Math Review Notes—Convex Optimization

Gregory Faletto

Contents

Last updated May 23, 2019

1 Convex Optimization

These are my notes from taking EE 588 at USC taught by Mahdi Soltanolkotabi and the textbook *Convex Optimization* (Boyd and Vandenberghe) 7th printing [Boyd et al., 2004], as well as Math 541A at USC taught by Steven Heilman.

Need to cover:

- Update rules for optimization problems (e.g. gradient descent, be able to write down gradient, etc.)
- Know which algorithms are useful in which settings
- Homework-like problems from first part of class (no proofs though) (Boyd homework is good practice)
- Understand how to derive algorithms
- Understand how to calculate gradients, proximal functions, etc.
- Understand examples, how to run algorithms
- Only conceptual thing: duality question (write down dual)
- Formulate problems as convex optimization problems

Do not need to cover:

- ADMM
- Proofs from 2nd half of class (rates of convergence, etc.)
- Coding

1.1 Convex Functions

Definition 1.1 (Math 541A definition). Let $\phi : \mathbb{R} \to \mathbb{R}$. We say that ϕ is **convex** if, for any $x, y \in \mathbb{R}$ and for any $t \in [0, 1]$, we have

$$\phi(tx + (1-t)y) \leq t\phi(x) + (1-t)\phi(y).$$

Definition 1.2 (Strict convexity, Math 541A notes definition 6.6). Let $\phi : \mathbb{R} \to \mathbb{R}$. We say that ϕ is **strictly convex** if, for any $x, y \in \mathbb{R}, x \neq y$ and for any $t \in (0, 1)$, we have

$$\phi(tx + (1-t)y) < t\phi(x) + (1-t)\phi(y).$$

Definition 1.3 (Convex function in $\mathbb{R}^n$, Math 541A Definition). Let $\phi : \mathbb{R}^n \to \mathbb{R}$. We say that ϕ is **convex** if, for any $x, y \in \mathbb{R}^n$ and for any $t \in [0, 1]$, we have

$$\phi(tx + (1-t)y) \leq t\phi(x) + (1-t)\phi(y). \tag{1}$$

Lemma 1 (Result from Math 541A Homework 2)**.** The slope of a convex function is nondecreasing. More formally, let $\phi : \mathbb{R} \to \mathbb{R}$ be a convex function. For any $x \in \mathbb{R}$, let

$$M_R := \left\{ \frac{\phi(c) - \phi(x)}{c - x} : c > x \right\}, \quad M_L := \left\{ \frac{\phi(x) - \phi(b)}{x - b} : b > x \right\}$$

be the slopes of the secant lines through ϕ using points to the right and left of x, respectively. Then for any $m \in M_R$, $p \in M_L$ we have $m \geq p$.

Proof. Fix $x \in \mathbb{R}$. Let $m \in M_R, p \in M_L$. By definition, there exist $b < x < c$ such that

$$m = \frac{\phi(c) - \phi(x)}{c - x}, \quad p = \frac{\phi(x) - \phi(b)}{x - b}.$$

Let $t \in (0, 1)$ such that

$$tb + (1 - t)c = x. \tag{2}$$

Then we have

$$m \geq p \iff \frac{\phi(c) - \phi(x)}{c - x} \geq \frac{\phi(x) - \phi(b)}{x - b} \iff (x - b)(\phi(c) - \phi(x)) \geq (c - x)(\phi(x) - \phi(b))$$

$$\iff (x - b)\phi(c) + b\phi(x) \geq c\phi(x) - (c - x)\phi(b)$$

From (2), we have $x - b = tb + (1 - t)c - b = (t - 1)b + (1 - t)c = (1 - t)(c - b)$ and $t(b - c) = x - c \iff t(c - b) = c - x$. Therefore

$$(x - b)\phi(c) + b\phi(x) \geq c\phi(x) - (c - x)\phi(b) \iff (1 - t)(c - b)\phi(c) + b\phi(x) \geq c\phi(x) - t(c - b)\phi(b)$$

$$\iff (1 - t)(c - b)\phi(c) \geq (c - b)\phi(x) - t(c - b)\phi(b) \iff (1 - t)\phi(c) + t\phi(b) \geq \phi(x)$$

But $t\phi(b) + (1 - t)\phi(c) \geq \phi(x)$ since ϕ is convex. Therefore $m \geq p$.

□

Theorem 2 (**Result from 541A Homework 2; equivalent conditions for convexity**)**.** Let $\phi : \mathbb{R} \to \mathbb{R}$. Then ϕ is convex if and only if: for any $y \in \mathbb{R}$, there exists a constant a and there exists a function $L : \mathbb{R} \to \mathbb{R}$ defined by $L(x) = a(x - y) + \phi(y)$, $x \in \mathbb{R}$, such that $L(y) = \phi(y)$ and such that $L(x) \leq \phi(x)$ for all $x \in \mathbb{R}$. (In the case that ϕ is differentiable, the latter condition says that ϕ lies above all of its tangent lines.)

Proof. $\implies$: As in Lemma 1, let

$$M_R := \{\frac{\phi(c) - \phi(y)}{c - y} : c > y\}, \qquad M_L := \{\frac{\phi(y) - \phi(b)}{y - b} : b > y\}$$

be the slopes of the secant lines through ϕ using points to the right and left of y, respectively. Then by Lemma 1, for any $m \in M_R$, $p \in M_L$ we have $m \geq p$, so we can choose some $a_0 \in \mathbb{R}$ such that $p \leq a_0 \leq m$ for all $p \in M_L, m \in M_R$. Then let $L : \mathbb{R} \to \mathbb{R}$ be defined by $L(x) = a_0(x - y) + \phi(y)$, $x \in \mathbb{R}$. Note that $L(y) = \phi(y)$.

We argue that $L(x) \leq \phi(x)$ for all $x \in \mathbb{R}$ by contradiction. Suppose there is some $z \in \mathbb{R}$ with $L(z) > \phi(z)$. Note that $z \neq y$ because we have already shown that $L(y) = \phi(y)$. Then we have

$$L(z) > \phi(z) \iff a_0(z - y) + \phi(y) > \phi(z) \tag{3}$$

If $z > y$, then we can solve (3) for a_0 as follows:

$$a_0 > \frac{\phi(z) - \phi(y)}{z - y}$$

But $z \in M_R$, so we have $\frac{\phi(z)-\phi(y)}{z-y} > a_0$. Contradiction. If $z < y$, we solve (3) for a_0 as follows:

$$a_0 < \frac{\phi(z) - \phi(y)}{z - y} = \frac{\phi(y) - \phi(z)}{y - z}$$

But $z \in M_L$, so we have $\frac{\phi(y)-\phi(z)}{y-z} < a_0$. Contradiction. Therefore for every $z \in \mathbb{R}$ we have $L(z) \leq \phi(z)$ as desired.

$\Longleftarrow$: Now suppose that for any $y \in \mathbb{R}$ there exists a constant a and a function $L : \mathbb{R} \to \mathbb{R}$ defined by $L(x) = a(x - y) + \phi(y)$, $x \in \mathbb{R}$, such that $L(y) = \phi(y)$ and such that $L(x) \leq \phi(x)$ for all $x \in \mathbb{R}$.

Fix $b, c \in \mathbb{R}$ and let $t \in (0, 1)$. Set $y := tb + (1 - t)c$. Then by assumption we can write

$$a(b - y) + \phi(y) \leq \phi(b), \qquad a(c - y) + \phi(y) \leq \phi(c)$$

Multiply by $t > 0$ and $(1 - t) > 0$ respectively to yield

$$ta(b - y) + t\phi(y) \leq t\phi(b), \qquad (1 - t)a(c - y) + (1 - t)\phi(y) \leq (1 - t)\phi(c) \tag{4}$$

Note that

$$ta(b - y) + (1 - t)a(c - y) = a(tb - ty + (c - y - ct + yt)) = a(tb + c - y - ct)$$

$$= a(tb + c(1-t) - tb - (1-t)c) = 0$$

So adding the inequalities in (4) yields

$$t\phi(y) + (1-t)\phi(y) \leq t\phi(b) + (1-t)\phi(c) \iff \phi(y) \leq t\phi(b) + (1-t)\phi(c)$$

$$\iff \phi(tb + (1-t)c) \leq t\phi(b) + (1-t)\phi(c).$$

□

My original proof from submitted homework. If ϕ is differentiable at y, let $a = \phi'(y)$. If not, let a be any subgradient of ϕ at y. Then $L(x)$ is (a) tangent line to ϕ at y, which should be lesser than or equal to ϕ for all $x \in \mathbb{R}$ if ϕ is convex. If and only if this is true at every $y \in \mathbb{R}$ (the tangent line is a global underestimator at y for every $y \in \mathbb{R}$), then ϕ must be convex. We proceed to show this formally:

$\implies$: We will show that if ϕ is convex; that is, if for any $x, y \in \mathbb{R}$ and for any $t \in [0,1]$, we have

$$\phi(tx + (1-t)y) \leq t\phi(x) + (1-t)\phi(y) \tag{5}$$

then the inequality

$$\phi'(y)(x-y) + \phi(y) \leq \phi(x) \quad \forall x \in \mathbb{R} \tag{6}$$

holds. Starting from (5) note that

$$\phi(tx + (1-t)y) \leq t\phi(x) + (1-t)\phi(y) \implies \phi(tx + (1-t)y) - \phi(x) \leq (1-t)(\phi(y) - \phi(x))$$

Suppose $y > x$. Then $tx + (1-t)y - x = (1-t)(y-x) > 0$, so we can divide by it on both sides:

$$\implies \frac{\phi(tx + (1-t)y) - \phi(x)}{tx + (1-t)y - x} \leq \frac{(1-t)(\phi(y) - \phi(x))}{(1-t)(y-x)} \implies \frac{\phi(tx + (1-t)y) - \phi(x)}{tx + (1-t)y - x} \leq \frac{\phi(y) - \phi(x)}{y-x}$$

Taking the limit as $t \to 1$ yields

$$\phi'(x) \leq \frac{\phi(y) - \phi(x)}{y-x}$$

if ϕ is differentiable, which is (equivalent to) what we hoped to prove. The case where $x > y$ is analogous.

$\impliedby$: We will show that if (6) holds then ϕ is convex; that is, (5) holds for any $x, y \in \mathbb{R}$ and for any $t \in [0,1]$. Starting from (6) note that

$$\phi'(y)(x-y)+\phi(y)\leq \phi(x) \iff \phi'(y) \leq \frac{\phi(x)-\phi(y)}{x-y} \quad \forall\ x,y \in \mathbb{R}$$

□

Theorem 3 (Global minimum of convex functions; Math 541A Homework problem). Let $f : \mathbb{R}^n \to \mathbb{R}$ be a convex function. Let $x \in \mathbb{R}^n$ be a local minimum of f. Then

(a) x is a global minimum of f.

(b) If f is strictly convex, then there is at most one global minimum of f.

(c) If f is a C^1 function (all derivatives of f exist and are continuous), and $x \in \mathbb{R}^n$ satisfies $\nabla f(x) = 0$, then x is a global minimum of f.

Proof. (a) Since x is a local minimum, we have that there exists $\epsilon > 0$ such that $f(x) \leq f(y)\ \forall\ y \in B(x,\epsilon)$ where $B(x,\epsilon) \subseteq \mathbb{R}^n$ is an n-dimensional L_2 ball of radius ϵ centered at x. Suppose there exists some $z \in \mathbb{R}^n$ such that $f(z) < f(x)$. Then by convexity of f, for $t \in [0,1]$,

$$f(tx+(1-t)z) \leq tf(x)+(1-t)f(z) < tf(x)+(1-t)f(x) = f(x)$$

which when $t = 1$ leads to the contradiction $f(x) < f(x)$. (Also, for $t = 1-\delta$ with δ sufficiently small, we get $f(x') \leq tf(x)+(1-t)f(z) < f(x)$ where $x' = tx+(1-t)z$ such that $x' \in B(x,\epsilon)$, contradicting the fact that $f(x)$ is a local minimum.) Therefore there is no $z \in \mathbb{R}^n$ such that $f(z) < f(x)$, so x is a global minimum.

(b) If f is strictly convex, for any $z \in \{\mathbb{R}^n \setminus x\}$ we have

$$f(tx+(1-t)z) < tf(x)+(1-t)f(z), \qquad \forall x,z \in \mathbb{R}^n, x \neq z \tag{7}$$

We have already shown that there exists no $z \in \{\mathbb{R}^n \setminus x\}$ such that $f(z) < f(x)$. Suppose there is more than one global minimum of f; that is, there exists $z \in \{\mathbb{R}^n \setminus x\}$ such that $f(z) = f(x)$. That is, for all $y \in \{\mathbb{R}^n \setminus \{x,z\}\}$,

$$f(x) = f(z) \leq f(y). \tag{8}$$

But then by strict convexity,

$$f\left(\frac{x+z}{2}\right) < \frac{1}{2}f(x)+\frac{1}{2}f(z) = \frac{1}{2}f(x)+\frac{1}{2}f(x) = f(x)$$

which contradicts (8) if $y = (x+z)/2$. Therefore the global minimum of f is unique.

(c) Recall from Exercise 4 in Homework 2 that f is convex if and only if for any $x \in \mathbb{R}^n$ there exists a constant $a \in \mathbb{R}^n$ and a function $L : \mathbb{R}^n \to \mathbb{R}$ defined by $L(y) = a^T(y-x)+f(x), y \in \mathbb{R}^n$ such that $L(x) = f(x)$ and $L(y) \leq f(y)$ for all $y \in \mathbb{R}^n$. Further, if f is a C^1 function then this function exists for $a = \nabla f(x)$. That is,

$$f(y) \geq f(x) + \nabla f^T(x)(y-x),\ \forall y \in \mathbb{R}^n.$$

Since $\nabla f(x) = 0$, if we plug in $y = x$ we get

$$f(y) \geq f(x), \quad \forall y \in \mathbb{R}^n.$$

□

Theorem 4 (Jensen's Inequality, from Math 541A). Let $X : \Omega \to [-\infty, \infty]$ be a random variable. Let $\phi : \mathbb{R} \to \mathbb{R}$ be convex. Assume that $\mathbb{E}|X| < \infty$ and $\mathbb{E}|\phi(X)| < \infty$. Then

$$\phi(\mathbb{E}X) \leq \mathbb{E}\phi(X).$$

Proof. Note that from Theorem 2, for any $y \in \mathbb{R}$ there exists a constant a and a function L such that

$$a(x - y) + \phi(y) \leq \phi(x) \quad \forall x \in \mathbb{R}$$

Letting $y = \mathbb{E}(X)$ we have

$$a(X - \mathbb{E}X) + \phi(\mathbb{E}X) \leq \phi(X)$$

Since expectations preserve inequalities,

$$\mathbb{E}[a(X - \mathbb{E}X) + \phi(\mathbb{E}X)] \leq \mathbb{E}\phi(X)$$

But

$$\mathbb{E}[a(X - \mathbb{E}X) + \phi(\mathbb{E}X)] = a(\mathbb{E}X - \mathbb{E}X) + \mathbb{E}(\phi(\mathbb{E}X)) = \phi(\mathbb{E}X)$$

which yields

$$\phi(\mathbb{E}X) \leq \mathbb{E}\phi(X).$$

□

Corollary 4.1 (Jensen's Inequality: EE 588 Formulation). f is convex if and only if

$$f\left(\frac{a+b}{2}\right) \leq \frac{f(a) + f(b)}{2}$$

for all $a, b \in \mathbf{dom}(f)$.

Proof. Follows from Theorem 4 if X is a discrete random variable that equals a or b each with probability $1/2$ and $\phi(X) = f(X)$. Note that $\phi(X)$ is convex.

□

Corollary 4.2 (**Triangle Inequality**)**.** Let $X : \Omega \to [-\infty, \infty]$ be a random variable with $\mathbb{E}|X| < \infty$. Then

$$\|EX| \leq \mathbb{E}|X|.$$

Proof. Note that $\phi(x) = |x|$ is convex by the definition of convexity: for any $x, y \in \mathbb{R}$ and for any $t \in (0, 1)$, we have

$$\phi(tx + (1-t)y) = |tx + (1-t)y| \leq \ldots = t|x| + (1-t)|y| = t\phi(x) + (1-t)\phi(y).$$

Then the result follows immediately from Jensen's Inequality (Theorem 4) using $\phi(X) = |X|$:

$$|\mathbb{E}X| \leq \mathbb{E}|X|$$

□

Theorem 5 (**Conditional Jensen Inequality**)**.** Let $X, Y : \Omega \to \mathbb{R}$ be random variables that are either both discrete or both continuous. Let $\phi : \mathbb{R} \to \mathbb{R}$ be convex. Then

$$\phi(\mathbb{E}(X|Y)) \leq \mathbb{E}(\phi(X)|Y).$$

If ϕ is strictly convex, then equality holds only if X is constant on any set where Y is constant. That is, (by an Exercise from the previous homework) equality holds only if X is a function of Y.

Proof. Recall that from Exercise 4 in Homework 2 that since ϕ is convex, for any $y \in \mathbb{R}$ there exists a constant a and a function L such that

$$a(x - y) + \phi(y) \leq \phi(x) \quad \forall x \in \mathbb{R}$$

Letting $x = X$ and $y = \mathbb{E}(X \mid Y)$ we have

$$a(X - \mathbb{E}(X \mid Y)) + \phi(\mathbb{E}(X \mid Y)) \leq \phi(X)$$

Since by Lemma **??** conditional expectations preserve inequalities,

$$\mathbb{E}[a(X - \mathbb{E}[X \mid Y]) + \phi(\mathbb{E}[X \mid Y]) \mid Y] \leq \mathbb{E}(\phi(X) \mid Y)$$

But

$$\mathbb{E}[a(X - \mathbb{E}[X \mid Y]) + \phi(\mathbb{E}[X \mid Y]) \mid Y] = a(\mathbb{E}[X \mid Y] - \mathbb{E}[\mathbb{E}(X \mid Y) \mid Y]) + \mathbb{E}[\phi(\mathbb{E}[X \mid Y]) \mid Y].$$

By Corollary **??** (letting $h(Y) = \phi(\mathbb{E}[X \mid Y])$), $\mathbb{E}[\phi(\mathbb{E}[X \mid Y]) \mid Y] = \phi(\mathbb{E}[X \mid Y])$. By Corollary **??**, $\mathbb{E}[\mathbb{E}(X \mid Y)) \mid Y] = \mathbb{E}(X \mid Y)$. Therefore we have

$$= a(\mathbb{E}[X \mid Y] - \mathbb{E}[X \mid Y]) + \phi(\mathbb{E}[X \mid Y]) = \phi(\mathbb{E}(X \mid Y))$$

which yields

$$\phi(\mathbb{E}(X \mid Y)) \leq \mathbb{E}(\phi(X) \mid Y).$$

□

Proposition 6 (Convexity of affine functions). Let $c \in \mathbb{R}^n$ and $d \in \mathbb{R}$ be fixed. Let $x \in \mathbb{R}^n$. Then the function $f(x) = c^T x + d$ is convex.

Proof. We will show that f satisfies (1) for any $x, y \in \mathbb{R}^n$ and any $t \in [0, 1]$:

$$f(tx + (1-t)y) = c^T(tx + (1-t)y) + d = c^T(tx + (1-t)y) + d = tc^T x + td + (1-t)c^T y + (1-t)d$$

$$= t[c^T x + d] + (1-t)[c^T y + d] = tf(x) + (1-t)f(y).$$

In particular, (1) is satisfied with equality.

□

Proposition 7 (Convexity of quadratic forms). Let $A \in \mathbb{R}^{m \times n}$ and $k > 0$ be fixed. Let $x \in \mathbb{R}^n$. Then the function $f(x) = kx^T A^T Ax$ is convex.

Proof. We will show that f satisfies the definition of convexity. Plugging $\phi(x) = kx^T A^T Ax$ into the left side of (1), we have

$$k(tx + (1-t)y)^T A^T A(tx + (1-t)y) = k(tx^T + (1-t)y^T)A^T A(tx + (1-t)y)$$

$$= k\big[tx^T A^T Atx + tx^T A^T A(1-t)y + (1-t)y^T A^T Atx + (1-t)y^T A^T A(1-t)y\big]$$

$$= k\big[t^2 x^T A^T Ax + t(1-t)x^T A^T Ay + t(1-t)y^T A^T Ax + (1-t)^2 y^T A^T Ay\big]$$

Note that $x^T A^T Ay \in \mathbb{R} = [x^T A^T Ay]^T = y^T A^T Ax$, so we have

$$= k\big[t^2 x^T A^T Ax + 2t(1-t)x^T A^T Ay + (1-t)^2 y^T A^T Ay\big] \tag{9}$$

Plugging $\phi(x) = kx^T A^T Ax$ into the right side of (1) yields

$$ktx^T A^T Ax + k(1-t)y^T A^T Ay \tag{10}$$

We can verify the inequality in (1) by subtracting (10) from (9) to see if a negative number results:

$$kt^2x^TA^TAx + 2kt(1-t)x^TA^TAy + k(1-t)^2y^TA^TAy - ktx^TA^TAx - k(1-t)y^TA^TAy$$

$$= kt(t-1)x^TA^TAx + 2kt(1-t)x^TA^TAy + k(1-t)[1-t-1]y^TA^TAy$$

$$= -kt(1-t)[x^TA^TAx - 2x^TA^TAy + y^TA^TAy] = -kt(1-t)[x^TA^T - y^TA^T][Ax - Ay]$$

$$= -kt(1-t)[A(x-y)]^TA(x-y) \leq 0$$

for all $x, y \in \mathbb{R}^n$ and any $t \in [0,1]$ since $-kt(1-t) \leq 0$ (with equality only when $t = 0$ or $t = 1$) and $[A(x-y)]^TA(x-y) \geq 0$ (with equality only when $x = y$). This verifies the inequality in (1), which proves that kx^TA^TAx is convex.

□

Proposition 8 (Sum of convex functions is convex). Let $f_1, \ldots, f_n : \mathbb{R}^n \to \mathbb{R}$ be (strictly) convex functions. Then the function $g(x) := \sum_{i=1}^n f_i(x)$ is (strictly) convex.

Proof. Since f_i is convex for all $i \in \{1, \ldots, n\}$, f_i satisfies

$$f_i(tx + (1-t)y) \leq tf_i(x) + (1-t)f_i(y), \qquad \forall i \in \{1, \ldots, n\}.$$

We make use of these inequalities to show that g satisfies (1) for any $x, y \in \mathbb{R}^n$ and any $t \in [0,1]$:

$$\begin{aligned} g(tx + (1-t)y) &= \sum_{i=1}^n f_i(tx + (1-t)y) \leq \sum_{i=1}^n [tf_i(x) + (1-t)f_i(y)] \\ &= t\sum_{i=1}^n f_i(x) + (1-t)\sum_{i=1}^n f_i(y) = tg(x) + (1-t)g(y) \end{aligned}$$

which proves the result. (Note that if the initial inequality is strict then strict convexity follows.)

□

Proposition 9 (Exercise 6.43 in Math 541A Lecture Notes). Let $f_1, \ldots, f_n : \mathbb{R} \to \mathbb{R}$ be n strictly convex functions. Define $g : \mathbb{R}^n \to \mathbb{R}$ by

$$g(x_1, \ldots, x_n) := \sum_{i=1}^n f_i(x_i), \qquad \forall (x_1, \ldots, x_n) \in \mathbb{R}^n.$$

Then $g : \mathbb{R}^n \to \mathbb{R}$ is convex.

Proof. Since f_i is strictly convex for all $i \in \{1, \ldots, n\}$, we have that for any $x_i, y_i \in \mathbb{R}$, for all $t \in (0,1)$

$$f_i(tx_i + (1-t)y_i) < tf_i(x_i) + (1-t)f_i(y_i).$$

Therefore for any $x, y \in \mathbb{R}^n$ (where $x = (x_1, \ldots, x_n), y = (y_1, \ldots, y_n)$), for all $t \in (0,1)$

$$g(tx + (1-t)y) = \sum_{i=1}^{n} f_i(tx_i + (1-t)y_i) < \sum_{i=1}^{n} [tf_i(x_i) + (1-t)f_i(y_i)] = t\sum_{i=1}^{n} f_i(x_i) + (1-t)\sum_{i=1}^{n} f_i(y_i)$$
$$= tg(x) + (1-t)g(y).$$

□

Proposition 10 (Exercise 6.44 from Math 541A lecture notes). Let $f : \mathbb{R}^n \to \mathbb{R}$. Suppose that for any fixed $i \in \{1, \ldots, n\}$ and for any $x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n$, the function

$$x_i \mapsto f(x_1, \ldots, x_n)$$

is strictly convex. Then f has at most one global minimum.

Proof. An equivalent statement to our assumption is that for any i, f is strictly convex in x_i keeping $(x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n)$ fixed. That is, if we let $h_i : \mathbb{R} \to \mathbb{R}$ be defined for all $i \in \{1, \ldots, n\}$ by

$$h_i\big(x_i \mid (x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n)\big) := f\big((x_1, \ldots, x_n)\big) \qquad \forall i \in \{1, \ldots, n\},$$

then h_i is strictly convex for all $i \in \{1, \ldots, n\}$. That is, for any $x_i, y_i \in \mathbb{R}$, for all $t \in (0,1)$

$$h_i\big(tx_i + (1-t)y_i \mid (tx_1 + (1-t)y_1, \ldots, tx_{i-1} + (1-t)y_{i-1}, tx_{i+1} + (1-t)y_{i+1}, \ldots, tx_n + (1-t)y_n)\big)$$

$$< th_i\big(x_i \mid (x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n)\big) + (1-t)h_i\big(y_i \mid (y_1, \ldots, y_{i-1}, y_{i+1}, \ldots, y_n)\big). \tag{11}$$

By Theorem 3(b), there is at most one global minimum of f if it is strictly convex, so all we need to show is that f is strictly convex. That is, we must show that for all $(x_1, \ldots, x_n), (y_1, \ldots, y_n) \in \mathbb{R}^n$, for all $t \in (0,1)$,

$$f\big(t(x_1, \ldots, x_n) + (1-t)(y_1, \ldots, y_n)\big) < tf\big((x_1, \ldots, x_n)\big) + (1-t)f\big((y_1, \ldots, y_n)\big). \tag{12}$$

We will argue by contradiction. Suppose that for some $(x_1^*, \ldots, x_n^*) \in \mathbb{R}^n$ and $(y_1^*, \ldots, y_n^*) \in \mathbb{R}^n$, (12) does not hold. That is,

$$f\big(t^*(x_1^*, \ldots, x_n^*) + (1-t^*)(y_1^*, \ldots, y_n^*)\big) \geq t^*f\big((x_1^*, \ldots, x_n^*)\big) + (1-t^*)f\big((y_1^*, \ldots, y_n^*)\big).$$

for some $t^* \in (0,1)$. This is equivalent to

$$\begin{aligned} &h_1\big(t^*x_1^* + (1-t^*)y_1^* \mid (t^*x_2^* + (1-t^*)y_2^*, \ldots, t^*x_n^* + (1-t^*)y_n^*)\big) \\ &\quad \geq t^*h_1\big(x_1^* \mid (x_2^*, \ldots, x_n^*)\big) + (1-t^*)h_1\big(y_1^* \mid (y_2^*, \ldots, y_n^*)\big). \end{aligned}$$

But this contradicts (11). Therefore (12) holds for all $(x_1, \ldots, x_n), (y_1, \ldots, y_n) \in \mathbb{R}^n$, for all $t \in (0,1)$, so f is strictly convex, which means (by Theorem 3(b)) that f has at most one global minimum.

□

Proposition 11. Let A be a real $m \times n$ matrix. Let $x \in \mathbb{R}^n$ and let $b \in \mathbb{R}^m$. Then the function $f\colon \mathbb{R}^n \to \mathbb{R}$ defined by $f(x) = \frac{1}{2}\|Ax - b\|^2$ is convex.

Proof. We have

$$f(x) = \frac{1}{2}\|Ax-b\|^2 = \frac{1}{2}(Ax-b)^T(Ax-b) = \frac{1}{2}(x^TA^T - b^T)(Ax-b)$$

$$= \frac{1}{2}(x^TA^TAx - b^TAx - x^TA^Tb + b^Tb) = \frac{1}{2}x^TA^TAx - b^TAx + \frac{1}{2}b^Tb$$

where the last step follows because $b^TAx \in \mathbb{R} = (b^TAx)^T = x^TA^Tb$ since a real number equals its transpose. The affine function $-b^TAx + \frac{1}{2}b^Tb$ is convex by Proposition 6, and the quadratic form $\frac{1}{2}x^TA^TAx$ is convex by Proposition 7. Since the sum of convex functions is convex by Proposition 8, the result follows.

□

Remark. Moreover,

$$\nabla f(x) = A^T(Ax-b), \qquad D^2f(x) = A^TA.$$

(Here D^2f denotes the matrix of second derivatives of f.)

So, if $\nabla f(x) = 0$, i.e. if $A^TAx = A^Tb$, then x is the global minimum of f. And if A has full rank, then A^TA is invertible, so that $x = (A^TA)^{-1}A^Tb$ is the global minimum of f.

Proposition 12 (**Convexity of norms**)**.** Every norm on $\mathbb{R}^n$ is convex.

Proof. Suppose we have a generic norm $\|\cdot\|_*$ in $\mathbb{R}^n$. Because $\|\cdot\|_*$ is a norm, it satisfies the triangle inequality; that is, for all $x, y \in \mathbb{R}^n$,$\|x+y\|_* \leq \|x\|_* + \|y\|_*$. Further, for any $t \in [0,1]$, we have

$$\|tx + (1-t)y\|_* \leq \|tx\|_* + \|(1-t)y\|_* = t\|x\|_* + (1-t)\|y\|_*$$

where the last step also follows from a property of all norms.

□

Proposition 13 (**Mentioned in in-class 541A review; might have been on HW?**)**.** If ϕ is strictly convex and $\mathbb{E}(\phi(X)) = \phi(\mathbb{E}(X))$ then X is almost surely constant.

1.2 Schur Complement Trick

1.2.1 Definition

For a matrix $X \in \boldsymbol{S}^n$ partitioned as

$$X = \begin{bmatrix} A & B \\ B^T & C \end{bmatrix}$$

the Schur complement is (if $\mathbf{det}(A) \neq 0$)

$$S = C - B^T A^{-1} B$$

The Schur complement has two useful properties in convex analysis.

Theorem 14. (a) $X \succ 0$ if and only if $A \succ 0$ and $S \succ 0$.

(b) If $A \succ 0$, then $X \succeq 0$ if and only if $S \succeq 0$.

1.2.2 The Trick

Suppose we are trying to express a problem as a semidefinite program (SDP); that is, in the form

$$\begin{array}{ll} \text{minimize} & c^T x \\ \text{subject to} & x_1 F_1 + \ldots + x_n F_n + G \preceq 0 \\ & Ax = b \end{array}$$

where $G, F_1, \ldots, F_n \in \boldsymbol{S}^k$ and $A \in \mathbb{R}^{p \times n}$. If we have a constraint of the form $c^T F(x)^{-1} c \leq t$ where $F(x)$ is symmetric and positive definite and $t \in \mathbb{R}$, by Theorem 14(b) we can write

$$c^T F(x)^{-1} c \leq t \iff \begin{bmatrix} F(x) & c \\ c^T & t \end{bmatrix} \succeq 0$$

in order to get our constraint in the form required for an SDP.

1.2.3 Example 1: Last Year's Final, Question 2(b)

Suppose we have the constraints

$$Ax + b \geq 0$$
$$\frac{(c^T x)^2}{d^T x} \leq t$$

which we would like to express in an SDP. By Theorem 14(b) we can write

$$\frac{(c^Tx)^2}{d^Tx} \leq t \iff d^Tx - (c^Tx)^T t^{-1} c^Tx \geq 0 \iff \begin{bmatrix} t & c^Tx \\ c^Tx & d^Tx \end{bmatrix} \succeq 0$$

Since

$$Ax + b \geq 0 \iff \mathbf{diag}(Ax+b) \succeq 0$$

we can finally write our constraints as

$$\begin{bmatrix} \mathbf{diag}(Ax+b) & 0 & 0 \\ 0 & t & c^Tx \\ 0 & c^Tx & d^Tx \end{bmatrix} \succeq 0$$

1.2.4 Example 2: Last Year's Final, Question 4(b)

Suppose we have the constraints

$$\begin{gathered} Ax + b \geq 0 \\ \frac{(c^Tx)^2}{d^Tx} \leq t \end{gathered}$$

which we would like to express in an SDP. By Theorem 14(b) we can write

$$\frac{(c^Tx)^2}{d^Tx} \leq t \iff d^Tx - (c^Tx)^T t^{-1} c^Tx \geq 0 \iff \begin{bmatrix} t & c^Tx \\ c^Tx & d^Tx \end{bmatrix} \succeq 0$$

Since

$$Ax + b \geq 0 \iff \mathbf{diag}(Ax+b) \succeq 0$$

we can finally write our constraints as

$$\begin{bmatrix} \mathbf{diag}(Ax+b) & 0 & 0 \\ 0 & t & c^Tx \\ 0 & c^Tx & d^Tx \end{bmatrix} \succeq 0$$

1.3 Duality

Theorem 15. Slater's condition/constraint qualification: Strong duality holds for a convex problem

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

if it is strictly feasible, i.e., there exists at least one x in the domain of f_0 such that $f_i(x) < 0, \; i = 1, 2, \ldots, m$, $Ax = b$.

1.4 MLE estimates

For linear estimates with iid noise

$$y_i = a_i^T x + v_i, i = 1, \ldots, m$$

where a is observed and $x \in \mathbb{R}^n$ are the parameters to be estimated, the likelihood function is

$$p_x(y) = \prod_{i=1}^m \Pr(v_i = y_i - a_i^T x \mid x)$$

Therefore the log likelihood function is:

$$\ell_x(y) = \sum_{i=1}^m \log[\Pr(v_i = y_i - a_i^T x \mid x)]$$

1.5 Practice Final (2017 Final)

(1) (a) Strictly convex. Multiply by x/x (allowed in this case since $x > 0$) to get $\frac{x^2}{x+1}$ which is a quadratic over linear, which is convex in $\mathbb{R}^{++}$ according to CVX rules.

(b) Not convex, it is convex for $x \geq -1$, but there is a boundary problem at $x = -1$. Note that Jensen's inequality (Theorem 4.1)

$$\frac{f(a) + f(b)}{2} \geq f\left(\frac{a+b}{2}\right)$$

is violated because

$$\frac{f(-1.3) + f(-0.9)}{2} = \frac{2.3 + 0}{2} = 1.15 \leq 2.2 = f(-1.1) = f\left(\frac{-1.3 + -0.9}{2}\right)$$

(c)

(d)

$$f(x) = \sup \log\left(\frac{p(t)}{q(t)}\right) = \sup\{\log p(t) - \log q(t)\} = \sup\{\log\left(\sum_{i=1}^n \exp(x_i \sin(it))\right) - \sum_{i=1}^n x_i \sin(it)\}$$

(e) The proximal mapping is

$$\operatorname{prox}_{\mathcal{R}}(z) = \arg\min_y \frac{1}{2}\|z-y\|_2^2 + \mathcal{R}(y) = \arg\min_y \frac{1}{2}\sum_{i=1}^n (z_i - y_i)^2 + \sum_{i=1}^n w_i|y_i|$$

$$= \arg\min_y \frac{1}{2}\sum_{i=1}^n \left[(z_i - y_i)^2 + w_i|y_i|\right]$$

Taking the gradient of the inside quantity with respect to y, we have

$$\nabla(y) = \begin{pmatrix} \frac{1}{2}\cdot 2(z_1 - y_1) + \textbf{sign}(y_1)w_1 \\ \frac{1}{2}\cdot 2(z_2 - y_2) + \textbf{sign}(y_2)w_2 \\ \vdots \\ \frac{1}{2}\cdot 2(z_n - y_n) + \textbf{sign}(y_n)w_n \end{pmatrix} = \begin{pmatrix} z_1 - y_1 + \textbf{sign}(y_1)w_1 \\ z_2 - y_2 + \textbf{sign}(y_2)w_2 \\ \vdots \\ z_n - y_n + \textbf{sign}(y_n)w_n \end{pmatrix}$$

Setting equal to 0, we have

$$y = \begin{pmatrix} z_1 \pm w_1 \\ z_2 \pm w_2 \\ \vdots \\ z_n \pm w_n \end{pmatrix}$$

(2) (a) The constraint is convex (affine). The denominator is affine. Since $c^Tx = x^Tc$, the numerator

$$(c^Tx)^2 = (c^Tx)(c^Tx) = x^Tcc^Tx = x^T(cc^T)x$$

is convex since cc^T is positive semidefinite.

(b) We start by using the epigraph trick to transform the problem:

$$\begin{aligned} \text{minimize} \quad & t \\ \text{subject to} \quad & \frac{(c^Tx)^2}{d^Tx} \leq t \\ & Ax + b \geq 0 \end{aligned}$$

We are trying to express this problem as a semidefinite program (SDP); that is, in the form

$$\begin{aligned} \text{minimize} \quad & c^Tx \\ \text{subject to} \quad & x_1F_1 + \ldots + x_nF_n + G \preceq 0 \\ & Ax = b \end{aligned}$$

where $G, F_1, \ldots, F_n \in \boldsymbol{S}^k$ and $A \in \mathbb{R}^{p\times n}$. The first constraint

$$\frac{(c^Tx)^2}{d^Tx} \leq t$$

can be expressed in the form

$$(c^Tx)^2 \leq td^Tx \iff (c^Txc^T - td^T)x \leq 0$$

We have a constraint

$$Ax + b \geq 0$$

which can be expressed in the form

$$Ax \geq -b$$

$$c^T F(x)^{-1} c \leq t$$

where $F(x)$ is symmetric and positive definite and $t \in \mathbb{R}$, by Theorem 14(b) we can write

$$c^T F(x)^{-1} c \leq t \iff \begin{bmatrix} F(x) & c \\ c^T & t \end{bmatrix} \succeq 0$$

in order to get our constraint in the form required for an SDP.

(3) (a) Yes, g is convex over $\mathcal{X}$ since it is quadratic over linear.

(b) The only points satisfying the constraint have $x_1 = 0$. Therefore the primal optimal value (the only feasible value) is $e^0 = \boxed{1}$.

(c) Lagrangian:

$$L(x, \lambda) = e^{-x_1} + \lambda(x_1^2/x_2)$$

The Lagrangian obtains its minimum value of 0 when $x_2 = x_1^3$ and $x_1 \to \infty$. Thus, its dual function ($g(\lambda) = \min_x L(x, \lambda)$) is

$$g(\lambda) = 0$$

The dual problem is then

$$\boxed{\begin{array}{ll} \text{maximize} & 0 \\ \text{subject to} & \lambda \geq 0 \end{array}}$$

(d) The optimal value of the dual problem is 0. Strong duality does not hold since the optimum of the dual problem is less than the optimum of the primal problem. We can also tell this because Slater's Condition (Theorem 15) is violated; that is, there is no (x_1, x_2) that is strictly feasible since x_1 must equal 0, which is on the boundary of the feasible region.

(e) Now for the primal problem, instead of $x_1 = 0$, we have

$$\frac{x_1^2}{x_2} \leq u \iff x_1^2 \leq ux_2 \implies -\sqrt{ux_2} \leq x_1 \leq \sqrt{ux_2}$$

Since e^{-x_1} is minimized as $x_1 \to \infty$, our optimal solution is $x_2 \to \infty, x_1 = \sqrt{ux_2} \to \infty$ yielding a primal optimal value of $\boxed{0}$. For the dual problem, we have

$$L(x, \lambda) = e^{-x_1} + \lambda\left(\frac{x_1^2}{x_2} - u\right)$$

Dual function ($g(\lambda) = \min_x L(x, \lambda)$):

$$\frac{x_1^2}{x_2} - u = 0 \implies x_2 = \frac{x_1^2}{u}$$

and let $x_1 \to -\infty$ to yield

$$g(\lambda) = 0$$

The dual problem is then

$$\boxed{\text{maximize} \quad 0}$$

with optimal value 0, so there is no longer a duality gap. We can also tell this because Slater's Condition (Theorem 15) is satisfied; that is, there exists an (x_1, x_2) which is strictly feasible (say $(x_1, x_2) = (\sqrt{u}, 10)$.

(4) (a) Yes, the set is convex. If $(u_i, v_i) = \boldsymbol{u}_i$, each

$$\sqrt{(x-u_i)^2 + (y-v_i)^2} = \|\boldsymbol{x} - \boldsymbol{u}_i\|_2$$

is convex in $\boldsymbol{x}$. Therefore the function

$$\sum_{i=1}^{k} \|\boldsymbol{x} - \boldsymbol{u}_i\|_2$$

is convex. For any fixed d, this set is a sublevel set of this function, which is convex since the function is convex.

(b) This is a feasibility problem:

$$\begin{array}{ll} \text{find} & \boldsymbol{x} \\ \text{subject to} & \displaystyle\sum_{i=1}^{k} \|\boldsymbol{x} - \boldsymbol{u}_i\| \leq d \\ & \displaystyle\sum_{i=1}^{j} \|\boldsymbol{x} - \boldsymbol{v}_i\| \leq e \end{array}$$

or

$$\begin{array}{ll} \text{minimize} & 0 \\ \text{subject to} & \displaystyle\sum_{i=1}^{k} \|\boldsymbol{x} - \boldsymbol{u}_i\| \leq d \\ & \displaystyle\sum_{i=1}^{j} \|\boldsymbol{x} - \boldsymbol{v}_i\| \leq e \end{array}$$

for two sets of points in $\mathbb{R}^2$ $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_k$, $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_j$. We would like to express these constraints as matrix inequalities in order to have an SDP. To do this, first rewrite the problem as

$$\begin{array}{ll} \text{minimize} & 0 \\ \text{subject to} & \|\boldsymbol{x} - \boldsymbol{u}_i\| \leq t_i, i = 1, \ldots, k \\ & \|\boldsymbol{x} - \boldsymbol{v}_i\| \leq s_i, s = 1, \ldots, j \\ & \mathbf{1}^T t \leq d \\ & \mathbf{1}^T s \leq e \end{array}$$

Then note that we can use the Schur trick:

$$(\boldsymbol{x}-\boldsymbol{u}_i)^T I(\boldsymbol{x}-\boldsymbol{u}_i) \leq t_i \iff \begin{bmatrix} I & \boldsymbol{x}-\boldsymbol{u}_i \\ (\boldsymbol{x}-\boldsymbol{u}_i)^T & t_i \end{bmatrix} \succeq 0$$

and write the optimization problem as an SDP:

$$\boxed{\begin{array}{ll} \text{minimize} & 0 \\ \text{subject to} & \begin{bmatrix} I & \boldsymbol{x}-\boldsymbol{u}_i \\ (\boldsymbol{x}-\boldsymbol{u}_i)^T & t_i \end{bmatrix} \succeq 0, i=1,\ldots,k \\ & \begin{bmatrix} I & \boldsymbol{x}-\boldsymbol{v}_i \\ (\boldsymbol{x}-\boldsymbol{v}_i)^T & s_i \end{bmatrix} \succeq 0, s=1,\ldots,j \\ & \mathbf{1}^T t \leq d \\ & \mathbf{1}^T s \leq e \end{array}}$$

(5) (a) To minimize the MSE:

$$\mathcal{L}(z) = \sum_r (y_r - |a_r^T x|^2)^2$$

For MLE estimate:

$$p_x(y) = \prod_{r=1}^{m} \Pr(w_r = y_r - (a_r^T x)^2 \mid x) = \frac{1}{(y_r - (a_r^T x)^2)!} \cdot \exp\big(-(a_r^T x)^2\big) \cdot (a_r^T x)^{2[y_r - (a_r^T x)^2]}$$

Therefore the log likelihood function is:

$$\ell_x(y) = \sum_{i=1}^{m} \log[\Pr(y_i - a_i^T x \mid x)] = \sum_{i=1}^{m} \log\left[\frac{1}{(y_r - (a_r^T x)^2)!} \cdot \exp\big(-(a_r^T x)^2\big) \cdot (a_r^T x)^{2[y_r - (a_r^T x)^2]}\right]$$

$$= \sum_{i=1}^{m} \log\left[\frac{1}{(y_r - (a_r^T x)^2)!}\right] - (a_r^T x)^2 + 2[y_r - (a_r^T x)^2] \cdot \log\big[(a_r^T x)\big]$$

(b) b

(c) c

(d) d

(e) e

References

S. Boyd, S. Boyd, L. Vandenberghe, and C. U. Press. *Convex Optimization*. Berichte über verteilte messysteme. Cambridge University Press, 2004. ISBN 9780521833783. URL https://books.google.com/books?id=mYm0bLd3fcoC.