10-725/36-725: Convex Optimization **Spring 2015**

Lecture 26: April 22nd

Lecturer: Ryan Tibshirani *Scribes: Eric Wong, Jerzy Wieczorek, Pengcheng Zhou*

Note: *LaTeX template courtesy of UC Berkeley EECS dept.*

Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications. They may be distributed outside this class only with the permission of the Instructor.*

The set of convex optimization is only a small subset of all convex optimizations problems. Most problems are non-convex! Non-convex problems are typically harder to solve and analyze and have higher variance, but some can be solved exactly (to global optimality).

26.1 Non-convex problems

Non-convex problems typically have higher variance. A heuristic argument for this is as follows: Suppose we have a convex problem and a non-convex problem. If we perturb the input data by some smooth function, a non-convex surface could result in substantially different local minima depending on the realization of the input data. On the other hand, a convex surface is more likely to result in the same or similar (perturbed) minima.

Today's lecture is largely about a few special cases of non-convex problems that happen to be tractable. It does not cover general strategies for arbitrary non-convex problems.

26.1.1 Definition

A nonconvex problem is defined as

$$\begin{aligned} \min_x \quad & f(x) \\ \text{subject to} \quad & h_i(x) \le 0, i = 1, \dots, m \\ & l_j(x) = 0, j = 1, \dots, r \end{aligned} \tag{26.1}$$

where at least one of the conditions for a convex problem is not met (f and h_i are convex, l_j are affine).

This definition includes non-convex formulations of convex problems. For example

$$\begin{aligned} \min_x \quad & f(x) + g(x) \\ \text{subject to} \quad & g(x) = 0 \end{aligned} \tag{26.2}$$

where f is convex, and g is non-convex can be reduced to a convex problem, but is a non-convex formulation. From this point on, we consider non-convex problems that are not trivially equivalent to convex ones.

Nonconvex problems can have local minima. By solving a nonconvex problem, we mean finding the global minimizer. We also implicitly mean doing it efficiently, i.e., in polynomial time.

26.1.2 Linear fractional programs

A linear fraction program is

$$\begin{aligned} \min_{x} \quad & \frac{c^T x + d}{e^T x + f} \\ \text{subject to} \quad & Gx \leq h,\ e^T x + f > 0 \\ & Ax = b \end{aligned} \tag{26.3}$$

This is not convex because the ratio of linear functions is not convex, but quasiconvex. Using a transformation, this is actually equivalent to a linear program. Interestingly, a place where this shows up are the knots in the LASSO solution path. Every knot is the solution of some linear-fractional program.

(A quasiconvex function is one where all the sublevel sets are convex. For example, $f(x) = \sqrt{|x|}$ is not convex, but is quasiconvex: for any $t > 0$, the set $\{x : f(x) \leq t\}$ is a convex set.)

Takeaway: try adding a variable to transform the non-convex problem into a convex one.

26.1.3 Geometric programs

A geometric program is

$$\begin{aligned} \min_{x} \quad & f(x) \\ \text{subject to} \quad & h_i(x) \leq 1, i = 1, \ldots, m \\ & l_j(x) = 1, j = 1, \ldots, r \end{aligned} \tag{26.4}$$

where f, h_i are posynomials and l_j are monomials. We can transform this to a convex problem with $y_i = \log x_i$.

Takeaway: try transforming your variables to convert the non-convex problem into a convex one.

26.1.4 Convex equality constraints

If equality constraints are convex but not affine, we can relax them from $l(x) = 0$ to $l(x) \leq 0$. This is very useful when $l(x)$ is convex but not affine. Then, if $l(x^*) = 0$ at any solution, then the relaxation is equivalent. This idea is used in the maximum utility problem, which we also saw in HW1 Q1.

Takeaway: try a convex relaxation of your constraints to turn the non-convex problem into a convex one.

26.1.5 Two quadratic functions

Consider the following problem

$$\begin{aligned} \min_{x} \quad & x^T A_0 x + 2b_0^T x + c_0 \\ \text{subject to} \quad & x^T A_1 x + 2b_1^T x + c_1 \leq 0 \end{aligned} \tag{26.5}$$

Note this is not convex since A_0, A_1 are not necessarily positive definite. However, its dual problem can be cast as a semidefinite program and is convex, and it can be shown that strong duality holds.

Takeaway: try deriving the dual, which will be convex; if strong duality holds, solving the (convex) dual may solve your (non-convex) primal problem.

26.2 Eigen Problems

Now, we move on from classical non-convex problems.

26.2.1 Principal Component Analysis

PCA is the problem of given a matrix Y:

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

This has the well-known solution of taking a singular value decomposition, and truncating all but the top k singular values.

This is not a convex problem due to the constraint (the objective is convex). The rank constraint is non-convex: for example, let J_{11}, J_{22} be 2×2 zero matrices with a single 1 in the upper left corner and lower right corner respectively. Then, each of these have rank 1, but their average has rank 2.

This rank constraint is a matrix analogue to the l_0 norm on vectors (which is also non-convex). The l_0 norm counts the number of nonzero entries; the rank constraint counts the number of nonzero singular values.

Here is another (equivalent) characterization of the SVD. Given a symmetric matrix S (which you can think of as Y^TY or $\text{Cov}(Y)$ for the Y in the problem above):

$$\min_Z ||S - Z||_F^2 \text{ subject to } \text{rank}(Z) = k,\ Z \text{ is a projection}$$

This is also non-convex, since both rank and projection constraints are non-convex. The solution is $\hat{Z} = V_k V_k^T$ where the columns of V_k are the first k eigenvectors of S. This is equivalent to a convex problem: first re-express the constraint set C as

$$C = \{Z \in \mathbb{R}^{p \times p} : Z = Z^T,\ \lambda_i(Z) \in \{0, 1\} \text{ for } i = 1, \ldots, p,\ \text{tr}(Z) = k\}$$

The only offending (non-convex) constraint here is $\lambda_i(Z) \in \{0, 1\}$. We can relax this to $\lambda_i(Z) \in (0, 1)$ (this is equivalent to relaxing the problem to its convex hull). Now we have

$$C = \{Z \in \mathbb{R}^{p \times p} : Z = Z^T,\ 0 \preceq Z \preceq 1,\ \text{tr}(Z) = k\}$$

which is a convex set called the Fantope of order k. Relaxing this Fantope to its convex hull actually admits the same solution to the original, but it is now a convex formulation.

Vu et al. (2013) use this convex relaxation via the Fantope to derive a sparse version of PCA. Adding a l_1 penalty term to the original non-convex problem is intractable. Adding a l_1 penalty term to the Fantope-based convex relaxation is still convex, and hence tractable. However, the solution to this penalized-and-relaxed problem is no longer necessarily the same as the solution to the original penalized problem. . . but maybe that is not a big concern since the original wasn't tractable anyway.

26.2.2 Other problems

MDS (multidimensional scaling) and generalized eigenvalue problems are other eigen problems, but we will skip them for the sake of time.

26.3 Graph problems

26.3.1 Min cut

The min-cut problem is given a graph $G = (V, E)$, two nodes $s, t \in V$, and costs $c_{ij} \geq 0$ on the edges $(i, j) \in E$, minimizes the cost of a partition between s and t.

This is not convex; in fact is is an integer program. If we throw out some of the hard constraints, we end up with a linear program which is the dual of the max flow linear program (see lecture 12 for more detail).

26.3.2 Shortest paths

Finding the shortest path between nodes is also a non-convex problem. Dijkstra's algorithm solves this and more, in a reasonable time $O(|E| \log |V|)$. We skip details for sake of time.

26.4 Non-convex proximal operators

Consider a weighted l_0 norm:

$$\min_x \sum_{i=1}^{n} (y_i - \beta_i)^2 + \sum_{i=1}^{n} \lambda_i 1\{\beta_i \neq 0\} \tag{26.6}$$

By inspection, the solution is to hard-threshold y: $\beta_i = y_i$ if $y_i^2 > \lambda_i$, and 0 otherwise.

If we switch the loss to $||y - X\beta||_2^2$, this is called best subset selection. This is generally NP hard!

26.4.1 l_0 segmentation

Given a sequence y_i, find a sequence β_i that minimizes the number of "jumps". specifically,

$$\min_\beta \sum_{i=1}^{n} (y_i - \beta_i)^2 + \lambda \sum_{i=1}^{n-1} 1\{\beta_i \neq \beta_{i+1}\} \tag{26.7}$$

Recall that we get fused lasso if we replace $1\{\beta_i \neq \beta_{i+1}\}$ with $|\beta_i - \beta_{i+1}|$.

Both of these can be solved with dynamic programming. Note that this is different from "typical" dynamic programming since this is over a continuous range of variables. An algorithm by Johnson is more efficient, and a version by Bellman is more general.

26.4.2 Tree-leaves projection

Given a target $u \in \mathbb{R}^n$, tree g on $\mathbb{R}^n$ and label $y \in \{0, 1\}$, consider

$$\min_z ||u - z||_2^2 + \lambda \cdot 1\{g(z) \neq y\}$$

In other words, we want to find a z such that the the label under the tree is the same as y.

We can argue that the solution is either $\hat{z} = u$ or $\hat{z} = P_S(u)$, the projection onto the set $S = g^{-1}(1) = \{z : g(z) = y\}$. This is the projection onto a set of tree leaves, which is very non-convex.

S is a union of boxes (tree leaves), but projection onto a set of boxes is non-convex. We could project onto each box individually, but this is expensive.

Instead, label every node with the list of its leaves and the bounding box of its leaves. Now perform DFS and prune nodes that do not contain a leaf labeled y, or if the bounding box is further away from the current closest.

26.5 Discrete problems

Skipped.

26.6 Infinite-dimensional problems

Skipped.

26.7 Statistical problems

26.7.1 Sparse underdetermined linear systems

Given X such that $||X_i||_2 = 1$ for $i = 1, \dots, n$ and y, consider

$$\min_\beta ||\beta||_0 \text{ subject to } X\beta = y$$

This is non-convex and also NP hard without additional restrictions on the form of X. The natural approach is to relax to the l_1 norm:

$$\min_\beta ||\beta||_1 \text{ subject to } X\beta = y$$

These two problems are interestingly connected.

If $p > n$ very large, there is a threshold $\rho(p/n)$ such that for most matrices X if we solve the l_1 problem and find a solution with less than ρn non-zero components, then this is the unique solution of the l_0 problem.

If we have greater than ρn non-zero components, then there is no solution with less than ρn non-zero components. In this case, maybe we shouldn't care about it.

In this case, "most" refers to with high probability when drawing X uniformly at random.