10-725/36-725: Convex Optimization **Fall 2018**

Lecture 3: September 5, 2018

Lecturer: Ryan Tibshirani *Scribes: Ruochi Zhang, Hongyu Zheng, Zirui Wang*

3.1 Announcements

Homework 1 is now posted on the website. Also the proposal guideline is avaliable now. Three options avaliable.

If you are on the waitlist, please DO NOT sign up for scribe. Leave that to the people who is enrolled.

Deadline for quiz is always midnight(Pittsburgh local time). Some browsers say it's 4AM(UTC), please don't believe that.

3.2 Convex Functions (continued)

3.2.1 Converse of Hyperplane Theorems

Theorem 3.1 (Converse of Separating Hyperplane Theorem.) *If you have two convex sets, one is open and one is closed and they are separated by a hyperplane, then they are disjoint.*

Theorem 3.2 (Converse of Supporting Hyperplane Theorem.) *If a set is closed and have nonempty interior, and has a supporting hyperplane at every boundary point, then it's convex.*

3.2.2 Composition Rules

Theorem 3.3 (Affine Composition.) *If $g(x)$ is a convex function, $f(x) = g(Ax + b)$ is convex.*

Usually you can use affine composition rule to avoid using chain rule at proving convexity.

Theorem 3.4 (Composition Rules of Scalar Convex Functions.) *if $h : \mathbb{R}^n \to \mathbb{R}$ is convex, $g : \mathbb{R} \to \mathbb{R}$ is convex and non-decreasing, then their composition $f : \mathbb{R}^n \to \mathbb{R} = g \cdot h$ is convex. If g is instead non-increasing, f is concave.*

One way to remember all these are to remember the chain rule when h is a scalar function, that is, $h : \mathbb{R} \to \mathbb{R}$:

$$\begin{aligned} f(x) &= h(g(x)) \\ f'(x) &= h'(g(x)) \times g'(x) \\ f''(x) &= h''(g(x))(g'(x))^2 + h'(g(x))g''(x) \end{aligned}$$

If h' is non-negative(h is non-decreasing), h'' and g'' are non-negative(h and g are convex), we have non-negative f''(f is convex). Note that the theorem itself does not assume the differentiability of the functions.

Theorem 3.5 (Composition Rules of Vector Convex Functions.) *if $h : \mathbb{R}^n \to \mathbb{R}^m$ is convex, $g : \mathbb{R}^m \to \mathbb{R}$ is convex and non-decreasing in each argument, then their composition $f : \mathbb{R}^n \to \mathbb{R} = g \cdot h$ is convex. If g is instead non-increasing(in each argument), f is concave.*

You can still use the chain rule to remember this.

An example is the log-sum-exp function. We first use affine rule to reduce the problem to $f(x) = \log(\sum \exp x_i)$, then one can calculate its second derivative and realize that the Hessian is diagonally dominant, thus positive semidefinite. The following lemma is useful here:

Lemma 3.6 (Diagonally Dominant Matrices.) *If a Hermitian matrix $A \in \mathbb{R}^{n\times n}$ satisfies $A_{ii} > \sum_{j\neq i} |A_{ij}|$, A is positive semidefinite.*

3.3 Optimization Basics

3.3.1 Terminology

Recall the difinition of a convex problem:

$$\begin{array}{ll} \min_{x\in\mathcal{D}} & f(x) \\ \text{subject to} & g_i(x) \leq 0, i = 1, ..., m \\ & Ax = b \end{array}$$

These are the nomenclature we use in convex optimization:

- convex objective or criterion: f.
- domain: $\mathcal{D}$.
- inequality constraints: g_i.
- equality constraints: $Ax = b$.
- feasible points x: $x \in \mathcal{D}$ and it satisfy both inequality and equality constraints.
- optimal value f^*: the minimum of $f(x)$ over all feasible points.
- optimal/solution/minimizer x: if x is feasible and $f(x) = f^*$(these might not exist in some cases, for example, $f(x) = e^{-x}$ for unconstrained x, but we are not going into this).
- ϵ-suboptimal x: if x is feasible and $f(x) \leq f^* + \epsilon$.
- activity: $g_i(x) = 0$ for feasible x.
- solution set: $X_{opt} = \arg\min f(x)$ subject to all constraints.

(One question: why we don't have generic convex functions in equality constraints? Because equality can be seen as two inequalities together(that is, $h(x) \leq 0$ and $-h(x) \leq 0$). For both $h(x)$ and $-h(x)$ to be convex, h has to be affine. Ask Ryan at office hours if you are interested in sufficiency conditions for convex problems to have solutions.)

Convex problem can be reposed as concave function.

Theorem 3.7 (Convexity of Solution Set.) *X_{opt} is always a convex set.*

Proof: We can mostly follow the definition of convex sets. For $x, y \in X_{opt}, 0 \leq t \leq 1$, first note that $tx + (1-t)y$ is feasible since the set of feasible solutions is convex, and we have $f(tx + (1-t)y) \geq f^*$. Then by convexity of f we have $f(tx+(1-t)y) \leq tf(x)+(1-t)f(y) = f^*$. We conclude with $f(tx+(1-t)y) = f^*$, and by definition of X_{opt}, $tx + (1-t)y \in X_{opt}$. ■

Theorem 3.8 (Solution Set for Strongly Convex Objectives.) *X_{opt} is unique(contains at most one point).*

Proof: We can mostly follow last proof. If the set is not unique, for $x, y \in X_{opt}, 0 < t < 1$, first note that $tx + (1-t)y$ is feasible since the set of feasible solutions is convex, and we have $f(tx + (1-t)y) \geq f^*$. Then by strong convexity of f we have $f(tx+(1-t)y) < tf(x)+(1-t)f(y) = f^*$, which contradicts with previous statement. ■

3.3.2 Examples

(1) We first look at the constrained form of Lasso Optimization.

$$\begin{array}{ll} \min_\beta & \|y - X\beta\|_2^2 \\ \text{subject to} & \|\beta\|_1 \leq s \end{array}$$

Usually for these type of problems, we use regularized form($f(x) = ||y - X\beta||_2^2 - C||\beta||_1$), but later in this course(when we cover KKT condition) we will see these two types are in some sense equivalent to each other.

Is this a convex problem? Yes, since $f(x)$ is a squared loss function which is quadratic function with quadratic term X^TX, positive semi-definite by definition.

Is the solution unique? Depends.

If $n \geq p$ and X has full column rank, X^TX is now invertible which means its smallest eigenvalue is positive, thus it's now positive definite, and f is now strictly convex, so solution will be unique. Note that this has nothing to do with constraints.

If $p > n$, X^TX can't have full rank, so f can't be strictly convex and this means the solution will in general not be unique. However, using KKT conditions and probability theory, we can prove that "very often" we can achieve the uniqueness.

(2) The second example we look at is SVM.

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

This is the inseparable version of SVM, and some calculations will show that $\frac{1}{2}||\beta||_2^2$ corresponds to the margin of the hyperplane separator.

Is this a convex problem? Yes, since everything is convex here.

Is the solution unique? No since there is a linear component in criterion value, however one can argue the β is unique because the objective is strictly convex in β.

3.3.3 Rewriting Constraints

We first look at how we rewrite the constraints. Recall that the indicator function $I_C(x) = 0$ if $x \in C$, and $-\infty$ otherwise.

For the first step, we can always get rid of explicit constraints by setting the domain of x to be the set that all constraints are satisfied, i.e. $\min_{x \in C} f(x)$ where $C = \{x : g_i(x) \leq 0, Ax = b\}$. Further, we can write $f(x) = \min f(x) + I_C(x)$ and let x to be unconstrained. All these transformations are fully general.

3.3.4 First-Order Optimality Condition

Theorem 3.9 (First-Order Optimality.) *For a convex optimization problem* $\min f(x)$*, if* f *is differentible and* x *is feasible,* x *is optimal if and only if* $\nabla f(x)^T(y - x) \geq 0$ *for all* $y \in C$*.*

Theorem 3.10 (Unconstrained First-Order Condition.) *For the above optimization problem, if* x *is unconstrained(*$x \in \mathbb{R}^n$*), the condition is* $\nabla f(x) = 0$*.*

Proof: We only need to show if $\nabla f(x)^T(y - x) \geq 0$ for all $y \in \mathbb{R}^n$, $\nabla f(x) = 0$. This is obvious since we can show $\nabla f(x)^T v = 0$ for any $v \in \mathbb{R}^n$, by $\nabla f(x)^T v \geq 0$ and $\nabla f(x)^T(-v) \geq 0$. ■

3.3.5 Lagrangian Multiplier

Consider the equality-constrained problem:

$$\min_x f(x) \text{ subject to } Ax = b$$

with f differentiable.

Then, using the first-order optimality condition we have that $\nabla f(x)^T(y - x) \geq 0$ for solution x. This implies that $\nabla f(x)^T v = 0$ for some $v \in null(A)$. Since row space of A is the orthogonal complement of the null-space of A, this implies that $\nabla f(x)$ must be in the row space and hence we must be able to write $\nabla f(x) = -A^T u$ for some u. Thus, we have the Lagrangian multiplier optimality condition as:

$$\nabla f(x) + A^T u = 0$$

3.3.6 Partial Optimization

We can always partially optimize a convex problem and retain convexity by reformatting the domain of partial variables.

Example: Hinge Form of SVMs

We argue the exact value of ξ_i can be determined easily, and plug this to eliminate ξ_i (but we introduced a max function now and that's why you don't see this often).

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

where $[x]_+ = \max(x, 0)$, the hinge function.

3.3.7 Transformations and Change of Variables

Theorem 3.11 (Transformations under Monotone Functions.) *If h is a monotone increasing function,* $\min f(x)$ *has the same solution set as* $\min h(f(x))$.

The most common use is to reveal some hidden convexity; For example, for many distributions in exponential family, the likelihood function is not convex, but we can take logarithm and the resulting LL(log-likelihood) function sometimes can be convex.

Change of variables is also used at times, the most popular example being linear transformations:

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

3.3.8 Eliminating Linear Constraints (special case of change of variables)

Theorem 3.12 (Solutions to Linear Constraints.) *We can eliminate the linear constraint using the following equivalence condition:*

$S = \{x : Ax = b\} = \{My + x_0 : Ax_0 = b, col(M) = null(A)\}$

Here $null(A)$ represents the solution space of $Ax = 0$, that is, if x satisfies $Ax = 0$, the it's linear combination of $null(A)$.

Why we don't use this in practice? First, calculating the null space can be very costly. Second, there are many situations where A is very sparse, but its null space is usually very dense.

3.3.9 Slack Variables

We can eliminate the inequalities by introducing slack variables. Saying $g(x) \leq 0$ is equivalent to say $g(x) + \xi = 0$ and $\xi \geq 0$(You see this in SVM). In other words, we can transform the following problem:

$$\begin{array}{ll} \min_x & f(x) \\ \text{subject to} & g_i(x) \leq 0, i = 1, ..., m \\ & Ax = b \end{array}$$

into:

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

However, the problem now is usually no longer convex, since $g(x) + \xi = 0$ does not in general define a convex set, unless $g(x)$ is affine.

Note that we use this technique in linear programming, since all constraints are affine, so slack formation does not lose any useful property.

3.3.10 Relaxation of Nonaffine Inequalities

This is the last technique we cover this lecture. Suppose you have an optimization problem with convex $g_i(x)$, but the constraint is now equality, that is $g_i(x) = 0$, so it's no longer convex. We can relax the constraint to $g_i(x) \leq 0$ and solve with convex optimization, but now the solution might no longer be feasible; We can guarantee that it's a lower bound of the original problem, that is, $f^{*'} \leq f^*$.

An important application is the maximum utility problem we will cover in next guest lecture.

3.4 Contributions

- **Ruochi Zhang**: Revisions and polishing.
- **Hongyu Zheng**: First draft of the scribe.
- **Zirui Wang**: Revisions and polishing.