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

Lecture 1: January 12

Lecturer: Ryan Tibshirani *Scribes: Seo-Jin Bang, Prabhat KC, Josue Orellana*

1.1 Review

We begin by going through some examples and key properties of convex functions we discussed in the last lecture *Convexity I: Sets and functions, Aug 31.* In the review section, only the important examples and properties that the instructor mentioned again would be scribed.

1.1.1 Examples of Convex Functions

We start to go over examples of convex function we mentioned last time.

- Convexity of **univariate functions** such as Exponential function, Power function, Logarithmic function can be checked easily by drowing the functions.
- **Affine function** ($a^T x + b$) is both convex and concave.
- **Quadratic function** $\frac{1}{2}x^T Qx + b^T x + c$ is convex provided that $Q \succeq 0$ (i.e. positive semidefinite) Using the second-order characteristic of convexity, it can be derived easily.
- **Least squares loss** is always convex because $\|y - Ax\|_2^2$ is a type of the quadratic function having $Q = A^T A$ and $A^T A$ is always positive semidefinite.
- Every **norm** is convex including operator (spectral) and trace (nuclear) norms. The proof can be done by using the definition of convexity. Operator norm is the largest singular value of matrix X. It also has basic properties of norm such as theh triangle inequality.
- Convexity of **indicator function provided that C, support function, max function** can be checked from the definition of the convexity.

1.1.2 Key properties of convex functions

In this section, we go over the key properties of convex functions.

- **Epigraph characterization**: $epi(f)$ is a set of every points that lie on above the function f:

  $$epi(f) = \{(x, t) \in dom(f) \times \mathbb{R} : f(x) \leq t\}$$

  A function f is convex if and only if its epigraph $epi(f)$ is a convex set. It is useful properties because we can derive convexity of a function from convexity of a set.
- **Convex sublevel sets**: a sublevel set of a function f is a set of points in domain of f such that its value $f(x)$ is not larger than any fixed point $t \in \mathbb{R}$:

  $$\{x \in dom(f) : f(x) \leq t\}$$

If function f is convex, then its sublevel sets are convex for any choice of t. Note that the conver is not true. The counter example is $f(x) = \sqrt{|x|}$. When it sublevel sets are convex, we call f quasiconvex function.

- **First-order characterization**: if a function is differentiable, then f is convex if and only if its domain is convex and satisfies a condition such that:

$$f(y) \geq f(x) + \nabla f(x)^T (y - x)$$

for all points x and y in its domain. We can understand the property easily through an one dimensional funciton $f(x) = x^2$:

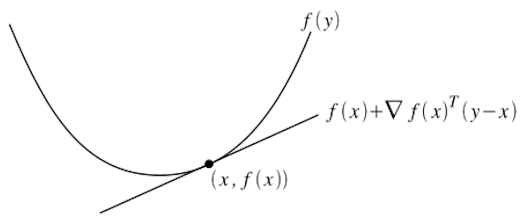

Figure 1.1: Illustration of the first-order condition for convexity. `http://funktor.github.io/2015/07/03/Convex-Optimization/`

We can drow a tangent line at any fixed point x such that $g(y) = f(x) + \nabla f(x)^T (y - x)$. The tangent line always lies below the convex function. Therefore, $f(y) \geq g(y) = f(x) + \nabla f(x)^T (y - x)$.

The first-order characterizaiton of strict convexity of a function f can be seen similar way: if a cuntion is differentiable, then f is strictly convex if and only if

$$f(y) > f(x) + \nabla f(x)^T (y - x)$$

- **Second-order characterization**: if a function is twice differentiable, then f is convex if and only if its domain is convex and satisfies a condition such that:

$$\nabla^2(f) \succ 0$$

for all points x in its domain.

If a function f is strictly convex, then $\nabla^2(f) \succeq 0$ (positive definite). However the converse is not true. An counter example is $f(x) = x^2$. It is strictly convex but it has zero second derivative (i.e. $f^{''}(0) = 0$)

- **Nonnegative linear combination** of convex functions is convex.

- **Pointwise maximization**: If we define a new function $f(x)$ at x as the maximum value of (countable infinite) convex functions at x, f is convex. It implies that we can always maximize a bunch of functios in pointwise fashion.

- **Partial minimization** If $g(x, y)$ is convex in x and y and C is a convex set, then partially minimized function on any variable over the convex set C, i.e. $f(x) = \min_{y \in C} g(x, y)$ and $f(y) = \min_{x \in C} g(x, y)$, is convex.

1.1.3 More operations preserving convexity

This section we go over examples of composition of functions that preserves convexity. some examples of composition

- **Affine composition** in a convex function f is always convex. That is, if f is convex, then $g(x) = f(Ax+b)$ is convex.
- **General composision** is about convexity of $f(x) = h(g(x))$ when the outside function $h : \mathbb{R} \to \mathbb{R}$ is monotone and the inside function $g : \mathbb{R}^n \to \mathbb{R}$ is convex/concave:

  f is convex if h is convex and nondecreasing, g is convex

  f is convex if h is convex and nonincreasing, g is concave

  f is concave if h is convex and nondecreasing, g is concave

  f is concave if h is convex and nonincreasing, g is convex
- **Vector composition** is similar manner with the general composition in pointwise fashion.

1.1.4 Example: log-sum-exp function

Log-sum-exp function is:

$$g(x) = \log\left(\sum_{i=1}^{k} \exp(a_i^T x + b_i)\right)$$

for fixed a_i and b_i. It is a nice example of convex function which convexity can be shown by the operations preserving convexity.

Since affine composition preserves convexity, it is enough to show $f(x) = \log\left(\sum_{i=1}^{k} \exp(x_i)\right)$. Using the second-order characteristic of $f(x)$ we can show the convexity of $g(x)$.

1.1.5 Is $\max\left\{\log\left(\frac{1}{(aTx+b)^7}\right), \|Ax+b\|_1^5\right\}$ convex?

We will make use of operations that preserve convexity to determine the curvature of following function

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

We begin by realizing that $(a^T x + b)$ and $Ax + b$ are affines and so are convex functions. Accordingly, the problem can be reformulated as determining the convexity of the problem, $\max\{-7\log(x), ||y||_1^5\}$. Here, log is concave function and so -7log becomes convex. Likewise, norm is convex function and so norm raised to power of 5 is convex. Finally, the fact that maximum of convex functions is convex deduces that the given problem is, indeed, a convex function.

This lecture will comprise of following topics:

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

1.2 Optimization terminology

We begin by defining a convex optimization problem (or program) as follows:

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

where objective (or criterion) function, f, and inequality constraint functions, g_i, are all convex. Likewise, the equality constraint is linear. Also, we do not often discuss this but it is implicitly implied that the domain is $D = \text{dom}(f) \cap \bigcap_{i=1}^{m} \text{dom}(g_i)$.

Furthermore, any x that satisfies all the constraints of the optimization problem is called a feasible point. The minimum of our criterion, $f(x)$, over all feasible points, x, is called the optimal value, $f^\star$. Likewise, if $x^\star \epsilon x$ s.t. $f(x^\star) = f^\star$, then $x^\star$ is called optimal or a solution. Next, a feasible point, x, is called $\epsilon-$suboptimal, if it has the property $f(x) \leq f^\star + \epsilon$. Similarly, if x is feasible and $g_i(x) = 0$, then we say that g_i is active at x. In contrast, if $g_i(x) < 0$, then we say g_i is inactive at x. Finally, any convex minimization can be reposed as concave maximization. This is primarily owing to the fact that minimizing $f(x)$ subject to some constraints is equivalent to maximizing -$f(x)$ over the same constraint, in the sense that they both have same solution.

1.3 Convex solution set

Consider X_{opt} to be the set of all solutions of a convex problem. Then it can be expressed as:

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

Here, we can quickly check the convexity of X_{opt} by considering two solutions x, y. Then for $0 \leq t \leq 1$, $tx + (1 - y)y \epsilon D$. Likewise, the two solutions satisfy inequality and equality constraints. Next,

$$f(tx + (1-t)y) \leq tf(x) + (1-t)f(y) = tf^\star + (1-t)f^\star = f^\star.$$

Hence, $tx + (1 - t)y$ is also a solution and X_{opt} is a convex set. This outcome is due to the property of convex functions that their solutions are convex. However, as mentioned in previous lectures, just because X_{opt} is a convex set, does not mean that it is unique. This is to say that, even though a local solution is also globally minimum, there could still be multiple solutions to a convex optimization problem. In particular, these optimization problems may have 0, 1 or infinitely many solutions. One obtains a unique solution if f is strictly convex.

Some of the examples of convex optimization problem include:

1.3.1 Lasso

Lasso is a common problem that people look at in machine learning and statistics. Basically, it is a regression problem. Given $y \epsilon R^n$, and $X \epsilon R^{n \times p}$, a lasso problem can be formulated as:

$$\min_{\beta} \ ||y - X\beta||_2^2$$
$$\text{subject to} \ ||\beta||_1 \leq s$$

Lasso is a convex optimization problem because the objective function is a least squared loss which is convex and the constraint is a norm minus a constant which in itself is convex. Moreover, the problem has only inequality, $g_i(\beta) = ||\beta||_1 - s$, and no equality constraint. The feasible set is $\beta \epsilon R^p$ that satisfy the L_1-norm bound (or L_1 ball) of $|\beta||_1 \leq s$.

- $n \geq p$ and X has full column rank

  Here, $\nabla^2 f(\beta) = 2X^T X$. Given that X is a full rank $\Rightarrow X^T X$ is invertible $\Rightarrow X^T X$ is positive definite. Hence, $\nabla^2 f(\beta) \succ 0$ and so the solution in this case is unique because strictly convex functions have only one solution.

- $p > n$ (high-dimension) case.

  In this case, $X^T X$ is singular. Then $f(\beta) = \beta^T X^T X \beta - 2y^T X \beta + y^T y$ and for some $\beta \neq 0$ and $X\beta = 0$, we get $\beta^T X^T X \beta = 0$. This would mean that the function, $f(\beta)$, is linear. Thus, we get multiple solutions and cannot guarantee a unique solution. However, later in the course, we will see that in most cases where $n > p$, we still get unique solution with lasso.

If a function f is strictly convex, that implies uniqueness, otherwise we cannot say anything about uniqueness. But we can still evaluate the particular circumstances of a problem on a case by case basis.

1.3.2 Example: support vector machines

This is a way to produce a linear classification function. Linear in the sense that the decision boundary is still linear in the variables.

Given labels $y \in \{-1, 1\}^n$, and features $X \in \mathbb{R}^{nxp}$ with rows $x_1, ..., x_n$

There is really only two variables (β and ξ). The intercept β_0 is a single dimensional parameter. C is a chosen constant.

Here is the SVM criterion:

$$\underset{\beta,\beta_0,\xi}{min} \frac{1}{2}\|\beta\|_2^2 + C\Sigma_{i=1}^n \xi_i \tag{1.1}$$

subject to the following constraints

$\xi_i \geq 0, i = 1, ..., n$

$y_i(x_i^T \beta + \beta_0) \geq 1 - \xi_i,\ i = 1, ..., n$

This problem is **convex** because the criterion is just a quadratic plus a linear function. We can also rewrite the constrain $-\xi_i \leq 0$ which is affine and a convex function. In the same way we can rewrite the full inequality constrain as $-y_i(x_i^T \beta + \beta_0) + 1 - \xi_i \leq 0$.

This problem is **not strictly convex** because the criterion is a linear function of ξ's. Thus based on what we know so far, we cannot say anything about uniqueness.

Special case: If we fix all of the other variables, and only treat the component β, which determines the hyperplane, then the criterion is strictly convex. The criterion becomes just the squared error loss, and strict convexity implies uniqueness.

1.3.3 Rewriting constraints

Consider this optimization problem

$$\min_{x} f(x) \tag{1.2}$$

subject to $g_i(x) \leq 0, i = 1, ..., m$ and $Ax = b$

Without loss of generality, the constraints can be encapsulated into a set C.

subject to $x \in C$. Where $C = \{x : g_i(x) \leq 0, i = 1, ..., m, Ax = b\}$, the feasible set.

Saying that $x \in C$ is therefore equivalent to saying that all of the constraints described in C are met. Furthermore we can use I_C as the indicator of C to rewrite the problem as:

$$\min_{x} f(x) + I_C(x) \tag{1.3}$$

The indicator function I_C is 0 when $x \in C$ and infinity when $x \notin C$. When C is convex, this is going to be a convex function. Using the definitions from convex functions, if $g_i(x)$ is convex, then $g_i(x) \leq 0$ is a convex set because it is a sub level of a convex function. An intersection of convex sets is created when we assert that $g_i(x) \leq 0$ must be true for all $i = 1, ..., m$. Intersection is also an operation that preserves convexity for sets. Thus C is a convex set made out of convex constraints.

$$C = \cap_{i=1}^{m} \{x : g_i(x) \leq 0\} \cap \{x : Ax = b\} \tag{1.4}$$

1.3.4 First-order optimality

First-order optimality is a necessary and sufficient condition for convex functions. The statement is similar for convex problems.

$$\min_{x} f(x) \tag{1.5}$$

subject to $x \in C$

Let f be differentiable (smooth), then a feasible point x is optimal if and only if:

$$\nabla f(x)^T (y - x) \geq 0 \tag{1.6}$$

for all $y \in C$

All feasible directions from x are aligned with the gradient $\nabla f(x)$

Interpretation: Assume you are at a feasible point x, and you are thinking of moving to a feasible point y. Then if the gradient is aligned with the vector from x to y, the function should increase because you are going in the direction in which the gradient is increasing. If that is true for all feasible points y, then the point x must be the solution.

Special case: $C = \mathbb{R}^n$. This is the case of unconstrained optimization, in which we are just trying to minimize a convex smooth function f. The solution must be at the point where the gradient is zero $\nabla f(x) = 0$.

1.3.5 Example: quadratic minimization

Consider minimizing:

$$f(x) = \frac{1}{2} x^T Q x + b^T x + c \tag{1.7}$$

Where $Q \succeq 0$. The first order condition says that the solution satisfies:

$$\nabla f(x) = Qx + b = 0 \tag{1.8}$$

There are 3 possible solutions which depend on Q:

- if $Q \succeq 0$, i.e. positive definite, then there is a **unique solution** at $x = -Q^{-1}b$
- if Q is singular, i.e. not invertible, and $b \notin col(Q)$, then there is **no solution**.
- if Q is singular, i.e. not invertible, and $b \in col(Q)$, then there are **infinitely many solutions** of the form $x = Q^{+}b + z$ where $z \in null(Q)$ and Q^{+} is a pseudo-inverse of Q

1.3.6 Example: equality-constrained minimization

Consider minimizing the equality constrained convex problem:

$$\min_{x} f(x) \tag{1.9}$$

subject to $Ax = b$ with f being differentiable.

We can write a Lagrange multiplier optimality condition

$$\nabla f(x) + A^T u = 0 \tag{1.10}$$

for some u.

We will come back to this derivation when we cover topics in duality. For now we can state how to prove this according to first order optimality. The solutions x satisfies $Ax = b$

$$\nabla f(x)^T (y - x) \geq 0 \tag{1.11}$$

for all y such that $Ay = b$

Because $null(A)^{\perp} = row(A)$. This is equivalent to

$$\nabla f(x)^T v = 0 \tag{1.12}$$

for all $v \in null(A)$

1.3.7 Partial optimization

We can always partially optimize a convex problem and retain convexity.

This stands on the fact that we can always partially minimize a function over some of its variables as long as the set being minimized is convex. Formally: $g(x) = min_{y \in C} f(x, y)$ is convex in x, provided that f is convex in (x, y) and C is a convex set.

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

$$\min_{x_1, x_2} f(x_1, x_2) \tag{1.13}$$

subject to $g_1(x_1) \leq 0$ and $g_2(x_2) \leq 0$

Partially we can also

$$\min_{x_1} \tilde{f}(x_1) \tag{1.14}$$

subject to $g_1(x_1) \leq 0$ where $\tilde{f}(x_1) = min\{f(x_1, x_2) : g_2(x_2) \leq 0\}$

The second problem is convex if the first problem is convex.

1.3.8 Example: hinge form of SVMs

Refer to the optimization problem given in a previous section of this lecture.

Let us rewrite the constrains as $0 \leq \xi_i$, and $1 - y_i(x_i^T\beta + \beta_0) \leq \xi_i$ or equivalenty $\xi_i \geq max\{0, 1 - y_i(x_i^T\beta + \beta_0)\}$. We can argue that this inequality is exactly larger during optimization, and exactly equal only at the solution. This means we can eliminate ξ_i because we have identified what it exactly is at the solution: $\xi_i = max\{0, 1 - y_i(x_i^T\beta + \beta_0)\}$.

Thus plugging in for optimal ξ we can rewrite the problem in its hinge form:

$$\min_{\beta,\beta_0} \frac{1}{2}\|\beta\|_2^2 + C\Sigma_{i=1}^n[1 - y_i(x_i^T\beta + \beta_0)]_+ \tag{1.15}$$

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

1.3.9 Transformations and change of variables

Transforming variables

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

$min_x\, f(x)$ subject to $x \in C \iff min_x\, h(f(x))$ subject to $x \in C$

Similarly, inequality or equality constraints can be transformed and yield equivalent optimization problems. Can use this to reveal the "hidden" convexity of a problem. We do this often in statistics, when we optimize the log likelihood instead of the likelihood because Log is monotonically increasing.

Changing variables

If $\phi : \mathbb{R}^n \to \mathbb{R}^m$ is one to one, and its image covers feasible set C.

$min_x\, f(x)$ subject to $x \in C \iff min_y\, f(\phi(y))$ subject to $\phi(y) \in C$