Convex Optimization

Sonia Reilly

October 15, 2023

These notes are based on Boyd and Vandenberghe's *Convex Optimization*, David Rosenberg's "Extreme Abridgement" of Boyd and Vandenberghe, and Paul Beckman and Dave Connelly's orals notes.

1 Convexity

1.1 Notation

- $\mathbb{R}_+$ represents nonnegative reals, $\mathbb{R}_{++}$ represents positive reals
- $a \succeq b$ means componentwise inequality for a vector or positive semidefiniteness for a matrix

1.2 Convex sets

- $C \subseteq \mathbb{R}^n$ is affine if for any $x_1, x_2 \in C$, $\theta \in \mathbb{R}$, $\theta x_1 + (1-\theta)x_2 \in C$ (i.e. line through any two points in C is in C). Affine sets are points, lines, planes, hyperplanes.
- C is convex if for any $x_1, x_2 \in C$, $0 \leq \theta \leq 1$, then

  $$\theta x_1 + (1-\theta)x_2 \in C,$$

  i.e. if the line segment between any two points in C is in C.
- If C, D are bounded, nonempty convex sets, then there is a hyperplane that separates them, i.e. $a^T x \leq b$ for one and $a^T x \geq b$ for the other.
- The convex hull of a set S is the set of all convex combinations $\theta_1 x_1 + \ldots + \theta_n x_n$, $\sum_i \theta_i = 1$, $\theta_i \geq 0$ of points $x_i \in S$.
- A cone is a set for which $x \in S \implies \theta x \in S$ for any $\theta \geq 0$. Two important cones are
  - the second order cone $\{(x,t) \in \mathbb{R}^{n+1} \mid ||x||_2 \leq t\}$
  - the cone of positive semidefinite matrices.
- Convexity is preserved under intersection and under transformation by an affine function.

1.3 Convex functions

- A function is convex if its domain is convex and for any x, y in its domain and $\theta \in [0, 1]$,

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

  This is equivalent to the epigraph (region above the function) being convex.

- A function f is concave if $-f$ is convex.

- If f is differentiable, f is convex iff for all x, y in the domain,

$$f(x) + \nabla f(x)^T(y-x) \leq f(y) \quad \text{or} \quad \nabla^2 \succeq 0,$$

  i.e. that the first order Taylor expansion lies under the graph or that the Hessian is positive semidefinite.

- Convex function examples:
  - exponential e^{ax} on $\mathbb{R}$
  - powers x^a on $\mathbb{R}$ except $0 < a < 1$ which are concave
  - $\log x$ on $\mathbb{R}_{++}$
  - any norm on $\mathbb{R}^n$
  - $\log\det X$ is concave on S^n_{++} (positive definite matrices)

- Operations that preserve convexity:
  - weighted sums with positive weights
  - composition with an affine mapping $g(x) = f(Ax+b)$
  - pointwise maximum of convex functions
  - composition $f = h \circ g$ of a nondecreasing convex h and a convex g

1.4 Convex optimization problems

- A standard convex optimization problem has the form

$$\begin{array}{rl} \text{minimize} & f_0(x) \\ \text{subject to} & f_i(x) \leq 0, \quad i = 1, \dots m \\ & a_i^T x = b_i, \quad i = 1, \dots p \end{array}$$

  where $f_0, \dots f_m$ are convex functions. This adds three requirements over a general optimization problem: the objective function is convex, the inequality constraint functions are convex, and the equality constraints are affine. The result is that the feasible region is convex.

- *A local optimizer for a convex optimization problem is a global optimizer.*

- Proof: Consider some x that is locally optimal in an ϵ-neighborhood and not globally optimal, i.e. there is a feasible y with $f(y) < f(x)$. Then there is a z along the line between x and y that is in the ϵ-neighborhood and also feasible, because the feasible region is convex. By convexity of the objective function, $f(z) < f(x)$, contradicting local optimality of x.

- A point x is optimal iff
$$\nabla f_0(x)^T(y-x) \geq 0$$
for all feasible y, or $\nabla f_0(x)^T = 0$ if there are no constraints.

- A linear program (LP) is a convex optimization problem of the form
$$\begin{array}{rl} \text{minimize} & c^T x \\ \text{subject to} & Ax = b \\ & x \succeq 0. \end{array}$$

- A semidefinite program (SDP) is a convex optimization problem of the form
$$\begin{array}{rl} \text{minimize} & \mathrm{tr}(CX) \\ \text{subject to} & \mathrm{tr}(A_i X) = b_i, \ \ i = 1, \dots p \\ & X \succeq 0. \end{array}$$

2 Duality

2.1 Lagrange dual function

- For a generic optimization problem with constraints $g_i(x) \leq 0, h_i(x) = 0$, the Lagrangian is given by
$$L(x, \lambda, \nu) := f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) + \sum_{i=1}^{p} \nu_i h_i(x).$$
The vectors λ and ν are called dual variables or Lagrange multipliers.

- The supremum of the Lagrangian over $\lambda \succeq 0, \nu$ is the objective function f_0 for feasible x and ∞ otherwise. So the optimal value can be written as
$$p^* = \inf_x \sup_{\lambda \succeq 0, \nu} L(x, \lambda, \nu).$$

- The Lagrange dual function is given by
$$g(\lambda, \nu) := \inf_{x \in \mathcal{D}} L(x, \lambda, \nu)$$

- The Lagrange dual function is concave (even when the problem is not convex) because it is the pointwise minimum of affine functions.

2.2 Lagrange dual problem

- The Lagrange dual function is a lower bound for the minimum value p^*, because for any feasible y, $L(y, \lambda, \nu) \leq f(y)$, so
$$g(\lambda, \nu) = \inf_{x \in \mathcal{D}} L(x, \lambda, \nu) \leq L(y, \lambda, \nu) \leq f(y)$$
for all feasible y, so in particular $g(\lambda, \nu) \leq p^* = f(x^*)$.

- Since $g(\lambda, \nu)$ is a lower bound on the minimum, we are interested in the best possible lower bound. This is the Lagrange dual problem,

$$\begin{array}{rl} \text{maximize} & g(\lambda, \nu) \\ \text{subject to} & \lambda \succeq 0. \end{array}$$

- Some (λ, ν) are dual feasible if $g(\lambda, \nu) > -\infty$. The λ is called the dual slack variable.
- The dual problem is convex even if the primal is not.
- Weak duality: $d^* \leq p^*$, as we have shown. The duality gap is $p^* - d^*$.
- Strong duality: $d^* = p^*$. Strong duality holds for convex problems that satisfy Slater's condition, i.e. that there is a strictly feasible point (one that satisfies the inequality constraints strictly).

2.3 Examples of dual problems

- For the standard LP, the Lagrange dual function is

$$g(\lambda, \nu) = \begin{cases} -b^T \nu & A^T \nu - \lambda + c = 0 \\ -\infty & \text{otherwise.} \end{cases}$$

Therefore the dual problem is

$$\begin{array}{rl} \text{maximize} & -b^T \nu \\ \text{subject to} & A^T \nu - \lambda + c = 0 \\ & \lambda \succeq 0, \end{array}$$

which can easily be written (by removing λ) as another LP.

- For the standard SDP, the Lagrange dual function is

$$g(\lambda, \nu) = \begin{cases} -b^T \nu & \sum_{i=1}^p A_i \nu_i - \Lambda + C = 0 \\ -\infty & \text{otherwise} \end{cases}$$

and the dual problem is

$$\begin{array}{rl} \text{maximize} & -b^T \nu \\ \text{subject to} & \displaystyle\sum_{i=1}^p A_i \nu_i - \Lambda + C = 0 \\ & \Lambda \succeq 0, \end{array}$$

which as above can be written as an SDP.

- The dual of the dual of an LP or SDP is the primal (easy to check).

3 Optimality conditions

3.1 Complementary slackness

- Let x^* and λ^*, ν^* be primal optimal and dual optimal, respectively, and assume that strong duality holds. Then

  $$\begin{aligned} f_0(x^*) &= g(\lambda^*, \nu^*) \\ &= \inf_x \left(f_0(x) + \sum_{i=1}^m \lambda_i^* f_i(x) + \sum_{i=1}^p \nu_i^* h_i(x) \right) \\ &\le f_0(x^*) + \sum_{i=1}^m \lambda_i^* f_i(x^*) + \sum_{i=1}^p \nu_i^* h_i(x^*) \\ &\le f_0(x^*), \end{aligned}$$

  by strong duality, the definition of the dual function, choosing $x = x^*$, and applying $\lambda_i^* \ge 0$, $f_i(x^*) \le 0$, and $h_i(x^*) = 0$. So both inequalities are inequalities, and we get conditions on x^*.

- The first condition,

  $$\sum_{i=1}^m \lambda_i^* f_i(x^*) = 0 \qquad \Longrightarrow \qquad \lambda_i^* = 0 \text{ or } f_i(x^*) = 0 \ \forall i,$$

  is called complementary slackness. It says that the i-th optimal Lagrange multiplier is 0 unless the i-th inequality constraint is active.

3.2 KKT optimality conditions

- The calculation above suggests the Karush-Kuhn-Tucker (KKT) conditions for optimality,

  $$\begin{aligned} \textbf{primal feasibility} \quad & f_i(x^*) \le 0, h_i(x^*) = 0 \\ \textbf{dual feasibility} \quad & \lambda_i^* \ge 0 \\ \textbf{complementary slackness} \quad & \lambda_i^* f_i(x^*) = 0 \\ \textbf{stationarity} \quad & \nabla f_0(x^*) + \sum_{i=1}^m \lambda_i^* \nabla f_i(x^*) + \sum_{i=1}^p \nu_i^* \nabla h_i(x^*) = 0. \end{aligned}$$

  For convex problems, any x^* and (λ^*, ν^*) that satisfy KKT are primal and dual optimal, with no duality gap. (Easy to verify by evaluating $g(\lambda^*, \nu^*)$). KKT is necessary but not sufficient in the nonconvex case.

- The stationarity condition comes from the fact that x^* minimizes $L(x, \lambda^*, \nu^*)$ (follows from calculation above), so the gradient must be 0.

4 Convex optimization methods

4.1 General descent methods

- A general descent method for minimizing a convex function has the form

  $$x_{k+1} = x_k + t_k \Delta x_k,$$

where we choose the step length t_k and the descent direction Δx_k so that $f(x_{k+1}) < f(x_k)$.

- To find a good step length, we often use backtracking line search, where we start by setting $t_k = 1$ and then update $t_k = \beta t_k$ until

$$f(x_k + t\Delta x_k) \le f(x_k) + \alpha t_k \nabla f(x_k)^T \Delta x_k,$$

for some constants $0 < \alpha < 0.5$ and $0 < \beta < 1$.

4.2 Gradient descent

- Gradient descent simply chooses $\Delta x_k = -\nabla f(x_k)$, which is guaranteed to locally be a descent direction:

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

- The convergence of gradient descent with backtracking line search is linear:

$$f(x_k) - p^* = c^k (f(x_0) - p^*)$$

for c a constant that depends on α, β.

4.3 Newton's method

- Newton's method comes from Taylor expanding

$$f(x + \Delta x) = f(x) + \nabla f(x)^T \Delta x + \frac{1}{2} \Delta x^T \nabla^2 f(x) \Delta x.$$

The second-order Taylor expansion is minimized when

$$\nabla^2 f(x) \Delta x = -\nabla f(x).$$

- Newton's method has quadratic convergence

$$f(x_{k+1}) - p^* = c\,(f(x_k) - p^*)^2$$

- Assumes twice differentiable f with Lipschitz-continuous Hessian that is locally spd.

5 ADMM

- Alternating Direction Method of Multipliers: for problems of the form

$$\begin{array}{rl} \text{minimize} & f(x) + g(z) \\ \text{subject to} & Ax + Bz = c. \end{array}$$

- More or less alternates a method of multipliers optimization in x and then in y.

- Takes advantage of separability of objective.

6 Interior point methods

- The idea is that you want to be able to use Newton's method for constrained convex optimization, so you find a way to make the discontinuous "barriers" at the inequality constraints continuous.

- For each inequality constraint we add a log barrier that blows up near the wall. We do a few Newton steps and get a solution. Then we replace the log barrier with a steeper one and repeat. In case the solution is on or near the constraint, this recursion gets us closer and closer to allowing solutions at the constraint.