x^\star) + \sum_{j=1}^{p} \nu_j^\star \nabla h_j(x^\star) = 0$$

($x^\star$ critical point of the Lagrangian)

- Remark: for convex problem, last condition implies $\mathcal{L}(x^\star, \lambda^\star, \nu^\star) \leq \mathcal{L}(x, \lambda^\star, \nu^\star)$ for feasible x.

Least-norm solution of linear equation (example)

Lagrange dual

$$\min.\ \|x\|_2^2 \quad \text{s.t. } Ax = b$$

- Lagrangian:

$$\mathcal{L}(x,\nu) = x^\top x + \nu^\top (Ax - b)$$

- Dual function: (minimum of $\mathcal{L}$ w.r.t. x when $\nabla_x \mathcal{L}(x,\nu) = 0$)

$$\begin{aligned}\mathcal{L}_D(\nu) &= \mathcal{L}(-\frac{1}{2}A^\top \nu, \nu) \\ &= -\frac{1}{4}\nu^\top A A^\top \nu - b^\top \nu \leq \inf\{\|x\|_2^2 \mid Ax = b\}\end{aligned}$$

- Primal and dual problems:

$$p^\star : \begin{cases}\min.\ x^\top x \\ \quad \text{s.t. } Ax = b\end{cases} \qquad d^\star : \max.\ -\frac{1}{4}\nu^\top A A^\top \nu - b^\top \nu$$

Least-norm solution of linear equation (example)

KKT conditions and solution

$$\min. \ \|x\|_2^2 \quad \text{s.t. } Ax = b$$

- Lagrangian: $\mathcal{L}(x, \nu) = x^\top x + \nu^\top (Ax - b)$
- Dual function: $\mathcal{L}_D(\nu) = -\frac{1}{4}\nu^\top AA^\top \nu - b^\top \nu$
- KKT conditions:

$$\begin{cases} Ax^\star = b \\ 2x^\star + A^\top \nu^\star = 0 \end{cases}$$

- Solution (when $AA^\top$ invertible):

$$\begin{cases} x^\star = A^\top (AA^\top)^{-1} b \\ \nu^\star = -2(AA^\top)^{-1} b \end{cases}$$

LP (standard form) (example)

Lagrange dual

$$\min.\ c^\top x \quad \text{s.t. } Ax = b, x \succeq 0$$

- Lagrangian:

$$\begin{aligned}\mathcal{L}(x,\lambda,\nu) &= c^\top x - \lambda^\top x + \nu^\top (Ax - b)\\ &= -b^\top \nu + (c + A^\top \nu - \lambda)^\top x\end{aligned}$$

- Dual function:

$$\mathcal{L}_D(\lambda,\nu) = \begin{cases} -b^\top \nu & \text{if } A^\top \nu - \lambda + c = 0 \\ -\infty & \text{otherwise.}\end{cases}$$

- Primal and dual problems:

$$p^\star : \begin{cases} \min.\ c^\top x \\ \quad \text{s.t. } Ax = b \\ \qquad\quad x \succeq 0 \end{cases} \qquad d^\star : \begin{cases} \max.\ -b^\top \nu \\ \quad \text{s.t. } A^\top \nu + c \succeq 0 \end{cases}$$

LP (standard form) (example)

KKT conditions

$$\min.\ c^\top x \quad \text{s.t. } Ax = b, x \succeq 0$$

- Lagrangian:

$$\begin{aligned}\mathcal{L}(x, \lambda, \nu) &= c^\top x - \lambda^\top x + \nu^\top (Ax - b)\\ &= -b^\top \nu + (c + A^\top \nu - \lambda)^\top x\end{aligned}$$

- KKT conditions:

$$\begin{cases} Ax^\star = b, \quad x^\star \succeq 0\\ \lambda^\star \succeq 0\\ \lambda_i^\star x_i^\star = 0, \quad i = 1, \dots, n\\ A^\top \nu^\star + c - \lambda^\star = 0\end{cases}$$

Equality constr. convex quad. minimization (example)

KKT conditions

$$\begin{cases} \min. \dfrac{1}{2}x^\top P x + q^\top x + r \\ \quad \text{s.t. } Ax = b \end{cases} \text{ with } P \in \mathbb{S}_+^n.$$

- Lagrangian: $\mathcal{L}(x,\nu) = \frac{1}{2}x^\top P x + q^\top x + r + \nu^\top (Ax - b)$
- KKT conditions:

$$Ax^\star = b, \quad Px^\star + q + A^\top \nu^\star = 0$$

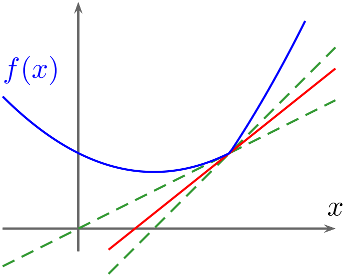
can be written as:

$$\begin{bmatrix} P & A^\top \\ A & 0 \end{bmatrix} \begin{bmatrix} x^\star \\ \nu^\star \end{bmatrix} = \begin{bmatrix} -q \\ b \end{bmatrix}$$

Subgradient

A **subgradient** of f at $\overline{x}$ is any vector ϕ such that:

$$f(\overline{x}) + \phi^\top (x - \overline{x}) \leq f(x) \text{ for all } x$$



- $x \mapsto f(\overline{x}) + \phi^\top (x - \overline{x})$ is a *linear underestimator* of f.
- if f convex and differentiable, $\nabla f(\overline{x})$ is (unique) subgradient.

Subdifferential

Definition

The **subdifferential** of f at $\overline{x}$ is the set of all subgradients, denoted by:

$$\partial f(\overline{x}) = \{\phi \in \mathbb{R}^n \,|\, f(\overline{x}) + \phi^\top (x - \overline{x}) \leq f(x) \text{ for all } x\}$$

- $\partial f(\overline{x})$ is a closed convex set (always).
- ∂f is a multi-function / set-valued map. $\partial f : \mathbb{R}^n \to 2^{\mathbb{R}^n}$
- $\mathrm{dom}\, \partial f = \{x \in \mathbb{R}^n \,|\, \partial f(x) \neq \emptyset\}$.
- If f convex and differentiable at $\overline{x} \in \mathrm{int}\,\mathrm{dom}\, f$, then $\partial f(\overline{x}) = \{\nabla f(\overline{x})\}$.

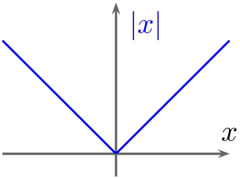
Subdifferential

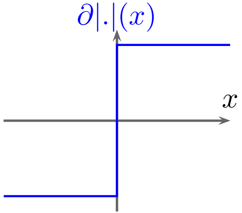
Examples

The **subdifferential** of f at $\overline{x}$ is the set of all subgradients:

$$\partial f(\overline{x}) = \{\phi \in \mathbb{R}^n \mid f(\overline{x}) + \phi^\top(x - \overline{x}) \leq f(x) \text{ for all } x\}$$

- If f convex and differentiable at $\overline{x} \in \operatorname{int}\operatorname{dom} f$, then $\partial f(\overline{x}) = \{\nabla f(\overline{x})\}$.
- Absolute value: $\partial|.|(x) = \begin{cases} \{-1\} & x<0, \\ \{1\} & x>0, \\ [-1,1] & x=0 \end{cases}$





- Indicator function: $\partial\iota_C(\overline{x}) = \mathcal{N}_C(\overline{x})$ (normal cone operator)

Fermat's rule

Characterization of global minimizer:

$$x^\star \textbf{ global minimizer } \text{of } f \quad \Leftrightarrow \quad 0 \in \partial f(x^\star)$$

Proof: Use definition of subdifferential:

$$0 \in \partial f(x^\star) \Leftrightarrow f(x^\star) + \langle 0, x - x^\star \rangle \leq f(x) \text{ for all } x$$

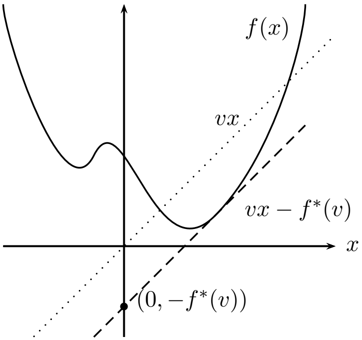
Remark:

- holds also for nonconvex f.

The Fenchel conjugate function (definition)

Let $f : \mathbb{R}^n \to \mathbb{R}$ (with $f(x) = \infty$ for $x \notin \operatorname{dom} f$).
Fenchel conjugate:

$$f^*(v) = \sup_{x \in \mathbb{R}^n} (v^\top x - f(x))$$



The Fenchel conjugate function (first properties)

$$f^*(v) = \sup_{x \in \mathbb{R}^n} (v^\top x - f(x))$$

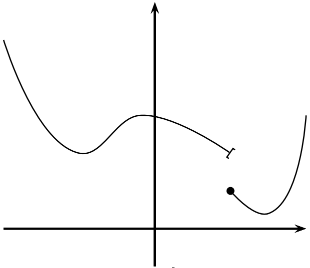
- f^* is convex
  (because $\sup$ of affine functions. True for non convex f also.)
- $f \geq g$ implies $f^* \leq g^*$
- If $\operatorname{dom} f \neq \emptyset$, f^* never takes the value $-\infty$
- f^* is l.s.c. (lower semi-continuous)
  (because epigraph closed)

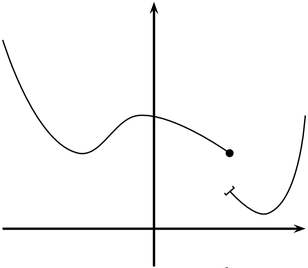
Lower semi-continuous function (l.s.c.)

f is l.s.c. if and only if at any point x:

$$x_n \xrightarrow[n\to\infty]{} x \quad \Rightarrow f(x) \leq \lim_{n\to\infty} f(x_n)$$

f is l.s.c. $\Leftrightarrow$ epigraph $\{(x,t) \in \mathbb{R}^n \times \mathbb{R} \,|\, f(x) \leq t\}$ is a closed set


l.s.c.


non l.s.c.

Fenchel conjugate function

Examples

- $f(x) = ax + b$ $$f^*(v) = \begin{cases} -b & \text{if } v = a, \\ +\infty & \text{otherwise.} \end{cases}$$
- $f(x) = e^x$ $$f^*(v) = \begin{cases} v \log v - v & \text{if } v \geq 0, \\ +\infty & \text{otherwise.} \end{cases}$$
- $f(x) = -\log x$ $$f^*(v) = \begin{cases} -\log(-v) - 1 & \text{if } v < 0, \\ +\infty & \text{otherwise.} \end{cases}$$

Fenchel conjugate function

Examples (continued)

- $f(x) = \frac{1}{2}x^\top Qx$ with $Q \succ 0$ $\qquad f^*(v) = \frac{1}{2}v^\top Q^{-1}v$.
- $f(x) = \frac{1}{2}\|x\|_2^2$ $\qquad f^*(v) = \frac{1}{2}\|v\|_2^2$.
- $f(x) = \|x\|$ $\qquad f^*(v) = \begin{cases} 0 & \text{if } \|v\|_* \leq 1, \\ +\infty & \text{otherwise.} \end{cases}$

Dual norm

Let $\|.\|$ be a norm on $\mathbf{E}$.
Associated **dual norm** $\|.\|_*$:

$$\|z\|_* := \sup_{\|x\| \leq 1} \langle z, x \rangle$$

- $\langle z, x \rangle \leq \|z\|_* \|x\|$
- Dual norm of $\|.\|_2$ is itself.
- $\|.\|_\infty$ and $\|.\|_1$ are dual norms of each other.
- Dual of ℓ_p-norm is ℓ_q norm with $\frac{1}{p} + \frac{1}{q} = 1$.
- $\|.\|_{**} = \|.\|$ (need not hold in infinite dimensional spaces)

Fenchel-Young inequality

- For any x and v in $\mathbb{R}^n$: $\quad f(x) + f^*(v) \geq v^\top x$
- Equality case:

$$f(x) + f^*(v) = v^\top x \Leftrightarrow v \in \partial f(x)$$

- For f **convex, l.s.c., proper**, equality case:

$$\begin{aligned} f(x) + f^*(v) = v^\top x &\Leftrightarrow v \in \partial f(x) \\ &\Leftrightarrow x \in \partial f^*(v) \end{aligned}$$

Fenchel biconjugate

- The biconjugate $f^{**} = (f^*)^*$ is convex l.s.c.
  (from properties of $f^*(v) = \sup_{x \in \mathbb{R}^n} (v^\top x - f(x))$)
- f^{**} is a minorant of f
  (follows from Fenchel-Young inequality $f(x) \geq v^\top x - f^*(v)$)

Theorem

For any function $f : \mathbb{R}^n \to]-\infty, +\infty]$:

$$\begin{aligned} f = f^{**} &\Leftrightarrow f \text{ is closed (l.s.c.) and convex} \\ &\Leftrightarrow \text{For all points in } \mathbb{R}^n, \\ &\qquad f(x) = \sup\{\alpha(x) \mid \alpha \text{ an affine minorant of } f\} \end{aligned}$$

For proper closed convex functions, the conjugacy operation induces a bijection.

Fenchel duality

Let $f : \mathbb{R}^n \to]-\infty, +\infty]$ and $g : \mathbb{R}^m \to]-\infty, +\infty]$ be given function and $A \in \mathbb{R}^{m \times n}$.

$$p^\star := \inf_{x \in \mathbb{R}^n} \{f(x) + g(Ax)\} \qquad \textbf{(primal value)}$$

$$d^\star := \sup_{v \in \mathbb{R}^m} \{-f^*(A^\top v) - g^*(-v)\} \qquad \textbf{(dual value)}$$

We have:

- Weak duality: $d^\star \leq p^\star$ (proof: Fenchel-Young inequality)
- Strong duality: if f and g are **convex**, under qualification constraints[1]: $p^\star = d^\star$
  and the supremum in the dual problem is attained if finite.

[1] $0 \in \mathrm{core}(\mathrm{dom}\, g - A\, \mathrm{dom}\, f)$ or stronger condition $A\, \mathrm{dom}\, f \cap \mathrm{cont}\, g \neq \emptyset$

Fenchel and Lagrange duality

- Primal problem (as in Fenchel: previous slide): $\min_{x \in \mathbb{R}^n} . \ f(x) + g(Ax)$
- Equivalent constrained problem:

$$\min_{(x,y) \in \mathbb{R}^n \times \mathbb{R}^m} . \ f(x) + g(y) \quad \text{s.t. } y = Ax$$

- Lagrangian and dual function:

$$\begin{aligned} L(x, y, \nu) &= f(x) + g(y) + \nu^\top (y - Ax) \\ \inf_{x,y} L(x, y, \nu) &= - \sup_{x \in \mathbb{R}^n} \{x^\top A^\top \nu - f(x)\} - \sup_{y \in \mathbb{R}^m} \{(-\nu)^\top y - g(y)\} \\ &= -f^*(A^\top \nu) - g^*(-\nu) \end{aligned}$$

- Dual problem: $\max_{\nu \in \mathbb{R}^m} . \ - f^*(A^\top \nu) - g^*(-\nu)$ is exactly Fenchel dual! (see previous slide)

Part IV

Algorithms

Unconstrained minimization

With $f : \mathbb{R}^n \to \mathbb{R}$ convex, twice differentiable, find solution to:

$$p^\star : \qquad \min_x . \; f(x)$$

- Optimality condition: $\nabla f(x^\star) = 0$
- Produce a **sequence of points** $x^{(k)} \in \operatorname{dom} f$ such that:

$$f(x^{(k)}) \to p^\star$$

- Starting point $x^{(0)}$ required, such that:

  $x^{(0)} \in \operatorname{dom} f$
  sublevel set $\{x \mid f(x) \leq f(x^{(0)})\}$ is closed

Descent methods

Starting from $x^{(0)}$ repeat for $k = 0, 1, 2, \dots$:

$$x^{(k+1)} = x^{(k)} + t\Delta x^{(k)} \quad \text{with } f(x^{(k+1)}) < f(x^{(k)})$$

- $t \geq 0$ is the **step size** or step length
- $\Delta x^{(k)}$ is the **search direction** or step and must satisfy:

$$\nabla f(x^{(k)})^\top \Delta x^{(k)} < 0$$

(because $f(x^{(k)}) + \nabla f(x^{(k)})^\top (t\Delta x^{(k)}) \leq f(x^{(k+1)})$ from convexity)

- Simplified notation:

current point: x, search direction: Δx

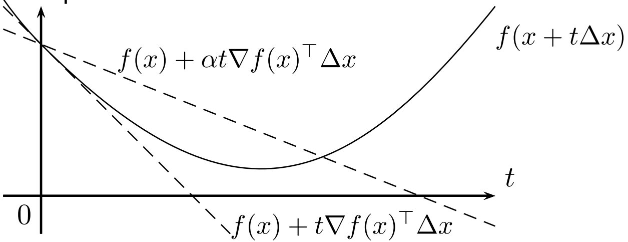
next point: $x^+ = x + t\Delta x$ with: $f(x^+) < f(x)$

Step size and line search

- **Constant step size** $t > 0$ chosen and fixed.
- **Exact line search** $t = \mathrm{argmin}_{t \geq 0} f(x + t\Delta x)$
- **Backtracking** (with parameters $\alpha \in]0, 1/2[, \beta \in]0, 1[$)
  starting at $t = 1$, repeat $t := \beta t$ until:

$$f(x + t\Delta x) < f(x) + \alpha t \nabla f(x)^\top \Delta x$$

(also known as Armijo's rule)
graphical interpretation:



Unconstrained descent method

given starting point $x^{(0)} \in \operatorname{dom} f$, tolerance $\epsilon > 0$,
repeat:

1. Compute search direction $\Delta x^{(k)}$
2. Stopping criterion: **quit** if it is smaller than ϵ.
3. Choose step size t (backtracking, line search, constant, ...)
4. Update: $x^{(k+1)} = x^{(k)} + t\Delta x^{(k)}$

Possible search directions for a descent method:

- gradient: $\Delta x_{\text{grad}}^{(k)} = -\nabla f(x^{(k)})$
- (normalized) steepest descent: $\Delta x_{\text{nsd}}^{(k)} = \operatorname{argmin}_v \{\nabla f(x^{(k)})^\top v \mid \|v\| \leq 1\}$
- Newton: $\Delta x_{\text{nt}}^{(k)} = -\nabla^2 f(x^{(k)})^{-1} \nabla f(x^{(k)})$

Gradient descent

Gradient descent direction (at point x):

$$\Delta x_{\mathrm{grad}} = -\nabla f(x)$$

Stopping condition: usually $\|\nabla f(x^{(k)})\|_2 < \epsilon$.

Strongly convex function

f is **strongly convex** iff $f - \frac{m}{2}\|x\|_2^2$ is convex for an $m > 0$.
For twice continuously differentiable f, equivalent to $\nabla^2 f(x) \succeq m\mathbf{Id}$

Implications:

- $f(y) \geq f(x) + \nabla f(x)^\top (y-x) + \frac{m}{2}\|y-x\|_2^2$ (convexity)
- $p^\star \geq f(x) - \frac{1}{2m}\|\nabla f(x)\|_2^2$ (minimize above r.h.s. w.r.t. y)
- Sublevel sets are bounded (because of the first inequality above). On $\{x \,|\, f(x) \leq f(x^{(0)})\}$, Hessian max. eigenvalue bounded: $\nabla^2 f(x) \preceq M\mathbf{Id}$.

$\rightarrow$ $f(y) \leq f(x) + \nabla f(x)^\top (y-x) + \frac{M}{2}\|y-x\|_2^2$

- M/m is an upper-bound on the condition number of $\nabla^2 f(x)$.

$$m\mathbf{Id} \preceq \nabla^2 f(x) \preceq M\mathbf{Id}$$

Convergence
(Gradient with exact line search)

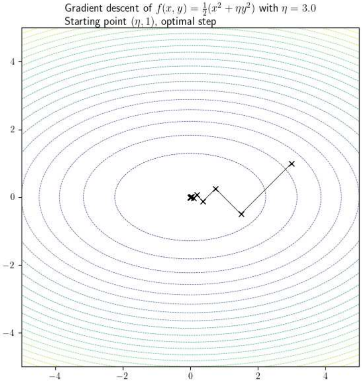
For strongly convex f:

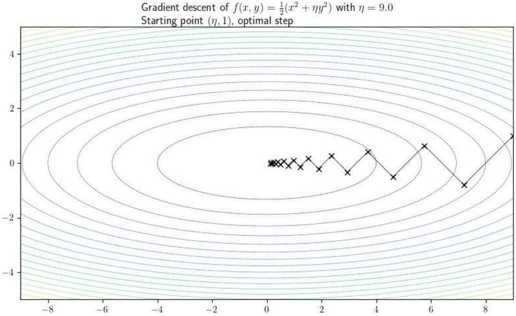
$$f(x^{(k)}) - p^\star \leq c^k (f(x^{(0)}) - p^\star)$$

- $c \in]0,1[$ is a constant, depends on $x^{(0)}$ and the function f.
- $c = 1 - \frac{m}{M}$ if $m\mathbf{Id} \preceq \nabla^2 f(x) \preceq M\mathbf{Id}$.
- $f(x^{(k)}) - p^\star \leq \epsilon$ after at most $\frac{\log((f(x^{(0)} - p^\star)/\epsilon)}{\log 1/c}$ iterations.

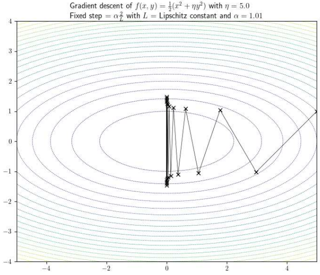
$\rightsquigarrow$ gradient very simple but very slow, rarely used in practice.

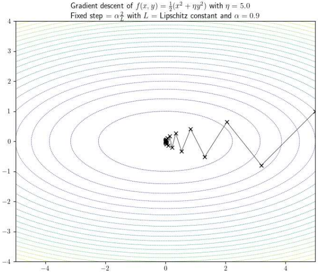
Gradient with optimal step

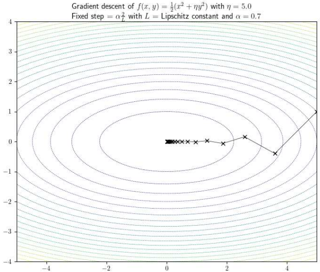


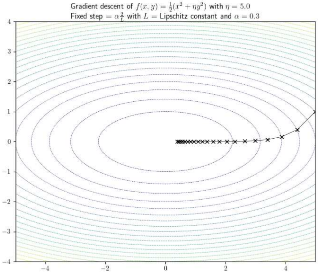


Gradient with fixed step









Steepest descent

Normalized direction (at x for given $\|.\|$)

$$\Delta x_{\text{nsd}} = \text{argmin}\{\nabla f(x)^\top v \mid \|v\| \leq 1\}$$

Unnormalized direction: $\Delta x_{\text{sd}} = \|\nabla f(x)\|_* \Delta x_{\text{nsd}}$

- For Euclidian norm, $\Delta x_{\text{sd}} = \Delta x_{\text{grad}}$.
- For the norm $\|z\|_P = (z^\top P z)^{1/2}$ with $P \in \mathbb{S}_+^n$, $\Delta x_{\text{sd}} = -P^{-1}\nabla f(x)$.
- For ℓ_1 norm, $\Delta x_{\text{sd}} = -\frac{\partial f(x)}{\partial x_i} e_i$ where e_i is i-th standard basis vector and i such that $\|\nabla f(x)\|_\infty = |[\nabla f(x)]_i|$.

Newton step

Newton method: general descent method with search direction

$$\Delta x_{\mathrm{nt}} = -\nabla^2 f(x)^{-1} \nabla f(x) .$$

- $x + \Delta x_{\mathrm{nt}}$ minimizes second order approximation

$$\hat{f}_2(x+v) = f(x) + \nabla f(x)^\top v + \frac{1}{2} v^\top \nabla^2 f(x) v$$

- $x + \Delta x_{\mathrm{nt}}$ solves linearized optimality condition

$$\begin{aligned} \nabla f(x+v) &\approx \nabla f(x) + \nabla^2 f(x) v \\ &= 0 \end{aligned}$$

- Δx_{nt} is steepest descent direction at x in local Hessian norm

Newton decrement

Measure of the proximity of x to $x^\star$:

$$\lambda(x) = \left(\nabla f(x)^\top \nabla^2 f(x)^{-1} \nabla f(x)\right)^{1/2}$$

- gives an estimate of $f(x) - p^\star$, using quadratic approximation $\widehat{f}$:

$$f(x) - \inf_y \widehat{f}(y) = \frac{1}{2}\lambda(x)^2$$

- equal to the norm of the Newton step in the quadratic Hessian norm

$$\lambda(x) = \left(\Delta x_{\mathrm{nt}}{}^\top \nabla^2 f(x) \Delta x_{\mathrm{nt}}\right)^{1/2}$$

- directional derivative in the Newton direction: $\nabla f(x)^\top \Delta x_{\mathrm{nt}} = -\lambda(x)^2$
- affine invariant (unlike $\|\nabla f(x)\|_2$)

Unconstrained Newton method

given starting point $x^{(0)} \in \mathrm{dom}\, f$, tolerance $\epsilon > 0$,
repeat:

1. Compute the Newton step $\Delta x_{\mathrm{nt}}^{(k)}$ and decrement $\lambda(x^{(k)})$.
2. Stopping criterion: **quit** if $\lambda^2/2 \leq \epsilon$
3. Choose step size t by backtracking line search.
4. Update: $x^{(k+1)} = x^{(k)} + t\Delta x_{\mathrm{nt}}^{(k)}$

- descent method: for all k, $f(x^{(k+1)}) < f(x^{(k)})$
- affine invariant: Newton iterates for $\tilde{f}(y) = f(Ty)$ with starting point $y^{(0)} = T^{-1}x^{(0)}$ are $y^{(k)} = T^{-1}x^{(k)}$.

Convergence

Newton method

For f strongly convex ($\nabla^2 f(x) \succeq m\mathbf{Id}$) and Hessian L-Lipschitz, there exist η, γ with $0 < \eta \leq m^2/L$, $\gamma > 0$:

- if $\|\nabla f(x^{(k)})\|_2 \geq \eta$ (**damped phase**):

$$f(x^{(k+1)}) - f(x^{(k)}) \leq -\gamma$$

- if $\|\nabla f(x^{(k)})\|_2 \geq \eta$ (**quadratically convergent phase**), bactracking selects unit step and:

$$\frac{L}{2m^3}\|\nabla f(x^{(k+1)})\|_2 \leq \left(\frac{L}{2m^3}\|\nabla f(x^{(k+1)})\|_2\right)^2$$

$\rightarrow$ number of iterations until $f(x) - p^\star \leq \epsilon$ bounded above by:

$$\frac{f(x^{(0)}) - p^\star}{\gamma} + \log_2 \log_2 \left(\frac{2m^3}{L^2\epsilon}\right)$$

Equality constrained minimization

With $f : \mathbb{R}^n \to \mathbb{R}$ convex, twice differentiable, find solution to:

$$\begin{cases} \text{min. } f(x) \\ \quad \text{s.t. } Ax = b \end{cases}$$

- Optimality condition: there exists a $\nu^\star$ such that:

$$\begin{bmatrix} Ax^\star = b \\ \nabla f(x^\star) + A^\top \nu^\star = 0 \end{bmatrix}$$

Equality constr. convex quad. minimization (example)

KKT conditions

$$\begin{cases} \min. \ \frac{1}{2}x^\top P x + q^\top x + r \\ \text{s.t. } Ax = b \end{cases} \quad \text{with } P \in \mathbb{S}^n_+.$$

- Lagrangian: $\mathcal{L}(x,\nu) = \frac{1}{2}x^\top P x + q^\top x + r + \nu^\top (Ax - b)$
- KKT conditions:

$$Ax^\star = b, \quad Px^\star + q + A^\top \nu^\star = 0$$

can be written as:

$$\begin{bmatrix} P & A^\top \\ A & 0 \end{bmatrix} \begin{bmatrix} x^\star \\ \nu^\star \end{bmatrix} = \begin{bmatrix} -q \\ b \end{bmatrix}$$

Equality constrained Newton method (1/2)

- Newton step at feasible point x is given by:

$$\begin{bmatrix} \nabla^2 f(x) & A^\top \\ A & 0 \end{bmatrix} \begin{bmatrix} \Delta x_{\mathrm{nt}} \\ w \end{bmatrix} = \begin{bmatrix} -\nabla f(x) \\ 0 \end{bmatrix}$$

  Interpretation:
  - Δx_{nt} solves second order approximation.
  - Linearized optimality conditions.

- Newton decrement (expression differs from unconstrained case, same interpretation):

$$\lambda(x) = \left(\Delta x_{\mathrm{nt}}{}^\top \nabla^2 f(x)^{-1} \Delta x_{\mathrm{nt}}\right)^{1/2} = \left(-\nabla f(x)^\top \Delta x_{\mathrm{nt}}\right)^{1/2}$$

Equality constrained Newton method (2/2)

given starting point $x^{(0)} \in \mathrm{dom}\, f$ with $Ax^{(0)} = b$ (feasible), tolerance $\epsilon > 0$, **repeat**:

1. Compute the Newton step Δx_{nt} and decrement $\lambda(x)$.
2. Stopping criterion: **quit** if $\lambda^2/2 \leq \epsilon$
3. Choose step size t by backtracking line search.
4. Update: $x^{(k+1)} = x^{(k)} + t\Delta x_{\mathrm{nt}}$

- feasible descent method: for all k, $f(x^{(k+1)}) < f(x^{(k)})$ and $x^{(k)}$ feasible
- affine invariant

Infeasible start Newton method (1/2)

Newton method can be generalized to infeasible x (*i.e.* $Ax \neq b$)
Newton step at infeasible point x is given by:

$$\begin{bmatrix} \nabla^2 f(x) & A^\top \\ A & 0 \end{bmatrix} \begin{bmatrix} \Delta x_{\text{nt}} \\ w \end{bmatrix} = - \begin{bmatrix} \nabla f(x) \\ Ax - b \end{bmatrix}$$

primal-dual interpretation

- write optimality conditions as $r(y) = 0$, where:

$$y = (x, \nu) \qquad r(y) = (\nabla f(x) + A^\top \nu, Ax - b)$$

- linearizing $r(y) = 0$ gives $r(y + \Delta y) \approx r(y) + Dr(y)\Delta y = 0$ and yields the above equation with $w = \nu + \Delta \nu_{\text{nt}}$.

Infeasible start Newton method (2/2)

given starting point $x^{(0)} \in \operatorname{dom} f$, $\nu^{(0)}$,
tolerance $\epsilon > 0$, $\alpha \in]0, 1/2[, \beta \in]0, 1[$

repeat:

1. Compute primal and dual Newton steps $\Delta x_{\mathrm{nt}}, \Delta \nu_{\mathrm{nt}}$
2. Bactracking line search on $\|r\|_2$.
   $t := 1$
   while $\|r(x + t\Delta x_{\mathrm{nt}}, \nu + t\Delta \nu_{\mathrm{nt}}\|_2 > (1 - \alpha t)\|r(x, \nu)\|_2$, $t := \beta t$
3. Update: $x^{(k+1)} = x^{(k)} + t\Delta x_{\mathrm{nt}}$, $\nu^{(k+1)} = \nu^{(k)} + t\Delta \nu_{\mathrm{nt}}$

until $Ax = b$ and $\|r(x, \nu)\|_2 \leq \epsilon$

- not a descent method: $f(x^{(k+1)}) > f(x^{(k)})$ is possible

Inequality constrained minimization

Notations and assumptions

With functions f_i convex, twice continuously differentiable and $A \in \mathbb{R}^{p \times n}$, $\operatorname{rank} A = p$, find solution to:

$$p^* : \qquad \begin{cases} \text{min. } f_0(x) \\ \quad \text{s.t. } f_i(x) \leq 0, \quad i = 1, \ldots, m \\ \qquad\quad Ax = b \end{cases}$$

Assumptions:

- $p^\star$ is finite and attained
- problem is strictly feasible: there exist $\tilde{x}$ with
  $\tilde{x} \in \operatorname{dom} f_0 \qquad f_i(\tilde{x}) < 0, i = 1, \ldots, m, \qquad A\tilde{x} = b$

$\rightarrow$ strong duality holds, dual optimum is attained.

Inequality constrained minimization

Reformulation

Original problem reads also:

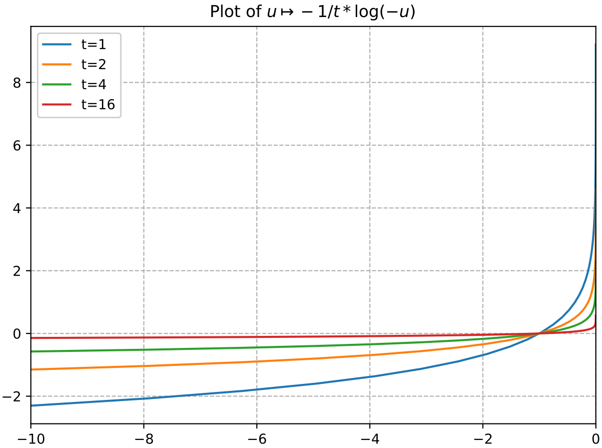
$$p^* : \qquad \begin{cases} \text{min. } f_0(x) \\ \quad \text{s.t. } f_i(x) \leq 0, \quad i = 1, \ldots, m \\ \qquad\quad Ax = b \end{cases}$$

Using **indicator function** ($\imath_{\mathbb{R}_-}(u) = 0$ if $u \leq 0$ and $+\infty$ otherwise)
$\rightsquigarrow$ equality constrained problem:

$$p^\star : \qquad \begin{cases} \text{min. } f_0(x) + \sum_{i=1}^{m} \imath_{\mathbb{R}_-}(f_i(x)) \\ \quad \text{s.t. } Ax = b \end{cases}$$

$\rightsquigarrow$ Find an approximation for $\imath_{\mathbb{R}_-}$.

Logarithmic barrier



- For $t > 0$, $u \mapsto -\frac{1}{t}\log(-u)$ is a smooth approximation of $\imath_{\mathbb{R}_-}$
- Approximation improves as $t \to \infty$

Approximate problem

$$p^\star : \qquad \begin{cases} \text{min. } f_0(x) + \sum_{i=1}^{m} \imath_{\mathbb{R}_-}(f_i(x)) \\ \text{s.t. } Ax = b \end{cases}$$

Approximation with **logarithmic barrier** $\phi(x) = -\sum_{i=1}^{m} \log(-f_i(x))$

$$\begin{cases} \text{min. } f_0(x) - \frac{1}{t} \sum_{i=1}^{m} \log(-f_i(x)) \\ \text{s.t. } Ax = b \end{cases}$$

$\rightsquigarrow$ equality constrained problem

$\rightsquigarrow$, can be solved by Newton method for increasing values of t

Central path

For $t > 0$, define $x^\star(t)$ as the solution of

$$\begin{cases} \min.\ f_0(x) - \dfrac{1}{t}\displaystyle\sum_{i=1}^{m} \log(-f_i(x)) \\ \quad \text{s.t. } Ax = b \end{cases}$$

Central path is $\{x^\star(t) \,|\, t > 0\}$

One can prove:

$$p^\star \geq f_0(x^\star(t)) - \frac{m}{t}$$

$\rightsquigarrow x^\star(t)$ converges to optimal point as $t \to \infty$

Central path: proof of suboptimality bound

From previous slide, $x^\star(t)$ satisfies for a $\hat{\nu}$:

$$\begin{cases} Ax^\star(t) = b, \quad f_i(x^\star(t)) < 0 \\ \nabla f_0(x^\star(t)) + \frac{1}{t}\sum_{i=1}^{m} \frac{1}{-f_i(x^\star(t))} \nabla f_i(x^\star(t)) + A^\top \hat{\nu} = 0 \end{cases}$$

Last equation reads $\nabla f_0(x^\star(t)) + \sum_{i=1}^m \lambda_i^\star(t) \nabla f_i(x^\star(t)) + A^\top \nu^\star(t) = 0$ with $\lambda_i^\star(t) = 1/(-t f_i(x^\star(t))) \geq 0$ and $\nu^\star(t) = \hat{\nu}$. Since $x^\star(t)$ minimizes original Lagrangian at $\lambda^\star(t), \nu^\star(t)$, the latter are dual feasible and:

$$\begin{aligned} p^\star &\geq \mathcal{L}_D(\lambda^\star(t), \nu^\star(t)) = \mathcal{L}(x^\star(t), \lambda^\star(t), \nu^\star(t)) \\ &\geq f_0(x^\star(t)) + \sum_{i=1}^{m} \lambda_i^\star(t) f_i(x^\star(t)) + \nu^\star(t)^\top (Ax^\star(t) - b) \\ &\geq f_0(x^\star(t)) - \frac{m}{t} \end{aligned}$$

Barrier method

Given strictly feasible x, $t = t^{(0)}$, $\mu > 1$, tolerance $\epsilon > 0$,
repeat:

1. Centering step. Compute $x^\star(t)$ by minimizing $tf_0 + \phi$ subject to $Ax = b$.
2. Update. $x := x^\star(t)$.
3. Stopping criterion. **quit** if $m/t \leq \epsilon$.
4. Increase t. $t := \mu t$.

- Terminates with $f_0(x) - p^\star \leq \epsilon$
- Centering usually done using Newton's method, starting at current x
- Choice of μ involves a trade-off: large μ means fewer outer iterations, more inner (Newton) iterations; typical values: $\mu = 10 - 20$.
- Several heuristics for choice of $t^{(0)}$

Feasibility and phase I methods

Feasibility problem: find x such that

$$f_i(x) \le 0, \quad i = 1, \dots, m, \quad Ax = b \tag{1}$$

Phase I: computes strictly feasible point for barrier method
Basic phase I method

$$\begin{cases} \text{min. } s \\ \quad \text{s.t. } f_i(x) \le s, \quad i = 1, \dots, m \\ \qquad\quad Ax = b \end{cases} \tag{2}$$

- If x, s feasible with $s < 0$, then x strictly feasible for (1).
- If optimal value $\bar{p}^\star$ of (2) is positive, then (1) infeasible.
- If $\bar{p}^\star = 0$ in (2) and attained, then (1) feasible (but not strictly).
  if $\bar{p}^\star = 0$ in (2) and not attained, then (1) infeasible.

Generalized inequalities

$$\begin{cases} \min.\ f_0(x) \text{s.t.} \quad f_i(x) \prec_{K_i} 0, \quad i = 1, \dots, m \\ \qquad Ax = b \end{cases}$$

- f_0 convex
- $f_i : \mathbb{R}^n \to \mathbb{R}^{k_i}$ convex with respect to proper cones $K_i \subset \mathbb{R}^{k_i}$
- f_i twice continuously differentiable
- $A \in \mathbb{R}^{p \times n}$ with $\mathrm{rank} A = p$
- We assume $p^\star$ is finite and attained
- We assume proble is strictly feasible; hence strong duality holds and dual optimum is attained

$\rightsquigarrow$ Ex: SOCP, SDP

(A few words about) Convergence

Number of outer (centering) iterations: exactly

$$\left\lceil \frac{\log(m/(\epsilon t^{(0)}))}{\log \mu} \right\rceil$$

plus the initial centering step (to compute $x^\star(t^{(0)})$)

Centering problem: see convergence analysis of Newton's method

Part V

Proximal methods

Generalities about proximal methods

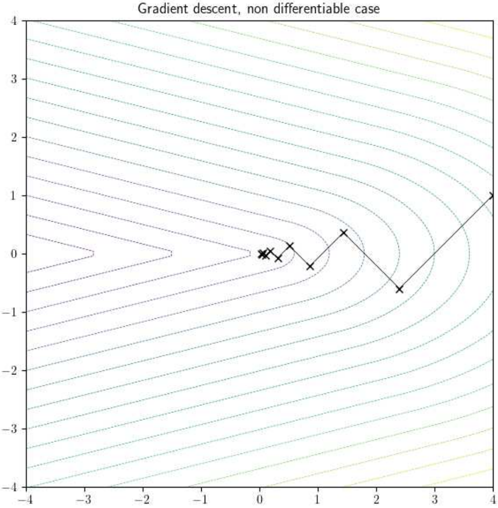
Gradient and **Newton** methods:

- smooth functions (differentiable once or twice),
- medium size problems (Newton), sometimes larger (gradient)

Proximal methods:

- suitable for **smooth** and **non-smooth** functions,
- suitable for **constrained and unconstrained** problems,
- large size and distributed implementations,
- based on high level "prox" operation, which is itself an optimization problem.

(Sub)-gradient in non differentiable case



$$f(x,y) = \begin{cases} \sqrt{x^2+\eta y^2} & \text{if } |y| \leq x, \\ \frac{x+\eta|y|}{\sqrt{1+\eta}} & \text{if } |y| \geq x\,. \end{cases}$$

Optimal step size

Starting point: $(x^{(0)}, y^{(0)}) = (\eta, 1)$

Proximal operator

Let f be a **closed proper convex** function.

Proximal operator

$$\text{prox}_f(v) = \text{Arg}\min_x f(x) + \frac{1}{2}\|x - v\|_2^2$$

Proximal operator of the **scaled** function (with $\lambda > 0$)

$$\text{prox}_{\lambda f}(v) = \text{Arg}\min_x f(x) + \frac{1}{2\lambda}\|x - v\|_2^2$$

Projection and prox

With $\imath_C$ indicator function of convex set C, **proximal operator generalizes projection** Π_C:

$$\begin{aligned}
\operatorname{prox}_{\lambda \imath_C}(v) &= \operatorname{Arg\,min}_x \imath_C(x) + \frac{1}{2\lambda}\|x-v\|_2^2 \\
&= \operatorname{Arg\,min}_{x \in C} \|x-v\|_2^2 \\
&= \Pi_C(v)
\end{aligned}$$

- Ex: for C an affine subset $C = \{x \mid Ax = b\}$:

$$\operatorname{prox}_{\imath_{\{x \mid Ax=b\}}}(v) = (\mathbf{Id} - A^\top (AA^\top)^{-1} A)v + A^\top (AA^\top)^{-1} b$$

Prox: examples

Affine function: $f(x) = b^\top x + c$:

$$\operatorname{prox}_{\lambda f}(v) = v - \lambda b$$

Quadratic function: $f(x) = \frac{1}{2}x^\top A x + b^\top x + c$ with $A \in \mathbb{S}_+^n$

$$\operatorname{prox}_{\lambda f}(v) = (\mathbf{Id} + \lambda A)^{-1}(v - \lambda b)$$

Indeed: above expression(s) obtained by setting derivative to zero
$\nabla f(x) + \frac{1}{\lambda}(x - v) = Ax + b + \frac{1}{\lambda}(x - v) = 0$

- **Shrinkage** operator: $\operatorname{prox}_{\frac{\lambda}{2}(.)^2}(v) = \frac{1}{1+\lambda}v$ or more generally:

$$\operatorname{prox}_{\frac{\lambda}{2}\|.\|_2^2}(v) = \frac{1}{1+\lambda}v$$

- For 1st order approximation $\hat{f}_1(x) = f(x_0) + \nabla f(x_0)^\top (x - x_0)$:

$$\operatorname{prox}_{\lambda \hat{f}_1}(x_0) = x_0 - \lambda \nabla f(x_0)$$

- For 2nd order approximation $\hat{f}_2(x) = f(x_0) + \nabla f(x_0)^\top (x - x_0) + \frac{1}{2}(x - x_0)^\top \nabla^2 f(x_0)(x - x_0)$:

$$\operatorname{prox}_{\lambda \hat{f}_2}(x_0) = x_0 - \left(\frac{1}{\lambda} \mathbf{Id} + \nabla^2 f(x_0) \right)^{-1} \nabla f(x_0)$$

Interpretation of prox

$$\mathrm{prox}_{\lambda f}(v) = \mathrm{Arg}\min_x f(x) + \frac{1}{2\lambda}\|x - v\|_2^2$$

- $\mathrm{prox}_{\lambda f}(v)$ moves from v towards the minimum of f, penalized by the cost of staying near to v depending on λ
- Connection with gradient step (under some assumptions, for small λ):

$$\mathrm{prox}_{\lambda f}(v) \approx v - \lambda \nabla f(v)$$

Prox and subdifferential

From $\operatorname{prox}_{\lambda f}(v) = \operatorname{Arg\,min}_x f(x) + \frac{1}{2\lambda}\|x - v\|_2^2$, it follows:

$$\begin{aligned} p = \operatorname{prox}_{\lambda f}(v) \Leftrightarrow &\quad 0 \in \partial f(p) + \frac{1}{\lambda}(p - v) \\ \Leftrightarrow &\quad v \in p + \lambda \partial f(p) \\ \Leftrightarrow &\quad v \in (\mathbf{Id} + \lambda \partial f)(p) \end{aligned}$$

Resolvent

For an operator T, the resolvent of T is $(\mathbf{Id} + \lambda T)^{-1}$.

Resolvent of subdifferential

$$\operatorname{prox}_{\lambda f} = (\mathbf{Id} + \lambda \partial f)^{-1}$$

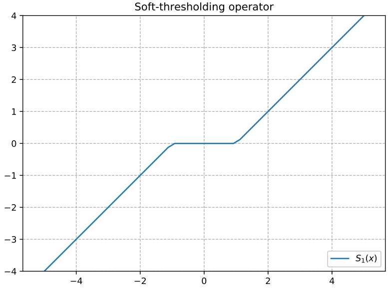
In addition, $\operatorname{prox}_{\lambda f}$ is single-valued.

Soft thresholding

(Scalar case)

$\mathrm{prox}_{\lambda|.|}(.)$ of absolute value is the **soft thresholding** operator:

$$S_\lambda(v) = \mathrm{sign}(v)\left[|v| - \lambda\right]_+ = \begin{cases} v - \lambda & \text{if } v \geq \lambda, \\ 0 & \text{if } -\lambda \leq v \leq \lambda, \\ v + \lambda & \text{if } v \leq -\lambda. \end{cases}$$



prox of separable sum

If $f(x) = \sum_{i=1}^{n} f_i(x_i)$,

$$\mathrm{prox}_f(v) = \begin{bmatrix} \mathrm{prox}_{f_1}(v_1) \\ \vdots \\ \mathrm{prox}_{f_n}(v_n) \end{bmatrix}$$

- For $f(x) = \|x\|_1$:

$$\left[\mathrm{prox}_{\lambda\|.\|_1}(v)\right]_i = S_\lambda(v_i)$$

- For $f(x) = \frac{1}{2}\|x\|_2^2$:

$$\mathrm{prox}_{\frac{\lambda}{2}\|.\|_2^2}(v) = \left(\frac{1}{1+\lambda}\right) v$$

Other properties of prox

- Precomposition: if $\tilde{f}(x) = f(\alpha x + \beta)$,

$$\operatorname{prox}_{\lambda \tilde{f}}(v) = \frac{1}{\alpha}\left[\operatorname{prox}_{\alpha^2 \lambda f}(\alpha v + \beta) - \beta\right]$$

- Postcomposition: if $\tilde{f}(x) = \alpha f(x) + b$ with $\alpha > 0$,

$$\operatorname{prox}_{\lambda \tilde{f}}(v) = \operatorname{prox}_{\alpha \lambda f}(v)$$

- Affine addition: if $\tilde{f}(x) = f(x) + a^\top x + b$,

$$\operatorname{prox}_{\lambda \tilde{f}}(v) = \operatorname{prox}_{\lambda f}(v - \lambda a)$$

- Regularization: if $\tilde{f}(x) = f(x) + \rho/2\|x - a\|_2^2$,

$$\operatorname{prox}_{\lambda \tilde{f}}(v) = \operatorname{prox}_{\tilde{\lambda} f}((\tilde{\lambda}/\lambda)v + (\rho\tilde{\lambda})a) \text{ where } \tilde{\lambda} = \lambda/(1 + \lambda\rho)$$

Moreau decomposition

Let $f^*(v) = \sup_x \langle v, x \rangle - f(x)$ be the Fenchel conjugate of f.

Moreau decomposition

$$v = \mathrm{prox}_f(v) + \mathrm{prox}_{f^*}(v)$$

Moreau decomposition

Let $f^*(v) = \sup_x \langle v, x\rangle - f(x)$ be the Fenchel conjugate of f.

Moreau decomposition

$$v = \operatorname{prox}_f(v) + \operatorname{prox}_{f^*}(v)$$

Proof: Let $p = \operatorname{prox}_f(v)$ and define $q = v - p$. By definition of prox, $q \in \partial f(p)$ and hence $p \in \partial f^*(q)$, which means $v - q \in \partial f^*(q)$ and hence $q = \operatorname{prox}_{f^*}(v)$.

Moreau decomposition

Let $f^*(v) = \sup_x \langle v, x\rangle - f(x)$ be the Fenchel conjugate of f.

Moreau decomposition

$$v = \mathrm{prox}_f(v) + \mathrm{prox}_{f^*}(v)$$

Proof: Let $p = \mathrm{prox}_f(v)$ and define $q = v - p$. By definition of prox, $q \in \partial f(p)$ and hence $p \in \partial f^*(q)$, which means $v - q \in \partial f^*(q)$ and hence $q = \mathrm{prox}_{f^*}(v)$.

- generalizes orthogonal decomposition:
  - take L a subspace and $f = \imath_L$:

$$\begin{aligned} \imath_L^*(v) &= \sup_x (v^\top x - \imath_L(x)) = \sup_{x\in L} v^\top x \\ &= \begin{cases} +\infty & \text{if } v^\top x_0 \neq 0 \text{ for an } x_0 \in L \\ 0 & \text{if } v^\top x = 0 \text{ for all } x \in L \end{cases} = \imath_{L^\perp}(v) \\ &\text{where } L^\perp = \{y \mid y^\top x = 0 \text{ for all } x \in L\} \end{aligned}$$

  - The Moreau decomposition reads: $v = \Pi_L(v) + \Pi_{L^\perp}(v)$

Fixed points of prox

Minimizers of f are fixed points of prox_f:

$$x^\star \text{ minimizes } f \Leftrightarrow x^\star = \operatorname{prox}_f(x^\star)$$

Proof:

- $\Rightarrow$ $f(x) \geq f(x^\star)$ for any x hence $f(x) + \frac{1}{2}\|x - x^\star\|_2^2 \geq f(x^\star) + \frac{1}{2}\|x^\star - x^\star\|_2^2$ which proves that $x^\star$ minimizes the l.h.s. expression.
- $\Leftarrow$ $\tilde{x} = \operatorname{prox}_f(v)$ if and only if $\tilde{x}$ minimizes $f(x) + \frac{1}{2}\|x - v\|_2^2$, that is if and only if $0 \in \partial f(\tilde{x}) + (\tilde{x} - v)$. With $\tilde{x} = v$, we get $0 \in \partial f(\tilde{x})$ and thus $\tilde{x} = v = x^\star$.

Proximal point algorithm

Proximal minimization algorithm

$$x^{(k+1)} = \mathrm{prox}_{\lambda f}\left(x^{(k)}\right)$$

- Convergence can be justified, few applications.

➤ Iterative refinement method for solving $Ax = b$ ($A \in \mathbb{S}_+^n$):

$$x^{(k+1)} = x^{(k)} + (A + \epsilon \mathbf{Id})^{-1}(b - Ax^k)$$

$\leftrightarrow$ Proximal point minimization of $g(x) = \frac{1}{2}x^\top Ax - b^\top x$:

$$\begin{aligned}\mathrm{prox}_{\lambda g}(v) &= (\mathbf{Id} + \lambda A)^{-1}(v + \lambda Av - \lambda Av + \lambda b)\\ &= v - (\frac{1}{\lambda}\mathbf{Id} + A)^{-1}(Av - b)\end{aligned}$$

Proximal gradient

- Split objective:

$$\min.\ f(x) + g(x)$$

$f : \mathbb{R}^n \to \mathbb{R}$, $g : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ are l.s.c., proper, convex;
f is **differentiable** and g can be **nonsmooth**

- Proximal gradient method:

$$x^{(k+1)} := \operatorname{prox}_{\lambda_k g}\left(x^{(k)} - \lambda_k \nabla f(x^{(k)})\right)$$

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

  - ▷ Converges with fixed step size $\lambda_k = \lambda \in]0, 2/L]$ when ∇f is Lipschitz continuous with constant L.

LASSO (Least Absolute Shrinkage and Selection Operator)

(Proximal gradient algorithm)

$$\min.\ \frac{1}{2}\|Ax-b\|_2^2+\gamma\|x\|_1$$

- Splitting:

$$f(x)=\frac{1}{2}\|Ax-b\|_2^2 \qquad g(x)=\gamma\|x\|_1$$
$$\nabla f(x)=A^\top(Ax-b) \qquad \mathrm{prox}_{\lambda g}(x)=S_{\lambda\gamma}(x)$$

- Proximal algorithm:

$$x^{(k+1)}:=S_{\lambda\gamma}\left(x^{(k)}-\lambda A^\top(Ax^{(k)}-b)\right)$$

where fixed step-size $0<\lambda\leq\frac{1}{\|A^\top A\|_2}$

▷ Sometimes called ISTA (Iterative Shrinkage-Thresholding Algorithm), accelerated version called FISTA (Fast ISTA).

Alternating Direction Method of Multipliers (ADMM)
(seen as a proximal algorithm)

- Split objective:

$$\min.\ f(x) + g(x)$$

$f, g : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ are l.s.c., proper, convex.
f **and** g can be **nonsmooth**.

- Alternating direction method of multipliers (**ADMM**):

$$\left[\begin{aligned} x^{(k+1)} &:= \operatorname{prox}_{\lambda f}(z^{(k)} - u^{(k)}) \\ z^{(k+1)} &:= \operatorname{prox}_{\lambda g}(x^{(k+1)} + u^{(k)}) \\ u^{(k+1)} &:= u^{(k)} + x^{(k+1)} - z^{(k+1)} \end{aligned}\right.$$

▷ Also known as Douglas-Rachford splitting.

Augmented Lagrangian and prox operator

- $\min.\ f(x) + g(x)$ equivalent to:

$$\begin{cases} \min.\ f(x) + g(z) \\ \quad \text{s.t. } x - z = 0 \end{cases}$$

- Augmented Lagrangian (with parameter $\rho > 0$):

$$L_\rho(x, z, y) = f(x) + g(z) + y^\top (x - z) + \frac{\rho}{2} \|x - z\|_2^2$$

can be written with $u = \frac{1}{\rho} y$:

$$L_\rho(x, z, y) = f(x) + g(z) + \frac{\rho}{2} \|x - z + u\|_2^2 - \frac{\rho}{2} \|u\|_2^2$$

$\Rightarrow$ $\operatorname{Arg\,min}_x L_\rho(x, z, y) = \operatorname{prox}_{\lambda f}(z - u)$

$\Rightarrow$ $\operatorname{Arg\,min}_z L_\rho(x, z, y) = \operatorname{prox}_{\lambda g}(x + u)$ where $\lambda = \frac{1}{\rho}$.

Alternating Direction Method of Multipliers (ADMM)

(seen as an augmented Lagrangian method)

- $\min.\ f(x) + g(x)$ equivalent to:

$$\begin{cases} \min.\ f(x) + g(z) \\ \text{s.t. } x - z = 0 \end{cases}$$

- Augmented Lagrangian (with parameter $\rho > 0$):

$$L_\rho(x, z, y) = f(x) + g(z) + y^\top (x - z) + \frac{\rho}{2}\|x - z\|_2^2$$

- Alternate Direction Method of Multipliers (**ADMM**) iterations:

$$\left[\begin{aligned} x^{(k+1)} &:= \operatorname{Arg\,min}_x L_\rho(x, z^{(k)}, y^{(k)}) \\ z^{(k+1)} &:= \operatorname{Arg\,min}_z L_\rho(x^{(k+1)}, z, y^{(k)}) \\ y^{(k+1)} &:= y^{(k)} + \rho(x^{(k+1)} - z^{(k+1)}) \end{aligned}\right.$$

Basis pursuit
(ADMM algorithm)

$$\begin{cases} \min. \ \|x\|_1 \\ \text{ s.t. } Ax = b \end{cases}$$

- Equivalent to:

$$\begin{cases} \min. \ \imath_{\{x \mid Ax=b\}}(x) + \|z\|_1 \\ \text{ s.t. } x - z = 0 \end{cases}$$

- ADMM iterations (derived from slide 154):

$$\left[\begin{array}{l} x^{(k+1)} := \Pi_{\{x \mid Ax=b\}}(z^{(k)} - u^{(k)}) \\ z^{(k+1)} := S_\lambda(x^{(k+1)} + u^{(k)}) \\ u^{(k+1)} := u^{(k)} + x^{(k+1)} - z^{(k+1)} \end{array}\right.$$

with S_λ: a soft thresholding and $\Pi_{\{x \mid Ax=b\}}$: projection.

LASSO (Least Absolute Shrinkage and Selection Operator) (ADMM algorithm)

$$\min.\ \frac{1}{2}\|Ax-b\|_2^2+\gamma\|x\|_1$$

- Equivalent to:

$$\begin{cases}\min.\ \dfrac{1}{2}\|Ax-b\|_2^2+\gamma\|z\|_1\\ \text{s.t. } x-z=0\end{cases}$$

- ADMM iterations (derived from slide 154):

$$\left[\begin{aligned}x^{(k+1)}&:=(\lambda A^\top A+\mathbf{Id})^{-1}((z^{(k)}-u^{(k)})+\lambda A^\top b)\\ z^{(k+1)}&:=S_{\lambda\gamma}(x^{(k+1)}+u^{(k)})\\ u^{(k+1)}&:=u^{(k)}+x^{(k+1)}-z^{(k+1)}\end{aligned}\right.$$

with $S_{\lambda\gamma}$: soft thresholding.

Alternating Direction Method of Multipliers (ADMM)

(seen as an augmented Lagrangian method)

$$\begin{cases} \text{min. } f(x) + g(z) \\ \quad \text{s.t. } Ax + Bz = c \end{cases}$$

- Augmented Lagrangian (with parameter $\rho > 0$):

$$L_\rho(x, z, y) = f(x) + g(z) + y^\top (Ax + Bz - c) + \frac{\rho}{2} \|Ax + Bz - c\|_2^2$$

- Alternate Direction Method of Multipliers (**ADMM**) iterations:

$$\left[\begin{array}{l} x^{(k+1)} := \operatorname{Arg\,min}_x L_\rho(x, z^{(k)}, y^{(k)}) \\ z^{(k+1)} := \operatorname{Arg\,min}_z L_\rho(x^{(k+1)}, z, y^{(k)}) \\ y^{(k+1)} := y^{(k)} + \rho(Ax^{(k+1)} + Bz^{(k+1)} - c) \end{array}\right.$$

Generalized LASSO

(ADMM algorithm)

$$\min.\ \frac{1}{2}\|Ax-b\|_2^2 + \gamma\|Fx\|_1$$

- Equivalent to:

$$\begin{cases} \min.\ \frac{1}{2}\|Ax-b\|_2^2 + \gamma\|z\|_1 \\ \text{s.t. } Fx - z = 0 \end{cases}$$

- ADMM iterations (derived from slide 159 with $\rho = 1/\lambda$, compare with slide 158):

$$\left[\begin{array}{l} x^{(k+1)} := (A^\top A + \rho F^\top F)^{-1}(A^\top b + \rho F^\top (z^{(k)} - u^{(k)})) \\ z^{(k+1)} := S_{\gamma/\rho}(Fx^{(k+1)} + u^{(k)}) \\ u^{(k+1)} := u^{(k)} + Fx^{(k+1)} - z^{(k+1)} \end{array}\right.$$