Solvers Principles and Architecture (SPA)

Lectures 6 and 8

Convex Optimization

Master Sciences Informatique (Sif)
October 23rd, 2017
Rennes

Khalil Ghorbal
khalil.ghorbal@inria.fr

$$\begin{aligned} \min \quad & f_0(x) \\ s.t. \quad & f_i(x) \leq 0, \quad i = 1, \ldots, m \\ & h_j(x) = 0, \quad j = 1, \ldots, p \end{aligned}$$

- $x \in \mathbb{R}^n$: optimization variable
- $f_0 : \mathbb{R}^n \to \mathbb{R}$, objective or cost function
- $f_i : \mathbb{R}^n \to \mathbb{R}$, $i = 1, \ldots, m$: inequality constraint functions
- $h_j : \mathbb{R}^n \to \mathbb{R}$, $j = 1, \ldots, p$: equality constraint functions

Optimal value:

$$p^\star = \inf\{f_0(x) \mid f_i(x) \leq 0, i = 1, \ldots, m, h_j(x) = 0, j = 1, \ldots, p\}$$

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

- Find an optimum of a function over a constrained set
- Optimum: minimum or maximum
- The function to optimize is called the objective or cost function
- The constrained set is called the feasible set
- A solution to the problem is called an objective vector
- An optimum is called an objective value

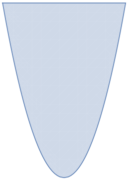
Convex Optimization

- The objective function is convex
- The feasible set is convex

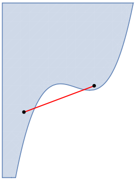
What is convexity?

Outline

- **Intuition**: A set C is convex if and only if, for any two points in C, *the shortest path* that links these two points is also entirely in C.
- In this course: we will restrict ourselves to the flat Euclidean space $\mathbb{R}^n$, thus $C \subseteq \mathbb{R}^n$.
- As a vector space, a point $p \in \mathbb{R}^n$ is a vector and one can define scalar multiplication and addition over vectors.
- In these settings, C is convex if and only if, for all $c_1, c_2 \in C$, for all $\lambda \in [0,1], \lambda c_1 + (1-\lambda)c_2$ is also in C.



Convex



Non convex

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

$$\text{epi}(f) := \{(x, y) \mid f(x) \leq y\} \subset \mathbb{R}^n \times \mathbb{R}$$

- f is convex if and only if its epigraph is a convex set
- f is concave if and only if $-f : x \mapsto -f(x)$ is convex

Examples:

- $f : x \mapsto x^2$ is convex (cf. left figure in the previous slide)
- $f : x \mapsto x^3 + x^2$ is not convex (cf. right figure in the previous slide)

- $\forall \lambda \in [0,1].\ \forall x, y. \quad f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda) f(y)$
- **Intuition**: the image of a point in the segment joining x and y is somewhere below the segment joining $f(x)$ and $f(y)$
- Any local minimum of f is also a global minimum
- One can define a weak notion of differentiability over convex functions
- The sub-differential of $f : \mathbb{R}^n \to \mathbb{R}$ at x is defined by the following set:

$$\partial f(x) := \{z \in \mathbb{R}^n \mid \forall t \in \mathbb{R}^n. \quad f(t) \geq f(x) + z \cdot (t - x)\}$$

  where $x \cdot y$ denotes the usual scalar product over $\mathbb{R}^n$
- **Intuition**: the sub-differential at x is the set of all affine functions that touches the graph of f only at x
- **Example**: the absolute value function is non-differentiable at 0 in the usual sense, but it is sub-differentiable, $\partial f(0) = [-1, 1]$

$$\begin{aligned} \min \quad & f_0(x) \\ s.t. \quad & f_i(x) \leq 0, \quad i = 1, \ldots, m \\ & a_j \cdot x = b_j, \quad j = 1, \ldots, p \end{aligned}$$

- $f_0, f_1, \ldots, f_m$ are convex functions
- h_j (equality constraints) are affine in x

Important property: The feasible set of convex optimization problem is convex

Find the minimum of a convex function $f : \mathbb{R}^n \to \mathbb{R}$, over a convex set C

$$\begin{aligned} \min \quad & f(x) \\ s.t. \quad & x \in C \\ & x \geq 0 \end{aligned}$$

$$\boxed{\begin{aligned} \min \quad & f(x) \\ s.t. \quad & Ax \leq b \qquad (\mathcal{P}) \\ & x \geq 0 \end{aligned}}$$

- **This lecture**: C will be a polyhedron, or a linear convex set
- A is an $m \times n$ matrix of rank m (no redundant rows)
- Inequalities are component-wise
- x is in the positive orthant $(\mathbb{R}^n_+)$
- Any set of linear non-strict constraints can be put in this form: if $x \leq a$ then introduce $y \geq 0$ and substitute x by $a - y$ everywhere

Let $\mathrm{val}(\mathcal{P})$ denote the objective value of $\mathcal{P}$

Existence

- The feasible set C could be empty: the problem is infeasible
- $\mathrm{val}(\mathcal{P})$ could be unbounded: the problem is degenerated

Uniqueness

- For feasible, non-degenerated problems, $\mathrm{val}(\mathcal{P})$ is unique for convex objective functions because every local minimum is global
- The objective value could be reached by several objective vectors

Intuition: inject the constraint into the objective function

The Lagrangian associated to $(\mathcal{P})$ is defined by ($y \in \mathbb{R}^m$, $y \geq 0$)

$$L(x, y) = f(x) + y \cdot (Ax - b)$$

If there exists an index i such that the ith component of the vector $Ax - b$ is positive, then $L(x, y)$ is unbounded since y_i can be chosen arbitrarily big. Thus

$$\sup_y L(x, y) = \begin{cases} f(x) & \text{if } Ax \leq b \\ +\infty & \text{otherwise} \end{cases}$$

Solving $(\mathcal{P})$ is then equivalent to minimizing $\sup_y L(x, y)$ over x:

$$\mathrm{val}(\mathcal{P}) = \inf_x \sup_y L(x, y)$$

In general (no convexity assumptions are needed here), if L is a real valued function defined over the product $X \times Y$, then

$$\sup_y \inf_x L(x, y) \leq \inf_x \sup_y L(x, y)$$

Proof. Let $(\bar{x}, \bar{y}) \in X \times Y$, then, by definition of inf and sup

$$\inf_x L(x, \bar{y}) \leq L(\bar{x}, \bar{y}) \leq \sup_y L(\bar{x}, y)$$

So the quantity $\sup_y L(\bar{x}, y)$ is an upper bound of $\inf_x L(x, \bar{y})$. Since the sup is the smallest upper bound by definition, one gets

$$\sup_y \inf_x L(x, \bar{y}) \leq \sup_y L(\bar{x}, y)$$

But then $\sup_y \inf_x L(x, \bar{y})$ is a lower bound for $\sup_y L(\bar{x}, y)$. Since, dually, the inf is the biggest lower bound, one gets the desired result.

Weak duality applied to $(\mathcal{P})$

By the weak duality, we get a lower bound of the objective value

$$\mathrm{val}(\mathcal{D}) := \sup_y \inf_x L(x, y) \leq \inf_x \sup_y L(x, y) = \mathrm{val}(\mathcal{P})$$

where $\mathrm{val}(\mathcal{D})$ if the objective value of a distinct, yet related, optimization problem, $(\mathcal{D})$, called the dual problem, and defined by $\sup_y \inf_x L(x, y)$, for the exact same Lagrangian L of $(\mathcal{P})$, called the primal problem.

Suppose $f(x) = c \cdot x$ for some vector $c \in \mathbb{R}^n$.
The Lagrangian L could be rearranged as follows (recall that $Ax \cdot y = x \cdot A^t y$, where A^t denotes the transpose of the matrix A):

$$L(x,y) = -b \cdot y + x \cdot (A^t y + c)$$

Since $x \geq 0$, we get

$$\inf_x L(x,y) = \begin{cases} -b \cdot y & \text{if } A^t y + c \geq 0 \\ -\infty & \text{otherwise} \end{cases}$$

and hence, the dual problem is (the primal is recalled on the right)

$$\begin{array}{lll} \max & -b \cdot y & \\ s.t. & A^t y \geq -c & (\mathcal{D}) \\ & y \geq 0 & \end{array} \qquad \begin{array}{lll} \min & c \cdot x & \\ s.t. & Ax \leq b & (\mathcal{P}) \\ & x \geq 0 & \end{array}$$

- The evaluation of the dual cost function on any feasible point of the dual problem bounds from below the evaluation of the primal cost function on any feasible point of the primal problem:

$$\forall y.\ A^t y \geq -c, \quad \forall x.\ Ax \leq b. \quad -b \cdot y \leq c \cdot x$$

- If the primal problem is degenerated then the dual is unfeasible
- If the dual problem is degenerated then the primal is unfeasible
- The primal and dual cannot be degenerated simultaneously, but
- The primal and the dual can be both unfeasible:

$$\begin{array}{lll} \max & y & \\ s.t. & 0y \geq 1 & (\mathcal{D}) \\ & y \geq 0 & \end{array} \qquad \begin{array}{lll} \min & -x & \\ s.t. & 0x \leq -1 & (\mathcal{P}) \\ & x \geq 0 & \end{array}$$

Strong duality occurs when $\text{val}(\mathcal{D}) = \text{val}(\mathcal{P})$

Theorem
When the cost function is linear, $(\mathcal{P})$ and $(\mathcal{D})$ are either strongly dual or both unfeasible.

Proof. By Farkas' Lemma (upcoming lectures)

Any polyhedron $Ax \leq b$ can be seen as the projection of a polyhedron $A'x' = b$ of higher dimension: simply augment the vector $x \in \mathbb{R}^n$ with the slack vector $s \in \mathbb{R}^m$, $s \geq 0$, and the matrix A with the identity matrix I_m of $\mathbb{R}^m$ to saturate all inequalities:

$$A' := \begin{pmatrix} A & I_m \end{pmatrix}; \quad x' := \begin{pmatrix} x \\ s \end{pmatrix} \in \mathbb{R}^{n+m}$$

Thus $A'x' = Ax + I_m s = b$.
(If x_i is unconstrained, one adds two nonnegative variables s_{1i}, s_{2i} and substitute x_i by $s_{1i} - s_{2i}$)

So there is no loss of generality if one considers the saturated form

$$A'x' = b \quad \text{instead of} \quad Ax \leq b$$

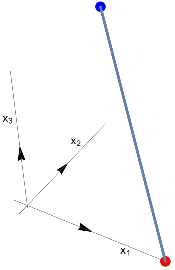
Consider the convex polyhedron $Ax = b$, where the rank of A is m

- If $m > n$, the problem is unfeasible
- Thus, in the sequel, we will suppose that $m \leq n$

Vertex of convex polyhedron. For $x \geq 0$, define

$$\mathfrak{B} := \{i \mid x_i > 0\} \qquad \mathfrak{N} := \{i \mid x_i = 0\}$$

Then, rearrange the matrix A accordingly: $A = \begin{pmatrix} A_{\mathfrak{B}} & A_{\mathfrak{N}} \end{pmatrix}$.
x is a vertex if $\operatorname{rank}(A_{\mathfrak{B}}) \leq m$.



- degenerated vertex $\operatorname{rank}(A_{\mathfrak{B}}) < m$
- non-degenerated vertex $\operatorname{rank}(A_{\mathfrak{B}}) = m$
- **Example**: $\begin{pmatrix} 1 & 1 & 0 \\ 0 & 1 & -1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$
- $(1, 0, 0)$ (red point) is degenerated
- $(0, 1, 1)$ (blue point) is non-degenerated

Let $\hat{x} = \begin{pmatrix} \hat{x}_{\mathfrak{B}} \\ \hat{x}_{\mathfrak{N}} \end{pmatrix}$ be a feasible non-degenerated vertex of $Ax = b$

- $\hat{x}_{\mathfrak{B}} > 0$, $\hat{x}_{\mathfrak{N}} = 0$ and $A_{\mathfrak{B}}$ is invertible (because $\hat{x}$ is non-degenerated)
- Let x be a feasible point in an open ball centered at $\hat{x}$ and separating $\hat{x}$ from all the other feasible vertices
- Such a ball is non-empty because vertices are discrete
- For all x in this ball the constraint $Ax = b$, rewritten as $A_{\mathfrak{B}} x_{\mathfrak{B}} + A_{\mathfrak{N}} x_{\mathfrak{N}} = b$, links $x_{\mathfrak{B}}$ to $x_{\mathfrak{N}}$, namely
- Thus $x_{\mathfrak{B}} = A_{\mathfrak{B}}^{-1}(b - A_{\mathfrak{N}} x_{\mathfrak{N}})$
- So locally, the cost function $c \cdot x$ depends only on $x_{\mathfrak{N}}$, indeed
- $c \cdot x = \begin{pmatrix} c_{\mathfrak{B}} \\ c_{\mathfrak{N}} \end{pmatrix} \cdot \begin{pmatrix} A_{\mathfrak{B}}^{-1}(b - A_{\mathfrak{N}} x_{\mathfrak{N}}) \\ x_{\mathfrak{N}} \end{pmatrix} = \underbrace{\left(c_{\mathfrak{N}} - A_{\mathfrak{N}}^t {A_{\mathfrak{B}}^{-1}}^t c_{\mathfrak{B}}\right)}_{r} \cdot x_{\mathfrak{N}} + \underbrace{c_{\mathfrak{B}} \cdot A_{\mathfrak{B}}^{-1} b}_{a}$
- Thus, $c \cdot x$ has the form $r \cdot x_{\mathfrak{N}} + a$
- To locally decrease the value of $c \cdot x$, only $x_{\mathfrak{N}}$ counts

- The function $r \cdot x_{\mathfrak{N}}$ is called the reduced cost function
- We seek a displacement that locally decreases the reduced cost function
- Suppose that there exists a index j such that the jth component of the vector r is negative, $r_j < 0$
- Consider a displacement along e_j (where e_j denotes jth vector of the canonical orthonormal basis of $\mathbb{R}^{n-m}$), ρ is a positive real number
- $r \cdot (\hat{x}_{\mathfrak{N}} + \rho e_j) = r \cdot \hat{x}_{\mathfrak{N}} + \rho r \cdot e_j = r \cdot \hat{x}_{\mathfrak{N}} + \rho r_j < r \cdot \hat{x}_{\mathfrak{N}}$, and
- $\hat{x}_{\mathfrak{N}} + \rho e_j \geq 0$ (feasible)

Optimality criterion: $r \geq 0$
No possible local improvement for $r \cdot x_{\mathfrak{N}}$, hence we reached a local minimum, which is also global by convexity.

- The displacement ρ could be bounded
- Recall that locally $x_{\mathfrak{B}} = A_{\mathfrak{B}}^{-1}(b - A_{\mathfrak{N}} x_{\mathfrak{N}})$
- So the update $x_{\mathfrak{N}} \leftarrow \hat{x}_{\mathfrak{N}} + \rho e_j$ leads to $x_{\mathfrak{B}} \leftarrow \hat{x}_{\mathfrak{B}} + \rho d_{\mathfrak{B}}$, where $d_{\mathfrak{B}} = -A_{\mathfrak{B}}^{-1} A_{\mathfrak{N}} e_j$
- Since $x_{\mathfrak{B}} \geq 0$, we get $\hat{x}_{\mathfrak{B}} \geq \rho(-d_{\mathfrak{B}})$
- So ρ should not exceed the minimum of the quotient of the ith component of $\hat{x}_{\mathfrak{B}}$ over the $i$the component of $-d_{\mathfrak{B}}$ (whenever defined)

Unboundedness criterion: $d_{\mathfrak{B}} \geq 0$
In this case, ρ can be chosen arbitrarily big and the minimum is $-\infty$

- As seen earlier, a basis corresponds to a vertex (degenerated or not)
- When $\hat{x}$ is updated, one reaches a new basis $\mathfrak{B}$: one component of $x_{\mathfrak{B}}$ will vanish and gets replaced by a component of $x_{\mathfrak{N}}$ which becomes strictly positive after the update
- So an update moves from one vertex to another
- Moreover, such updates make $\hat{x}$ moves along the edges of the convex polyhedron $Ax = b$ since the displacement of $x_{\mathfrak{N}}$ are performed along one component at a time (namely the jth for the update $x_{\mathfrak{N}} \leftarrow \hat{x}_{\mathfrak{N}} + \rho e_j$)
- Recall: an edge is a facet of dimension 1

1. Start at a vertex (basis)
2. If the optimality criterion is satisfied, halt: the problem is solved
3. Otherwise, move along an edge that minimizes the reduced cost function
4. If the unboundedness criterion is satisfied, halt: the problem is unbounded
5. Otherwise, we reach a new basis and we loop back to the first step

Does it always terminate?

$$\begin{aligned} \min \quad & x_1 + x_2 \\ s.t. \quad & x_1 + x_2 = 0 \\ & x \geq 0 \end{aligned}$$

- Start with the basis $\mathfrak{B} = \{1\}$, $\mathfrak{N} = \{2\}$
- $\hat{x}^t = \begin{pmatrix} 0 & 0 \end{pmatrix}$, $A_{\mathfrak{B}} = A_{\mathfrak{N}} = \begin{pmatrix} 1 \end{pmatrix}$
- $r = c_{\mathfrak{N}} - A_{\mathfrak{N}}^t A_{\mathfrak{B}}^{-t} c_{\mathfrak{B}} = \begin{pmatrix} -1 \end{pmatrix}$ and $d_{\mathfrak{B}} = -A_{\mathfrak{B}}^{-1} A_{\mathfrak{N}} e_j = \begin{pmatrix} -1 \end{pmatrix}$
- update $x_{\mathfrak{N}} \leftarrow 0 + \rho$, $x_{\mathfrak{B}} \leftarrow 0 - \rho$, so $\rho = 0$
- So the algorithm is updating the basis without changing the vertex
- Similarly, the algorithm can cycle forever between several bases without actually halting

$$\begin{aligned} \min \quad & f_0(x) \\ s.t. \quad & f_i(x) \leq 0 \\ & Ax = b \end{aligned}$$

- f_i convex and 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 the problem is strictly feasible: there exists $\tilde{x}$ with

$$\tilde{x} \in \mathrm{dom} f_0, \qquad f_i(\tilde{x}) < 0, i = 1, \ldots, m, \qquad A\tilde{x} = b$$

hence, strong duality holds and dual optimum is attained

Example: Linear, Quadratic, Geometric Programming (LP, QP, GP)

Reformulate the problem via indicator functions:

$$\begin{aligned} \min \quad & f_0(x) + \sum_{i=1}^{m} I_-(f_i(x)) \\ s.t. \quad & Ax = b \end{aligned}$$

where $I_-(u) = 0$ if $u \leq 0$, $I_-(u) = 0$ otherwise (indicator function of $\mathbb{R}_-$)
Approximation via logarithmic barrier

$$\begin{aligned} \min \quad & f_0(x) - \left(\frac{1}{t}\right) \sum_{i=1}^{m} \log(-f_i(x)) \\ s.t. \quad & Ax = b \end{aligned}$$

- for $t > 0$, $-\left(\frac{1}{t}\right) \log(-u)$ is a smooth approximation of I_-
- approximation improves as $t \to +\infty$

$$\phi(x) = -\sum_{i=1}^{m} \log(-f_i(x)), \qquad \mathrm{dom}\phi = \{x \mid f_1(x) < 0, \ldots, f_m(x) < 0\}$$

- convex (composition rules)
- twice continuously differentiable

$$\nabla\phi(x) = \sum_{i=1}^{m} \frac{1}{-f_i(x)} \nabla f_i(x)$$

$$\nabla^2\phi(x) = \sum_{i=1}^{m} \frac{1}{-f_i(x)^2} \nabla f_i(x) \nabla f_i(x)^t + \sum_{i=1}^{m} \frac{1}{-f_i(x)} \nabla^2 f_i(x)$$

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

$$\begin{aligned} \min \quad & tf_0(x) + \phi(x) \\ s.t. \quad & Ax = b \end{aligned}$$

  (assuming $x^\star(t)$ exists and is unique for each $t > 0$)
- **Central path** is $\{x^\star(t) \mid t > 0\}$

Example: the central path for an LP: $\min c^t x$, subject to $a_i^t x \leq b_i$, for $i = 1, \ldots, 6$ is defined by the successive tangent points of the hyperplane $c^t x = c^t x^\star(t)$ to the level curves of ϕ.

x is equal to $x^\star(t)$ if there exists a w such that

$$t\nabla f_0(x) + \nabla\phi(x) + A^t w = 0, \qquad Ax = b.$$

Therefore, $x^\star(t)$ minimizes the Lagrangian

$$L(x, \lambda^\star(t), \mu^\star(t)) = f_0(x) + \sum_{i=1}^{m} \lambda_i^\star(t) f_i(x) + \mu^\star(t)^t (Ax - b)$$

where we define $\lambda_i^\star(t) = \frac{1}{-tf_i(x^\star(t))}$ and $\mu^\star(t) = \frac{w}{t}$
This confirms the intuitive idea that $f_0(x^\star(t)) \to p^\star$ if $t \to +\infty$:

$$\begin{aligned} p^\star &\geq g(\lambda^\star(t), \mu^\star(t)) \\ &= L(x^\star(t), \lambda^\star(t), \mu^\star(t)) \\ &= f_0(x^\star(t)) - \frac{m}{t} \end{aligned}$$

Start with a strictly feasible x, $t := t^{(0)} > 0$, $\mu > 1$, and $\epsilon > 0$

1. Compute $x^\star(t)$ by minimizing $tf_0(x) + \phi(x)$ subject to $Ax = b$ (Centering method)
2. Update $x := x^\star(t)$
3. Quit if $\frac{m}{t} < \epsilon$ (Stopping criterion)
4. increase t: $t := \mu t$
5. repeat

- terminates with $f_0(x) - p^\star \leq \epsilon$
- Newton's method is used for centering
- choice of μ involves a trade-off: large μ means fewer outer iterations and more inner (Newton) iterations. Typically $10 \leq \mu \leq 20$
- Several heuristics are used for the choice of $t^{(0)}$

$$\begin{aligned} \min \quad & f_0(x) \\ s.t. \quad & f_i(x) \preceq 0, i = 1, \ldots, m \\ & Ax = b \end{aligned}$$

- f_0 convex, $f_i : \mathbb{R}^n \to \mathbb{R}^{k_i}$, $i = 1, \ldots, m$, convex w.r.t. proper cones $K_i \in \mathbb{R}^{k_i}$
- f_i twice continuously differentiable
- $A \in \mathbb{R}^{p \times n}$ with rank p
- assume $p^\star$ finite and attained
- assume the problem is strongly feasible and hence strong duality holds and dual optimum is attained

$\psi : \mathbb{R}^q \to \mathbb{R}$ is a generalized logarithm for the proper cone $K \subseteq \mathbb{R}^q$ if:

- $\mathrm{dom}\psi = \mathrm{int}K$ and $\nabla^2\psi(y) \prec 0$ for $0 \prec_K y$
- $\psi(sy) = \psi(y) + \theta log(s)$ for $0 \prec_K y$ and $s > 0$ (θ is the degree of ψ)

Examples:

- When K is the nonnegative orthant $K = \mathbb{R}^n_+$: $\psi(y) = \sum_{i=1}^n log(y_i)$ $(theta = n)$
- positive semidefinite cone $K = S^n_+$:

$$\psi(Y) = \log(\det Y) \qquad (\theta = n)$$

- Duality is key to approximate optimal solutions of hard nonconvex problems
- Almost all optimality conditions are more or less a rewriting of the KKT conditions
- Almost all state-of-the-art solvers implement interior point methods and its generalized version (for SDP)

This course is largely based on:

- *Convex Optimization*. Stephen Boyd and Lieven Vandenberghe. Cambridge University Press
- *Recherche Opérationnelle: aspects mathématiques et applications*. Frédéric Bonnans and Stéphane Gaubert. Ecole Polytechnique
- EE364A (Stephen Boyd, Stanford), EE236B (UCLA), Convex Optimization
    - www.stanford.edu/class/ee364a
    - www.ee.ucla.edu/ee236b/