Convexity II: Optimization Basics

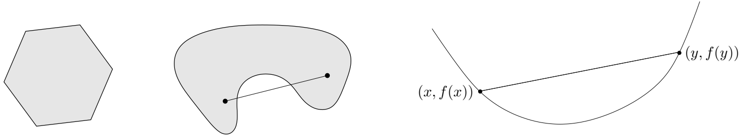
Lecturer: Aarti Singh
Co-instructor: Pradeep Ravikumar

Convex Optimization 10-725/36-725

Last time: convex sets and functions

"Convex calculus" makes it easy to check convexity. Tools:

- Definitions of convex sets and functions, classic examples



- Key properties (e.g., first- and second-order characterizations for functions)
- Operations that preserve convexity (e.g., affine composition)

E.g., is $\max\left\{\log\left(\frac{1}{(a^T x + b)^7}\right), \|Ax + b\|_1^5\right\}$ convex?

Outline

Today:

- Optimization terminology
- Properties and first-order optimality
- Equivalent transformations

Convex optimization problems

Optimization problem:

$$\begin{aligned} \min_{x \in D} \quad & f(x) \\ \text{subject to} \quad & g_i(x) \le 0,\ i = 1, \dots m \\ & h_j(x) = 0,\ j = 1, \dots p \end{aligned}$$

Here $D = \mathrm{dom}(f) \cap \bigcap_{i=1}^m \mathrm{dom}(g_i) \cap \bigcap_{j=1}^p \mathrm{dom}(h_j)$, common domain of all the functions

This is a convex optimization problem provided the functions f and $g_i, i = 1, \dots m$ are convex, and $h_j, j = 1, \dots p$ are affine:

$$h_j(x) = a_j^T x + b_j, \quad j = 1, \dots p$$

Optimization terminology

Reminder: a convex optimization problem (or program) is

$$\begin{aligned} \min_{x \in D} \quad & f(x) \\ \text{subject to} \quad & g_i(x) \leq 0,\ i = 1, \dots m \\ & h_j(x) = 0\ j = 1, \dots p \end{aligned}$$

where f and g_i, $i = 1, \dots m$ are all convex, h_j, $j = 1, \dots, p$ are affine, and the optimization domain is D (often we do not write D) where $D = \mathrm{dom}(f) \cap \bigcap_{i=1}^m \mathrm{dom}(g_i) \cap \bigcap_{j=1}^p \mathrm{dom}(g_j)$

- f is called criterion or objective function
- g_i is called inequality constraint function
- h_j is called equality constraint function
- If $x \in D$, $g_i(x) \leq 0$, $i = 1, \dots m$, and $h_j(x) = 0$, $j = 1, \dots, p$ then x is called a feasible point
- The minimum of $f(x)$ over all feasible points x is called the optimal value, written $f^\star$

- If x is feasible and $f(x) = f^\star$, then x is called optimal; also called a solution, or a minimizer[1]
- If x is feasible and $f(x) \le f^\star + \epsilon$, then x is called ϵ-suboptimal
- If x is feasible and $g_i(x) = 0$, then we say g_i is active at x
- Convex minimization can be reposed as concave maximization

$$\begin{array}{ll} \min\limits_{x} & f(x) \\ \text{subject to} & g_i(x) \le 0, \\ & \;\; i = 1, \ldots m \\ & Ax = b \end{array} \quad \Longleftrightarrow \quad \begin{array}{ll} \max\limits_{x} & -f(x) \\ \text{subject to} & g_i(x) \le 0, \\ & \;\; i = 1, \ldots m \\ & Ax = b \end{array}$$

Both are called convex optimization problems

[1] Note: a convex optimization problem need not have solutions, i.e., need not attain its minimum, but we will not be careful about this

Convex solution sets

Let $X_{\sf opt}$ be the set of all solutions of convex problem, written

$$\begin{array}{rcll} X_{\sf opt} & = & \text{argmin} & f(x) \\ & & \text{subject to} & g_i(x) \le 0, \; i = 1, \ldots m \\ & & & Ax = b \end{array}$$

Key property: $X_{\sf opt}$ is a convex set

Proof: use definitions. If x, y are solutions, then for $0 \le t \le 1$,

- $tx + (1-t)y \in D$
- $g_i(tx + (1-t)y) \le tg_i(x) + (1-t)g_i(y) \le 0$
- $A(tx + (1-t)y) = tAx + (1-t)Ay = b$
- $f(tx + (1-t)y) \le tf(x) + (1-t)f(y) = f^\star$

Therefore $tx + (1-t)y$ is also a solution

Another key property: if f is strictly convex, then the solution is unique, i.e., $X_{\sf opt}$ contains one element

Example: lasso

Given $y \in \mathbb{R}^n$, $X \in \mathbb{R}^{n \times p}$, consider the lasso problem:

$$\begin{aligned} \min_\beta \quad & \|y - X\beta\|_2^2 \\ \text{subject to} \quad & \|\beta\|_1 \le s \end{aligned}$$

Is this convex? What is the criterion function? The inequality and equality constraints? Feasible set? Is the solution unique, when:

- $n \ge p$ and X has full column rank?
- $p > n$ ("high-dimensional" case)?

Example: lasso

Given $y \in \mathbb{R}^n$, $X \in \mathbb{R}^{n \times p}$, consider the lasso problem:

$$\begin{aligned} \min_{\beta} \quad & \|y - X\beta\|_2^2 \\ \text{subject to} \quad & \|\beta\|_1 \le s \end{aligned}$$

Is this convex? What is the criterion function? The inequality and equality constraints? Feasible set? Is the solution unique, when:

- $n \ge p$ and X has full column rank?
- $p > n$ ("high-dimensional" case)?

How do our answers change if we changed criterion to Huber loss:

$$\sum_{i=1}^n \rho(y_i - x_i^T \beta), \quad \rho(z) = \begin{cases} \frac{1}{2} z^2 & |z| \le \delta \\ \delta|z| - \frac{1}{2}\delta^2 & \text{else} \end{cases} \quad ?$$

Example: support vector machines

Given $y \in \{-1, 1\}^n$, $X \in \mathbb{R}^{n \times p}$ with rows $x_1, \ldots x_n$, consider the support vector machine or SVM problem:

$$\begin{array}{ll} \displaystyle\min_{\beta, \beta_0, \xi} & \displaystyle\frac{1}{2}\|\beta\|_2^2 + C\sum_{i=1}^n \xi_i \\ \text{subject to} & \xi_i \geq 0, \ i = 1, \ldots n \\ & y_i(x_i^T\beta + \beta_0) \geq 1 - \xi_i, \ i = 1, \ldots n \end{array}$$

Is this convex? What is the criterion, constraints, feasible set? Is the solution (β, β_0, ξ) unique?

Local minima are global minima

For convex optimization problems, local minima are global minima

Local minimum: If x is feasible ($x \in D$, and satisfies all constraints) and minimizes f in a local neighborhood, i.e. for some $\rho > 0$

$$f(x) \leq f(y) \text{ for all feasible } y, \ \|x - y\|_2 \leq \rho$$

Local minima are global minima

For convex optimization problems, local minima are global minima

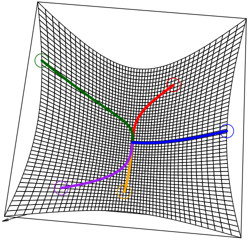
Local minimum: If x is feasible ($x \in D$, and satisfies all constraints) and minimizes f in a local neighborhood, i.e. for some $\rho > 0$

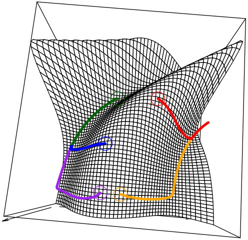
$$f(x) \le f(y) \text{ for all feasible } y, \ \|x - y\|_2 \le \rho$$

For convex problems, x is also a global minimum

$$f(x) \le f(y) \text{ for all feasible } y$$

This is a very useful fact and will save us a lot of trouble!





Rewriting constraints

The optimization problem

$$\begin{aligned} \min_x \quad & f(x) \\ \text{subject to} \quad & g_i(x) \le 0, \ i = 1, \ldots m \\ & Ax = b \end{aligned}$$

can be rewritten as

$$\min_x \ f(x) \ \text{ subject to } \ x \in C$$

where $C = \{x : g_i(x) \le 0, \ i = 1, \ldots m, \ Ax = b\}$, the feasible set. Hence the above formulation is completely general

Rewriting constraints

The optimization problem

$$\begin{aligned} \min_x \quad & f(x) \\ \text{subject to} \quad & g_i(x) \le 0, \ i = 1, \dots m \\ & Ax = b \end{aligned}$$

can be rewritten as

$$\min_x \ f(x) \ \text{ subject to } \ x \in C$$

where $C = \{x : g_i(x) \le 0, \ i = 1, \dots m, \ Ax = b\}$, the feasible set. Hence the above formulation is completely general

With I_C the indicator of C, we can write this in unconstrained form

$$\min_x \ f(x) + I_C(x)$$

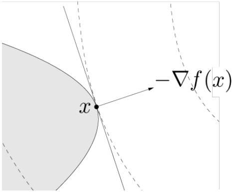
First-order optimality condition

For a convex problem

$$\min_x \; f(x) \;\; \text{subject to} \;\; x \in C$$

and differentiable f, a feasible point x is optimal if and only if

$$\nabla f(x)^T (y - x) \geq 0 \quad \text{for all } y \in C$$



This is called the first-order condition for optimality

In words: all feasible directions from x are aligned with gradient $\nabla f(x)$

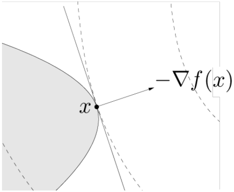
First-order optimality condition

For a convex problem

$$\min_x \ f(x) \ \text{ subject to } \ x \in C$$

and differentiable f, a feasible point x is optimal if and only if

$$\nabla f(x)^T (y - x) \geq 0 \quad \text{for all } y \in C$$



This is called the first-order condition for optimality

In words: all feasible directions from x are aligned with gradient $\nabla f(x)$

Important special case: if $C = \mathbb{R}^n$ (unconstrained optimization), then optimality condition reduces to familiar $\nabla f(x) = 0$

Example: quadratic minimization

Consider minimizing the quadratic function

$$f(x) = \frac{1}{2}x^T Qx + b^T x + c$$

where $Q \succeq 0$. The first-order condition says that solution satisfies

$$\nabla f(x) = Qx + b = 0$$

Cases:

- if $Q \succ 0$, then there is a unique solution $x = -Q^{-1}b$
- if Q is singular and $b \notin \mathrm{col}(Q)$, then there is no solution (i.e., $\min_x \ f(x) = -\infty$)
- if Q is singular and $b \in \mathrm{col}(Q)$, then there are infinitely many solutions

  $$x = -Q^+ b + z, \quad z \in \mathrm{null}(Q)$$

  where Q^+ is the pseudoinverse of Q

Example: projection onto a convex set

Consider projection onto convex set C:

$$\min_x \ \frac{1}{2}\|a - x\|_2^2 \ \text{ subject to } \ x \in C$$

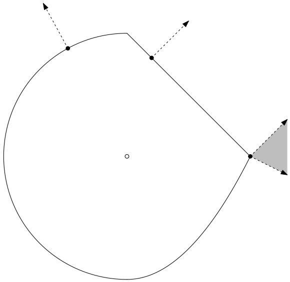
First-order optimality condition says that the solution x satisfies

$$\nabla f(x)^T(y - x) = (x - a)^T(y - x) \geq 0 \quad \text{for all } y \in C$$

Equivalently, this says that

$$a - x \in \mathcal{N}_C(x)$$

where recall $\mathcal{N}_C(x)$ is the normal cone to C at x



Partial optimization

Reminder: $g(x) = \min_{y \in C} \; f(x,y)$ is convex in x, provided that f is convex in (x,y) and C is a convex set

Therefore we can always partially optimize a convex problem and retain convexity

E.g., if we decompose $x = (x_1, x_2) \in \mathbb{R}^{n_1+n_2}$, then

$$\begin{array}{ll} \displaystyle\min_{x_1,x_2} & f(x_1,x_2) \\ \text{subject to} & g_1(x_1) \leq 0 \\ & g_2(x_2) \leq 0 \end{array} \quad \Longleftrightarrow \quad \begin{array}{ll} \displaystyle\min_{x_1} & \tilde{f}(x_1) \\ \text{subject to} & g_1(x_1) \leq 0 \end{array}$$

where $\tilde{f}(x_1) = \min\{f(x_1,x_2) : g_2(x_2) \leq 0\}$. The right problem is convex if the left problem is

Example: hinge form of SVMs

Recall the SVM problem

$$\begin{aligned}\min_{\beta,\beta_0,\xi} \quad & \frac{1}{2}\|\beta\|_2^2 + C\sum_{i=1}^n \xi_i \\ \text{subject to} \quad & \xi_i \geq 0,\ y_i(x_i^T\beta + \beta_0) \geq 1 - \xi_i,\ i = 1, \ldots n\end{aligned}$$

Rewrite the constraints as $\xi_i \geq \max\{0, 1 - y_i(x_i^T\beta + \beta_0)\}$. Indeed we can argue that we have $=$ at solution

Therefore plugging in for optimal ξ gives the hinge form of SVMs:

$$\min_{\beta,\beta_0} \ \frac{1}{2}\|\beta\|_2^2 + C\sum_{i=1}^n \big[1 - y_i(x_i^T\beta + \beta_0)\big]_+$$

where $a_+ = \max\{0, a\}$ is called the hinge function

Transformations and change of variables

If $h : \mathbb{R} \to \mathbb{R}$ is a monotone increasing transformation, then

$$\begin{aligned} &\min_x \; f(x) \;\; \text{subject to} \;\; x \in C \\ \iff &\min_x \; h(f(x)) \;\; \text{subject to} \;\; x \in C \end{aligned}$$

For example, maximizing log likelihood instead of maximizing likelihood

If $\phi : \mathbb{R}^n \to \mathbb{R}^m$ is one-to-one, and its image covers feasible set C, then we can change variables in an optimization problem:

$$\begin{aligned} &\min_x \; f(x) \;\; \text{subject to} \;\; x \in C \\ \iff &\min_y \; f(\phi(y)) \;\; \text{subject to} \;\; \phi(y) \in C \end{aligned}$$

Introducing slack variables

Simplifying inequality constraints. Given the problem

$$\begin{aligned} \min_{x} \quad & f(x) \\ \text{subject to} \quad & g_i(x) \leq 0, \ i = 1, \ldots m \\ & Ax = b \end{aligned}$$

we can transform the inequality constraints via

$$\begin{aligned} \min_{x,s} \quad & f(x) \\ \text{subject to} \quad & s_i \geq 0, \ i = 1, \ldots m \\ & g_i(x) + s_i = 0, \ i = 1, \ldots m \\ & Ax = b \end{aligned}$$

Note: this is no longer convex unless g_i, $i = 1, \ldots, n$ are affine

Example: SVM derivation (hard margin constraint)

The hard-margin SVM problem is originally cast as:

$$\begin{aligned} \min_{\beta,\beta_0} \quad & \frac{1}{2}\|\beta\|_2^2 \\ \text{subject to} \quad & y_i(x_i^T\beta + \beta_0) \geq 1,\ i = 1, \ldots n \end{aligned}$$

Introducing slack variables ξ_i, we get

$$\begin{aligned} \min_{\beta,\beta_0,\xi} \quad & \frac{1}{2}\|\beta\|_2^2 \\ \text{subject to} \quad & \xi_i \geq 0,\ y_i(x_i^T\beta + \beta_0) = 1 - \xi_i,\ i = 1, \ldots n \end{aligned}$$

Relaxing nonaffine equality constraints

Given an optimization problem

$$\min_x \ f(x) \ \text{ subject to } \ x \in C$$

we can always take an enlarged constraint set $\tilde{C} \supseteq C$ and consider

$$\min_x \ f(x) \ \text{ subject to } \ x \in \tilde{C}$$

This is called a relaxation and its optimal value is always smaller or equal to that of the original problem

Relaxing nonaffine equality constraints

Given an optimization problem

$$\min_x \; f(x) \;\; \text{subject to} \;\; x \in C$$

we can always take an enlarged constraint set $\tilde{C} \supseteq C$ and consider

$$\min_x \; f(x) \;\; \text{subject to} \;\; x \in \tilde{C}$$

This is called a relaxation and its optimal value is always smaller or equal to that of the original problem

Important special case: relaxing nonaffine equality constraints, i.e.,

$$h_j(x) = 0, \; j = 1, \ldots r$$

where h_j, $j = 1, \ldots r$ are convex but nonaffine, are replaced with

$$h_j(x) \leq 0, \; j = 1, \ldots r$$

Example: principal components analysis

Given $X \in \mathbb{R}^{n \times p}$, consider the low rank approximation problem:

$$\min_R \|X - R\|_F^2 \ \text{ subject to } \ \text{rank}(R) = k$$

Here $\|A\|_F^2 = \sum_{i=1}^n \sum_{j=1}^p A_{ij}^2$, the entrywise squared ℓ_2 norm, and $\text{rank}(A)$ denotes the rank of A. Also called principal components analysis or PCA problem.

This problem is not convex. Why?

Example: principal components analysis

Given $X \in \mathbb{R}^{n \times p}$, consider the low rank approximation problem:

$$\min_R \; \|X - R\|_F^2 \;\; \text{subject to} \;\; \mathrm{rank}(R) = k$$

Here $\|A\|_F^2 = \sum_{i=1}^n \sum_{j=1}^p A_{ij}^2$, the entrywise squared ℓ_2 norm, and $\mathrm{rank}(A)$ denotes the rank of A. Also called principal components analysis or PCA problem.

This problem is not convex. Why?

Given $X = UDV^T$, singular value decomposition or SVD, the solution is

$$R = U_k D_k V_k^T$$

where U_k, V_k are the first k columns of U, V and D_k is the first k diagonal elements of D. I.e., R is reconstruction of X from its first k principal components

We can recast the PCA problem in a convex form. First rewrite as

$$\begin{aligned}&\min_{Z \in \mathbb{S}^p} \|X - XZ\|_F^2 \ \text{ subject to } \ \text{rank}(Z) = k, \ Z \text{ is a projection} \\ \iff &\max_{Z \in \mathbb{S}^p} \text{tr}(SZ) \ \text{ subject to } \ \text{rank}(Z) = k, \ Z \text{ is a projection}\end{aligned}$$

where $S = X^T X$. Hence constraint set is the nonconvex set

$$C = \Big\{ Z \in \mathbb{S}^p : \lambda_i(Z) \in \{0, 1\}, \ i = 1, \ldots p, \ \text{tr}(Z) = k \}$$

where $\lambda_i(Z)$, $i = 1, \ldots n$ are the eigenvalues of Z. Solution in this formulation is

$$Z = V_k V_k^T$$

where V_k gives first k columns of V

Now consider relaxing constraint set to $\mathcal{F}_k = \mathrm{conv}(C)$, its convex hull. Note

$$\begin{aligned}\mathcal{F}_k &= \{Z \in \mathbb{S}^p : \lambda_i(Z) \in [0,1],\ i = 1, \ldots p,\ \mathrm{tr}(Z) = k\} \\ &= \{Z \in \mathbb{S}^p : 0 \preceq Z \preceq I,\ \mathrm{tr}(Z) = k\}\end{aligned}$$

Recall this is called the Fantope of order k

Hence, the linear maximization over the Fantope, namely

$$\max_{Z \in \mathcal{F}_k} \ \mathrm{tr}(SZ)$$

is convex. Remarkably, this is equivalent to the nonconvex PCA problem (admits the same solution)!

(Famous result: Fan (1949), "On a theorem of Weyl conerning eigenvalues of linear transformations")