Sparse Optimization
Lecture 1: Review of Convex Optimization

Instructor: Wotao Yin

July 2013

online discussions on piazza.com

Those who complete this lecture will know

- convex optimization background
- various standard concepts and terminology
- reformulating ℓ_1 optimization and its optimality conditions

Resources for convex optimization

- Book: Convex Analysis by T. Rockafellar
- Book: Convex Optimization by S. Boyd and L. Vandenberge, along with online videos and slides
- Book: Introductory Lectures on Convex Optimization: A Basic Course by Y. Nesterov
- A large number of online lecture slides, notes, and videos online

Review: mathematical optimization

Formulation

$$\begin{aligned} \underset{\mathbf{x}}{\text{minimize}}\ & f_0(\mathbf{x}) \\ \text{subject to}\ & f_i(\mathbf{x}) \leq 0,\ i = 1, \ldots, m, \\ & h_j(\mathbf{x}) = 0,\ j = 1, \ldots, p. \end{aligned}$$

- **decision variables**: $\mathbf{x} = (x_1, \ldots, x_n)$
- **objective function**: $f_0 : \mathbb{R}^n \to \mathbb{R}$
- functions defining **inequality constraints**: $f_i : \mathbb{R}^n \to \mathbb{R},\ i = 1, \ldots, m$
- functions defining **equality constraints**: $h_j : \mathbb{R}^n \to \mathbb{R},\ j = 1, \ldots, p$

Terminology

- **feasible solutions**: all points x satisfying the constraints $f_i(\mathbf{x}) \leq 0$ $(i = 1, \ldots, m)$ and $h_j(\mathbf{x}) = 0$ $(j = 1, \ldots, p)$.
- **feasible set**: the set of all feasible solutions, often denoted by $\mathcal{X}$.
- **(global) (optimal) solution**: feasible solution $\mathbf{x}^*$ that achieves the minimum objective value among all feasible solutions.
- **local (optimal) solution**: feasible solution $\mathbf{x}^*$ that achieves the minimal objective value among a neighborhood around $\mathbf{x}^*$, say, the set $\{\mathbf{x} : \|\mathbf{x} - \mathbf{x}^*\| \leq \delta\} \cap \mathcal{X}$ for some $\delta > 0$

Some examples

- Find two nonnegative numbers whose product up to 9 and so that the sum of the two numbers is a maximum.
- Find the largest area a rectangular region provided that its perimeter is not great than 100.
- Given a sequences of nonnegative numbers, find a start point and an end point so that the partial sum of the sequence between the two points is a maximum.

Solving optimization problems

In general, everything is optimization, but optimization problems are generally *not solvable*, even by the most powerful computers.

Some classes of problems can be solved efficiently and reliably, for example:

- least-squares problems
- linear programming problems
- quadratic programming problems
- convex optimization problems
- a subclass of network-flow problems
- submodular function minimization
  (.... more, but not much more...)
- some sparse optimization problems

Least squares

$$\underset{\mathbf{x}}{\text{minimize}} \quad \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2$$

- analytic solution $\mathbf{x}^* = (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{b}$ if $\mathbf{A}$ has independent columns
- reliable and efficient algorithms and software packages
- computation time proportional to n^2k ($\mathbf{A} \in \mathbb{R}^{k\times n}$), less if structured
- a mature technology (unless $\mathbf{A}$ is huge and/or distributed)

Linear programming (LP)

$$\begin{aligned} &\underset{\mathbf{x}}{\text{minimize}}\ \mathbf{c}^T\mathbf{x} \\ &\text{subject to}\ \mathbf{a}_i^T\mathbf{x} \leq b_i,\ i = 1, \ldots, m \end{aligned}$$

- no analytic formula for solutions
- reliable and efficient algorithms and software packages
- computation time proportional to $n^2 m$ if $m \geq n$, less with structured data
- a mature technology
- a few standard tricks used to convert problems (with ℓ_1 or ℓ_∞, piecewise linear functions) into linear programs

Convex optimization

$$\begin{aligned} &\underset{\mathbf{x}}{\text{minimize}} && f_0(\mathbf{x}) \\ &\text{subject to} && f_i(\mathbf{x}) \leq 0, \ i = 1, \ldots, m, \\ &&& \mathbf{A}\mathbf{x} = \mathbf{b}. \end{aligned}$$

where objective and constraint functions are *convex*, i.e.,

$$f_i(\theta \mathbf{x}^1 + (1-\theta)\mathbf{x}^2) \leq \theta f_i(\mathbf{x}^1) + (1-\theta) f_i(\mathbf{x}^2)$$

for all $i = 0, 1, \ldots, m$, $\theta \in (0, 1)$ and $\mathbf{x}^1, \mathbf{x}^2 \in \mathrm{dom} f_i$.

- no analytic solution
- relatively reliable and efficient algorithms and software packages
- computation time (roughly) proportional to $\max\{n^3, n^2 m, F\}$, where F is cost of evaluating f_i's and their first and second derivatives.
- almost a technology

Least-squares and linear programs are special convex programs.

Non-convex optimization problems

General optimization problems are non-convex

$$\begin{aligned} &\underset{\mathbf{x}}{\text{minimize}} \; f_0(\mathbf{x}) \\ &\text{subject to } f_i(\mathbf{x}) \leq 0, \; i = 1, \ldots, m \end{aligned}$$

Local optimization methods

- find a solution which minimizes f_0 among feasible solutions near it
- fast and handle large problems
- require initial guess
- provide no information about the distance to global optima

Global optimization methods

- find the global solution
- worst-case complexity grows exponentially with problem size.

These methods are often based on solving convex subproblems.

Brief history of convex optimization

theory (convex analysis): 1900–1970s

algorithms

- 1947: simplex algorithm for linear programming (Dantzig)
- 1960s: early interior-point methods (Fiacco & McCormick, Dikin, . . .)
- 1970s: ellipsoid method and other subgradient methods
- 1980s: polynomial-time interior-point methods for linear programming (Karmarkar 1984)
- late 1980s-2000s: polynomial-time interior-point methods for nonlinear convex optimization (Nesterov & Nemirovski 1994)
- recently: revived interests in first-order (gradient-based) algorithms, solving big-data problems

applications

- before 1990: mostly in operations research; few in engineering
- since 1990: many new applications in engineering (control, signal processing, communications, circuit design, . . .); new problem classes (semidefinite and second-order cone programming, robust optimization, sparse optimization)

Convex set

A set $\mathcal{C}$ is called **convex** if the segment between any two points in $\mathcal{C}$ lies entirely in $\mathcal{C}$.

Formally, $\mathcal{C}$ is convex if for any $\mathbf{x}^1, \mathbf{x}^2 \in \mathcal{C}$ and $\theta \in (0, 1)$, we have $\theta\mathbf{x}^1 + (1-\theta)\mathbf{x}^2 \in \mathcal{C}$.

Examples:

- **Euclidean balls:** $B(\mathbf{x}_c, r) = \{\mathbf{x} : \|\mathbf{x} - \mathbf{x}_c\|_2 \leq r\}$
- **ellipsoid:** $\{\mathbf{x} : (\mathbf{x} - \mathbf{x}_c)^T \mathbf{P}^{-1}(\mathbf{x} - \mathbf{x}_c) \leq 1\}$ with $\mathbf{P}$ being symmetric positive definite
- **polyhedra:** $\{\mathbf{x} : \mathbf{A}\mathbf{x} \leq \mathbf{b}, \quad \mathbf{C}\mathbf{x} = \mathbf{d}\}$ with $\mathbf{A} \in \mathbb{R}^{m \times n}$, $\mathbf{C} \in \mathbb{R}^{p \times n}$
- several operations preserving convexity: intersection; affine function; perspective function; linear-fractional functions.

In most time, recognizing a convex set is not a problem.

Convex functions

A function $f : \mathbb{R}^n \to \mathbb{R}$ is **convex** if $\mathrm{dom} f$ is convex and for any $\mathbf{x}^1, \mathbf{x}^2 \in \mathrm{dom} f$ and $\theta \in (0, 1)$, we have

$$f(\theta \mathbf{x}^1 + (1-\theta)\mathbf{x}^2) \leq \theta f(\mathbf{x}^1) + (1-\theta) f(\mathbf{x}^2).$$

f is **concave** if $(-f)$ is convex.

f is **strictly convex** if $\mathrm{dom} f$ is convex and

$$f(\theta \mathbf{x}^1 + (1-\theta)\mathbf{x}^2) < \theta f(\mathbf{x}^1) + (1-\theta) f(\mathbf{x}^2).$$

Examples of convex functions

Examples in $\mathbb{R}^n$

- affine function $f(\mathbf{x}) = \mathbf{a}^T\mathbf{x} + b$
- norms: $\|\mathbf{x}\|_p = (\sum_{i=1}^n |x_i|^p)^{1/p}$ for $p \geq 1$; $\|\mathbf{x}\|_\infty = \max_i |x_i|$.

Examples in $\mathbb{R}^{m\times n}$

- affine function

$$f(\mathbf{X}) = \mathrm{tr}(\mathbf{A}^T\mathbf{X}) + b = \sum_{i=1}^m \sum_{j=1}^n A_{ij}X_{ij} + b$$

- spectral norm (maximum singular value)

$$f(\mathbf{X}) = \|\mathbf{X}\|_2 = \sigma_{\max}(\mathbf{X}) = (\lambda_{\max}(\mathbf{X}^T\mathbf{X}))^{1/2}$$

- nuclear norm

$$f(\mathbf{X}) = \|\mathbf{X}\|_* = \sum_{i=1}^{\min\{m,n\}} \sigma_i$$

Terminology

- **extended value**: f may take on value $+\infty$, reduce the need of $\mathrm{dom} f$
- **proper**: exists $\mathbf{x}$ so that $f(\mathbf{x})$ is finite
- **lower semi-continuous** (LSC): $\liminf_{\mathbf{x}\to\mathbf{x}_0} f(\mathbf{x}) \geq f(\mathbf{x}_0)$
- **closed**: f has a *closed* epigraph

$$\mathrm{epi} f = \{(\mathbf{x}, \mu) : \mu \in \mathbb{R}, \mu \geq f(\mathbf{x})\}$$

- Lemma: a proper convex function is closed if and only if its is LSC
- **subdifferential**

$$\partial f(\mathbf{x}) = \{\mathbf{p} : f(\mathbf{y}) \geq f(\mathbf{x}) + \langle \mathbf{p}, \mathbf{y} - \mathbf{x}\rangle \ \forall \mathbf{y}\}$$

  - each $\mathbf{p} \in \partial f(\mathbf{x})$ is called a subgradient
  - if $f \in C^1$ near $\mathbf{x}$, then $\partial f(\mathbf{x}) = \{\nabla f(\mathbf{x})\}$

First-order condition

f is differentiable if the derivative

$$\nabla f(\mathbf{x}) = \left[\frac{\partial f(\mathbf{x})}{\partial x_1}, \frac{\partial f(\mathbf{x})}{\partial x_2}, \ldots, \frac{\partial f(\mathbf{x})}{\partial x_n}\right]^T$$

exists at every $\mathbf{x} \in \mathrm{dom} f$.

first-order condition: differentiable f with convex domain is convex iff

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla f(\mathbf{x})^T(\mathbf{y} - \mathbf{x}) \text{ for all } \mathbf{x}, \mathbf{y} \in \mathrm{dom} f$$

first-order condition: subdifferentiable f with convex domain is convex iff

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \mathbf{p}^T(\mathbf{y} - \mathbf{x}) \text{ for all } \mathbf{x}, \mathbf{y} \in \mathrm{dom} f,\ \mathbf{p} \in \partial f(\mathbf{x})$$

first-order optimality condition: $\mathbf{x}^* \in \arg\min f(\mathbf{x}) \iff 0 \in \partial f(\mathbf{x}^*)$

Second-order condition

f is twice differentiable if Hessian $\nabla^2 f(\mathbf{x}) \in \mathcal{S}^n$ defined by

$$\nabla^2 f(\mathbf{x})_{ij} = \frac{\partial^2 f(\mathbf{x})}{\partial x_i \partial x_j}, \quad i, j = 1, \ldots, n,$$

exists at every $\mathbf{x} \in \mathrm{dom} f$.

second-order condition: twice differentiable f with convex domain is convex iff

$$\nabla^2 f(\mathbf{x}) \succeq 0, \text{ for all } \mathbf{x} \in \mathrm{dom} f.$$

Furthermore, if $\nabla^2 f(\mathbf{x}) \succ 0$ for all $\mathbf{x} \in \mathrm{dom} f$, then f is strictly convex.

Very useful in general convex optimization but not so in sparse optimization

Convex optimization formulation

Standard-form convex optimization problem

$$\begin{aligned} &\underset{\mathbf{x}}{\text{minimize}} \; f_0(\mathbf{x}) \\ &\text{subject to } f_i(\mathbf{x}) \leq 0, \; i = 1, \ldots, m, \\ &\mathbf{A}\mathbf{x} = \mathbf{b}. \end{aligned}$$

- the feasible set of a convex optimization problem is convex.
- $f_0, f_1, \ldots, f_m$ are convex; equality constraints are affine.

Local and global solutions

Theorem

Any local solution of a convex problem is a global solution.

Proof.

Suppose that $\mathbf{x}$ is a local solution and $\mathbf{y}$ is a global solution and that

$$f_0(\mathbf{y}) < f_0(\mathbf{x}).$$

Consider $\mathbf{z} = \theta\mathbf{y} + (1-\theta)\mathbf{x}$. Since

$$f_0(\mathbf{z}) \leq \theta f_0(\mathbf{x}) + (1-\theta) f_0(\mathbf{y}) < f_0(\mathbf{x})$$

for any $\theta \in (0,1)$ and $\|\mathbf{x}-\mathbf{z}\|$ can be arbitrary small, $\mathbf{x}$ cannot be a local solution. □

Optimality criterion for differentiable f_0

Since the feasible set is convex and

$$f_0(\mathbf{y}) \geq f_0(\mathbf{x}) + \nabla f_0(\mathbf{x})^T(\mathbf{y} - \mathbf{x}),$$

$\mathbf{x}$ is optimal iff it is feasible and

$$\nabla f_0(\mathbf{x})^T(\mathbf{y} - \mathbf{x}) \geq 0 \text{ for all feasible } \mathbf{y}.$$

- **unconstrained problem:** $\mathbf{x}$ is optimal if and only if

$$\mathbf{x} \in \mathrm{dom} f_0, \quad \nabla f_0(\mathbf{x}) = 0$$

- **equality constrained problem:**

$$\underset{\mathbf{x}}{\mathrm{minimize}}\, f_0(\mathbf{x}) \text{ subject to } \mathbf{A}\mathbf{x} = \mathbf{b}$$

$\mathbf{x}$ is optimal if and only if their exist a vector ν such that

$$\mathbf{x} \in \mathrm{dom} f_0, \quad \mathbf{A}\mathbf{x} = \mathbf{b}, \quad \nabla f_0(\mathbf{x}) + \mathbf{A}^T \nu = 0$$

- **minimization over nonnegative orthant**

$$\underset{\mathbf{x}}{\mathrm{minimize}}\, f_0(\mathbf{x}) \text{ subject to } \mathbf{x} \geq 0$$

$\mathbf{x}$ is optimal if and only if

$$\mathbf{x} \in \mathrm{dom} f_0, \quad \mathbf{x} \geq 0, \quad \begin{cases} \nabla f_0(\mathbf{x})_i \geq 0 & x_i = 0 \\ \nabla f_0(\mathbf{x})_i = 0 & x_i > 0 \end{cases}$$

Unconstrained problem with nondifferentiable f_0

$\mathbf{g}$ is a **subgradient** of a convex function f at $\mathbf{x} \in \mathrm{dom} f$ if

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \mathbf{g}^T(\mathbf{y} - \mathbf{x}), \quad \forall \mathbf{y} \in \mathrm{dom} f.$$

the **subdifferential** $\partial f(\mathbf{x})$ of f at $\mathbf{x}$ is the set of all subgradients:

$$\partial f(\mathbf{x}) = \{\mathbf{g} : \mathbf{g}^T(\mathbf{y} - \mathbf{x}) \leq f(\mathbf{y}) - f(\mathbf{x}) \ \forall \mathbf{y} \in \mathrm{dom} f\}$$

$\mathbf{x}^*$ minimizes $f_0(\mathbf{x})$ if and only if

$$0 \in \partial f_0(\mathbf{x}^*)$$

Optimality criteria in the general case

Standard form problem (not necessarily convex)

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

domain $\mathcal{D}$, optimal value p^*

Lagrangian: $L : \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^p \to \mathbb{R}$ with $\mathrm{dom} L = \mathcal{D} \times \mathbb{R}^m \times \mathbb{R}^p$,

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

- λ_i is Lagrange multiplier associated with $f_i(\mathbf{x}) \leq 0$
- ν_j is Lagrange multiplier associated with $h_j(\mathbf{x}) = 0$

Lagrange dual function

Lagrange dual function: $g : \mathbb{R}^m \times \mathbb{R}^p \to \mathbb{R}$,

$$\begin{aligned} g(\lambda, \nu) &= \inf_{\mathbf{x} \in \mathcal{D}} L(\mathbf{x}, \lambda, \nu) \\ &= \inf_{\mathbf{x} \in \mathcal{D}} \left(f_0(\mathbf{x}) + \sum_{i=1}^{m} \lambda_i f_i(\mathbf{x}) + \sum_{j=1}^{p} \nu_j h_j(\mathbf{x}) \right) \end{aligned}$$

g is concave, can be $-\infty$ for some λ, ν.

Lower bound property: if $\lambda \succeq 0$, then $g(\lambda, \nu) \leq p^*$.

Dual problem

Lagrange dual problem

$$\begin{aligned} &\underset{\lambda,\nu}{\text{maximize}} \; g(\lambda,\nu) \\ &\text{subject to } \lambda \succeq 0 \end{aligned}$$

- finds the best lower bound n p^*
- a convex optimization problem; optimal value denoted d^*
- λ, ν are dual feasible if $\lambda \succeq 0$, $(\lambda,\nu) \in \mathrm{dom} g$

Strong duality: $d^* = p^*$

- does not hold in general
- (usually) holds for convex problems
- conditions that guarantee strong duality in convex problems are called **constraint qualifications**

Slater's constraint qualification

Strong duality holds for a convex problem

$$\begin{aligned} &\underset{x}{\text{minimize}} \ f_0(\mathbf{x}) \\ &\text{subject to } f_i(\mathbf{x}) \leq 0, \ i = 1, \ldots, m \\ &\mathbf{A}\mathbf{x} = \mathbf{b} \end{aligned}$$

if it is strictly feasible, i.e.,

$$\exists \mathbf{x} \in \text{int}\mathcal{D} : \quad f_i(\mathbf{x}) < 0, \ i = 1, \ldots, m, \quad \mathbf{A}\mathbf{x} = \mathbf{b}$$

- also guarantees that the dual optimum is attained (if $p^* > -\infty$)
- linear inequalities do not need to hold with strict inequality
- there are many other types of constraint qualifications
- some non-convex optimization problems may have strong duality

Complementary slackness

Assume strong duality holds, x^* is primal optimal, and (λ^*, ν^*) is dual optimal

$$\begin{aligned} f_0(\mathbf{x}^*) = g(\lambda^*, \nu^*) &= \inf_{\mathbf{x}} \left(f_0(\mathbf{x}) + \sum_{i=1}^{m} \lambda_i^* f_i(\mathbf{x}) + \sum_{j=1}^{p} \nu_j^* h_j(\mathbf{x}) \right) \\ &\leq f_0(\mathbf{x}^*) + \sum_{i=1}^{m} \lambda_i^* f_i(\mathbf{x}^*) + \sum_{j=1}^{p} \nu_j^* h_j(\mathbf{x}^*) \\ &\leq f_0(\mathbf{x}^*) \end{aligned}$$

- $\mathbf{x}^*$ minimizes $L(\mathbf{x}, \lambda^*, \nu^*)$
- $\lambda_i^* f_i(\mathbf{x}^*) = 0$ for $i = 1, \ldots, m$ (complementary slackness)

Karush-Kuhn-Tucker (KKT) conditions

KKT conditions for a problem with differentiable f_i, h_j:

- primal constraints: $f_i(\mathbf{x}) \leq 0,\ i = 1, \ldots, m,\ h_j(\mathbf{x}) = 0,\ j = 1, \ldots, p$
- dual constraints: $\lambda \succeq 0$
- complementary slackness: $\lambda_i f_i(\mathbf{x}) = 0,\ i = 1, \ldots, m$
- gradient of Lagrangian with respect to $\mathbf{x}$ vanishes:

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

If $\tilde{\mathbf{x}}$, $\tilde{\lambda}$, $\tilde{\nu}$ satisfy KKT for a convex problem, then they are optimal.

Exercise: constrained/unconstrained ℓ_1 problem

Consider two ℓ_1 problems

$$\begin{aligned} \underset{\mathbf{x}}{\text{minimize}} \ & \|\mathbf{x}\|_1 \\ \text{subject to } & \mathbf{A}\mathbf{x} = \mathbf{b} \\ & \mathbf{l} \leq \mathbf{x} \leq \mathbf{u} \end{aligned}$$

and

$$\begin{aligned} \underset{\mathbf{x}}{\text{minimize}} \ & \|\mathbf{x}\|_1 + \frac{\lambda}{2}\|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2 \\ \text{subject to } & \mathbf{l} \leq \mathbf{x} \leq \mathbf{u} \end{aligned}$$

Exercises: derive their

- LP or QP formulations
- Lagrange dual problems
- KKT conditions

Exercise: total variation problem*

The discrete total variation of a vector $\mathbf{x} \in \mathbb{R}^n$ is

$$\mathrm{TV}(\mathbf{x}) = \sum_{i=1}^{n-1} |x_{i+1} - x_i|.$$

Consider problem

$$\begin{aligned} &\underset{\mathbf{x}}{\text{minimize}} \;\; \mathrm{TV}(\mathbf{x}) + \frac{\lambda}{2}\|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2 \\ &\text{subject to } \mathbf{l} \leq \mathbf{x} \leq \mathbf{u} \end{aligned}$$

Exercises: derive its

- SOCP formulation (refer to Sec.4.2.2 of Boyd&Vandenberghe)
- Lagrange dual problem
- KKT conditions