Convexity and Optimization

Statistical Machine Learning, Spring 2015

Ryan Tibshirani (with Larry Wasserman)

1 An entirely too brief motivation

1.1 Why optimization?

- Optimization problems are ubiquitous in statistics and machine learning. A huge number of problems that we consider in these disciplines (and, other disciplines) can indeed be posed as optimization tasks
- But why bother studying the details? Other people have already understood the importance of optimization and have provided us with fast software for various optimization algorithms
- Two major reasons: (1) different algorithms can perform (sometimes drastically) better or worse in different scenarios, and an understanding of why this happens requires an understanding of optimization; (2) often times, understanding a problem from the optimization perspective can contribute to our statistical understanding of the problem as well
- Since this is a theoretical course, we will ignore reason (1), and focus on reason (2)

1.2 Why convexity?

- Simply: because we can broadly understand and solve convex optimization problems. Nonconvex ones are understood and solved more on a case by case basis (this isn't entirely true)
- Historically, linear programs were the focus in the optimization community, and initially, it was thought that the major divide was between linear and nonlinear optimization problems; later people discovered that some nonlinear problems were much harder than others, and the "right" divide was between convex and nonconvex problems

1.3 Two great references

- There are many great books on convexity and optimization. They can be roughly divided into books focused on convex analysis (the turf of mathematicians) and books focused on convex optimization (the turf of engineers). Here are two such books:
  - Boyd & Vandenberghe (2004)
  - Rockafellar (1970)

  Our presentation here is based on these two excellent books, especially Boyd & Vandenberghe (2004). Combined, the two provide a pretty complete coverage

1.4 A shameless plug

- You should take our Convex Optimization course (10-725/36-725), you'll learn a lot more

2 Convex sets

2.1 Basic definitions

- A set $C \subseteq \mathbb{R}^n$ is *convex* provided that, for any $x, y \in C$ and $\theta \in [0, 1]$, we have

  $$\theta x + (1 - \theta)y \in C,$$

  i.e., the line segment joining x, y lies entirely in C

- In a more general probabilistic form: if $X \in \mathbb{R}^n$ is a random variable supported on a convex set $C \subseteq \mathbb{R}^n$, then $\mathbb{E}(X) \in C$

- A *convex combination* of points $x_1, \ldots x_k \in \mathbb{R}^n$ is a combination of the form

  $$\theta_1 x_1 + \theta_2 x_2 + \ldots + \theta_k x_k,$$

  where $\theta_i \geq 0$, $i = 1, \ldots k$ and $\sum_{i=1}^k \theta_i = 1$

- The *convex hull* of a set C is the set of all convex combinations of points in C,

  $$\text{conv}(C) = \Big\{ \theta_1 x_1 + \theta_2 x_2 + \ldots + \theta_k x_k \ : \ \ x_i \in C, \, \theta_i \geq 0 \text{ for } i = 1, \ldots k, \text{ and } \sum_{i=1}^k \theta_i = 1 \Big\}$$

- If you've forgotten, you should remind yourself of the basics of point-set topology: open and closed sets, closure of a set (written $\text{cl}(S)$), and interior and boundary of a set (written $\text{int}(S)$ and $\text{bd}(S)$)

2.2 Some examples

- The empty set $\emptyset$ is convex

- Lines, rays, line segments, linear spaces, and affine spaces are all convex

- A *hyperplane* is convex: this is a set of the form $\{x : a^T x = b\}$

- A *halfspace* is convex: this is a set of the form $\{x : a^T x \leq b\}$

- A *norm ball* is convex: given a norm $\|\cdot\|$ on $\mathbb{R}^n$ (e.g., the ℓ_p norm, $\|\cdot\|_p$, for $p \geq 1$) this has the form $\{x : \|x\| \leq t\}$

- A *polyhedron* is convex: this is the intersection of some finite number of halfspaces, as in

  $$\{x : a_i^T x \leq b_i, \, i = 1, \ldots m\}.$$

  We can abbreviate this as $\{x : Ax \leq b\}$, where $b = (b_1, \ldots b_m) \in \mathbb{R}^m$, $A \in \mathbb{R}^{m \times n}$ with rows a_i, $i = 1, \ldots m$, and the inequality $Ax \leq b$ is interpreted componentwise

- Note: we can also write a polyhedron as $\{x : Ax \leq b, \, Cx = d\}$. (Why?)

- Any bounded polyhedron (called a *polytope*) can also be written as the convex hull of a finite set of points; this is called its V-representation. The original representation given above, as an intersection of halfspaces, is called its H-representation

- *Simplexes* are a special case of polyhedra that are given by taking the convex hull of a set of points $\{x_0, \ldots x_k\} \subseteq \mathbb{R}^n$ that are *affinely independent*, which means that $x_1 - x_0, \ldots x_k - x_0$ are linearly independent. In particular, this is a k-dimensional simplex in $\mathbb{R}^n$

- The canonical simplex is the *probability simplex*, given by the convex hull of $\{e_1, \ldots e_n\} \subseteq \mathbb{R}^n$, the standard basis vectors in $\mathbb{R}^n$, which can be written as

$$\{\theta : \theta \geq 0,\, 1^T\theta = 1\}.$$

  Note again that the inequality $\theta \geq 0$ is to be interpreted componentwise (and we will, without distinction, use 1 to denote the vector of 1s whenever convenient)

- Consider the set of symmetric $n \times n$ matrices,

$$\mathbb{S}^n = \{X \in \mathbb{R}^{n \times n} : X = X^T\}.$$

  Think of this as a vector space of dimension $n(n+1)/2$. Now consider the subset of this vector space

$$S_+^n = \{X \in \mathbb{S}^n : X \succeq 0\},$$

  where $X \succeq 0$ means that X is positive semidefinite. We call S_+^n the *positive semidefinite cone*, and it is a convex set (again, think of it as a set in the ambient $n(n+1)/2$ vector space of symmetric matrices)

2.3 Key properties

- *Separating hyperplane theorem:* if C, D are nonempty, and disjoint ($C \cap D = \emptyset$) convex sets, then there exists $a \neq 0$ and b such that $C \subseteq \{x : a^T x \leq b\}$ and $D \subseteq \{x : a^T x \geq b\}$

- *Supporting hyperplane theorem:* if C is a nonempty convex set, and $x_0 \in \mathrm{bd}(C)$, then there exists a supporting hyperplane to C at x_0, i.e., there exists $a \neq 0$ and b such that $a^T x_0 = b$ and $C \subseteq \{x : a^T x \leq b\}$

- *Closed halfspace representation:* if C is a closed convex set, then it can be represented as the intersection of all halfspaces that contain it,

$$C = \bigcap \big\{H : H \text{ is a halfspace, and } H \supseteq C\big\}$$

2.4 Operations that preserve convexity

- Convexity of all sets in Section 2.2 can be verified directly from the definition. Often though, to check that a set S is convex, it is easier to start with a set of basic sets that we know are convex (such as those in Section 2.2), and recognize that our set S of interest is given by a transformation of one of these basic sets, via an operation that preserves convexity

- The *intersection* of any number of convex sets is convex. This even holds for an (uncountably) infinite number of sets

  E.g., from this we can show that the positive semidefinite cone S_+^n is convex, because

$$\begin{aligned} S_+^n &= \big\{X \in \mathbb{S}^n : a^T X a \geq 0 \text{ for all } a \in \mathbb{R}^n\big\} \\ &= \bigcap_{a \in \mathbb{R}^n} \{X \in \mathbb{S}^n : a^T X a \geq 0\}. \end{aligned}$$

  Note that, for a fixed $a \in \mathbb{R}^n$, the term $a^T X a = \sum_{i,j=1}^n a_i a_j X_{ij}$ is actually a linear function in X, so the above is an intersection of halfspaces in X, and therefore convex

- *Affine images* and *affine preimages* of convex sets are convex. I.e., if $f : \mathbb{R}^n \to \mathbb{R}^m$ is an affine function, meaning that $f(x) = Ax + b$, and $S \subseteq \mathbb{R}^n$, $T \subseteq \mathbb{R}^m$ are convex, then both

  $$f(S) = \{f(x) : x \in S\}$$

  and

  $$f^{-1}(T) = \{x : f(x) \in T\}$$

  are convex

  E.g., from this we can show that the solution set of a *linear matrix inequality*

  $$\big\{x \in \mathbb{R}^k : x_1A_1 + x_2A_2 + \ldots + x_kA_k \preceq B\big\},$$

  where $A_1, \ldots A_k, B \in \mathbb{S}^n$ is convex. To see this, note that this set is the inverse image of S^n_+ under the affine function $f : \mathbb{R}^k \to \mathbb{S}^n$,

  $$f(x) = B - (x_1A_1 + x_2A_2 + \ldots + x_kA_k)$$

- Note in particular that both *scaling* and *translation* preserve convexity (special cases of affine images), i.e., if $S \subseteq \mathbb{R}^n$ is convex then

  $$\alpha S = \{\alpha x : x \in S\}$$

  is convex for any $\alpha \in \mathbb{R}$, and

  $$S + c = \{x + c : x \in S\}$$

  is convex for any $c \in \mathbb{R}^n$

- *Perspective images* and *perpsective preimages* of convex sets are also convex. The perspective function $P \in \mathbb{R}^{n+1}$, with domain $\mathrm{dom}(P) = \mathbb{R}^n \times \mathbb{R}_{++}$ (here $\mathbb{R}_{++}$ denotes the set of positive reals), is defined as

  $$P(x, t) = x/t = (x_1/t, \ldots x_n/t).$$

  Then for any convex $S \subseteq \mathrm{dom}(P) \subseteq \mathbb{R}^{n+1}$, the image

  $$P(S) = \{P(z) : z \in S\}$$

  is convex, and for any convex $T \subseteq \mathbb{R}^n$, the preimage

  $$P^{-1}(T) = \{(x, t) : t > 0,\, x/t \in T\}$$

  is also convex

- *Linear-fractional images* and *linear-fractional preimages* are convex. A *linear-fractional* function is the perspective function composed with an affine function, i.e., if $g : \mathbb{R}^n \times \mathbb{R}^{m+1}$ is affine,

  $$g(x) = \begin{bmatrix} A \\ c^T \end{bmatrix} x + \begin{bmatrix} b \\ d \end{bmatrix},$$

  and $P : \mathbb{R}^{m+1} \to \mathbb{R}^m$ is the perspective map, then $f = P \circ g : \mathbb{R}^n \to \mathbb{R}^m$ is a linear-fractional function. Note

  $$f(x) = \frac{Ax + b}{c^Tx + d},$$

  with domain $\mathrm{dom}(f) = \{x : c^Tx + d > 0\}$. From what we know already, if $S \subseteq \mathrm{dom}(f) \subseteq \mathbb{R}^n$ is convex, then the image $f(S) = P(g(S))$ is convex, and also, if $T \subseteq \mathbb{R}^{m+1}$ is convex, the the preimage $f^{-1}(T) = g^{-1}(P^{-1}(T))$ is convex

E.g., using this we can show the following fact. Let U, V be random variables taking discrete values in $\{1, \dots n\}$ and $\{1, \dots m\}$, respectively, and let $S \subseteq \mathbb{R}^{nm}$ be a set of joint probabilities for U, V. In other words, each $p \in C$ defines a probability distribution over U, V, as in $p_{ij} = \mathbb{P}(U = i, V = j)$. If S is convex, then the set of conditional probabilities of U given V is also convex.

Why? The set of conditional probabilities of U given V is

$$\Big\{ q \in \mathbb{R}^{nm} : q_{ij} = \frac{p_{ij}}{\sum_{k=1}^n p_{kj}}, \text{ for some } p \in C \Big\}.$$

This is the image of C under a linear-fractional function, and is hence convex provided that C is convex

3 Convex functions

3.1 Basic definitions

- In a rough sense, convex functions are even more important than convex sets, because we use them more (though this sounds funny, because the two are intimately related)
- A function $f : \mathbb{R}^n \to \mathbb{R}$ is *convex* if its domain $\mathrm{dom}(f)$ is convex, and for any $x, y \in \mathrm{dom}(f)$ and $\theta \in [0, 1]$,
$$f\big(\theta x + (1 - \theta)y\big) \leq \theta f(x) + (1 - \theta) f(y).$$
In words, the function lies below the line segment joining its evaluations at x, y. A function is *strictly convex* if this same inequality holds strictly for $x \neq y$ and $\theta \in (0, 1)$,
$$f\big(\theta x + (1 - \theta)y\big) < \theta f(x) + (1 - \theta) f(y)$$
- A function f is *concave* or *strictly concave* if $-f$ is convex or strictly convex, respectively
- Affine functions, i.e., such that $f(x) = a^T x + b$, are both convex and concave (conversely, any function that is both convex and concave is affine)
- A function f is *strongly convex* with parameter $m > 0$ (written m-strongly convex) provided that
$$f(x) - \frac{m}{2}\|x\|_2^2$$
is a convex function. In rough terms, this means that f is "as least as convex" as a quadratic function. This is the strongest form of convexity (hence its name), so that strong convexity implies strict convexity implies convexity

3.2 Some examples

- Examples on $\mathbb{R}$: the exponential function e^{ax} is convex for any $a \in \mathbb{R}$; the power function x^a is convex on $\mathbb{R}_{++}$ for any $a \geq 1$ or $a \leq 0$; the negative entropy function $x \log x$ is convex on $\mathbb{R}_{++}$; the log function $\log x$ is concave on $\mathbb{R}_{++}$; the power function x^a is concave on $\mathbb{R}_{++}$ for any $0 \leq a \leq 1$; the affine function $ax + b$ is both convex and concave
- Norms are convex, i.e., $f : \mathbb{R}^n \to \mathbb{R}$ defined by $f(x) = \|x\|$ is a convex function, for any norm $\|\cdot\|$

- An important special case is the ℓ_p norm, $\|\cdot\|_p$, for $p \geq 1$. Recall that this is defined as

$$\|x\|_p = \Big(\sum_{i=1}^n |x_i|^p \Big)^{1/p}$$

when $p < \infty$, and

$$\|x\|_\infty = \max_{i=1,\ldots n} |x_i|.$$

- Two other important special cases are the operator (spectral) and trace (nuclear) norms for matrices. Recall that if $X \in \mathbb{R}^{m \times n}$ has singular values $\sigma_1(X) \geq \sigma_2(X) \geq \ldots \geq \sigma_r(X) \geq 0$, where $r = \text{rank}(X) \leq \min\{m, n\}$, then its operator (or spectral) norm is

$$\|X\|_{\text{op}} = \sigma_1(X),$$

and its trace (or nuclear) norm is

$$\|X\|_{\text{tr}} = \sum_{i=1}^r \sigma_r(X)$$

- The indicator function $f(x) = I_C(x)$ of a convex set $C \subseteq \mathbb{R}^n$ is convex. This is defined as

$$I_C(x) = \begin{cases} 0 & x \in C \\ \infty & x \notin C \end{cases}$$

- The quadratic function $f(x) = \frac{1}{2} x^T Q x + c^T x + b$ is convex provided that $Q \succeq 0$
- The least squares criterion $f(x) = \|Ax - b\|_2^2 = x^T A^T A x - 2b^T A x + b^T b$ is hence convex
- The max function $f(x) = \max\{x_1, \ldots x_n\}$ is convex
- The function $f(x) = \log(\sum_{i=1}^n \exp(x_i))$ is convex, and called the *log-sum-exp* function. This is often viewed as an (infinitely differentiable) approximation to the max function, since

$$\max\{x_1, \ldots x_n\} \leq f(x) \leq \max\{x_1, \ldots x_n\} + \log n$$

3.3 Key properties

- A convex function is continuous on the relative interior of its domain; it can only have points of discontinuity on its relative boundary
- A function is convex if and only its restriction to any line is convex. That is, $f : \mathbb{R}^n \to \mathbb{R}$ is convex if and only if $g(t) = f(x + tv)$ is convex in $t \in \mathbb{R}$ (on its domain $\{t : x + tv \in \text{dom}(f)\}$), for all $v \in \mathbb{R}^n$

  E.g., from this we can show that the function $f : \mathbb{S}^n \to \mathbb{R}$, $f(X) = \log\det X$, with $\text{dom}(f) = S_{++}^n = \{X \in \mathbb{S}^n : X \succ 0\}$, is concave
- *First-order characterization:* suppose that f is differentiable (and write ∇f for its gradient). Then f is convex if and only if $\text{dom}(f)$ is convex, and for all $x, y \in \text{dom}(f)$,

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

  In words, the function always dominates its first order (linear) Taylor approximation. It's an analogous story for strict convexity: the condition is that for all $x \neq y$,

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

E.g., from the first-order characterization, we can deduce a useful property: if $\nabla f(x) = 0$ for a convex function f, then $f(y) \geq f(x)$ for all y, so x is a minimizer of f. Further, if f is strictly convex, then x is the unique minimizer

- *Second-order characterization:* suppose that f is twice differentiable (and we write $\nabla^2 f$ for its Hessian). Then f is convex if and only if $\mathrm{dom}(f)$ is convex, and

$$\nabla^2 f(x) \succeq 0$$

for all $x \in \mathrm{dom}(f)$. Note that $\nabla^2 f(x) \succ 0$ for all $x \in \mathrm{dom}(f)$ implies strict convexity—but the converse is not true!

E.g., using this second-order characterization, we can verify the convexity of the quadratic function $f(x) = x^T Q x + c^T x + b$ when $Q \succeq 0$ (and strict convexity when $Q \succ 0$). We can also verify that $f : \mathbb{R}^n \to \mathbb{R}$ defined by

$$f(x) = \log\Big(\sum_{i=1}^n \exp(x_i)\Big)$$

is convex

- *Strong convexity characterizations:* if f is differentiable, then m-strong convexity is equivalent to $\mathrm{dom}(f)$ being convex and

$$f(y) \geq f(x) + \nabla f(x)^T (y - x) + \frac{m}{2}\|y - x\|_2^2$$

for all $x, y \in \mathrm{dom}(f)$. I.e., f is lower bounded by its second order (quadratic) approximation, rather then only its first order (linear) approximation, which is implied by regular convexity

If f is twice differentiable, then m-strong convexity is equivalent to $\mathrm{dom}(f)$ being convex and

$$\nabla^2 f(x) \succeq mI$$

for all $x \in \mathrm{dom}(f)$, i.e., the smallest eigenvalue of the Hessian is lower bounded by m, everywhere

- A convex function f has convex *level sets*,

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

for any $t \in \mathbb{R}$. The converse is not true

- *Epigraph characterization:* a function f is convex if and only if its *epigraph*

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

is a convex set. This ties together convexity for functions and sets; in fact many properties of convex functions can be proven from those for convex sets

- *Jensen's inequality:* if f is convex, and X is a random variable supported on $\mathrm{dom}(f)$, then

$$f\big(\mathbb{E}(X)\big) \leq \mathbb{E}\big(f(X)\big).$$

This is a more general probabilistic form of the basic inequality for convexity

3.4 Operations that preserve convexity

- As was true for convex sets, the easiest way to prove convexity of a function is often to show that it can be built up from simple convex functions, using operations that preserve convexity
- *Nonnegative linear combinations:* if $f_1, \ldots f_m$ are convex, then

  $$a_1 f_1 + \ldots + a_m f_m$$

  is convex for any $a_1, \ldots a_m \geq 0$
- *Affine composition:* if f is convex, then $g(x) = f(Ax + b)$ is convex
- *Pointwise maximum:* if $f_1, \ldots f_m$ are convex, then

  $$f(x) = \max\{f_1(x), \ldots f_m(x)\}$$

  are convex. This extends to (uncountably) infinitely many functions: if $f_s(x)$ is convex for any $s \in S$, then

  $$f(x) = \max_{s \in S} f_s(x)$$

  is convex

  We can use this to show some fairly nonobvious functions are convex. E.g., the maximum distance to an arbitrary set $C \subseteq \mathbb{R}^n$,

  $$f(x) = \max_{y \in C} \|x - y\|,$$

  in any norm $\|\cdot\|$, is convex. This is because $\|x - y\|$ is convex in x for any fixed y. Also, the optimal weighted least squares cost,

  $$f(w) = \min_{\beta \in \mathbb{R}^p} \sum_{i=1}^n w_i (y_i - x_i^T \beta)^2,$$

  is concave as a function of the weights $w \in \mathbb{R}^n$, with domain

  $$\mathrm{dom}(f) = \Big\{ w : \min_{\beta \in \mathbb{R}^p} \sum_{i=1}^n w_i (y_i - x_i^T \beta)^2 > -\infty \Big\}.$$

  This is because each $\sum_{i=1}^n w_i (y_i - x_i^T \beta)^2$, for fixed β, is affine and hence concave in w, so the pointwise minimum f is also concave
- *Partial minimization:* if $f(x, y)$ is convex in x, y, and C is a convex and nonempty set, then

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

  is convex in x, provided that $g(x) > -\infty$ for all x in its domain,

  $$\mathrm{dom}(g) = \{x : (x, y) \in \mathrm{dom}(f) \text{ for some } y \in C\}$$

  Again, we can use this to show some fairly nonobvious properties. E.g., the minimum distance to a convex set C,

  $$f(x) = \min_{y \in C} \|x - y\|,$$

in any norm $\|\cdot\|$, is convex. This is because $\|x-y\|$ is convex in x, y, and we have assumed that C is convex. (N.B. the maximum distance to *any* set is a convex function.) Also, suppose that

$$\begin{bmatrix} A & B \\ B^T & C \end{bmatrix} \succeq 0$$

and so

$$f(x,y) = x^T A x + 2x^T B y + y^T C y$$

is convex is x, y. Then we know that

$$g(x) = \min_{y \in \mathbb{R}^n} f(x,y)$$

is convex in x; a simple calculation shows that, assuming C is invertible,

$$g(x) = x^T (A - BC^{-1}B^T)x.$$

Therefore, for g to be convex, we know that the *Schur complement* has to be positive semidefinite,

$$A - BC^{-1}B^T \succeq 0$$

- *Composition:* this is a bit tricky, as composition rules that preserve convexity (or concavity) rely on monotonicity conditions. Here are a few results to remember, in the setting $f = h \circ g$, where $g : \mathbb{R}^n \to \mathbb{R}$, $h : \mathbb{R} \to \mathbb{R}$, so $f : \mathbb{R}^n \to \mathbb{R}$. (We assume for simplicity that $\mathrm{dom}(g) = \mathbb{R}^n$ and $\mathrm{dom}(h) = \mathbb{R}$.)
  - f is convex provided that h is convex and nondecreasing, and g is convex
  - f is convex provided that h is convex and nonincreasing, and g is concave
  - f is concave provided that h is concave and nondecreasing, and g is concave
  - f is concave provided that h is concave and nonincreasing, and g is convex

  While these rules hold without assuming differentiability of h, g, in order to remember them, it may help to think of the chain rule on $\mathbb{R}$:

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

  Now we can see directly that if, e.g., h is convex and nondecreasing, then $h'' \geq 0$ and $h' \geq 0$, and if g is convex, then $g'' \geq 0$, so altogether $f'' \geq 0$

4 Optimization problems

4.1 Basic definitions

- An *optimization problem* has the form

$$\begin{array}{ll} \min\limits_{x \in D} & f(x) \\ \text{subject to} & h_i(x) \leq 0, \;\; i = 1, \ldots m \\ & \ell_j(x) = 0, \;\; j = 1, \ldots r. \end{array} \tag{1}$$

  Here $D = \mathrm{dom}(f) \cap \bigcap_{i=1}^m \mathrm{dom}(h_i) \cap \bigcap_{j=1}^r \mathrm{dom}(\ell_j)$, the common domain of all functions. The function f is called the *objective* or *criterion*. A *feasible point* x is a point in D such that all inequality and equality constraints are met. A *solution* or *minimizer* $x^\star$ is a feasible point that achieves the minimal criterion value. We will often denote the minimum criterion value by $f^\star$. (We will also often stop explicitly writing $x \in D$, and consider this requirement implicit)

- A *convex optimization problem* is an optimization problem in which all functions $f, h_1, \ldots h_m$ are convex, and all functions $\ell_1, \ldots \ell_r$ are affine. (Think: why affine?) Hence, we can express it as

$$\begin{aligned} \min_x \quad & f(x) \\ \text{subject to} \quad & h_i(x) \leq 0, \;\; i = 1, \ldots m \\ & Ax = b \end{aligned} \tag{2}$$

- The problem (2) is of course equivalent to a concave maximization problem, as in

$$\begin{aligned} \max_x \quad & -f(x) \\ \text{subject to} \quad & -h_i(x) \geq 0, \;\; i = 1, \ldots m \\ & Ax = b. \end{aligned}$$

  Often we will not make any distinction, and still call the above a convex optimization problem

4.2 Some examples

- When f is affine, and all $h_1, \ldots h_m$ are affine, problem (2) becomes

$$\begin{aligned} \min_{x \in \mathbb{R}^n} \quad & c^T x \\ \text{subject to} \quad & Gx \leq h \\ & Ax = b, \end{aligned} \tag{3}$$

  and is known as a *linear program* (LP). This problem is always convex, and is a well-studied topic (there are entire courses, and entire books about linear programming alone). The feasible set in (3) is the polyhedron $\{x : Gx \leq h, \, Ax = b\}$; it is not hard to see that a solution in (3) always lies at a vertex (exposed point) of this polyhedron

- When f is quadratic, and still all $h_1, \ldots h_m$ are affine, problem (2) becomes

$$\begin{aligned} \min_{x \in \mathbb{R}^n} \quad & \frac{1}{2} x^T Q x + c^T x \\ \text{subject to} \quad & Gx \leq h \\ & Ax = b, \end{aligned} \tag{4}$$

  and is called a *quadratic program* (QP). This problem is convex provided that $Q \succeq 0$

- Given $y_i \in \{0, 1\}$, $x_i \in \mathbb{R}^p$, $i = 1, \ldots n$, the problem

$$\begin{aligned} \max_{\beta \in \mathbb{R}^p} \quad & \prod_{i=1}^n \Big(\frac{\exp(x_i^T \beta)}{1 + \exp(x_i^T \beta)} \Big)^{y_i} \cdot \Big(\frac{1}{1 + \exp(x_i^T \beta)} \Big)^{1 - y_i} \\ \text{subject to} \quad & \|\beta\|_1 \leq t, \end{aligned}$$

  is an ℓ_1 regularized logistic regression problem. Because log is monotone increasing, we can take the log of the criterion value, and flip its sign, to yield the equivalent problem

$$\begin{aligned} \min_{\beta \in \mathbb{R}^p} \quad & \sum_{i=1}^n \Big(-y_i (x_i^T \beta) + \log \big(1 + \exp(x_i^T \beta) \big) \Big) \\ \text{subject to} \quad & \|\beta\|_1 \leq t. \end{aligned}$$

  This is a convex problem because the criterion is a sum of affine functions and log-sum-exp functions (composed with affine functions), and the ℓ_1 norm is convex

4.3 Key properties

- Perhaps the most important property of convex optimization problems is that *any local minimizer is a global minimizer*. To see this, suppose that x is feasible for (2), and there exists some $R > 0$ such that

$$f(x) \leq f(y) \quad \text{for all feasible } y \text{ with } \|x - y\|_2 \leq R.$$

  Such a point x is called a local minimizer. For the sake of contradiction, suppose that x was not a global minimizer, i.e., there exists some feasible z such that $f(z) < f(x)$. By convexity of the constraints (and the domain D), the point $\theta z + (1-\theta)x$ is feasible for any $0 \leq \theta \leq 1$. Furthermore, by convexity of f,

$$f\big(\theta z + (1-\theta)x\big) \leq \theta f(z) + (1-\theta) f(x) < f(x)$$

  for any $0 < \theta < 1$. Finally, we can choose $\theta > 0$ small enough so that $\|x - (\theta z + (1-\theta)x)\|_2 = \theta \|x - z\|_2 \leq R$, and we obtain a contradiction

- Beware of a common misconception: this does not mean that a convex optimization problem must have a unique minimizer! Simply consider an unconstrained convex problem with $f(x) = c$, a constant

- However, we do know that the set of solutions of a convex problem forms a convex set. This is true because if x and z are solutions, then $\theta x + (1-\theta)z$ is feasible for any $0 \leq \theta \leq 1$, and by convexity

$$f\big(\theta x + (1-\theta)z\big) \leq \theta f(x) + (1-\theta) f(z) = f^\star$$

  for any $0 \leq \theta \leq 1$, i.e., $f(\theta x + (1-\theta)z) = f^\star$ as $f^\star$ is optimal, which means that $\theta x + (1-\theta)z$ is also a solution

- Furthermore, a convex problem with a strictly convex criterion function f does have a *unique solution*. This follows because if x and z were both solutions with $x \neq z$, then $\theta x + (1-\theta)z$ is feasible for any $0 \leq \theta \leq 1$, and by strict convexity

$$f\big(\theta x + (1-\theta)z\big) < \theta f(x) + (1-\theta) f(z) = f^\star$$

  for any $0 < \theta < 1$, which cannot be the case, because $f^\star$ is the optimal criterion value

- It can happen that a nonconvex optimization problem, i.e., a problem of the form (1) where at least one of $f, h_1, \ldots f_m$ is not convex, or at least one of $\ell_1, \ldots \ell_r$ is not affine, actually reduces to a convex optimization problem. So think carefully about whether you can manipulate the form of the particular problem in your favor

5 Subgradients

5.1 Basic definitions

- Remember that for a convex function $f : \mathbb{R}^n \to \mathbb{R}$, we have

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

  for all $x, y \in \text{dom}(f)$. I.e., the linear approximation always underestimates f. A *subgradient* of f at $x \in \text{dom}(f)$ is any $g \in \mathbb{R}^n$ such that

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

  for all $y \in \text{dom}(f)$

- Subgradients always exist for convex functions (to be precise, this is only true on the relative interior of $\mathrm{dom}(f)$). One can prove this by representing a convex function via its epigraph, and using the supporting hyperplane theorem
- If f is convex and differentiable at x, then $g = \nabla f(x)$ is the unique subgradient at x
- The same definition for subgradients also applies to nonconvex functions f; but in this case, subgradients need not exist (even when f is differentiable)
- The set of all subgradients of a f at x is called its *subdifferential* at x, denoted

$$\partial f(x) = \big\{g : g \text{ is a subgradient of } f \text{ at } x\big\}.$$

  This set $\partial f(x)$ is closed and convex (even when f is nonconvex). For a convex function f (and x in the relative interior of $\mathrm{dom}(f)$), the set $\partial f(x)$ is nonempty. Note that if f is convex and differentiable at x, then $\partial f(x) = \{\nabla f(x)\}$; conversely, if f is convex and $\partial f(x) = \{g\}$, then f is differentiable at x and $\nabla f(x) = g$

5.2 Some examples

- Consider the absolute value function, $f : \mathbb{R} \to \mathbb{R}$, $f(x) = |x|$. When $x \neq 0$, f has a unique unique subgradient $g = \mathrm{sign}(x)$. When $x = 0$, subgradient g can be any element of $[-1, 1]$
- Consider the ℓ_2 norm, $f : \mathbb{R}^n \to \mathbb{R}$, $f(x) = \|x\|_2$. When $x \neq 0$, f has a unique subgradient $g = x/\|x\|_2$. When $x = 0$, subgradient g can be any element of the ℓ_2 ball, $\{z : \|z\|_2 \leq 1\}$
- Consider the ℓ_1 norm, $f : \mathbb{R}^n \to \mathbb{R}$, $f(x) = \|x\|_1$. When $x_i \neq 0$, a subgradient g of f has the unique ith component $g_i = \mathrm{sign}(x_i)$. When $x_i = 0$, g_i can be any element of $[-1, 1]$
- Let $f_1, f_2 : \mathbb{R}^n \to \mathbb{R}$ be convex, differentiable, and consider $f(x) = \max\{f_1(x), f_2(x)\}$.
  - When $f_1(x) > f_2(x)$, f has a unique subgradient $g = \nabla f_1(x)$
  - When $f_2(x) > f_1(x)$, f has a unique subgradient $g = \nabla f_2(x)$
  - When $f_1(x) = f_2(x)$, g can be any point on the line segment joining $\nabla f_1(x)$ and $\nabla f_2(x)$
- Consider $f(x) = I_C(x)$, the indicator function of a convex set $C \in \mathbb{R}^n$. Then subgradients of f at a point $x \in C$ are exactly the *normal cone* of C at x, written $\partial I_C(x) = \mathcal{N}_C(x)$, where

$$\mathcal{N}_C(x) = \big\{g : g^T x \geq g^T y \text{ for any } y \in C\big\}$$

5.3 Key properties

- *Subgradient calculus:* here are several basic rules for subgradients of convex functions.
  - *Scaling:* $\partial(af) = a \cdot \partial f$ provided $a > 0$
  - *Addition:* $\partial(f_1 + f_2) = \partial f_1 + \partial f_2$
  - *Affine composition:* if $g(x) = f(Ax + b)$, then

$$\partial g(x) = A^T \partial f(Ax + b)$$

  - *Finite pointwise maximum:* if $f(x) = \max_{i=1,\ldots m} f_i(x)$, then

$$\partial f(x) = \mathrm{conv}\Big(\bigcup_{i : f_i(x) = f(x)} \partial f_i(x)\Big),$$

    the convex hull of union of subdifferentials of all active functions at x

- *General pointwise maximum:* an extension of the finite pointwise maximum rule. If $f(x) = \max_{s \in S} f_s(x)$, then

$$\partial f(x) \supseteq \mathrm{cl}\Big\{\mathrm{conv}\Big(\bigcup_{s: f_s(x) = f(x)} \partial f_s(x)\Big)\Big\},$$

  and under some regularity conditions on S, f_s, we get an equality above (a sufficient condition, e.g., if that S is compact, and the function $s \mapsto f_s(x)$ are continuous in s for each fixed x)

- *Subgradients of norms:* an important special case of the above rule. Let $f(x) = \|x\|$ for some arbitrary norm $\|\cdot\|$, and let $\|\cdot\|_*$ denote its dual norm—we will return to this later, but for now, you can think of $\|\cdot\|_p$ and $\|\cdot\|_q$ being dual, where $1/p + 1/q = 1$. Then

$$\|x\| = \max_{\|z\|_* \leq 1} z^T x,$$

  and hence

$$\partial \|x\| = \Big\{y : \|y\|_* \leq 1 \text{ and } y^T x = \max_{\|z\|_* \leq 1} z^T x\Big\}$$

- *Optimality characterization:* certainly one of the most important facts to know about subgradients. For any f (convex or not),

$$x \text{ minimizes } f \iff 0 \in \partial f(x).$$

  Why? This is very easy to show: $g = 0$ being a subgradient means that for all $y \in \mathrm{dom}(f)$,

$$f(y) \geq f(x) + 0^T(y - x) = f(x).$$

  Note the connection to the convex and differentiable case, in which $\partial f(x) = \{\nabla f(x)\}$

  This optimality characterization can be very helpful. Here are two examples of putting it to use. First, consider a closed, convex set $C \subseteq \mathbb{R}^n$. For $y \in \mathbb{R}^n$, we define its *projection* onto C by

$$P_C(y) = \operatorname*{argmin}_{x \in C} \|y - x\|_2.$$

  Using our optimality characterization, we can show that $x = P_C(y)$ if and only if

$$\langle y - x, x - u \rangle \geq 0 \quad \text{for all } u \in C,$$

  which is sometimes called the *variational inequality*. How to see this? Note that $x = P_C(y)$ minimizes the criterion

$$f(x) = \frac{1}{2}\|y - x\|_2^2 + I_C(x)$$

  where I_C is the indicator function of C. Hence we know this is equivalent to

$$0 \in \partial f(x) = -(y - x) + \mathcal{N}_C(x),$$

  i.e.,

$$y - x \in \mathcal{N}_C(x),$$

  which exactly means that

$$(y - x)^T x \geq (y - x)^T u \quad \text{for all } u \in C.$$

  Rearranging gives the result.

As a second example, consider the ℓ_1 penalized least squares problem

$$\min_{\beta \in \mathbb{R}^n} \ \frac{1}{2}\|y - \beta\|_2^2 + \lambda \|\beta\|_1.$$

(This is a lasso problem with identity predictor matrix.) We claim that the solution of this problem is $\hat{\beta} = S_\lambda(y)$, where S_λ is the *soft-thresholding operator*, defined as

$$[S_\lambda(y)]_i = \begin{cases} y_i - \lambda & \text{if } y_i > \lambda \\ 0 & \text{if } -\lambda \leq y_i \leq \lambda \\ y_i + \lambda & \text{if } y_i < -\lambda \end{cases}, \qquad i = 1, \ldots n.$$

Why? Subgradients of $f(\beta) = \frac{1}{2}\|y - \beta\|_2^2 + \lambda\|\beta\|_1$ are

$$g = \beta - y + \lambda s,$$

where $s_i = \text{sign}(\beta_i)$ if $\beta_i \neq 0$ and $s_i \in [-1, 1]$ if $\beta_i = 0$. Now just plug in $\beta = S_\lambda(y)$ and check that we can get $g = 0$

6 Duality

6.1 Basic definitions and properties

- Duality is one of the most useful, and most beautiful, concepts in optimization. It provides us with an equivalent optimization problem to inspect, that often has complementary properties to the original (called the primal) problem
- Recall that a general optimization problem has the form

$$\begin{aligned} \min_x \quad & f(x) \\ \text{subject to} \quad & h_i(x) \leq 0, \ i = 1, \ldots m \\ & \ell_j(x) = 0, \ j = 1, \ldots r. \end{aligned} \tag{5}$$

If $f, h_1, \ldots h_m$ are convex and $\ell_1, \ldots \ell_r$ are affine, then this problem is a convex optimization problem. For now we will *not assume convexity* and just stick with the general problem (5). We first define the *Lagrangian* associated with (5) by

$$L(x, u, v) = f(x) + \sum_{i=1}^m u_i h_i(x) + \sum_{j=1}^r v_j \ell_j(x).$$

Note that L is a function of three (blocks of) variables: x, and u, v which are new variables, called *dual variables*, that we have just introduced. In particular, we have $u \in \mathbb{R}^m, v \in \mathbb{R}^r$, with the implicit domain $u \geq 0$ (i.e., implicitly, we define $L(x, u, v) = -\infty$ for $u < 0$). A trivial but important property is that for any primal feasible x (i.e., x satisfying the constraints in (5)) and dual feasible u, v (i.e., such that $u \geq 0$), we have

$$\begin{aligned} L(x, u, v) &= f(x) + \sum_{i=1}^m u_i h_i(x) + \sum_{j=1}^r v_j \ell_j(x) \\ &\leq f(x) + \sum_{i=1}^m u_i \cdot 0 + \sum_{j=1}^r v_j \cdot 0 \\ &= f(x). \end{aligned}$$

In other words, $L(\cdot, u, v)$ provdes a lower bound on f for any dual feasible u, v

- Now let C denote primal feasible set,

$$C = \{x : h_i(x) \leq 0,\ i = 1, \ldots m,\ \ell_j(x) = 0,\ j = 1, \ldots r\},$$

and $f^\star$ denote the optimal primal criterion value. Then minimizing $L(x, u, v)$ over all x gives a lower bound on $f^\star$

$$f^\star \geq \min_{x \in C} L(x, u, v) \geq \min_x L(x, u, v) := g(u, v).$$

We call $g(u, v)$ the *Lagrange dual function*, and it provides a lower bound on $f^\star$ for any dual feasible u, v

- Finally, the *dual problem* associated with (5) is given by maximizing this lower bound over all feasible points,

$$\begin{aligned} \max_{u,v} \quad & g(u, v) \\ \text{subject to} \quad & u \geq 0. \end{aligned} \tag{6}$$

A key property, called *weak duality*: if we write $g^\star$ as the dual optimal value, then

$$f^\star \geq g^\star.$$

Note that this always holds (even if the primal problem is nonconvex)

- Another key property: the Lagrange dual function g is *always concave*, regardless of the primal problem (5). This makes the dual problem (6) always a concave maximization problem, i.e., a convex optimization problem. Why? According to its definition,

$$g(u, v) = \underbrace{-\max_x \Big\{ -f(x) - \sum_{i=1}^m u_i h_i(x) - \sum_{j=1}^r v_j \ell_j(x) \Big\}}_{\text{pointwise maximum of convex functions in } (u, v)}$$

6.2 Some examples

- Consider the quadratic program

$$\begin{aligned} \min_{x \in \mathbb{R}^n} \quad & \frac{1}{2} x^T Q x + c^T x \\ \text{subject to} \quad & Ax = b,\ x \geq 0 \end{aligned}$$

where $Q \succ 0$. The Lagrangian is

$$L(x, u, v) = \frac{1}{2} x^T Q x + c^T x - u^T x + v^T (Ax - b)$$

The Lagrange dual function is

$$g(u, v) = \min_{x \in \mathbb{R}^n} L(x, u, v) = -\frac{1}{2}(c - u + A^T v)^T Q^{-1} (c - u + A^T v) - b^T v$$

The dual problem is

$$\begin{aligned} \max_{u \in \mathbb{R}^n,\, v \in n\mathbb{R}^m} \quad & -\frac{1}{2}(c - u + A^T v)^T Q^{-1} (c - u + A^T v) - b^T v \\ \text{subject to} \quad & u \geq 0, \end{aligned}$$

which is another quadratic program, whose optimal value is $g^\star \leq f^\star$, where $f^\star$ is the optimal value of the primal problem

Figure 1 shows an an example of a QP in 2 dimensions, with no equality constraints (so the dual QP is also 2 dimensional). Note: it looks like $g^\star = f^\star$. Is this a coincidence?

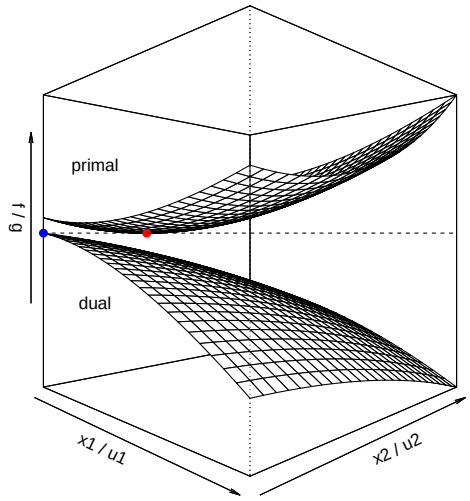

Figure 1: *The primal and dual criterion surfaces for the quadratic minimization example.*

- As another example, consider the quartic minimization problem

$$\begin{aligned} \min_{x \in \mathbb{R}} \quad & x^4 - 50x^2 + 100x \\ \text{subject to} \quad & x \geq -4.5. \end{aligned}$$

This is nonconvex (because its criterion is nonconvex). Although it is a pretty messy calculation, here the dual function g can be derived explicitly (via the analytic formula for roots of a cubic equation):

$$g(u) = \min_{i=1,2,3} F_i^4(u) - 50F_i^2(u) + 100F_i(u),$$

where for $i = 1, 2, 3$,

$$F_i(u) = \frac{-a_i}{12 \cdot 2^{1/3}} \Big(432(100-u) - \big(432^2(100-u)^2 - 4 \cdot 1200^3\big)^{1/2} \Big)^{1/3} - 100 \cdot 2^{1/3} \frac{1}{\Big(432(100-u) - \big(432^2(100-u)^2 - 4 \cdot 1200^3\big)^{1/2} \Big)^{1/3}},$$

and $a_1 = 1$, $a_2 = (-1 + i\sqrt{3})/2$, $a_3 = (-1 - i\sqrt{3})/2$. Without the context of duality it would be difficult to tell whether or not g is concave ... but we know it must be!

Figure 2 displays the primal and dual criterion functions for this quartic example. Note: it is evident that $g^\star < f^\star$, i.e., the lower bound constructed from the dual problem is strictly less than the primal optimal value. Why?

6.3 Strong duality and Slater's condition

- We have seen an example above in which the dual problem delivers a tight lower bound, in that $g^\star = f^\star$; this phenomenon is called *strong duality*
- When does strong duality this hold in general? A sufficient (but not necessary) condition is called *Slater's condition*: if the primal problem (5) is a convex (i.e., $f, h_1, \ldots h_m$ are convex,

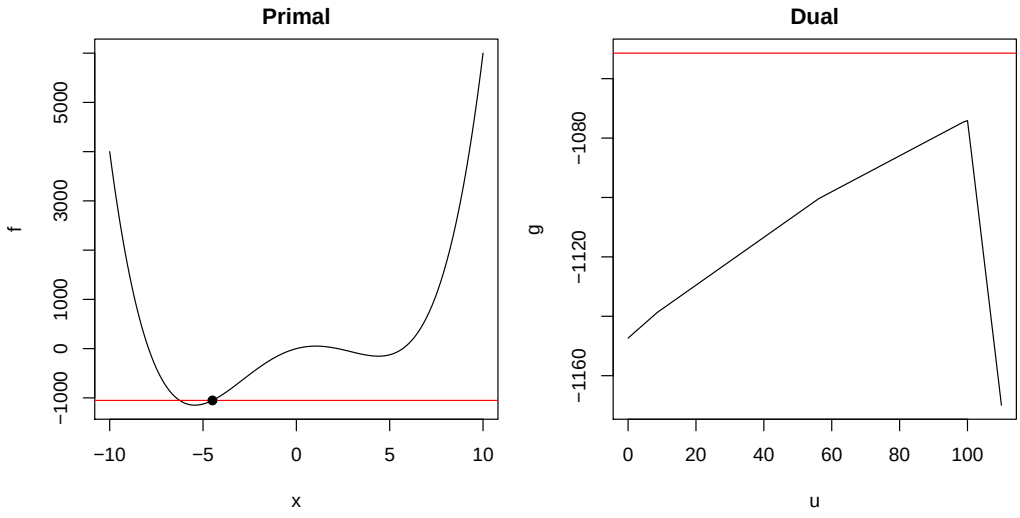

Figure 2: *The primal and dual criterion functions for the quartic minimization example.*

$\ell_1, \ldots \ell_r$ are affine), and there exists at least one strictly feasible x, meaning that

$$h_1(x) < 0, \ldots h_m(x) < 0 \quad \text{and} \quad \ell_1(x) = 0, \ldots \ell_r(x) = 0,$$

then strong duality holds

- This is a pretty weak condition. And it can be further refined: we actually only need strict inequalities only over functions h_i that are not affine

6.4 Duality gap

- In general, given primal feasible x and dual feasible u, v, the quantity

$$f(x) - g(u, v)$$

is called the *duality gap* between x and u, v. Note that

$$f(x) - f^\star \leq f(x) - g(u, v),$$

so if the duality gap is zero, then x is primal optimal, and similarly, u, v are dual optimal

- This plays a key role in establishing optimality conditions for problems under strong duality, as we will cover next

7 The KKT conditions

7.1 Statement of conditions

- Given the general problem

$$\begin{aligned} \min_x \quad & f(x) \\ \text{subject to} \quad & h_i(x) \leq 0, \quad i = 1, \ldots m \\ & \ell_j(x) = 0, \quad j = 1, \ldots r, \end{aligned}$$

the *Karush-Kuhn-Tucker conditions* (or *KKT conditions*) are

- $0 \in \partial f(x) + \sum_{i=1}^m u_i \partial h_i(x) + \sum_{j=1}^r v_j \partial \ell_j(x)$ (stationarity)
- $u_i \cdot h_i(x) = 0$ for all $i = 1, \ldots m$ (complementary slackness)
- $h_i(x) \leq 0$, $\ell_j(x) = 0$ for all $i = 1, \ldots m$, $j = 1, \ldots r$ (primal feasibility)
- $u_i \geq 0$ for all $i = 1, \ldots m$ (dual feasibility)

Note that the KKT conditions are a statement about a triplet of variables x, u, v, where x is a primal variable, and u, v are dual variables, i.e., u, v are associated with the dual problem

$$\begin{aligned} \max_{u,v} \quad & g(u, v) \\ \text{subject to} \quad & u \geq 0. \end{aligned}$$

But importantly, we don't to need to form the dual function g to examine the KKT conditions

7.2 Necessity

- Let $x^\star$ and $u^\star, v^\star$ be primal and dual solutions with zero duality gap (this means that strong duality holds, e.g., under Slater's condition). Then the KKT conditions must hold
- Proof: we have

$$\begin{aligned} f(x^\star) &= g(u^\star, v^\star) \\ &= \min_x \; f(x) + \sum_{i=1}^m u_i^\star h_i(x) + \sum_{j=1}^r v_j^\star \ell_j(x) \\ &\leq f(x^\star) + \sum_{i=1}^m u_i^\star h_i(x^\star) + \sum_{j=1}^r v_j^\star \ell_j(x^\star) \\ &\leq f(x^\star). \end{aligned}$$

  In other words, all these inequalities are actually equalities. Two things to learn from this:

  1. The point $x^\star$ minimizes $L(x, u^\star, v^\star)$ over $x \in \mathbb{R}^n$. Hence the subdifferential of $L(x, u^\star, v^\star)$ must contain 0 at $x = x^\star$; this is exactly the stationarity condition
  2. We must have $\sum_{i=1}^m u_i^\star h_i(x^\star) = 0$, and because each term here is ≤ 0, this implies that $u_i^\star h_i(x^\star) = 0$ for $i = 1, \ldots m$; this is exactly complementary slackness

  Primal and dual feasibility obviously hold. Hence, we've verified the KKT conditions
- For this direction of the problem, we have assumed nothing a priori about the convexity of our optimization problem; we have rather assumed strong duality (a zero duality gap) directly

7.3 Sufficiency

- Suppose that $x^\star, u^\star, v^\star$ satisfy the KKT conditions. Then $x^\star$ is a primal solution and $u^\star, v^\star$ is a dual solution
- Proof: we have

$$\begin{aligned} g(u^\star, v^\star) &= f(x^\star) + \sum_{i=1}^m u_i^\star h_i(x^\star) + \sum_{j=1}^r v_j^\star \ell_j(x^\star) \\ &= f(x^\star) \end{aligned}$$

  where the first equality holds from the stationarity condition, and the second equality holds from complementary slackness and primal feasibility. Therefore the duality gap is zero (and $x^\star$ and $u^\star, v^\star$ are primal and dual feasible) so $x^\star$ and $u^\star, v^\star$ are primal and dual optimal

7.4 Putting it together

- The KKT conditions are always sufficient for optimality, and necessary under strong duality
- Hence, for a problem with strong duality—e.g., under Slater's condition: the problem is convex and there exists a point x strictly satisfying its nonaffine inequality contraints—we have

$$\begin{aligned} & x^\star \text{ and } u^\star, v^\star \text{ are primal and dual solutions} \\ \iff & x^\star \text{ and } u^\star, v^\star \text{ satisfy the KKT conditions} \end{aligned}$$

- A warning, concerning the stationarity condition: for a differentiable function f, we do not know that $\partial f(x) = \{\nabla f(x)\}$ unless f is convex. This is a common mistake!

7.5 Some examples

- Consider for $Q \succeq 0$, the quadratic program

$$\begin{aligned} \min_{x \in \mathbb{R}^n} \quad & \frac{1}{2} x^T Q x + c^T x \\ \text{subject to} \quad & Ax = 0 \end{aligned}$$

  This is a convex problem, with no inequality constraints, so by the KKT conditions: x is a solution if and only if

$$\begin{bmatrix} Q & A^T \\ A & 0 \end{bmatrix} \begin{bmatrix} x \\ u \end{bmatrix} = \begin{bmatrix} -c \\ 0 \end{bmatrix}$$

  for some u. Linear system combines stationarity, primal feasibility (complementary slackness and dual feasibility are vacuous)

- When the primal problem convex, and unconstrained, the KKT conditions are necessary and sufficient for optimality. But in this case, the KKT conditions just reduce to the stationarity condition:

$$0 \in \partial f(x),$$

  which we already know is necessary and sufficient for optimality, from the subgradient characterization

- Consider the ℓ_1 penalized problem

$$\min_{\beta \in \mathbb{R}^p} \; f(X\beta) + \lambda \|\beta\|_1,$$

  where $f : \mathbb{R}^n \to \mathbb{R}$ is a convex, differentiable function, and $X \in \mathbb{R}^{n \times p}$ is a matrix of predictors (columns) $X_1, \ldots X_p \in \mathbb{R}^n$. This problem is convex and unconstrained, so the stationarity condition is necessary and sufficient for optimality, which is

$$-X^T \nabla f(X\beta) = \lambda s,$$

  where $s \in \partial \|\beta\|_1$, i.e.,

$$s_i \in \begin{cases} \{1\} & \text{if } \beta_i > 0 \\ \{-1\} & \text{if } \beta_i < 0 \\ [-1, 1] & \text{if } \beta_i = 0. \end{cases}$$

  Now we can read off an important fact: if $|X_i^T \nabla f(X\beta)| < \lambda$, then $\beta_i = 0$

- Consider the graphical lasso problem

$$\min_{\Theta \in \mathbb{S}^p_{++}} \; -\log\det\Theta + \mathrm{tr}(S\Theta) + \lambda\|\Theta\|_1,$$

where $S \in \mathbb{S}^p_+$ is a positive semidefinite sample covariance matrix, and $\|\Theta\|_1 = \sum_{i,j=1}^p |\Theta_{ij}|$. The KKT conditions again reduce to the stationary condition,

$$-\Theta^{-1} + S + \Gamma = 0,$$

where $\Gamma \in \mathbb{R}^{p\times p}$ has elements $\Gamma_{ij} \in \partial|\Theta_{ij}|$, i.e.,

$$\Gamma_{ij} \in \begin{cases} \{1\} & \text{if } \Theta_{ij} > 0 \\ \{-1\} & \text{if } \Theta_{ij} < 0 \\ [-1,1] & \text{if } \Theta_{ij} = 0. \end{cases}$$

This stationarity condition actually tells us whole lot about the structure of Θ at optimality. Let $\tilde{S}$ denote the componentwise soft-thresholded version of S, i.e., with components

$$\tilde{S}_{ij} = \begin{cases} S_{ij} - \lambda & \text{if } S_{ij} > \lambda \\ 0 & \text{if } -\lambda \le S_{ij} \le \lambda \\ S_{ij} + \lambda & \text{if } S_{ij} < -\lambda \end{cases}.$$

Observe:

  - If Θ is block diagonal, then so is Θ^{-1}, with the same block structure. Hence, for all i, j in different blocks, we must have $|S_{ij}| \le \lambda$, so $\tilde{S}$ has the same block structure
  - If $\tilde{S}$ is block diagonal, then the stationarity condition is satisfied with $\Gamma_{ij} = 0$ for all i, j in different blocks, and Θ^{-1} block diagonal. Hence Θ has the same block structure

Therefore, we have shown that the block structure of the minimizer $\hat{\Theta}$ is exactly the same as the block structure of $\tilde{S}$, the thresholded sample covariance matrix. This makes the graphical lasso look very simple-minded, in a way!

7.6 Constrained and Lagrange forms

- Often in statistics and machine learning, we'll switch back and forth between the *constrained* form of a problem, where $t \in \mathbb{R}$ is a tuning parameter,

$$\min_x \; f(x) \;\; \text{subject to} \;\; h(x) \le t, \tag{C}$$

and the *Lagrange* form, where $\lambda \ge 0$ is a tuning parameter,

$$\min_x \; f(x) + \lambda \cdot h(x), \tag{L}$$

and claim these are equivalent. Is this true (assuming convex f, h)?

- (C) to (L): if problem (C) is strictly feasible, then strong duality holds, and there exists some $\lambda \ge 0$ (dual solution) such that any solution $x^\star$ in (C) minimizes

$$f(x) + \lambda \cdot (h(x) - t),$$

so $x^\star$ is also a solution in (L)

- (L) to (C): if $x^\star$ is a solution in (L), then the KKT conditions for (C) are satisfied by taking $t = h(x^\star)$, so $x^\star$ is a solution in (C)
- Conclusion:

$$\bigcup_{\lambda \geq 0} \{\text{solutions in (L)}\} \quad \subseteq \quad \bigcup_{t} \{\text{solutions in (C)}\}$$
$$\bigcup_{\lambda \geq 0} \{\text{solutions in (L)}\} \quad \supseteq \quad \bigcup_{\substack{t \text{ such that (C)} \\ \text{is strictly feasible}}} \{\text{solutions in (C)}\}$$

Strictly speaking this is not a perfect equivalence (albeit minor nonequivalence). Note: when the only value of t that leads to a feasible but not strictly feasible constraint set is $t = 0$, i.e.,

$$\{x : h(x) \leq t\} \neq \emptyset,\ \{x : h(x) < t\} = \emptyset \quad \Longrightarrow \quad t = 0$$

(e.g., this is true if g is a norm), then we do get perfect equivalence

7.7 Solving the primal via the dual

- Recall that under strong duality, given dual optimal $u^\star, v^\star$, any primal solution minimizes $L(x, u^\star, v^\star)$ over x (it satisfies the stationarity condition). In other words, any primal solution $x^\star$ solves

$$\min_x\ f(x) + \sum_{i=1}^m u_i^\star h_i(x) + \sum_{j=1}^r v_i^\star \ell_j(x).$$

Often, solutions of this unconstrained problem can be expressed explicitly, giving an explicit characterization of primal solutions from dual solutions

- As an exmaple, consider the lasso problem

$$\min_{\beta \in \mathbb{R}^p}\ \frac{1}{2}\|y - X\beta\|_2^2 + \lambda\|\beta\|_1.$$

Its dual function is just a constant (equal to $f^\star$). Hence we reparametrize the primal problem

$$\min_{\beta \in \mathbb{R}^p,\, z \in \mathbb{R}^n}\ \frac{1}{2}\|y - z\|_2^2 + \lambda\|\beta\|_1 \ \text{ subject to } \ z = X\beta,$$

so the dual function is now

$$\begin{aligned} g(u) &= \min_{\beta \in \mathbb{R}^p,\, z \in \mathbb{R}^n}\ \frac{1}{2}\|y - z\|_2^2 + \lambda\|\beta\|_1 + u^T(z - X\beta) \\ &= \frac{1}{2}\|y\|_2^2 + \min_{z \in \mathbb{R}^n}\ \Big(\frac{1}{2}\|z\|_2^2 - (y - u)^T z\Big) + \min_{\beta \in \mathbb{R}^p}\ \Big(\lambda\|\beta\|_1 - (X^T u)^T \beta\Big) \\ &= \frac{1}{2}\|y\|_2^2 - \frac{1}{2}\|y - u\|_2^2 - I\big\{\|X^T u\|_\infty \leq \lambda\big\}. \end{aligned}$$

Above, we used the fact that the minimum over β is $-\infty$ if $\|X^T u\|_\infty > \lambda$, and 0 otherwise. Therefore, the *lasso dual* is

$$\max_{u \in \mathbb{R}^n}\ \frac{1}{2}\Big(\|y\|_2^2 - \|y - u\|_2^2\Big) \ \text{ subject to } \ \|X^T u\|_\infty \leq \lambda,$$

or equivalently

$$\min_{u \in \mathbb{R}^n}\ \|y - u\|_2^2 \ \text{ subject to } \ \|X^T u\|_\infty \leq \lambda.$$

Strong duality holds here (Slater's condition), and given a dual solution $\hat{u}$, any lasso solution $\hat{\beta}$ satisfies (from the z block of the stationarity condition)

$$\hat{z} - y + \hat{\beta} = 0,$$

i.e.,

$$X\hat{\beta} = y - \hat{u},$$

So the lasso fit is just the dual residual. Looking back at the dual problem, we can express the dual solution as $\hat{u} = P_C(y)$, the projection of y onto the convex polyhedron

$$C = \{u : \|X^T u\|_\infty \leq \lambda\}.$$

Hence, the lasso fit is the residual from projecting y onto the convex polyhedron C. This is actually quite a fruitful perspective, and we can use it to establish several nontrivial properties of the lasso. See Figure 3 for a geometric picture

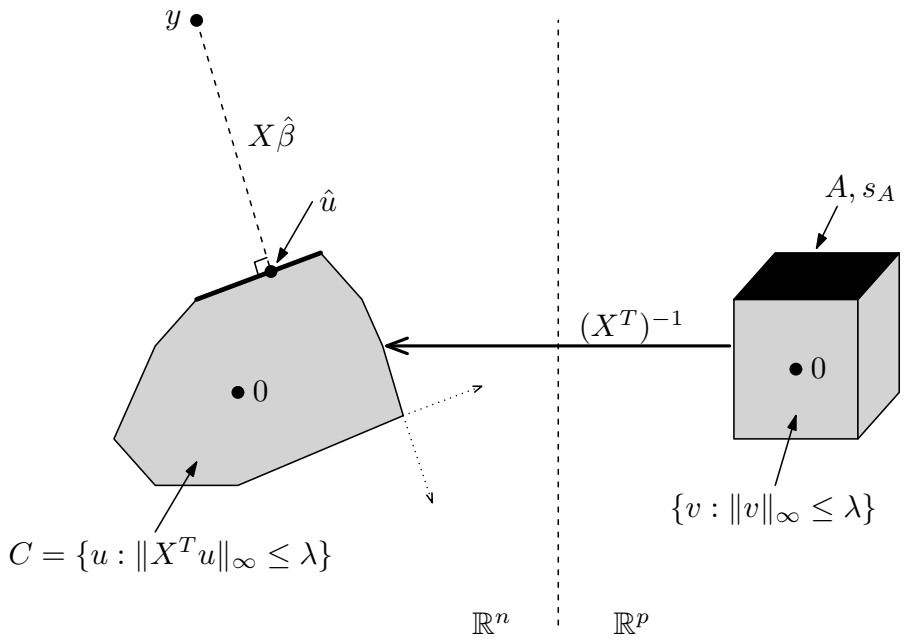

Figure 3: *An illustration of the primal-dual relationship for the lasso.*

References

Beck, A. & Teboulle, M. (2009), 'A fast iterative shrinkage-thresholding algorithm for linear inverse problems', *SIAM Journal on Imaging Sciences* **2**(1), 183–202.

Boyd, S. & Vandenberghe, L. (2004), *Convex Optimization*, Cambridge University Press, Cambridge.

Nesterov, Y. (1983), 'A method of solving a convex programming problem with convergence rate $O(1/k^2)$', *Soviet Mathematics Doklady* **27**(2), 372–376.

Rockafellar, R. T. (1970), *Convex Analysis*, Princeton University Press, Princeton.