3. Optimization

3.1. **Introduction to optimization.** *Note:* This lecture owes most of its material to the book by Boyd and Vandenberghe (2009), and various lecture notes which have appeared on the web and which are based on that book.

The line segment between x_1 and x_2 in a vector space V is all points

$$x = \theta x_1 + (1-\theta)x_2 \text{ with } 0 \leq \theta \leq 1.$$

A set C is **convex** if it contains the line segment between any two points in the set, i.e.

$$x_1, x_2 \in C, 0 \leq \theta \leq 1 \Rightarrow \theta x_1 + (1-\theta)x_2 \in C.$$

A **convex combination** of $x_1, \ldots, x_k$ is any point x of the form

$$x = \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_k x_k$$

with $\theta_1 + \cdots + \theta_k = 1, \theta_i \geq 0$. The **convex hull** of a set S, denoted conv(S), is the set of all convex combinations of points in S.

Practical methods for establishing convexity of a set C include the following. One can always try to apply the definition directly. A faster way might be to show that C is obtained from simple convex sets (hyperplanes, halfspaces, norm balls, etc.) by operations that preserve convexity:

- intersection
- affine functions
- perspective function
- linear-fractional functions

Before continuing, we establish some notation for matrices. S^n denotes the space of $n \times n$ symmetric, real matrices. S^n_+ denotes the subspace of S^n with non-negative eigenvalues, or equivalently, the subspace $\{A \in S^n : v^T A v \geq 0 \ \forall v\}$. S^n_{++} is defined similarly, but for *strictly positive* matrices, i.e. those satisfying $v^T A v > 0$ for any nonzero v.

If $f : \mathbb{R}^n \to \mathbb{R}^m$ is affine ($f(x) = Ax + b$ with $A \in \mathbb{R}^{m \times n}$, $b \in \mathbb{R}^m$) then the image (or inverse image) of a convex set under f is convex. Examples include scaling, translation, projection and the solution set of a linear matrix inequality

$$\{x \mid x_1 A_1 + \cdots + x_m A_m \preceq B\} \text{ with } A_i, B \in S^p.$$

The perspective function $P : \mathbb{R}^{n+1} \to \mathbb{R}^n$ is defined by

$$P(x,t) = x/t, \quad \text{dom} P = \{(x,t) \mid t > 0\}$$

It is a fact that images and inverse images of convex sets under perspective are convex.

A linear-fractional function $f : \mathbb{R}^n \to \mathbb{R}^m$ is defined as

$$f(x) = (Ax + b)/(c^T x + d),$$

with $\mathrm{dom}(f) = \{x \mid c^T x + d > 0\}$. Images and inverse images of convex sets under linear-fractional functions are convex.

Definition 3.1. A function $f : \mathbb{R}^n \to \mathbb{R}$ is **convex** if $\mathrm{dom} f$ is a convex set and

$$f(\theta x + (1-\theta)y) \leq \theta f(x) + (1-\theta) f(y) \tag{3.1}$$

for all $x, y \in \mathrm{dom} f, 0 \leq \theta \leq 1$. Function f is concave if $-f$ is convex. A function f is strictly convex if the inequality in (3.1) is strict.

Examples on $\mathbb{R}$ include affine, exponential e^{ax} for any $a \in \mathbb{R}$, powers x^α on $\mathbb{R}_{++}$ for $\alpha \geq 1$ or $\alpha \leq 0$, powers of absolute value $|x|^p$ on $\mathbb{R}$, for $p \geq 1$, and negative entropy $x \log x$ on $\mathbb{R}_{++}$.

On $\mathbb{R}^n$ affine functions are convex and concave; all norms on any normed vector space are convex. Examples on $\mathbb{R}^{m \times n}$ include $f(X) = \mathrm{tr}(A^T X) + b$, and the spectral (maximum singular value) norm

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

A function $f : \mathbb{R}^n \to \mathbb{R}$ is convex if and only if the function $g : \mathbb{R} \to \mathbb{R}$,

$$g(t) = f(x + tv) \quad \text{with} \quad \mathrm{dom}(g) = \{t \mid x + tv \in \mathrm{dom} f\}$$

is convex (in t) for any $x \in \mathrm{dom} f$ and $v \in \mathbb{R}^n$. Hence one can check convexity of f by checking convexity of functions of one variable.

A function f is *differentiable* if $\mathrm{dom} f$ is open and the gradient exists at each $x \in \mathrm{dom} f$. The following basic fact is known as the *first-order condition* for convexity: differentiable f with convex domain is convex iff

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

In other words the first-order approximation of f is global underestimator. A function f is *twice differentiable* if $\mathrm{dom} f$ is open and the Hessian $\nabla^2 f(x) \in S^n$ exists at each $x \in \mathrm{dom} f$.

There are also *second-order conditions* for checking convexity of differentiable functions. For twice differentiable f with convex domain f is convex if and only if

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

If $\nabla^2 f(x) \succ 0$ for all $x \in \text{dom} f$, then f is strictly convex.

Let's give some examples of multivariate functions that are convex. For our first example, consider a quadratic function:

$$f(x) = (1/2)x^T Px + q^T x + r \text{ (with } P \in S^n)$$

Then one calculates $\nabla f(x) = Px + q$, $\nabla^2 f(x) = P$. The function $f(x)$ is convex if $P \succeq 0$. A second example is the familiar least-squares objective: $f(x) = \|Ax - b\|^2$, which has

$$\nabla f(x) = 2A^T(Ax - b), \quad \nabla^2 f(x) = 2A^T A$$

This is convex for any A. The quadratic-over-linear function $f(x, y) = x^2/y$ is convex for $y > 0$ (just compute the Hessian). The log-sum-exp function $f(x) = \log \sum_k \exp x_k$ is convex.

$$\nabla^2 f(x) = \frac{1}{\mathbb{1}'z} \text{diag}(z) - \frac{1}{(\mathbb{1}'z)^2} zz^T \quad \text{for } z_k = \exp x_k$$

The geometric mean $f(x) = (\prod_k x_k)^{1/n}$ on R^n_{++} is concave.

Definition 3.2. Define the **α-sublevel** set of $f : \mathbb{R}^n \to \mathbb{R}$ as

$$C_\alpha = \{x \in \text{dom} f \mid f(x) \leq \alpha\}$$

Fact: sublevel sets of convex functions are convex (converse is false).

Definition 3.3. Define the **epigraph** of $f : \mathbb{R}^n \to \mathbb{R}$ as

$$\text{epi} f = \{(x, t) \in \mathbb{R}^{n+1} \mid x \in \text{dom} f, f(x) \leq t\}$$

A function f is convex if and only if epi$f$ is a convex set.

If f is convex, then

$$f(\mathbb{E}z) \leq \mathbb{E}f(z)$$

for any random variable z. This is known as *Jensen's inequality*. The basic inequality (3.1) that defines convexity is a special case with a discrete distribution.

In practice a great way to prove convexity of f is to show that f is obtained from simple convex functions by operations that preserve convexity:

- nonnegative weighted sum
- composition with affine function
- pointwise maximum and supremum
- composition
- minimization
- perspective

Examples include the log barrier for linear inequalities

$$f(x) = -\sum \log(b_i - a_i^T x), \text{ with dom} f = \{x \mid a_i^T x < b_i, i = 1, \ldots, m\}$$

and any norm of affine function: $f(x) = \|Ax + b\|$.

If $f_1, \ldots, f_m$ are convex, then $f(x) = \max\{f_1(x), \ldots, f_m(x)\}$ is convex. For example, a piecewise-linear function

$$f(x) = \max_{i=1,\ldots,m} (a_i^T x + b_i)$$

is convex. If $f(x, y)$ is convex in x for each $y \in A$, then $g(x) = \sup_{y \in A} f(x, y)$ is convex.

Examples include distance to farthest point in a set C:

$$f(x) = \sup_{y \in C} \|x - y\|.$$

Also the maximum eigenvalue of symmetric matrix: for $X \in S^n$,

$$\lambda_{max}(X) = \sup_{\|y\|_2 = 1} y^T X y.$$

Theorem 3.1. *Consider the composition of $g : \mathbb{R}^n \to \mathbb{R}$ and $h : \mathbb{R} \to \mathbb{R}$:*

$$f(x) = h(g(x))$$

Then f is convex if g convex, h convex, $\tilde{h}$ nondecreasing. Alternatively if g is concave and $\tilde{h}$ nonincreasing, then again f is convex.

Proof. For $n = 1$ and differentiable g, h:

$$f''(x) = h''(g(x))g'(x)^2 + h'(g(x))g''(x).$$

This needs to be ≥ 0. So for example, if h is convex then the first term is nonnegative, and if g is convex and $h' \geq 0$ at all points in its domain then the second term is nonnegative too. This should actually be covered in everyone's calculus-one course. □

Example applications of Theorem 3.1: $\exp g(x)$ is convex if g is convex. Also $1/g(x)$ is convex if g is concave and positive.

If $f(x, y)$ is convex in (x, y) and C is a convex set, then

$$g(x) = \inf_{y \in C} f(x, y)$$

is convex. The distance to a set $d(x, S) = \inf_{y \in S} \|x - y\|$ is convex if S is convex.

Definition 3.4. The **convex conjugate** of a function f is

$$f^*(y) = \sup_{x \in \mathrm{dom} f} \left(y^T x - f(x) \right). \tag{3.2}$$

It is also known as Legendre-Fenchel transformation or Fenchel transformation (after Adrien-Marie Legendre and Werner Fenchel).

This is another way of generating convex functions: f^* is convex (even if f is not). More importantly, this conjugation operation is fundamental to the construction of the *dual problem*, which we will discuss a bit later on. The dual is a very important practical tool for actually solving optimization problems numerically.

Now consider the basic optimization problem in standard form:

$$\begin{aligned} &\text{minimize } f_0(x) \\ &\text{subject to } f_i(x) \le 0,\ i = 1, \ldots, m \\ &\qquad\qquad h_i(x) = 0,\ i = 1, \ldots, p \end{aligned}$$

The optimal value will be denoted

$$p^* = \inf \left\{ f_0(x) \mid f_i(x) \le 0,\ h_j(x) = 0\ (\forall i, j) \right\}$$

By convention, $p^* = \infty$ if the problem is infeasible (no x satisfies the constraints), and $p^* = -\infty$ if the problem is unbounded below.

Definition 3.5. A point x is said to be **feasible** if $x \in \mathrm{dom} f_0$ and it satisfies the constraints. The **feasible set** is the set of all feasible x. A feasible x is said to be **optimal** if $f_0(x) = p^*$. We will use X_{opt} to denote the set of optimal points. A point z is **locally optimal** if it's optimal for the original problem with added constraint that $\|x - z\| \le R$ for some $R > 0$.

The standard form optimization problem has an implicit constraint

$$x \in \mathscr{D} = \bigcap_{i=0}^{m} \mathrm{dom}(f_i) \ \cap \ \bigcap_{j=1}^{p} \mathrm{dom}(h_j)$$

This $\mathscr{D}$ is called the domain of the problem. A problem is said to be unconstrained if it has no explicit constraints ($m = p = 0$). For example, minimization of $f_0(x) = -\sum_i \log(b_i - a_i^T x)$ is an unconstrained problem with implicit constraints $a_i^T x < b_i$.

The standard form *convex* optimization problem is the special case of the above in which $f_0, f_1, \ldots, f_m$ are convex and equality constraints are affine. In other words, the p equality constraints can be written as $Ax = b$ where $b \in \mathbb{R}^p$.

Definition 3.6. The **standard form convex optimization problem** is

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

where all f_i are convex.

The feasible set of a convex optimization problem is convex.

Theorem 3.2. *Any locally optimal point of a convex problem is (globally) optimal.*

Proof. Suppose x is locally optimal and there exists y with $f_0(y) < f_0(x)$. Now x locally optimal means $\exists\, R > 0$ such that

$$z \text{ feasible}, \; \|z - x\|_2 \leq R \;\Rightarrow\; f_0(z) \geq f_0(x).$$

Consider $z = \theta y + (1-\theta)x$ with $\theta = R/(2\|y - x\|_2)$. Since $\|y - x\|_2 > R$, one has $0 < \theta < 1/2$. Note z is a convex combination of two feasible points, hence also feasible. $\|z - x\|_2 = R/2$ and

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

which contradicts our assumption that x is locally optimal. □

Theorem 3.3. *If f is differentiable and convex, then any point $x_0 \in \mathrm{dom} f$ satisfying $\nabla f(x_0) = 0$ is a global minimum of $f(x)$.*

Proof. By the first-order condition for convexity,

$$f(y) \geq f(x) + \nabla f(x)^T (y - x) \;\forall\, y.$$

At $x = x_0$, $\nabla f(x_0) = 0$ and hence this reduces to $f(y) \geq f(x_0) \;\forall\, y$. □

Two problems are (informally) **equivalent** if the solution of one is readily obtained from the solution of the other, and vice-versa. Convex problems can be equivalent to non-convex ones. There are some common transformations that preserve convexity, and which sometimes transform the problem into one that's easier to deal with. The standard-form convex problem in Def. 3.6 is equivalent to the following one:

$$\begin{aligned} \text{minimize (over } z) : \;\; & f_0(Fz + x_0) \\ \text{subject to} \;\; : \;\; & f_i(Fz + x_0) \leq 0, \; i = 1, ..., m \end{aligned}$$

where F and x_0 are such that

$$Ax = b \iff x = Fz + x_0 \;\text{ for some }\; z.$$

Similarly, there is a technique known as introducing slack variables for linear inequalities. The problem

$$\text{minimize } f_0(x) \text{ subject to } a_i^T x \le b_i, \quad i = 1, \dots, m$$

is equivalent to

$$\begin{aligned} \text{minimize (over } x, s) : \ & f_0(x) \\ \text{subject to } : \ & a_i^T x + s_i = b_i, \ \text{ and } \ s_i \ge 0, \ \ i = 1, \dots, m, \end{aligned}$$

Definition 3.7. For a general optimization problem in standard form (not necessarily convex), define the **Lagrangian** $L : \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^p \to \mathbb{R}$, with dom$L = \mathscr{D} \times \mathbb{R}^m \times \mathbb{R}^p$ as follows:

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

The elements of λ, ν are called Lagrange multipliers. Define the **Lagrange dual** function: $g : \mathbb{R}^m \times \mathbb{R}^p \to R$, as follows:

$$g(\lambda, \nu) = \inf_{x \in \mathscr{D}} L(x, \lambda, \nu)$$

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

Theorem 3.4. *If $\lambda \succeq 0$, then $g(\lambda, \nu) \le p^*$.*

Proof. For any feasible $\tilde{x}$, one has

$$f_0(\tilde{x}) \ge L(\tilde{x}, \lambda, \nu) \ge \inf_{x \in \mathscr{D}} L(x, \lambda, \nu) = g(\lambda, \nu)$$

where we used that $\lambda \succeq 0$. Minimizing over all feasible $\tilde{x}$ gives $p^* \ge g(\lambda, \nu)$. □

The preceding theorem is aptly named the *lower bound property*.

Consider the minimum-norm solution of a set of linear equations:

$$\text{minimize } x^T x \ \text{ subject to } \ Ax = b$$

The Lagrangian is $L(x, \nu) = x^T x + \nu^T (Ax - b)$. To minimize L over x, set gradient equal to zero:

$$\nabla_x L(x, \nu) = 2x + A^T \nu = 0 \ \Rightarrow \ x = -\frac{1}{2} A^T \nu$$

Then

$$g(\nu) = L\left(-\frac{1}{2} A^T \nu, \nu\right) = -\frac{1}{4} \nu^T A A^T \nu - b^T \nu$$

which we can see explicitly is a concave function of ν. The lower bound property in this case states that

$$p^* \geq -\frac{1}{4}\nu^T AA^T \nu - b^T \nu$$

for all ν.

The Lagrange dual is closely related to the conjugate function which we introduced above in (3.2). Consider the problem

$$\begin{aligned} &\text{minimize } f_0(x) \\ &\text{subject to } Ax \preceq b, Cx = d \end{aligned}$$

The Lagrange dual function is

$$\begin{aligned} g(\lambda, \nu) &= \inf_{x \in \mathrm{dom} f_0} \left(f_0(x) + (A^T\lambda + C^T\nu)^T x - b^T\lambda - d^T\nu \right) \\ &= -f_0^*(-A^T\lambda - C^T\nu) - b^T\lambda - d^T\nu \end{aligned}$$

Hence it greatly simplifies the computation of the dual if the conjugate of f_0 is known.

Recall that by Theorem 3.4, if $\lambda \succeq 0$, then $g(\lambda, \nu) \leq p^*$. This provides all sorts of lower bounds for p^*, but some are better than others. This suggests an idea: what if we maximized $g(\lambda, \nu)$ over the space of all $\lambda \succeq 0$? This is a fruitful idea. This leads us to formulate the **Lagrange dual problem**:

$$\text{maximize } g(\lambda, \nu) \text{ subject to } \lambda \succeq 0.$$

The idea is that if we could solve this problem, we would know the best lower bound on p^* that could be obtained by applying the lower bound property of the dual function. This is generally a convex optimization problem, even if the original one wasn't! The optimal value of the dual problem is typically denoted d^*. The pair λ, ν are said to be dual feasible if $\lambda \succeq 0$, and $(\lambda, \nu) \in \mathrm{dom}(g)$.

As an example, consider the standard form LP and its dual:

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

The statement $d^* \leq p^*$ always holds, and is known as *weak duality*. It is sometimes useful to find nontrivial lower bounds for difficult problems. The equality $d^* = p^*$ does not usually hold. When it holds, it is known as *strong duality*. It usually does hold for convex problems, under conditions that we'll make precise very soon. The difference $p^* - d^*$ is known as the *duality gap*. It is always positive.

Any condition that guarantees strong duality in convex problems is called a *constraint qualification*. The most well-known one is Slater's constraint qualification.

Theorem 3.5 (Slater's condition)**.** *Strong duality holds for a convex problem (cf. Definition 3.6) if it is strictly feasible, i.e.,*

$$\exists\, x \in \operatorname{int}(\mathscr{D}) \;:\; f_i(x) < 0 \text{ for } \; i = 1, \ldots, m,\; Ax = b$$

Under these conditions, the dual optimum is attained if $p^* > -\infty$.

The theorem can be sharpened in various ways. For example, linear inequalities do not need to hold with strict inequality. Note that there exist non-convex problems with strong duality, it just isn't guaranteed in general.

REFERENCES

Boyd, Stephen and Lieven Vandenberghe (2009). *Convex optimization*. Cambridge university press.